Synthèse audiovisuelle de la parole expressive : modélisation des émotions par apprentissage profond

Sara Dahmani

UNIVERSITÉ DE LORRAINE

École doctorale IAEM Lorraine

Synthèse audiovisuelle de la parole expressive : modélisation des émotions par apprentissage profond

(Audiovisual synthesis of expressive speech : modeling of emotions with deep learning)

THÈSE

présentée et soutenue publiquement le 13 novembre 2020

pour l'obtention du

Doctorat de l'Université de Lorraine

(mention informatique)

par

Sara DAHMANI

Composition du jury

Président :	Mathieu CONSTANT, Professeur, Université de Lorraine
Rapporteurs :	Magalie OCHS, Maître de conférences, Aix-Marseille Université Damien LOLIVE, Maître de conférences, Université de Rennes
Examinateurs :	Catherine PELACHAUD, Directeur de recherche, Université Pierre-et-Marie-Curie Yannick ESTÈVE, Professeur, Université d'Avignon Anne BOYER, Professeur, Université de Lorraine
Directeurs de thèse :	Slim OUNI, Maître de conférences, Université de Lorraine Vincent COLOTTE, Maître de conférences, Université de Lorraine

Laboratoire Lorrain de Recherche en Informatique et ses Applications — UMR 7503

Mis en page avec la classe thesul.

Remerciements

Je tiens à adresser tout d'abord mes respectueux remerciements à Mme Magalie OCHS et M. Damien LOLIVE pour avoir accepté de rapporter cette thèse, puis à l'ensemble des membres du jury, Mme Catherine PELACHAUD, Mme Anne BOYER, M. Mathieu CONSTANT et M. Yannick ESTÈVE pour la qualité et le soin de leurs observations et leurs remarques.

J'adresse également mes vifs remerciements à mes deux directeurs de thèses Slim OUNI et Vincent COLOTTE qui m'ont encadré durant mon contrat d'ingénieure à l'INRIA Nancy, puis durant mes années de thèse. Merci pour votre confiance, votre patience et votre accompagnement.

Je tiens à remercier les membres de l'équipe Multispeech pour leur sympathie et tous les échanges intéressants que nous avons pu avoir durant ces longues années au LORIA.

Je veux remercier les deux acteurs, grâce à qui, nous avons pu enregistrer nos corpus de données et mener à bien ces travaux. J'en profite pour remercier toute personne qui a participé aux expériences perceptives que nous avons réalisées.

Je remercie mes amis Valérian, Karima, Amine, Pierre et Ameur pour leur encouragement et pour tous les moments que nous avons pu partager, les meilleures comme les plus dures.

Enfin, je remercie ma famille pour leur soutien sans faille dans tout ce que j'entreprends.

Sommaire

Table des figures

Introduction générale

Les humains ont la particularité de communiquer avec un langage et d'engager des conversations avec d'autres personnes. Cette capacité de communication englobe différents signaux émanant du corps humain comme les gestes des mains, la posture, les mouvements de la tête, le regard ou encore l'intonation de la voix pour appuyer ou clarifier certains propos. Lorsque nous regardons le visage de quelqu'un qui parle, nous recevons deux flux d'informations distincts : un signal acoustique constitué d'une série de sons et un signal visuel constitué des variations visibles du visage. Les informations acoustiques de la parole sont en réalité la conséquence du mouvement des articulateurs et de la circulation de l'air dans le système de production de la parole. Il faut donc accéder à ces deux modalités acoustique et visuelle pour avoir une représentation complète de la parole.

Aujourd'hui, la technologie s'intègre de plus en plus dans nos vies. Les écrans et les interfaces utilisateurs sont partout autour de nous et une grande partie des services sont automatisés. Dans cette tendance d'automatisation, des efforts considérables sont déployés pour développer des agents conversationnels virtuels. Pour remplir au mieux leur rôle, ces agents artificiels doivent être capables de communiquer au moyen de comportements verbaux et non verbaux (gestes et expressions faciales) [Cassell et al., 2000, Ruttkay et al., 2004, Mancini et al., 2017]. Les agents virtuels couvrent aujourd'hui un large éventail d'applications dans les domaines des télécommunications, de la robotique humanoïde, du multimédia, de la médecine, de l'éducation, des assistants virtuels et de l'industrie du divertissement (jeux vidéo, films, etc.) [Sproull et al., 1996, Pandzic et al., 1999, Dehn and Van Mulken, 2000, Ochs and Blache, 2016, Dworkin et al., 2018, Falconer et al., 2019, Beskow, 2019].

Si l'on veut animer un agent virtuel par l'intermédiaire d'une tête parlante par exemple, il est possible de se reposer sur des séquences préenregistrées d'audio ou de mouvements faciaux. Cependant, cette technique ne s'avère suffisante que dans le cas de scénarios de communication très basiques et limités. Lorsque l'échange avec l'agent virtuel devient plus complexe, il faut prévoir des fonctionnalités de génération de nouvelles séquences de parole. À cette fin, les techniques de synthèse de la parole prennent tout leur sens. Ces techniques permettent de générer de nouvelles séquences de parole à partir d'un texte et peuvent être conçues pour générer des sons (acoustique) tout comme produire des gestes articulatoires. La synthèse de la parole à partir du texte (TTS : Text-To-Speech) est un processus automatique de conversion d'un texte en un signal de parole et consiste généralement en deux phases : d'abord une phase d'analyse du texte pour générer la transcription symbolique de sa prononciation adéquate, puis d'une phase de synthèse où les symboles générés sont convertis en signaux de la parole. La TTS a commencé par la synthèse par diphones.

Le premier système de synthèse vocale date de 1791 avec "la machine parlante" [Kempelen, 1791]. Cette machine consiste en un système mécanique qui imite le système phonatoire humain. Ce système a connu de nombreuses améliorations durant le 19ème siècle. En arrivant à la première moitié du 20ème siècle, et avec l'avènement de la synthèse vocale à base de règles et de formants

[Dudley, 1939, Fant, 1953], la génération de la parole est passée des systèmes mécaniques à des systèmes électriques. Concernant la synthèse visuelle, les premiers systèmes sont apparus vers les années 90s avec des méthodes basées sur l'interpolation des images clés et d'autres basées sur des règles articulatoires [Massaro and Cohen, 1990, Beskow, 1995]. Les recherches entreprises ces dernières années en informatique ont permis d'améliorer sensiblement la qualité de la parole synthétique.

Cependant, le but ultime d'un système de TTS est de générer une parole synthétique imperceptible de celle que pourrait prononcer une personne. Bien que le problème de la synthèse vocale de bonne qualité a été considéré comme résolu suite à la synthèse par concaténation. Récemment, et grâce aux progrès dus aux techniques de l'apprentissage profond, la qualité de la synthèse vocale par réseaux de neurones a gagné énormément de naturel et a même dépassé la qualité des systèmes par concaténation [Oord et al., 2016]. Or, les résultats manquent encore de réalisme à la fois dans l'expressivité de la parole que dans la cohérence entre les modalités acoustique et visuelle générées, et cette problématique reste toujours ouverte. La dépendance/complémentarité entre la modalité acoustique et visuelle a été démontrée également pour des visages synthétiques [McGurk and MacDonald, 1976, Andersen, 2010, Fu et al., 2007, 2008], mais la synthèse de l'aspect visuel de la parole n'a toujours pas atteint un niveau de naturel suffisant pour être intégrée dans les supports de communication à une échelle aussi large que celle de la synthèse vocale. En réalité, la modélisation du visage humain doit être sans faute. Comme l'explique l'artiste Faigin [2012] dans son livre "The Artist's Complete Guide to Facial Expression" :

> *« Il n'y a pas de paysage que nous connaissions aussi bien que le visage humain. Les vingt-cinq pouces carrés sont le territoire le plus minutieusement examiné qui existe, examiné constamment et soigneusement, avec bien plus qu'un intérêt intellectuel. Chaque détail du nez, des yeux et de la bouche, chaque régularité en proportion, chaque variation d'un individu à l'autre, sont des sujets sur lesquels nous sommes tous maîtres. »*

Souvent, les tentatives de synthèse visuelle réaliste se heurtent à la théorie de la Vallée dérangeante *(Uncanny Valley)* [Mori et al., 1970] qui stipule que plus un agent virtuel est proche d'un aspect humain, plus ses imperfections nous paraissent monstrueuses. Aussi, la plupart des travaux de modélisation des agents virtuels réalistes s'intéressent à la création des agents virtuels dans un contexte statique en négligeant l'aspect plus complexe et dynamique relatif à la parole et à la synchronisation labiale.

À toutes les difficultés que rencontrent la TTS audiovisuelle s'ajoute la volonté d'augmenter les agents conversationnels avec une dimension expressive. En effet, les modalités acoustique et visuelle peuvent être suffisantes pour transmettre un message clair et intelligible, cependant, l'expression des émotions est d'une importante majeure pour réguler les interactions et remplir des fonctions sociales importantes comme la communication des convictions, des désirs et des intentions aux autres [Frijda and Mesquita, 1994, Keltner and Haidt, 1999, Keltner and Kring, 1998, Morris and Keltner, 1999].

Plusieurs travaux ont intégré la dimension expressive dans l'animation des agents virtuels. Il a été établi que les agents virtuels expressifs, dans une interaction humain-machine, sont jugés plus naturels, plus crédibles et plus réalistes par rapport aux agents virtuels non expressifs [Bates et al., 1994]. D'autres expériences ont démontré un changement du comportement des utilisateurs face à des agents virtuels expressifs que face à des agents non-expressifs. Les utilisateurs deviennent plus engagés et les émotions exprimées par l'agent virtuel ont même une influence sur leur prise de décision [Walker et al., 1994, de Melo et al., 2012].

Dans le souhait de créer des agents virtuels expressifs, il faut connaître la contribution de chaque modalité de la parole dans la communication des sentiments. Le travail de Mehrabian [1968] montre l'importance majeure de la modalité non-verbale dans cette communication. En effet, une formule précise a été proposée, dans cette étude, pour quantifier la contribution de chaque composante de la communication. 7% de l'information sur notre état affectif est communiquée par les mots, 38% est communiquée par la voix et 55% par les expressions faciales et le langage corporel. Cette étude est très importante et montre que les mots ne suffisent pas pour communiquer ses sentiments. De ce fait, pour avoir une modélisation complète des émotions, il est crucial d'inclure la modalité visuelle dans les systèmes et les supports de communication.

Les premières tentatives de TTS vocale et visuelle expressive sont des systèmes à base de règles augmentées avec des modules spécifiques aux émotions. Ensuite, ont suivi les techniques de TTS par diphones puis par concaténation, en passant par les HMMs (Hidden Markov Models) jusqu'aux systèmes à base de réseaux de neurones largement adoptés de nos jours. Comme pour la TTS visuelle non-expressive, les études concernant l'expressivité faciale, négligent souvent l'aspect dynamique relatif de l'articulation. En raison de la coopération complexe des activités musculaires du visage, les mouvements du bas du visage sont contrôlés conjointement par les gestes liés à la parole et aux expressions faciales. Bailly et al. [2008] ont montré que les émotions étudiées (joie et dégoût), perturbent considérablement les mouvements de certains articulateurs (lèvres et mâchoire inférieure) durant la parole. Les expressions faciales imposent un effort de compensation et de réorganisation de l'articulation. En réalité, les stratégies articulatoires mises en œuvre par les locuteurs pour faire face à la production concomitante de la parole et des expressions faciales entraînent parfois à la résolution d'instructions contradictoires, par exemple l'étirement des lèvres pour sourire lors de la production de voyelles ou de consonnes arrondies. Cet article [Bailly et al., 2008] conclu que ces perturbations ne sont pas simplement additives, et qu'elles dépendent de l'articulation. De ce fait, l'ajout de certaines expressions spécifiques à une émotion à des données visuelles neutres n'est pas suffisant pour modéliser correctement la parole visuelle dans un contexte expressive. Fónagy [1976] et Nordstrand et al. [2004] ont aussi étudié les effets de l'expressivité sur les mouvements articulatoires, et ont trouvé que les effets des émotions sur l'articulation différent d'une émotion à l'autre. Vu la nature très dynamique des gestes de la parole et des expressions faciales, les interactions et contributions exactes de ces deux sources sont inconnues. À cette fin, les systèmes à base de réseaux de neurones sont de plus en plus adoptés pour apprendre cette relation complexe et non linéaire entre l'articulation et l'expressivité. Cependant, les recherches actuelles n'ont pas encore atteint le stade où les agents virtuels parlants sont capables d'exprimer les émotions de manière aussi naturelle que l'être humain.

Si la qualité de la TTS expressive vocale a pu connaître aujourd'hui plusieurs améliorations, ce n'est pas uniquement grâce aux techniques utilisées mais surtout grâce aux ressources largement disponibles. Malheureusement, ce n'est pas le cas pour la synthèse visuelle. Les bases de de données expressives existantes ne contiennent souvent que la modalité acoustique (SynPaFlex, AlloSat, PAVOQUE, etc). Pour les bases de données visuelles, dans leur majorité, la modalité visuelle est sous forme d'enregistrements vidéo 2D (GEMEP, CVSP-EAV, eNTERFACE'05, etc). Bien qu'ils soient faciles et moins coûteux à construire, dans ces corpus, l'information sur la profondeur de la scène est perdue. De ce fait, certains gestes liés à la parole, comme la protrusion des lèvres, ne peuvent pas êtres suivis/prédits avec précision. Heureusement, quelques bases de données audiovisuelles expressives contenant des données 3D existent. Toutefois, elles ne contiennent généralement qu'une seule émotion (AV-LASYN le rire), ou alors contiennent une quantité très faible de données (IEMOCAP 30 minutes de parole, toutes émotions confondues ou alors Biwi 3D avec environ 80 phrases par locuteur) ce qui est insuffisant pour entraîner des systèmes de

synthèse de la parole à base de réseaux de neurones par exemple. L'autre problématique avec ces corpus est que le nombre de phrases par émotion, n'est pas équilibré (IEMOCAP neutre 28%, frustration 24%, excitation 17%, tristesse 15%, colère 7%, joie 7%, surprise 2%, dégoût 1%, les autres $< 1\%$) ce qui ne permet pas de comparer la performance des systèmes de synthèse pour les différentes classes d'émotions. De ce fait, et avant de se lancer dans la mise en place d'un système de TTS visuel, nous avons commencé par la construction et l'étude de deux corpus audiovisuels expressifs pour répondre à toutes ces exigences.

Par ailleurs, une piste prometteuse aujourd'hui est celles des techniques non-supervisées capables d'apprendre sur des corpus non-annotés. En réalité, les corpus annotés ne sont pas toujours disponibles et quand ils sont disponibles, ils peuvent contenir des annotations peu fiables qui risquent de nuire à l'apprentissage. Ce genre de techniques permet de profiter d'une quantité plus large de données sans se soucier de la tâche laborieuse de son étiquetage. Dans ce travail de thèse, nous adoptons une technique non-supervisée pour entraîner des modèles neuronaux avec des données sans étiquette d'émotions, et cela pour modéliser les émotions dans un contexte acoustique et visuel.

Toutes ces techniques permettent d'améliorer la qualité du signal produit en profitant de bases de données plus larges. Or, la qualité du signal n'est pas à elle seule garante du réalisme et du naturel de la parole de synthèse puisqu'elle n'est pas suffisante pour couvrir toute la diversité de la parole humaine. La théorie des émotions par catégorie postule que le système affectif se compose de six émotions universelles basiques (bonheur, surprise, peur, tristesse, colère et dégoût) [Ekman, 1992]. Pourtant, la diversité des émotions humaines peut générer de nombreux états affectifs complexes et subtils tels que la désapprobation, la dépression et le mépris qui ne peuvent pas être couverts par ces catégories d'émotions de base. De plus, certaines recherches confirment que les états affectifs ne sont pas des entités isolées, mais sont plutôt systématiquement connectés [Russell, 1980, Plutchik, 1984, Laresn and Diener, 1992]. Par conséquent, les modèles dimensionnels considèrent l'expérience affective comme un continuum d'états non extrêmes et fortement interconnectés, similaire au spectre de couleur [Posner et al., 2005, Russell and Fehr, 1994]. De ce fait, pour avoir un résultat naturel et réaliste, il est crucial de pouvoir modéliser des mélanges d'émotions et différentes émotions avec des intensités contrôlables. Une des contributions majeures de cette thèse est la proposition d'une méthode non-supervisée de modélisation des émotions qui permet de modéliser différentes émotions avec des intensités contrôlables ainsi que de modéliser des mélanges d'émotions. Nous avons utilisé les données synthétiques prédites avec nos modèles pour animer un agent virtuel 3D expressif.

Les travaux de thèse présentés dans ce document ont pour objet d'étudier et de modéliser les émotions dans un contexte audiovisuel. La modélisation des émotions doit nous permettre, à la fois, de générer une parole expressive, intelligible et naturelle et d'un autre côté nous permettre d'avoir un contrôle sur le spectre des émotions pour pouvoir les nuancer et les mélanger afin d'imiter la complexité du système émotionnel humain.

Organisation du document

Pour présenter les travaux réalisés dans le cadre de cette thèse, ce document s'articule autour de trois parties. Tout d'abord, nous présentons un état de l'art concernant la TTS, et plus spécifiquement la TTS audiovisuelle et expressive. La seconde partie est réservée à la présentation des protocoles d'acquisition des données expressives et les corpus de données enregistrés. Enfin, la dernière partie présente les architectures de réseaux de neurones utilisées ainsi que les expériences menées pour répondre à la problématique.

Dans la première partie, intitulée « Analyse de l'état de l'art », nous présentons le contexte

scientifique de ce travail de thèse. Nous commençons par le chapitre 1 (« La parole ») qui introduit les principes de traitement automatique de la parole et sa transformation en paramètres pertinents pour la synthèse de la parole. Le chapitre 2 (« Synthèse de la parole audiovisuelle ») introduit les différentes techniques de TTS que ça soit pour la synthèse acoustique ou visuelle. Le chapitre 3 (« Synthèse expressive de la parole ») se focalise sur les techniques de synthèse expressive de la parole et permet de situer notre travail de thèse par rapport aux travaux de la littérature.

Dans la deuxième partie, intitulée « Corpus audiovisuels expressifs » nous présentons deux corpus audiovisuels expressifs de la parole que nous avons enregistrés auprès de deux acteurs de théâtre. Le chapitre 4 (« Étude d'un corpus expressif ») présente la démarche de collecte d'un corpus prototype. Les analyses objectives et perceptives effectuées sur ce corpus nous ont permis de valider notre protocole d'acquisition des données audiovisuelles expressives. Dans le chapitre 5 (« Acquisition d'un corpus expressif pour la synthèse de la parole »), et en reposant sur les conclusions du chapitre précédent, nous enregistrons et étudions le contenu expressif d'un corpus de taille plus importante qui sera dédié à la TTS audiovisuelle expressive.

Dans la dernière partie, intitulée « Synthèse audiovisuelle expressive de la parole », nous présentons deux architectures neuronales pour la TTS audiovisuelle expressive de la parole. Dans le chapitre 6 (« Synthèse par architecture entièrement connectée ») nous présentons une architecture entièrement connectée qui nous a permis de valider certains choix de paramètres et de configurations lors de l'entraînement des réseaux de neurones. Dans le chapitre 7 (« Synthèse par CVAE ») nous présentons la seconde architecture neuronale reposant sur un Auto-Encodeur Variationnel. Cette seconde architecture nous a permis de passer d'une modélisation discrète des émotions à une représentation continue et complètement malléable.

Pour conclure ce document, nous dressons une synthèse des résultats du travail réalisé, et proposons quelques perspectives de recherches.

Première partie

Analyse de l'état de l'art

1

La parole

Sommaire

1.1 Introduction

La communication humaine est à la fois sociale et cognitive. C'est un processus par lequel les individus échangent des informations à travers un système commun de codes et de signes. La communication humaine repose sur des intentions fondamentalement liées à l'aide et au partage [Tomasello, 2010]. Les humains communiquent pour demander de l'aide, pour informer les autres de choses utiles et pour partager des attitudes comme moyen de s'intégrer au sein d'un groupe d'individus. La communication est également cognitive car elle permet d'échanger du savoir et de la connaissance à propos de structures cognitives abstraites ou symboliques.

Par ailleurs, tout animal communique avec ses semblables, cette communication peut prendre plusieurs formes : par des signaux visuels, acoustiques, chimiques, tactiles ou même électriques chez certains animaux. Les éléphants font usage des d'infrasons pour communiquer entre eux à de grandes distances. Certaines races de singes, et de corbeaux se servent de cris spécifiques pour signaler la présence de danger en utilisant différents codes pour différents prédateurs. Les fourmis utilisent des phéromones comme outils de repérage de pistes olfactives destinées à guider les fourmis vers des sources de nourriture. Les bonobos font aussi grand usage d'une gestuelle subtile pour communiquer, nous pouvons trouver plein d'autres exemples dans la nature.

Bien que, comme les exemples l'illustrent, les animaux soient en capacité de communiquer, cette communication reste limitée et ne porte que sur des structures cognitives communes ou sur des expériences partagées. Les animaux sont dans l'incapacité de distribuer des informations abstraites ou nécessitant de nouvelles structures cognitives [Puppel and Puppel, 1995]. La communication humaine, en revanche, est plus complexe, elle permet de communiquer des situations

présentes, passées et futures. Elle permet également d'échanger à propos de choses absentes ou abstraites.

Le principale mode de communication des humains est la parole, en fait, la capacité des humains à parler, est essentiellement anatomique [Gärdenfors, 2006]. Chez les humains, le larynx est placé dans une position plus basse permettant ainsi la production de sons plus variés. Aussi, le cerveau humain, notamment le lobe frontal, a des connections neuronales adaptées pour gérer et coordonner des mouvements très rapides et complexes nécessaires pour créer un débit fluide de la parole.

Dans ce chapitre, nous présentons différentes facettes de la communication humaine par la parole. Nous détaillons quelques concepts essentiels liés à la production de la parole et leurs caractéristiques respectives. Nous allons aussi exposer les moyens de représentation numérique de ces facettes de communication dans un contexte de traitement automatique de la parole et de l'apprentissage profond.

1.2 Qu'est-ce que la parole ?

La communication humaine est composée d'une partie verbale et d'une autre non-verbale. La partie verbale de la communication est appelée la parole. La partie non-verbale concerne généralement les aspects visibles comme les expressions faciales, le regard, les mouvements de la tête, la posture et les gestes corporels ainsi que d'autres signes (conscients ou inconscients).

Les sons produits par l'humain durant la communication verbale, ne sont pas aléatoires. En fait, ces sons suivent un certain nombre de règles imposées par le langage utilisé. Ces règles cadrent l'enchaînement des sons afin de former des unités linguistiques (les mots) porteuses de sens, et s'organisant selon une grammaire propre. Ces règles peuvent être connues oralement au sein d'un groupe d'individus pour former un dialecte. Elles peuvent également évoluer et être standardisées pour former un **langage**. Les langages existent aussi sous une forme écrite. Ils ont un système de transcription stricte qui permet de les préserver et de les transmettre. Chaque langue a des sons propres à elle, appelés **les phonèmes**. L'étude et la classification des phonèmes relève du domaine de la phonétique.

Le mot parole (« speech » en anglais) est défini dans la langue française comme étant : « la faculté d'exprimer et de communiquer la pensée au moyen du système des sons du langage articulé émis par les organes phonateurs » [1]. Cependant, et par abus de langage, le mot « parole » est souvent utilisé dans le monde scientifique comme équivalent au mot « communication ». Nous pouvons donc distinguer la **« parole acoustique »** [Chibelushi et al., 1993] de la **«parole visuelle»** [Cohen and Massaro, 1993] ou encore les combiner dans la **«parole audiovisuelle»** [Dupont and Luettin, 2000]. Nous avons choisi, dans ce manuscrit, de garder cette même métonymie afin d'être alignés avec les termes utilisés dans la littérature et dans les travaux que nous allons citer et discuter.

La parole audiovisuelle est une série complexe de sons et de mouvements qui modulent le ton de la voix pour créer un signal audiovisuelle intelligible. La parole est produite par une coordination précise entre les muscles de la tête, le cou, la poitrine et l'abdomen. Le développement de la parole est un processus graduel qui requiert des années d'entraînement pour réussir à réguler ses muscles pour produire une parole compréhensible.

1. Le Centre National de Ressources Textuelles et Lexicales : https ://www.cnrtl.fr

1.3 La parole acoustique

La parole acoustique est un signal. Ce signal, émit de la bouche du locuteur, représente les changements dans le temps de la pression de l'air.

Le fonctionnement de l'appareil vocal humain a quelques similarités avec certains instruments de musique. Par exemple, pour la guitare, les mouvements des cordes génèrent des vibrations de l'air. À chaque fois que l'on pince une corde, cette dernière émet un son, mais ce son est à peine audible car le diamètre de la corde est trop petit pour produire un son audible. C'est à ce niveau là qu'intervient la caisse de résonance de la guitare qui est une cavité dans le corps de la guitare. Le rôle de la caisse de résonance est d'amplifier les vibrations de l'air générées par le pincement des cordes pour produire un son audible et de ce fait générer le signal musical final.

Le système vocal humain est aussi composé de cordes vocales et de caisses de résonances. Toutefois, les vibrations des cordes vocales humaines ne se font pas à travers des pincements. Alors, comment est-il possible de créer des vibrations ?

La figure 1.1 reprend les différents organes impliqués dans le système de phonation de l'être humain. Le corps humain compte deux cordes vocales situées dans le larynx juste au-dessus des poumons. Pour parler, nous pompons de l'air des poumons vers le larynx. Ce flux d'air exerce une pression sur les cordes vocales qui entraîne une vibration de ces dernières. Elles effectuent ainsi un travail de « battement ». C'est cette discontinuité dans la colonne d'air qui génère une onde sonore, nous parlons ainsi de **source du signal acoustique**. La fréquence de battement des cordes vocales par seconde détermine **la fréquence fondamentale de la parole**. La fréquence fondamentale est notée **F0** et son unité est le **hertz (Hz)**. Les variations de la fréquence fondamentale dans le temps constituent **l'intonation** de la voix. Il est essentiel de noter que tous les sons de la parole ne sont pas produits au niveau de la source (vibration des cordes vocales), mais peuvent résulter des frictions en un point du conduit vocal pour former les consonnes fricatives ([f], [v], [s]...), ou par un blocage-relâchement de l'air produisant les consonnes occlusives ([p], [t], [k]...). Dans ce cas, ces sons sont appelés des **sons non-voisés**. Les sons produits aux niveaux des cordes vocales quant à eux s'appellent les **sons voisés**.

En revanche, nos poumons ne se vident pas d'un seul coup comme pour un ballon gonflable, et heureusement, car la grande pression risque d'abîmer grièvement nos cordes vocales. Le corps humain est doté d'un certain nombre de nerfs et de muscles qui permettent de gérer inconsciemment et avec une grande précision le flux d'air des poumons. Avant de parler, nous estimons la longueur de la phrase que nous souhaitons prononcer et nous inspirons la quantité d'air nécessaire pour la produire. À l'aide des muscles impliqués dans la respiration nous pouvons donc retenir notre souffle et surtout de gérer **le débit** de la parole en jouant sur **les durées** des sons que nous émettons.

Les vibrations créées au niveau des cordes vocales se dirigent vers les caisses de résonances du système phonatoire. Le corps humain compte trois caisses de résonances comme montré dans la figure 1.1 : Le pharynx (arrière gorge), la cavité buccale et la cavité nasale. Les caisses de résonances vont alors amplifier les vibrations pour créer du son audible. Chaque cavité résonne à une fréquence différente. En effet, la taille et la forme des cavités permettent de rendre certaines fréquences plus proéminentes que d'autres. Ainsi, les résonateurs agissent comme des **filtres de fréquences** qui amplifient certaines fréquences tout en affaiblissant d'autres.

Lorsque nous parlons, nous modifions successivement la forme des cavités vocales pour produire les différents sons de la parole. Nous sommes aussi capables de choisir d'inclure ou pas certaines cavités dans l'émission d'un son (cavité nasale par exemple). C'est grâce à cette grande flexibilité du larynx et de la cavité buccale (forme et position de la langue et des lèvres par exemple) que nous pouvons produire une large variété de fréquences. Ces différentes fréquences

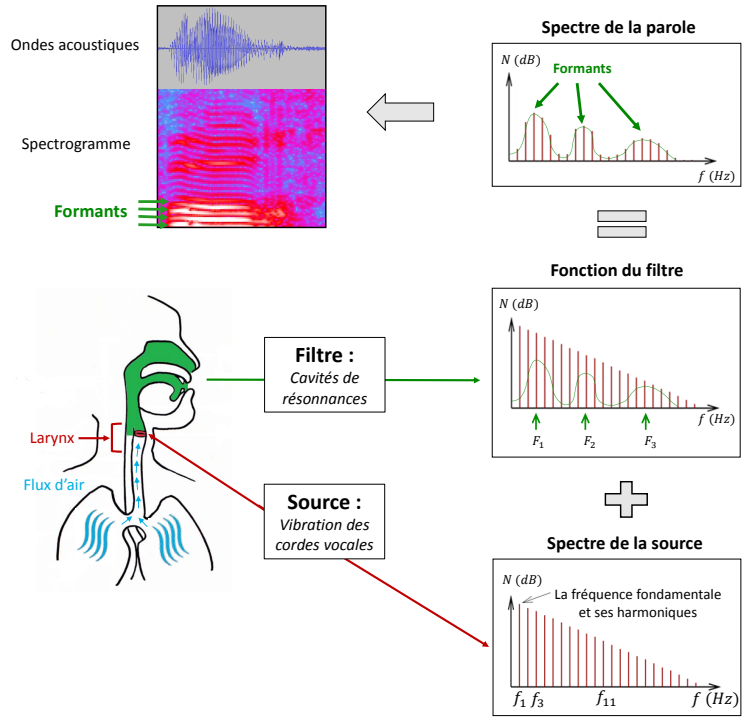


FIGURE 1.1 – *Les mécanismes de production de la voix humaine.*

sont notées **F1, F2, F3, etc**. Elles sont appelées **les formants**. Par exemple, pour le phonème français [u], F1 se situe aux alentours de 300 Hz, F2 à environ 800 Hz et F3 est proche de 2000 Hz [Calliope and Fant, 1989, Georgeton et al., 2012].

Dans un son pur, le signal a une forme sinusoïdale, il a une fréquence et une amplitude constantes. Mais dans la nature un son avec de telles caractéristiques ne peut pas exister, sauf à l'aide d'un générateur électronique. Contrairement à un son pur, le signal que nous obtenons à partir de la parole peut être composé de plusieurs "sous-signaux" sinusoïdaux d'amplitudes différentes et qui oscillent à des fréquences différentes. Ces fréquences sont des multiples de la fréquence fondamentale F0 et sont appelées **des harmoniques** (la F0 étant la première harmonique du signal). Les harmoniques ont des amplitudes différentes puisque ces amplitudes ont été modulées par les cavités de résonances pour former les formants. C'est le physicien Joseph Fourier qui a découvert qu'un son non sinusoïdal, soit un son complexe, pouvait être décomposé en harmoniques [Bracewell and Bracewell, 1986]. Il est donc possible d'estimer **le spectre de fréquences** relatif au signal de la parole.

Dans le domaine du traitement et de la synthèse de la parole acoustique, un certain nombre de paramètres sont généralement utilisés pour représenter le signal. L'extraction de ces paramètres repose sur le passage d'une représentation temporelle du signal acoustique à sa représentation dans le domaine de fréquences. La transformation de fourrier peut identifier la fréquence fondamentale F0 et ses harmoniques. La F0 est généralement considérée avec une échelle logarithmique **log(F0)**.

Le deuxième paramètre très largement utilisé en traitement de la parole sont les coefficients **MFCC** *(Mel-Frequency Cepstral Coefficients)* [Davis and Mermelstein, 1980]. Ils sont obtenus

par filtrage du spectre d'amplitudes en suivant l'échelle de Mel puis en passant dans le domaine cepstral en appliquant une DCT *(Discret Cosinus Transform)*. Cette série de transformations a pour but de compresser le signal en un nombre réduit de coefficients qui contiennent des informations pertinentes sur le timbre du son. D'autres représentations du spectre de la parole ont été adoptés dans la littérature, notamment les LSPs (Line-spectral coefficients) [Itakura, 1975] et les coefficients MGCs *(Mel-Generalized Cepstral)* [Tokuda et al., 1994].

Un paramètre binaire est aussi utilisé qui concerne la nature voisée ou non-voisée du son. Son calcul se base généralement sur la présence/absence de la F0 dans la fenêtre du signal analysée. En réalité, même pour les sons voisés, les vibrations des cordes vocales ne sont pas parfaitement périodiques. De ce fait, la quantité de dévoisement, qui se manifeste notamment dans les hautes fréquences, est décrite par l'apériodicité du signal [Fujimura, 1968]. Inclure le paramètre de l'apériodité moyenne pour chaque bande de fréquence dans un processus de synthèse augmente significativement la qualité du son généré [Yoshimura et al., 2001]. Ce paramètre est appelé **BAP** *(band aperiodicity)*.

Bien que l'utilisation de ces paramètres statiques (log(F0), MFCC et BAP, voisé/non-voisé) soit primordiale dans la génération de la parole, ils ne sont pas suffisants pour capturer la dynamique de la parole. Sans caractéristiques dynamiques, la discontinuité des spectres provoque des sauts dans la parole synthétisée. D'autre part, avec des fonctionnalités dynamiques, la synthèse vocale devient plus lisse et naturelle [Masuko et al., 1996]. Les paramètres dynamiques sont les dérivées temporelles des paramètres statiques, ils permettent de mesurer les changements dans le spectre de la parole sur plusieurs frames pour modéliser les caractéristiques à long-terme du signal. Ils ont prouvé leur utilité dans les systèmes HMMs et les DNNs de type *Feed Forward* (voir chapitre 2) en permettant de générer des trajectoires acoustiques plus lisses.

L'outil qui permet d'obtenir les paramètres statiques et dynamiques d'un signal acoustique est appelé **le vocodeur** [Dudley, 1939]. Le vocodeur permet aussi de reconstruire un signal acoustique à partir de cet ensemble de paramètres, on parle ainsi de signal acoustique **re-synthétisé** ou son de re-synthèse. Les vocodeurs les plus utilisés de nos jours sont STRAIGHT [Kawahara et al., 1999] et WORLD [Morise et al., 2016]. La dernière version de WORLD génère un signal de meilleure qualité que STRAIGHT [Morise and Watanabe, 2018] de plus WORLD est un logiciel libre et son code source est ouvert contrairement à STRAIGHT qui est sous une licence commerciale.

Dernièrement les vocodeurs traditionnels sont de plus en plus remplacés par des vocodeurs basés sur des réseaux de neurones (WaveNet). La qualité de la parole par ce type de vocodeurs a gagné énormément de naturel et a même dépassé la qualité des systèmes par concaténation [Oord et al., 2016].

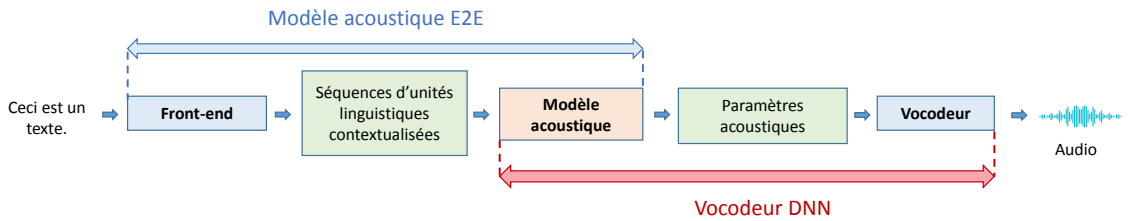


FIGURE 1.2 – *La pipeline complète de la synthèse acoustique de la parole. Les modèles acoustiques E2E et les vocodeurs basés sur des DNNs y sont situés par rapport à leurs entrées et sorties.*

Quelques étapes du processus de la synthèse sont fusionnés pour s'approcher de plus en plus

d'un système de synthèse de bout en bout (*end-to-end* ou E2E) comme représenté dans la figure 1.2. Le modèle acoustique et le vocoder sont combinés pour former un seul système convertissant les paramètres linguistiques en un signal acoustique directement. Les modèles acoustiques tel que Tacotron [Wang et al., 2017], permettent de générer des paramètres acoustiques à partir du texte brute ou des graphèmes éliminant ainsi le besoin d'utiliser un front-end, mais ces techniques nécessitent une importante quantité de données.

1.4 La parole visuelle

La parole visuelle est la manifestation visible des mouvements des articulateurs impliqués dans la production des sons. Comme les informations acoustiques de la parole sont en réalité la conséquence du mouvement des articulateurs du système de production de la parole, le mouvement des articulateurs est fortement corrélée à la série des sons prononcés.

Comme expliqué dans la section 1.3, le système vocale humain est constitué d'organes de phonation qui génèrent les vibrations nécessaires pour la parole (poumons et larynx) et des articulateurs (vélum, langue, lèvres, mâchoire) qui modulent la résonance pour former les sons de la parole. Quelques parties du système de production de la parole sont constamment visibles (lèvres et placement de la mâchoire) et d'autres peuvent l'être dans certains contextes (langue, dents).

Contrairement à l'aspect acoustique qui semble avoir un consensus sur les techniques d'acquisition et de représentation des données, les techniques dédiées aux données visuelles quant à elles sont beaucoup plus diversifiées. Les données visuelles sont soit sous forme d'enregistrements vidéo 2D soit sous forme de données spatiales 3D. L'acquisition de données visuelles peut se faire à l'aide d'un :

- Algorithme de traitement d'image et de suivi de mouvements faciaux, car il est aussi possible d'extraire des informations visuelles de la parole à partir des vidéos 2D sous forme de positions et de déplacements de points 2D dans le temps en utilisant ce genre d'algorithmes.

- Scanner 3D qui génère une surface complète du visage du locuteur plusieurs fois par seconde [Fanelli et al., 2010]. Si la qualité du scanner le permet, ces systèmes sont capables de capturer des détails très fins du visage. Toutefois, les données sont très volumineuses et le nombre d'images par seconde reste très faible (environ 25 fps) ce qui n'est pas suffisant pour capturer la parole avec fluidité.

- Système de capture de mouvement (optique ou magnétique). Un certain nombre de capteurs sont collés sur le visage du locuteur et les mouvements de ces capteurs sont récupérés sous forme d'évolution des trajectoires de points 3D dans le temps. Certes, ces systèmes couvrent moins de surface faciale mais offrent une fréquence d'échantillonnage plus élevée (environ 100 fps).

- Système de capture de mouvement sans capteurs. Ces systèmes se basent sur des algorithmes de suivit de mouvements faciaux et sur des caméras de profondeur pour capturer les informations dans un environnement 3D.

1.4.1 Données 2D

Les vidéos 2D sont en réalité des séquences d'images, la fréquence d'échantillonnage d'une vidéo est le nombre d'images par seconde qu'elle contient. Chaque image est composée d'un certain nombre de pixel. Plus une image en contient, plus grande est sa définition.

Des techniques de traitement d'images sont largement utilisées pour extraire les changements de couleurs des différents pixels en fonction des sons prononcés. L'un des algorithmes les plus utilisés est le AAM (Active Appearance Model) [Cootes et al., 2001]. Ce modèle permet de

localiser des objets non-rigides qui ont une grande variabilité (le visage par exemple). Il permet de suivre un nombre de points virtuels sur le visage et d'en extraire des paramètres relatifs à leurs déplacement. La position des points virtuels est standardisée par le protocole MPEG-4 [Pandzic and Forchheimer, 2002] qui découpe les mouvements du visage en un ensemble de 68 unités de mouvements appelées **les FAPs** *(Facial Action Parameters)* à partir du mouvement de 84 points bien définis. Chacune des unités de mouvements est liée à une action de déformation du visage vers un état éloigné de son état au repos.

Les données extraites des vidéos 2D manquent de précision en ce qui concerne la profondeur de la scène, toutefois cette information est cruciale pour modéliser les mouvements de reculement/protrusion des lèvres (mouvements largement utilisés dans la langue française) comme pour prononcer le son "ou". Les données des vidéos 2D sont majoritairement utilisées dans un contexte d'animation basée sur l'exemple, où plusieurs segments préenregistrés sont regroupés et collés ensemble pour former une nouvelle vidéo. En fait, la génération d'animations à partir de données de vidéos 2D souffre de plusieurs limitations. D'abord la sensibilité aux changements des conditions de l'environnement comme le changement de la lumière et de l'angle de la caméra/locuteur pendant l'enregistrement, et aussi la difficulté de l'estimation des mouvements de la tête dans les données enregistrées. Cependant, dans des conditions idéales, les résultats des animations 2D sont photo-réalistes puisqu'ils se basent sur des images et des textures réelles.

Les modèles 3D, quant à eux, sont plus flexibles puisque l'agent virtuel 3D peut interagir avec son environnement et changer de pose ou de conditions lumineuses facilement. De nos jours, la barrière entre l'animation 2D et 3D devient de plus en plus floue, puisque certains systèmes actuels projettent les transformations faciales des vidéos sur des modèles 3D intermédiaires. Ces derniers permettent de calculer les déformations avec plus de réalisme avant de les re-projeter dans le résultat final de la vidéo en 2D. Le modèle 3D intermédiaire est généralement estimé à partir de quelques images ou de extraits vidéos très courts d'une même personne [Wang et al., 2011].

1.4.2 Données 3D

Afin de créer des modèles 3D à animer, les systèmes de modélisation graphique représentent la surface faciale comme un ensemble de points 3D (nommés vertex). En connectant les différents points entre eux, nous obtenons une approximation de la surface faciale à modéliser. Pour créer une animation 3D, un ensemble de règles sont utilisées pour déformer le modèle 3D d'une manière contrôlée. Plusieurs approches de l'animation faciale ont été présentées dans la littérature.

Les données récupérées peuvent être utilisées à leur état brut pour animer un modèle 3D. C'est le cas des systèmes d'animation par algorithmes de déformation de forme libre [Kalra et al., 1992, Noh et al., 2000, Rhee et al., 2011]. Ces méthodes considèrent les données 3D capturées comme un maillage de basse résolution qui guidera la déformation du maillage de haute résolution qui est le modèle 3D à animer. L'idée est de projeter les données 3D sur les vertex les plus proches du modèle et de définir des règles précises pour ces points de contrôle gérant la force et la direction de leur mouvement. Bien que cette méthode soit rapide est adaptée pour le temps réel, elle nécessite une configuration rigoureuse et ne convient pas pour intégrer des contraintes liées à la structure osseuse et musculaire du visage humain ainsi que les détails subtiles (par exemple les renflements et ridules de la peau) nécessaires pour un résultat réaliste.

Le modèle musculaire proposé par Waters [1987], décrit des champs de déformations de la peau délimités par des actions musculaires. Cette approche modélise la structure anatomique du visage et sa dynamique sous-jacente par simulation des effets des interactions entre les muscles et en respectant la physique de la déformation des tissus. Il faut d'abord créer un modèle 3D de

chair et de muscles de haute résolution et anatomiquement précis, conçu pour un sujet spécifique. Ensuite, il faut traduire les trajectoires des données 3D capturées en **signaux d'activation musculaire** pour créer une animation comme proposé par Choe et al. [2001]. Bien qu'elles résultent en des animations anatomiquement crédibles, la mise en place de ce genre de modèles nécessite beaucoup de travail, de connaissance anatomique et de talent de modélisation. De plus, puisque la morphologie du visage peut être très différente d'une personne à l'autre, le transfert des animations d'un modèle à un autre ne peut pas se faire sans effort élevé d'adaptation. Le coût de calcul est aussi élevé pour cette approche d'animation, vu le nombre de paramètres à calculer pour obtenir la déformation pour chaque image de l'animation.

Une autre approche d'animation qui permet de créer des modèles 3D sans modéliser toute la structure osseuse et musculaire est la méthode dite des ***Blendshapes***. Chaque image de l'animation est le résultat d'une somme pondérée de la position des vertex d'un ensemble de modèles 3D pré-modélisés. Où le vertex représente un point 3D qui fait partie d'une liste ordonnée définissant la géométrie du modèle 3D. Cet ensemble de modèles suit les conventions du système de codage d'action faciale (FACS mis en place par Ekman and Friesen [1978]) qui décrit les mouvements faciaux par un ensemble d'unités d'actions **UAs**. Chaque unité d'action représente le mouvement d'un muscle du visage. La liste des différentes UAs et leurs explications sont présentées dans l'annexe B. Pour chaque UA ou sous-ensemble d'UAs, il faut créer un modèle 3D correspondant appelé une *Blendshape*. Au moment de l'animation, il faut calculer pour chaque image un **vecteur de poids** de l'ensemble des *Blendshapes* utilisées. Une interpolation pondérée entre ces dernières génère la forme finale du modèle à un instant donné [Chuang and Bregler, 2002]. Le système de Ekman est adapté pour l'animation globale du visage mais a fait l'objet de critiques pour son utilisation dans la modélisation de l'articulation. En fait, l'ensemble d'UAs proposées est adapté à la modélisation des expressions faciales mais manque de finesse quant à la modélisation des gestes fins de l'articulation de la parole. Il est donc nécessaire de le compléter avec un ensemble de *Blendshapes* spécifiques à la parole appelées **les visèmes**, qui sont les équivalents visuels des phonèmes. Plusieurs phonèmes peuvent être représentés par un même visème, par exemple les phonèmes b, p et m peuvent être représentés par le même visème représentant des lèvres pressées, d'autres exemples sont présentés dans le tableau 5.6 du chapitre 5. L'utilisation des *Blendshapes* offre un niveau d'abstraction permettant d'animer n'importe quel modèle 3D ayant le même ensemble de *Blendshapes*, quel que soit sa morphologie, ainsi cela permet de transférer une animation d'un modèle à un autre en toute simplicité.

Il est aussi commun d'utiliser des *Blendshapes* issues d'une **Analyse par composantes principales (ACP)** [Hong et al., 2002, Chuang and Bregler, 2002]. Dans ce cas les *Blendshapes* ne sont pas créées suivant le FACS mais sont calculées par une ACP appliquée sur l'ensemble des données 3D pré-capturées pour définir les axes les plus dominants pour les mouvements faciaux. Ces axes sont appelés des unités de mouvement ou MUs *(Motion Units)*. N'importe quelle déformation faciale peut être approximée par une combinaison linéaire des MUs. Bien qu'il a été prouvé que les MUs produisent une erreur de reconstruction plus petite que les UAs, mais l'usage de l'ACP reste limité dans le domaine de l'animation puisque les MUs générées sont liées aux mouvements de plusieurs muscles simultanément. De plus ces MUs sont étroitement liées à la base de données et manquent d'interprétabilité pour les artistes qui vont créer les modèles 3D correspondants. Il existe une approche hybride qui combine les *Blendshapes* standards et celles issues de l'ACP. Ces dernières permettent de raffiner le modèle 3D et de le rendre plus adapté à un locuteur spécifique [Li et al., 2013].

1.5 La transcription de la parole

La parole est largement présentée sous sa forme textuelle. Un texte est une série écrite d'unités constituant un ensemble cohérent, porteur de sens et respectant les structures propres à une langue. Afin de transformer un texte en parole, il faut tout d'abord l'analyser pour le comprendre et décider de la manière dont il sera prononcé. Puisque, dans plusieurs langues, notamment la langue française, certains mots qui s'écrivent de la même manière peuvent être prononcés de façons distinctes (hétéronymes), aussi dans certains contextes syntaxiques, certaines lettres sont ignorées ou rattachées à d'autres par des liaisons. Le but de cette analyse est donc de convertir l'écriture en texte brute en une écriture qui décrit la prononciation exacte du mot. Cette représentation ce traduit par une chaîne de symboles représentant les sons distinctifs de la langue (les phonèmes). Pour faire cela, trois étapes s'imposent. Tout d'abord une étape de normalisation et de pré-traitement du texte brute. Cette étape consiste en un nettoyage et gestion de certaines anomalies. Les ressources textuelles contiennent souvent des chiffres, des abréviations, des dates, des références, etc. Il faut donc retranscrire en toutes lettres les chaînes de caractères non-lexicales (hors dictionnaire), ou inconnues. La deuxième étape, est celle de l'analyse lexicale et morpho-syntaxique. C'est l'étape de recherche des informations associées aux différents lexèmes. En se basant sur une base de données spécifiques, les informations sur la classe grammaticale du mot (nom, adjectif, verbe, pronom, etc.) et ses propriétés grammaticales (genre, nombre, infinitif, verbe d'état, verbe transitif, etc.) sont ajoutées. Toutefois, certains mots peuvent s'écrire de la même manière et avoir des propriétés grammaticales différentes (par exemple, le nom masculin "est" et la forme conjuguée de l'auxiliaire être "est"). Afin d'enlever cette ambiguïté, le choix de la catégorie de chaque mot se fait au moyen de règles contextuelles heuristiques, ou de modèles statistiques, prenant en compte les catégories grammaticales des mots adjacents. Nous pouvons donc rechercher dans l'ensemble des successions possibles de catégories grammaticales la succession de catégories la plus probable. La troisième et dernière étape est celle de la transcription graphème-phonème. Il s'agit de déterminer la prononciation du texte ou de phonétisation du texte. Pour cela, une liste de symboles qui représente les différents phonèmes a été établie dans l'alphabet phonétique international (API)[Decker et al., 1999]. Pour le français, un tel alphabet comporte 36 sons. Un système de phonétisation est un automate paramétré appliquant un ensemble de règles de réécriture, à ceux-là s'ajoutent des règles pour gérer les exceptions comme les noms propres, les mots empruntés de d'autres langues, liaisons, etc. Ainsi, en suivants des règles spécifiques pour combiner les phonèmes, nous pouvons former des syllabes qui sont importantes pour modéliser la prosodie. Nous présentons sur la figure 1.3 le processus de phonétisation d'une phrase, en construisant graduellement le contexte linguistique nécessaire pour la synthèse de la parole. Le système qui effectue la phonétisation d'une langue donnée est appelé un front-end.

La représentation du texte en symboles phonétiques et textualisés est nécessaire pour obtenir un alignement entre les sons de la parole et leur transcription. Les données textuelles et acoustiques/visuelles alignées représentent respectivement les entrées et sorties pour entraîner les systèmes de synthèse de la parole (TTS) à partir du texte. Nous détaillons dans le chapitre 2 et 3 les aspects relatifs à la création du modèle acoustique, visuel et audiovisuel de la parole (nommé le *back-end*).

La prise en compte du contexte dans la modélisation des phonèmes est cruciale pour capturer les effets suprasegmentaux (comme l'accent, l'intonation et le rythme) et les effets de **la coarticulation** [Abry and Lallouache, 1991, Bell-Berti and Harris, 1979, 1982]. En fait, la prononciation d'un phonème change selon son contexte (les phonèmes qui l'entourent). Ce phénomène de coarticulation résulte du fait que le conduit vocal a besoin d'un temps de transition pour passer

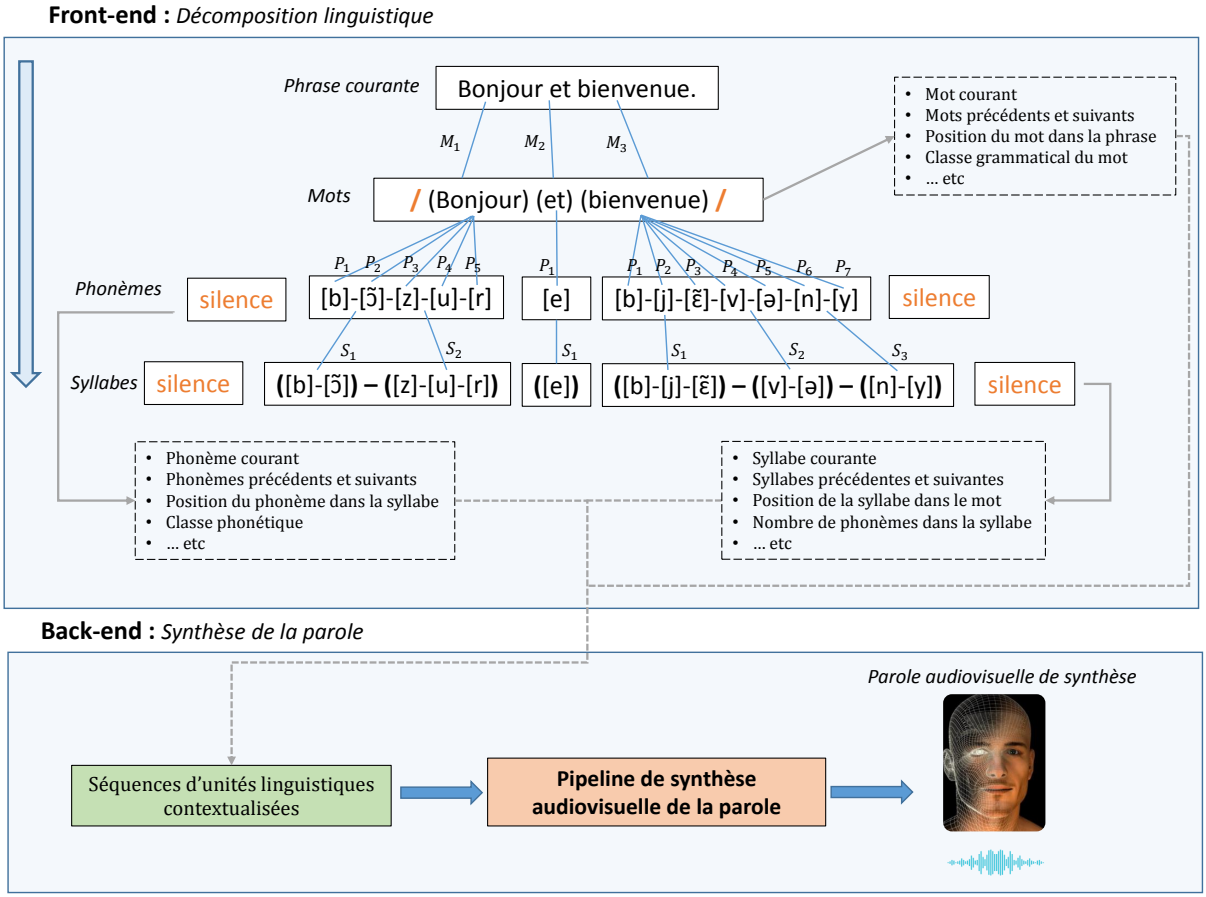


FIGURE 1.3 – *Décomposition du texte en unités linguistiques contextualisées. Ces unités sont les entrée du système de TTS audiovisuelle.*

d'une configuration à une autre. Cette phase de transition génère un très grand nombre de configurations transitoires possibles des articulateurs. La modélisation de la coarticulation est très importante pour que le son produit soit compréhensible et intelligible.

Les algorithmes d'apprentissage profond ne peuvent pas opérer sur des données sous forme de caractères, c'est pour cela qu'il faut les convertir sous une forme numérique. Il est très commun d'utiliser l'encodage one-hot qui affecte un chiffre (0 ou 1) à chaque information linguistique utilisée. Si le phonème courant satisfait un certain nombre de conditions linguistiques et contextuelles, alors seules les champs correspondants à ces conditions prennent la valeur "1", les autres prennent tous "0".

Une autre manière de représenter les informations textuelles consiste en un modèle vectoriel (*vector space model*, VSM) [Lu et al., 2013] ou des vecteurs latents [Watts et al., 2013]. Ces représentations du texte proviennent de modèles pré-entraînés de manière non-supervisée et permettent d'extraire des informations relatives à la sémantiques et aux contextes des mots [Mikolov et al., 2013, Pennington et al., 2014, Bengio et al., 2003]. Ces vecteurs de mots peuvent être ajoutés en tant qu'entrées auxiliaires à un modèle TTS pour transmettre des connaissances textuelles supplémentaires qui ne peuvent pas être apprises à partir des données de texte d'origine. Ces informations ont montré leur utilité dans un contexte d'apprentissage semi-supervisé. Dans le travail de Chung et al. [2019], les chercheurs ont réussi à générer de la parole intelligible

avec moins d'une demi-heure de données <texte,parole> appariées et 40 heures de parole non appariée en ajoutant des représentations latentes des mots dans le processus d'apprentissage. Dans l'article de Li et al. [2016b], l'incorporation de la représentation latente des phonèmes est appliquée à la synthèse des agents virtuels à partir de la parole à partir d'un DNN. Les résultats expérimentaux montrent que l'ajout de ces vecteurs implique une amélioration de 10,2% dans le test objectif.

1.6 Conclusion

Dans ce chapitre, nous avons présenté plusieurs définitions relatives à la communication et à la synthèse de la parole audiovisuelle. Nous avons ensuite exposé plus en détails les deux modalités essentielles composant la communication humaine : les mécanismes de production de la parole acoustique et visuelle. Pour chaque modalité, nous avons présenté les différentes approches de la littérature de la synthèse à partir du texte pour extraire les caractéristiques primordiales et pour paramétrer les flux pour la synthèse acoustique et visuelle. Nous avons évoqué également les techniques de préparation et de conversion du texte brut vers un format compatible avec les algorithmes d'apprentissage profond.

Dans le chapitre suivant, nous présentons les différentes approches de la littérature pour la synthèse audiovisuelle de la parole, qui profitent des techniques de paramétrisation expliquées dans ce chapitre.

2

Synthèse de la parole audiovisuelle

Sommaire

2.1 Introduction

La synthèse de la parole a connu un progrès remarquable durant les deux dernières décennies. De nos jours, la qualité des résultats de certains systèmes de synthèse vocale est très proche de la qualité de la voix originale [Shen et al., 2018].

Ce progrès est dû principalement à l'augmentation de la capacité de stockage et de la puissance de calcul qui ont engendré la réémergence des DNNs, Aussi, le besoin croissant d'avoir plus de contrôle et de flexibilité/variation de la parole a permis à de nouvelles techniques d'apprentissage de voir le jour, notamment les méthodes paramétriques statistiques vers la fin des années 90.

Toutefois, la communication parlée est essentiellement bimodale, elle est composée d'un aspect vocal et d'un aspect visuel. Les informations acoustiques de la parole sont en réalité la conséquence du mouvement des articulateurs du système de production de la parole.

Par conséquent, lorsque nous regardons le visage de quelqu'un qui parle, nous recevons deux flux d'informations distincts : un signal acoustique de parole constitué d'une série de sons et un signal visuel constitué des variations visibles du visage occasionnées par les articulateurs au niveau de la bouche. C'est la combinaison de ces deux flux d'informations qui est appelée le signal de parole audiovisuel.

Dans le cadre de la synthèse de la parole, il est donc essentiel d'inclure la modalité visuelle pour avoir une représentation complète de la parole. La perception des informations acoustiques est influencée par celle des informations visuelles et inversement. Cette dépendance/complémentarité a été prouvée par plusieurs travaux, le plus connu est l'effet McGurk [McGurk and MacDonald, 1976]. Cela se produit lors de l'affichage de fragments de discours audiovisuels dont les informations acoustiques et visuelles proviennent de sources différentes. Par exemple, lorsqu'une piste sonore contenant un / ba / et une piste vidéo contenant un / ga / sont

présentées, les participants déclarent avoir entendu un / da /, qui est une combinaison entre le contenu de la modalité acoustique et celle visuelle. Un autre effet de ce type, appelé capture visuelle *(Visual capture)*, se produit lorsque les participants qui perçoivent un extrait audio-visuel avec un visuel et un son qui ne correspondent pas, entendent la syllabe présentée visuellement au lieu de la syllabe présente dans la bande sonore [Andersen, 2010].

Le besoin de compléter la modalité acoustique par celle visuelle est d'autant plus important lorsque les utilisateurs finaux sont des malentendants ou lorsque l'environnement est bruyant [Pandzic et al., 1999, Le Goff et al., 1994]. L'ensemble du visage naturel restitue les deux tiers de l'intelligibilité acoustique manquante lorsque la transmission acoustique est dégradée ou manquante ; le modèle facial (mouvements de la langue exclus) en restitue la moitié ; et le modèle des lèvres à lui seul en restitue un tiers [Le Goff et al., 1994]. Par ailleurs, l'ajout d'une modalité visuelle rend le système plus attractif, plus divertissant pendant le temps d'attente, pousse les utilisateurs à réagir de manière plus positive, à être plus coopératifs [Pandzic et al., 1999, Ostermann and Millen, 2000] et rend le système plus digne de confiance [Van Mulken et al., 1999]. La modalité visuelle peut être générée à partir d'une piste audio si cette dernière est disponible comme elle peut l'être à partir du texte [Simons, 1990, Fan et al., 2016, Zhou et al., 2018]. Dans le cadre de ce travail, nous nous focalisons sur la deuxième approche en considérant que seul le texte est disponible pour générer une parole audiovisuelle de synthèse.

Les méthodes de synthèse de la parole sont passées des méthodes basées sur les règles aux méthodes basées sur les corpus. Dans cette partie, nous présentons l'historique des méthodes de synthèse vocale et visuelle en détaillant les techniques les plus récentes de l'état de l'art.

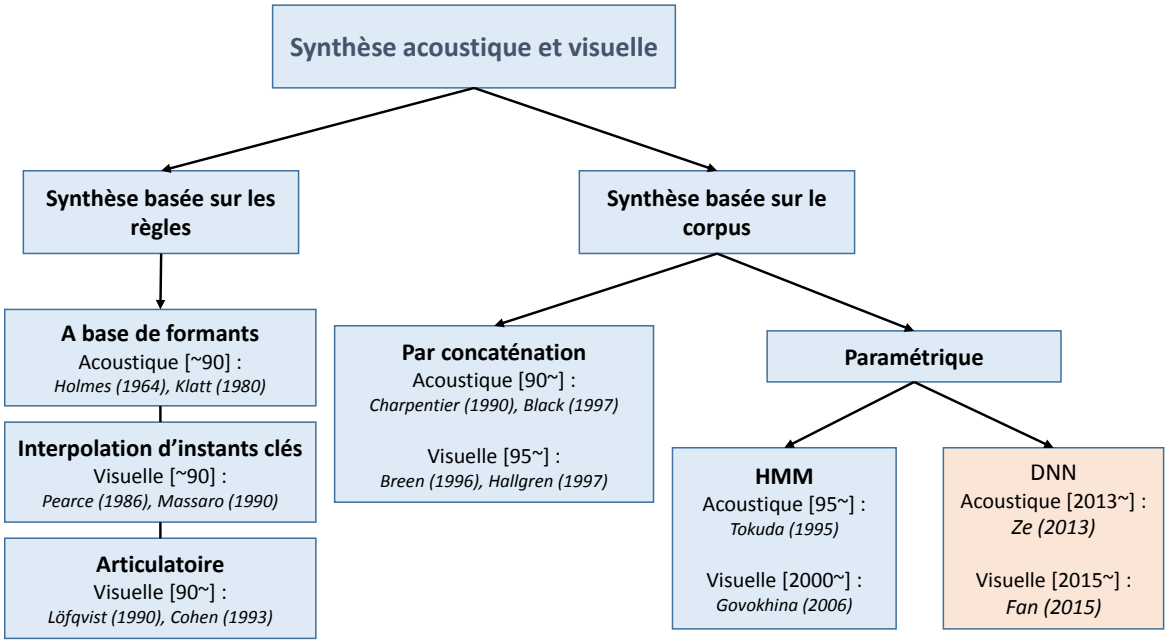


FIGURE 2.1 – *Les différentes techniques de synthèse acoustique et visuelle dans la littérature.*

2.2 Synthèse à base de règles

Pour la synthèse à base de règles à partir du texte TTS (Text-To-Speech), un certain nombre de règles sont définies pour émuler la parole humaine. Ces règles sont définies par des experts du domaine. Elles permettent de définir un ensemble de valeurs cibles et de trajectoires pour modéliser les paramètres représentant le signal acoustique ou visuel associé à une séquence phonétique.

Avant les années 90, la technique de synthèse vocale basée sur les règles était largement étudiée et utilisée [Holmes et al., 1964, Klatt, 1980, Antonov, 1981]. Cette technique de synthèse se base sur la modification des paramètres fondamentaux du signal pour simuler le spectrogramme de la parole. Elle repose uniquement sur le traitement du signal sans aucune base de données ou échantillon de voix. L'hypothèse principale derrière ce type de synthèse, est la possibilité de simuler le spectre de la parole et produire un signal intelligible à partir d'un petit nombre de formants (voir section 1.3) et de règles. Étant des générateurs de fréquences, ce genre de synthétiseurs permet de générer un nombre infini de sons, ce qui le rend plus flexible que les autres systèmes de synthèse. En revanche, le résultat de synthèse manque cruellement de naturel et la voix de synthèse a un timbre très robotique très facilement discernable de la parole naturelle humaine.

Dans les travaux de Massaro and Cohen [1990] et Beskow [1995], les premiers synthétiseurs audiovisuels à base de règles ont été présentés. Ils combinent une synthèse vocale à base de formants avec des systèmes de synthèse visuelle basée sur des règles simplistes de frames clés et de technique d'interpolation. Ainsi, dans la synthèse visuelle basée sur les règles, les propriétés du signal visuel synthétique sont prédites pour certains instants clés (généralement milieu du phonème/visème). Ces systèmes doivent implémenter des stratégies pour générer un signal avec des transitions fluides et naturelles entre les instants clés générés. Pour arriver à cette fin, des modèles de coarticulation doivent être définis. Plus tard, le modèle de Cohen and Massaro [1993] a été largement utilisé pour modéliser la coarticulation. Ce modèle est inspiré du modèle de Löfqvist [1990]. Le modèle de Cohen-Massaro suggère que les segments de la parole ont une valeur cible et une dominance sur les articulateurs qui croît puis décroît de manière exponentielle dans le temps. Les segments adjacents vont générer un chevauchement des fonctions de dominances qui, fusionnées dans le temps, créent le phénomène de la coarticulation. Ce modèle suggère aussi que chaque segment n'a pas qu'une seule fonction de dominance, mais plusieurs, chacune relative à un articulateur donné. Pour chaque instant de la parole, la valeur des paramètres de ce modèle est donnée par une somme pondérée de toutes les valeurs cibles et leurs fonctions de dominance. Les valeurs cibles ainsi que les taux de croissance/décroissance des fonctions de dominances dans le temps sont des paramètres à définir pour chaque phonème et pour chaque articulateur. Ces trois paramètres étaient choisis de façon empirique avec un processus essai-erreur pour trouver les paramètres optimaux. Goff [1997] et contrairement à la méthode de Cohen-Massaro, a utilisé une approche automatique pour identifier les paramètres optimaux du modèle. Cette approche automatique se base sur la minimisation de la distance euclidienne entre les trajectoires de la parole réelle et celles de la parole synthétisée. En revanche, le modèle de Cohen-Massaro ne permet pas d'atteindre certaines cibles articulatoires, notamment les fermetures des lèvres dans les phonèmes bilabiaux (b, p et m). Pour pallier à cette problématique, Cosi et al. [2002] ont augmenté le modèle Cohen-Massaro par une fonction de résistance pour supprimer la dominance des segments adjacents et favoriser l'atteinte des cibles. Les paramètres de ce modèle modifié ont été estimés à partir d'une base de données de VCV (Voyelle-Consonne-Voyelle) enregistrée à l'aide d'un système de capture de mouvements.

L'intérêt principal de cette méthode est qu'elle ne nécessite pas de capacité de stockage car

n'utilise pas directement une base de données. Cet avantage était d'une importance capitale avant l'augmentation des capacités de stockage des terminaux que nous connaissons aujourd'hui. Ce genre de synthèse peut être utilisé sur des terminaux sans capacité de stockage (téléphone portable) ou dans des systèmes embarqués (comme les synthétiseurs vocaux de Stephen Hawking : DECtalk et CallText 5010). Aujourd'hui les algorithmes de suivi des mouvements faciaux dans les vidéos et les systèmes de capture de mouvements sont de plus en plus accessibles. Aussi, les outils de stockage et de traitements des données visuelles permettent de gérer un grand volume de données visuelles. Ainsi, il n'est plus nécessaire d'utiliser des modèles à base de formants ou articulatoires conçus manuellement à partir de quantité limitée de données. Il est possible de stocker des échantillons acoustiques et des trajectoires visuelles réelles, les sélectionner et les concaténer pour obtenir une parole de synthèse qui respecte les règles de l'articulation et de la coarticulation.

2.3 Synthèse par concaténation

Cette méthode de synthèse se base sur la concaténation d'unités de parole préexistantes. Elle nécessite donc d'avoir un grand corpus de parole d'un même locuteur. Ce corpus peut aller jusqu'à quelques dizaines d'heures pour un résultat de haute qualité [Guennec, 2016]. Selon le texte à synthétiser, des unités de la parole sont sélectionnées dans une base de données préenregistrée. Ces unités sont choisies selon un certain nombre de critères et de coûts définis par l'algorithme de sélection adopté. Ces unités sont ensuite concaténées pour former des séquences complètes de parole. Cette approche ne nécessite aucun modèle de coarticulation puisque les trajectoires de la parole sont obtenues à partir d'exemples réels de coarticulation. Ainsi, plus la base de données est grande et riche, plus il y aura de chance que les segments sélectionnés correspondent aux besoins du texte à synthétiser. De ce fait, les interpolations et les lissages entre les segments seront limités préservant ainsi une coarticulation naturelle.

La synthèse par diphones [Charpentier and Stella, 1986] est le premier type de synthèse par concaténation. Un exemple unique de chaque diphone est stocké et concaténé avec les autres lors de la synthèse. Un diphone représente une succession entre deux moitiés de phonèmes et représente les transitions entre un couple de phonème. Le choix du diphone comme unité de concaténation est motivé par le fait que les frontières entre les diphones sont similaires puisque le centre du phonème (et donc le point de concaténation) et le point le plus stable du phonéme [O'Shaughnessy et al., 1988], ce qui réduit les aléas de concaténation. Cette technique ne nécessite pas beaucoup d'espace de stockage puisque les diphones ne sont pas très nombreux (environ 1200 pour la langue française ce qui fait environ 3 minutes de parole), de plus, plusieurs d'entre eux sont rares et nous ne pouvons les trouver qu'aux frontières entre deux mots [Dutoit, 1997]. Le résultat de cette méthode n'est pas de bonne qualité car cette base de données limitée ne peut pas représenter la variété des contextes phonétiques de la parole. En fait, elle ne peut représenter que partiellement les effets de la coarticulation, puisqu'un phonème n'est pas uniquement influencé par son voisin gauche et droit (précédant et suivant), mais l'effet de la coarticulation peut s'étendre jusqu'à cinq voyelle ou même quatre syllabes [Kochetov and Neufeld, 2013]. Pour minimiser les différences entre les diphones à concaténer, certaines modifications sont appliquées aux paramètres prosodiques du signal acoustique avec la méthode appelée TD-PSOLA *(time-domain pitch-synchronous overlap-and-add)* [Moulines and Charpentier, 1990, Colotte and Laprie, 2002]. Lors de la phase de synthèse, cette méthode agit sur le signal original pour rendre sa durée et sa F0 identique à celles prédites. Toutefois, cette méthode entraîne des distorsions qui détériorent la qualité du signal original. Pour l'aspect visuel, une approche similaire de concaténation de

demi-syllabes suédoises a été introduite par Hällgren and Lyberg [1997] comme premier système de synthèse visuel se basant sur des données 3D. Ce système a été augmenté d'un synthétiseur vocal pour générer de la parole audiovisuelle à partir d'une séquence de demi-syllabes [Hallgren and Lyberg, 1998].

Avec l'augmentation des espaces de stockage dans les années 80, le stockage de plusieurs exemples de chaque diphone est devenu possible. La multi-représentation des diphones combinés avec les bons algorithmes de sélection d'unités ont permis de minimiser les distorsions et les artefacts de concaténation [Sagisaka, 1988]. Le choix des unités à concaténer se fait par une fonction de coût que l'algorithme de sélection cherche à minimiser comme montré sur la figure 2.2.

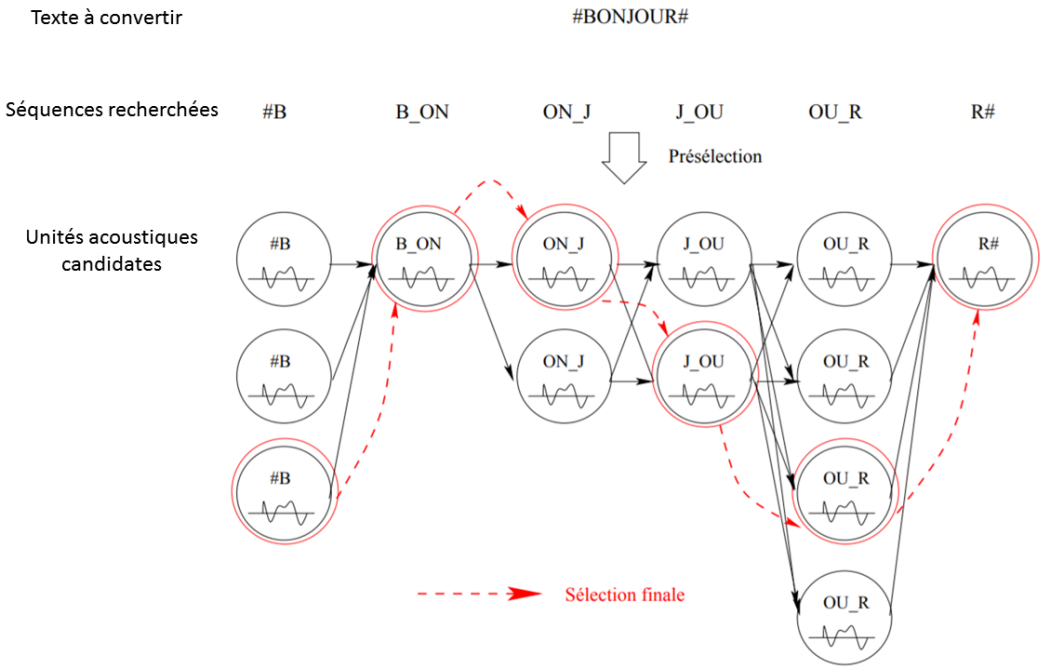


FIGURE 2.2 – *Sélection des unités pour la synthèse vocale par concaténation [Rouibia, 2006].*

L'avantage de cette méthode de synthèse (lorsque la base de données est suffisamment large) est son résultat naturel jusqu'alors inégalé par toutes les autres méthodes de synthèse. En fait, le coté naturel du résultat vient du fait que les unités de la parole concaténées proviennent de la parole humaine originale préservée, sans avoir subit de modifications particulières. En revanche, pour avoir un résultat d'une aussi grande qualité, il est nécessaire d'avoir un corpus de taille conséquente pour couvrir toutes les variétés phonétiques de la langue dans différents contextes linguistiques. En cas d'absence d'une unité de la base de données, le résultat de la synthèse peut être facilement dégradé à cause de la grande différence entre les unités concaténées qui sera très perceptible au niveau des transitions et des frontières des unités. Ces artefacts ponctuels de concaténation pénalisent lourdement l'intelligibilité du signal généré.

Les systèmes de synthèse par concaténation adoptent différentes tailles d'unités de concaténation [Kishore and Black, 2003]. Plus l'unité est longue, plus la base de données doit être conséquente pour couvrir tous les contextes possibles. Les unités les plus utilisées dans la littérature pour la synthèse vocale sont, le phonème [Black, 1996], le diphone [Black and Taylor, 1997], le demi-phonème [Beutnagel et al., 1999] et la syllabe [Kishore and Black, 2003]. D'autres travaux ont utilisés des frames de 5ms d'écart comme unité [Hirai and Tenpaku, 2004], des états HMMs

[Donovan and Woodland, 1995] ou des unités plus longues ou non-uniformes [Taylor and Black, 1999]. En ce qui concerne la synthèse visuelle et selon le système à mettre en place, les données visuelles peuvent être sous forme de séquences vidéos 2D ou alors des données spatiales 3D issues de systèmes de capture de mouvements ou de séquences de scans et de modèles 3D. Comme pour la synthèse vocale, les différents systèmes de synthèse visuelle par concaténation adoptent différentes unités de concaténation. Certains utilisent les phonèmes [Minnis and Breen, 2000], la plupart utilisent les di-phonème/di-visème [Breen et al., 1996, Ouni et al., 2013], d'autres préfèrent des unités plus grandes (tri-phonèmes, syllabes ou mots) pour limiter les concaténations [Bregler et al., 1997], et certains systèmes utilisent des unités hybrides ou dynamiques, où le choix de l'unité se fait au moment de la sélection en favorisant le choix des utilités les plus longues d'abord [Cao et al., 2004, Minnis and Breen, 2000]. Des techniques de programmation dynamiques sont utilisées pour déterminer le meilleur chemin qui minimise le coût global de la concaténation des unités sélectionnées.

Un système de synthèse visuelle par concaténation où le processus de sélection d'unité est basé sur des HMMs a été présenté par Govokhina et al. [2006]. Ce système augmente considérablement la corrélation entre les trajectoires visuelles synthétiques et celles originales. Dans la plupart des systèmes de synthèse audiovisuelle, les deux modalités acoustique et visuelle sont modélisées séparément puis synchronisées ultérieurement. Dans Minnis and Breen [2000], les modalités acoustique et visuelle ont été conjointement synthétisées avec un système de synthèse par concaténation de phonèmes. Ce système définit les règles de concaténation des unités visuelles et permet de modéliser les effets visuels coarticulatoires comme une extension du système de synthèse vocale par sélection d'unité. Les travaux de Ouni et al. [2013] et Musti [2013] présentent une technique de synthèse audiovisuelle qui génère simultanément le signal vocal et une animation 3D du visage d'un personnage. Cela se fait en concaténant des diphones bimodaux naturellement synchronisés. La technique proposée surmonte les problèmes d'asynchronie et d'incohérence inhérents aux approches classiques de la synthèse audiovisuelle. Les différentes étapes de synthèse sont similaires à la synthèse vocale concaténative classique mais sont généralisées au domaine audiovisuel. Un nouveau terme a été introduit dans la fonction de coût qui permet de pénaliser les discontinuités importantes entre les diphones à concaténer. Des évaluations perceptuelles montrent que cette technique de synthèse bimodale fournit une parole intelligible que ça soit pour la modalité acoustique ou visuelle.

La synthèse par diphone permet d'obtenir un résultat visuel naturel de haute qualité, toutefois, cette approche souffre d'une série de limitations. Tout d'abord, il est nécessaire d'avoir une grande quantité de données pour obtenir de bons résultats de synthèse, car le moindre défaut de concaténation nuit énormément à la qualité de la synthèse. Dans les années 90s le système proposé par AT&T [Beutnagel et al., 1999] pour la synthèse vocale par concaténation avait tellement surpris la communauté par sa haute qualité, que le problème de la synthèse vocale avait été considéré par certains chercheurs comme résolu. Néanmoins, le naturel de la parole humaine ne vient pas seulement de la qualité des sons mais aussi de la variabilité des intonations et des styles de la voix. La synthèse par concaténation manque de flexibilité. En effet, elle permet uniquement de générer la parole dans le même style de parole présent dans la base de données. Pour pouvoir générer de nouveaux styles, différentes émotions ou de nouvelles voix, il est nécessaire d'enregistrer d'autres corpus pour chaque variation souhaitée.

2.4 Synthèse paramétrique

Les systèmes paramétriques, n'utilisent pas le signal acoustique/visuel dans son état brut (comme c'est le cas pour la synthèse par concaténation), le signal est transformé en une série de paramètres pertinents comme expliqué dans le chapitre 1. Lors du processus de l'apprentissage, l'optimisation de ces systèmes se fait par des approches du domaine de l'apprentissage statistique et non en identifiant les règles manuellement.

Les méthodes de synthèse paramétriques se basent sur des modèles préalablement entraînés sur une ou plusieurs bases de données. Ces approches nécessitent moins de mémoire lors de la phase de synthèse, puisqu'il faut garder en mémoire uniquement les paramètres du modèle entraîné et non la base de données elle-même. Ces méthodes sont aussi plus flexibles et contrôlables, permettant plus facilement de changer le style de la parole ou la voix par adaptation, interpolation ou mixage de différents modèles entraînés [Yoshimura et al., 2000, Schabus et al., 2012]. Deux approches de synthèse paramétrique vont être détaillées ci-dessous : HMM (Hidden Markov Models) et DNN (Deep Neural Networks).

2.4.1 Synthèse par HMM

La synthèse vocale par HMMs fut très largement adoptée à partir des années 90. Cette approche statistique paramétrique se base sur la modélisation des paramètres acoustiques/visuelles avec des modèles de Markov cachés. La synthèse par HMMs repose sur des arbres de décisions. Plusieurs paramètres linguistiques sont considérés avec des facteurs contextuels (linguistiques, prosodiques et phonologiques). D'abord un modèle à états (généralement trois ou cinq états) est produit pour chaque phonème de manière indépendante, ces modèles portent le nom de CI-HSMM (pour *Context Independant Hidden State Markov Models*). Ensuite, ces modèles (CI-HSMM) sont utilisés pour initialiser les modèles CD-HSMM (pour *Context Dependant Hidden State Markov Models*) qui sont estimés en prenant en compte le contexte de chaque phonème. Une liste de questions binaires correspondant aux paramètres linguistiques est alors parcourue et les états des HMMs sont regroupés, étape par étape, suivant les réponses à ces questions. Le choix de la question à chaque étape, se fait de telle sorte que la question sélectionnée soit celle qui maximise le gain de la vraisemblance des données générés avec les données d'apprentissage. Une illustration du fonctionnement de l'arbre de décision est présentée dans la figure 2.3. Finalement, pour créer les modèles contextuels, les états sont partagés suivant un processus nommée *"state-tying"* [Young et al., 1994]. Cette étape consiste à regrouper les contextes linguistiques qui sont susceptibles d'être associés à des observations proches et de ré-estimer les modèles CD-HSMM pour obtenir des TCD-HSMM *(Tied-state Context-Dependent)*.

Pour la synthèse vocale, les paramètres du signal acoustique et les informations linguistiques du texte sont utilisées comme entrées et sorties pendant la phase de l'entraînement. Communément, les trois flux (spectre de fréquences (conduit vocal), la fréquence fondamentale (source de la voix) et la durée des phonèmes) de la parole sont modélisés simultanément à l'aide de HMMs basés sur le critère de vraisemblance maximale (MLPG) [Tokuda et al., 2000, 1995a, Masuko et al., 1996, Tokuda et al., 1995b]. Chaque flux est modélisé par un arbre de décision dédié. La motivation derrière ce choix vient du fait que différentes informations linguistiques sont importantes pour différents flux de données. Ce qui veut dire que la combinaison de plusieurs flux de données dans un même arbre de décision est problématique, puisque des questions pertinentes pour un flux peuvent ne pas l'être pour un autre et va donc partager les données de manière sous-optimale si elle est posée très tôt, près de la racine de l'arbre. On parle alors de modèles HMM multi-flux. Les paramètres liés au spectre de la parole sont modélisées par

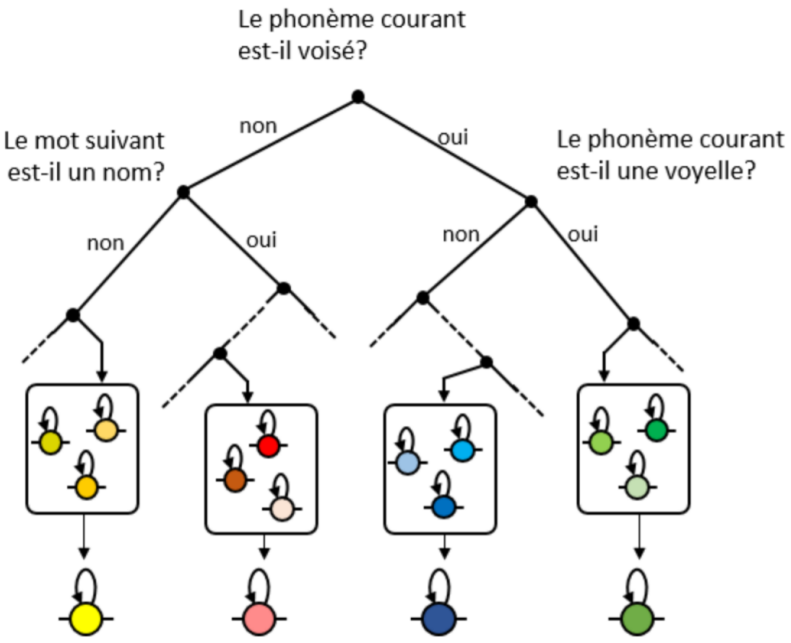


FIGURE 2.3 – *Illustration du partitionnement de l'espace d'apprentissage grâce à un arbre de décision [Pouget, 2017].*

des HMMs à densité continue, les durées sont modélisées par des distributions gaussiennes multidimensionnelles [Yoshimura et al., 1998], les paramètres relatifs à la fréquence fondamentale sont modélisés par une probabilité de distribution multi-espace MSD (Multi-Space probability Distribution) [Tokuda et al., 1999, 2002]. En fait, les paramètres liés à la fréquence fondamentale ne sont pas définis de façon continue (ils ne le sont que pour les sons voisés). La modélisation d'une trajectoire discontinue nécessite une adaptation des procédures d'apprentissage des HMMs. L'approche par MSD décompose chaque densité de probabilité comme une combinaison d'une distribution discrète (pour modéliser le caractère voisé/non-voisé de chaque trame) et d'une distribution continue (pour modéliser les variations de la fréquence fondamentale). HTS [Zen et al., 2007] est le système de synthèse à base de HMMs le plus répandu, il a été publié en 2002 avec un code source ouvert et a été largement utilisé par la communauté scientifique de synthèse de la parole [Zen et al., 2004b,a, Wu and Wang, 2006].

Les HMMs ont eu aussi beaucoup de succès dans le domaine de la synthèse visuelle. Le premier système TTS visuel par HMMs a été proposé par Brooke and Scott [1994]. Ce prototype, applique l'ACP (15 axes principaux) sur des vidéos monochromes contenant une bouche prononçant des chiffres. Masuko et al. [1998] ont augmenté ce système avec des paramètres dynamiques (les dérivées temporelles des paramètres visuels). Ces paramètres ont permis d'incorporer des informations contextuelles sur les frames passées et futures. Par conséquent, cet ajout a permis d'obtenir des trajectoires visuelles plus lisses, naturelles et qui intègrent les règles de la coarticulation. Govokhina et al. [2006] ont aussi trouvé que l'utilisation des paramètres dynamiques pendant l'entraînement du système basé sur les HMMs augmente significativement la qualité de trajectoires visuelles générées (notamment la première dérivée temporelle). Dans ce travail, une comparaison entre le système basé sur les HMMs et un système par concaténation a été réalisée. La qualité de la synthèse par HMMs a dépassé celle de la synthèse par concaténation. Cette constatation a été confirmée par une évaluation subjective et une autre objective. Une approche

hybride entre les HMMs et la concaténation a été présentée par Wang et al. [2010]. Dans ce travail la trajectoire du mouvement des lèvres est prédite en utilisant un modèle HMMs préalablement entraîné. Au moment de la synthèse, le système se base sur les trajectoires générées par les HMMs et sur le principe de minimisation de l'erreur de génération pour sélectionner les unités les plus adéquates et les concaténer. Ce système a obtenu la première place dans la compétition LIPS2009 dans le contexte audiovisuel, après une évaluation perceptive par des humains.

Dans le travail de Sako et al. [2000], un système de synthèse audiovisuelle a été proposé avec deux modèles HMMs séparés, un dédié pour la modalité acoustique et l'autre pour celle visuelle. Afin de garantir la synchronisation entre les deux modalités, les mêmes durées de phonèmes ont été utilisées pour les deux modèles HMMs. D'abord, les durées et les paramètres acoustiques sont générés par le modèle acoustique, ensuite ses durées sont passées au modèle visuelle pour générer l'animation des lèvres. Xie et al. [2014] proposent un système de synthèse audiovisuelle vidéo-réaliste basée sur les HMMs pour animer la partie inférieure du visage. Schabus et al. [2013] ont montré qu'un entraînement joint des deux modalités acoustiques et visuelles résulte en un modèle de durées plus naturel que celui entraîné sur des données acoustiques ou visuelles seulement. De plus, les données générées par le modèle joint sont d'aussi bonne qualité que celles générées par les modèles entraînés séparément. Dans les systèmes TTS basés sur les HMMs, la continuité des trajectoires acoustiques et visuelles générées a été assurée par l'algorithme de génération de paramètres de vraisemblance maximale MLPG [Tokuda et al., 2000]. MLPG est utilisé pour prendre en compte les contraintes des paramètres dynamiques. Il génère, au moment de la synthèse, la séquence la plus probable compte tenu des statistiques des paramètres statiques et dynamiques ce qui résulte en des trajectoires plus lisses.

La synthèse par HMMs est flexible. Contrairement à la synthèse par concaténation qui se base sur des unités vocales préenregistrées, ce système génère la parole à partir des HMMs eux mêmes, ce qui permet de modifier facilement les caractéristiques de la parole en appliquant des changements sur les modèles HMMs directement. Plusieurs travaux ont été publiés pour l'adaptation/imitation de voix [Tamura et al., 1998, 2001, Yamagishi et al., 2003b], interpolation/mixage de voix [Yoshimura et al., 1997, 2000], production de voix avec des caractéristiques particulières [Shichiri et al., 2002, Kazumi et al., 2010] ou contrôler le degré d'expressivité d'une voix ou le style de la parole [Nose et al., 2007]. De plus, lorsque la couverture phonétique du corpus est faible ou lorsque sa qualité d'enregistrement est mauvaise, les systèmes de synthèse par HMMs restent robustes [Zen and Toda, 2005]. Pour l'aspect visuel, quelques travaux ont prouvé la flexibilité des modèles HMMs. Filntisis et al. [2017] ont adapté un modèle HMMs entraîné sur des données neutres avec une petite quantité de données expressives pour générer des animations expressives audiovisuelles de bonne qualité. Ils ont également montré la possibilité de générer différents niveaux d'intensité pour les émotions et des styles de parole intermédiaires par interpolation des modèles HMMs. Dans Schabus et al. [2012], il a été démontré que lorsque la quantité de données visuelles d'un locuteur donné est petite, il est possible d'augmenter la qualité de la synthèse visuelle pour ce locuteur en adaptant un modèle HMM entraîné sur plusieurs bases de données avec les données visuelles spécifique à ce locuteur.

Toutefois, sous des conditions idéales la qualité de synthèse par concaténation reste supérieure à celle par HMMs comme il a été démontré dans les résultats du "Blizzard Challenge" [2] [Black et al., 2010, King and Karaiskos, 2011, 2012]. En effet, cette dégradation de la qualité de synthèse vient du fait que les systèmes à HMMs appliquent une série de lissages à différents endroits du processus de l'apprentissage [Merritt et al., 2015]. La manière dont les HMMs sont paramétrés implique systématiquement un regroupement de modèles lorsque les contextes linguistiques sont

2. Lien contenant l'historique des résultats du "Blizzard challenge" : http ://www.festvox.org/blizzard/

similaires. Cependant, la moyenne sur plusieurs contextes dégrade considérablement la qualité de la parole synthétisée. De surcroît, l'approche standard de la synthèse paramétrique statistique utilise des HMMs avec un nombre fixe d'états émetteurs. Chaque état contient une distribution gaussienne multivariée. Lorsqu'on génère d'un tel modèle en utilisant l'algorithme MLPG, une séquence de frames est émise de chaque état. En revanche, la moyenne sur la durée de l'état est toujours la même. Cette moyenne introduit un lissage temporel sur les paramètres générés, et la force du lissage dépend de la durée de l'état. De plus, puisque le corpus d'apprentissage n'est pas suffisant pour couvrir toutes les classes (selon leur contexte linguistique), la probabilité des classes non observées va alors être nulle causant le problème de dispersion de données (*data sparsity*) et l'estimation des modèles contextuels ne sera pas robuste. Des exemples tirés de contextes différents doivent alors être regroupés et moyennés afin de pouvoir estimer de manière fiable la moyenne des états et leur variance. Ceci introduit le lissage spectral interclasse suivant un algorithme particulier de lissage [Chen and Goodman, 1999]. D'autres facteurs qui peuvent dégrader la qualité de la synthèse par HMMs peuvent venir du fait que la variance des trajectoires générées ne correspond pas exactement à celle de la parole naturelle à cause de l'estimation des paramètres du modèle d'un nombre limité de données ou un modèle inadéquat.

D'un autre côté, les arbres de décisions représentent un certain nombre de limitations. En fait, ils ne sont pas efficaces pour modéliser les dépendances contextuelles complexes tel que l'opération logique XOR (OU exclusif) par exemple [Esmeir and Markovitch, 2007]. Pour représenter de tels cas, les arbres de décision seront très larges. Deuxièmement, cette approche divise les données et utilise des paramètres distincts pour chaque région, chaque région étant associée à un noeud terminal de l'arbre de décision. Cela aboutit à la fragmentation des données d'apprentissage et à la réduction de la quantité de données pouvant être utilisée pour regrouper les autres contextes et estimer les distributions [Yu et al., 2011]. Le fait de disposer d'un arbre de taille prohibitive et de fragmenter les données d'apprentissage entraînera à la fois une sur-adaptation et une dégradation de la qualité de la parole synthétisée. Il est donc nécessaire de trouver un modèle de régression plus puissant que l'arbre de décision.

2.4.2 La réémergence des réseaux de neurones

Le domaine de l'intelligence artificielle (IA) a connu une période d'intérêt très vif suivit d'une déception et de critiques sévères dans les années 70. Cette période communément connu comme "l'hiver de l'intelligence artificielle", a été marquée par le pessimisme dû aux résultats insatisfaisants des systèmes basés sur l'IA de l'époque. Jugés trop lents, moins précis et beaucoup plus coûteux que l'humain et ne répondant pas aux promesses exagérées faites par les scientifiques, la plupart des financements ont été coupés. Dans le début des années 2000, la technologie basée sur l'IA a recommencé à gagner de l'intérêt. Elle a été appliquée avec succès à de nombreux problèmes grâce à l'accès à de grandes quantités de données, à des ordinateurs plus rapides et à des techniques avancées d'apprentissage automatique.

En 2013 les premiers travaux de synthèse acoustique par réseaux de neurones profonds DNNs ont été publiés. Ces systèmes utilisent des DNNs à propagation vers l'avant DNNs-FF (FeedForward DNNs) pour modéliser le lien entre les paramètres linguistiques et acoustiques directement [Ze et al., 2013, Lu et al., 2013, Qian et al., 2014, Wu et al., 2015b]. Les DNNs peuvent être vus comme un remplacement des arbres de décisions utilisées par les HMMs. Les DNNs simulent la production de parole humaine par une structure en couches afin de transformer les informations textuelles linguistiques en une sortie vocale finale.

L'utilisation des DNNs a permis de surmonter certaines limitations liées à l'utilisation des arbres de décision. Comme discuté dans la section précédente (2.4.1). Les arbres de décisions

ne sont pas efficaces pour traiter des problèmes de grande complexité et chaque paramètre du modèle n'est appris que sur un sous-ensemble de données d'apprentissage alors que les DNNs sont capables de représenter des fonctions de transformations hautement complexes de manière compacte. L'utilisation des arbres de décision introduit une moyenne sur plusieurs contextes linguistiques, ce qui dégrade sensiblement le naturel du discours synthétisé. Les DNNs, en revanche, peuvent facilement représenter des fonctions avec des dépendances complexes, et chaque paramètre du modèle est optimisé pour tous les échantillons d'apprentissage. En évitant le partitionnement des données dans les arbres de décision, l'utilisation des DNNs peut atténuer les effets néfastes du moyennage sur plusieurs contextes linguistiques. L'utilisation des DNNs a permis d'obtenir une amélioration significative de la qualité de la synthèse par rapport aux résultats des systèmes par HMMs [Ze et al., 2013, Qian et al., 2014, Wu et al., 2015b, Hashimoto et al., 2015] cela a pu être démontré avec des évaluations objectives sur les paramètres spectraux et d'excitation, et aussi avec des tests perceptifs en utilisant les mêmes paramètres d'entraînement [Ze et al., 2013]. En revanche, il était difficile de savoir d'où venaient ces améliorations et si elles étaient liées à l'utilisation des DNNs ou à d'autres facteurs. En effet, les systèmes par HMMs se basent sur une représentation en états du signal avec un arbre pour chaque flux de données (paramètres du filtre de la source et du conduit vocaux) alors que l'approche par DNNs apprend à représenter les différents flux simultanément. Aussi, les DNNs se basent sur une représentation en frames des données et non en états comme les HMMs. Pour comprendre la source de l'amélioration significative attribuée à l'utilisation des DNNs, une étude a été réalisée pour identifier l'apport de chaque facteur de manière isolée [Watts et al., 2016]. Cette étude a montré que l'apprentissage sur des flux de données joints au lieu de flux séparés n'a pas d'impact significatif sur la qualité de la synthèse. En revanche, trois changements ont amélioré la qualité de la synthèse : 1) Le passage d'une architecture basée sur des arbres de décision à une architecture basée sur des DNNs, 2) le passage d'une représentation par états des données à une représentation par frames et 3) l'enrichissement des données contextuelles avec des informations relatives aux durées. Cette étude démontre bien que dans le cadre de la synthèse vocale, le remplacement des arbres de décisions par des DNNs résulte en une parole synthétique plus naturelle et valide donc le choix de l'utilisation des DNNs dans ce genre de système. Il faut noter aussi que les modèles HMMs ne supportent que les paramètres binaires, alors que Ze et al. [2013] ont trouvé que l'utilisation de paramètres numériques améliore la performance des DNNs et que ces derniers sont plus efficaces puisqu'ils permettent de réduire la taille du vecteur d'entrée en le représentant de manière plus compacte.

Bien que les résultats des DNNs-FF soient meilleurs que ceux des HMMs, ces réseaux effectuent le lien entre les paramètres linguistiques et acoustiques frame par frame sans prendre en compte les contraintes contextuelles du passé ou du futur [Schuster and Paliwal, 1997].

Dans la synthèse basée sur les DNNs, l'algorithme de lissage MLPG a été hérité des HMMs, et est utilisé avec les modèles DNNs-FF pour la prédiction des paramètres acoustiques [Klimkov et al., 2018] et visuels [Filntisis et al., 2017] plus lisses et pour diminuer leurs discontinuités.

Les réseaux de neurones récurrents RNNs (Recurrent Neural Network) sont capables d'incorporer automatiquement des informations contextuelles provenant des entrées passées, ce qui leur permet de modéliser des liens séquence à séquence. Schuster and Paliwal [1997] proposent les RNNs bidirectionnels (BRNNs) pouvant incorporer des informations contextuelles provenant d'entrées passées et futures. L'idée de base de cette structure bidirectionnelle est de présenter chaque séquence en avant et en arrière avec deux couches cachées récurrentes distinctes qui sont toutes les deux connectées à la même couche suivante. Cela fournit au réseau un contexte complet, symétrique, passé et futur pour chaque point de la séquence d'entrée. Mais les RNNs conventionnels ne peuvent pas bien modéliser les relations à longue portée dans les données

séquentielles à cause du problème du gradient disparaissant *(Vanishing Gradient Problem)*. Hochreiter and Schmidhuber [1997] ont constaté que l'architecture LSTM, qui utilise des cellules de mémoire spécialement conçues pour stocker des informations, permet de mieux rechercher et exploiter un contexte à longue portée. Les LSTMs bidirectionnels ou BLSTMs (Bidirectionnel Long Short-Term Memory) [Hochreiter and Schmidhuber, 1997] peuvent accéder à un contexte à longue portée dans les deux sens (passé vers le futur et futur vers le passé). Pour inclure ces contraintes contextuelles, le réseau récurent RNN bidirectionnel de type BLSTM a été utilisé dans [Fan et al., 2014] pour faire le lien entre des séquences de paramètres linguistiques avec des séquences de paramètres acoustiques tout en préservant les informations contextuels des frames passées et futures. Comme précisé plus haut, l'utilisation des HMMs et des DNNs-FF implique l'utilisation de paramètres statiques et dynamiques des données acoustiques pour obtenir des trajectoires lisses en sortie [Tokuda et al., 2000]. En utilisant les BLSTMs, il est possible de ne pas utiliser des paramètres dynamiques tout en obtenant des trajectoires acoustiques bien lisses [Fan et al., 2014], puisque l'évolution dynamique est prise en compte par les frames passées et futures. Il est aussi possible de réduire les paramètres d'entrée relatifs au contexte linguistique à l'information concernant la frame courante uniquement. Ceci permet de réduire le nombre de paramètres d'entrée et de sortie, même si l'utilisation des BLSTMs augmente le nombre de paramètres du réseau à entraîner et nécessite plus de ressources et de puissance d'entraînement.

Inspiré du ToolKit HTS [Zen et al., 2007] et du front-end Festival, un nouveau ToolKit de synthèse vocale par DNNs nommé MERLIN-TTS [Wu et al., 2016] est apparu fin 2016. Cet outil propose une structure pour entraîner un système de synthèse acoustique à partir d'un certain nombre de paramètres d'entrée et de sortie. Il permet d'entraîner un modèle de durée et un autre modèle acoustique séparément comme détaillé dans la figure 2.4.

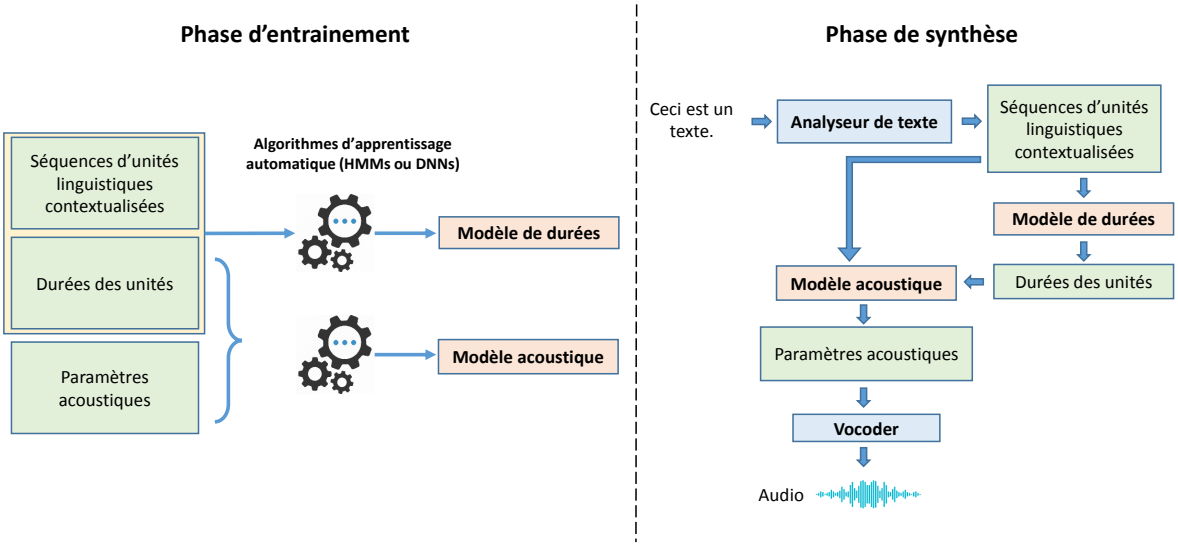


FIGURE 2.4 – *Les deux phases d'entraînement et de synthèse pour la synthèse vocale par DNNs. Les deux modèles de durées et acoustique sont entraînés séparément.*

Les deux modèles de durées et acoustique, prennent en entrée des informations linguistiques contextualisées. Concernant les paramètres de sorties, le modèle de durées génère uniquement une information relative à la durée de chaque phonème. Quant au modèle acoustique, il permet de générer les paramètres : MFCC, BAP, Log-F0 (plus une valeur binaire concernant la nature voisé/non-voisé de la frame courante) ainsi que leurs paramètres dynamiques (première

et deuxième dérivées). Ces paramètres sont générés par un Vocodeur qui permet de les extraire à partir d'un fichier audio pour qu'ils soient utilisés pour l'entraînement du modèle acoustique. Le Vocodeur intervient également dans la phase de la synthèse pour reconstruire un fichier audio à partir de paramètres générées par les modèles acoustiques.

Les expériences menées par Fan et al. [2015] ont montré qu'un système TTS visuel par réseau BLSTM surpasse largement (préféré à 61.5%) un système équivalent à base de HMMs (préféré à 23%, les 15.5% restants sont donnés à une réponse neutre). Ils ont également étudié différentes architectures de réseaux de neurones pour générer des données visuelles à partir du texte. Plusieurs architectures de différentes largeurs (nombre de neurones par couche) et profondeurs (nombre de couche cachées) ont été analysées. Ils ont trouvé que l'architecture qui donne les meilleures résultats est celle contenant trois couches cachées, une à propagation vers l'avant suivie de deux couches BLSTMs. Filntisis et al. [2017] ont utilisé des DNNs-FF pour entraîner un système de synthèse audiovisuelle. Dans ce travail deux configurations ont été adoptées. La première contient deux modèles DNNs séparés (DNN-S), le premier dédié à la modalité acoustique et le second à celle visuelle. La deuxième configuration contient un seul modèle DNN entraîné sur les deux modalités conjointement (DNN-J). L'évaluation comparative perceptive sur les résultats de ces modèles a montré que les deux systèmes DNN-S et DNN-J se valent, aucune différence statistiquement significative n'a été obtenue pour les stimuli audiovisuels générés. Par ailleurs, des comparaisons avec un système à base de HMMs et un autre par concaténation, tous entraînés sur les mêmes données, ont montré que les systèmes par DNNs (DNN-S et DNN-J) présentent plus de réalisme que ce soit pour la modalité visuelle ou pour la modalité acoustique.

En plus de générer des résultats naturels et intelligibles, la synthèse par réseaux de neurones offre aussi une très grande flexibilité. Les DNNs ont montré des capacités pour l'adaptation à la voix d'un nouveau locuteur [Wu et al., 2015a], pour changer les caractéristiques vocales liées à l'âge ou au genre du locuteur [Luong et al., 2017] et pour l'adaptation à différents styles de parole Wu et al. [2018] ou différentes émotions [Parker et al., 2017]. Nous présenterons en détails les applications des DNNs pour la synthèse expressive audiovisuelle dans le chapitre suivant (chapitre 3).

2.5 Conclusion

Nous avons présenté dans ce chapitre les différentes techniques de synthèse vocale et visuelle dans la littérature. De nos jours, c'est la synthèse paramétrique par réseaux de neurones qui concentre toute l'attention. Les différents travaux et comparaisons ont montré que les techniques basées sur cette approche offrent plus de flexibilité alors que les résultats des techniques par sélection d'unités sont plus naturels.

Les méthodes paramétriques furent longtemps critiquées puisque le passage par un vocodeur impose une limitation intrinsèque à la qualité de la parole même avant le passage par un modèle acoustique. Pour la modalité visuelle, le mouvement des articulateurs est jugé convainquant et de meilleure qualité. En revanche, l'aspect visuel souffre aussi de résultats flous et de la perte de précision introduite par la paramétrisation des données visuelles (à cause de la réduction de dimension par ACP par exemple). Pour cela, plusieurs travaux de synthèse hybride ont été proposés pour l'audio [Kominek and Black, 2006, Black et al., 2007, Merritt et al., 2016] et le visuel [Tao et al., 2009, Wang et al., 2010]. Ces systèmes tirent partie des avantage de la synthèse par concaténation et de celle paramétrique. La sélection d'unité guidée par des modèle HMMs puis DNNs ont permis d'obtenir de meilleures résultats que les deux systèmes pris indépendamment.

Mais aujourd'hui, et grâce aux Vocodeurs basés sur des DNNs (WaveNet) présentés dans le

chapitre précédent (chapitre 1), la qualité de la parole a gagné énormément de naturel et a même dépassé la qualité des systèmes par concaténation [Oord et al., 2016].

De nos jours la question de l'interprétabilité des résultats générés par les réseaux de neurones est soulevée. Avec les systèmes à base de formants ou par concaténation les règles sont écrites par des humains. Pour les systèmes HMMs, il est possible d'accéder à tout l'acheminement réalisé par le système pour générer les résultats et peut être lu et compris par l'humain facilement. En revanche, les réseaux de neurones sont des boites noires, et on ne peut pas expliquer pourquoi le réseau est arrivé à une certaine solution. Récemment, quelques travaux dans "l'intelligence artificielle explicable" (*explainable AI* ou *XAI*) ont été réalisés pour comprendre ce qui se cache derrière des systèmes de réseaux de neurones entraînés pour des tâches diverses. Actuellement, ce genre de recherches concernent principalement les systèmes médicaux et juridiques où toutes décisions doivent être justifiées et transparentes [Gunning, 2017, Holzinger et al., 2017]. Il est très probable que les outils d'analyse qui résulteront de ces recherches seront adaptables pour le domaine de la synthèse de la parole et permettront une meilleure compréhension du fonctionnement des modèles DNNs dédiés.

Finalement, la synthèse à l'aide des réseaux de neurones représente aujourd'hui le futur de la synthèse audiovisuelle. Elle permet de générer des résultats naturels et intelligibles avec une très grande flexibilité, il n'y a actuellement aucune raison d'hésiter entre la synthèse par réseaux de neurones et les autres techniques de TTS. Cependant, un domaine qui reste à explorer est celui de la synthèse audiovisuelle expressive de la parole. Les travaux réalisés dans ce domaine sont discutés dans le chapitre suivant.

3

Synthèse expressive de la parole

Sommaire

3.1 Introduction

Dans son quotidien, l'être humain est confronté à plusieurs situations qui peuvent déclencher des états émotionnels divers. Les émotions étant des réactions spontanées à une situation donnée, elles peuvent entraîner des manifestations physiques (pâleur, rougissement, agitation, accélération cardiaque et du respiratoire, transpiration, etc.) et psychologiques (pensées négatives ou positives, changement d'humeur) qui durent quelques secondes, quelques minutes ou même quelques heures Verduyn et al. [2015], Frijda et al. [1991]. Les recherches ont reconnu que l'expression des émotions aide à réguler l'interaction sociale et remplie des fonctions sociales importantes comme la communication des convictions, des désirs et des intentions aux autres Frijda and Mesquita [1994], Keltner and Haidt [1999], Keltner and Kring [1998], Morris and Keltner [1999].

La parole expressive de synthèse peut être utilisée dans des applications de narration pour enfants où, différentes expressions doivent être générées dans différents contextes de l'histoire Theune et al. [2006]. Elle peut être utilisée aussi dans un système de dialogue qui rend l'interaction humain-machine plus naturelle et efficace Bates et al. [1994]. Plusieurs travaux ont intégré la dimension expressive dans l'animation des agents virtuels. Il a été établi que les agents virtuels expressifs sont jugés plus naturels, plus crédibles et plus réalistes par rapport aux agents virtuels non expressifs Bates et al. [1994]. Dans l'expérience de Walker et al. [1994] les participants ont montré plus d'engagement et ont consacré plus de temps pour répondre aux questions lorsque l'agent virtuel présenté avait une expression faciale en colère (que quand l'agent virtuel était neutre). de Melo et al. [2012] ont exploré la manière dont les émotions exprimées par les agents virtuels influent sur la prise de décision des personnes dans la négociation humain-agent. Ils

ont étudié les effets de la démonstration de la joie, de la tristesse, de la colère et de la culpabilité des agents virtuels sur la décision des personnes de contrer, d'accepter ou d'abandonner la négociation. Les résultats indiquent que le comportement des participants est très impacté par l'état émotionnel perçu des agents virtuels. Ils ont trouvé, par exemple, que les participants ont tendance à accepter de renégocier avec un agent souriant mais de refuser une renégociation avec un agent en colère. Les résultats ont indiqué aussi que les personnes concédaient plus à un agent en colère qu'à un agent heureux. La raison ici est que les personnes pensent que l'agent en colère est moins flexible et qu'il peut facilement quitter la négociation, donc les participants sont obligés de concéder ; contrairement à un agent souriant qui semble avoir de faibles aspirations et les participants se permettre donc d'être stratégiquement plus exigeants.

Le travail de Mehrabian [1968] montre l'importance majeure de la modalité non-verbale dans la communication des sentiments. En effet, une formule précise a été proposée, dans cette étude, pour quantifier la contribution de chaque composante de la communication. 7% de l'information sur notre état affectif est communiquée par les mots, 38% est communiquée par la voix et 55% par les expressions faciales et le langage corporel. Cette étude est très importante et montre que pour avoir une modélisation complète des émotions, il est crucial d'inclure la modalité visuelle dans les systèmes et les supports de communication.

Dans l'approche par catégorie, Ekman [1979] dresse une liste de six émotions qu'il qualifie d'émotions "basiques" : la joie, la tristesse, la colère, le dégoût, la surprise et la peur. Ces émotions sont parfois dites simples dans le sens où elles ne sont pas composées d'autres états affectifs. Contrairement aux émotions complexes comme le remord qui, par exemple, est un mélange de dégoût et de tristesse, et la honte qui est un mélange de peur et de colère retournée contre soi. La théorie des émotions basiques capture ce qui est unique dans les émotions et ce qu'elles ont en commun qui les distinguent des autres états affectifs. Ekman and Cordaro [2011] présentent treize caractéristiques trouvées dans la plupart des émotions basiques. L'une des plus importantes caractéristiques est l'aspect distinctif et universel des signaux de chacune de ces émotions. Cette approche par catégories, permet de modéliser les émotions de manière discrète, en considérant chaque émotion comme une classe indépendante et isolée des autres états affectifs.

D'un autre côté, les approches dimensionnelles considèrent les émotions comme un continuum ou une transition progressive. En cartographiant les émotions dans un espace défini, les relations entre les émotions peuvent être capturées. Les émotions peuvent être représentées dans un espace continu multidimensionnel comme dans le modèle circumplex de Russell [1980] ou la roue de Plutchik [1984] présentée dans la figure 3.1. Cette modélisation permet de mieux refléter la complexité et les variations des expressions contrairement au système de catégories. Par exemple, un espace tridimensionnel, suffisant pour décrire tout état émotionnel, a été suggéré par Osgood et al. [1957] et a été défini par les trois axes : l'évaluation *(evaluation)*, la puissance *(potency)* et l'activité *(activity)*. Dans une étude de Davitz [1964], la parole expressive a été corrélée à ces axes et certains paramètres acoustiques ont pu être associés aux catégories émotionnelles de l'étude. Dans des recherches récentes sur la parole et les émotions, l'espace tridimensionnel se compose souvent de l'excitation *(arousal)*, de la valence *(valence)* et du contrôle *(control)*, comme le suggère Schröder [2003].

Dans la perspective de modéliser les émotions pour la synthèse expressive audiovisuelle de la parole, il faut prendre en compte l'impact de l'expressivité sur l'articulation. Les expressions faciales imposent un effort de compensation et de réorganisation de l'articulation. En réalité, les stratégies articulatoires mises en oeuvre par les locuteurs pour faire face à la production concomitante de la parole et des expressions faciales entraînent parfois à la résolution d'instructions contradictoires. Par exemple l'étirement des lèvres pour sourire lors de la production de voyelles ou de consonnes arrondies. Bailly et al. [2008] ont montré que les émotions étudiées (joie et dé-

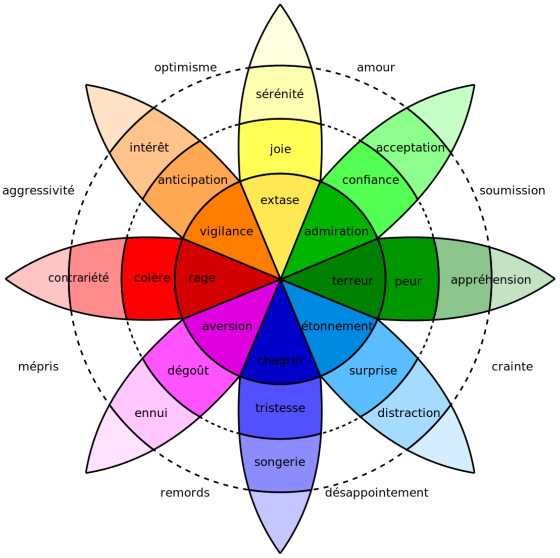


FIGURE 3.1 – *La roue des émotions de Plutchik [1984].*

goût), perturbent considérablement les mouvements de certains articulateurs (lèvre et mâchoire inférieure) durant la parole. Cet article [Bailly et al., 2008] conclu que ces perturbations ne sont pas simplement additives, et qu'elles dépendent de l'articulation. De ce fait, l'ajout de certaines expressions spécifiques à une émotion à des données visuelles neutres n'est pas suffisant pour modéliser correctement la parole visuelle dans un contexte expressive. Fónagy [1976] et Nordstrand et al. [2004] ont aussi étudié les effets de l'expressivité sur les mouvements articulatoires, et ont trouvé que les effets des émotions sur l'articulation différent d'une émotion à l'autre. Mais encore, il n'est pas clair comment les gestes liés à la parole sont combinés dynamiquement avec les expressions faciales pour assurer des apparences naturelles, réalistes et cohérentes. Il n'y a pas de fondement théorique pour déterminer les interactions et les contributions relatives de ces deux catégories de mouvements qui, ensemble, provoquent la totalité des déformations faciales. Certains chercheurs [Kshirsagar et al., 2001, Cao et al., 2005, Shaw and Theobald, 2016] ont utilisé des techniques telles que l'analyse en composantes indépendantes (ICA) et l'ACP pour séparer et modéliser les visèmes et l'espace des expressions faciales. Cependant, aucun des deux n'est basé sur les bases théoriques pour déterminer les interactions et les contributions relatives à ces deux catégories de mouvements.

Dans le travail de cette thèse, nous nous intéressons aux manifestations visibles des émotions sur les muscles faciaux (expressions faciales et articulation) et les changements que cela génère sur la voix. Dans le chapitre précédent (chapitre 2), nous avons présenté les différentes approches de la littérature pour la synthèse de la parole neutre. Dans ce chapitre, nous expliquerons les techniques d'extension de ces approches vers des systèmes expressifs. Les approches visant à ajouter de l'expressivité à la parole synthétique ont considérablement changé au cours des 30 dernières années [Schröder, 2001]. Les premiers systèmes de synthèse expressives étaient axés sur des modèles de «contrôle explicite», d'autres se basent sur plusieurs corpus expressifs préenregistrés. Aujourd'hui, les approches de «contrôle implicite» permettent de contrôler l'expressivité en combinant et en interpolant des modèles pré-entraînés sur différentes bases de données expressives. Le présent chapitre donne un aperçu des approches présentes dans la littérature pour

la synthèse acoustique, visuelle et audiovisuelle expressive.

3.2 Modélisation explicite des émotions

Les approches explicites de la modélisation de l'expressivité se basent sur la transformation de la parole neutre en une parole expressive en suivant les règles établies pour chaque émotion cible. Les systèmes expressifs à base de règles nécessitent l'adaptation ou la création de nouvelles règles à chaque ajout d'une nouvelle émotion.

La synthèse par formants, qui est la technique de synthèse la plus ancienne, permet de contrôler plusieurs paramètres, y compris les aspects glottaux et supraglottaux de la production de la parole. Beaucoup de ces paramètres sont potentiellement pertinents pour la modélisation de la parole expressive. Les premiers systèmes de synthèse vocale expressive ont été créés à partir du synthétiseur de formants commercial DECTalk : Affect Editor [Cahn, 1990] et HAMLET [Murray and Arnott, 1995] ont ajouté des modules spécifiques aux émotions à ce système. Pour chaque catégorie d'émotion, ils ont mis en place un modèle explicite de synthèse acoustique, en s'appuyant sur la littérature existante et en affinant les règles dans une procédure d'essais et erreurs. Dans une étude plus récente, Burkhardt and Sendlmeier [2000] ont utilisé la synthèse de formants pour faire varier systématiquement les paramètres acoustiques afin de trouver des valeurs optimales pour un certain nombre de catégories d'émotions. Au lieu de dériver les règles acoustiques de la littérature, des expériences perceptives ont été menées pour trouver des valeurs optimales pour les différents paramètres. Bien qu'un succès partiel ait été obtenu, un aspect naturel réduit a été signalé en raison des règles imparfaites utilisées.

Dans l'approche de synthèse par diphones, un exemple de chaque diphone est enregistré avec une voix neutre et un pitch monotone. Au moment de la synthèse, les paramètres prosodiques (la durée des unité et le contour F0) sont modifiés avec des techniques de traitement de signal (MBROLA [Dutoit et al., 1996]). Dans la plupart des systèmes de synthèse par diphone, uniquement la F0 et la durée peuvent être contrôlées. Mais les études montrent que ces deux paramètres ne sont pas suffisants pour exprimer les émotions. Des résultats très différents pour les systèmes par diphones ont été rapportés dans la littérature. Quelques travaux [Vroomen et al., 1993, Edgington, 1997, Montero et al., 1999] ont montré que les émotions synthétisées sont plutôt bien reconnues alors que d'autres [Heuft et al., 1996, Rank and Pirker, 1998] rapportent des taux de reconnaissances non significatifs.

Bevacqua and Pelachaud [2004] proposent un algorithme pour la synthèse visuelle expressive basé sur les règles. Cet algorithme détermine les visèmes appropriés en appliquant des règles de coarticulation pour tenir compte des contextes acoustiques ainsi que des phénomènes musculaires tels que la compression des lèvres et l'étirement des lèvres. Les cibles associées aux voyelles et aux consonnes ont été extraites des données réelles, les cibles consonantiques sont ensuite modifiées en fonction des contextes vocaux pour simuler l'effet de la coarticulation (notamment pour les consones bilabiales). L'algorithme simule aussi le degré d'influence des voyelles sur les consonnes. La simulation des comportements musculaires spécifiques aux émotions (par exemple, les lèvres tendues de la colère) a été intégrée au modèle à travers un ensemble de règles. Ce modèle de mouvement des lèvres a été appliqué sur un modèle facial 3D. Dans un autre travail, Beskow and Nordenberg [2005] ont étendu le modèle articulatoire de Cohen and Massaro [1993] pour créer un système de synthèse visuelle en entraînant des modèles articulatoires pour chacune des dix composantes principales extraites et pour chacune des cinq émotions étudiées (joie, colère, surprise, tristesse et neutre). Cette expérience a rapporté un taux de reconnaissance significatif pour la joie, la colère et la tristesse. Toujours à partir du modèle de Cohen and Massaro [1993],

Wu et al. [2006] ont généré des animation avec des visèmes coarticulés sous l'effet de l'émotion. Six messages écrits porteurs d'une charge émotionnelle et six animations expressives de synthèses ont été présentés aux évaluateurs dans un ordre aléatoire. Ces derniers ont été capables de trouver quel texte correspond à quelle animation avec 85% de bonnes réponses.

Un système de synthèse audio-visuelle a été proposé par Tang et al. [2008a,b]. Ce système combine une synthèse acoustique par diphone avec une synthèse visuelle par interpolation d'images clés. Ce système utilise une simple combinaison linéaire à poids égaux entre les visèmes relatifs à la parole et ceux relatifs à l'expressivité. Cette combinaison suppose que les mouvements articulatoires et expressifs sont simplement additifs, ce qui est une hypothèse qui s'avèrent beaucoup trop simpliste et conduit à de mauvais résultats (par exemple, bouche constamment ouverte durant la parole sous l'effet du sourire) [Bailly et al., 2008].

3.3 Modélisation implicite des émotions

Le système de parole expressive basé sur une modélisation implicite de l'expressivité ne nécessite pas une intervention humaine pour établir les règles spécifiques à chaque émotion. Ces règles sont implicitement incluses dans les données ou les modèles statistiques appris sur ces données.

3.3.1 Modélisation discrète des émotions

Le but des premiers travaux de synthèse expressive était de modéliser les émotions basiques telles quelles sont enregistrées dans les bases de données. Les systèmes présentés dans ces travaux sont capables de générer uniquement les émotions préenregistrées de manière isolée et sans possibilité de générer des émotions d'intensités variables ou des mélanges d'émotions.

Les systèmes de synthèse par concaténation d'unités de la parole dans un corpus furent les premiers systèmes de synthèse expressive discrète. Contrairement à l'approche par diphones, ce type de synthèse s'appuie sur une grande base de données. Si toutes les unités de la parole sont disponibles dans la base de données, aucune modification du signal original n'est nécessaire pour obtenir un résultat très naturel. Concernant la parole expressive, ce point fort de l'approche devient son point faible. En fait, uniquement les émotions préenregistrées dans la base de données peuvent être générées, il devient donc nécessaire d'enregistrer une nouvelle base de données pour chaque émotion cible. Iida and Campbell [2003] ont enregistré une base de données vocales pour chacune des trois émotions : joie, colère et tristesse. Au moment de la synthèse, et selon l'émotion choisie, les unités sont sélectionnées dans la base de données correspondante uniquement. Les émotions générées par cette méthodes ont donné un bon taux de reconnaissance des émotions de synthèse (50-80%). Dans une méthode similaire Johnson et al. [2002] ont présenté un système de synthèse vocale pour générer un discours militaire expressif convaincant. Les styles enregistrés contiennent des commandes criées, des conversations criées, des commandes parlées et des conversations parlées. Dans le même esprit, Pitrelli et al. [2006] a enregistré une grande base de données de parole neutre contenant onze heures de données acoustique ainsi que plusieurs bases de données acoustique de parole expressive relativement plus petites (1 heure de données acoustique pour chaque base de données). Au lieu de sélectionner des unités d'une base de données particulière au moment de la synthèse, ils ont fusionné toutes les bases de données ensemble et ont sélectionné des unités de cette grande base de données selon certains critères. Ils ont supposé que de nombreux segments d'une phrase, prononcée de manière expressive, pouvaient provenir de la base de données neutre. Cette approche a donné des résultats intéressants.

Concernant la modalité visuelle, un système de synthèse expressive visuelle par concaténation de disèmes (équivalent visuel du diphone) a été proposé par Henton and Litwinowicz [1994]. Ce système permet de générer des animations expressives de plusieurs personnages 3D on utilisant des disèmes préenregistrés dans plusieurs états émotionnels.

Les systèmes par HMMs ont eu beaucoup de succès dans le domaine de la synthèse expressive de la parole. Le système "voice puppetry" proposé par Brand [1999] est basé sur des HMMs et permet de générer une animation vidéo à partir d'une séquence audio. La vidéo contient une animation complète du visage qui correspond à l'état émotionnel contenu dans la séquence acoustique. Ding and Pelachaud [2015] proposent un système hybride, basé sur des GMMs et des HMMs pour synthétiser en temps réel les animations des lèvres qui parlent et qui rient pour les agents virtuels. Ce modèle de synthèse d'animation labiale prend en entrée la décomposition d'un texte parlé en phonèmes ainsi que leurs durées respectives pour générer des paramètres visuels sous formes de FAPs (Facial Action Parameters). Quatre paramètres ont été pris en compte pour évaluer objectivement leur modèle : l'ouverture des lèvres (distance entre le milieu de la lèvre supérieure et le milieu de la lèvre inférieure), étirement des lèvres (distance entre le coin gauche et droit des lèvres), protrusion de la lèvre supérieure et protrusion de la lèvre inférieure. Les résultats de cette étude objective ont prouver l'efficacité de cette méthode.

Quelques travaux ont adopté des architectures DNNs pour modéliser quelques catégories d'émotions. Dans l'étude de Xue et al. [2018a], trois approches de synthèse acoustique expressive par DNNs de type LSTM ont été proposées comme présenté dans la figure 3.2. La première (a) consiste en un réentrainement d'un modèle neutre, avec des données relatives à une émotion donnée, de cette manière ils obtiennent plusieurs modèles DNNs spécialisés chacun dans une émotion donnée. La deuxième approche (b) augmente le vecteur d'entrée avec les étiquettes (codes) des émotions, et la troisième approche (c) utilise une couche de sortie indépendante pour chaque émotion tout en gardant les couches cachées partagées entre les différentes classes d'émotions. Les évaluations objectives et subjectives ont montré que la troisième approche surpasse les deux autres approches avec une synthèse expressive plus naturelle.

Parker et al. [2017] et Filntisis et al. [2017] ont tous les deux utilisé des DNNs de type Feed-Forward pour la synthèse audiovisuelle expressive de la parole. Les deux systèmes ont obtenus des résultats subjectifs satisfaisants et ont montré que la qualité des résultats des systèmes de synthèse à base de DNNs dépasse significativement celle des systèmes par HMMs.

Li et al. [2016a] comparent plusieurs architecture de DNNs de types BLSTM pour l'adaptation d'un modèle entraîné sur un corpus neutre de grande taille avec une petite quantité de données expressives. Les cinq systèmes proposés génèrent une animation visuelle expressive à partir d'un fichier audio. Les résultats des expériences objectives et subjectives sur les données visuelles expressives générées montrent que le système le plus performant est celui qui prend en entrée un vecteur résultant de la concaténation entre les paramètres acoustiques et des données visuelles neutres.

3.3.2 Modélisation continue des émotions

Actuellement, diverses approches sont envisagées pour accroître la flexibilité de la dimension expressive tout en maintenant la qualité des systèmes de synthèse. Les émotions peuvent également être représentées dans un espace continu multidimensionnel comme dans le modèle circumplex de Russell [1980]. Cette modélisation permet de mieux refléter la complexité et les variations des expressions, contrairement au système par catégories. La modélisation continue des émotions permet de contrôler l'expressivité en combinant et interpolation entre les modèles entraînés sur différentes bases de données expressives.

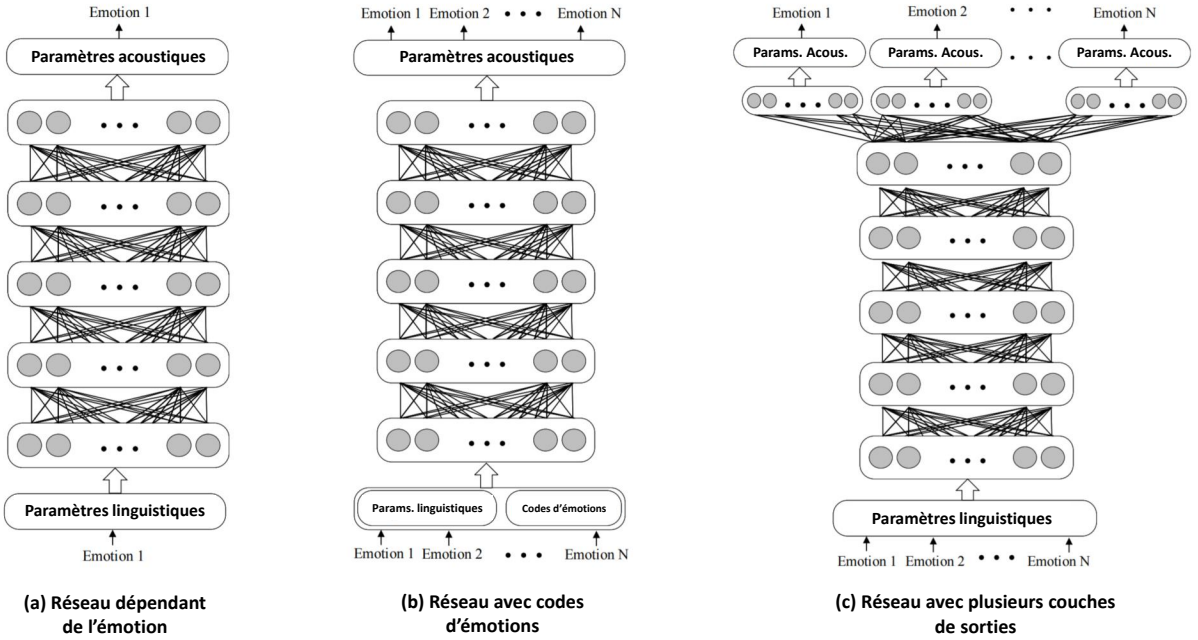


FIGURE 3.2 – *Figure tirée de l'article deXue et al. [2018a] représentant leurs trois approches pour la synthèse acoustique expressive de la parole par DNNs.*

Synthèse acoustique

Les systèmes de synthèse expressive par HMMs ont fait preuve d'une très grande flexibilité. Grâce au travail de Yoshimura et al. [1999], un système de synthèse de la parole dans lequel le spectre, le pitch et la durée des états HMMs sont modélisés simultanément dans un même modèle HMM, a été présenté. La motivation derrière la création de ce système vient du fait que pour changer arbitrairement la voix et le style et/ou l'émotion de la parole synthétique en conservant son naturel, il faut contrôler les caractéristiques prosodiques et spectrales simultanément en tenant compte de la relation entre le spectre et la prosodie puisqu'ils sont plus ou moins liées. Cette technique a été reprise dans les travaux de Yamagishi et al. [2003a, 2004], Tachibana et al. [2004]. Dans ces travaux deux méthodes de modélisation des styles de la parole acoustique ont été comparées. Dans la première méthode, les différents styles de la parole ont été modélisés de manière séparée, chacun avec un modèle HMM indépendant. Dans la deuxième méthode tous les styles ont été modélisés par un même modèle HMM en considérant des labels différents pour chaque style. Les résultats ont montré que les deux modèles se valent et qu'ils ont les mêmes performances en termes de modélisation et de naturel des styles synthétisés. Ils ont aussi étudié une méthode de synthèse de styles intermédiaires en appliquant une technique d'interpolation des modèles HMMs. Une technique d'adaptation des modèles HMMs a aussi été présentée dans Yamagishi et al. [2004] pour générer un style en utilisant une petite quantité de données d'apprentissage. Le travail de El Haddad et al. [2015] propose de contrôler l'intensité du sourire dans la parole acoustique synthétique par interpolation des modèle HMMs. Après la création d'un modèle pour la parole neutre et un autre pour la parole avec un sourire, une interpolation pondérée a été effectuée pour générer de la parole synthétique avec différents degrés de sourire.

Une approche différente pour l'interpolation des modèles acoustiques expressifs et leurs in-

tensités a été proposée par Masuko et al. [2004]. Cette approche permet de contrôler le degré d'expressivité et le style de la parole en utilisant un vecteur appelé le "vecteur de contrôle de style". Les vecteurs de contrôles sont représentés dans un espace bi-dimensionnel où chaque style est défini par ses propres coordonnées : lecture = (0, 0), sévère = (0, 1), heureux = (1, 0), triste = (−1, 0). À partir des résultats des tests subjectifs, ils ont réussi à contrôler les styles en choisissant le vecteur de contrôle de style de manière appropriée. De plus, il est possible de générer n'importe quel style de parole synthétique souhaité en spécifiant le vecteur de contrôle de style qui représente un point dans l'espace des styles.

Synthèse Audiovisuelle

Shaw and Theobald [2016] ont étudié des données visuelles extraites par l'algorithme AAM *(Active Appearance Model)* [Cootes et al., 2001] d'un corpus de vidéos expressives. En supposant que la parole expressive est une combinaison linéaire entre les expressions d'émotion et des mouvements articulatoires, le but de ce travail était de créer une représentation dans laquelle les mouvements relatifs à la parole et ceux relatifs à l'expressivité puissent être manipulés séparément. Les données visuelles extraites par AAM sont transformées en données de dimension inférieure par une ACP pour faciliter les calculs. Un seul modèle ICA est alors construit pour toutes les émotions de la base de données. La création du modèle ICA suppose que le contenu phonétique de toutes les émotions de la base soit similaire à celui de l'état neutre pour qu'un alignement et qu'une estimation des paramètres de l'ICA soient possibles. Ils définissent également un ensemble d'opérations d'édition qui peuvent être effectuées sur les données dans l'espace linéaire défini par l'ICA. De ce fait, cette méthode permet de générer des vidéos avec différentes émotions mais permet également de générer de nouveaux styles expressifs en naviguant dans le spectre expressif du modèle ICA.

Jia et al. [2010] présente une approche de synthèse expressive audiovisuelle. Les émotions sont représentées dans un espace tridimensionnel dans lequel chaque émotion est décrite et quantifiée selon trois dimensions : Plaisir – Déplaisir *(**P**leasure–Displeasure)*, Excitation – Non Excitation *(**A**rousal–Non Arousal)*, et Dominance – Soumission *(**D**ominance–Submissiveness)* notés PAD. D'abord, le texte à synthétiser et les valeurs PAD cibles sont donnés comme entrées au système, ensuite, en utilisant le moteur de synthèse, la parole neutre est générée. La parole neutre est par la suite transformée en parole expressive en utilisant des GMMs *(Gaussian Mixture Model)*. Les modèles GMMs préalablement entraînés génèrent un vecteur représentant la différence entre les paramètres acoustiques de l'état neutre et l'état émotionnel cible. En se basant sur ce vecteur de différence, l'algorithme TD-PSOLA est utilisé pour modifier le pitch et les durées de la parole neutre pour obtenir une parole expressive. Les expressions faciales sont générées à partir des paramètres PAD et des paramètres acoustiques sous formes de FAPs du standard MPEG-4 et en visèmes puis transformés en animation 3D.

Wan et al. [2013] et Anderson et al. [2013] ont proposé une méthode pour générer des animations audiovisuelles expressives à partir du texte en utilisant la méthode du CAT (Cluster Adaptive Training) [Gales, 2000] des modèles HMMs. En règle générale, dans les HMMs, un arbre de décision est utilisé pour regrouper et sélectionner les quinphones. L'intérêt du CAT est l'utilisation de plusieurs arbres de décision pour capturer des informations dépendantes de chaque émotion, chaque cluster a son propre arbre de décision. Dans le travail de Wan et al. [2013] le CAT et l'algorithme AAM a été adopté pour la paramétrisation des données visuelles et permet de modéliser des expressions de différentes émotions, ainsi que de générer des combinaisons d'émotions. Ce même travail a été repris par Parker et al. [2017], en remplaçant les modèles CAT-HMMs par un DNN de type Feed-Forward. Le DNN transforme les paramètres

linguistiques en paramètres acoustiques et visuelles. Les paramètres acoustiques et visuels sont modélisés conjointement. Les auteurs de ce papier affirment qu'il existe une grande corrélation entre la parole acoustique et visuelle, et qu'en les modélisant ensemble dans un modèle unique, les informations mutuelles entre la parole acoustique et visuelle peuvent être exploitées. Afin de produire un discours expressif, toutes les émotions sont modélisées ensemble dans un seul DNN avec plusieurs sorties, une par émotion. Ainsi, toutes les couches du DNN, à l'exception de la couche de sortie, sont partagées entre les différentes classes d'émotions et bénéficient d'un entraînement sur l'ensemble du corpus expressif. De plus, ce travail présente une méthode d'adaptation du DNN préalablement entraîné pour inclure une nouvelle expression en utilisant une petite quantité de données d'apprentissage. Les expériences montrent que le système de synthèse basé sur les DNNs est préféré de 57,9% par rapport au à celui basé sur les CAT-HMM.

Filntisis et al. [2017] présentent également une comparaison entre les résultats des HMMs et ceux des DNNs pour la synthèse audiovielle expressive. Un système basé sur des DNNs de type Feed-Forward est évalué et comparé à un autre système basé sur les HMMs, et à un autre basé sur la sélection d'unités concaténative, à la fois sur le réalisme et l'expressivité de la tête parlante générée. Les résultats montrent que les résultats du système par DNNs surpasse significativement les résultats des HMMs et par concaténation d'unité, que ce soit pour le réalisme ou l'expressivité de l'animation audiovisuelle générée. Ce travail traite également l'adaptation des modèles HMMs pour générer des animations audiovisuelles expressives à partir d'un modèle neutre pré-entraîné sur un grand corpus et un petit corpus expressif. Les auteurs affirment avoir réussi à générer des séquences expressives acoustiques et visuelles de bonne qualité. Ils ont également montré la possibilité de générer différents niveaux d'intensités pour les émotions ainsi que des styles de parole intermédiaires par interpolation des modèles HMMs.

3.3.3 Approches non-supervisées

Les approches couvertes dans les sections précédentes reposent majoritairement sur un contrôle conçu manuellement ou appris de manière supervisée à partir de données annotées, car pour entraîner un système de synthèse, les labels des émotions sont nécessaires. Cependant, l'annotation et la préparation des labels des émotions est une tâche très laborieuse et sujette à l'erreur. En fait, les bases de données sont labélisées par des annotateurs humains qui peuvent se tromper ou avoir des avis divergents. De plus, lorsque les émotions sont regroupées sous un nombre limité de catégorie émotionnelle, la notion de nuance et de graduation des émotions peut être perdue.

De ce fait, une nouvelle architecture de DNN est de plus en plus adoptée pour modéliser les émotions avec des bases de données non-annotées ou partiellement-annotées. Ces approches sont appelées des approches non-supervisées ou semi-supervisées respectivement. La plupart des travaux sur la synthèse expressives non-supervisés s'intéressent à l'aspect acoustique uniquement, pour le moment très peu de travaux traitent de la synthèse visuelle ou audiovisuelle expressive et non-supervisée qui reste encore un domaine pas suffisamment exploré.

Aujourd'hui, les architectures encodeur-décodeur permettent de générer des représentations latentes des émotions. Ces architectures sont souvent établies pour séparer les différentes informations contenues dans les données d'entraînement. Elles ont d'abord été utilisées dans le domaine de la reconnaissance automatique de la parole et plus précisément pour l'adaptation des locuteurs dans [Abdel-Hamid and Jiang, 2013]. L'architecture encodeur-décodeur utilisée est capable de transformer les caractéristiques de chaque locuteur en un espace de caractéristiques génériques et indépendants du locuteur (espace latent) nous parlons alors de vecteurs de plongement ou *embedding vectors*. L'adaptation à un nouveau locuteur peut se faire simplement en apprenant le code (représentation latente) du nouveau locuteur. Cette technique a été adaptée

pour la synthèse de la parole avec plusieurs locuteurs, d'abord par Luong et al. [2017] qui ont créé un système de synthèse acoustique avec un contrôle sur l'identité du locuteur et plus récemment par Gibiansky et al. [2017] et Taigman et al. [2017]. Avec cette technique, il est possible d'interpoler de façon linéaire les codes des locuteurs en changeant progressivement les valeurs du vecteur d'un locuteur vers celles d'un autre. Cela permet de réaliser l'interpolation des locuteurs. Luong et al. [2017] ont également trouvé qu'en manipulant manuellement les paramètres du vecteur du locuteur, certaines caractéristiques de la voix synthétique peuvent être altérées (l'âge et le genre du locuteur). An et al. [2017] se sont inspiré de la technique du code du locuteur utilisée par Huang et al. [2016] pour créer le code d'émotion. En utilisant les labels des émotions pour entraîner les codes des émotions, An et al. [2017] ont réussi a normaliser et unifier ces derniers dans un espace latent unique, permettant ainsi de contrôler efficacement le type et l'intensité de l'émotion dans la parole synthétique.

Henter et al. [2017] ont décrit comment les nuances des émotions peuvent être apprises pour la synthèse vocale avec une base de données non-annotée en degré ou nuance des émotions. Dans ce travail, uniquement les labels des émotions sont utilisés pendant l'entraînement d'un DNN de type BLSTM. La particularité de ce système est qu'il utilise à la fois des observations annotées (étiquettes des classes des émotions) et non-annotées (étiquettes des nuances des émotions) afin d'effectuer la synthèse acoustique expressive. L'apprentissage des nuances non-annotées des émotions est réalisé dans un espace de vecteurs latents. Les expériences effectuées confirment que ce système de synthèse est capable de générer des émotions tout en permettant d'ajuster leurs degrés d'intensités.

Dans le travail de Watts et al. [2015b] un système de synthèse vocale expressive à partir du texte a été proposé. Ce système à base de DNN-FF a été entraîné sur des données provenant de livres audio. Contrairement aux approches classiques, les labels des émotions n'ont pas été fournis durant l'entraînement. Des vecteurs bi-dimensionnels, appelés vecteurs de contrôle, ont été appris de façon non-supervisée pour absorber les informations relatives à l'état émotionnel pour chaque phrase. Ils ont montré, qu'en ajustant les paramètres des vecteurs de contrôle, les paramètres prosodiques des phrases générées sont modifiés de manière simple et robuste durant la phase de synthèse.

Récemment, les autoencodeurs variationnels (VAEs) [Kingma and Welling, 2013], sont de plus en plus convoités dans le domaine de la synthèse expressive de la parole. Les VAEs représentent une architecture plus complète pour l'apprentissage des variables latentes. Ces architectures se basent sur des réseaux de neurones pour apprendre à la fois la façon dont les observations dépendent des variables latentes, ainsi que la façon d'inférer les distributions des variables latentes à partir des observations. Les VAEs sont considérés comme des auto-encodeurs car le processus d'inférence peut être considéré comme un encodage d'une observation en une variable latente tandis que la génération peut être considérée comme un décodage de cette variable latente vers le domaine initial des observations. De plus, les deux processus (encodage et décodage) peuvent être appris conjointement par descente de gradient dans une architecture unique [Doersch, 2016].

Le premier travail considérant les VAEs pour la synthèse vocale expressive est le travail de Akuzawa et al. [2018]. Dans cet article, une architecture VAE a été combinée avec VoiceLoop [Taigman et al., 2017], afin de permettre à ce modèle de synthèse autorégressif d'être plus expressif. La méthode proposée peut modéliser les émotions dans le processus de synthèse de la parole d'une manière non-supervisée.

Dans Henter et al. [2018] les méthodes supervisées et non supervisées pour l'apprentissage des modèles acoustiques contrôlables sur un large corpus expressif ont été comparées. Les résultats objectifs et subjectifs montrent que les méthodes non supervisées apprennent et reproduisent avec succès les classes d'émotions de la base de données et arrive même à surpasser la méthode super-

visée. En plus des avantages qu'offrent les approches non-supervisées, ces résultats représentent un argument de plus en leur faveur pour la synthèse de parole expressive non-supervisée.

Les paramètres de contrôle sont souvent constitués de variables latentes et restent complexes à interpréter. Bien que les travaux cités ci-dessus montrent qu'il est possible de construire un espace latent conduisant à des variables pouvant être utilisées pour contrôler le style dans la synthèse vocale. Cependant, ces travaux ne fournissent pas d'informations sur les relations entre l'espace latent résultant et les caractéristiques acoustiques qu'il est possible de contrôler. L'article de Tits et al. [2019], propose une analyse de l'espace latent obtenu par un entraînement non-supervisé d'un VAE sur une base de données acoustique expressive pour la synthèse TTS. Cette analyse intéressante montre comment certains paramètres acoustiques changent de manière linéaire en traversant l'espace latent dans des directions bien choisies. Cette relation interprétable entre l'espace latent et les paramètres acoustiques est très adaptée pour construire des systèmes de synthèse vocale contrôlables avec un comportement compréhensible.

3.4 Conclusion

Dans ce chapitre, nous avons présenté les différentes approches de synthèse expressive acoustique et audiovisuelle de la parole dans la littérature. Nous avons d'abord abordé les approches explicites pour la modélisation des émotions, ces dernières se basent généralement sur la transformation de la parole neutre en une parole expressive en suivant les règles établies pour chaque émotion cible. Cependant, ces systèmes sont peu flexibles et nécessitent l'adaptation ou la création de nouvelles règles à chaque ajout d'une nouvelle émotion. Les approches de modélisation implicite de l'expressivité ont ensuite été exposées, que ce soit pour représenter les catégories des émotions de manière discrète seulement ou de manière continue. L'avantage de la représentation continue des émotions est sa capacité à mieux refléter la complexité et les variations des émotions humaines. Elle permet de contrôler l'expressivité en combinant et interpolation entre différentes émotions pour créer une panoplie d'états émotionnels mixtes ou intermédiaires.

La problématique des données annotées a également été abordée dans ce chapitre et nous avons décrit l'état de l'art sur ce sujet en présentant les approches non-supervisées. Les architectures encodeur-décodeur, notamment les VAEs, ont montré leur efficacité dans quelques travaux récemment publiés. Ces derniers permettent d'apprendre des représentations latentes des émotions sans avoir besoin des labels des émotions pendant la phase de l'apprentissage. Ces systèmes représentent l'état de l'art de la synthèse expressive audiovisuelle en ce moment puisqu'ils regroupent tous les avantages des techniques précédemment citées. Ils permettent d'avoir une modélisation implicite, continue, contrôlable et non-supervisée des émotions humaines. Bien que les travaux dans le domaine acoustique soient bien avancés, le domaine de la synthèse expressive audiovisuelle contrôlable et non-supervisée restent encore, pour le moment, pas suffisamment exploré. C'est dans ce cadre précis que se situe une contribution importante de ce travail de thèse.

Dans la partie suivante de ce manuscrit, nous exposons notre protocole d'acquisition, de traitement et d'analyse d'un corpus audiovisuel expressif qui sera utilisé plus tard pour la synthèse de la parole audiovisuelle expressive.

Deuxième partie

Corpus audiovisuels expressifs

4

Étude d'un corpus expressif

Sommaire

4.1 Introduction

Dans le contexte de la synthèse audiovisuelle expressive de la parole, la qualité des données utilisées dans l'entraînement des modèles est corrélée à la qualité de la parole de synthèse générée. De ce fait, il est important de s'assurer que les émotions du corpus sont bien perçues par les humains. De plus, l'entraînement d'un modèle de synthèse nécessite une base de données de taille conséquente, contenant au moins quelques heures de parole [Guennec, 2016]. Les bases de de données expressives existantes ne contiennent souvent que la modalité acoustique (SynPaFlex, AlloSat, PAVOQUE, etc). Pour les bases de données audiovisuelles, dans leur majorité, la modalité visuelle est sous forme d'enregistrements vidéos (GEMEP, CVSP-EAV, eNTERFACE'05, MSP-IMPROV, VAM-Video, SAVEE, MODALITY, etc). Bien qu'ils soient faciles et moins coûteux à enregistrer, dans ces enregistrement, l'information sur la profondeur de la scène est perdue. De ce fait, certains gestes liés à la parole, comme la protrusion des lèvres, ne peuvent pas êtres suivis/prédits avec précision. Heureusement, quelques bases de données audiovisuelles expressives contenant des données 3D existent. Par exemple, la base de données AV-LASYN [Cakmak et al., 2014] qui contient un corpus synchrone de données audio et de trajectoires de marqueurs faciaux 3D, cependant, cette base ne contient qu'une seule émotion et est dédiée pour la synthèse audiovisuelle du rire seulement. La base de données IEMOCAP [Busso et al., 2008],

quant à elle, contient des séquences audiovisuelles enregistrées avec des systèmes de capture de mouvement. Cette base de données contient des enregistrements de dix acteurs et plusieurs émotions : état neutre, colère, joie, excitation, tristesse, frustration, peur, surprise, etc. Toutefois, chaque locuteur n'a enregistré que 30 minutes de parole scriptée (toutes émotions confondues) ce qui est insuffisant pour entraîner des systèmes de synthèse de la parole. De plus, le nombre de phrases par émotions, n'est pas équilibré (neutre 28%, frustration 24%, excitation 17%, tristesse 15%, colère 7%, joie 7%, surprise 2%, dégoût 1%, les autres $< 1\%$) ce qui ne permet pas de comparer la performance des systèmes de synthèses pour les différentes classes d'émotions. La base de données Biwi 3D [Fanelli et al., 2010] propose des enregistrements audiovisuelles sous formes de séquences de scans 3D et d'audio. Ce corpus est très intéressant car il fournit une information complète sur la déformation du visage en entier (et pas qu'une sélection de points), mais il est aussi de petite taille (1109 phrases en totales, 14 locuteurs, environ 80 phrases par locuteur) et ne peut pas être utilisé dans un processus de synthèse.

Pour toutes les raisons évoquées plus haut, nous avons décidé d'enregistrer notre propre corpus qui répond à nos exigences :

1. Le corpus doit contenir une modalité acoustique et visuelle synchronisée,
2. La modalité visuelle doit être capturée en 3D,
3. Le corpus doit contenir plusieurs classes d'émotions,
4. Le corpus doit être équilibré : les classes d'émotions doivent avoir le même contenu linguistique et le même nombre de phrases.
5. Le corpus doit être suffisamment grand pour entraîner des modèles de synthèse de la parole (quelques heures).

Sachant, que l'enregistrement d'un corpus audiovisuel expressive d'une grande taille est une tâche laborieuse, nous avons décidé d'enregistrer un corpus prototype qui nous permettra de vérifier notre protocole d'acquisition et d'analyser son contenu avant d'enregistrer un corpus de grande taille. Ce dernier sera dédié à la synthèse comme nous le détaillerons dans les chapitres suivants.

Dans ce chapitre [3] nous présentons notre démarche pour enregistrer un mini-corpus expressif. Nous analysons ensuite le corpus obtenu pour vérifier sa qualité et nous assurer de son contenu émotionnel. Nous avons effectué deux analyses différentes : (1) une évaluation de la production de la parole expressive et (2) une évaluation perceptive. La première évaluation permet d'identifier les caractéristiques spécifiques à chaque contexte expressif. L'étude perceptive a été effectuée sous forme de tâche de reconnaissance des émotions par des humains en utilisant différents types de stimuli.

4.2 Description et acquisition du corpus

Ce corpus a été enregistré avec un acteur de 27 ans, français natif. Rappelons que l'objet de ce mini-corpus est de servir de corpus prototype pour la validation de notre protocole d'acquisition de données multimodales. De toute évidence, ce corpus ne peut pas être utilisé pour développer un système de synthèse de parole audiovisuelle, car nous avons probablement besoin de plus de 20 ou 50 fois la taille de ce corpus.

L'idée de combiner différents systèmes d'acquisition dans une plateforme d'acquisition multimodale, est née de notre volonté d'obtenir un corpus audiovisuel expressif 3D de haute qualité,

3. Les travaux présentés dans ce chapitre ont fait l'objet d'un article de revue qui a été accepté dans la revue internationale LREV.

en utilisant une technique bien adaptée pour chaque partie du visage. La capture de la dynamique des expressions faciales était d'une grande importance, mais l'obtention d'un mouvement d'articulation très précis était également cruciale. Dans une étude précédente, nous avons évalué la précision de ces trois systèmes. Nous avons découvert que le système AG501 a la précision temporelle et spatiale la plus élevée, suivi du système Vicon puis du système RealSense [Ouni and Dahmani, 2016]. De plus, nous voulons tester les différents systèmes de capture de mouvement pour envisager leur utilisation ou non pour un corpus dédié à la synthèse audiovisuelle.

4.2.1 Système d'acquisition multimodale

Notre plateforme d'acquisition multimodale est composée de (1) un système de capture de mouvement (VICON) utilisant des marqueurs optiques réfléchissants, (2) un articulographe (AG501) utilisant des capteurs électromagnétiques et (3) un système de capture de mouvement sans marqueurs physiques (Intel RealSense) :

1. VICON : Ce système se base sur des caméras infrarouges et des capteurs rétro-réflectifs. Dans notre configuration, nous avons à notre disposition 4 caméras (MX3+) avec des objectifs modifiés pour les courtes distances. Nous avons collé des capteurs de 9mm de diamètre sur un bonnet serré que l'acteur devait porter sur sa tête durant l'acquisition. Ces capteurs nous permettent d'obtenir des informations liées aux translations et rotations de la tête. Nous avons collé aussi des capteurs de 3mm de diamètre sur la partie supérieure du visage. Ces capteurs nous permettent de capturer les expressions faciales. Les caméras Vicon ont été placée à une distance d'environ 150 centimètres du locuteur. Le logiciel propriétaire Vicon Nexus nous a permis de suivre les capteurs sous forme de trajectoires de points 3D à une fréquence d'échantillonage de 100Hz.
2. Articulograph (AG501) : Les capteurs de l'articulograph permettent de traquer les mouvements des lèvres. La technologie utilisée dans le AG501 permet de suivre les gestes les plus fins des lèvres. Ce système est connu pour sa grande précision et pour sa capacité de traquer les mouvements de la partie interne des lèvres [Berry, 2011, Stella et al., 2013, Yunusova et al., 2009], surtout dans le cas de la fermeture de la bouche. En fait, la gestion de l'occlusion des lèvres n'est pas possible avec les systèmes à base de caméras, car les capteurs doivent toujours être dans le champ de vision de ces dernière [Ouni and Gris, 2018]. Le logiciel AG501 fournit également la position spatiale 3D de chaque capteur à une fréquence d'échantillonnage de 250Hz.
3. Intel RealSense (f200) : Les caméras RealSense sont équipées d'une caméra RVB et d'une caméra de profondeur. Les informations collectées par ces caméras utilisées par un algorithme de suivi des mouvements faciaux pour extraire la position en temps réel d'un certain nombre de points 3D du visage. Cette technologie de capture de mouvement est non invasive et nous a permis, dans ce travail, d'avoir un suivi de l'ouverture/fermeture des yeux. La fréquence d'échantillonage de ce système est de 50 Hz.

La figure 4.1 présente la plate-forme multimodale composée des systèmes AG501, Vicon et RealSense. Chaque système a des conditions d'utilisation et des contraintes de placement différentes. Le système EMA doit être placé à l'écart de tout équipement contenant des matériaux métalliques ferromagnétiques, tels que les caméras Vicon et les trépieds, pour éviter la déformation du champ électromagnétique. Le système Vicon doit être placé à une distance raisonnable de l'acteur pour pouvoir suivre les capteurs réfléchissants de 3 mm collés sur son visage. Concernant la RealSense, le capteur de profondeur a une courte portée (entre 20 cm et 120 cm) et doit donc être placé plus près que Vicon de l'acteur, avec un risque de couvrir le champ de vision

FIGURE 4.1 – *La plate-forme multimodale utilisée pour enregistrer les données.*

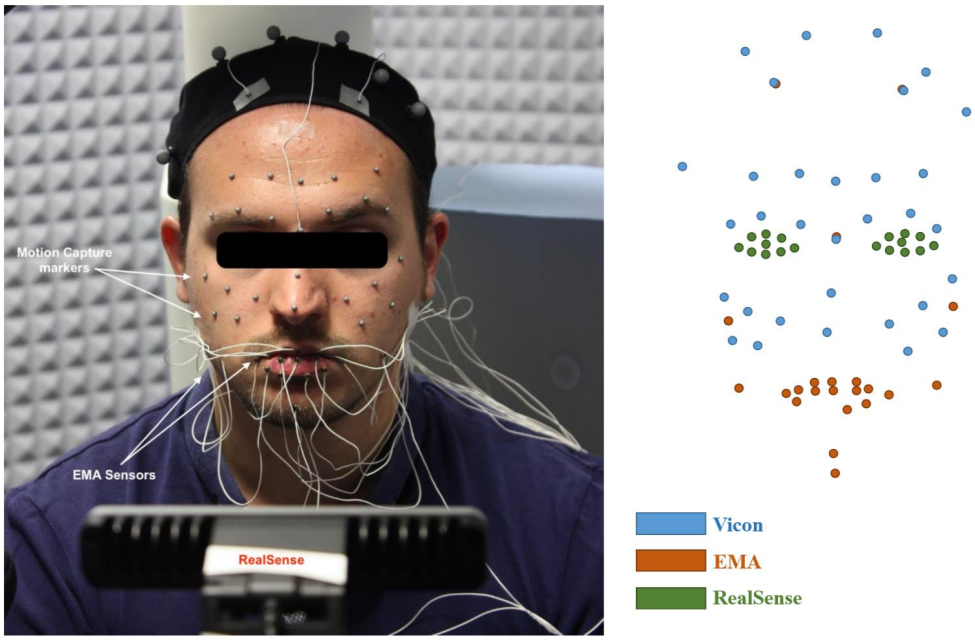


FIGURE 4.2 – *Les positions des marqueurs Vicon, EMA et RealSense sur le visage de l'acteur et la représentation minimaliste du visage obtenue après la fusion des données des trois systèmes.*

des caméras Vicon. Les fils du système EMA ne doivent pas masquer les marqueurs Vicon et ne doivent pas empêcher l'algorithme de suivi de la RealSense de reconnaître la forme du visage. De plus, les synchronisations spatiales et temporelles entre les différents canaux imposent des combinaisons supplémentaires et une disposition particulière des capteurs dont nous discuterons dans la sous section dédiée au post-traitement (4.2.3).

La figure 4.2 montre la configuration des marqueurs Vicon et EMA sur le visage, et la position de la RealSense devant l'acteur. La plupart des emplacements des marqueurs ont été inspirés de la norme MPEG-4 [Pandzic and Forchheimer, 2002]. Les capteurs EMA sont concentrés autour de la bouche. Les autres régions du visage sont complétées par des capteurs Vicon. Trois marqueurs Vicon et trois capteurs EMA ont été placés dans les mêmes positions (chacun au-dessus de l'autre, voir figure 4.3). Ces trois capteurs sont utilisés comme capteurs de référence pour calculer l'alignement spatial entre les deux systèmes. Nous avons placé cinq marqueurs supplémentaires sur le dessus de la tête pour supprimer le mouvement de la tête. On peut obtenir les coordonnées relatives de ces dernières, par soustraction des trajectoires des marqueurs de la tête. De cette façon, nous ne gardons que les mouvements faciaux. La figure 4.2 présente la disposition des marqueurs physiques et virtuels sur le visage de l'acteur.

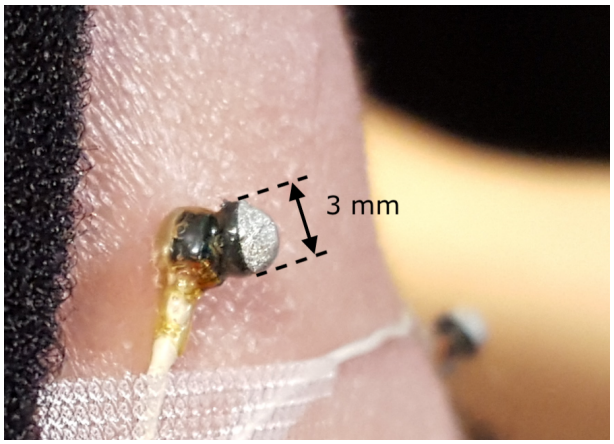


FIGURE 4.3 – *Un marqueur Vicon collé au dessus d'un capteur EMA. Ils sont utilisés comme points de référence pour fusionner les données (voir la section 4.2.3).*

La modalité acoustique a été acquise simultanément avec les données spatiales à l'aide d'un microphone cardioïde (Rode NT3) avec une fréquence d'échantillonnage de 48 kHz. Pour synchroniser le canal audio et le canal visuel, nous avons utilisé un appareil électronique fabriqué en interne. Il déclenche simultanément une lampe infrarouge capturée par les système Vicon et RealSense, et un son aigu généré par un vibreur piézoélectrique pour l'audio capturé par l'articulographe (AG501 a un déclencheur électronique intégré qui synchronise l'audio et les mouvements des capteurs). Nous avons développé plusieurs outils et techniques pour traiter et fusionner les données avec précision, qui seront présentés dans la section 4.2.3.

Avant d'enregistrer l'acteur, plusieurs préparatifs ont été faits. Le système Vicon et le système EMA ont été calibrés. Les marqueurs Vicon et les capteurs EMA ont été collés sur le visage de l'acteur. Les différents systèmes ont été testés avant de démarrer l'enregistrement. Nous avons également placé une caméra pour avoir une référence vidéo pour les données enregistrées.

4.2.2 Déroulement de l'acquisition

Nous avons demandé à un acteur semi-professionnel de 27 ans de prononcer des phrases dans sept états émotionnels différents (neutre, joie, surprise, peur, colère, tristesse et dégoût). L'acteur a utilisé la technique des *exercices de style* où il dissocie la sémantique de la syntaxe des phrases et joue les mêmes phrases dans des styles différents. Il se conditionne pour être dans un état émotionnel donné et prononce ensuite les différentes phrases sans prêter attention à leur sens. Les phrases ont été présentées, une par une, sur un écran positionné devant l'acteur. Dans ce contexte, les émotions doivent être considérées comme jouées car elles sont un peu exagérées comme dans le cas d'une pièce de théâtre. Nous avons fait ce choix car il a été observé dans le domaine des personnages animés [Bates et al., 1994] que l'expression des émotions humaines doit être exagérée pour les agents virtuels pour qu'elles puisse être convaincante. L'étude de Kätsyri et al. [2003] montre également que les expressions d'émotions sont moins bien identifiées pour un agent virtuel que pour un vrai visage.

Le mini-corpus, enregistré dans cette expérience, contient 30 phrases en langue française. Ces phrases sont de différentes longueurs : 10 phrases courtes, 10 phrases moyennes et 10 phrases longues. Les mêmes phrases sont utilisées pour enregistrer l'état neutre et chacune des six émotions. Le nombre total des phrases est de 210. Les phrases longues contiennent en moyenne 27 mots et durent environ 10 secondes chacune. Les phrases courtes ont une longueur moyenne de 4 mots et durent environ 1,3 seconde. Les phrases de longueur moyenne n'ont pas été utilisées dans les différentes analyses présentées dans ce chapitre.

4.2.3 Post-traitement

Les données visuelles ont été obtenues directement sous forme de coordonnées de points 3D pour les trois systèmes. Chaque système fournit des données 3D sur sa propre référence spatiale (position de l'origine et direction des axes). Pour ces raisons, il est nécessaire de résoudre le problème de la synchronisation temporelle, car chaque système a sa propre fréquence d'échantillonnage. Il est également nécessaire de définir un référentiel unique pour y fusionner les données des différents systèmes.

Synchronisation des données

Les trois systèmes d'acquisition des données 3D ont différentes fréquences d'échantillonnage : EMA (250Hz), Vicon (100Hz) et RealSense (50Hz). La fusion des données commence par l'unification de leurs fréquences d'échantillonnage. Nous choisissons de conserver la fréquence d'échantillonnage la plus élevée pour conserver une meilleure précision. Nous effectuons un suréchantillonnage, en utilisant une interpolation linéaire, sur les données de Vicon et RealSense pour atteindre 250Hz.

Pour synchroniser les différents flux de données, nous utilisons une méthode en deux étapes. Tout d'abord, nous utilisons les informations fournies par le déclencheur que nous avons fabriqué en interne. Ce dernier génère simultanément un spot lumineux infrarouge et un son aigu. Nous cherchons manuellement la première frame du signal de chaque flux : pour Vicon et RealSense, nous cherchons la première frame où apparaît le marqueur artificiel (représentant le spot infrarouge), pour AG501, et comme il a son propre déclencheur, nous cherchons la première frame acoustique où le son aigu apparaît. Cette technique offre un moyen raisonnable de synchroniser les différents flux. Cependant, le résultat peut être décalé de quelques frames, en raison de la différence de fréquence d'échantillonage de chaque système (même après le sur-échantillonage). Pour cette raison, nous avons combiné cette étape avec une seconde pour affiner la synchronisation.

Au cours de cette deuxième étape, qui est également manuelle, nous synchronisons spatialement les données Vicon et EMA, en utilisant les marqueurs de référence EMA et Vicon (collés dans les mêmes positions). Notre logiciel de visualisation Visartico [Ouni et al., 2012] nous permet de visualiser les trajectoires d'un marqueur donné. Nous avons choisi un marqueur Vicon au hasard, puis nous pouvons faire glisser interactivement sa trajectoire pour correspondre parfaitement à la trajectoire EMA correspondante. À ce stade, nous considérons EMA, Vicon et les données acoustiques synchronisées. Nous répétons ce processus pour les données RealSense. Comme nous n'avons pas de marqueurs de référence pour la RealSense, nous avons choisi un ensemble de marqueurs virtuels proposés par le logiciel de suivi facial RealSense pour les associer aux marqueurs des sourcils, de la bouche et des yeux. Nous visualisons à nouveau l'un des marqueurs de la RealSense utilisés pour l'alignement et nous essayons d'adapter sa trajectoire à celle d'un capteur EMA. Il est important de noter que le décalage est appliqué à tous les marqueurs du Vicon et de la RealSense même si un seul marqueur a été choisi pour déplacer les trajectoires.

Fusion des données

Étant donné que chaque système fournit les données 3D sur son propre repère de référence, l'étape suivante du post-traitement consiste à définir une frame de référence et à fusionner les données des différents systèmes dans cette frame. Pour ce faire, nous avons utilisé les marqueurs de référence EMA et VICON que nous avons collés au même endroit (voir figure 4.3). Trois marqueurs, différents de l'origine du repère spatial, sont nécessaires pour construire trois vecteurs non-planaires. Ensuite, nous calculons la translation et la rotation nécessaires pour que les données Vicon correspondent à la position des données EMA. Ces paramètres sont appliqués à l'ensemble des capteurs Vicon et à toutes les frames de l'enregistrement. Nous n'avons pas eu besoin de redimensionner les données, car les trois systèmes fournissent des données 3D à la même échelle.

Retirer les mouvements de la tête

Enfin, nous avons utilisé des marqueurs Vicon supplémentaires sur la tête de l'acteur et trois autres capteurs EMA (deux derrière les oreilles et un entre les yeux) pour supprimer les mouvements de la tête. Ce type de données sera utilisé dans la synthèse vocale audiovisuelle, où la suppression du mouvement de la tête est nécessaire pour pouvoir générer des expressions faciales et des gestes vocaux cohérents et non ambigus. Il convient de noter qu'il est toujours possible de réintégrer ces mouvements ultérieurement si nécessaire. Nous utilisons les marqueurs supplémentaires pour construire les vecteurs de transformation. La première frame de l'enregistrement est utilisée comme référence où la tête sera fixée à cette pose particulière. Ensuite, nous calculons les transformations de translation et de rotation, en fonction de la configuration de la première frame. Après cela, nous appliquons la transformation calculée frame par frame.

Comme nous pouvons le voir sur la figure 4.2, le résultat final consiste en une configuration de points 3D représentant le visage : la bouche, les joues, le nez, le contour des yeux, les pupilles des yeux, les sourcils et le front. Ainsi, les données multimodales consolidées sont représentées par un groupe de points 3D à une fréquence d'échantillonnage de 250Hz, pour toutes les phrases du mini-corpus.

4.3 Analyse de la production

Dans cette analyse, nous avons utilisé les mêmes phrases pour étudier les différentes caractéristiques émotionnelles visuelles et acoustiques. Nous n'avons utilisé que les phrases longues. Dans une étude précédente, nous avions analysé les données acoustiques et visuelles des phrases courtes dans des conditions similaires (en utilisant uniquement le système Vicon [Ouni et al., 2016]). Dans les sections suivantes, nous faisons référence aux résultats de cette étude. Dans ce chapitre, les deux types de phrases, courtes et longues, n'ont été utilisés que lors de l'évaluation perceptuelle. Les analyses ont été effectuées sur les données obtenues, composées de sept fichiers d'enregistrement (un pour chaque émotion), chaque enregistrement contenant dix phrases.

4.3.1 Analyse visuelle

Comme les données visuelles du corpus enregistré sont constituées de 48 points spatiaux 3D représentant un espace de grande dimension, l'analyse peut être longue et difficile à mener vu le nombre de dimensions à explorer. De plus, contrairement à l'étude des émotions dans un contexte statique (images ou vidéos sans parole), nous considérons la dynamique des émotions représentées par des séquences de parole. Dans ce cas particulier, nous devons analyser ces séquences pour extraire les mouvements les plus importants sans les analyser image par image. Pour ces raisons, nous avons effectué une analyse en composantes principales (ACP) sur les données afin de réduire le nombre de dimensions à analyser. Le corpus a été divisé en sept fichiers, chacun contenant dix phrases d'une émotion donnée. Nous avons appliqué l'ACP à chaque petit corpus pour identifier les directions principales du mouvement lorsqu'une émotion donnée est présente. Nous avons également calculé plusieurs mesures faciales (ouverture des yeux, mouvements des sourcils, ouverture de la bouche et étirement de la bouche). Le but est de quantifier les différentes variations trouvées avec l'analyse PCA et de les représenter avec les unités d'action du manuel FACS [Ekman et al., 2002]. Le FACS est le système de codage d'action faciale (mis en place par Ekman and Friesen [1978]) et décrit les mouvements faciaux par un ensemble d'unités d'actions (UAs), nous présentons dans l'annexe B la liste des différentes UAs accompagnées de leur description. En représentant nos résultats en termes de UAs nous serons en mesure de les comparer avec les autres études qui se basent sur ce même encodage.

Le tableau 4.1 montre le pourcentage de variance des cinq premières composantes principales (CPs) des différents états émotionnels. La variance est répartie sur les CPs , mais CP1 ne capture pas un geste dominant pour les différentes émotions (plus de 50% de la variation). Nous atteignons un pourcentage cumulé de variance supérieur à 50% avec les trois premières composantes. Dans une étude précédente sur les phrases courtes, ce taux a été atteint avec les 2 premiers composantes uniquement [Ouni et al., 2016].

Dans la figure 4.4, nous représentons la variation des trois premières CPs pour chaque émotion. Afin de représenter toute la variation pour chaque CP, nous avons représenté leur valeur minimale et maximale sur chaque figure. La déformation du visage est présentée lorsque la CP correspondante a une valeur de -3 (bleu) ou +3 (rouge) écarts-types (nous supposons qu'ils sont la limite inférieure et supérieure de la variation CP). Nous présentons dans le tableau 4.2 les différentes UAs sollicitées pour chacune des trois premières CPs de chaque émotion. L'état neutre représente des mouvements très légers, essentiellement l'ouverture des lèvres (CP1 avec UA25 et CP3 avec UA26) et la protrusion des lèvres (CP2 avec UA18 et UA22). Ces mouvements articulatoires sont importants pour la production de la parole. Cette variation ne concerne que la partie inférieure du visage. La partie supérieure, qui est généralement importante pour l'expression des émotions, bouge à peine.

Emotion	CP1	CP2	CP3	CP4	CP5
Neutre	20 (*20*)	16 (*36*)	12 (*49*)	8 (*57*)	6 (*63*)
Joie	34 (*34*)	18 (*52*)	7 (*60*)	6 (*66*)	5 (*72*)
Surprise	34 (*34*)	18 (*53*)	11 (*64*)	6 (*71*)	5 (*76*)
Colère	40 (*40*)	13 (*53*)	9 (*63*)	6 (*70*)	4 (*74*)
Peur	26 (*26*)	13 (*39*)	12 (*52*)	8 (*60*)	7 (*68*)
Tristesse	23 (*23*)	15 (*38*)	13 (*51*)	9 (*60*)	6 (*67*)
Dégoût	31 (*31*)	11 (*42*)	10 (*53*)	8 (*61*)	8 (*70*)

TABLE 4.1 – *Pourcentages de la variance des cinq premières composantes principales pour l'état neutre et les 6 émotions. Le nombre entre parenthèse est le pourcentage cumulé de variance.*

Émotion	Unités d'action		
	CP1	CP2	CP3
Neutre	UA25	UA18, UA22	UA26
Joie	UA1, UA2, UA5, UA6, UA9, UA12, UA26	UA12, UA18, UA22, UA26	UA26
Surprise	UA1, UA2, UA5, UA26	UA5, UA26	UA18, UA22, UA26
Colère	UA1, UA2, UA5, UA26	UA18, UA22, UA23, UA24, UA26	UA26
Peur	UA1, UA2, UA5, UA26	UA25	UA5, UA18, UA22, UA26
Tristesse	UA1, UA4, UA7, UA15, UA25	UA18, UA22	UA7, UA15, UA26
Dégoût	UA1, UA2, UA4, UA7, UA10, UA25	UA7, UA9, UA10, UA20, UA25	UA7, UA18, UA22, UA10, UA20, UA25

TABLE 4.2 – *Les trois composantes principale des données faciales pour chaque émotion et les unités d'action correspondantes.*

Pour la première composante, et contrairement aux autres émotions où la forme des yeux varie de grand ouvert (UA5) à légèrement fermé (UA7), pour la *surprise* et la *peur* les yeux sont grands ouverts en continu. L'autre caractéristique notable est que la *tristesse* et le *dégoût* se distinguent par la taille des yeux la plus petite.

Pour la *tristesse*, la forme des sourcils, des yeux et de la bouche se courbe vers le bas. En ce qui concerne le *dégoût*, la bouche est également courbée vers le bas (UA15), mais l'ouverture est plus importante. Un important mouvement nasal est également présent dans la deuxième composante du dégoût (UA9). 18% de la variance de la *joie* (CP2) représente un étirement des lèvres vers le haut, qui est caractéristique de la forme familière du sourire (UA12). Il convient de noter que le mouvement du visage lié à la parole (ouverture de la bouche et protrusion) est également important. Globalement pour toutes les émotions, les trois premières CPs sont liées à l'ouverture des lèvres (UA25 et UA26) et à l'étirement / protrusion des lèvres (UA18 et UA22).

Sur la base de la représentation réduite du visage de l'acteur (voir figure 4.5), quelques mesures ont été calculées et comparées en fonction de certains capteurs et distances spécifiques. Dans la figure 4.5, la distance euclidienne entre les capteurs (a) et (b) a été utilisée pour calculer l'ouverture / fermeture de l'oeil (UA5, UA7 et UA45). Pour le mouvement des sourcils, le capteur (d) représente le capteur central du sourcil gauche au repos. Les coordonnées de ce capteur ont été utilisées comme référence pour calculer l'élévation / froncement des sourcils (UA2 et UA4) en utilisant les coordonnées du capteur (c) qui représente le capteur central du sourcil gauche

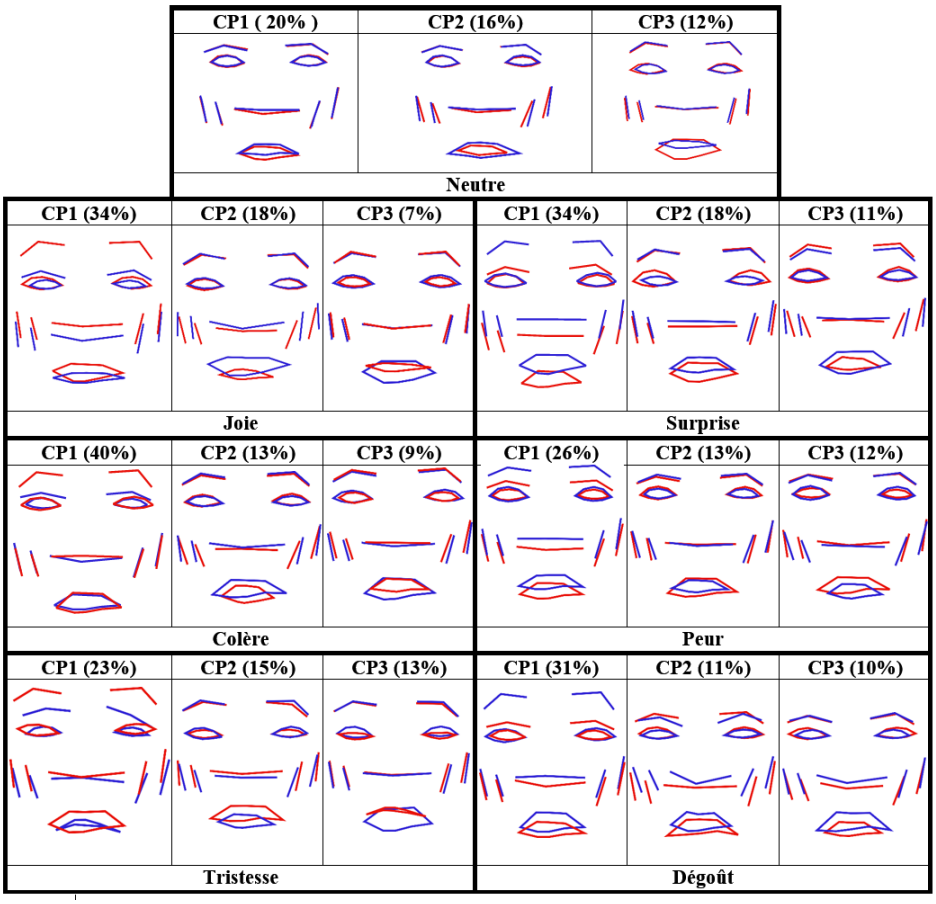


FIGURE 4.4 – *Les trois premières composantes principales des données visuelles et leur pourcentage de variance pour l'état neutre et les 6 émotions. Chaque paire de couleurs montre la déformation du visage lorsque les composantes prennent des valeurs entre -3 (bleu) et +3 (rouge) de déviation standard.*

pour les différentes émotions. Les capteurs (e) et (f) ont été utilisés pour mesurer l'étirement de la bouche, tandis que les capteurs (g) et (h) ont été utilisés pour mesurer l'ouverture de la bouche.

La figure 4.6 montre le résultat de l'ouverture / fermeture moyenne des yeux (UA5, UA7 et UA45) pour chaque émotion. La valeur moyenne et l'écart type l'ouverture de chaque oeil (droite et gauche) ont été calculés. Le résultat de la figure 4.6 est cohérent avec nos résultats en utilisant l'ACP. La *surprise* et la *peur* sont les émotions ayant la valeur d'ouverture oculaire la plus élevée, la *tristesse* et le *dégoût* sont les plus faibles et les émotions restantes ont une valeur d'ouverture oculaire modérée.

Les figures 4.9 et 4.10 montrent respectivement l'étirement et l'ouverture moyens des lèvres pour chaque émotion. L'étirement des lèvres (lié aux unités d'action UA12, UA13, UA14, UA15 et UA20) a été calculé sur la base de la distance e-f et l'ouverture des lèvres (liée aux unités d'action UA25, UA26 et UA27) est basée sur la distance g-h. Pour le mouvement des sourcils (lié aux unités d'actions UA1, UA2 et UA4) les résultats sont présentés sur la figure 4.7. Cette mesure a été calculée sur la base du point central des sourcils (c). Une frame représentant le visage au repos a été sélectionnée pour servir de référence. La distance c-d a été calculée pour

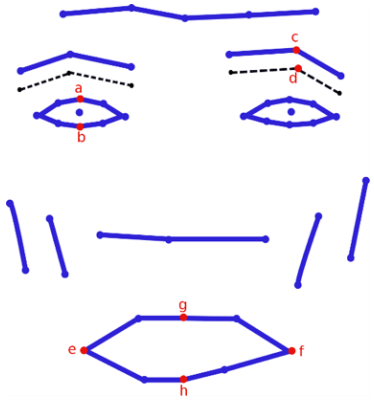


FIGURE 4.5 – *La configuration des marqueurs utilisés pour calculer les mesures faciales.*

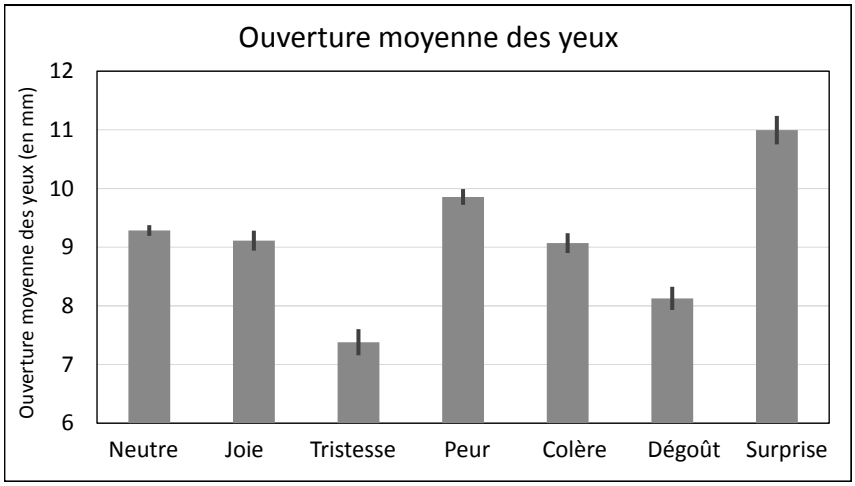


FIGURE 4.6 – *L'ouverture moyenne des yeux (en mm) pour chaque émotion et leur écart type. Cette mesure est représentée sur la figure 4.5 avec la distance a-b. La valeur moyenne et l'écart-type d'ouverture de chaque oeil (droite et gauche) ont été calculés, puis nous avons calculé leur valeur moyenne.*

les deux sourcils lors de l'enregistrement de chaque émotion. La distance euclidienne moyenne c-d a été calculée pour les deux sourcils, puis la valeur moyenne des résultats des deux sourcils a été calculée.

Les figures 4.6 et 4.10 montrent que les émotions avec l'ouverture moyenne des yeux la plus élevée ont également une grande valeur d'ouverture de la bouche et inversement. Les tendances globales de l'ouverture des yeux et des lèvres sont très similaires. Comme le montre la figure 4.8, pratiquement pour toutes les émotions, les mouvements des sourcils ne concernent que les haussements (UA1 et UA2), à l'exception du *dégoût* qui contient plusieurs froncements des sourcils (UA4). Certaines émotions se distinguent par leurs caractéristiques visuelles tandis que d'autres partagent des caractéristiques similaires :

Neutre : Les mesures de l'état neutre ont été utilisées comme référence pour comparer toutes les autres caractéristiques des émotions. Pour cette état, les sourcils bougent à peine et les autres mesures ont les valeurs les plus basses ou très basses.

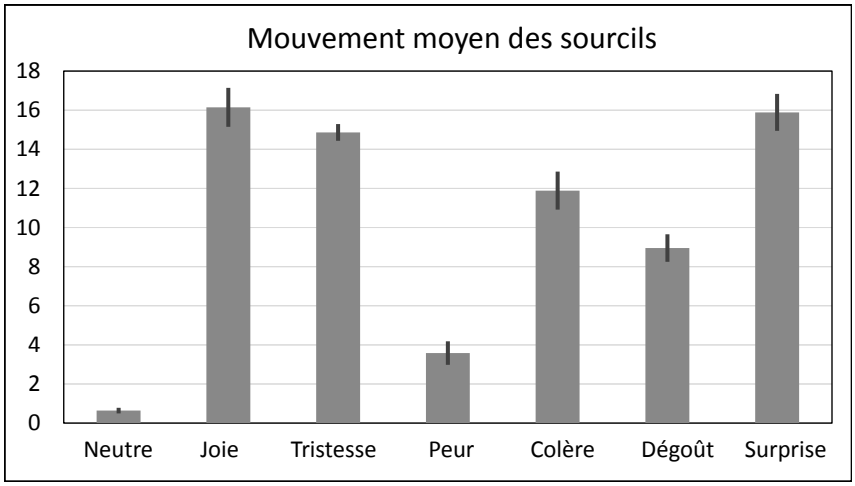


FIGURE 4.7 – *Mouvement moyen des sourcils (en mm) pour les 7 émotions et leur écart type. Calculé sur la base du point central des sourcils (c). Une frame en position de repos a été sélectionnée comme référence.*

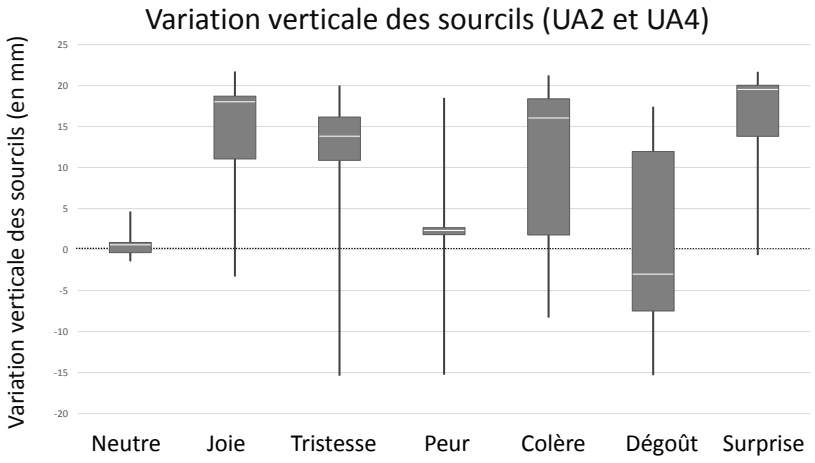


FIGURE 4.8 – *Valeurs de l'axe vertical (en mm) pour le capteur central du sourcil gauche. Les rectangles représentent les premier et troisième quartiles. La ligne blanche horizontale représente la médiane et les extrémités des lignes verticales représentent les valeurs min et max de la position du capteur. Les valeurs positives représentent l'élévation des sourcils (UA2), celles négatives représentent le froncement des sourcils (UA4).*

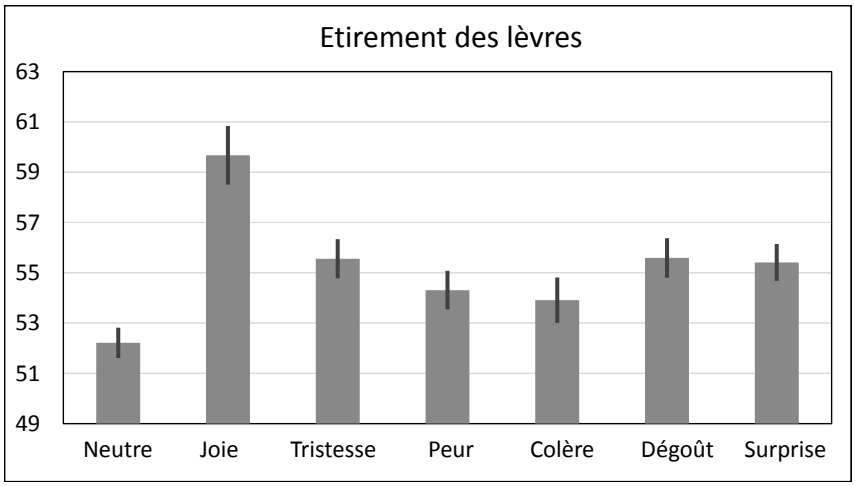


FIGURE 4.9 – *L'étirement moyen des lèvres (en mm) pour chaque émotion et leur écart type. L'étirement des lèvres a été calculé sur la base de la distance e-f.*

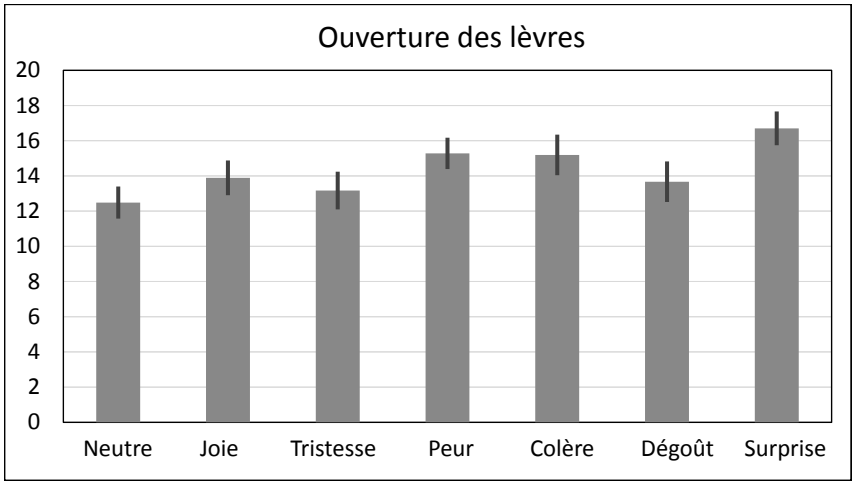


FIGURE 4.10 – *L'ouverture moyenne des lèvres UA25/UA26 (en mm) pour chaque émotion et leur écart type. L'ouverture des lèvres a été calculée sur la base de la distance g-h.*

Surprise : Cette émotion se distingue sur plusieurs niveaux des autres émotions. L'ouverture des yeux et des lèvres est la plus grande (UA5 et UA26). Cette émotion a également un étirement des lèvres important par rapport au neutre. De plus, les valeurs de mouvement des sourcils sont les plus élevées (UA1 et UA2). Ces remarques sont cohérentes avec le résultat de l'ACP, puisque 34% de la variation représente une bouche ouverte en permanence et des sourcils haussés.

Joie : Cette émotion est caractérisée par l'étirement des lèvres le plus élevé (UA12). Ce résultat peut être expliqué par la forme du sourire qui nécessite un étirement important des lèvres, cette forme est capturée par la deuxième CP de la *joie*. le froncement du nez (UA9) est également présent dans CP1 et CP2. La *joie* nécessite également un grand mouvement des sourcils (UA1 et UA2). Pour l'ouverture des lèvres et des yeux (UA25, UA26 et UA5) les valeurs obtenues sont relativement modérées.

Tristesse et dégoût : Lors de l'analyse de la *tristesse*, une certaine ressemblance avec le *dégoût* a été repéré. elles ont toutes les deux l'ouverture des yeux la plus basse (UA7) et un étirement des lèvres important lié aux UA15 et UA20 (en cohérence avec les résultats de l'ACP, où ces deux émotions étaient caractérisées par un mouvement d'étirement des lèvres vers le bas). De plus, ces deux émotions ont une ouverture de lèvre modérée (UA25). Pour le mouvement des sourcils, la *tristesse* se caractérise par un mouvement constant d'élévation des sourcils (UA1 et UA2). Cependant, pour le *dégoût*, les mouvements varient d'une montée rapide des sourcils à une position stable de froncement des sourcils (UA1, UA2 et UA4).

Colère : Cette émotion représente une ouverture marquée des yeux que nous pouvons constater dans la CP1. La CP2 concerne principalement le serrage des lèvres (UA23) et les mouvements de protrusion des lèvres (UA18 et UA22), ce qui est cohérent avec la faible valeur d'étirement des lèvres de la *colère* dans la figure 4.9. De plus, dans le corpus de la *colère*, l'ouverture des lèvres (UA26) était remarquablement élevée (deuxième après *surprise*). Le froncement du nez (UA9) peut être remarqué dans la CP2 et aucun pattern dominant n'a été remarqué pour les mouvements des sourcils. Les mouvements des sourcils alternent entre une position de repos stable et des élévations rapides ou stables (UA1 et UA2).

Peur : Cette émotion présente l'ouverture des yeux la plus importante (UA5) après la *surprise*. De la même manière que la *colère*, et malgré leur faible valeur d'étirement des lèvres, ces émotions compensent par un mouvement d'ouverture des lèvres élevé (UA26). Les figures 4.7 et 4.8 montrent que les mouvements de haussements des sourcils (UA1 et UA2) sont présents mais extrêmement faibles. En décrivant les différentes caractéristiques visuelles des émotions en termes d'unités d'action, nous concluons que nos résultats sont similaires à ce qui peut être généralement trouvé dans la littérature [Ekman and Friesen, 1976, Tian et al., 2001, Wiggers, 1982, Lucey et al., 2010]. En plus de reproduire les résultats précédents, la présente étude a démontré que la principale caractéristique faciale de ces émotions est maintenue même pendant l'activité de la parole. De plus, des unités d'action supplémentaires sont présentes dans les trois premières composantes principales de toutes les émotions (UA26, UA26, UA18, UA22). Ces actions sont liées à l'activité de la parole (ouverture de la bouche, protrusion).

4.3.2 Analyse acoustique

Les données acoustiques ont été enregistrées en même temps que les données visuelles et les deux flux ont été synchronisés. Nous avons commencé le post-traitement en faisant un aligne-

ment de la parole pour chaque phrase à différents niveaux : mots, syllabes et phonèmes. avec cet alignement, nous avons cherché à obtenir des caractéristiques acoustiques à un niveau très fin. Nous avons utilisé CMUSphinx (une boîte à outils open source pour la reconnaissance et l'alignement de la parole par Lamere et al. [2003]) pour effectuer un premier alignement phonétique, puis nous avons effectué une vérification manuelle, pour corriger les éventuelles erreurs et imperfections. Nous avons utilisé le logiciel PRAAT (un logiciel gratuit pour l'analyse de la parole en phonétique [Boersma et al., 2002]), pour calculer les différents paramètres acoustiques. Le corpus étant relativement petit, nous nous sommes concentrés sur les caractéristiques globales, calculées sur la totalité de la phrase. Nous calculons les caractéristiques les plus courantes [Pell et al., 2009] : 1) F0 et énergie : moyenne, minimum, maximum, plage, 2) Durée : débit de l'articulation. Pour évaluer les caractéristiques vocales, le *jitter* et le *shimmer* sont généralement pris en compte comme paramètres dans les systèmes du traitement automatique de la parole. Le *jitter* (respect. *shimmer*) mesure la perturbation de la longueur (respect. amplitude) entre deux périodes de pitch consécutives. En d'autres termes, Le *jitter* représente la micro-variation du pitch dans la voix et *shimmer* signifie la variation de l'intensité de la voix. Ces caractéristiques sont calculées pour chaque phrase et la valeur moyenne est utilisée pour chaque émotion.

La figure 4.11 et la figure 4.12 (détaillées dans le tableau 4.3) montrent les caractéristiques de la F0 pour chaque émotion. L'intervalle de confiance à 95% montre que la moyenne de la F0 est assez stable pour toutes les phrases, c'est donc une caractéristique robuste.

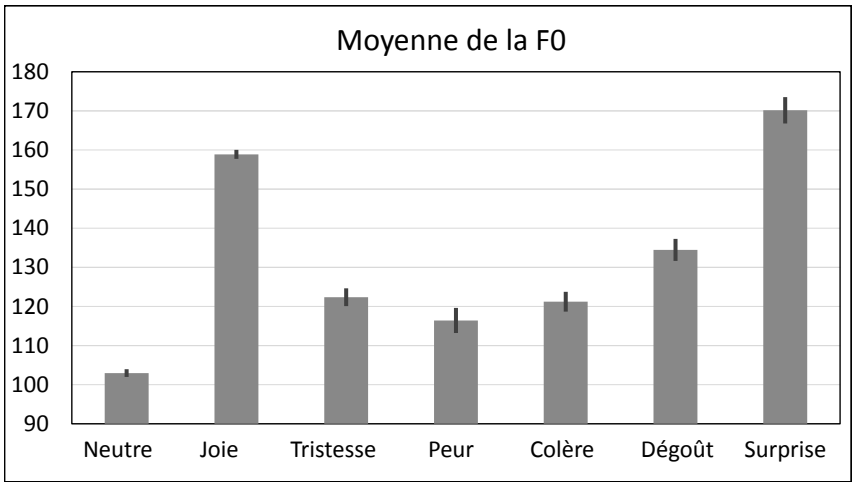


FIGURE 4.11 – *Valeurs moyennes de la F0 (en Hz) pour les 7 émotions et leur intervalle de confiance à 95%.*

De plus, l'éthologue Eugene Morton a trouvé un modèle inter-espèces, dans lequel les hautes fréquences sont corrélées avec les comportements d'affiliation, tandis que les basses fréquences sont associées aux comportements agressifs [Morton, 1977, 1994]. De plus, dans une étude transculturelle de la prosodie de la parole, Bolinger [1978] ont trouvé une association similaire chez l'homme : une F0 élevée est associée à des comportements amicaux tandis qu'une F0 faible est associé à des comportements agressifs. Les résultats tracés dans les figures 4.11 et 4.12 révèlent que la joie et la surprise ont une F0 plus grande que les autres émotions, comme suggéré précédemment. Quant à la peur et à la colère, elles ont la F0 la plus basse puisqu'elles sont les plus éloignées d'une attitude sympathique.

Néanmoins, résumer l'intonation de l'émotion par des valeurs min/max/moyenne est assez restrictif. L'intonation peut également être considérée comme un geste. Le contour de l'intonation

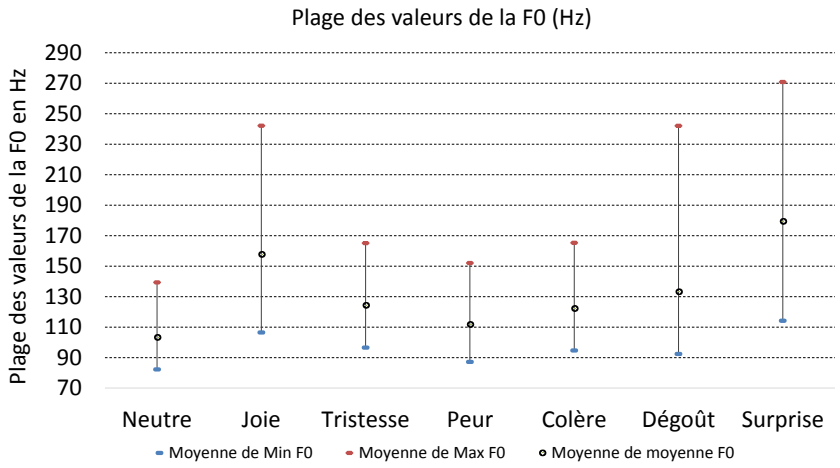


FIGURE 4.12 – *Plage des valeurs de la F0 (en Hz) pour les 7 émotions.*

est également significatif pour comparer les émotions. Un exemple du contour typique de la F0 est donnée pour la phrase : "Mais les gens ne se mettent pas en grève par plaisir, tu devrais le savoir, si les grenouilles avaient des ailes, elles ne s'embêteraient pas à sauter." dans la figure 4.13. La valeur moyenne de la F0 a été calculée pour chaque syllabe à l'aide de PRAAT. La durée de la syllabe n'est pas représentée sur la figure, car la durée de la phrase pour chaque émotion est différente. Les contours des 10 phrases de toutes les émotions ont été calculés. Après cela, la moyenne F0 de chaque énoncé a été calculée et présentée sur la figure 4.14. Cette figure capture la même tendance moyenne que ce qui a été trouvé pour le contour complet d'une phrase (figure 4.13). Le classement des émotions en termes de valeurs F0 suit pratiquement le même ordre pour toutes les phrases. Il est à noter que les tendances dominantes des contours F0 sont correctement capturées par la mesure moyenne et le rang F0 global de chaque émotion est préservé.

	Neutre	**Joie**	**Tristesse**	**Peur**	**Colère**	**Dégoût**	**Surprise**
Moyenne Min F0	82.19	106.49	96.53	87.16	94.68	92.37	114.18
Moyenne Max F0	139.33	242.08	165.07	152.01	165.31	242.03	270.77
Moyenne Moyenne F0	103.20	157.71	124.24	111.731	122.21	133.14	179.37

TABLE 4.3 – *Plage des valeurs de la F0 pour les 7 émotions.*

Globalement, les résultats obtenus confirment ce qui a été trouvé dans d'autres études [Scherer, 1986, Paeschke et al., 1999] sur la corrélation entre les émotions et la F0 moyenne. Toutes les émotions ont une moyenne globale F0 plus élevée que l'état neutre. Nos résultats sont similaires à ce qui est attendu, à l'exception de la tristesse qui a été considérée comme ayant une F0 inférieure à celle du neutre. Cette différence peut provenir de la nature jouée (exagérée) de notre corpus qui peut transmettre un degré d'émotion plus fort que la parole spontanée.

Débit : La figure 4.15 et le tableau 4.4 montrent les statistiques du débit de la parole (phonèmes par seconde). Pour calculer cette mesure, la longueur totale des sons (les silences n'ont pas été pris en compte) a été divisée par le nombre total de sons par phrase. Cette information (la longueur d'un seul son dans une émotion donnée) a été utilisée pour calculer le nombre de sons par seconde. La valeur moyenne a été calculée pour les dix phrases pour toutes les émotions. Toutes les émotions ont un débit de parole plus élevé que le neutre. Dans la figure 4.15, le taux

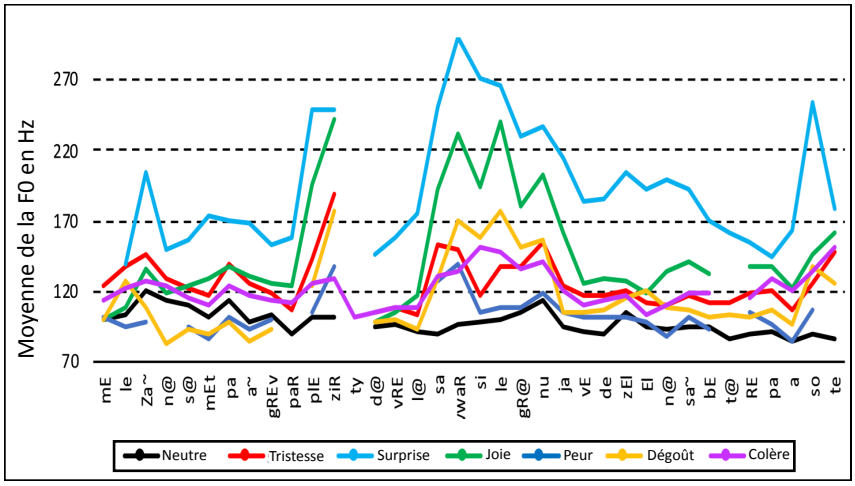


FIGURE 4.13 – *Contours de la F0 d'une phrase du corpus pour les 7 émotions (par syllabes).*

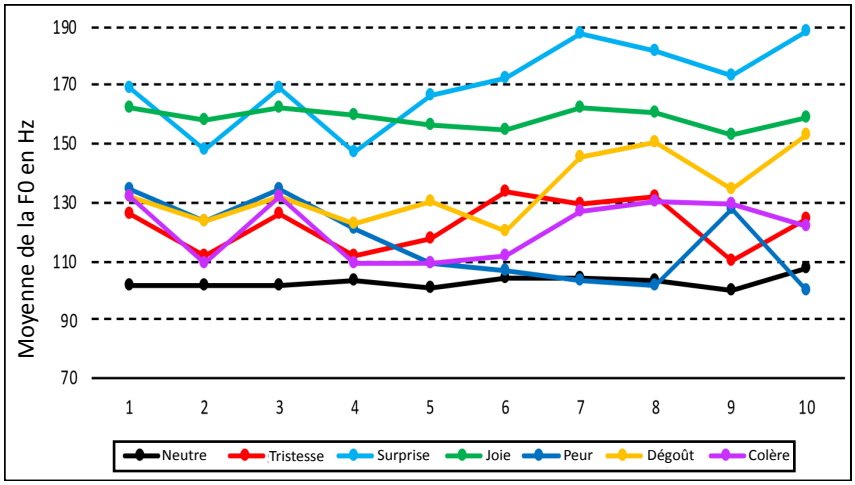


FIGURE 4.14 – *Valeurs moyennes de la F0 pour les 7 émotions (par phrases)*

d'articulation est exprimé en pourcentage par rapport au débit de l'articulation de l'état neutre. Par exemple, +20% signifie que la vitesse est 20% plus rapide que la vitesse de l'état neutre neutre.

Le débit de l'articulation distingue la *colère* et le *dégoût* des autres émotions et est nettement plus rapide que le *neutre*. La *joie* et la *peur* ont un débit très similaire environ 10% plus rapide que le neutre. La *tristesse* et la *surprise* ont une vitesse plus faible (7% et 3% respectivement). Cependant, contrairement aux études antérieures, dans notre analyse, la *tristesse* et le *dégoût* se sont avérés être associés à un débit de parole plus rapide, plutôt qu'à un rythme plus lent, par rapport au *neutre*.

Jitter/shimmer : Les figures 4.16 et 4.17 présentent les résultats de nos données pour les paramètres de *jitter* et *shimmer*. La différence entre les émotions est très subtile pour ces paramètres. Les résultats ont montré que des valeurs de *jitter* et *shimmer* les plus élevées sont associées à la *peur*, au *dégoût* et à la *colère*. Les valeurs de *jitter* et *shimmer* les plus faibles ont été trouvées pour la *surprise*. Pour les autres émotions (la *tristesse* et la *joie*), elles ont une valeur

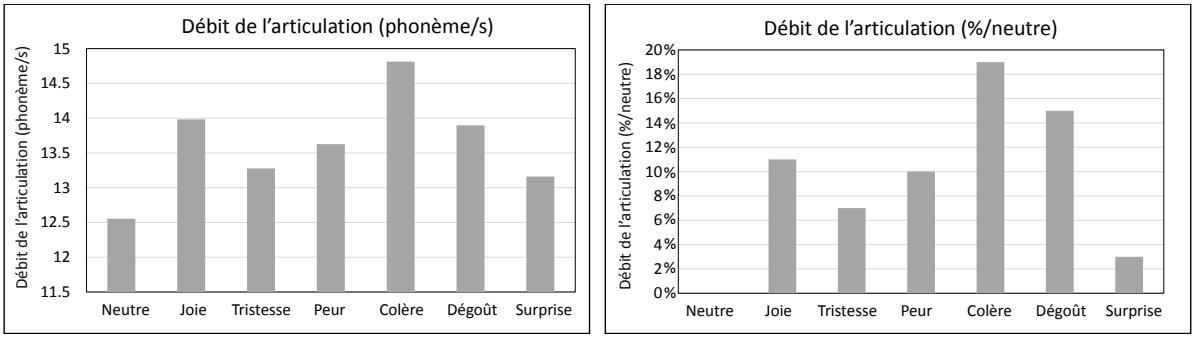


FIGURE 4.15 – *A gauche : Débit de l'articulation par émotion (nombre de sons par seconde), calculé à partir des 10 phrases. A droite : Débit de l'articulation par émotion par rapport au taux d'articulation de l'état neutre, calculé à partir des 10 phrases.*

	Colère	**dégoût**	**peur**	**Joie**	**Neutre**	**Tristesse**	**Surprise**
(phonème/s)	15.24	14.73	14.1	14.19	12.82	13.66	13.22
(%/neutre)	19%	15%	10%	11%	0%	7%	3%

TABLE 4.4 – *Débit de l'articulation par émotion calculé à partir des 10 phrases.*

similaire au *neutre* (*tristesse* avait une valeur légèrement supérieure à la *joie*). De plus, Nunes [2013] ont constaté que *jitter* et *shimmer* sont plus élevés pour les émotions les plus négatives et faibles pour les émotions positives. Nos résultats sont alignés avec ces conclusions, à l'exception de la tristesse qui représente des valeurs faible de *jitter* et de *shimmer* (proches de celles du *neutre*).

Certaines émotions se distinguent par leurs caractéristiques acoustiques tandis que d'autres partagent des caractéristiques similaires :

Neutre : L'émotion neutre a été utilisée à nouveau comme référence pour comparer le comportement des autres émotions. Les valeurs des paramètres les plus faibles ont été trouvées pour cette état émotionel (F0 la plus basse, débit d'articulation, *jitter* et *shimmer*).

Surprise : Cette émotion présente la valeur de la F0 la plus élevée. La *Surprise* était également la seule à avoir des valeurs de *jitter* et *shimmer* inférieures au *neutre*.

Joie : Les valeurs de la F0 étaient également importantes pour cette émotion. Mais, contrairement à la *surprise*, le débit de d'articulation de la joie était élevé. Les valeurs de *jitter* et *shimmer* étaient identiques à celle du *neutre*.

Tristesse : Cette émotion a une moyenne de F0 et un débit d'articulation modérés. En ce qui concerne le *jitter* et *shimmer*, les valeurs de la *tristesse* étaient très proches du *neutre*.

Dégoût : Cette émotion a une valeur de F0 moyenne supérieure au *neutre* mais la plage des valeurs de la F0 est clairement l'une des plus élevées. Le débit d'articulation est également l'un des plus rapides.

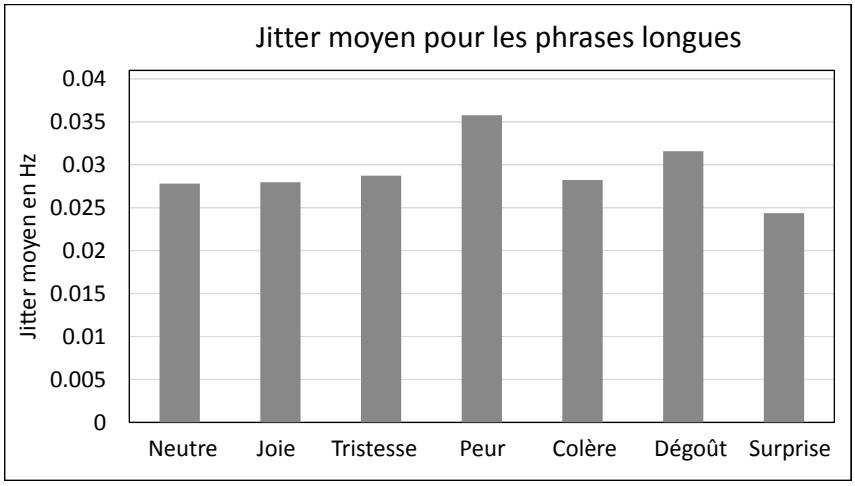


FIGURE 4.16 – *Valeur moyenne du jitter pour les 7 émotions.*

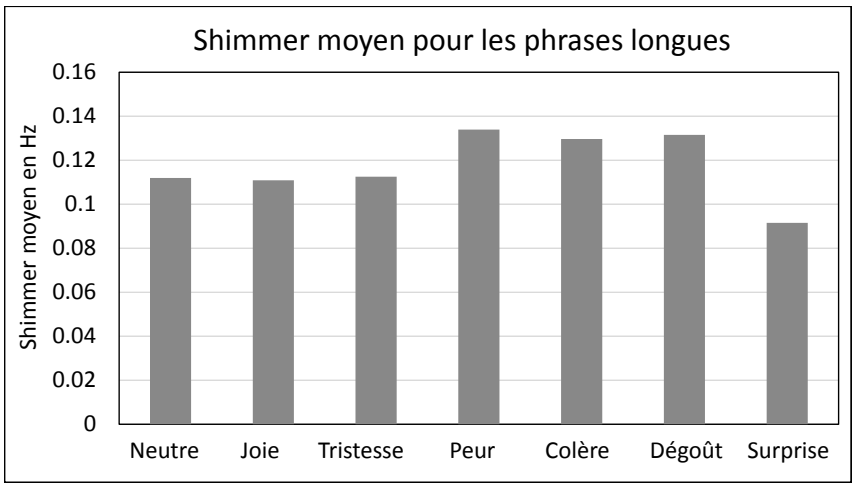


FIGURE 4.17 – *Valeur moyenne du shimmer pour les 7 émotions.*

Colère : Lors de l'analyse des phrases de la *colère*, la plage des valeurs de la F0 était relativement faible, mais le débit d'articulation s'est avéré être le plus élevé parmi toutes les autres émotions. Sa valeur de *jitter* est identique au *neutre* mais la valeur de *shimmer* est plus importante.

Peur : La valeur de la F0 était très faible, légèrement supérieure au *neutre*. Contrairement au *neutre*, les pics des valeurs de la F0 pour la *peur* atteignent des niveaux plus élevés. D'un autre côté, une valeur modérée du débit d'articulation a été trouvée pour la *peur*. Cette émotion a les valeurs de *jitter* et *shimmer* les plus élevées.

4.4 Étude perceptive du corpus

L'évaluation perceptuelle a pour objectif de déterminer si l'acteur a été capable de transmettre les différentes émotions correctement. Cela est essentiel pour décider de la qualité du corpus audiovisuel expressif acquis. Comme les données finales seront utilisées pour développer

un système de synthèse audiovisuelle expressive de la parole, il est important d'évaluer la configuration des marqueurs 3D pour voir si cette représentation minimale du visage est suffisante pour modéliser l'expressivité faciale et si leurs nombre et les positions sont capables de capturer correctement les caractéristiques d'expressivité.

Une technique d'évaluation similaire a été utilisée dans le passé pour évaluer un système de synthèse audiovisuelle [Bailly et al., 2002]. Dans un travail précédent, nous avons présenté une étude détaillée mettant l'accent sur les aspects perceptuels de l'expressivité. Nous avons également discuté de l'influence de chaque modalité et de sa contribution dans la perception des émotions [Ouni et al., 2017]. En particulier, nous avons comparé deux modalités : bimodale (avec audio) et unimodale (sans audio), à travers trois présentations : points 3D sans mouvement de la tête, points 3D avec mouvements de la tête puis une vidéo du visage de l'acteur. Les stimuli utilisés étaient 10 phrases courtes pour les différentes émotions (tous les détails se trouvent dans [Ouni et al., 2017]). Dans la présente évaluation, nous nous concentrerons uniquement sur deux présentations et deux modalités : (1) phrases courtes vs. (2) phrases longues et (3) visage de l'acteur (vidéo) vs. (4) ensemble de points 3D (représentation minimaliste du visage).

4.4.1 Stimuli

Nous avons utilisé le même jeu de phrases que présenté en 4.2.2 : Vingt phrases (10 courtes et 10 longues) prononcées par un acteur de 27 ans, semi-professionnel, avec 7 états émotionnels différents (neutre, joie, surprise, peur, colère, tristesse et dégoût), l'acteur utilisant la technique des *exercices de style*. Comme les participants n'étaient pas tous de la même culture, nous avons choisi d'utiliser les émotions de base (neutre, joie, surprise, peur, colère, tristesse et dégoût), car il s'est avéré qu'elles étaient universelles [Ekman and Friesen, 1971]. Nous avons découpé les enregistrements en une phrase par fichier de stimuli, puis trois types de stimulus ont été présentés aux participants : (1) le stimulus audiovisuel du visage de l'acteur (video) (2) le stimulus audiovisuel avec représentation minimaliste du visage (3) le stimulus visuel uniquement avec représentation minimaliste du visage. Pour chaque partie de l'expérience, les stimuli ont été présentés de manière aléatoire.

4.4.2 Participants

Les expériences perceptuelles ont été menées avec deux groupes de participants. Les expériences perceptuelles des phrases longues comptent douze participants (adultes, 5 femmes, 7 hommes). Pour les phrases courtes, un autre groupe de treize participants (adultes, 3 femmes et 10 hommes) ont passé les tests. Les participants des deux groupes n'étaient pas des français natifs, mais ils vivaient en France durant la période de l'étude.

4.4.3 Méthode

Nous avons mis en place une application web sur laquelle les participants pouvaient se connecter pour effectuer les tests perceptifs. Une série de stimulus ont été présentés un par un et chaque participant devait choisir, selon lui, et parmi une liste de sept possibilités (neutre, joie, surprise, peur, colère, tristesse et dégoût) l'émotion exprimée dans les stimuli. Le participant devait sélectionner une réponse et valider pour pouvoir voir le stimulus suivant. Les participants avaient la possibilité de rejouer les stimuli autant de fois qu'ils le souhaitaient. Nous n'avons pas commenté ni imposé de définition des émotions pour éviter de biaiser la perception des participants. Les participants ont effectué le test avec les stimuli acoustiques uniquement, puis visuels uniquement, avant de pouvoir passer les tests audiovisuels.

		Émotion perçue						
		Joie	**Tristesse**	**Colère**	**Peur**	**Dégoût**	**Surprise**	**Neutre**
Émotion produite	**Joie**	**95.00(*)**	0	0	0	0	5	0
	Tristesse	0	**83.33(*)**	0	3.33	10	0	3.33
	Colère	0	8.33	**60(*)**	1.67	25	3.33	1.67
	Peur	0	1.67	8.33	**71.67(*)**	1.67	16.67	0
	Dégoût	0	0	48.33	0	**51.67(*)**	0	0
	Surprise	10	0	0	1.67	0	**86.67(*)**	1.67
	Neutre	0	5	0	0	0	0	**95(*)**

TABLE 4.5 – *La matrice de confusion du taux de reconnaissance des 7 émotions avec **la vidéo du visage de l'acteur pour les phrases longues**. Les lignes représentent la distribution des réponses données par les participants.*

		Émotion perçue						
		Joie	**Tristesse**	**Colère**	**Peur**	**Dégoût**	**Surprise**	**Neutre**
Émotion produite	**Joie**	**97.12(*)**	0	0	0	0	2.88	0
	Tristesse	0	**83.65(*)**	0	0	14.42	0	1.92
	Colère	0	2.88	**70.19(*)**	2.88	9.62	7.69	6.73
	Peur	0	2.88	0	**92.31(*)**	0	4.81	0
	Dégoût	0	4.81	4.81	0.96	**89.42(*)**	0	0
	Surprise	4.81	0	15.38	0	4.81	**75(*)**	0
	Neutre	0	0	0	0	0.96	0	**99.04(*)**

TABLE 4.6 – *La matrice de confusion du taux de reconnaissance des 7 émotions avec **la vidéo du visage de l'acteur pour les phrases courtes**. Les lignes représentent la distribution des réponses données par les participants.*

Résultats

Après la collecte des résultats des différentes expériences, nous avons calculé le niveau de significativité statistique en utilisant la valeur-p avec un test-t. Puisque nous testons le taux de reconnaissance de sept émotions en parallèle nous avons corrigé les résultats obtenus avec la méthode de Holm–Bonferroni [Holm, 1979]. Cette correction est conçue pour les tests à hypothèses multiples et réduit la possibilité d'obtenir un résultat statistiquement significatif lors de l'exécution de plusieurs tests parallèles. Pour chaque expérience, nous avons utilisé un degré de liberté égale au nombre de participants moins 1 (12 pour les phrases courtes et 11 pour les longues). Nous avons utilisé une valeur critique correspondant au risque (alpha) égale à 5% et un niveau de hasard de 14% correspondant à la probabilité qu'une émotion donnée soit choisie au hasard parmi les sept choix possible. Nous avons ajouté le symbole (*) pour les résultats statistiquement significatifs et (-) pour ceux statistiquement non-significatifs. Les résultats sont présentés dans les tableau 4.5– 4.10.

En ce qui concerne les stimuli audiovisuels des phrases longues contenant des vidéos de l'acteur (table 4.5), la majorité des émotions ont été très bien reconnues (plus de 70%). Toutefois, la colère et le dégoût ont été beaucoup confondues. Cette constatation a été signalée par des résultats similaires dans la littérature [Ekman and Friesen, 1986]. De plus, quelques recherches ont aussi révélé qu'il n'est pas correct de supposer que le public partage des significations similaires des noms des émotions et que la compréhension commune du mot dégoût reflète une combinaison

		Émotion perçue						
		Joie	**Tristesse**	**Colère**	**Peur**	**Dégoût**	**Surprise**	**Neutre**
Émotion produite	**Joie**	**81.67(*)**	0	0	1.67	0	10	6.67
	Tristesse	0	**76.67(*)**	0	3.33	15	1.67	3.33
	Colère	1.67	5	**46.67(*)**	6.67	20	16.67	3.33
	Peur	0	5	5	**66.67(*)**	5	18.33	0
	Dégoût	0	0	65	6.67	**23.33(-)**	1.67	3.33
	Surprise	16.67	1.67	0	5	3.33	**68.33(*)**	5
	Neutre	0	16.67	0	1.67	5	0	**76.67(*)**

TABLE 4.7 – *La matrice de confusion du taux de reconnaissance des 7 émotions avec* ***la représentation minimaliste du visage de l'acteur pour les phrases longues (avec audio)****. Les lignes représentent la distribution des réponses données par les participants.*

		Émotion perçue						
		Joie	**Tristesse**	**Colère**	**Peur**	**Dégoût**	**Surprise**	**Neutre**
Émotion produite	**Joie**	**84.62(*)**	0	0	0.96	1.92	11.54	0.96
	Tristesse	0	**58.65(*)**	0.96	0	25.96	0	14.42
	Colère	0.96	1.92	**40.38(*)**	10.58	18.27	7.69	20.19
	Peur	0.96	4.81	0	**75.00(*)**	4.81	13.46	0.96
	Dégoût	0.96	2.88	39.42	1.92	**47.12(*)**	1.92	5.77
	Surprise	3.85	0	3.85	9.62	3.85	**77.88(*)**	0.96
	Neutre	0	3.85	2.88	0	4.81	0	**88.46(*)**

TABLE 4.8 – *La matrice de confusion du taux de reconnaissance des 7 émotions avec* ***la représentation minimaliste du visage de l'acteur pour les phrases courtes (avec audio)****. Les lignes représentent la distribution des réponses données par les participants.*

entre le dégoût et la colère [Nabi, 2002]. Nous rappelons que nous n'avons pas commenté ni imposé de définition des émotions pour éviter de biaiser la perception des participants. Comme le montre le tableau 4.6, la présentation audiovisuelle par vidéos des phrases courtes a des taux de reconnaissance plus élevés pour presque toutes les émotions. Cela peut s'expliquer par le fait que l'acteur doit produire une émotion plus intense, car la durée de délivrance des émotions est courte.

Concernant le corpus audiovisuel (les vidéos de l'acteur), les résultats des deux expériences précédentes montrent que la majorité des participants valident la performance de l'acteur et confirment la bonne qualité du corpus audiovisuel produit.

Nous rappelons que lors du développement de la tête parlante expressive, nous n'utiliserons pas directement la vidéo de l'acteur, mais les points 3D correspondant aux marqueurs sur le visage de l'acteur (figure 4.2). Les points 3D seront utilisés pour animer un modèle 3D. Le but de cette évaluation perceptive est également de voir si la représentation minimaliste est suffisante pour exprimer les différentes émotions.

Pour les phrases longues, toutes les émotions ont été correctement reconnues sauf le dégoût (voir Table 4.7). Les taux de reconnaissance les plus faibles étaient ceux de la colère et du dégoût qui ont été confondues l'une avec l'autre (46% et 23% respectivement). Une baisse moyenne de 15% du taux de reconnaissance des émotions a été constatée lorsque la représentation minimale a été présentée plutôt que le vrai visage de l'acteur. Cette baisse peut s'expliquer par l'absence

d'une partie des informations dans la représentation minimaliste notamment l'aspect musculaire, certains changements de la peau comme le rougissement, l'apparition de ridules et de déformations de la peau dues à l'expressivité. Pour les phrases courtes, les résultats présentés dans le tableau 4.8 montrent que la tristesse et la colère ont un taux de reconnaissance inférieur à celui des phrases longues. Cependant, la joie, la peur, le dégoût, la surprise et le neutre ont un taux de reconnaissance plus élevé. La colère et la tristesse, dans leur présentation minimaliste, peuvent nécessiter un peu plus de temps pour être identifiées par les participants. Néanmoins, le taux de reconnaissance suit globalement la même tendance que pour les phrases longues, mais il est difficile de confirmer que la durée des phrase est un facteur dans la perception des émotions.

Pour les expériences utilisant la représentation minimaliste audiovisuelle du visage, il n'est pas clair si le taux de reconnaissance observé est principalement déterminé par les informations contenues dans la voix ou par les caractéristiques du visage minimaliste. Pour clarifier cela, nous avons effectué d'autres tests avec la représentation minimaliste du visage mais cette fois sans audio. Les résultats sont présentés dans les tableaux 4.9 et 4.10. Le taux de reconnaissance a chuté de façon drastique pour la colère, le dégoût et la peur, pour les phrases longues et courtes, et est devenu statistiquement non-significatif. Cela montre que pour ces émotions, la modalité acoustique contient une part importante des informations émotionnelles.

		Émotion perçue						
		Joie	**Tristesse**	**Colère**	**Peur**	**Dégoût**	**Surprise**	**Neutre**
Émotion produite	**Joie**	**73.33(*)**	0	0	1.67	0	15	10
	Tristesse	1.67	**50(*)**	1.67	11.67	23.33	0	11.66
	Colère	0	8.33	**18.33(-)**	10	13.33	33.33	16.67
	Peur	5	10	5	**30(-)**	6.67	23.33	20
	Dégoût	0	3.33	70	1.67	**21.67(-)**	0	3.33
	Surprise	15	3.33	6.67	16.67	1.67	**38.33(*)**	18.33
	Neutre	0	25	1.67	10	11.67	0	**51.67(*)**

TABLE 4.9 – *La matrice de confusion du taux de reconnaissance des 7 émotions avec* ***la représentation minimaliste du visage de l'acteur pour les phrases longues (sans audio)****. Les lignes représentent la distribution des réponses données par les participants.*

		Émotion perçue						
		Joie	**Tristesse**	**Colère**	**Peur**	**Dégoût**	**Surprise**	**Neutre**
Émotion produite	**Joie**	**75(*)**	0.96	0.96	2.88	0.96	15.38	3.85
	Tristesse	0.96	**69.23(*)**	0	4.81	14.42	0	10.58
	Colère	2.88	13.46	**5.77(-)**	16.35	4.81	47.12	9.62
	Peur	2.88	8.65	7.69	**21.15(-)**	14.42	7.69	37.5
	Dégoût	0	4.81	72.12	0.96	**17.31(-)**	2.88	1.92
	Surprise	5.77	4.81	4.81	34.62	5.77	**40.38(*)**	3.85
	Neutre	1.92	9.62	8.65	5.77	11.54	0	**62.5(*)**

TABLE 4.10 – *La matrice de confusion du taux de reconnaissance des 7 émotions avec* ***la représentation minimaliste du visage de l'acteur pour les phrases courtes (sans audio)****. Les lignes représentent la distribution des réponses données par les participants.*

Au final, ces résultats montrent que la représentation minimale du visage porte des infor-

mations émotionnelles pour la plupart des émotions étudiées notamment si elle est associée à l'audio. Cela valide la qualité de la représentation minimale audiovisuelle car ses taux de reconnaissances reflètent la même tendance que pour l'enregistrement vidéo de l'acteur, même si le taux de reconnaissance reste plus faible.

4.5 Conclusion

Dans ce chapitre, nous avons présenté une technique d'acquisition multimodale pour collecter des données audiovisuelles dans le but de développer une tête parlante virtuelle. Nous avons combiné trois systèmes d'acquisition pour obtenir des informations pertinentes pour chaque partie du visage (EMA pour le mouvement des lèvres lié à la parole, RealSense pour les yeux et caméras VICON pour les expressions faciales).

Dans l'analyse de la production de la parole, nous avons étudié les caractéristiques visuelles et acoustiques de nos données. Pour la modalité visuelle, nous avons effectué une ACP pour extraire les mouvements faciaux dominants de chaque émotion pendant la parole. La présente étude a démontré que les principales caractéristiques faciales de ces émotions sont maintenues même pendant l'activité de la parole. De plus, des unités d'action liées à l'activité de la parole (ouverture de la bouche et protrusion), sont présentes dans les trois premières CPs de toutes les émotions.

Nous avons présenté des expériences perceptives dont les résultats montrent que la majorité des participants valide la performance de l'acteur. Ce résultat confirme la bonne qualité du corpus audiovisuel. Nous avons également constaté qu'il est nécessaire de partager la même définition des émotions (notamment le dégoût) pour éviter une confusion basée sur une mauvaise compréhension de la signification des noms des émotions. La représentation minimaliste du visage semble être suffisante pour transmettre les émotions, elle a pu exprimer correctement certaines émotions avec un taux de reconnaissance étonnement élevé. Ce travail nous a donc permis de valider la configuration des capteurs utilisée pour cette expérience, nous pouvons donc l'utiliser dans les enregistrements futurs. La longueur des phrases a eu un certain effet sur l'expression des émotions, notamment la présentation audiovisuelle par vidéos où les phrases courtes ont eu un taux de reconnaissance plus élevés que les phrases longues. Cela peut s'expliquer par le fait que l'acteur doit produire une émotion plus intense, car la durée de délivrance des émotions est courte et que l'émotion est probablement diluée pour les phrases longues, car l'acteur ne peut pas maintenir la même intensité durant toute la phrase. Cependant, il est difficile de confirmer que la durée des phrases est un facteur dans la perception des émotions. Ce point nécessiterait une étude plus approfondie.

En se basant sur l'expérience de ce petit corpus, dans le chapitre suivant, nous présentons notre démarche pour acquérir un corpus audiovisuel expressive de taille plus grande. Ce corpus plus complet, sera utilisé dans le reste du manuscrit dans le processus de synthèse expressive de la parole.

5

Acquisition d'un corpus expressif pour la synthèse de la parole

Sommaire

5.1 Introduction

Dans ce chapitre nous présentons les différentes phases d'acquisition d'un corpus [4] audiovisuel expressif dédié à la synthèse de la parole. Nous détaillons les phases de préparations, d'acquisition et de post-traitement ainsi que la technique d'animation adoptée pour animer les personnages audiovisuels 3D. Dans le chapitre précédent (chapitre 4) nous avons discuté des conditions que notre corpus doit satisfaire. Le corpus que nous souhaitons enregistrer doit contenir les modalités acoustique et visuelle avec des données visuelles en 3D. Le corpus doit également contenir plusieurs catégories d'émotions. Nous avons constaté dans le corpus précédent que pour certaines émotions les définitions n'ont pas été comprises de la même manière par tous les participants. De plus, le nouveau corpus va être utilisé pour la synthèse de la parole, nous ajoutons quatre autres conditions :

1. Le locuteur et les participants doivent partager une compréhension commune des définitions des noms des émotions,

4. Le corpus présenté dans ce chapitre a été utilisé dans des publications dans des conférences internationales (Interspeech 2019 : https ://hal.inria.fr/hal-02175776/document et https ://hal.inria.fr/hal-02175780/document), européenes (Eusipco 2020 : https ://hal.inria.fr/hal-02573885/document) et nationales (JEP 2020 : https ://hal.archives-ouvertes.fr/hal-02798526/document).

2. Le corpus doit être équilibré : les classes d'émotions doivent avoir le même contenu linguistique et le même nombre de phrases pour avoir une couverture phonétique identique,
3. Le corpus doit être suffisamment grand pour entraîner des modèles de synthèse de la parole (quelques heures),
4. Notre protocole d'acquisition doit nous permettre d'enregistrer le corpus sur plusieurs sessions.

Nous expliquons dans ce chapitre comment nous avons traité chacune de ces nouvelles conditions.

5.2 Analyse linguistique

Comme les acquisitions audiovisuelles et leur post-traitement prennent beaucoup de temps, le but de cette analyse est de créer un corpus ayant la plus grande couverture phonétique possible tout en conservant un nombre raisonnable de phrases [François and Boëffard, 2001, Bozkurt et al., 2003, Barbot et al., 2015]. Chevelu and Lolive [2015] ont montré que le choix du contenu phonologique d'un corpus a beaucoup d'influence sur la qualité de la TTS. Le protocole que nous avons adopté consiste tout d'abord en une collecte d'un maximum de phrases afin de créer un premier grand corpus en langue française. Ce vaste corpus garantit une grande couverture linguistique initiale et sera traité ultérieurement afin de réduire sa taille. Pour ce faire, nous avons construit un premier corpus d'environ 7900 de phrases non redondantes et résulte de la fusion d'un corpus textuel français libre de droit SIWIS [5] et un autre interne à l'équipe (Lisa).

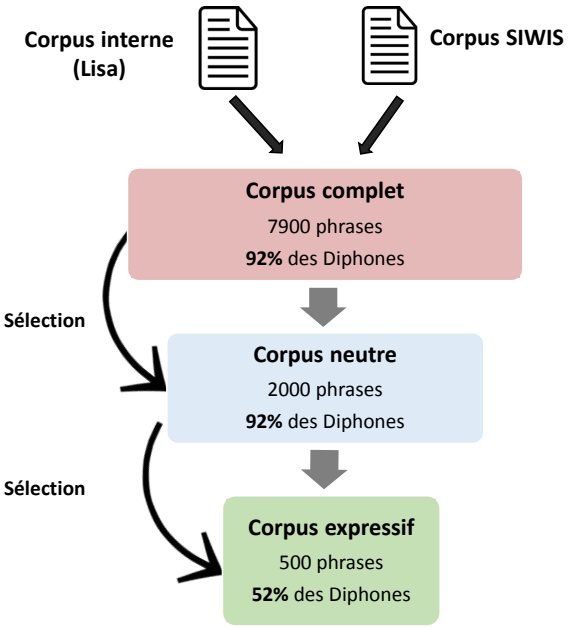


FIGURE 5.1 – *Les étapes que nous avons suivies pour construire deux corpus de tailles raisonnables avec une haute couverture diphonétique. Le premier corpus de 2000 phrases sera utilisé pour enregistrer l'état neutre et le deuxième de 500 phrases sera utilisé pour enregistrer les différents états expressifs.*

5. La base de données SIWIS : https ://www.unige.ch/lettres/linguistique/research/latl/siwis/database/

L'analyse linguistique consiste à découper toutes les phrases en une séquence de phonèmes à l'aide d'un phonétiseur et de calculer ensuite leurs couverture diphonétique. Ce traitement a été réalisé par un système interne de synthèse vocale français nommé SOJA [6]. L'analyse du corpus initial des 7900 phrases nous a permis d'obtenir un taux de 92% de couverture diphonétique. Les 8% restants représentent un ensemble de paires de phonèmes rares ou inexistantes en langue française.

Afin de réduire la taille de ce corpus, notre stratégie, illustrée sur la figure 5.1, consiste à sélectionner le minimum de phrases qui nous donne idéalement le même taux de couverture phonétique que le corpus initial. Pour ce faire, nous avons utilisé un algorithme glouton qui prend en entrée la séquence des phonèmes et une liste de critères linguistiques, principalement la position du phonème et ses contextes gauche et droit. L'algorithme glouton fournit en sortie la liste de phrases classées par ordre de richesse en couverture de diphones ainsi que le cumule du pourcentage de la couverture diphonétique. De cette manière, nous avons pu sélectionner les 2000 premières phrases qui permettent de garder le même taux de couverture original de 92%. Sachant que le processus d'acquisition et de traitement des données audiovisuelles est un processus très laborieux, nous avons décidé d'utiliser les 2000 phrases pour enregistrer un grand corpus neutre et de créer un autre corpus plus petit pour les six émotions basiques. De ce fait, nous avons réitéré le processus avec l'algorithme glouton, mais cette fois sur la liste des 2000 phrases pour en sélectionner une liste avec au moins 50% de couverture. Finalement, nous avons sélectionné 500 phrases qui couvrent 52% des diphones.

Nous utilisons les 36 phonèmes français et un symbole représentant les pauses (#) pour représenter les données. Leur nombre d'occurrences des 36 phonèmes et le symbole représentant les pauses (#) varie de 66 pour les plus rares (ŋ, j, ɥ, ø) et plus de 7K pour les plus fréquents (ʁ, a, l) (voir Fig. 5.2).

5.3 Préparation et acquisition du corpus

Contrairement à l'acquisition du corpus précédent où nous avons combiné trois systèmes d'acquisition, pour ce corpus nous utilisons un seul système. Nous avons pu améliorer le matériel d'acquisition en nous procurant le système de capture de mouvement Optitrack^{TM} spécialisé dans l'acquisition des données de la région du visage / tête. Fort de notre expérience et post-traitement des données du corpus précédent (fusion et synchronisation des données) et puisqu'il s'agit d'acquérir un corpus de taille plus grande, il est plus judicieux de privilégier le confort des acteurs et la simplicité du post-traitement. En fait, l'utilisation de l'articulographe condamne l'acteur à rester assis durant toute la session d'acquisition sans pouvoir prendre de pause, ce qui est particulièrement pénible. Aussi, le processus de fusion des données des différents systèmes est particulièrement chronophage. Puisque la caméra Realsense a été utilisée pour suivre l'ouverture/fermeture des yeux dans le corpus précédent, nous avons choisi de mettre des marqueurs sur les paupières des acteurs de l'Optitrack^{TM} pour simplifier le post-traitement. De ce fait, nous avons choisi d'utiliser le système Optitrack^{TM} tout seul.

L'Optitrack^{TM} est composé de huit caméras (Flex 13) avec une vitesse de captures de 120 frames par seconde. Nous avons organisé huit caméras autour d'un écran de manière à ce que le visage de l'actrice soit toujours visible lors de la lecture des phrases à l'écran comme présenté sur la figure 5.3. Soixante-trois marqueurs réfléchissants de diamètres $3mm$ et $4mm$ ont été collés sur son visage. Pour suivre les mouvements de la tête, nous avons collé des marqueurs de $9mm$ sur le bonnet que l'actrice a porté durant les enregistrements. Nous avons utilisé un microphone

6. L'outil de synthèse vocale SOJA : https ://raweb.inria.fr/rapportsactivite/RA2010/parole/uid60.html

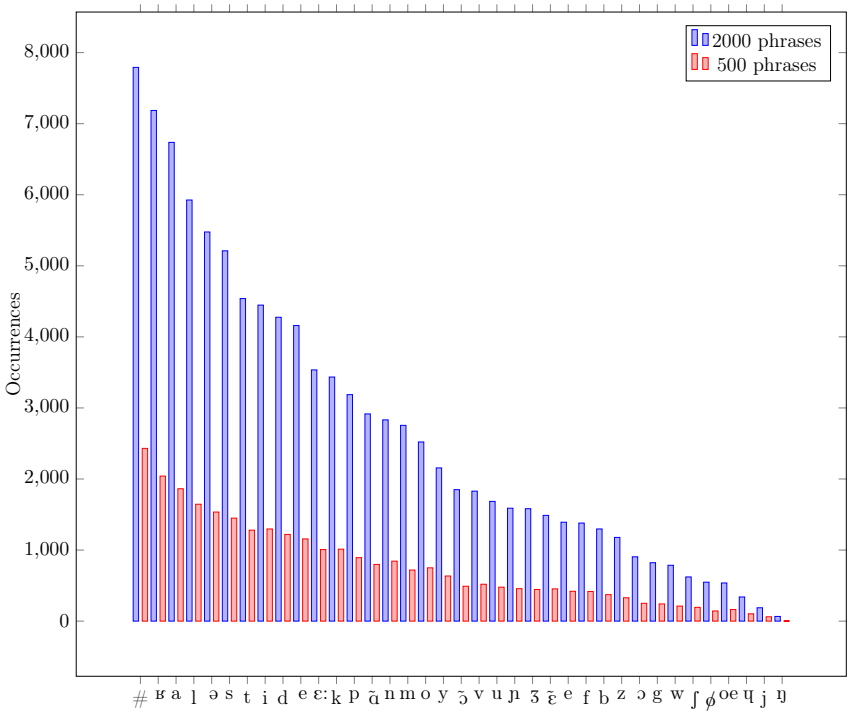


FIGURE 5.2 – *Le nombre d'occurrences des phonèmes dans le corpus de 2000 phrases et celui de 500 phrases.*

stéréo avec une fréquence d'échantillonnage de 48 KHz pour enregistrer la modalité acoustique. La disposition des marqueurs est présentée dans la figure 5.4. Une caméra vidéo a également été utilisée pour enregistrer l'ensemble de la performance.

Puisqu'il n'est pas possible d'enregistrer les 5000 (2000 neutres et 3000 expressives) en une seule session (calibration du système, pose des marqueurs, erreurs de prononciation / répétitions, pauses, ...), ce qui représente plus de 10h de parole, nous avons pensé à une solution pour garder la même disposition des marqueurs pour des différentes séances. Nous avons d'abord collé 63 marqueurs sur le visage de l'actrice puis nous en avons réalisé un scan 3D. Le choix des emplacements des marqueurs a été fait en s'inspirant de la norme MPEG-4 [Pandzic and Forchheimer, 2002]. Nous avons utilisé la texture du scan 3D pour identifier l'emplacement des marqueurs pour créer des trous dans le scan, puis nous l'avons imprimé en 3D comme présenté dans dans la figure 5.4-A. Ce masque a été utilisé au début de toutes les séances pour placer les marqueurs faciaux. Nous mettons ensuite 6 marqueurs sur la tête de l'actrice, leur disposition peut changer d'une séance à l'autre. Ce point n'est pas problématique puisqu'ils ne servent qu'à supprimer les mouvements de la tête. Quatre marqueurs ont été placés sur les paupières supérieures / inférieures pour suivre l'ouverture et le clignement des yeux, cette information n'a pas été utilisée dans ce travail.

Il était aussi important de garder la même configuration matérielle entre les différentes sessions, notamment la distance entre le micro et l'actrice. Comme nous pouvons voir sur la figure 5.3, nous avons placé des marqueurs blancs par terre pour marquer la position et la direction exactes de la chaise, le reste du matériel n'a pas été déplacé. Nous avons également pris les mesures nécessaires, pour pouvoir replacer le matériel au même endroit si nous souhaitons étendre le corpus avec plus de phrases ou avec de nouvelles émotions dans le futur. Nous avons effectué

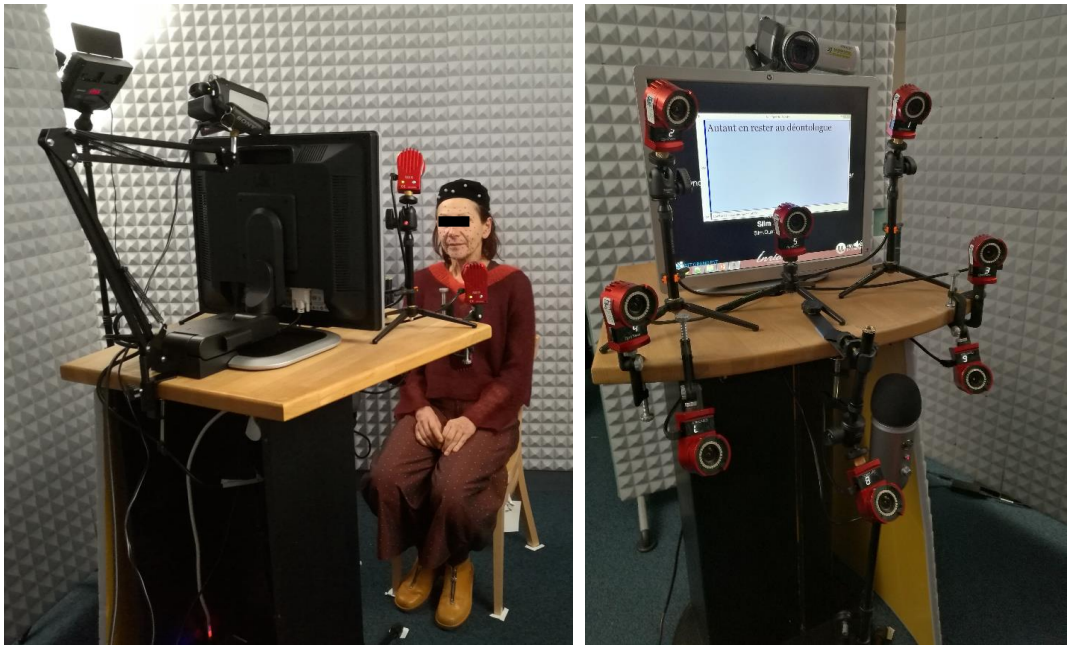


FIGURE 5.3 – *La configuration du système utilisé pour enregistrer le corpus expressif.*

quatre sessions d'acquisition, chaque session nous a permis d'enregistrer une partie du corpus neutre et quelques émotions. Les séances ont été organisées comme suit :

— Session 1 : enregistrement de 600 phrases neutres,
— Session 2 : enregistrement de 600 phrases neutres et 500 phrases de joie et 500 de tristesse,
— Session 3 : enregistrement de 400 phrases neutres, 500 de peur et 500 surprise,
— Session 4 : enregistrement de 400 phrases neutres, 500 colère et 500 phrases de dégoût.

Dans le chapitre précédent (chapitre 4), lors des tests perceptifs nous nous sommes aperçu que l'acteur et les participants ne partageaient pas la même définition de certaines émotions (dégoût par exemple). De ce fait, pour nous assurer d'une compréhension commune des émotions en utilisant une liste de scénario. Cette liste de scénario a été récupérée du corpus GEMEP [Bänziger and Scherer, 2010], elle est présentée dans l'annexe A. Cette liste contient trois scénarios possibles pour chaque émotion. Le but est de présenter ces scénarios à l'actrice, et c'est à elle de choisir le scénario qui lui parle le plus et donc le plus proche de sa mémoire affective. De cette manière l'actrice reproduira l'émotion de la même manière même si l'émotion n'est pas enregistrée sur une seule séance. Les scénarios choisis aideront aussi les utilisateurs pendant les tests perceptifs à mieux comprendre la définition des noms des émotions et ainsi d'éviter les confusions constatées lors de l'enregistrement du corpus précédant.

5.4 Post-traitement et alignement

La phase de post-traitement est passée par plusieurs étapes. Tout d'abord, nous avons traité la modalité visuelle pour supprimer le mouvement de la tête et unifier la position de la tête entre les différents enregistrements. Après cela, nous avons découpé les fichiers d'enregistrement contenant plusieurs phrases en fichiers contenant une seule phrase chacun. Ces derniers ont été utilisés dans le processus d'alignement texte-audio.

Annotation Semi-automatique des marqueurs : Cette étape consiste à utiliser le logiciel pro-

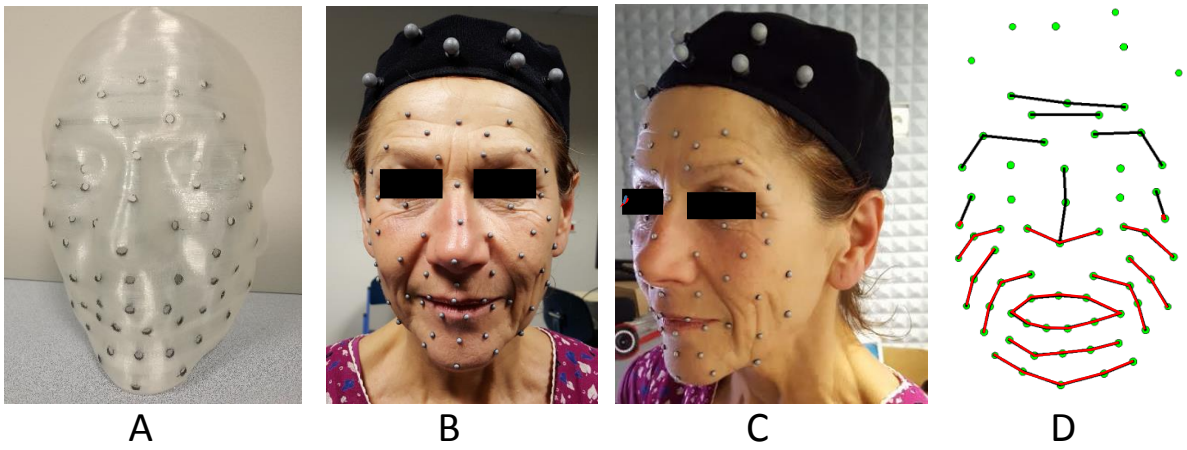


FIGURE 5.4 – *Les images B et C montent la configuration des marqueurs sur le visage et sur la tête de l'actrice vus de face et de profil. D montre la disposition des marqueurs après récupérations des données sous forme de trajectoires de points 3D, la zone marquée en rouge représente les marqueurs utilisés pour l'animation de la partie inférieure du visage. La figure A affiche le masque fabriqué pour garder la même position des marqueurs entre les différentes sessions d'enregistrement.*

priétaire de l'Optitrack™ nommé Motive pour attribuer une étiquette à chaque capteur. Cette phase est principalement automatique, mais certains problèmes doivent être corrigés manuellement (remplir les discontinuités lorsqu'un capteur disparaît, supprimer les faux marqueurs dus à des réflexions sur des objets, corriger les confusions entre deux capteurs proches etc).

Unification des données 3D : Cette tâche est cruciale pour garantir que la tête a la même position dans tous nos fichiers d'enregistrement. Pour chaque session, nous avons utilisé les 6 capteurs du chapeau pour supprimer les mouvements de la tête (comme nous l'avons expliqué dans le chapitre 4). De cette façon, nous obtenons une pose de tête statique le long de tous les fichiers, mais la tête peut avoir une pose de départ différente pour chaque fichier. Pour résoudre ce problème, nous choisissons une frame et nous l'utilisons comme référence pour remettre toutes les données des autres fichiers dans la même position spatiale.

Alignement text-audio : Après avoir généré la transcription phonétique de notre corpus textuel, nous avons utilisé Kaldi [7] pour aligner les phonèmes avec l'audio. Nous avons utilisé un modèle acoustique d'alignement entraîné avec des DNNs sur plus de 500 heures de parole française. Ce modèle a été utilisé pour générer un l'alignement phonétique entre le texte et l'audio.

Le tableau 5.1 montre les durées de chaque corpus après post-traitement. Les durées incluent les silences de début et de fin et les pauses de chaque phrase. Nous constatons que les différentes émotions ont des durées plus au moins proches mais que le dégoût a une durée égale à presque le double du reste des émotions. Contrairement au premier corpus où le dégoût représentait une émotion plus proche de la tristesse, dans ce corpus l'actrice a joué le dégoût comme un sentiment d'écœurement d'un objet répugnant qui peut provoquer l'aversion.

7. Kaldi est un outil open source pour l'alignement et la reconnaissance de la parole : https ://kaldi-asr.org/

	Nombre de phrases	Durées
Neutre	2000	4h :02min
Dégoût	500	1h :53min
Peur	500	1h :11min
Tristesse	500	1h :10min
Surprise	500	1h :04min
Joie	500	0h :58min
Colère	500	0h :55min

TABLE 5.1 – *La durées des données collectées par émotions. Ces durées correspondent aux données après nettoyage et découpage et contiennent les pauses et les silences de début et de fin de chaque phrase.*

5.5 Validation du corpus

Nous effectuons une évaluation perceptuelle pour déterminer si l'actrice a été capable de transmettre les différentes émotions correctement. Cette évaluation est essentielle pour valider la qualité du corpus audiovisuel expressif acquis. Nous avons effectué trois expériences perceptuelles portant sur les modalités visuelle et acoustique de notre corpus.

5.5.1 Stimuli

Nous avons utilisé les séquences vidéos que nous avons enregistrées en parallèle aux données 3D. Nous avons choisi 10 phrases avec le contenu linguistique le plus neutre possible et nous avons extrait les séquences audiovisuelles correspondantes pour chaque émotion. Trois types de stimuli ont été présentés aux participants pour chaque émotion : 1) stimuli acoustiques, 2) stimuli visuels et 3) stimuli audiovisuels.

5.5.2 Participants

Les expériences perceptuelles comptent une trentaine de participants naïfs pour chaque modalité : 1) 34 participants (20 hommes et 14 femmes) pour la modalité acoustique, 2) 31 participants (20 hommes et 11 femmes) pour la modalité visuelle et 3) 35 participants (23 hommes et 12 femmes) pour les tests audiovisuels. Les participants ne sont pas des français natifs, mais ils vivaient en France durant la période de l'étude.

5.5.3 Méthode

Nous avons mis en place une application web sur laquelle les participants peuvent se connecter pour effectuer les tests perceptifs. Une série de stimulus ont été présentés un par un et chaque participant devait choisir, selon lui, et parmi une liste de sept possibilités (neutre, joie, surprise, peur, colère, tristesse et dégoût) l'émotion exprimée dans les stimuli. Le participant devait sélectionner une réponse et valider pour pouvoir voir les prochains stimuli. Les participants avaient la possibilité de rejouer les stimuli autant de fois qu'ils le souhaitaient. Contrairement à l'expérience avec le corpus précédent, nous avons présenté aux participants, cette fois-ci, une définition de chaque émotion qui correspond au scénario choisi par l'actrice lors de l'enregistrement du corpus. Certains participants ont participé à deux ou trois expériences, dans ce cas ils ont dû respecter l'ordre suivant : 1) expérience avec des stimuli acoustiques uniquement, 2) expérience avec des stimuli visuels uniquement puis 3) expérience audiovisuelle.

5.5.4 Résultats

Comme pour la chapitre précèdent (chapitre 4), après la collecte des résultats des différentes expériences, nous avons calculé le niveau de significativité statistique en utilisant la valeur-p avec un test-t et nous avons corrigé les résultats obtenus avec la méthode de Holm–Bonferroni [Holm, 1979]. Pour chaque expérience, nous avons utilisé un degré de liberté égale au nombre de participants moins 1. Nous avons utilisé un alpha égal à 5% et un niveau de hasard de 14%. Nous avons ajouté le symbole (*) pour les résultats statistiquement significatifs et (-) pour les ceux statistiquement non significatifs. Les résultats sont présentés dans les tableau 5.2, 5.3 et 5.4. Nous pouvons constater que pour les trois expériences les taux de reconnaissance sont tous significatifs pour toutes les émotions. Nous remarquons également quelques confusions entre certaines émotions mais le taux de confusion reste statistiquement non significatif.

Pour le test acoustique, la peur et le dégoût ont été principalement confondus avec la tristesse et l'état neutre avec taux allant de 13% à 22%. La joie et la surprise ont aussi été légèrement confondues ensemble et également avec la colère. Il est à noter que la tristesse a été aussi confondue avec la peur.

		Émotion perçue (acoustique)						
		Colère	**Dégoût**	**Peur**	**Joie**	**Neutre**	**Tristesse**	**Surprise**
Émotion produite	**Colère**	**73.24(*)**	8.82(-)	3.82(-)	5.00(-)	3.24(-)	2.35(-)	3.53(-)
	Dégoût	4.41(-)	**48.82(*)**	8.53(-)	3.82(-)	17.35(-)	13.53(-)	3.53(-)
	Peur	10.00(-)	10.59(-)	**34.12(*)**	1.47(-)	16.76(-)	22.35(-)	4.71(-)
	Joie	15.59(-)	3.53(-)	5.00(-)	**50.00(*)**	7.35(-)	2.65(-)	15.88(-)
	Neutre	0.29(-)	1.76(-)	2.06(-)	2.94(-)	**81.18(*)**	8.53(-)	3.24(-)
	Tristesse	1.76(-)	2.65(-)	13.24(-)	1.18(-)	2.94(-)	**77.06(*)**	1.18(-)
	Surprise	9.71(-)	2.06(-)	3.53(-)	9.71(-)	3.53(-)	2.06(-)	**69.41(*)**

TABLE 5.2 – *La matrice de confusion du taux de reconnaissance des 7 émotions avec les stimuli acoustiques. Les lignes représentent la distribution des réponses données par les participants.*

En ce qui concerne les stimuli visuels, toutes les émotions sont mieux reconnues qu'avec la modalité acoustique seule la tristesse a moins été reconnue, le visuel de la tristesse semble plus proche du neutre alors sa modalité acoustique se distingue clairement de l'état neutre. Nous constatons que la peur et le dégoût sont plus portés par la modalité visuelle que par celle acoustique. Toutefois, plusieurs stimuli de la peur ont été confondus avec la surprise (même si le taux de la confusion est statistiquement non-significatif). Cette confusion peut s'expliquer par certains facteurs liés aux conditions de tournage. En fait, comme présenté précédemment, nous

		Émotion perçue (visuel)						
		Colère	**Dégoût**	**Peur**	**Joie**	**Neutre**	**Tristesse**	**Surprise**
Émotion produite	**Colère**	**80.00(*)**	4.84(-)	6.13(-)	0.65(-)	1.94(-)	1.29(-)	5.16(-)
	Dégoût	1.94(-)	**75.48(*)**	1.61(-)	2.58(-)	1.94(-)	15.81(-)	0.65(-)
	Peur	13.87(-)	1.94(-)	**63.55(*)**	0.00(-)	0.65(-)	0.32(-)	19.68(-)
	Joie	0.32(-)	0.32(-)	0.32(-)	**91.61(*)**	0.65(-)	0.00(-)	6.77(-)
	Neutre	0.00(-)	0.00(-)	1.29(-)	1.29(-)	**94.19(*)**	1.61(-)	1.61(-)
	Tristesse	0.32(-)	0.97(-)	8.39(-)	2.90(-)	28.71(-)	**56.45(*)**	2.26(-)
	Surprise	19.68(-)	0.65(-)	7.74(-)	0.32(-)	1.61(-)	1.61(-)	**68.39(*)**

TABLE 5.3 – *La matrice de confusion du taux de reconnaissance des 7 émotions avec les stimuli visuels. Les lignes représentent la distribution des réponses données par les participants.*

		Émotion perçue (audiovisuel)						
		Colère	**Dégoût**	**Peur**	**Joie**	**Neutre**	**Tristesse**	**Surprise**
Émotion produite	**Colère**	**92.57(*)**	2.00(-)	2.29(-)	0.00(-)	0.57(-)	0.29(-)	2.29(-)
	Dégoût	1.14(-)	**89.43(*)**	2.00(-)	0.29(-)	1.71(-)	3.43(-)	2.00(-)
	Peur	5.43(-)	3.43(-)	**73.43(*)**	0.29(-)	1.71(-)	3.43(-)	12.29(-)
	Joie	0.29(-)	0.57(-)	0.00(-)	**95.14(*)**	1.14(-)	0.00(-)	2.86(-)
	Neutre	0.00(-)	0.00(-)	0.57(-)	0.00(-)	**97.43(*)**	2.00(-)	0.00(-)
	Tristesse	0.57(-)	2.00(-)	4.86(-)	0.57(-)	1.14(-)	**90.86(*)**	0.00(-)
	Surprise	3.71(-)	0.86(-)	4.00(-)	0.86(-)	1.43(-)	0.29(-)	**88.86(*)**

TABLE 5.4 – *La matrice de confusion du taux de reconnaissance des 7 émotions avec les stimuli audiovisuels. Les lignes représentent la distribution des réponses données par les participants.*

avons enregistré notre corpus sur plusieurs séances. Durant ces séances nous avons configuré notre matériel et installation de telle sorte que la qualité du son et la position des points 3D soient identiques entre les différentes séances, toutefois pour l'aspect vidéo 2D, il y a d'autres facteurs qui sont entrés en compte. D'abord le changement des vêtements de l'actrice, puisque l'actrice ne s'est pas habillé de manière similaire durant les quatre séances, les participants ont probablement remarqué une corrélation entre les vêtements et les émotions jouées. L'autre facteur concerne les conditions de tournage, notamment l'éclairage que nous avons dû baisser durant certaines séances pour le confort visuel de l'actrice. La posture de l'actrice a été également très différente entre certaines émotions (penchée vers l'avant ou vers l'arrière). Toutefois, la présence de ces biais, ne semble pas affecter les résultats de tous les tests, en fait, la tristesse et la joie enregistrées dans des conditions de tournages similaires n'ont pas engendré de confusions notables, inversement, la colère enregistrée dans des conditions différentes de celles de la peur et la surprise a été pourtant remarquablement confondue avec ces dernières. À partir de ces constatations, nous pensons que les conditions de tournage peuvent avoir un impact sur les résultats des tests dès lors que les émotions confondues comportent déjà des similarités et ainsi que les conditions de tournages ne font qu'accentuer la confusion quand elle est déjà présente.

Il est intéressant de constater que la colère et la surprise sont aussi bien portées par la voix que par les expressions faciales et que la tristesse (contrairement à toutes les autres émotions) est plus exprimée par la voix que par les expressions faciales, contrairement au corpus précédent où c'était plutôt le cas de la colère. Nous remarquons aussi que grâce à l'introduction des scénarios pour partager une définition commune des émotions, le dégoût et la colère n'ont pas été confondues dans ce corpus.

Concernant les stimuli audiovisuels, les taux de reconnaissances sont très élevés (plus de 73%) pour toutes les émotions et montrent que les deux modalités acoustique et visuelle sont complémentaires. Ces résultats montrent que la majorité des participants valident la performance de l'actrice et confirment la bonne qualité du corpus expressif produit, nous pouvons donc l'utiliser pour des fins de synthèse expressive audiovisuelle de la parole.

5.6 Animation d'une tête parlante 3D

Le but de l'animation est de générer une séquence de poses du modèles 3D qui reproduisent la performance originale de l'acteur. Pour ce travail, un modèle 3D open-source d'un personnage féminin pour créer nos animations.

Dans ce travail de thèse nous animons la partie inférieure du visage qui correspond à la région des articulateurs (lèvres, joues et mâchoire et menton). Les marqueurs utilisés sont au nombre

de 44 points qui couvrent la partie inférieure du visage comme présenté sur la figure 5.4-D. Nous nous focalisons sur la partie inférieure du visage car les mouvements des articulateurs sont étroitement liés aux phonèmes produits, contrairement aux mouvements de la partie supérieure du visage, notamment les sourcils, qui dépendent d'autres facteurs. En fait, dans [Granström et al., 1999] et [House et al., 2001] les résultats des expériences indiquent que les mouvements des sourcils peuvent fonctionner comme un indice perceptuel de la proéminence d'un mot indépendamment des indices acoustiques et des indices visuels de la partie inférieure du visage. L'étude de Granström et al. [1999] montre aussi que les mouvements de sourcils ne sont pas simplement liés à la ponctuation de la phrase, mais pourraient être liés à la cohérence au sein d'une phrase. L'étude de Pelachaud et al. [1996] montre que l'élévation des sourcils peut être utilisée pour marquer une nouvelle information. De surcroît, les mouvements de la partie supérieure du visage ont peu d'influence sur les mouvements de la partie inférieure et vice versa [Ekman and Friesen, 1976, Donato et al., 1999, Zalewski and Gong, 2004], ce qui nous permet de traiter les mouvements de la partie inférieure du visage séparément de sa partie supérieure. Dans le chapitre suivant (chapitre 6) nous présentons les paramètres linguistiques dont nous disposons pour effectuer l'apprentissage des DNNs. Ainsi, il est possible de prédire les mouvements de la partie inférieure du visage à partir de ces données, alors que l'animation de la partie supérieure nécessite des informations prosodiques et sémantiques supplémentaires que nous n'avons pas.

Pour créer les animations 3D nous avons utilisé la technique de l'interpolation des *Blendshapes*. Nous rappelons qu'une *Blendshapes* est un modèle 3D qui représente le mouvement d'un muscle spécifique du visage. Au moment de l'animation, il faut calculer, pour chaque frame de données 3D de l'animation, un vecteur de poids de l'ensemble des *Blendshapes* utilisées. Une interpolation pondérée entre ces dernières génère la forme finale du modèle à un instant donné [Chuang and Bregler, 2002]. Ce processus est détaillé plus tard et nous le présentons sur la figure 5.5.

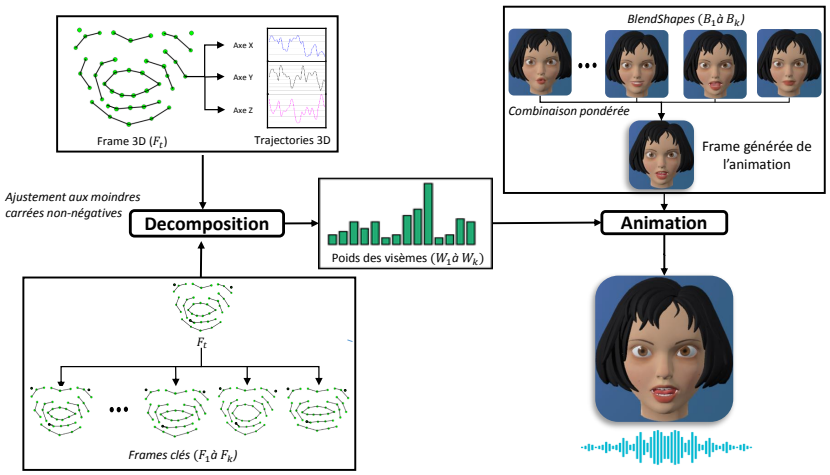


FIGURE 5.5 – *Le processus d'animation du personnage 3D en utilisant la technique de Chuang and Bregler [2002]. La frame 3D à un moment t est décomposée en un vecteur de poids en utilisant un ajustement de moindres carrées non-négatives. Ces poids sont ensuite affectés aux différentes Blendshapes qui seront interpolées pour former l'expression faciale du personnage à un moment t. La séquence des poses générées résulte en une animation fluide.*

Pour créer les *Blendshapes*, Benoit et al. [1992] proposent une liste de 17 visèmes comme

adaptation de la liste de 16 visèmes définie par la norme MPEG4 Pandzic and Forchheimer [2002] pour la langue française. Une autre liste de visèmes a été établie par Govokhina [2008] qui réduit la liste de *Blendshapes* à 8 visèmes seulement. Gonokhina a utilisé la distance de Bhattacharyya Mak and Barnard [1996] pour calculer les distances entre les distributions gaussiennes des cibles articulatoires des phonèmes correspondants. Les phonèmes les plus proches ont été regroupés pour former des catégories de visèmes. Ces deux groupes de visèmes ont été définis pour la langue française mais représentent différents niveaux de précision. Par exemple, dans la classification de Benoit et al. l'utilisation des articulateurs internes est prise en compte. Ainsi, les visèmes [t, d, n, ɲ, ŋ] se distinguent de [l] et [ʁ]. Pour la classification de Gonokhina, seuls les articulateurs externes / visibles ont été considérés, par conséquent [t, d, n, ɲ, ŋ, l, ʁ] sont regroupés. Concernant ce travail, les données des articulateurs non visibles n'ont pas été collectées. Ainsi, nous avons décidé de travailler avec la classification de Gonokhina et de l'affiner si besoin pour obtenir une animation de bonne précision et qui est fidèle au jeu de l'acteur. En fait, il n'existe pas de consensus sur la liste de visèmes optimale, pour l'anglais par exemple, il existe plus d'une quinzaine de classifications différentes [Bear and Harvey, 2017]. La liste des visèmes optimale est étroitement liée aux données elles mêmes et doit être modifiée et ajustée pour chaque nouveau corpus. Pour cette raison, l'objectif de la section suivante est de trouver la liste optimale de visèmes pour notre corpus.

Pour chaque visème de la liste de Gonokhina nous créons le modèle 3D correspondant (*Blendshape*) comme présenté sur la figure 5.5. L'ensemble de nos *Blendshapes* est donc $B = [B_1, ..., B_k]$. Chaque visème choisi correspond également à une frame (nuage de points 3D) de notre corpus. Ces frames clés seront notées $F = [F_1, ..., F_k]$ et permettrons de calculer le poids affecté à chaque *Blendshape* $W = [W_1, ..., W_k]$. Ce calcul de poids est basé sur le travail de Chuang and Bregler [2002]. En fait, les données visuelles sont décomposées en une combinaison pondérée de l'ensemble des frames clés :

$$F_t = \sum_{n=1}^{k} W_i F_i, W_i > 0 \tag{5.1}$$

Où k est le nombre des visèmes choisis et F_t est la frame des données visuelles à un moment t. F_i et W_i représentent la ième frame clé et son poids affecté. Cette décomposition est réalisée avec la méthode des moindres carrés non négatifs :

$$argmin_w ||F_t - F_w||_2,\ W \geq 0 \tag{5.2}$$

Où F_w est la frame résultante de la reconstruction de F_t en utilisant les poids calculés. Ce processus d'animation 3D par décomposition en poids de *Blendshapes* est présenté sur la figure 5.5.

Après avoir déterminer la série de poids, nous avons procédé à la reconstruction des trajectoires 3D à partir des poids des frames clés. L'algorithme de morphing utilisé dans notre travail est une combinaison linéaire (équation 5.1) où les poids sont déjà connus. Cette étape permet de vérifier la qualité de la reconstruction. En fait, après la reconstruction des fichiers 3D, nous calculons la RMSE et la corrélation de Pearson entre les données originales et celles reconstruites (voir Tableau 5.5).

Comme nous pouvons le constater dans le tableau 5.5, la reconstruction utilisant la liste de base de 8 visèmes n'était pas suffisante pour obtenir une bonne reconstruction. Nous avons donc identifié les visèmes avec la plus mauvaise reconstruction et nous avons ajouté un autre niveau de détails en créant de nouvelles *Blendshapes*. Les visèmes les plus problématiques sont ceux liés à la protrusion et pour lesquels le modèle de Gonokhina propose un même visème (y, u, ɸ, oe, o,

Mesure	Liste basique (8)	Liste enrichie (18)
RMSE_X	1.89891	0.86232
RMSE_Y	2.18001	0.97979
RMSE_Z	2.30564	1.08644
RMSE moyenne	2.12819	0.97618
PCorr_X	-0.11691	0.83264
PCorr_Y	0.15638	0.87013
PCorr_Z	0.28415	0.76353
PCorr moyenne	0.10787	0.82210

TABLE 5.5 – *RMSE (en mm) et corrélation de Pearson entre les données originales et celles reconstruites. Ces mesures ont été calculées pour les trois axes (X, Y, Z) puis nous avons calculé la valeur moyenne sur ces trois axes.*

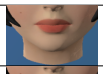
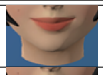
Visème	Phonème	BlendShape	Visème	Phonème	BlendShape
0	#, ə		7	i	
1	p, b, m		8	j	
2	f, v		9	o, ɔ, ɔ̃, ɑ̃	
3	ʃ , ʒ		10	y	
4	s, z		11	u	
5	t, d, n, ɲ, ŋ k, g, l, ʁ		12	w, ɥ	
6	a, e, ɛ, ɛ :, ɛ̃		13	ø, oe	
14	UA12		15	UA15	
16	UA9 et UA10		17	UA20	

TABLE 5.6 – *Liste des visèmes et de leurs phonèmes et UAs correspondants ainsi que leurs Blendshapes représentatives. Les symboles phonétiques sont tirés du International Phonetic Alphabet (IPA) Decker et al. [1999].*

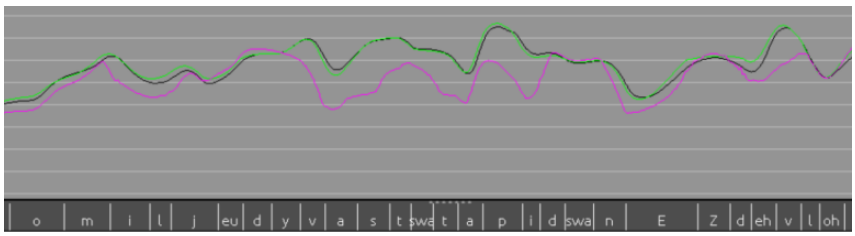


FIGURE 5.6 – *La trajectoire d'un capteur placé sur la lèvre inférieure sur l'axe y pour les données originales (en noir) et les données reconstruites. En rose : la reconstitution de la liste des visèmes de base (8 visèmes). En vert : la reconstitution de la liste des visèmes enrichis (18 visèmes).*

ɔ, ɔ̃, ɑ̃, w, ɥ). Pour obtenir des visèmes plus adaptés à chacun de ces sons, nous avons remplacé le visème unique par 5 visèmes différents (visèmes 9, 10, 11 , 12, 13 du tableau 5.6). Nous avons aussi créé un visème indépendant pour s et z puis pour i et j qui sont regroupés par Gonokhina dans un même visème. Pour prendre compte de la dimension expressive de notre corpus, et en se basant sur le système de codage des actions faciales (FACS) [Ekman et al., 2002], nous avons ajouté 4 unités d'actions (UAs) à notre liste de visèmes pour permettre au modèle 3D d'afficher différentes expressions sur la partie inférieure de son visage. La liste des différentes UAs et leurs explications sont présentées dans l'annexe B. La liste des visèmes conservée est celle présentée dans le tableau 5.6 contenant 18 *Blendshapes*. Comme nous pouvons le voir sur le tableau 5.5, la précision de reconstruction que nous avons obtenue avec notre liste de visèmes est très élevée pour les trois axes. Nous pouvons aussi voir une comparaison de la trajectoire 3D originale d'un capteur de lèvre inférieure sur l'axe y et le résultat de sa reconstruction en utilisant les deux listes de visèmes dans la figure 5.6, le résultat de notre liste suit très bien la trajectoire originale du capteur. Pour ces raisons, nous conservons cette liste de visèmes pour créer des animations 3D à partir des données de notre corpus. Cette liste de visèmes sera utilisée également pour créer des animations issues de données de synthèse.

5.7 Conclusion

Dans ce chapitre, nous avons présenté les différentes étapes effectuées pour l'acquisition et le post-traitement d'un corpus audiovisuel expressif. Nous avons consacré une partie importante du travail à la préparation et l'analyse de la composition linguistique de ce dernier pour qu'il soit phonétiquement équilibré. La réflexion sur la configuration du matériel impliqué a été réalisée dans le but de permettre un enregistrement sur plusieurs séances, ce qui est important si nous souhaitons étendre le corpus dans le futur. Nous avons aussi présenté le processus de post-traitement des données et de leur préparation pour la phase de synthèse ainsi que la technique adoptée pour générer des animations 3D à partir des données 3D du corpus.

En nous appuyant sur l'expérience d'acquisition du corpus du chapitre précédent (Chapitre 4) et en suivant les étapes citées plus haut, nous avons pu acquérir un corpus audiovisuel avec un contenu expressif correctement perçu même avec les modalités acoustique et visuelle présentées séparément. À ce niveau-là, nous estimons avoir mis tous les paramètres de notre côté pour avoir un corpus de bonne qualité dédié à la synthèse audiovisuelle expressive de la parole.

Dans les chapitres suivants, nous allons présenter nos approches de synthèse audiovisuelle de la parole basée sur les données de ce corpus.

Troisième partie

Synthèse audiovisuelle expressive de la parole

6

Synthèse audiovisuelle expressive par architecture entièrement connectée

Sommaire

6.1 Introduction

Dans ce chapitre [8] nous présentons l'architecture de référence que nous avons adoptée pour la synthèse audiovisuelle expressive de la parole. Dans ce travail nous utilisons le corpus que nous avons collecté et qui a été présenté dans la partie précédente de ce document (voir chapitre 5). Nous présentons d'abord les différents paramètres acoustiques et visuels utilisés et nous détaillons ensuite l'architecture entièrement connectée DNN-FC (FC pour *Fully-Connected*) adoptée pour l'entraînement du réseau. Nous finissons par exposer les mesures objectives des résultats obtenus par cette architecture que nous validons par une validation croisée sur les différentes émotions.

6.2 Préparation des paramètres d'entrée et de sortie

Quelques mois avant le début de ce travail de thèse, l'outil de synthèse acoustique à partir du texte et basé sur des réseaux de neurones nommé Merlin [Wu et al., 2016] a été publié. Le système prend les paramètres linguistiques en entrée et se base sur des réseaux de neurones pour prédire les durées des phonèmes et les caractéristiques acoustiques, qui sont ensuite transmises à

8. Les travaux présentés dans ce chapitre ont fait l'objet d'une publication dans la conférence JEP 2020 (https ://hal.archives-ouvertes.fr/hal-02798526/document) et une soumission d'un article à la conférence internationale Interspeech 2020.

un vocodeur pour produire un fichier audio. Diverses architectures de réseaux de neurones sont proposées sous forme de recettes, notamment un réseau DNN-FF et des réseaux de neurones récurrents tels que LSTM et BLSTM.

Merlin s'est beaucoup inspiré de l'outil HTS [Zen et al., 2007] qui est un outil de synthèse acoustique par HMMs. Merlin a repris sa manière de paramétriser les données textuelles et celles acoustiques. Comme HTS, l'outil Merlin n'est pas un système de synthèse acoustique complet. Il nécessite un vocoder (STRAIGHT ou WORLD) ainsi qu'un front-end externe (comme Festival ou Ossian) pour générer les données textuelles contextualisées nécessaires pour l'entraînement.

Paramètres linguistiques

La sortie du front-end doit être formatée sous forme de fichiers labels comme proposé par HTS. Les fichiers labels doivent contenir des informations temporelles précisant l'alignement phonétique entre le texte et l'audio. Merlin convertit ensuite ces fichiers labels en vecteurs de paramètres binaires et continus qui serviront d'entrée pour entraîner le réseau de neurones. Ces paramètres sont dérivées des fichiers labels à l'aide du fichier de "questions" proposé par HTS.

Puisque les front-end disponibles publiquement ne prennent pas en compte la langue française, nous avons utilisé le module NLP (Natural Language processing) de l'outil SOJA [9]. Ce dernier est un outil de synthèse par concaténation développé au sein de notre équipe. Cet outil contient un front-end intégré, et permet de générer pour chaque phonème les informations suivantes :

• Paramètres relatifs à l'alignement phonétique :
o Instant de début du phonème (t_début).
o Instant de fin du phonème (t_fin).

• Paramètres relatifs au phonème :
o Phonème courant (p_C).
o Phonème précédent (p_L) et suivant (p_R).
o Phonème avant précédent (p_LL) et après suivant (p_RR).
o Position du phonème courant dans la syllabe courante (calcul vers l'avant (p_posyl_fw) et vers l'arrière (p_posyl_bw)).

• Paramètres relatifs à la syllabe :
o Nombre de phonèmes dans la syllabe courante (syl_C_np).
o Nombre de phonèmes dans la syllabe précédente (syl_L_np) et suivante (syl_R_np).
o Voyelle centrale de la syllabe courante (syl_voy).
o Position de la syllabe courante dans le mot courant (calcul vers l'avant (syl_powo_fw) et vers l'arrière (syl_powo_bw)).
o Position de la syllabe courante dans la phrase courante (calcul vers l'avant (syl_posen_fw) et vers l'arrière (syl_posen_bw)).

• Paramètres relatifs au mot :
o Nombre de syllabes dans le mot courant (wo_C_n_syl).
o Nombre de syllabes dans le mot précédent (wo_L_n_syl) et suivant (wo_R_n_syl).
o Position du mot courant dans la phrase courante (calcul vers l'avant (wo_posen_fw) et vers l'arrière (wo_posen_bw)).

9. L'outil de synthèse vocale SOJA : https ://raweb.inria.fr/rapportsactivite/RA2010/parole/uid60.html

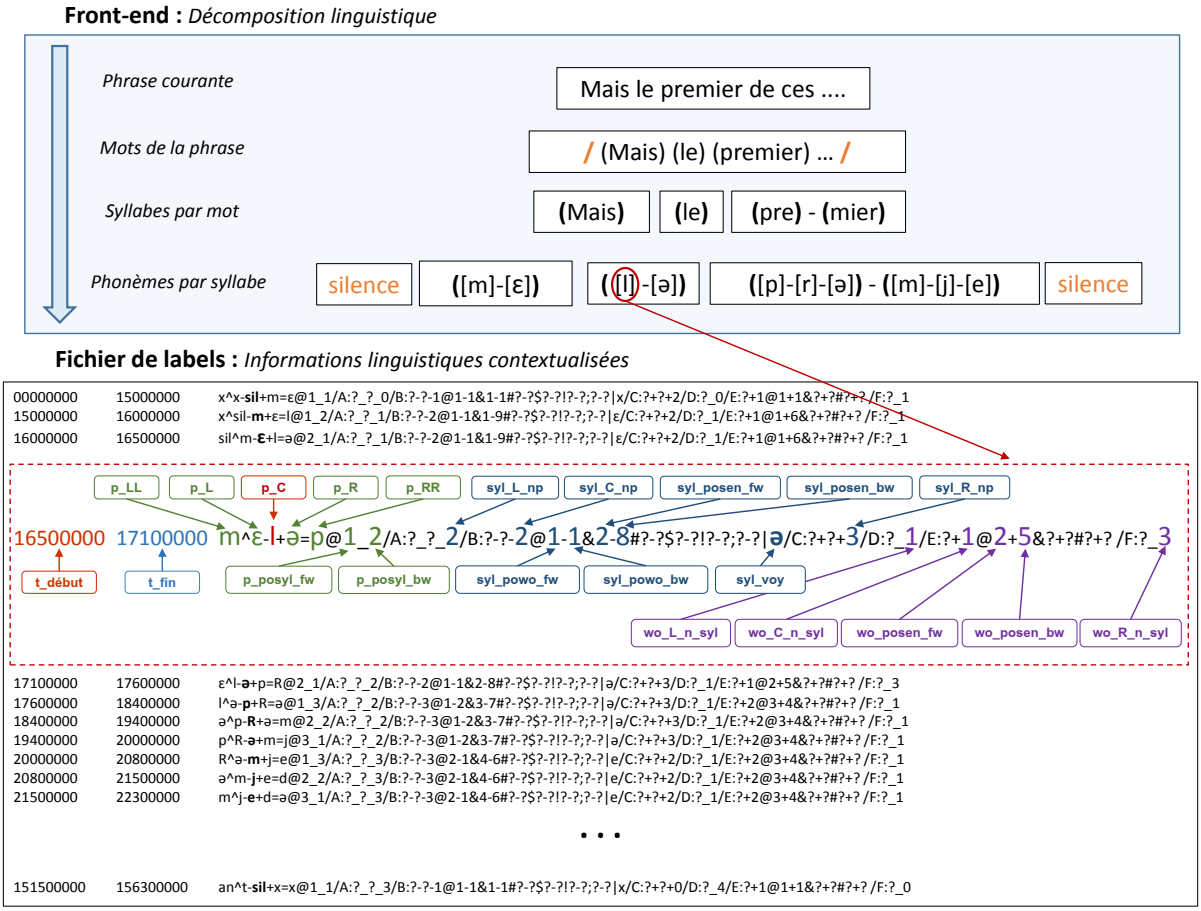


FIGURE 6.1 – *Composition du fichier de labels à partir de la décomposition des données linguistiques par un front-end. Les " ?" désignent des informations non renseignées.*

Ces informations générées par le *front-end* SOJA permettent de construire des fichiers de labels compatibles avec Merlin. Sa constitution est décrite dans la figure 6.1. Les deux premières colonnes contiennent l'instant de début et de fin des phonèmes. Le phonème courant et ses contextes gauches et droits sont ensuite présentés. La suite de chaque ligne contient les informations concernant la position du phonème dans la syllabe ainsi que des informations relatives aux syllabes et aux mots de la phrase courante comme décrit plus haut.

En utilisant le fichier de "questions", Merlin transforme les fichiers de labels en vecteurs binaires et numériques. Cette transformation se base sur un dictionnaire d'expressions régulières qui permet d'extraire les informations du fichier de labels. Pour la partie binaire, le but est d'obtenir un vecteur suivant l'encodage *one-hot* qui affecte un chiffre (0 ou 1) à chaque information linguistique utilisée. Si le phonème courant satisfait un certain nombre de conditions linguistiques et contextuelles, alors seuls les champs correspondants à ces conditions prennent la valeur 1, les autres prennent 0. Pour la partie numérique, les valeurs présentes dans le fichier des labels sont capturées par les expressions régulières et sont directement copiées dans le vecteur d'entrée.

D'autres informations binaires sont ajoutées au vecteur d'entrée. Ces informations sont relatives à l'appartenance ou non du phonème à une liste de 39 catégories spécifiées dans le fichier de questions, telles que la catégorie des voyelles, des consonnes, voyelles nasales, consonnes fricatives,

etc.

Au final, le vecteur contient 417 paramètres linguistiques et il est composé de :

— 190 paramètres binaires relatifs à la nature du phonème courant et de ses contextes gauches et droits (5x38 paramètres : 36 phonèmes et 2 codes supplémentaires, un pour les pauses et le deuxième pour les silences de début et de fin),
— 195 paramètres binaires relatifs à la catégorie phonétique du phonème courant et de ses contextes gauches et droits (5x39 paramètres),
— 2 paramètres numériques relatifs à la position du phonème courant dans la syllabe courante,
— 7 paramètres numériques relatifs à la syllabe courante précédente et suivante : le nombre de phonèmes qu'elles contiennent, la position dans la phrase et dans le mot courant,
— 18 paramètres binaires relatifs à la nature de la voyelle centrale dans la syllabe courante,
— 5 paramètres numériques relatifs aux nombres de syllabes dans le mot courant, précédent et suivant ainsi que la position du mot courant dans la phrase courante.

La séquence des vecteurs d'entrées (un pour chaque phonème) doit être alignée avec la séquence des vecteurs de sorties (acoustiques ou visuelles). Comme expliqué un peu plus bas dans cette section, les vecteurs des paramètres acoustiques sont extraits toutes les 5 millisecondes (ms). De ce fait, il faut appliquer la même fréquence d'échantillonnage sur les données linguistiques.

En se basant sur les deux premières colonnes du fichier des labels (instants de début et de fin de chaque phonème), la ligne correspondante à chaque phonème va être dupliquée autant de fois que sa durée comprend de frames de 5ms. La figure 6.2 donne un aperçu sur cette transformation.

L'outil Merlin TTS utilise deux modèles séparés pour la modélisation des durées et des paramètres acoustiques. Même si des travaux précédents ont étudié la possibilité d'une modélisation jointe des durées et des paramètres acoustiques [Watts et al., 2015a, Henter et al., 2016, Ronanki, 2019]. L'approche standard reste une modélisation séparée des deux aspects. De plus Ronanki [2019] déclarent que la performance d'une modélisation jointe sur la distorsion mel-cepstral et sur la corrélation de la F0 reste limitée puisque le modèle acoustique se base sur les informations des durées dans sa modélisation des paramètres acoustiques. Nous adoptons, donc, dans cette thèse l'approche standard comme montrée dans la figure 6.3.

Paramètres de durées

Un seul paramètre est considéré pour modéliser le paramètre de sortie du modèle des durées des phonèmes. Pour chaque ligne du fichier des labels (chaque phonème contextualisé), sa durée est extraite sous forme du nombre de frames qu'il couvre en considérant un pas de 5ms entre les frames consécutives.

Paramètres acoustiques

Nous avons utilisé le Vocodeur WORLD pour extraire les paramètres acoustiques suivants :

— 180 paramètres : 60 coefficients MFCC plus leurs deltas et delta-deltas,
— 15 paramètres : 5 paramètres BAP plus leurs deltas et delta-deltas,
— 3 paramètres : la fréquence fondamentale avec une échelle logarithmique (log F0) ainsi que son delta et delta-delta,
— 1 paramètre binaire pour préciser la nature voisée/non-voisée du son dans chaque frame.

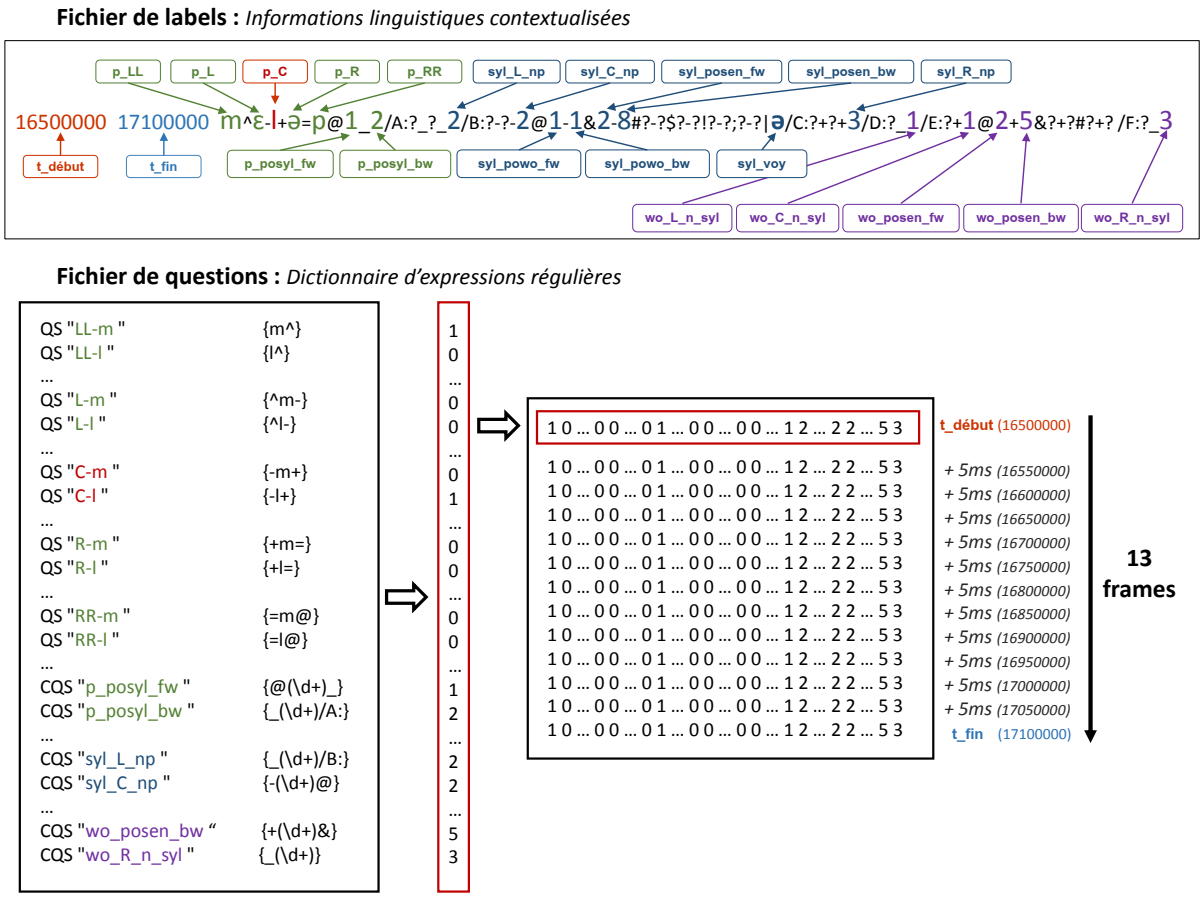


FIGURE 6.2 – *Exemple illustrant la transformation de chaque ligne du fichier des labels de la figure 6.1 en séquence de vecteurs de valeurs binaires et numériques. Cette transformation se fait en respectant la durée de chaque phonème et en se basant sur la fréquence d'échantillonnage des données acoustiques.*

Ces 199 paramètres ont été extraits des fichiers audio toutes les 5ms. Ces paramètres représentent la sortie du DNN qui sera entraîné pour prédire les paramètres acoustiques à partir des paramètres linguistiques.

Paramètres visuelles

Puisque le processus de paramétrisation résulte souvent en une perte d'information, nous avons décidé d'utiliser les paramètres visuels sous leurs forme brute pour conserver la totalité de l'information visuelle.

Les paramètres visuels utilisés dans ce travail sont les coordonnées X, Y et Z de chacun des 44 points de la partie inférieure du visage, soient 132 paramètres visuels comme montré dans la figure 5.5 du chapitre précédent (chapitre 5). Pour avoir le même nombre de frames que les deux autres aspects (durée et acoustique), nous avons augmenté la fréquence d'échantillonnage des données visuelles de 120fps (soit une frame toutes les 8.33ms) à 200fps (soit une frame toutes les 5ms) par une interpolation linéaire. En fait, l'étude de Trouvain et al. [2001], effectuée sur un corpus

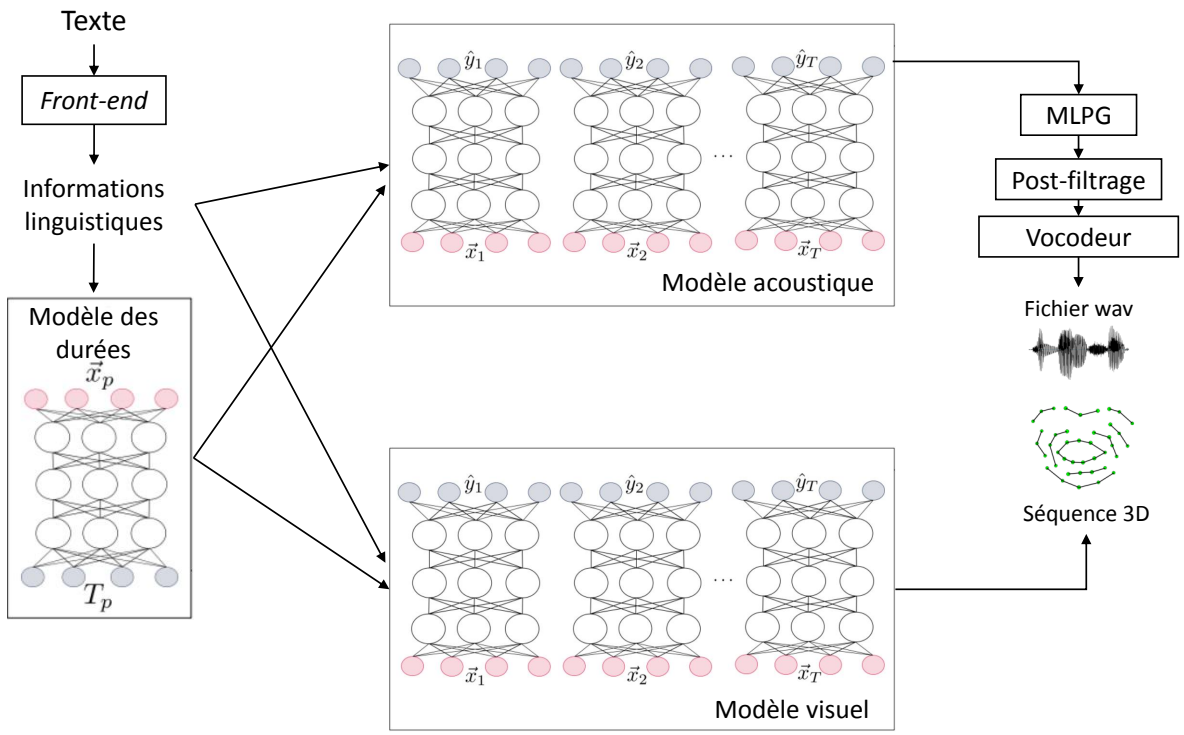


FIGURE 6.3 – *Synthèse audiovisuelle par modèles des durées, acoustiques et visuels séparés. Illustration inspirée de Fonseca De Sam Bento Ribeiro [2018]. Le modèle des durées génère les durées Tp pour chaque phonème à partir de ses spécifications linguistiques $\vec{x}_p$. Les modèles acoustique et visuel génèrent les paramètres de sortie pour chacune des T frames de la phrase et sont appliqués en boucle sur l'ensemble des frames de l'entrée $\vec{x}_t$.*

allemand, montre que le débit de la parole est d'environ 13 phonèmes par seconde en moyenne, soit à peu près un phonème toutes les 77ms. Dans le travail de Philippou-Hübner et al. [2012], le débit de la parole moyen de différentes émotions varie entre 9.9 phonèmes/seconde (dégoût et tristesse) et 16.5 phonèmes/seconde (peur) soit au maximum un phonème toute les 61ms. La fréquence d'échantillonnage de base de l'OptiTrack (120fps) permet d'obtenir au minimum 7 frames par phonème, ce qui permet de couvrir de manière correcte les gestes des articulateurs, surtout que notre corpus contient une articulation propre. De plus, l'étude que nous avons menée précédemment [Ouni and Dahmani, 2016] montre que les données articulatoires à 250fps et celles à 100fps sont hautement corrélées (91% pour l'axe X, 99% pour l'axe Y et 98% pour l'axe Z), et confirme ainsi que la perte d'information est minimale avec des fréquences d'échantillonnage aussi élevées. Il est donc peu probable de manquer une frame importante à cette vitesse de capture. Ainsi, les frames interpolées ne devraient pas nuire à la trajectoire des données originales.

6.3 Présentation de l'architecture utilisée

Dans ce chapitre nous considérons une architecture entièrement connectée DNN-FC. La sortie de chaque couche constitue l'entrée de la couche suivante et nous n'avons aucun contrôle sur la représentation intermédiaire des données à l'intérieur du réseau. Nous entraînons les trois aspects

de la parole avec trois modèles séparés :

1. Modèle des durées,
2. Modèle acoustique,
3. Modèle visuel.

Pour chaque modalité, nous considérons deux architectures : une architecture avec un réseau de type DNN-FF et une autre avec un réseau de type BLSTM. Les détails de ces architectures sont présentés dans la section 6.5.

Nous avons découpé le corpus pour définir trois sous-ensembles et chaque sous-ensemble est constitué de fichiers sélectionnés de manière aléatoire sur l'ensemble des données du corpus :

— Un ensemble d'apprentissage : constitué de 80% des données, soit 1600 sur 2000 pour l'état neutre et 400 sur 500 pour les émotions,
— Un ensemble de validation : constitué de 10% des données, soit 200 sur 2000 pour l'état neutre neutre et 50 sur 500 pour les émotions,
— Un ensemble de test : constitué de 10% des données, soit 200 sur 2000 pour l'état neutre neutre et 50 sur 500 pour les émotions.

Nous utilisons Keras [10], la librairie d'apprentissage profond de haut niveau. Keras est une librairie Python qui encapsule l'accès aux fonctions proposées par plusieurs librairies de machine learning, en particulier Tensorflow [11]. Ce dernier est une bibliothèque open-source qui implémente des méthodes d'apprentissage automatique basées des réseaux de neurones profonds.

La mise en correspondance des paramètres linguistiques avec les paramètres de sorties (durées, acoustiques ou visuels) constitue un problème de régression. Nous utilisons le carré moyen des erreurs MSE comme fonction d'erreur à minimiser en utilisant le principe de rétropropagation [Rumelhart et al., 1986] :

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2 \tag{6.1}$$

Où N est le nombre de paramètres de sortie, $\hat{y}_i$ est le i-ème paramètre prédit et y_i représente la donnée de sortie originale correspondante.

Il faut noter aussi que pour les trois modèles, l'entraînement s'est fait sur l'ensemble des phonèmes en plus des pauses et des silences de début et de fin. Mais au moment du calcul des métriques, nous avons écarté les pauses et les silences puisque ces derniers avaient un impact fort sur le taux d'erreur global du système (notamment pour le modèle des durées). Le manque de précision dans la prédiction des pauses et des silences est dû au fait que ces derniers ne suivent pas une règle précise d'articulation. Leurs durées ne sont pas fixes, celles des pauses peuvent dépendre de la prosodie et celles des silences peuvent dépendre de la manière dont les fichiers ont été découpés en phrases isolées. L'aspect acoustique peut contenir des sons de respiration et pour l'aspect visuel, durant les pauses et les silences, l'état de la bouche n'est pas connu, la locutrice peut garder la bouche ouverte entre les phrases, comme elle peut fermer la bouche pour avaler sa salive à l'intérieur d'une même phrase. Tous ces aspects rendent la prédiction des paramètres relatifs aux silences et aux pauses propices aux erreurs.

10. Librairie Keras : https ://keras.io/api/
11. Librairie Tensorflow : https ://www.tensorflow.org/

6.4 Mesures objectives

Nous présentons ici un ensemble d'expériences permettant d'apporter des réponses et des éclaircissements sur le comportement des DNNs face aux données linguistiques et audiovisuelles. Nous avons utilisé le vocodeur WORLD pour générer les paramètres acoustiques des données originales de test et nous présentons plus bas les mesures objectives pour mesurer la qualité de prédiction de nos modèles.

Erreur quadratique moyenne (RMSE) : Cette mesure est souvent utilisée pour mesurer la déviation entre les valeurs originales et celles prédites par un modèle. Dans cette thèse, elle est utilisée pour les durées, la F0 et les trajectoires visuelles. Elle est définie comme suit :

$$RMSE = \sqrt{\frac{\Sigma_{t=1}^{T}(x_t - \hat{x}_t)^2}{T}} \tag{6.2}$$

Où x_t et $\hat{x}_t$ dénotent les valeurs originales et prédites par nos modèles à un instant t, respectivement. La RMSE est calculée directement pour les durées en nombre de frames par phonèmes sur l'ensemble des phonèmes (à part les pauses et les silences). Pour les trajectoires visuelles, la RMSE est obtenue en millimètres d'écart. Il est d'abord calculé pour chaque fichier, pour chaque capteur et pour chacun des trois axes. Nous moyennons ensuite sur l'ensemble des fichiers, puis sur l'ensemble des capteurs, puis sur l'ensemble des axes pour obtenir la valeur finale. Pour la F0, la RMSE a été calculée sur l'échelle linéaire et non l'échelle logarithmique utilisée pour modéliser les valeurs de la F0.

Coefficient de corrélation linéaire (CORR) : Elle mesure la force et la direction de la relation linéaire entre deux variables. Elle est utilisée pour mesurer les performances de modélisation des durées, de la F0 et des trajectoires visuelles.

$$CORR = \frac{T\sum x_t\hat{x}_t - \sum x_t \sum \hat{x}_t}{\sqrt{T\sum x_t^2 - (\sum x_t)^2}\sqrt{T\sum \hat{x}_t^2 - (\sum \hat{x}_t)^2}} \tag{6.3}$$

Où x_t et $\hat{x}_t$ dénotent la valeur de référence et celle prédite respectivement. Si x_t et $\hat{x}_t$ ont une corrélation linéaire forte, CORR doit être proche de +1. Une valeur de CORR égale à +1 indique un ajustement positif parfait. Si x_t et $\hat{x}_t$ ont une corrélation linéaire négative forte, CORR doit être proche de -1. Une valeur de CORR égale à -1 indique un ajustement négatif parfait. Une valeur de CORR égale à 0 indique qu'il n'y a pas de corrélation entre les données.

Erreur sur les sons voisés/non-voisés (V/NV) : Cette mesure est souvent calculée avec la F0-RMSE car les valeurs de la F0 sont calculées par interpolation dans les régions non-voisées pour des raisons de modélisation. La décision sur le voisement est sauvegardée et plus tard prédite pour être utilisée pour la prédiction finale des valeurs de la F0. Par conséquent, l'erreur sur le voisement est calculée en pourcentage de frames étiquetées faussement comme voisée/non-voisée erronée.

La distorsion mel-cepstrale (MCD) : Kubichek [1993] mesure en décibels l'écart entre le vecteur original et celui prédit dans l'espace mel-cepstral. Elle est définit comme suit :

$$MCD = \frac{10}{Tln10}\sum_{t=1}^{T}\sqrt{2\sum_{d=1}^{D-1}(x_d(t) - \hat{x}_d(t))^2} \tag{6.4}$$

Où T est le nombre total de frames dans l'ensemble de test et D est la dimension des coefficients mel-cepstraux prédits pour chaque frame. Dans cette thèse, nous utilisons 60 coefficients pour chaque frame.

La distorsion de la bande d'apériodicité (BAPD) : Fonseca De Sam Bento Ribeiro [2018] mesure en décibels la distorsion des coefficients de la bande d'apériodicité. Elle est définit comme suit :

$$BAPD = \frac{1}{10T} \sum_{t=1}^{T} \sqrt{\sum_{d=1}^{D} (x_d(t) - \hat{x}_d(t))^2} \tag{6.5}$$

BAPD suit la même logique et les mêmes notations que la MCD. Pour chaque frame de l'ensemble de test, un vecteur de dimension D de paramètres BAP est prédit. La distorsion est ensuite calculée entre les paramètres originaux et ceux prédits. BAPD est calculée dans cette thèse sur les 25 coefficients de bande d'apériodicité.

6.5 Influence des paramètres linguistiques sur la qualité de la synthèse neutre

Dans cette section nous allons étudier l'impact des différentes informations linguistiques sur la qualité de la prédiction des données de durées, acoustiques et visuelle. Des études similaires ont été menées dans le passé sur des systèmes HMMs [Watts et al., 2010, Maguer et al., 2013, Cernak et al., 2013] mais peu d'études se sont intéressé à l'impact des paramètres linguistiques sur un système basé sur les DNNs et encore moins leur impact sur la modélisation de la parole visuelle. Maguer et al. [2013] ont étudié l'apport des différents paramètres linguistiques à la qualité de la synthèse du système HTS basé sur des HMMs. Cette étude menée sur un corpus acoustique de langue française a montré que l'utilisation du contexte phonétique améliore la modélisation du spectre de la parole et des durées, et que l'utilisation des informations sur les syllabes améliore la modélisation de la F0. Toutefois, le reste des facteurs contextuels semblent apporter une amélioration significative à la modélisation acoustique avec HTS. Cernak et al. [2013] ont également étudié les facteurs contextuels des données linguistiques pour la synthèse vocale par HMMs. Cette étude confirme que le contexte syllabique fait partie des facteurs contextuels les plus importants et que le contexte relatif aux mots de la phrase a peu d'importance comme préalablement établi dans l'étude de [Yu et al., 2010].

Pour la synthèse vocale par DNNs, Ribeiro et al. [2016] utilisent différents niveaux de contextes linguistiques pour entraîner un réseau de neurones à propagation vers l'avant. Les paramètres suprasegmentaux ont été traités par un DNN agissant au niveau des syllabes, et la sortie (sous forme de paramètres acoustiques) de ce dernier a été intégrée en tant qu'entrée supplémentaire à un DNN standard agissant au niveau des *frames*. Cette étude montre que l'ajout d'une représentation pré-entraînée des paramètres suprasegmentaux est bénéfique pour la modélisation acoustique. Par ailleurs, l'ajout des vecteurs de plongement (*embedding*) appris sur des mots ne montre aucune amélioration des performances du DNN. Récemment, Mametani et al. [2019] ont présenté une étude des paramètres contextuels appris automatiquement par un système de synthèse dit *End-to-End*. Ce genre de systèmes se base sur des DNNs et prend en entrée un texte brut (ou sa représentation phonétique) pour le convertir directement en paramètres vocaux. Les résultats expérimentaux montrent que les informations apprises par le réseau reflètent à la fois les contextes linguistiques et phonétiques, tels que l'identité du phonème et ses voisins gauches et droits, la réduction des voyelles, le stress et la position des syllabes dans le mot.

Pour analyser l'impact des paramètres linguistiques sur l'apprentissage des modèles de durées, acoustique et visuel sur les données neutres, nous testons différents paramètres d'entrées :

— *1_cont* : Uniquement l'information sur le phonème central ;

— *3_ cont* : L'information sur le phonème central son contexte gauche et droit immédiat ;
— *5_ cont* : L'information sur le phonème central ses deux contextes gauches et ses deux droits ;
— *5_ cont_ p* : Même informations que *5_ cont* avec en plus des informations sur la position du phonème courant et la catégorie phonétique des cinq phonèmes du contexte ;
— *5_ cont_ p_ s* : Même informations que *5_ cont_ p* avec en plus les informations sur les syllabes ;
— *5_ cont_ p_ s_ m* : Même informations que *5_ cont_ p_ s* avec en plus les informations sur les mots ;

Dans cette section nous utilisons uniquement les données du corpus neutre et adoptons deux architectures : une avec des DNN-FF et une autre avec des BLSTMs. Pour ces deux architectures, un DNN à deux couches a été retenu et nous avons essayé plusieurs largeurs de DNNs pour les différents vecteurs d'entrée (256, 512, 1024 et 2048). Les trois modèles ont été entraînés séparément et l'architecture qui donne le meilleur résultat sur l'ensemble de validation a été retenue pour chaque modèle et chaque configuration. Les meilleurs modèles ont été sélectionnés avec la technique du *early stopping*. La fonction d'activation des couches cachées est tanh et une fonction d'activation linéaire pour la couche de sortie. Nous avons utilisé l'optimiseur Adam et aucun dropout, BatchNorm ou régularisation spécifique n'a été utilisée.

Le calcul des différentes métriques a été effectué entre les paramètres prédits et ceux provenant du corpus original. Pour le modèle acoustique et visuel, les durées utilisées sont celles provenant du corpus original. Nous affichons la moyenne et les intervalles de confiance pour chaque métrique.

Nous présentons ci-dessous les résultats obtenus pour les trois modèles sur l'ensemble de test des données neutres contenant 200 phrases.

Modèle des durées

Le modèle des durées prend en entrée les informations linguistiques contextualisées et génère le nombre de frames relatif à chaque phonème. Nous calculons sur l'ensemble de test la CORR et RMSE en frames/phonème entre les durées prédites et les données originales.

	DNN-FF					
	1_cont [256, 256]	3_cont [256, 256]	5_cont [512, 512]	5_cont_p [512, 512]	5_cont_p_s [512, 512]	5_cont_p_s_m [512, 512]
RMSE (f/p)	8.672 (±0.018)	5.888 (±0.007)	5.532 (±0.006)	5.435 (±0.009)	5.259 (±0.008)	**5.256** (±0.007)
CORR	0.389 (±0.002)	0.781 (±0.0008)	0.811 (±0.0004)	0.822 (±0.0007)	0.827 (±0.0007)	**0.827** (±0.0007)
	BLSTM					
	1_cont [256, 256]	3_cont [256, 256]	5_cont [512, 512]	5_cont_p [512, 512]	5_cont_p_s [512, 512]	5_cont_p_s_m [512, 512]
RMSE (f/p)	7.237 (±0.011)	5.418 (±0.006)	5.413 (±0.006)	5.411 (±0.006)	5.301 (±0.007)	**5.247** (±0.006)
CORR	0.639 (±0.002)	0.821 (±0.0007)	0.824 (±0.00006)	0.826 (±0.0006)	0.827 (±0.0006)	**0.827** (±0.0006)

TABLE 6.1 – *Les résultats du RMSE en frames/phonème et de CORR sur l'ensemble de test générés par le modèle de durées sur les différentes configurations linguistiques lors de l'entraînement avec une architecture de type DNN-FF et BLSTM*

Modèle acoustique

Le modèle acoustique prend en entrée les informations linguistiques contextualisées et génère les paramètres acoustiques correspondants. Nous calculons ensuite sur l'ensemble de test la distorsion mel-cepstrale (MCD) et la distorsion de la bande-apériodicité (BAPD), le RMSE (F0-

RMSE) et la corrélation (F0-CORR) de la F0 ainsi que le pourcentage d'erreur sur la prédiction des frames voisées/non-voisées (V/NV).

	DNN-FF					
	1_cont [256, 256]	3_cont [512, 512]	5_cont [1024, 1024]	5_cont_p [1024, 1024]	5_cont_p_s [1024, 1024]	5_cont_p_s_m [1024, 1024]
MCD (dB)	6.653 (±0.026)	6.136 (±0.025)	6.132 (±0.024)	5.910 (±0.024)	5.901 (±0.024)	**5.900** (±0.024)
BAPD (dB)	0.327 (±0.004)	0.295 (±0.004)	0.292 (±0.004)	0.295 (±0.003)	0.288 (±0.003)	**0.287** (±0.003)
F0-RMSE (Hz)	35.226 (±1.105)	32.447 (±1.132)	31.334 (±1.106)	31.360 (±0.757)	30.648 (±0.733)	**30.555** (±0.743)
F0-CORR	0.341 (±0.021)	0.481 (±0.018)	0.529 (±0.017)	0.526 (±0.016)	0.557 (±0.016)	**0.563** (±0.015)
V/N-V (%)	14.250 (±0.424)	13.195 (±0.539)	13.090 (±0.553)	12.877 (±0.371)	12.872 (±0.375)	**12.848** (±0.371)
	BLSTM					
	1_cont [256, 256]	3_cont [512, 512]	5_cont [1024, 1024]	5_cont_p [1024, 1024]	5_cont_p_s [1024, 1024]	5_cont_p_s_m [1024, 1024]
MCD (dB)	5.304 (±0.029)	**5.099** (±0.023)	5.152 (±0.023)	5.103 (±0.025)	5.106 (±0.026)	5.146 (±0.024)
BAPD (dB)	0.282 (±0.004)	**0.242** (±0.003)	0.245 (±0.003)	0.242 (±0.003)	0.247 (±0.003)	0.247 (±0.003)
F0-RMSE (Hz)	32.580 (±0.818)	**27.934** (±0.622)	29.010 (±0.624)	28.460 (±0.690)	28.201 (±0.648)	28.207 (±0.865)
F0-CORR	0.471 (±0.016)	**0.640** (±0.013)	0.620 (±0.013)	0.628 (±0.014)	0.639 (±0.013)	0.637 (±0.014)
V/NV (%)	10.736 (±0.369)	**8.348** (±0.262)	8.822 (±0.257)	8.571 (±0.293)	8.566 (±0.303)	8.755 (±0.292)

TABLE 6.2 – *Les résultats sur l'ensemble de test générés par le modèle acoustique sur les différentes configurations linguistiques lors de l'entraînement avec une architecture de type DNN-FF et BLSTM.*

Modèle visuel

Le modèle visuel prend en entrée les informations linguistiques contextualisées et génère les paramètres visuels correspondants. Le modèle visuel génère une séquence de frames contenant les coordonnées X, Y et Z des 44 points impliqués dans le processus de l'entraînement. Nous calculons sur l'ensemble de test la corrélation de Pearson et le RMSE en millimètres entre les trajectoires visuelles prédites et originales.

	DNN-FF					
	1_cont [256, 256]	3_cont [256, 256]	5_cont [512, 512]	5_cont_p [1024, 1024]	5_cont_p_s [1024, 1024]	5_cont_p_s_m [1024, 1024]
RMSE (mm)	1.760 (±0.031)	1.458 (±0.029)	1.429 (±0.030)	1.427 (±0.030)	1.424 (±0.029)	**1.423** (±0.029)
CORR	0.574 (±0.007)	0.763 (±0.005)	0.778 (±0.006)	0.778 (±0.005)	0.779 (±0.005)	**0.779** (±0.005)
	BLSTM					
	1_cont [256, 256]	3_cont [256, 256]	5_cont [512, 512]	5_cont_p [1024, 1024]	5_cont_p_s [1024, 1024]	5_cont_p_s_m [1024, 1024]
RMSE (mm)	1.327 (±0.030)	**1.316** (±0.029)	1.328 (±0.031)	1.330 (±0.030)	1.331 (±0.031)	1.332 (±0.031)
CORR	0.823 (±0.005)	**0.828** (±0.005)	0.822 (±0.005)	0.822(±0.005)	0.821 (±0.005)	0.821 (±0.005)

TABLE 6.3 – *Les résultats sur l'ensemble de test générés par le modèle visuel sur les différentes configurations linguistiques lors de l'entraînement avec une architecture de type DNN-FF et BLSTM.*

Dans les tableaux 6.1, 6.2 et 6.3, il est très intéressant de constater qu'en utilisant une architecture DNN-FF, l'ajout de toutes les informations contextuelles améliore la qualité de la synthèse pour les trois aspects de la parole.

Toutefois, pour le réseau de type BLSTM les trois modèles n'ont pas tous le même comportement face aux informations linguistiques. Pour le modèle des durées, et de manière similaire au DNN-FF, l'ajout de l'ensemble des informations linguistiques améliore la prédiction des durées. Concernant les modèles acoustique et visuel, le réseau BLSTM atteint la meilleure qualité de synthèse avec les contextes gauche et droit immédiats uniquement. Cela peut s'expliquer par la capacité des BLSTMs du fait de leur architecture à accéder automatiquement aux contextes

passés et futurs de la frame courante, contrairement aux DNN-FF qui nécessitent que cette information soit explicitement donnée en entrée. Nous remarquons aussi pour BLSTM que pour le modèle acoustique, les informations sur la position et la catégorie des phonèmes ainsi que les informations sur les syllabes améliorent la modélisation de la F0, alors que l'ajout des informations relatives aux mots a un impact presque nul sur les mesures objectives. Ces constatations confirment les résultats des études précédentes [Maguer et al., 2013, Cernak et al., 2013, Yu et al., 2010]. Par ailleurs, force est de constater que, pour le modèle visuel avec BLSTM, l'ajout des informations contextuelles autres que les contextes gauche et droit immédiats n'améliore pas la qualité de l'apprentissage. Ce comportement peut être expliqué par le fait que la durée est une composante forte de la prosodie qui est fortement impactée par l'information sur les syllabes, ce qui explique l'importance de cette information pour la prédiction des durées. Une autre hypothèse concernant ce comportement, peut être liée la réduction du nombre d'exemples d'apprentissage avec l'augmentation du nombre de combinaisons possibles dans le vecteur d'entrée. En réalité, l'ajout de plus de contraintes contextuelles divise les données en classes de plus en plus petites, et réduit de ce fait le nombre d'exemples d'apprentissage de chaque classe. Pour le modèle des durées, ce comportement ne semble pas se produire. Nous pensons que cela vient du fait que le réseau doit prédire un seul et unique paramètre, qui est une tâche plus simple et qui nécessite donc moins d'exemples d'apprentissage.

6.6 Entraînement audiovisuel joint pour la synthèse expressive

Dans cette section nous étudions l'apport éventuel d'un entraînement joint des modalités acoustique et visuelle sur la qualité de la synthèse audiovisuelle expressive. Nous incluons les six catégories d'émotions dans le processus d'apprentissage et nous utilisons *3_ cont* comme informations linguistiques. Le vecteur de sortie pour le modèle joint est le résultat de la concaténation des paramètres acoustiques et visuels.

Nous avons entraîné toutes les émotions et l'état neutre conjointement en utilisant les étiquettes des émotions. Le calcul des différentes métriques a été effectué entre les paramètres prédits et ceux provenant du corpus original. Pour le modèle acoustique et visuel, les durées utilisées sont celles provenant du corpus original.

	Modèles Séparés							Modèles joints [2048, 2048]						
	Neu	Joi	Tri	Col	Sur	Peu	Dég	Neu	Joi	Tri	Col	Sur	Peu	Dég
	Acoustique [1024, 1024]							Acoustique						
MCD (dB)	**4.863**	**5.738**	**5.288**	**5.262**	**5.699**	**5.226**	**5.431**	5.305	6.135	5.740	5.691	6.157	5.669	5.844
BAP (dB)	**0.224**	**0.312**	**0.269**	**0.268**	**0.287**	**0.231**	**0.256**	0.265	0.359	0.304	0.322	0.335	0.269	0.304
F0-RMSE (Hz)	**26.172**	**46.723**	**36.943**	**39.514**	**32.203**	**40.617**	**37.972**	32.203	47.617	37.972	45.094	45.676	46.201	44.003
F0-CORR	**0.687**	**0.631**	**0.518**	**0.524**	**0.702**	**0.627**	**0.535**	0.683	0.627	0.514	0.513	0.683	0.488	0.518
V/N-V (%)	**6.900**	**10.167**	**7.692**	**8.082**	**9.874**	**7.711**	**9.137**	7.851	11.879	8.955	9.587	11.864	8.814	10.560
	Visuel [1024, 1024]							Visuel						
RMSE (mm)	**1.304**	**1.572**	**1.317**	**1.466**	**1.482**	**1.424**	**2.124**	1.309	1.581	1.320	1.475	1.504	1.429	2.132
Corrélation	**0.833**	**0.777**	**0.792**	**0.810**	**0.807**	**0.826**	**0.696**	0.829	0.776	0.790	0.808	0.803	0.825	0.689

TABLE 6.4 – *Les résultats pour les paramètres acoustiques et visuels sur l'ensemble de test générés en entraînant le DNN avec les modalités acoustiques et visuelles séparément puis conjointement.*

Le tableau 6.4 montre les résultats obtenus avec les deux architectures. Nous remarquons que l'entraînement joint des deux modalités dégrade toutes les mesures objectives, que ça soit pour la modalité acoustique ou visuelle. En effectuant une écoute informelle, nous avons constaté plus de distorsion et un son légèrement étouffé dans les résultats du modèle joint, mais pour les résultats visuels, nous n'avons constaté aucune différence humainement perceptible. Ce résultat rejoint celui de Filntisis et al. [2017] qui a montré via des tests perceptifs sur des données expressives

que les résultats des modèles séparés sont considérés comme légèrement plus réalistes, mais qu'aucune différence d'ordre significatif n'a été trouvée entre les résultats audiovisuels. Le même résultat que pour les données audiovisuelles a été trouvé pour les données visuelles. Cependant, les séquences acoustiques générées par le modèle séparé ont été considéré comme significativement plus réalistes que ceux générés par le modèle joint.

6.7 Validation-croisée des résultats de la synthèse expressive

Dans cette expérience nous souhaitons vérifier, via une étude objective, la capacité des modèles à apprendre des caractéristiques spécifiques à chaque émotion. Pour ce faire, nous entraînons le modèle de durées, le modèle acoustique et le modèle visuel avec l'ensemble des données neutres et expressives étiquetées puis nous procédons à une validation croisée. Dans cette expérience nous utilisons des modèles acoustique et visuel séparés avec la configuration *3_ cont* comme vecteur d'entrée, et *5_ cont_ p_ s_ m* comme entrée du modèle des durées.

		Durées						
		Neutre	Joie	Tristesse	colère	surprise	Peur	Dégoût
Neutre	RMSE (f/p)	**5.289**	6.110	6.001	5.917	5.652	6.378	15.280
	CORR	**0.831**	0.799	0.804	0.786	0.803	0.806	0.779
Joie	RMSE (f/p)	7.346	**7.136**	7.385	7.272(-)	7.206(-)	7.708	15.703
	CORR	0.752	**0.774**	0.756	0.760(-)	0.769(-)	0.751	0.720
Tristesse	RMSE (f/p)	6.886(-)	6.881	**6.606**	7.118	7.176	6.926(-)	14.856
	CORR	0.770(-)	0.765	**0.777**	0.755	0.754	0.770(-)	0.747
colère	RMSE (f/p)	6.879	7.130	7.222	**6.463**	7.597	6.578	16.195
	CORR	0.720	0.737(-)	0.728	**0.758**	0.729(-)	0.744	0.686
surprise	RMSE (f/p)	6.394	6.905	7.134	6.471(-)	**6.006**	7.532	16.582
	CORR	0.756	0.763(-)	0.741	0.753	**0.781**	0.749	0.708
Peur	RMSE (f/p)	7.573	7.468	7.287(-)	7.760	7.789	**7.174**	14.578
	CORR	0.767(-)	0.758	0.766(-)	0.756	0.763	**0.781**	0.753
Dégoût	RMSE (f/p)	19.614	18.709	17.669	19.162	19.361	18.146	**9.311**
	CORR	0.728	0.716	0.723	0.693	0.712	0.721	**0.741**

TABLE 6.5 – *Les résultats du RMSE en frames/phonème et de la CORR pour la validation croisée sur les résultats de prédiction des durées des données expressives de l'ensemble de test.*

		Visuel							
		Neutre	Joie	Tristesse	colère	surprise	Peur	Dégoût	Statique
Neutre	RMSE (mm)	**1.304**	2.392	1.635	2.464	1.945	2.245	2.377	2.170
	CORR	**0.833**	0.770	0.801(-)	0.769	0.782	0.805	0.739	—
Joie	RMSE (mm)	2.500	**1.572**	2.125	2.703	2.605	2.814	2.732	3.217
	CORR	0.727	**0.777**	0.734	0.722	0.736(-)	0.734(-)	0.712	—
Tristesse	RMSE (mm)	1.655	2.092	**1.317**	2.378	2.241	2.221	2.325	2.364
	CORR	0.775(-)	0.753	**0.792**	0.723	0.727	0.773	0.713	—
colère	RMSE (mm)	2.604	2.564	2.439	**1.466**	2.100	1.688	3.124	3.308
	CORR	0.732	0.735	0.714	**0.810**	0.783	0.774	0.716	—
surprise	RMSE (mm)	1.984	2.537	2.271	2.046	**1.482**	1.980	2.614	2.817
	CORR	0.750	0.744	0.723	0.785	**0.807**	0.771	0.727	—
Peur	RMSE (mm)	2.255	2.778	2.239	1.715	1.883	**1.424**	3.041	3.055
	CORR	0.794	0.772	0.791	0.795(-)	0.790	**0.826**	0.748	—
Dégoût	RMSE (mm)	2.823	3.160	2.822	3.414	3.063	3.460	**2.124**	3.530
	CORR	0.651	0.647	0.641	0.644	0.651	0.649	**0.696**	—

TABLE 6.6 – *Les résultats de validation croisée sur les résultats de prédiction des trajectoires visuelles des données expressives de l'ensemble de test. Statique représente un visage à l'état neutre avec une bouche constamment fermée.*

		Acoustique						
		Neutre	Joie	Tristesse	colère	surprise	Peur	Dégoût
Neutre	MCD (dB)	**4.863**	6.409	5.390	5.784	6.548	5.327	5.539
	BAP (dB)	**0.224**	0.304	0.243	0.268	0.263	0.229(-)	0.232
	F0-RMSE (Hz)	**26.172**	97.521	33.810	42.063	108.220	30.640	28.839
	F0-CORR	**0.687**	0.610	0.598	0.546	0.404	0.558	0.604
	V/N-V (%)	**6.900**	7.565	7.594	7.512	7.612	7.154(-)	7.486
Joie	MCD (dB)	7.010	**5.738**	6.696	6.417	6.132	7.227	7.045
	BAP (dB)	0.367	**0.312**	0.347	0.330	0.334	0.377	0.371
	F0-RMSE (Hz)	103.444	**46.723**	85.568	97.438	58.942	113.167	109.806
	F0-CORR	0.586	**0.631**	0.552	0.547	0.455	0.526	0.540
	V/N-V (%)	11.015	**10.167**	10.792	10.751	10.547(-)	10.812	10.908
Tristesse	MCD (dB)	5.688	6.442	**5.288**	5.825	6.642	5.660	5.817
	BAP (dB)	0.271(-)	0.301	**0.269**	0.284	0.284	0.271(-)	0.271
	F0-RMSE (Hz)	32.074	82.658	**36.943**	39.353	96.357	47.107	44.815
	F0-CORR	0.503(-)	0.476	**0.518**	0.496	0.284	0.509(-)	0.514
	V/N-V (%)	8.107(-)	8.246	**7.692**	8.167	8.258	7.984(-)	8.023
colère	MCD (dB)	6.177	6.069	5.793	**5.262**	6.171	6.039	6.040
	BAP (dB)	0.303	0.287	0.290	**0.268**	0.291	0.303	0.304
	F0-RMSE (Hz)	43.043	89.370	41.357(-)	**39.514**	97.935	44.919	43.601
	F0-CORR	0.440	0.454	0.497	**0.524**	0.357	0.505(-)	0.491
	V/N-V (%)	8.705	8.615	8.515	**8.082**	8.807	8.347(-)	8.495(-)
surprise	MCD (dB)	6.806	5.916	6.616	6.277	**5.699**	6.951	6.911
	BAP (dB)	0.305	0.300(-)	0.302(-)	0.301(-)	**0.287**	0.311	0.311
	F0-RMSE (Hz)	102.248	51.066	86.765	97.706	**32.203**	112.539	109.363
	F0-CORR	0.449	0.564	0.394	0.444	**0.702**	0.382	0.391
	V/N-V (%)	10.176	9.769(-)	9.967(-)	10.078(-)	**9.874**	10.007(-)	10.105
Peur	MCD (dB)	5.730	7.015	5.729	6.097	7.075	**5.226**	5.252(-)
	BAP (dB)	0.246(-)	0.334	0.262	0.297	0.286	**0.231**	0.234(-)
	F0-RMSE (Hz)	37.586	116.012	48.980	41.281(-)	128.206	**40.617**	32.505
	F0-CORR	0.435	0.404	0.487	0.477	0.242	**0.627**	0.494
	V/N-V (%)	7.951(-)	8.254	8.307	8.258	8.352	**7.711**	7.649(-)
Dégoût	MCD (dB)	5.995	7.124	5.949	6.250	7.269	5.641(-)	**5.431**
	BAP (dB)	0.272	0.338	0.279	0.328	0.296	0.265(-)	**0.256**
	F0-RMSE (Hz)	38.890(-)	117.422	46.779	42.528	132.273	36.477(-)	**37.972**
	F0-CORR	0.473	0.439	0.512	0.502	0.299	0.516(-)	**0.535**
	V/N-V (%)	9.350(-)	9.840	9.446(-)	9.837	9.828	9.245(-)	**9.137**

TABLE 6.7 – *Les résultats de validation croisée sur les résultats de prédiction des paramètres acoustiques des données expressives de l'ensemble de test.*

Dans cette expérience, nous évaluons la capacité de nos modèles à représenter les durées et les modalités acoustique et visuelle. Toutefois, la prononciation des phrases peut changer d'une émotion à l'autre (plus ou moins de pauses, suppression/ajout de voyelles) comme décrit par Qader et al. [2014]. Ce dernier point n'est pas étudié dans ce travail. Afin d'évaluer les différents modèles, l'idée est d'entraîner ces derniers avec les données étiquetées des différents états expressifs. Au moment de l'évaluation, nous générons les paramètres de sortie de chaque modèle et pour chaque émotion. Par la suite, nous comparons les résultats de prédiction de toutes les émotions avec les données originales de l'ensemble de test d'une émotion cible donnée. Si l'erreur sur les paramètres prédits de l'émotion cible est plus faible et que l'écart avec l'erreur des autres émotions est significatif, nous considérons que le modèle est plus spécialisé dans la prédiction de cette émotion cible que les autres émotions.

Pour le modèle des durées, nous avons utilisé les informations linguistiques de l'ensemble de test d'une émotion cible pour générer les durées de toutes les autres émotions et nous avons calculé les mesures de RMSE et de corrélation par rapport aux données originales de l'émotion cible. Pour

les modèles acoustique et visuel, nous avons utilisé les données linguistiques ainsi que les durées des données originales de l'ensemble de test de l'émotion cible. En utilisant ces informations, nous générons les paramètres acoustiques et visuels correspondants à chaque émotion et calculons les différentes mesures. Les résultats relatifs à chaque émotion traitée sont représentés dans les lignes des tableaux 6.5, 6.6 et 6.7.

Après la collecte des résultats, nous avons calculé avec un test-t le niveau de significativité statistique de l'écart des résultats des différentes émotions avec les résultats de l'émotion cible, les écarts non-significatifs sont présentés avec le signe (-). Les résultats affichés dans ces trois tableaux montrent que les trois modèles arrivent à se spécialiser dans la modélisation des différentes émotions puisque la majorité des écarts sont statistiquement significatifs. Pour le modèle des durées, même si les écarts sont généralement petits, mais une grande partie d'entre eux reste significative. Le dégoût semble être très différent des autres émotions. Cela peut s'expliquer par les durées des émotions dans le corpus utilisé. En fait, cette émotion a été jouée avec un débit remarquablement lent. La durée du corpus du dégoût (1h 53min) représente environ le double des durées des autres émotions (entre 55min et 1h 11min). Pour les résultats visuels et acoustiques, les écarts sont globalement significatifs, mais les petits écarts entre les résultats de certaines émotions laissent supposer l'existence de ressemblances entre quelques émotions. Pour le modèle visuel et dans le cas de l'état neutre, les mesures calculées, notamment le RMSE, montrent que la tristesse se distingue par des valeurs proches de celles du neutre. Nous pouvons constater la même chose entre la colère et la peur avec l'écart de RMSE le plus bas. En ce qui concerne le modèle acoustique, nous remarquons qu'il y a également une ressemblance entre l'état neutre et la tristesse puis entre la peur et le dégoût. De plus la joie et la surprise sont les émotions avec le plus grand écart de F0 par rapport au neutre et aux autres émotions.

Dans une étude perceptive présentée dans le chapitre suivant (chapitre 7), utilisant le même corpus et un type de couches cachées similaire (BLSTM), nous validons que les émotions synthétiques sont correctement reconnues par les participants (sauf la peur et la tristesse).

6.8 Conclusion

Dans ce chapitre, nous avons effectué une étude sur la synthèse audiovisuelle expressive de la parole, afin de donner des éclaircissements sur l'apport de certains paramètres sur les résultats générés. Pour atteindre cet objectif, nous avons adopté différentes architectures neuronales pour entraîner trois modèles : le modèle des durées, le modèle acoustique et le modèle visuel. Nous avons réalisé une comparaison directe entre ces architectures en variant les paramètres linguistiques utilisés. Les résultats obtenus montrent que, même si toutes les informations linguistiques sont bénéfiques pour le modèle des durées, seulement les informations sur les contextes gauche et droit immédiats ainsi que le contexte syllabique améliorent la prédiction pour le modèle acoustique. Toutefois pour le modèle visuel les informations autres que le contexte gauche et droit immédiat ne semblent pas apporter en précision à la qualité de la prédiction. Nous avons également comparé la qualité de la synthèse des modèles acoustique et visuel entraînés séparément puis conjointement. À l'aide d'une comparaison avec des mesures objectives, nous avons trouvé que les modèles entraînés séparément atteignent une meilleure précision de reconstruction.

Une hypothèse sur la baisse de précision des modèles acoustiques et visuels avec un contexte linguistique important ou un vecteur d'entrée de grande taille (audiovisuel), serait l'augmentation du nombre de combinaisons possibles pour les données, ce qui a pour effet de réduire le nombre d'exemples de chacune d'elles pendant l'apprentissage. Nous pensons que c'est pour cette raison que cet effet ne semble pas se produire avec les données des durées, qui ont un vecteur de très

petite taille contenant un seul paramètre. De ce fait, il serait intéressant, pour compléter cette étude, de faire une analyse sur un corpus de taille plus importante pour voir si ces conclusions peuvent être maintenues.

Finalement, les résultats objectifs de la validation-croisée effectuée sur les différentes émotions, montrent que les trois modèles arrivent à se spécialiser aux différentes émotions. Ces résultats nous ont aussi permis de constater des similarités et des différences entre certaines émotions. Une piste pour compléter cette étude serait de pousser plus loin l'évaluation du modèle des durées à un niveau plus fin pour voir si certains phonèmes sont mieux modélisés que d'autres suivant leur contexte linguistique ou émotionnel.

Nous souhaitons que cet ensemble d'expériences apporte plus de clarté sur le comportement des DNNs face aux différentes données linguistiques, des durées et audiovisuelles, et que cela nous facilite par la suite le choix de l'architecture neuronale la plus adaptée pour la synthèse audiovisuelle expressive de la parole.

7

Synthèse audiovisuelle expressive par CVAE

Sommaire

7.1 Introduction

Le système de synthèse avec une architecture DNNs entièrement connectée (DNN-FC) présenté dans le chapitre précédent (chapitre 6), et à l'instar des systèmes similaires dans la littérature [Li et al., 2016a, An et al., 2017, Zhang et al., 2017], permet de modéliser les classes d'émotions présentes dans la base de données d'entraînement. Chaque émotion modélisée est en réalité le résultat d'une moyenne sur l'ensemble des exemples vus pendant le processus d'entraînement. De ce fait, l'information sur les degrés des émotions, qui peuvent être présents dans l'ensemble d'apprentissage, vont disparaître et le contrôle de la variabilité de l'expressivité humaine va être perdu. Comme présenté dans le chapitre 3 de l'étude de l'état de l'art, et contrairement à la théorie catégorique des émotions [Ekman, 1992], les approches dimensionnelles et continues des

émotions [Russell, 1980, Plutchik, 1984, Laresn and Diener, 1992] considèrent le système émotionnel humain comme un spectre reliant les différents états émotionnels de manière continue [Posner et al., 2005, Russell and Fehr, 1994].

Afin d'éviter la représentation moyennée des émotions, dans le travail de Hofer et al. [2005], un système de synthèse vocale par sélection d'unités a été proposé pour générer des nuances d'émotions en utilisant une base de données annotée avec des degrés d'émotion. Dans une autre étude, le système de conversion de voix basé sur des règles présentées par Xue et al. [2018b] propose un espace bidimensionnel (valence et arousal) pour contrôler le degré des émotions. Le système de conversion de voix se base sur les règles et est réalisé par paramétrisation et remplacement des caractéristiques relatives aux émotions.

Toutefois, ces approches reposent entièrement sur les étiquettes des émotions dans le processus de l'apprentissage alors que la plupart des ressources expressives disponibles ne sont pas annotées. De surcroît, lorsque les étiquettes sont disponibles, elles ne sont pas totalement fiables à cause des éventuelles erreurs dues aux annotateurs ou à la mauvaise performance des acteurs. Henter et al. [2017] ont réussi à créer des degrés d'émotion sans des étiquettes relatives aux intensités des émotions en ayant toutefois utilisé des étiquettes de catégories d'émotions ce qui ne fait que déplacer le problème.

Dans ce chapitre [12], nous abordons ce problème de la synthèse expressive sans utilisation des étiquettes d'émotions durant le processus de l'apprentissage. Plus spécifiquement, nous abordons l'utilisation du *Variational Auto-Encoder* (VAE), approche introduite par Kingma and Welling [2013]. Nous explorons en particulier l'extension β-*Conditional Variational Auto-Encoder* (β-CVAE) et adapterons cette approche pour la synthèse expressive audiovisuelle de la parole. Nous montrerons également les possibilités offertes par cette approche rendant l'interpolation entre les émotions possible.

7.2 Présentation de l'architecture β-CVAE

Dans cette section, nous introduisons d'abord l'architecture encodeur-décodeur puis les principes de l'auto-encodeur variationnel *(Variational Auto-Encoder)* que nous appelons VAE et de l'auto-encodeur variationnel sous condition *(Conditional Variational Auto-Encoder)* que nous appelons CVAE. Ils représentent la base du travail présenté dans ce chapitre. Nous présentons par la suite l'architecture que nous avons adoptée pour les trois modèles de synthèse des durées, acoustique et visuel.

7.2.1 Architecture encodeur-décodeur

L'architecture d'un système encodeur-décodeur standard [Bengio et al., 2013] consiste en un encodeur et un décodeur comme présenté dans la figure 7.1. Le rôle de l'encodeur est d'extraire une représentation intermédiaire du vecteur d'entrée. Cette représentation est ensuite passée au décodeur pour générer le vecteur de sortie. Lorsque le but du système est d'encoder une information pour la reconstruire à la sortie, nous parlons alors d'un auto-encodeur. Ce système apprend une représentation intermédiaire z d'un vecteur d'entrée x en minimisant l'écart entre la sortie générée par le système $\hat{x}$ et l'entrée x. La fonction de perte d'un auto-encodeur *Loss* est

12. Les travaux présentés dans ce chapitre ont fait l'objet d'une publication dans la conférence internationale Interspeech 2019 (https ://hal.inria.fr/hal-02175776/document) et d'une soumission d'un article de revue à une session spéciale de la revue internationale Neural Networks (Special issue on Advances in Deep Learning Based Speech Processing).

définie uniquement par l'erreur de reconstruction *RE* entre les paramètres d'entrée et de sortie. Elle est définie dans ce travail par le carré moyen des erreurs :

$$Loss = RE = \frac{1}{N}\sum_{i=1}^{N}(\hat{x}_i - x_i)^2 \tag{7.1}$$

Où N est la taille de l'ensemble d'apprentissage.

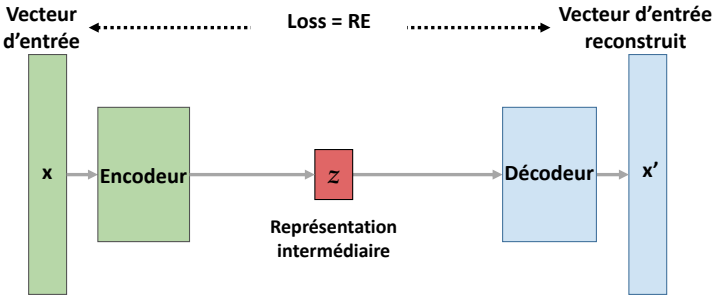


FIGURE 7.1 – *Architecture d'un encodeur-décodeur, la fonction d'erreur correspond à l'erreur de reconstruction RE entre les paramètres d'entrée et de sortie.*

7.2.2 Architecture VAE

L'architecture du VAE [Kingma and Welling, 2013] est similaire à celle d'un auto-encodeur standard, toutefois la représentation intermédiaire des données est construite différemment. En fait, pour un auto-encodeur standard, la représentation intermédiaire du vecteur d'entrée est apprise directement en encodant ce dernier en un vecteur latent et en le décodant par la suite. Aucune condition n'est imposée pendant le processus d'apprentissage pour structurer l'espace latent de manière intelligente et compatible avec un processus génératif. De ce fait, certains points de l'espace latent donneront un contenu incohérent et insensé une fois décodés. Pour un auto-encodeur variationnel, au lieu d'encoder une entrée comme un seul point latent, il l'encode plutôt comme une distribution sur l'espace latent. Le VAE apprend les paramètres d'une distribution de probabilité représentant les données (μ et σ). Nous pouvons donc utiliser ces paramètres pour remodeler l'espace latent et échantillonner par la suite à partir de cette distribution latente ($z \sim \mathcal{N}(\mu, \sigma^2)$) pour générer de nouvelles données cohérentes sans avoir au préalable des données originales correspondant aux vecteurs latents choisis. Il s'agit donc d'un modèle génératif. Comme présenté dans la figure 7.2, la fonction de perte d'un VAE contient un nouveau terme. En plus de la condition de réduction de l'erreur de reconstruction *RE*, le VAE introduit un terme supplémentaire de régularisation *KL* qui force la représentation latente z à suivre une distribution gaussienne normale. La fonction de perte d'un VAE est définie comme suit :

$$Loss = RE + KL \tag{7.2}$$

Le premier terme *RE* est l'erreur de reconstruction entre x et $\hat{x}$, elle encourage le décodeur à apprendre à reconstruire les données originales à partir de leur représentation latente. Le second terme *KL* représente la divergence de Kullback-Leibler [Kullback and Leibler, 1951] entre la distribution de l'espace latent et une distribution gaussienne normale :

$$KL = D_{KL}(\mathcal{N}(\mu, \sigma^2)||\mathcal{N}(0, I)) \tag{7.3}$$

Ce terme agit comme un régularisateur qui force la représentation latente à suivre une distribution normale, ce qui a pour effet de pousser les différents clusters de l'espace latent à être de plus en plus proches en se regroupant autour de la moyenne zéro. Nous définissons un cluster latent dans ce travail, par un regroupement de vecteurs latents appartenant à une émotion données. Ce comportement encourage une couverture maximale de l'espace latent en le lissant et supprimant les éventuelles zones mortes.

Dans cette approche et contrairement au chapitre précédent (chapitre 6), l'architecture n'est plus entièrement connectée, mais est scindée en deux réseaux neuronaux entraînés conjointement :

1. Réseau d'encodage (encodeur) : Réseau neuronal qui relie l'entrée x à une représentation latente z pour approximer la distribution postérieure insoluble des données d'entrée.
2. Réseau génératif de prédiction (décodeur) : Réseau neuronal qui reconstruit la variable d'entrée x à partir de sa représentation latente z.

Toutefois, une problématique s'impose et rend la rétropropagation de l'erreur à travers le VAE impossible. En fait, comme présenté plus haut, z est échantillonné de manière aléatoire à partir de la distribution latente ($z \sim \mathcal{N}(\mu, \sigma^2)$). En raison de la nature probabiliste de z aucune information ne peut être rétropropagée à travers ce noeud. Une astuce simple, appelée astuce de reparamétrisation proposée par Kingma and Welling [2013], est utilisée pour rendre possible la descente de gradient malgré l'échantillonnage aléatoire et consiste à déplacer la nature probabiliste vers un autre noeud en échantillonnant z comme suit :

$$z = \mu + \sigma^2 \odot \varepsilon \qquad avec \qquad \varepsilon \sim \mathcal{N}(0,\, I) \tag{7.4}$$

De cette manière z devient un noeud déterministe ce qui permet à l'erreur d'être rétropropagée dans le réseau.

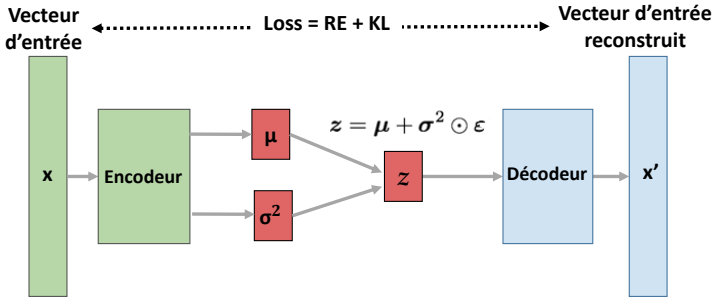


FIGURE 7.2 – *Architecture d'un VAE, la fonction d'erreur est augmenté d'un nouveau terme de régularisation KL et les vecteurs latents sont échantillonnés d'une distribution en utilisant une astuce de reparamétrisation [Kingma and Welling, 2013].*

7.2.3 VAE conditionnel (CVAE)

Le VAE conditionnel (CVAE) est un réseau de type VAE qui est conditionnée sur un paramètre supplémentaire $\boldsymbol{c}$. Dans ce travail la condition $\boldsymbol{c}$ représente les paramètres linguistiques correspondantes à l'entrée x. Un schéma représentatif de l'architecture du CVAE est présenté dans la figure 7.3.

Atanov et al. [2019] ont montré qu'en conditionnant le réseau sur une variable $\boldsymbol{c}$, la représentation latente des données devient indépendante de cette variable. Dans le travail de Skerry-Ryan et al. [2018] et dans le but de transférer la prosodie d'une phrase à une autre, les auteurs ont réussi à isoler la prosodie des autres composantes de la parole. Pour cela, ils ont conditionné leur réseau sur le contenu linguistique, l'identité du locuteur et les effets du canal (c'est-à-dire l'environnement d'enregistrement).

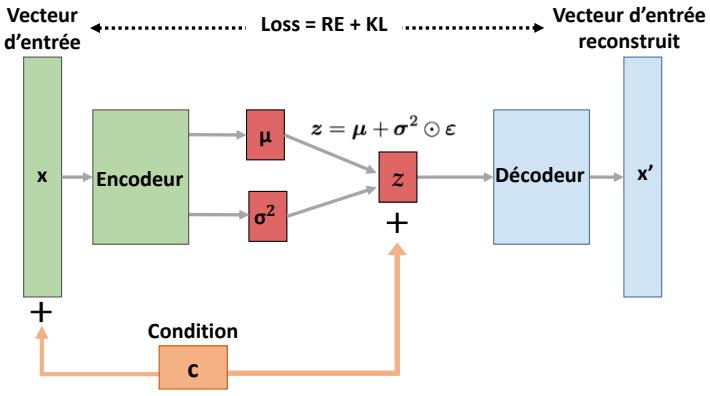


FIGURE 7.3 – *Architecture d'un CVAE, la fonction d'erreur est augmentée d'un nouveau terme de régularisation KL et les vecteurs latents sont échantillonnés d'une distribution en utilisant une astuce de paramétrisation [Kingma and Welling, 2013]. L'encodeur et le décodeur sont conditionnés par un vecteur* $\boldsymbol{c}$.

Ainsi, en conditionnant notre réseau sur $\boldsymbol{c}$, qui contient les paramètres linguistiques, nous forçons sa représentation latente à être indépendante du contenu linguistique des phrases. Le réseau doit apprendre à représenter les paramètres qui ne sont pas contenus dans les données textuelles. Puisque le corpus que nous utilisons (voir chapitre 5) compte une seule locutrice avec des conditions d'enregistrement identiques, nous capturons dans l'espace latent le reste de la variation, et donc principalement les informations relatives à l'état émotionnel de chaque exemple d'entraînement.

7.2.4 β-CVAE

Pendant l'implémentation d'un VAE, l'équilibre entre les deux termes de la fonction de coût n'est pas évident à trouver, ce qui rend l'entraînement d'un VAE particulièrement complexe. Généralement, l'un des deux termes de la fonction de coût, *RE* ou *KL*, prend le pas sur l'autre [Bowman et al., 2016, Higgins et al., 2017]. Lorsque le terme *KL* est très grand, le réseau se focalise uniquement sur lui en ignorant *RE*, empêchant ainsi le réseau d'apprendre à reconstruire les données. Dans le cas inverse, le terme *KL* est réduit à zéro et l'espace latent ne capture aucune information utile, le vecteur z sera simplement ignoré par le décodeur.

Pour régler cette problématique, Higgins et al. [2017] ont introduit un hyper-paramètre ajustable β dans la fonction de perte pour équilibrer entre le coût de la reconstruction et le coût de la régularisation. La nouvelle fonction de coût est présentée dans l'équation 7.5. Ce terme a été ensuite utilisé dans plusieurs travaux [Roche et al., 2019, Bowman et al., 2016, Yang et al., 2019, Wang et al., Alemi et al., 2017, Deng, 2012]. Une valeur élevée de β favorise la régularisation au détriment de la précision de la reconstruction. Nous expliquons la procédure de sélection du

paramètre β pour les différents aspects de la parole dans la section 7.3.2 de ce chapitre.

$$Loss = RE + \boldsymbol{\beta}\ KL \tag{7.5}$$

En utilisant cette variante nous pouvons contrôler la force du terme de régularisation KL pour remodeler l'espace latent sans que cela ne se fasse au détriment d'une bonne reconstruction des données. De cette manière nous seront capables de créer une continuité dans l'espace latent qui nous permettra de passer d'une émotion à l'autre en obtenant constamment des données de sorties cohérentes. Nous expliquons ce point plus en détails dans la section 7.3.2

7.2.5 Entraînement non-supervisé

Le rôle de l'encodeur est d'extraire une représentation latente compressée des données fournies à son entrée. En fait, l'encodeur effectue une tâche de réduction de dimensionnalité similaire à une ACP. Toutefois, dans le cas de l'encodeur basé sur les DNNs, cette tâche est effectuée de manière non-linéaire. Comme nous l'avons vu ci-dessus, l'encodeur est capable d'encoder les informations contenues dans les données d'entrée tout en ignorant les variations contenues dans la condition $\boldsymbol{c}$ utilisée.

L'architecture de l'encodeur et du décodeur sera présentée dans la section 7.3.1. La figure 7.4 montre l'évolution de la valeur de l'erreur d'entraînement du β-CVAE avec des données visuelles. Nous montrons le résultat d'une configuration où les étiquettes des émotions sont fournies comme entrée à l'encodeur, et dans le second cas sans fournir les étiquettes des émotions à ce dernier.

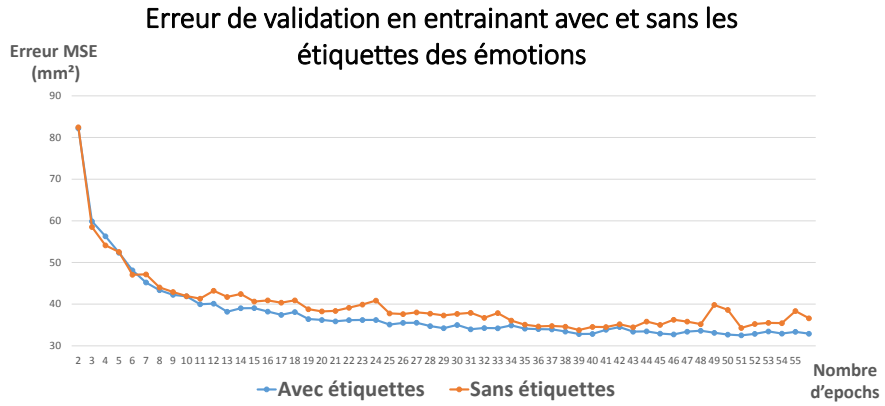


FIGURE 7.4 – *Impact du retrait des étiquettes des émotions sur le processus d'apprentissage du modèle visuel.*

Nous pouvons voir que le réseau arrive sensiblement à la même valeur après un certain nombre d'itérations (37 itérations), bien qu'un entraînement avec des étiquettes d'émotions rende l'apprentissage plus stable. Nous avons constaté un comportement similaire pour les données acoustiques et de durées également. Ce point est très intéressant car il nous permet de nous affranchir de l'information explicite (étiquette) sur les émotions. Nous pouvons donc adopter cette approche d'apprentissage non-supervisé des émotions pour pallier le problème lié aux annotations. Dans ce chapitre, nous allons présenter en détails les configurations et les résultats obtenus avec cette approche.

7.3 Architecture proposée

Nous avons utilisé une architecture β-CVAE pour prédire : 1) les durées 2) les paramètres acoustiques et 3) les paramètres visuels. Nous présentons l'architecture de chacun de ces trois modèles dans la figure 7.5 ci-dessous.

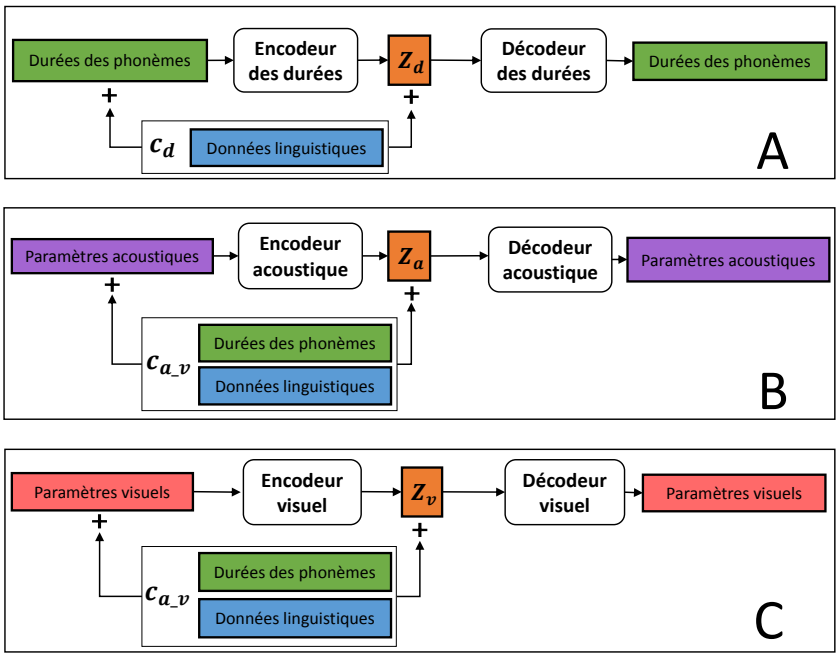


FIGURE 7.5 – *L'architecture encodeur-décodeur des trois modèles. A : Le modèle des durées conditionné sur les données linguistiques c_d uniquement. B et C : L'architecture encodeur-décodeur des modèles acoustique et visuel respectivement qui sont conditionnés sur les données linguistiques et sur les durées c_{a_v}.*

7.3.1 Configuration

Nous avons utilisé Merlin TTS comme outil basique de synthèse acoustique. Nous l'avons augmenté avec une architecture β-CVAE et un module visuel. Dans ce travail nous utilisons un β-CVAE asymétrique puisque le décodeur ne va pas uniquement décompresser la sortie de l'encodeur (le vecteur latent), mais il va effectuer une tâche plus complexe de prédiction non-linéaire. De ce fait, nous utilisons un réseau plus profond pour la partie du décodeur. L'optimiseur Adam a été utilisé pour entraîner les trois modèles, sans *dropout* ou paramètre de régularisation spécifique. Différentes architectures et valeurs de β ont été utilisées pour chaque modèle. Le choix du paramètre β pour chaque modèle est expliqué dans la section suivante (section 7.3.2). Les encodeurs et décodeurs ont été entraînés conjointement. Pour les trois modèles, nous avons utilisé une couche de 50 noeuds avec une fonction d'activation linéaire comme variable latente. Le choix de la taille de la variable latente (50) a été effectué après analyse des résultats obtenus avec différentes tailles : 25, 50 et 100 pour la modalité visuelle. Ayant obtenu l'erreur de reconstruction la plus faible avec une taille de 50, cette valeur a été retenue. Nous avons gardé cette même valeur pour le modèle des durées et acoustique, mais comme perspective, il conviendrait de chercher les valeurs optimales pour les durées et l'aspect acoustique également. En ce qui concerne les

paramètres linguistiques, pour les modèles acoustique et visuel, nous utilisons la configuration *3_ cont*, et pour le modèle des durées nous utilisons *5_ cont_ p_ s_ m*. Ces configurations ont été présentées dans le chapitre précédent (chapitre 6).

Acoustique

Nous concaténons les paramètres acoustiques avec les paramètres linguistiques pour former le vecteur d'entrée de l'encodeur. L'encodeur est un réseau d'une seule couche BLSTM de 1024 noeuds. Le décodeur est composé de deux couches BLSTM de 1500 noeuds suivies d'une couche de sortie linéaire. Il reçoit le vecteur latent et les paramètres linguistiques comme entrées. Au final, nous avons gardé un taux d'apprentissage de 10^{-4} et un paramètre $\beta = 5 \times 10^{-3}$.

Visuel

Ce module apprend à prédire la position des capteurs 3D (x,y,z). Nous donnons un vecteur contenant 132 paramètres (44 capteurs sur 3 axes chacun) concaténés avec les paramètres linguistiques comme entrée à l'encodeur. L'encodeur est un réseau d'une seule couche BLSTM de 1024 noeuds. Le décodeur est composé de deux couches BLSTM de 1024 noeuds suivies d'une couche de sortie linéaire, il reçoit le vecteur latent et les paramètres linguistiques comme entrées. Un taux d'apprentissage de 5×10^{-5} et un paramètre $\beta = 0.1$ ont été gardés.

Durées

Ce module apprend à prédire les durées des phonèmes. Nous concaténons le paramètre représentant la durée du phonème (en nombre de frames qu'il couvre) avec les paramètres linguistiques puis nous fournissons le vecteur résultant comme entrée de l'encodeur des durées. L'encodeur est composé d'une seule couche de type BLSTM avec 1024 noeuds. Le décodeur est composé aussi d'une seule couche linéaire de 256 noeuds avec tanh comme fonction d'activation suivie d'une couche linéaire comme couche de sortie. Il reçoit le vecteur latent et les paramètres linguistiques comme entrées. Le taux d'apprentissage sélectionné est de 5×10^{-4} avec $\beta = 2 \times 10^{-5}$.

7.3.2 Choix du paramètre β

Le choix du paramètre β est crucial pour obtenir un espace latent bien structuré. Le rôle de ce paramètre est de contrôler le poids du second terme de la fonction de coût du β-CVAE qui correspond au facteur de régularisation KL. Ce terme pousse la distribution de l'espace latent à se rapprocher d'une distribution gaussienne normale. En d'autres termes, il pousse les points de l'espace latent à se regrouper autour d'une moyenne de zéro tout en rapprochant la valeur de leur variance de un. Ceci a pour effet de rapprocher les points de l'espace latent, et donc diminuer les distances inter et intra clusters des émotions.

Un paramètre β très petit, ne permet pas au clusters des émotions d'être suffisamment proches pour pouvoir générer de nouveaux vecteurs latents par interpolation des vecteurs latents existants issus de clusters d'émotions différents. Toutefois, une valeur très grande de β résultera en un chevauchement très important des clusters, et empêchera le réseau d'apprendre à reconstruire les données des différentes émotions correctement. La figure 7.6 montre l'impact de l'augmentation de la valeur de β sur la qualité de la reconstruction des données visuelles.

Nous notons que cette expérience a été effectuée avec une base de donnée de taille réduite puisque l'entraînement du β-CVAE est particulièrement chronophage. Lorsque β est nul, le réseau est totalement focalisé sur la bonne reconstruction des données, mais en augmentant la valeur de

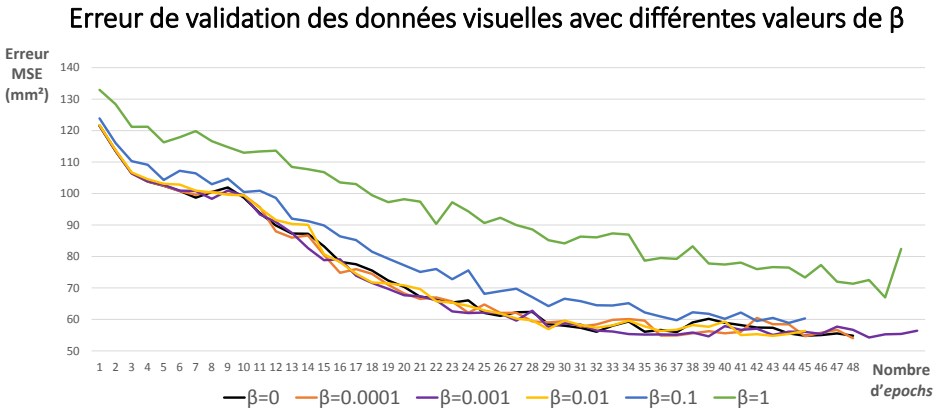


FIGURE 7.6 – *Impact de l'introduction graduelle du poids donné au terme de régularisation (voir équation 7.5) sur la qualité de la reconstruction des données visuelles de l'ensemble de validation.*

β, la reconstruction devient de plus en plus difficile, puisque les clusters se chevauchent de plus en plus jusqu'à se mélanger complètement (voir figure 7.7).

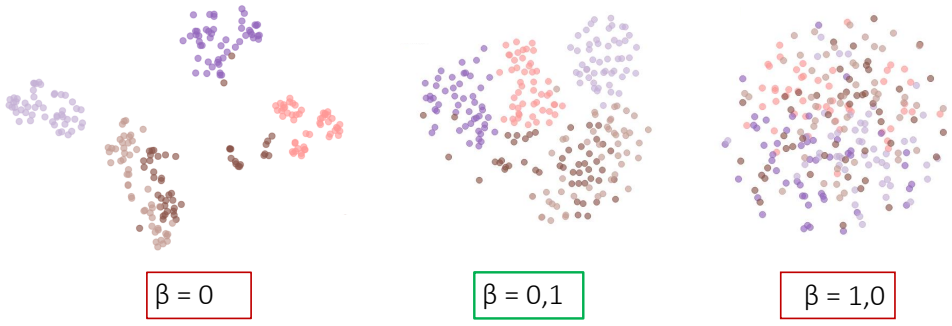


FIGURE 7.7 – *Impact de l'introduction graduelle du poids donné au terme de régularisation (voir équation 7.5), en augmentant la valeur de β, sur la structure de l'espace latent des données visuelles. Les clusters deviennent de plus en plus proches jusqu'à se mélanger complètement.*

Dans les travaux utilisant le β-CVAE, le choix de la valeur du paramètre β est fait selon l'objectif désiré. Dans les travaux de Higgins et al. [2017] et Yang et al. [2019] pour le démêlage des dimensions de l'espace latent, ce paramètre a été fixé par inspection visuelle et par une métrique qui permet de calculer le score quantitatif du démêlage des différentes dimensions de l'espace latent. Wang et al. ont utilisé un β-VAE pour obtenir des représentations latentes sémantiquement significatives, cependant le choix du paramètre β n'a pas été expliqué. Alemi et al. [2017] ont choisi le paramètre β en se basant sur les scores de classifications sur la base de données considérée (MNIST [Deng, 2012]).

Dans le travail de Roche et al. [2019] pour la synthèse de la musique, quatre valeurs du paramètre β ont été testées, et la valeur la plus petite a été sélectionnée. Néanmoins, ce choix n'a pas été validé par une métrique quantitative. Dans cette section, nous présentons notre procédure pour définir une valeur de β optimale pour atteindre notre objectif qui est d'obtenir des clusters suffisamment proches, voir en léger chevauchement. En réalité, les différentes techniques de réduction de dimensionnalité (ACP, t-SNE et U-map) permettent de visualiser l'espace latent

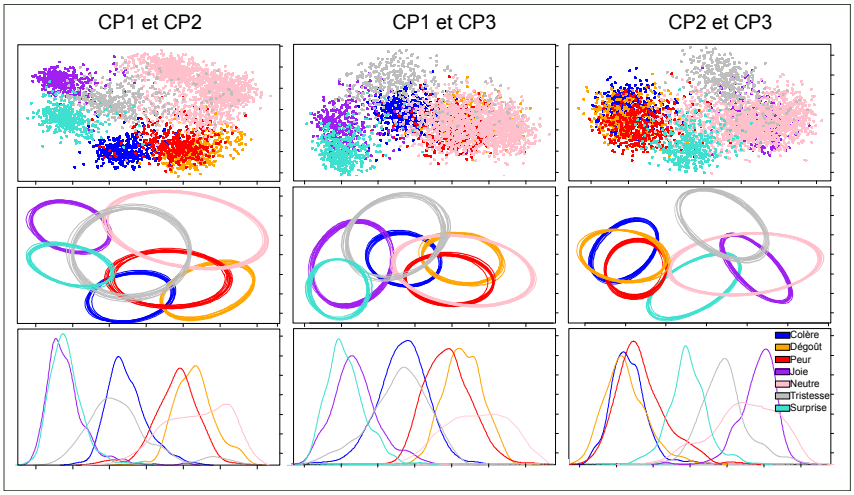


FIGURE 7.8 – *Dix projections elliptiques des régions de chaque cluster d'émotion pour chacune des combinaisons de paires de composantes principales des données acoustiques. Sont présentées aussi les projections unidimensionnelles des densités et les nuages de points en 2D.*

en 2D ou 3D.

La figure 7.8 montre les résultats de projection obtenus pour différentes paires de composantes principales. En utilisant une ACP, il est normal d'obtenir une projection différente selon les axes principaux sélectionnés. Par exemple, la projection dans l'espace formé par la première et la deuxième composantes principales est différente de celle obtenue dans l'espace formé par la première et la troisième composantes principales. Il est à noter que pour la modalité acoustique par exemple, les trois premières composantes principales n'expriment que 22.9% de la variation totale. Aussi, il n'est pas envisageable d'effectuer une projection pour chaque paire des cinquante dimensions de l'espace latent. Il est donc important d'avoir une projection qui prend en compte l'ensemble des dimensions de l'espace latent.

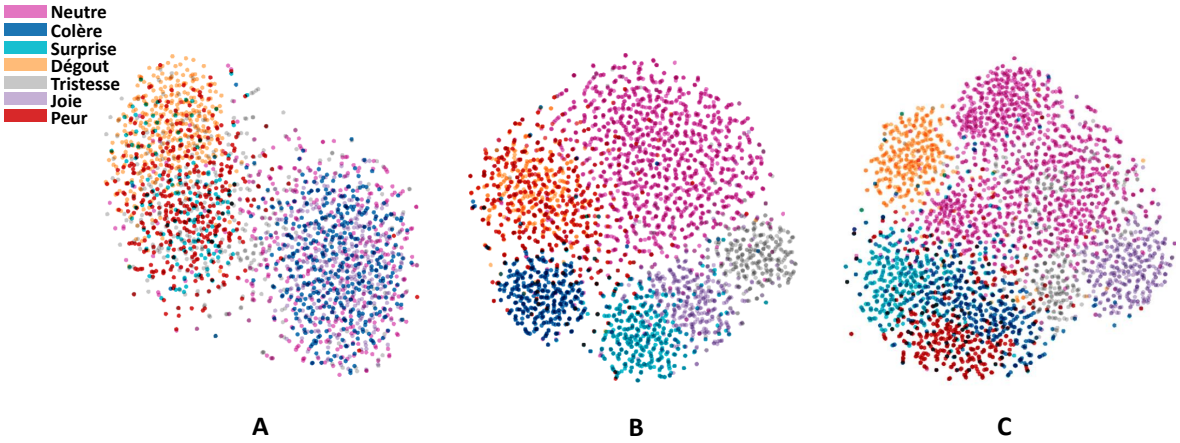


FIGURE 7.9 – *Carte t-SNE [Maaten and Hinton, 2008] des sept clusters de l'espace latent formés par la distribution des données d'entraînement des six émotions basiques et l'état neutre. Le terme de régularisation pousse les échantillons à se rassembler autour de zéro. Les échantillons ont été regroupés différemment selon la modalité étudié (A : Durées B : acoustique et C : visuelle).*

Comme pour une ACP, les algorithmes t-SNE *(t-distributed stochastic neighbor embedding)* [Maaten and Hinton, 2008] et U-map [McInnes et al., 2018] sont des techniques de réduction de dimension pour la visualisation de données à grande dimensions dans un espace de deux ou trois dimensions. Toutefois, ces deux techniques sont non-linéaires et permettent de conserver le voisinage entre les points et la structure globale des données originales dans l'espace à faible dimension, c'est à dire que deux points proches (resp. éloignés) dans l'espace d'origine devront être proches (resp. éloignés) dans l'espace de faible dimension.

Toutefois, ces techniques donnent toutes des résultats de projection légèrement différents (différence entre t-SNE et l'ACP en 2D pour la modalité acoustique dans les figure 7.9-B et 7.8). De plus, pour une même technique, le choix des paramètres de visualisation (nombres de voisins de chaque vecteur latent, le taux d'apprentissage, le nombre d'itérations,..) joue un rôle très important dans le résultat de la projection. De surcroît, le degré de chevauchement entre les clusters change aussi en fonction des dimensions latentes sélectionnées pour la visualisation.

Ces projections ne peuvent donc pas servir de moyen fiable pour évaluer l'état réel de l'espace latent. Ainsi, il est impératif d'utiliser une technique de quantification numérique du taux de chevauchement qui prend en considération la totalité des dimensions de l'espace latent et d'utiliser les projections seulement comme aide/accompagnement visuel des résultats numériques qui permettra de les interpréter en regardant l'emplacement des clusters par rapport aux autres. Pour ce faire, nous proposons d'utiliser la méthode probabiliste de Swanson et al. [2015] initialement proposée pour déterminer la région des niches des espèces animales et le chevauchement entre elles et qui peut être étendue au-delà de deux dimensions. Cette méthode est toujours activement utilisée dans le domaine de l'écologie pour calculer le chevauchement entre les habitats des espèces, pour évaluer le partage des ressources et la concurrence entre les espèces coexistantes [Chavarie et al., 2019, Jackson et al., 2016, McNicholl et al., 2018]. Cette méthode fournit des estimations directionnelles du chevauchement entre les niches et produit des projections uniques des données multivariées.

Swanson et al. [2015] définissent la région de niche comme une région de probabilité dans un espace multivarié. Le chevauchement est calculé comme la probabilité (0% à 100%) qu'un individu de l'espèce A de se trouver dans la région de niche de l'espèce B. Dans l'article original, cette méthode a été appliquée à des données isotopiques en trois dimensions. Dans ce travail de thèse nous allons l'appliquer à des vecteurs latents des modalités acoustique et visuelle et des durées dans un espace à cinquante dimensions. Nous commençons par un paramètre β nul, puis nous l'augmentons graduellement jusqu'à l'obtention d'un début de chevauchement entre plusieurs clusters. Les résultats obtenus avec cette métrique pour les différents modèles sont présentés dans les figures 7.10, 7.11, 7.12 et 7.13, dont les valeurs numériques sont dans l'Annexe C.

Les matrices présentées dans les figures 7.10, 7.11, 7.12 et 7.13 montrent la distribution de la métrique de probabilité de chevauchement (0% à 100%) des différents modèles et représentent la probabilité que les clusters des émotions présentées dans les colonnes de déborder sur des zones spécifiques aux émotions présentées dans les lignes. Les moyennes et les intervalles de confiance de 95% des distributions sont présentés en lignes continues et discontinues respectivement. Pour les données visuelles, nous voyons sur la figure 7.10 qu'avec $\beta = 0.05$, les clusters latents ne se chevauchent pas, ce qui indique la possible présence de discontinuités dans l'espace latent. Ces discontinuités, ou zones mortes, compromettent la possibilité d'interpolation entre les différents clusters.

Pour résoudre ce problème, nous augmentons la valeur de β pour pousser les clusters à se rapprocher. Nous pouvons voir, sur la même figure 7.10, qu'avec $\beta = 0.1$ les clusters sont suffisamment proches pour commencer à légèrement se chevaucher. Nous pouvons également

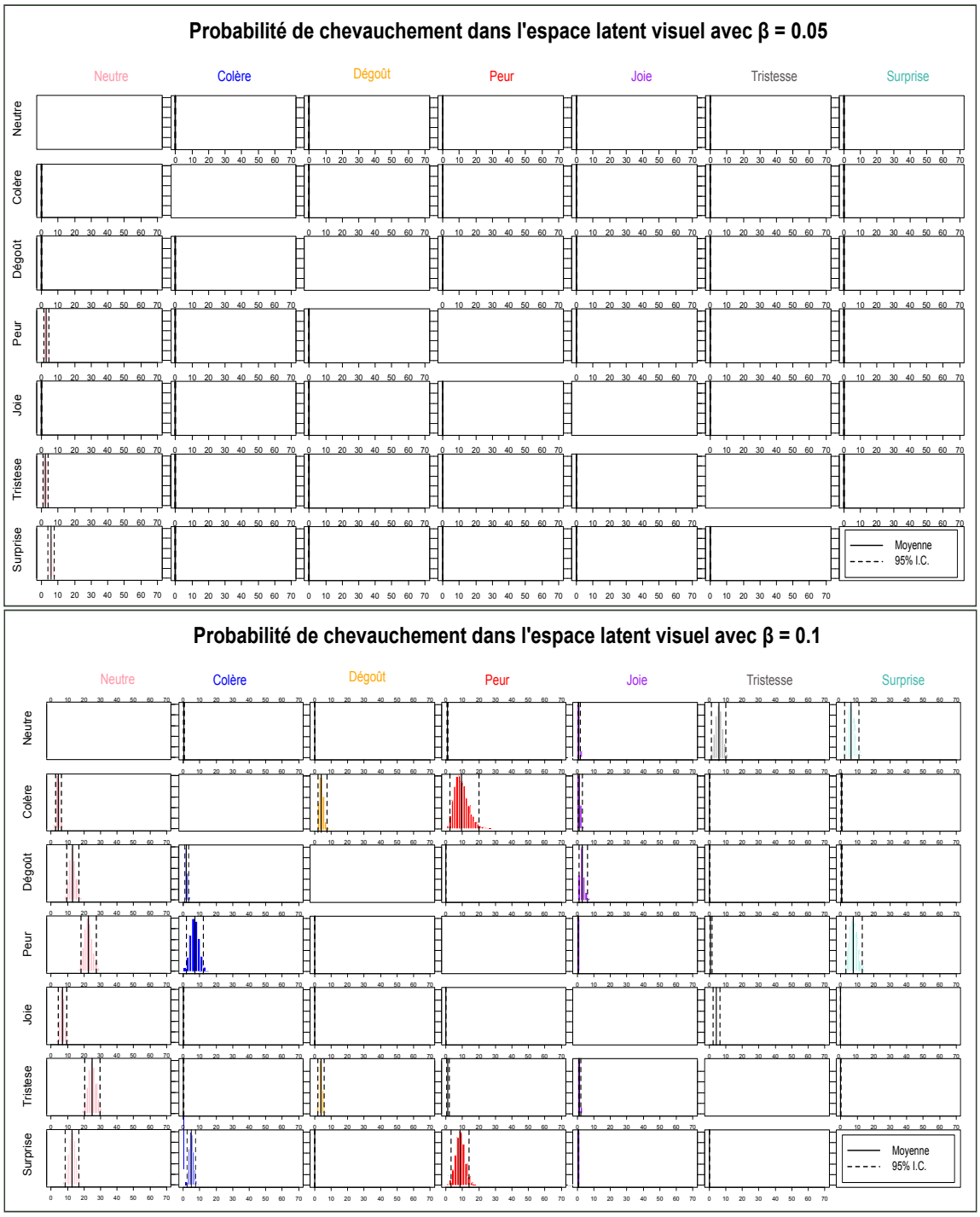


FIGURE 7.10 – *Distribution de la métrique de chevauchement probabiliste entre les clusters d'émotions pour la modalité* ***visuelle*** *(probabilité que les clusters des émotions présentées dans les colonnes débordent sur ceux des émotions présentées dans les lignes). Les moyennes et les intervalles de confiance de 95% des distributions sont présentés en lignes continues et discontinues respectivement. Les matrices sont présentées pour deux valeurs de β, 0.05 et 0.1.*

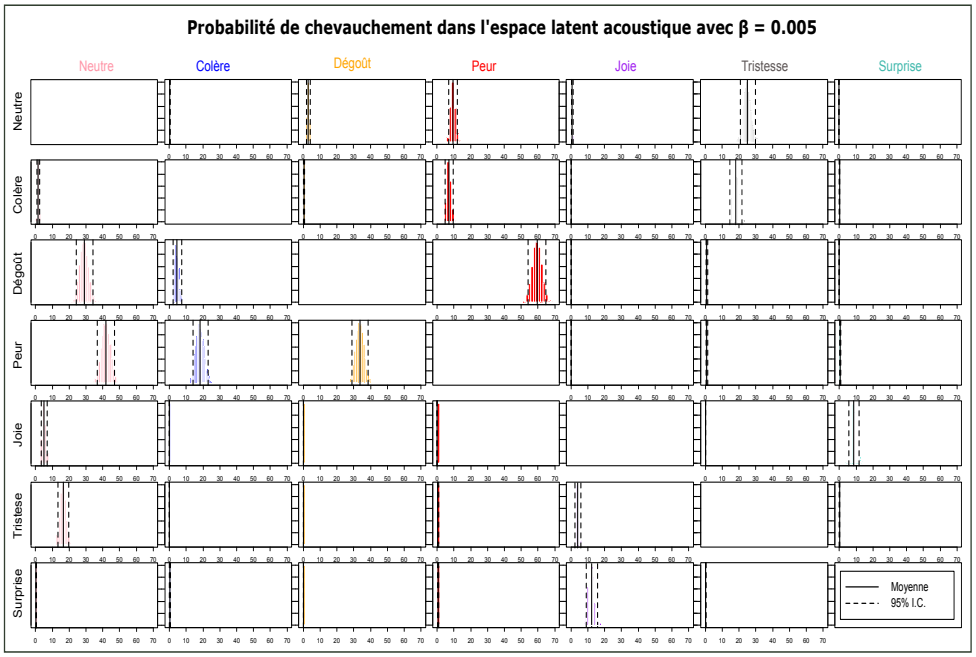


FIGURE 7.11 – *Distribution de la métrique de chevauchement probabiliste entre les clusters d'émotions pour la modalité* ***acoustique*** *avec* $\beta = 5 \times 10^{-3}$ *pour une intervalle de confiance de 95% (probabilité que les clusters des émotions présentées dans les colonnes débordent sur ceux des émotions présentées dans les lignes). Les moyennes et les intervalles de confiance de 95% des distributions sont présentés en lignes continues et discontinues respectivement.*

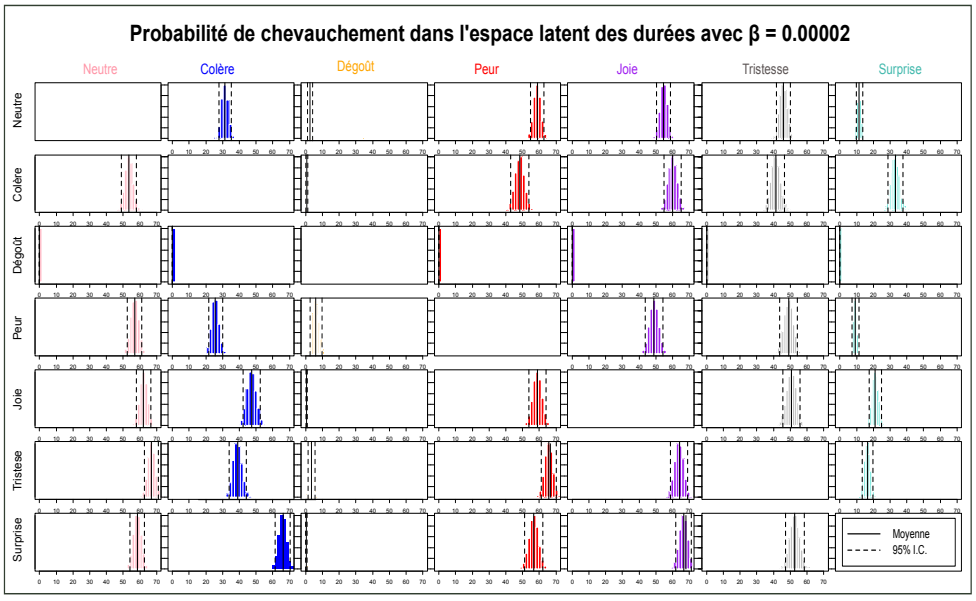


FIGURE 7.12 – *Distribution de la métrique de chevauchement probabiliste entre les clusters d'émotions pour le modèle* ***des durées*** *avec* $\beta = 2 \times 10^{-5}$ *pour une intervalle de confiance de 95% (probabilité que les clusters des émotions présentées dans les colonnes débordent sur ceux des émotions présentées dans les lignes). Les moyennes et les intervalles de confiance de 95% des distributions sont présentés en lignes continues et discontinues respectivement.*

voir qu'aucun cluster d'émotions ou sous-groupe de clusters n'est isolé, chaque cluster d'émotion se chevauche avec au moins trois autres clusters d'émotions. Les différents clusters d'émotions ont tendance à se rassembler autour du cluster de l'état neutre.

En ce qui concerne les données acoustiques (figure 7.11), une valeur plus petite ($\beta = 5 \times 10^{-3}$) était suffisante pour obtenir un bon chevauchement. Nous pouvons constater que le cluster latent de la surprise est uniquement lié au cluster latent de la joie. Toutefois l'augmentation de la valeur de β va empirer le degré de chevauchement entre le dégoût et la peur qui est déjà très élevé.

Lors de l'analyse de l'espace latent des durées (figure 7.12), le chevauchement entre les clusters était déjà très élevé avec $\beta = 0$, mais l'émotion de dégoût était isolée de tous les autres clusters d'émotions, ainsi, nous avons introduit une petite valeur de β (10^{-5} et $\beta = 2 \times 10^{-5}$) jusqu'à ce que le cluster du dégoût commence à se chevaucher avec les autres avec $\beta = 2 \times 10^{-5}$.

Impact d'un entraînement audiovisuel joint sur l'espace latent

Dans le chapitre précédent, et en utilisant une architecture DNN-FC, nous avons trouvé qu'un entraînement joint des deux modalités acoustique et visuelle détériore les performances du modèle par rapport à un entraînement séparé de ces deux modalités. Maintenant, nous allons effectuer un entraînement joint des deux modalités acoustique et visuelle mais, cette fois-ci, en utilisant notre architecture CVAE.

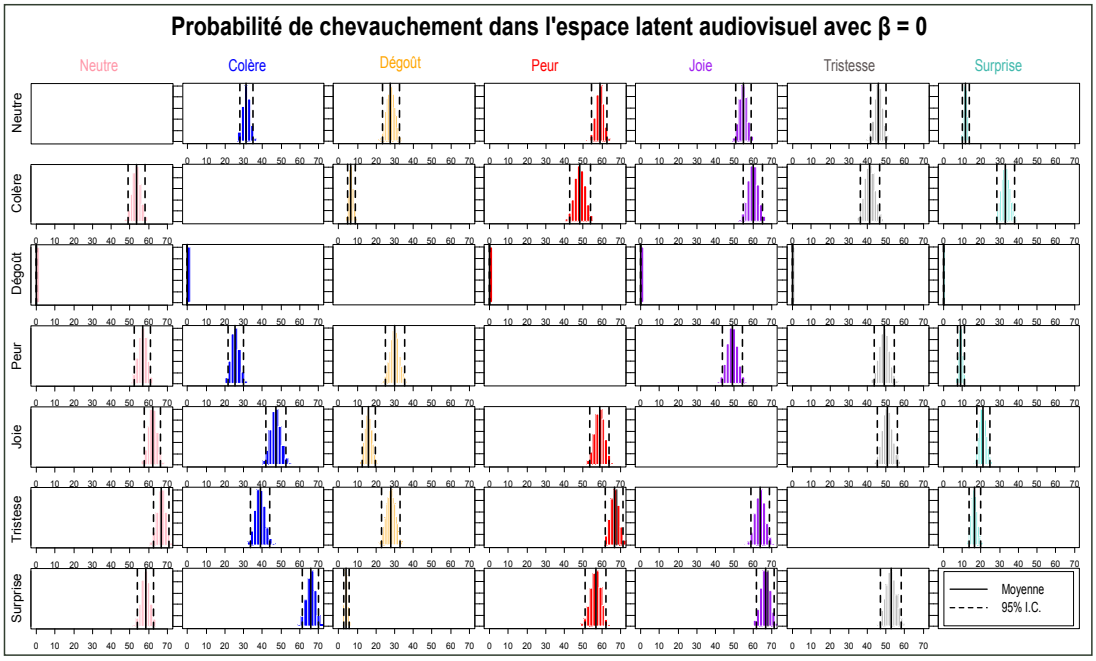


FIGURE 7.13 – *Distribution de la métrique de chevauchement probabiliste entre les clusters d'émotions pour le modèle* ***audiovisuel*** *avec $\beta = 0$ pour une intervalle de confiance de 95% (probabilité que les clusters des émotions présentées dans les colonnes débordent sur ceux des émotions présentées dans les lignes). Les moyennes et les intervalles de confiance de 95% des distributions sont présentés en lignes continues et discontinues respectivement.*

Le but est d'analyser l'impact d'un entraînement joint sur la structure de l'espace latent. Ce réseau est composé d'un encodeur avec une couche BLSTM de 1024 neurones, un vecteur latent de

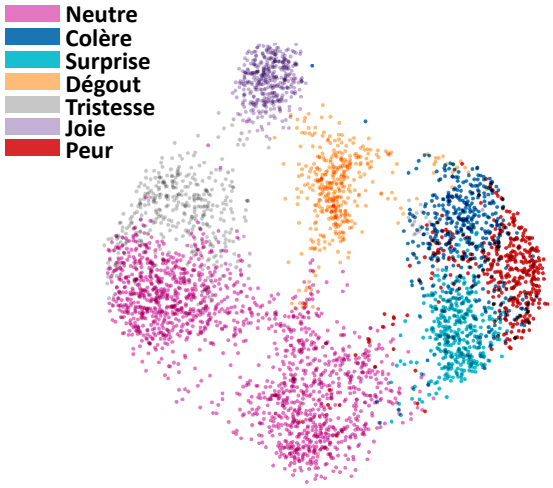


FIGURE 7.14 – *Représentation t-SNE de l'espace latent du modèle entraîné avec les données* ***audiovisuelles*** *conjointement. Les chevauchements sont déjà présents même sans l'introduction du terme de régularisation ($\beta = 0$).*

dimension 100 suivi d'un décodeur constitué de deux couches BLSTM de 1500 neurones chacune et une couche linéaire de sortie. Dans cette expérience, nous avons doublé la taille du vecteur latent (100) puisque pour les modèles séparés chaque modalité a été encodée dans un vecteur latent de dimension 50.

La figure 7.13 montre l'état de l'espace latent audiovisuel. Nous pouvons y voir que les clusters se chevauchent même sans introduire le terme de régularisation ($\beta = 0$). La représentation t-SNE de la figure 7.14 confirme également ces constatations. Ce résultat est conforme avec ceux trouvés dans l'expérience précédente avec les DNN-FC, où nous avons parlé d'une possible augmentation du nombre de combinaisons audiovisuelles. Cette augmentation résulte en une réduction du nombre d'échantillons d'apprentissage, et de ce fait détériore la qualité de la prédiction. Dans le cas des données audiovisuelles (figure 7.13), les clusters se chevauchent déjà avec $\beta = 0$, dans ce cas, il n'est donc pas nécessaire d'augmenter sa valeur.

7.3.3 Discussion

Dans cette section nous avons présenté l'architecture VAE que nous avons utilisé pour entraîner le modèle des durées, le modèle acoustique et le modèle visuel. L'apprentissage a été effectué sans utiliser les étiquettes des émotions et en conditionnant sur les paramètres linguistiques afin d'isoler la représentation des émotions dans l'espace latent. L'espace latent capturé par cette architecture semble porteur de sens, puisque nous pouvons y voir plusieurs clusters facilement identifiables, ce qui nous laisse penser que ces différents clusters représentent les différentes émotions du corpus original. En ajustant le paramètre β avec la métrique de chevauchement multidimensionnelle de Swanson et al. [2015], nous avons restructuré l'espace latent en le rendant plus lisse et cela en supprimant les éventuelles zones de discontinuités entre les clusters latents des émotions. Nous avons pu constater qu'en augmentant la valeur de β la structure de l'espace latent a changé de la manière que nous avons prévue en rapprochant les clusters de plus en plus.

De surcroît, les résultats de la métrique de chevauchement sont cohérents avec ce que nous avons obtenu dans le chapitre précédent avec l'architecture DNN-FC (chapitre 6). En fait, pour le modèle audiovisuel, le degré de chevauchement élevé entre les clusters latents, et sans introduction

du terme de régularisation ($\beta = 0$), explique la mauvaise performance du modèle audiovisuel joint. En combinant les données acoustiques et visuelles, l'espace latent échoue à former un cluster bien défini pour chaque émotion, ce que nous pensons est dû à la taille moyenne de notre corpus. Pour les autres modèles, nous pouvons remarquer une similarité entre les résultats des tables de validation croisée du chapitre précédent et les résultats de la métriques de chevauchement. Un chevauchement important entre deux clusters, signifie que leurs deux émotions respectives coexistent dans la même zone de l'espace latent, et donc qu'elles comportent des similarités. Pour les données visuelles, le tableau 6.6 du chapitre précédent et la figure 7.10 rapportent une similarité entre la tristesse et l'état neutre, entre la peur et la colère puis entre la surprise et la peur. Pour les données acoustiques, le tableau 6.7 du chapitre précédent et la figure 7.11 montrent des ressemblances entre la peur et le dégoût, entre l'état neutre et la tristesse puis entre la surprise et la joie. Concernant les données des durées, la figure 7.12 confirme que les durées de toutes les émotions sont très similaires, seule l'émotion de dégoût se démarque avec un débit de parole très faible. Ainsi, le cluster latent du dégoût chevauche à peine les autres clusters d'émotions.

L'étape suivante, est d'utiliser les modèles que nous avons entraînés pour générer des animations audiovisuelles 3D afin de les évaluer. Les résultats nous permettrons de juger l'efficacité de notre méthode. En fait, il faut évaluer la capacité de notre méthode à générer des animations expressives cohérentes avec des vecteurs latents fictifs et de ce fait profiter de l'aspect génératif du VAE.

7.4 Phase de synthèse

Comme présenté dans la figure 7.15, pendant la phase de synthèse, les encodeurs ne sont plus utilisés. Nous choisissons un vecteur z_d de l'espace latent des durées, nous le concaténons avec les informations linguistiques et nous le passons au décodeur pour prédire les durées correspondantes. Nous choisissons également les vecteurs z_a (respectivement z_v) de l'espace latent acoustique (respectivement visuel), en incorporant les informations linguistiques et les durées préalablement prédites par le modèle des durées puis nous synthétisons les données acoustiques (respectivement visuelles) en utilisant le décodeur acoustique (respectivement visuel). Les données prédites, acoustiques et visuelles, sont synchronisées puisqu'elles se basent sur les mêmes durées. Les trajectoires visuelles sont ensuite décomposées en poids de *blendshapes* pour animer l'agent virtuel 3D (voir section 5.6 du chapitre 5).

7.5 Évaluation

Pour évaluer notre méthode, nous avons réalisé quatre expériences perceptives pour valider différents résultats du CVAE. Pour chaque expérience, les données de durées, acoustiques et visuelles générées ont été utilisées pour créer des animations d'un agent virtuel 3D. Puisque nous animons uniquement la partie inférieure du visage de l'agent virtuel, nous avons délibérément flouté la partie supérieure de son visage pour éliminer tous biais involontaires relatifs à son manque d'expressivité. Pour ces expériences, nous avons aussi besoin des signaux originaux pour mener les comparaisons. Pour les données acoustiques originales (sans synthèse), nous avons utilisé le même Vocodeur WORLD que pour la synthèse afin de transformer les données en paramètres acoustiques, pour ensuite reconstruire les fichiers acoustiques. Cette étape est nécessaire pour éliminer le biais introduit par la dégradation de qualité dû à l'utilisation du vocodeur. La

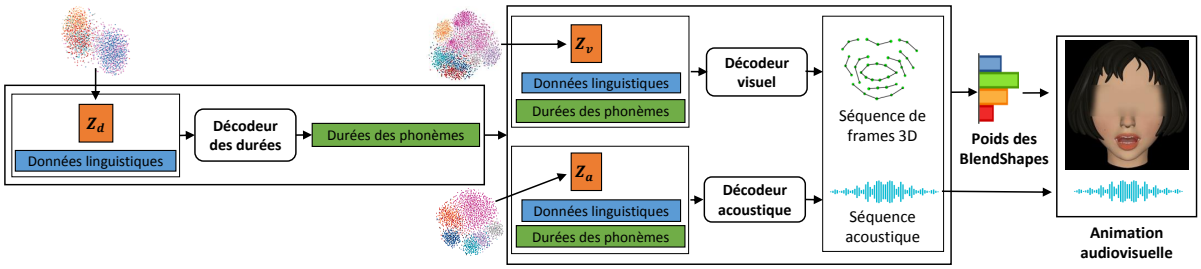


FIGURE 7.15 – *L'architecture du système d'animation audiovisuel à la phase de synthèse Les informations linguistiques ainsi que le vecteur z_d sont fournis au décodeur des durées pour prédire les durées. Les mêmes informations linguistiques, les durées prédites ainsi que les vecteurs z_a et z_v de l'espace latent acoustique et visuel sont donnés aux décodeurs acoustique et visuel pour générer une animation audiovisuelle synchronisée en utilisant la technique des blendshapes. La partie supérieure du visage de l'agent virtuel a été intentionnellement floutée.*

qualité du fichier audio généré par le Vocodeur représente la limite supérieure de la qualité qu'un système de synthèse peut atteindre.

7.5.1 Évaluation de la synthèse des émotions basiques

Dans cette expérience, nous évaluons la capacité de notre méthode à générer des émotions reconnaissables.

Stimuli

Dans cette expérience, nous avons généré 140 animations, divisées en deux ensembles, chacun contenant 70 animations (10 animations pour chaque classe d'émotion, neutre inclus). Le premier ensemble contient 70 animations 3D générées à partir des données originales : durées et trajectoires visuelles originales ainsi que les données acoustiques originales passées au Vocodeur. Le deuxième ensemble contient les animations 3D obtenues à partir des données synthétiques. Pour les obtenir, nous utilisons le vecteur moyen de chaque cluster d'émotion $z_{emo\_moyenne}$ pour générer les durées, les données acoustiques et visuelles. Ensuite nous utilisons ces données pour créer les 70 animations synthétiques. La représentation d'une émotion est définie comme le vecteur latent moyen du cluster latent de cette émotion. Nous le calculons en moyennant sur l'ensemble des vecteurs latents appartenant à une émotion donnée :

$$z_{emo\_moyenne} = \frac{1}{N}\sum_{i=1}^{N} z_i \tag{7.6}$$

Avec N la taille de l'ensemble d'apprentissage d'une émotion donnée. Nous définissons $z_{emo\_moyenne}$ pour chaque émotion et pour chacun des trois modèles entraînés (des durées, acoustique et visuel).

Méthode

À l'aide d'une application web, que nous avons mis en place, nous avons présenté à 12 participants 140 animations dans un ordre aléatoire. Pour chaque animation nous avons demandé aux participants de choisir dans une liste de sept choix (six émotions et le neutre) l'émotion qui,

selon eux, est exprimée par l'agent virtuel. L'interface utilisée pour cette expérience est illustrée dans la figure 7.16.

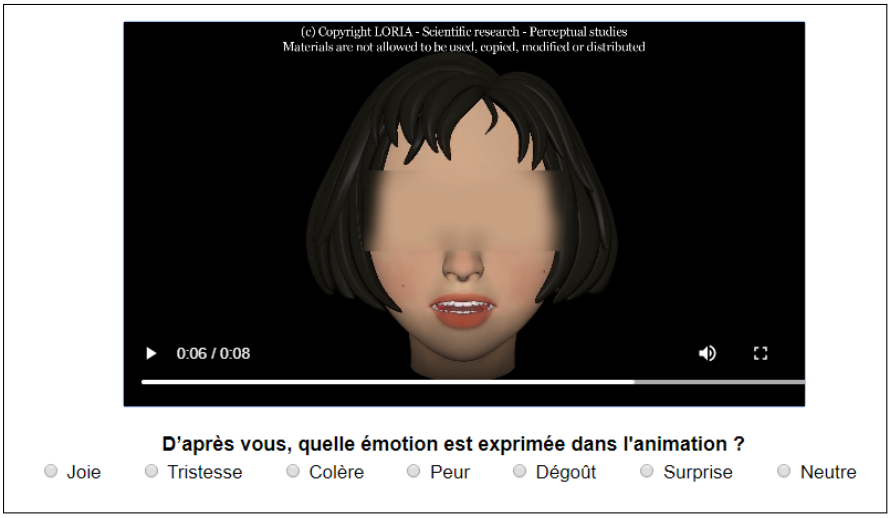


FIGURE 7.16 – *Interface de l'application web utilisée pour évaluer la capacité de notre système à générer des émotions reconnaissables. Les participants doivent choisir dans la liste des sept choix l'émotion exprimée, selon eux, par l'agent virtuel de l'animation.*

Nous n'avons pas pris en compte les 10 premiers tests dans le calcul des résultats, ces derniers étant considérés comme exemples d'accoutumance. Nous avons fait ce choix pour permettre aux participants de découvrir les différents styles de parole présents dans notre corpus et d'éviter les erreurs dues aux hésitations qui peuvent arriver au début des expériences. Les participants recrutés dans cette expérience ne comprennent pas tous la langue française, mais vivaient tous en France durant la période de participation aux expériences. Après la collecte des résultats, nous avons calculé les niveaux de significativité statistique en utilisant la valeur-p avec un test-t et nous avons corrigé les résultats obtenus avec la méthode de Holm–Bonferroni [Holm, 1979].

Résultats

Les résultats de cette expérience sont présentés dans les tableaux 7.1 et 7.2 sous forme de matrices de confusions des différentes émotions. Chaque ligne contient la répartition du taux de reconnaissances pour une émotion donnée. Le tableau 7.1 contient les résultats relatifs aux animations créées à partir des signaux originaux et le tableau 7.2 contient ceux relatifs aux animations provenant de la synthèse. Nous avons ajouté dans les diagonales des deux matrices le symbole (*) pour les résultats statistiquement significatifs et (-) pour ceux statistiquement non-significatifs.

Ces résultats confirment que les émotions présentées dans les animations audiovisuelles à partir des données de synthèse ont été correctement reconnues avec un taux supérieur à 71% pour la majorité des émotions, excepté pour la tristesse et la peur. En effet, ces deux émotions sont les plus dures à reconnaître, même pour les animations originales. Ce résultat était attendu, car la partie supérieure du visage est cruciale pour reconnaître ces émotions [Bassili, 1979, Costantini et al., 2005]. Nous constatons les mêmes tendances de confusion entre les émotions originales et celles synthétiques. Un grand pourcentage de confusion a été détecté entre la peur et la tristesse, entre la joie et la surprise, de plus, quelques animations contenant la peur et la tristesse ont été perçues comme appartenant à l'état neutre, ce qui explique leur faible taux de reconnaissance.

Notons que certaines émotions synthétiques ont été mieux reconnues que celles originales. C'est le cas du dégoût, la joie et légèrement la surprise. Toutefois, après vérification nous avons trouvé que la différence entre les résultats de ces trois émotions et ceux des animations originales n'est pas statistiquement significative. Nous pensons que cela est dû à l'utilisation du même vecteur latent z_{emo_100} pour la génération de toutes les animations synthétiques d'une émotion donnée. Les participants ont pu détecter le pattern lié au z_{emo_100} choisi et ainsi identifier plus facilement l'émotion synthétique. Cela montre également que la représentation latente a bien capturé la spécificité de chaque émotion. Nous rappelons ici qu'aucune étiquette d'émotion n'a été utilisée lors de la phase d'apprentissage. Les étiquettes des émotions sont uniquement utilisées pour calculer les z_{emo_100} lors de la phase de synthèse.

Nous pouvons aussi constater que les mêmes tendances de confusions entre les données synthétiques et celles originales. La confusion a été détectée entre la peur et la tristesse, entre la joie et la surprise, la peur et la tristesse ont également été perçues comme neutre, ce qui explique leur faible taux de reconnaissance.

		Émotion perçue						
		Colère	**Dégoût**	**Peur**	**Joie**	**Neutre**	**Tristesse**	**Surprise**
Émotion produite	**Colère**	**97.50(*)**	0.00(-)	0.00(-)	0.00(-)	0.00(-)	0.00(-)	2.50(-)
	Dégoût	0.83(-)	**67.50(*)**	8.33(-)	0.00(-)	0.83(-)	22.50(-)	0.00(-)
	Peur	15.00(-)	5.00(-)	**42.50(*)**	0.00(-)	12.50(-)	22.50(-)	2.50(-)
	Joie	18.33(-)	0.00(-)	0.00(-)	**69.17(*)**	1.67(-)	0.83(-)	10.00(-)
	Neutre	0.00(-)	0.00(-)	4.17(-)	12.50(-)	**77.50(*)**	4.17(-)	1.67(-)
	Tristesse	2.50(-)	0.00(-)	32.50(-)	5.00(-)	0.83(-)	**57.50(*)**	1.67(-)
	Surprise	16.67(-)	0.00(-)	0.83(-)	10.00(-)	0.00(-)	0.00(-)	**72.50(*)**

TABLE 7.1 – *La matrice de confusion des animations générées à partir des données audiovisuelles* ***originales****. La diagonale représente le pourcentage de réponses correctes. Les colonnes représentent la répartition des réponses fournies par les participants. Le symbole (*) indique que les résultats sont statistiquement significatifs et (-) qu'ils sont statistiquement non-significatifs.*

		Émotion perçue						
		Colère	**Dégoût**	**Peur**	**Joie**	**Neutre**	**Tristesse**	**Surprise**
Émotion produite	**Colère**	**71.67(*)**	15.83(-)	5.00(-)	0.00(-)	5.00(-)	0.83(-)	1.67(-)
	Dégoût	1.67(-)	**83.33(*)**	1.67(-)	0.00(-)	3.33(-)	10.00(-)	0.00(-)
	Peur	8.33(-)	11.67(-)	**11.67(-)**	0.83(-)	42.50(-)	20.83(-)	4.17(-)
	Joie	5.00(-)	0.00(-)	3.33(-)	**71.67(*)**	5.00(-)	4.17(-)	10.83(-)
	Neutre	0.00(-)	0.00(-)	1.67(-)	0.83(-)	**92.50(*)**	5.00(-)	0.00(-)
	Tristesse	5.00(-)	7.50(-)	15.00(-)	0.00(-)	45.00(-)	**26.67(-)**	0.83(-)
	Surprise	5.00(-)	0.00(-)	4.17(-)	9.17(-)	8.33(-)	0.00(-)	**73.33(*)**

TABLE 7.2 – *La matrice de confusion des animations générées à partir des données audiovisuelles* ***synthétique****. La diagonale représente le pourcentage de réponses correctes. Les colonnes représentent la répartition des réponses fournies par les participants. Le symbole (*) indique que les résultats sont statistiquement significatifs et (-) qu'ils sont statistiquement non-significatifs.*

7.5.2 Évaluation de la qualité de l'articulation

Dans cette expérience, nous évaluons la capacité de notre méthode à générer des sons et des gestes articulatoires cohérents.

Stimuli

Dans cette expérience, nous utilisons les mêmes 140 animations utilisées dans l'expérience précédente.

Méthode

À l'aide d'une application web, que nous avons mise en place, nous avons présenté à 19 participants, tous des francophones, 140 animations dans un ordre aléatoire. Pour chaque animation nous avons demandé aux participants de noter le degré de correspondance entre les sons prononcés et les mouvements des lèvres de l'agent virtuel. Les degrés de cohérence présentés sont les suivants : 1) jamais (0%), 2) rarement (25%), 3) moyennement (50%), 4) souvent (75%) et 5) tout le temps (100%). L'interface utilisée pour cette expérience est présentée dans la figure 7.17.

FIGURE 7.17 – *Interface de l'application web utilisée pour évaluer la capacité de notre système à générer des sons et des gestes articulatoires cohérents. Les participants doivent mettre le curseur à l'emplacement convenable.*

Nous n'avons pas pris en compte les 10 premiers tests dans le calcul des résultats, ces derniers étant considérés comme exemples d'accoutumance. Après la collecte des résultats, nous avons calculé avec un test-t le niveau de significativité statistique de la différence entre les résultats des animations synthétiques et celles originales.

Résultats

Les résultats de cette expérience sont présentés dans les tableaux 7.3. Nous y présentons les résultats pour chaque émotion des animations originale et celles synthétiques.

	Colère	**Dégoût**	**Peur**	**Joie**	**Neutre**	**Tristesse**	**Surprise**
Originales	72.53	72.76	77.53	68.03	75.44	72.67	69.67
Synthétiques	76.57	71.04	73.11	69.86	78.69	74.86	72.60

TABLE 7.3 – *Le degré de cohérence entre les sons prononcés et les mouvements des lèvres de l'agent virtuel pour les animations originales et celles synthétiques en considérant une échelle de 0 (jamais) à 100 (tout le temps).*

Ces résultats montrent que les animations originales et celles synthétiques contiennent des sons et des mouvements des lèvres cohérents. De plus, en utilisant un test-t pour comparer les résultats des deux groupes d'animations (originales et synthétiques), nous avons constaté que, pour toutes les émotions, il n'y a pas de différence statistiquement significative entre les résultats des deux groupes. Ces résultats confirment la qualité de l'articulation des animations synthétiques générées avec notre méthode et confirment qu'en plus d'exprimer correctement les différentes émotions, notre système est capable de générer des sons et des mouvements des lèvres cohérents dans un contexte expressif.

7.5.3 Évaluation de la synthèse des nuances d'émotions

Le but de cette deuxième expérience est d'évaluer la capacité de notre système à générer des nuances d'une émotion. Nous avons utilisé le vecteur qui correspond à une combinaison linéaire entre le vecteur moyen du cluster de l'état neutre et les vecteurs moyens des autres six clusters d'émotions.

Stimuli

Pour les données synthétiques, pour chaque émotion nous considérons z_{emo_100} le vecteur moyen de chaque cluster comme représentation à "100%" d'une émotion. Nous avons généré des animations à 100% d'intensité de toutes les émotions synthétiques z_{emo_100}. Ensuite, et en utilisant une combinaison linéaire entre les centres des différents clusters d'émotions et celui de l'état neutre, nous avons généré des nuances à 33% et 67% de chaque émotion comme suit :

$$z_{emo\_100} = z_{emo\_moyenne} \tag{7.7}$$

$$z_{emo\_67} = z_{neutre\_100} \times 0.33 + z_{emo\_100} \times 0.67 \tag{7.8}$$

$$z_{emo\_33} = z_{neutre\_100} \times 0.67 + z_{emo\_100} \times 0.33 \tag{7.9}$$

Nous pouvons noter également :

$$z_{emo\_0} = z_{neutre\_100} \tag{7.10}$$

Nous avons généré cinq exemples d'animations pour chaque émotion et pour l'état neutre, que ce soit pour les données originales ou synthétiques. Ensuite, nous avons généré, pour chacun de ces exemples synthétiques, leurs nuances à 33% et 67% d'intensité.

Méthode

Nous avons généré 210 comparaisons, chaque comparaison inclut deux animations à deux degrés d'intensité différents d'une émotion donnée. Les 7 comparaisons effectuées sont :

1. z_{emo_0} et z_{emo_33} ;
2. z_{emo_0} et z_{emo_67} ;

3. z_{emo_0} et z_{emo_100} ;
4. z_{emo_33} et z_{emo_67} ;
5. z_{emo_33} et z_{emo_100} ;
6. z_{emo_67} et z_{emo_100} ;
7. z_{emo_100} et $z_{emo\_originale}$.

Nous avons présenté les 210 comparaisons dans un ordre aléatoire à 10 participants et nous leur avons demandé, pour chacune, de choisir l'animation qui leur semble la plus expressive. Les participants n'ont pas été informés de la nature de l'émotion jouée par l'agent virtuel dans les différentes animations. Un exemple d'écran de comparaison de cette expérience est présenté dans la figure 7.18.

FIGURE 7.18 – *Capture d'écran du test perceptif d'évaluation de la capacité de notre système à générer des nuances d'émotions. Les participants doivent choisir des deux animations présentées, et selon eux, l'animation la plus expressive.*

Nous n'avons pas pris en compte les 10 premiers tests dans le calcul des résultats, ces derniers étant considérés comme exemples d'accoutumance. Nous avons fait ce choix pour permettre aux participants de découvrir les différents styles de parole présents dans notre corpus et d'éviter les erreurs dues aux hésitations qui peuvent arriver au début des expériences. Les participants recrutés dans cette expérience ne comprennent pas tous la langue française, mais vivaient tous en France durant la période de participation aux expériences. Après la collecte des résultats, nous avons calculé les niveaux de significativité statistique en utilisant la valeur-p avec un test-t et nous avons corrigé les résultats obtenus avec la méthode de Holm–Bonferroni [Holm, 1979]. Le symbole (-) indique un résultat statistiquement non-significatif, le reste des résultats sont tous statistiquement significatifs.

Résultats

Pour cette expérience, sur le tables 7.4 nous pouvons remarquer qu'en moyenne, le degré des nuances a été bien respecté (66% pour 0/33 et >80% pour les autres comparaisons). Les scores de comparaisons entre l'état neutre et les différentes nuances (0/33, 0/67 et 0/100) montrent que les émotions sont globalement bien perçues et facilement détectables, surtout pour la comparaison 0/100. La nuance subtile de 33% (z_{emo_33}) a été la plus difficile à détecter avec un score en dessous de 70% notamment pour la peur, la tristesse et le dégoût. Concernant les comparaisons des nuances entre elles (33/67, 33/100 et 67/100), les scores (> 80 %) montrent que la graduation

des émotions est représentée avec succès par la combinaison linéaire de vecteurs latents. Les participants ont pu percevoir la différence entre les différentes nuances des émotions et identifier correctement l'animation moins / plus expressive. Ces résultats sont très intéressants car ils prouvent que nous avons réussi à restructurer l'espace latent et à le rendre continu. En fait, les vecteurs utilisés pour générer les différentes nuances ne correspondent à aucune donnée réelle vue par le réseau de neurones pendant l'apprentissage. Ce sont en fait des vecteurs originaux et les nuances d'émotions que nous avons générées par combinaison linéaire sont complètement inventées. Les animations originales ont été globalement vues comme plus expressives que celles synthétiques (à 100%). Ce résultat peut s'expliquer par le fait qu'en mettant côte à côte les deux animations, les imperfections des données synthétiques deviennent plus visibles et facilement identifiables. D'autant plus que certains détails fins de la voix sont perdus au cours du processus d'apprentissage (tremblements, craquements de la voix). Les animations originales ont également une prosodie plus riche, contenant plus de variabilité au sein de la même phrase, tandis que notre modèle de durée semble moyenner les durées des phonèmes et résulte en un parole plus monotone.

	0/33	**0/67**	**0/100**	**33/67**	**33/100**	**67/100**	**100/originale**
Colère	82	94	90	94	96	88	82
Dégoût	52 (-)	80	82	92	86	70	86
Peur	58 (-)	56 (-)	80	66	72	80	88
Joie	74	92	96	90	90	90	91
Tristesse	56 (-)	70	88	74	76	86	95
Surprise	78	92	92	90	94	86	98
Moyenne	***66***	***80***	***88***	***84***	***85***	***83***	***90***

TABLE 7.4 – *Pourcentage des réponses choisissant correctement les animations les plus expressives lors de la comparaison des animations des nuances des émotions deux par deux. Les degrés des émotions comparées sont 0% (= 100% neutre), 33%, 67%, 100% et l'animation originale de chaque émotion. Le symbole (-) indique un résultat statistiquement non-significatif, le reste des résultats sont tous statistiquement significatifs.*

7.5.4 Évaluation de la synthèse des mélanges d'émotions

Dans cette expérience, nous évaluons la capacité de notre système à générer des mélanges d'émotions en interpolant les vecteurs latents entre eux.

Stimuli

Afin d'obtenir des mélanges d'émotions cohérents et humainement significatifs (par exemple, la colère et le dégoût entraînent le mépris), nous nous sommes inspiré de la roue de Plutchik [1984] . Nous avons défini les 4 scénarios de mélange d'émotions suivants :

1. Colère et dégoût (mépris) ;
2. Tristesse et dégoût (remord) ;
3. Tristesse et surprise (désappointement) ;
4. Peur et surprise (crainte).

Les mélanges sont effectués en calculant le vecteur latent situé à mi-chemin entre les deux émotions sources. Ce vecteur est calculé comme suit :

$$z_{melange} = z_{emo1\_moyenne} \times 0.5 + z_{emo2\_moyenne} \times 0.5 \tag{7.11}$$

Pour les 4 scénarios de mélanges, nous avons généré 5 exemples. Chaque exemple contient 5 animations :

— originale $emotion_1$,
— originale $emotion_2$,
— $emotion_1$ à 100%,
— $emotion_2$ à 100%,
— 50% de $emotion_1$ mélangé avec 50% de $emotion_2$.

Cette expérience contient un total de 100 animations.

Méthode

Nous avons présenté les 100 animations dans un ordre aléatoire à 12 participants. Nous avons demandé aux participants d'estimer la contribution des émotions mélangées avec un curseur qui a pour extrémités les deux émotions mélangées $emotion_1$ et $emotion_2$ (voir figure 7.19). Les participants avaient la possibilité de mettre le curseur à n'importe quelle valeur entre les deux extrémités.

FIGURE 7.19 – *Capture d'écran du test perceptif d'évaluation de la capacité de notre système à générer des mélanges d'émotions. Les participants doivent estimer, selon eux, la contribution des émotions mélangées avec un curseur qui a pour extrémités les deux émotions mélangées.*

Nous n'avons pas pris en compte les 10 premiers tests dans le calcul des résultats, ces derniers étant considérés comme exemples d'accoutumance. Les participants recrutés dans cette expérience ne comprennent pas tous la langue française, mais vivaient tous en France durant la période de participation aux expériences. Après la collecte des résultats, nous avons calculé les niveaux de significativité statistique en utilisant la valeur-p avec un test-t et nous avons corrigé les résultats obtenus avec la méthode de Holm–Bonferroni [Holm, 1979].

Résultats

Les résultats de cette expériences sont présentés dans la figure 7.20. Ces résultats montrent que pour les 4 scénario de mélange d'émotions, l'émotion créée par interpolation des vecteurs latents (à 50% chacun), et qui est représentée en vert, a été toujours perçue comme une émotion intérimaire entre les deux émotions mélangées.

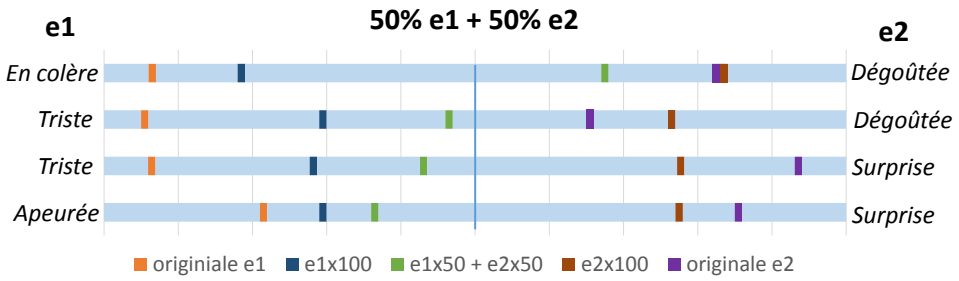


FIGURE 7.20 – *Les mélanges d'émotions (en vert) ont été perçus comme une émotion intermédiaire entre e_1 et e_2 pour les quatre scénario de mélange des émotions.*

Les résultats de cette expérience montrent que notre système a réussi à créer des mélanges d'émotions qui ont été correctement perçues comme des émotions intermédiaires dans les quatre scénarios de mélange considérés. Comme nous l'avons dit dans l'expérience précédente, les vecteurs que nous avons créés en combinant linéairement des vecteurs latents ne correspondent à aucune donnée réelle dans notre corpus. C'est le point fort du VAE en tant que modèle génératif, car il est capable de générer une sortie cohérente à partir de vecteurs latents fictifs / inventés. De plus, nous soulignons que même les vecteurs combinés linéairement (les centroïdes) sont eux-mêmes inexistants dans le corpus d'origine. Ce résultat valide à nouveau la continuité de l'espace latent restructuré. Ces résultats, rejoignent ceux de l'expérience précédente (générer des nuances d'émotion), et montrent que la structure de l'espace latent permet d'accéder à des zones intermédiaires aux clusters des émotions qui étaient inaccessibles avec une architecture standard du type DNN-FC ou avec une architecture CVAE sans introduction du terme de régularisation. En ce qui concerne les animations originales, nous pouvons voir qu'elles étaient principalement perçues comme plus proches de la définition de l'émotion que celles synthétiques à 100%. La seule exception est le dégoût, puisque ses animations synthétiques étaient considérées comme plus dégoûtées que les animations originales. Ces résultats confirment ceux de la première expérience (7.5.1 Évaluation de la synthèse des émotions basiques).

7.6 Conclusion

Dans ce chapitre, nous avons exploré les possibilités offertes par une architecture β-CVAE pour la synthèse expressive audiovisuelle de la parole à partir du texte. Nous avons entraîné trois modèles séparés pour chacun des aspects étudiés de la parole : un modèle CVAE pour les durées, un autre pour l'aspect acoustique et un dernier pour l'aspect visuel. Ces modèles ont été entraînés de manière non-supervisée, sans utiliser les étiquettes des émotions. En conditionnant le VAE sur les paramètres linguistiques, nous avons réussi à capturer la représentation des émotions dans l'espace latent qui s'est scindé en plusieurs clusters bien identifiables. Puisque les techniques de visualisations des espaces multidimensionnels sont peu robustes, nous nous sommes inspirés du domaine de l'écologie pour analyser de manière objective les niveaux de chevauchement des clusters des émotions dans l'espace latent. Nous avons utilisé une métrique probabiliste multidimensionnelle. Cette analyse nous a permis de choisir les paramètres β pour chacun de nos modèles pour restructurer de manière optimale les espaces latents obtenus. À travers cette restructuration, nous avons réussi à rendre l'espace latent complètement malléable et capable de générer de nouveaux styles de parole. Les résultats de notre système ont été validés par quatre expériences perceptuelles qui ont confirmé la capacité de notre système à générer des émotions

reconnaissables avec une articulation cohérente. De plus, la nature générative du CVAE nous a permis de générer des nuances bien détectées des six émotions et de mélanger différentes émotions. Ces résultats montrent que nous avons réussi à bien isoler la représentation des émotions dans l'espace latent du CVAE et de restructurer se dernier en le rendant complètement malléable. L'espace latent est devenu particulièrement robuste, car nous avons pu générer des sorties cohérentes à partir de vecteurs fictifs créés par interpolation linéaire des vecteurs latents réels.

Enfin, le travail présenté dans ce chapitre représente une feuille de route pour mettre en place un système de TTS audiovisuelle expressive permettant d'aller au delà des classes d'émotions du corpus original en créant de nouveaux styles de parole. cette approche peut être appliquée dans d'autres domaines pour profiter de l'existence de nombreuses sources de données non annotées. Elle permettra d'ajouter de la diversité dans les données générées et de surmonter, même partiellement, le problème de manque de données dans certains domaines qui reste l'un des grands freins devant l'utilisation des réseaux de neurones.

Conclusion et perspectives

Le travail réalisé au cours de cette thèse a porté sur la modélisation des émotions avec des DNNs et a permis d'élaborer les différentes étapes nécessaires à la réalisation d'un système de synthèse audiovisuel de parole expressive permettant de générer des mélanges et des nuances d'émotions. Dans cette dernière partie, nous allons souligner les différentes contributions et résultats importants de notre travail et proposer un certain nombre de pistes pour des recherches à venir.

Synthèse des contributions

La première partie de ce manuscrit a été consacrée aux études théoriques sur la synthèse audiovisuelle et expressive de la parole. Nous avons regroupé et analysé les connaissances qui nous semblent les plus pertinentes de l'état de l'art de la TTS audiovisuelle expressive.

La deuxième partie de ce manuscrit a porté sur la construction et l'étude de deux corpus audiovisuels expressifs. Nous avons effectué cette étape en raison de l'absence de corpus audiovisuels expressifs contenant des classes d'émotions parallèles et équilibrées notamment avec des données visuelles en trois dimensions. Le premier corpus étant un prototype, il nous a permis de valider notre protocole d'acquisition des données audiovisuelles expressives. De plus, et contrairement aux analyses menées habituellement sur les émotions dans un contexte statique (images ou vidéos sans parole), nous nous sommes intéressés dans cette étude aux émotions dans un contexte dynamique représenté par l'activité de la parole. Cette étude a démontré que les principales caractéristiques faciales des émotions sont maintenues même pendant la parole, et que les gestes liés à l'articulation sont présents dans les trois premières composantes principales de toutes les émotions étudiées. Après validation du contenu expressif du corpus via des tests perceptifs, nous avons constaté qu'une représentation minimaliste du visage (par capteurs faciaux uniquement) est capable de transmettre un taux important du contenu expressif et peut donc être utilisée dans l'acquisition d'autres corpus. En nous appuyant sur les résultats de ce corpus prototype, nous avons acquis un corpus de taille plus importante. En vue d'utiliser ce corpus dans un processus de TTS, nous avons effectué une analyse linguistique fine pour assurer une bonne couverture di-phonétique. Ce corpus a également été analysé par des tests perceptifs pour valider son contenu expressif. Les émotions ont été reconnues avec des taux statistiquement significatifs, que ça soit avec les séquences acoustiques et visuelles séparées qu'avec les séquences vidéo audiovisuelles. Les taux de reconnaissance pour les séquences audiovisuelles de toutes les émotions sont supérieurs à 73% et montrent que les deux modalités acoustique et visuelle sont complémentaires. Ces résultats montrent également que la majorité des participants valident les performances de l'actrice et confirment la bonne qualité du corps expressif produit. À partir de ces résultats, nous avons envisagé d'utiliser ce corpus à des fins de TTS audiovisuelle expressive.

Dans la troisième partie de ce travail, nous nous sommes penchés sur l'utilisation des réseaux de neurones dans la TTS audiovisuelle expressive. À cette fin, nous avons étudié deux archi-

tectures neuronales différentes. Tout d'abord, nous avons adopté une architecture entièrement connectée DNN-FC pour analyser le comportement des réseaux DNN-FF et BLSTM face aux facteurs contextuels et linguistiques pour la synthèse des durées des phonèmes et des modalités acoustique et visuelle. Les résultats montrent que le modèle des durées semble profiter de toutes les informations linguistiques, que ça soit pour DNN-FF ou BLSTM. Toutefois, pour les modèles acoustiques et visuels avec un réseau BLSTM, les informations autres que les contextes gauches et droits immédiats ne semblent pas améliorer la prédiction. Nous avons également comparé la qualité de la synthèse des modèles acoustique et visuel entraînés séparément puis conjointement. À l'aide d'une comparaison avec des mesures objectives, nous avons trouvé que les modèles entraînés séparément atteignent une meilleure précision de reconstruction. Une hypothèse sur la baisse de précision des modèles acoustiques et visuels avec un contexte linguistique important ou un vecteur d'entrée de grande taille (audiovisuel), serait l'augmentation du nombre de combinaisons possibles pour les données, ce qui a pour effet de réduire le nombre d'exemples de chacune d'elles pendant l'apprentissage. Nous pensons que c'est pour cette raison que cet effet ne semble pas se produire avec les données des durées, qui ont un vecteur de très petite taille contenant un seul paramètre. De ce fait, il serait intéressant, pour compléter cette étude, de faire une analyse sur un corpus de taille plus importante pour voir si ces conclusions peuvent être maintenues. Finalement, les résultats objectifs de la validation-croisée effectuée sur les différentes émotions montrent que les trois modèles arrivent à se spécialiser aux différentes émotions. Ces résultats nous ont aussi permis de constater des similarités et des différences entre certaines émotions.

L'architecture DNN-FC souffre néanmoins d'un certain nombre de limitations. Tout d'abord, elle nécessite des étiquettes des émotions pendant la phase de l'entraînement. De plus, elle ne permet de générer que les styles de parole présents dans le corpus d'apprentissage. Nous avons donc adopté une seconde méthode qui repose sur une architecture neuronale de type auto-encodeur variationnel sous condition CVAE. Cette nouvelle architecture a l'avantage d'être entraînable de façon non-supervisée (sans étiquettes d'émotions). En conditionnant le CVAE sur le contenu linguistique, nous avons réussi à capturer une représentation latente des émotions. L'intérêt d'isoler une représentation latente des émotions est de pouvoir la restructurer pour supprimer les éventuelles discontinuités et de ce fait pouvoir accéder à la totalité du spectre des émotions et donc accéder à des états émotionnels intermédiaires. Afin d'avoir une représentation latente optimale, nous avons proposé une méthode pour définir le paramètre β de la fonction de perte du CVAE. Cette méthode, inspirée du domaine de l'écologie, est une feuille de route pour choisir un paramètre β permettant aux clusters de l'espace latent d'être suffisamment proches pour assurer sa continuité tout en ayant un impact minimal sur la qualité de la reconstruction. En appliquant ce protocole, nous avons réussi à modifier la structure de l'espace latent du modèle des durées, du modèle acoustique et du modèle visuel. À l'aide de quatre expériences perceptives, nous avons pu valider les possibilités offertes par notre système. Tout d'abord, nous avons validé la capacité de ce dernier à générer des émotions basiques reconnaissables avec une articulation et des sons cohérents. Notre système nous a également permis de générer des nuances d'émotions et des mélanges d'émotions par interpolation linéaire des vecteurs latents.

Ces résultats montrent l'efficacité de notre approche et représentent un autre pas vers la synthèse TTS audiovisuelle expressive réaliste. Nous résumons dans la section suivante quelques perspectives se dégageant de ce travail.

Perspectives

Les travaux de thèse présentés dans ce manuscrit décrivent la construction et l'analyse de deux corpus expressifs. Toutefois, la partie de synthèse de la parole a été menée sur un seul corpus

d'une seule locutrice. Pour élargir la portée des résultats obtenus, le premier point intéressant serait d'acquérir d'autres corpus expressifs avec différents locuteurs. Ce travail nous permettra de déterminer si les conclusions établies peuvent être étendues à d'autres personnes ou si elles sont spécifiques au corpus utilisé. Il est également envisageable de combiner la représentation latente des émotions à une représentation latente des locuteurs. Nous pourrons par exemple transférer l'expressivité du corpus expressif d'un locuteur donné vers un autre locuteur avec un corpus neutre.

Pour approfondir l'analyse que nous avons effectuée avec une architecture DNN-FC, il serait intéressant de pousser les évaluations des différents modèles, notamment celui des durées, à un niveau plus fin. Au lieu d'étudier l'erreur de prédiction au niveau de la phrase en entier, nous pouvons approfondir l'étude, par exemple, au niveau des syllabes ou des phonèmes, pour voir si certaines syllabes ou phonèmes sont mieux modélisés que d'autres suivant leur contexte linguistique ou émotionnel.

En ce qui concerne l'architecture CVAE, nous proposons quelques pistes qui nous semblent pertinentes pour compléter nos travaux. Tout d'abord, dans notre approche, nous n'avons pas utilisé les étiquettes des émotions pendant la phase d'apprentissage, mais nous nous en sommes servis pour la phase de synthèse. Leur utilisation était nécessaire pour valider le bon fonctionnement de notre approche. Toutefois, si nous sommes amenés à entraîner notre système avec des corpus non-annotés, il serait intéressant d'utiliser des algorithmes tel que Elbow [Bholowalia and Kumar, 2014], qui déterminent un nombre optimal de clusters pour un jeu de données. Couplé à une méthode de visualisation, comme t-SNE par exemple, nous pouvons vérifier manuellement le contenu expressif de quelques exemples afin de déterminer l'étiquette de chaque cluster de données. Dans le cas d'un corpus faiblement annoté, nous pouvons nous passer de la phase de vérification manuelle. Si, idéalement, nous disposons d'un modèle pré-entraîné de reconnaissance des émotions, nous pouvons l'utiliser pour générer la totalité ou une partie des étiquettes du corpus.

Pour la représentation latente des émotions, nous avons utilisé les vecteurs latents obtenus pour générer des animations de synthèse. Cette représentation latente pourrait être analysée pour étudier son contenu comme effectué par Tits et al. [2019]. Nous pourrons ainsi, représenter sur l'espace latent les directions d'évolution des différents paramètres de durées, acoustiques et visuels afin de mieux comprendre les caractéristiques de chaque émotion.

Une autre idée serait l'utilisation des systèmes de synthèse de bout-en-bout (E2E). Sachant que ce genre de système est gourmand en ressources, après construction d'un plus grand corpus, nous souhaitons les utiliser pour tester notre méthode. Ce choix nous permettra de nous passer de la phase de phonétisation et d'alignement du texte, ce qui allégera remarquablement la phase de post-traitement du corpus.

Pour tenter d'améliorer la qualité des données prédites, une piste serait une architecture de type GAN (Generative Adversarial Network) [Goodfellow et al., 2014] ou une architecture hybride VAE-GAN [Larsen et al., 2016]. Ces réseaux ont été utilisés dans le domaine de génération d'images, et ont démontré de très bonnes performances [Denton et al., 2015, Radford et al., 2015] des fois même supérieures à celles d'un VAE [Larsen et al., 2016]. Les réseaux de types GAN ont été utilisés récemment pour la synthèse acoustique expressive [Ma et al., 2019] et leurs résultats représentent aujourd'hui l'état de l'art de la TTS acoustique expressive. Il serait donc intéressant d'appliquer ces architectures dans le cadre de la synthèse audiovisuelle expressive de la parole.

Dans ce travail nous nous sommes focalisés sur l'animation de la partie inférieure du visage de l'agent virtuel. Toutefois, certaines émotions, telles que la tristesse et la peur, sont essentiellement communiquées par la partie supérieure du visage [Bassili, 1979, Costantini et al., 2005]. De plus, dans le contexte de l'expression d'un mélange d'émotion, les expressions faciales résultantes sont

obtenues en utilisant une approche compositionnelle. Le visage est décomposé en zones faciales sur lesquelles les messages émotionnels peuvent être diffusés. De ce fait, l'expression résultante d'un mélange pourrait être mieux détectée lorsque le visage est entièrement animé [Pelachaud and Poggi, 2002, Bui et al., 2004, Martin et al., 2006, Niewiadomski et al., 2008]. Contrairement aux mouvements de la partie inférieure du visage, les mouvements de la partie supérieure du visage, notamment les sourcils, ne sont pas liés à l'articulation. Certaines études [Granström et al., 1999, House et al., 2001] lient les mouvements des sourcils à des indices perceptuels de la proéminence d'un mot, ou alors qu'ils ne sont pas simplement liés à la ponctuation de la phrase, mais pourraient être liés à la cohérence au sein d'une phrase [Granström et al., 1999]. L'étude de Pelachaud et al. [1996] montre que l'élévation des sourcils peut être utilisée pour marquer une nouvelle information. Bolinger and Bolinger [1989] ont trouvé une corrélation entre l'augmentation de la F0 et les mouvements de sourcils. Comme poursuite de ce travail, il serait intéressant d'entraîner un modèle DNN pour prédire les mouvements des sourcils. Il conviendrait donc d'inclure des informations sur la sémantique et la proéminence des mots dans une phrase ainsi que les informations acoustiques (prédites par notre modèle acoustique) pour générer les mouvements du haut du visage et obtenir une animation complète du visage.

Finalement, dans ce travail, nous traitons l'expression d'émotion comme manifestation superficielle de ce que l'humain peut ressentir, mais nous ne traitons pas les aspects complexes relatifs au ressenti profond, à l'analyse de l'affect dans un texte ou à l'adaptation aux comportements de l'autre durant une conversation. Dans une interaction humain-machine, l'agent virtuel doit être capable d'afficher une réponse émotionnelle appropriée pour réussir l'échange. Cette réponse émotionnelle doit correspondre au contexte de l'échange et aux émotions et à la personnalité de l'humain. Toutefois, la réalisation d'un tel système nécessite d'implémenter une vraie intelligence émotionnelle [Salovey et al., 2000, Kihlstrom and Cantor, 2000, Ochs et al., 2013] qui nécessite un effort pluridisciplinaire pour pouvoir espérer de le concrétiser un jour.

A

Première annexe

Scénarios utilisés pour enregistrer le corpus audiovisuel expressive

Ces scénarios sont choisis parmi une liste de plusieurs scénarios du corpus GEMEP [Bänziger and Scherer, 2010].

Colère : Je viens de surprendre deux adolescents en train de vandaliser ma voiture. Ils ont, non seulement, forcé la portière pour voler mon auto-radio, ils ont également rayé la carrosserie, arraché l'antenne et les rétroviseurs et crevé les pneus. J'arrive à rattraper l'un des deux qui se trouve être le fils de mes voisins de palier. Je n'ai pas beaucoup de sympathie pour ces voisins qui ne cessent de se disputer avec tous les habitants de l'immeuble et passent leur temps à créer des problèmes. Je ramène l'adolescent chez ses parents et j'exprime mes sentiments sur son comportement inqualifiable sans prendre de gants.

Dégoût : Pendant les vacances d'un ami, j'ai accepté d'entretenir son aquarium en venant changer l'eau et nettoyer les vitres une fois par semaine. Lorsque je suis arrivé la première semaine, un des poissons était mort (apparemment depuis plusieurs jours) et en train de se décomposer dans l'eau. Les autres poissons avaient commencé à le manger. J'ai dû sortir ce poisson à moitié décomposé de l'aquarium pour aller le jeter dans les toilettes. Je n'ai pas réussi à trouver une épuisette. J'ai donc dû plonger ma main dans l'aquarium pour sortir le poisson mort. L'odeur associée à la putréfaction était très forte. Je suis encore totalement écoeuré lorsque je repense à cette expérience aujourd'hui.

Joie : Lorsque j'étais à l'école, j'avais un ami dont j'étais très proche. Nous avons grandi ensemble et étions comme frère et soeur. A l'âge adulte, il a déménagé en Australie et nous sommes restés en contact par courrier et par téléphone. Mais nous ne nous sommes pas revus depuis plus de 5 ans. Je viens d'apprendre qu'il va venir passer ces vacances en Suisse. Je me réjouis énormément à l'idée de le revoir. Je l'appelle et nous faisons plein de projets pour son séjour.

Peur : Il est plus de minuit et je rentre chez moi à pied. Je suis seul pendant un moment, puis soudain je remarque qu'un homme me suit. J'entends ses pas derrière moi. J'accélère et il accélère également. Je me met à courir et il court après moi. Je sens qu'il m'agrippe par ma veste, je vois à présent qu'il tient de l'autre main un couteau à cran d'arrêt.

Surprise : En visite chez des amis, nous sommes à table quand soudain nous entendons un bruit de chute dans la pièce d'à côté. Il n'y a personne d'autre dans la maison. Nous courons voir

ce qui s'est passé et constatons qu'une étagère pleine de livres s'est effondrée sur elle-même.

Tristesse : Mon chien (ou chat) est très gravement malade. Il a déjà été opéré deux fois mais malgré cela il va de plus en plus mal. Le vétérinaire m'a expliqué ce matin qu'il n'y avait plus aucun espoir et qu'il valait mieux abréger les souffrances du pauvre animal en l'endormant. Je dois aller au cabinet du vétérinaire cet après-midi avec mon chien (chat) pour l'euthanasier.

B

Deuxième annexe

Liste des unités d'action du système de codage d'actions faciales et leurs descriptions

Numéro de l'UA	Descripion
0	Visage neutre
1	Remontée de la partie interne des sourcils
2	Remontée de la partie externe des sourcils
3	Rapprochement des coins internes des sourcils
4	Abaissement et rapprochement des sourcils
5	Ouverture entre la paupière supérieure et les sourcils
6	Remontée des joues
7	Tension de la paupière
8	Lèvres collées
9	Plissement de la peau du nez vers le haut
10	Remontée de la partie supérieure de la lèvre
11	Ouverture du nasolabial
12	Étirement du coin des lèvres
13	Étirement et rentrée des lèvres
14	Plissement externe des lèvres (fossettes)
15	Abaissement des coins externes des lèvres
16	Ouverture de la lèvre inférieure
17	Élévation du menton
18	Froncement central des lèvres
19	Sortie de la langue
20	Étirement externe des lèvres

TABLE B.1 – *Liste des unités d'action du système de codage d'actions faciales et leurs descriptions.*

Numéro de l'UA	Descripion
21	Tension du cou
22	Lèvres en "O" (protrusion)
23	Tension refermante des lèvres
24	Lèvres pressées (pincement des lèvres)
25	Ouverture de la bouche et séparation légère des lèvres
26	Ouverture de la mâchoire
27	Bâillement
28	Succion interne des lèvres
29	Poussée de la mâchoire
30	Déplacement de côté de la mâchoire
31	Serrement de la mâchoire
32	Morsure des lèvres
33	Gonflement des joues
34	Bouffée des joues
35	Aspiration des joues
36	Bombement de la langue
37	Essuyage des lèvres
38	Dilatation des nasaux
39	Compression des nasaux
40	Reniflement
41	Abaissement de la glabelle
42	Abaissement interne des sourcils
43	Yeux fermés
44	Rapprochement des sourcils
45	Clignotement de l'oeil
46	Clignement de l'oeil

TABLE B.2 – *Liste des unités d'action du système de codage d'actions faciales et leurs descriptions.*

C

Troisième annexe

Les intervalles de chevauchements entres les clusters des vecteurs latents des différentes émotions pour les modèles visuel, acoustique, audiovisuel et le modèle des durées.

A sur B		Moyenne	95% I.C.
Dégoût	Colère	0	(0, 0)
Peur	Colère	0.00025	(0, 0)
Joie	Colère	0	(0, 0)
Neutre	Colère	1e-06	(0, 0)
Tristesse	Colère	0	(0, 0)
Surprise	Colère	0.0036	(0, 0.02)
Peur	Dégoût	0	(0, 0)
Joie	Dégoût	0	(0, 0)
Neutre	Dégoût	0	(0, 0)
Tristesse	Dégoût	0	(0, 0)
Surprise	Dégoût	0	(0, 0)
Joie	Peur	0	(0, 0)
Neutre	Peur	5e-04	(0, 0.01)
Tristesse	Peur	0	(0, 0)
Surprise	Peur	0.05	(0.01, 0.11)
Neutre	Joie	0	(0, 0)
Tristesse	Joie	4e-06	(0, 0)
Surprise	Joie	0	(0, 0)
Tristesse	Neutre	0.00046	(0, 0.01)
Surprise	Neutre	1.6e-05	(0, 0)
Surprise	Tristesse	0	(0, 0)

B Surprise A		Moyenne	95% I.C.
Colère	Dégoût	0	(0, 0)
Colère	Peur	0.0046	(0, 0.02)
Colère	Joie	0	(0, 0)
Colère	Neutre	0	(0, 0)
Colère	Tristesse	0	(0, 0)
Colère	Surprise	0.012	(0, 0.04)
Dégoût	Peur	0	(0, 0)
Dégoût	Joie	0	(0, 0)
Dégoût	Neutre	8.4e-05	(0, 0)
Dégoût	Tristesse	0	(0, 0)
Dégoût	Surprise	0	(0, 0)
Peur	Joie	0	(0, 0)
Peur	Neutre	0	(0, 0)
Peur	Tristesse	0	(0, 0)
Peur	Surprise	0.026	(0, 0.07)
Joie	Neutre	0	(0, 0)
Joie	Tristesse	2.6e-05	(0, 0)
Joie	Surprise	0	(0, 0)
Neutre	Tristesse	0.25	(0.11, 0.43)
Neutre	Surprise	0.0058	(0, 0.03)
Tristesse	Surprise	0	(0, 0)

TABLE C.1 – *Les intervalles de chevauchements entres les clusters des vecteurs latents des différentes émotions pour la modalité visuelle avec* $\beta = 0.05$.

A sur B		Moyenne	95% I.C.
Dégoût	Colère	4.9	(1.44, 9.5)
Peur	Colère	9.7	(1.9, 19.5)
Joie	Colère	0.03	(0, 1.09)
Neutre	Colère	4.3	(2.8, 6.2)
Tristesse	Colère	0.078	(0.01, 0.22)
Surprise	Colère	0.41	(0.16, 0.81)
Peur	Dégoût	0.0018	(0, 0.01)
Joie	Dégoût	2.071	(0.01, 6.19)
Neutre	Dégoût	13	(9.5, 17)
Tristesse	Dégoût	0.11	(0.02, 0.28)
Surprise	Dégoût	0.49	(0.19, 0.95)
Joie	Peur	0	(0, 0)
Neutre	Peur	23	(18, 28)
Tristesse	Peur	0.83	(0.37, 1.5)
Surprise	Peur	8.6	(2.1, 12.3)
Neutre	Joie	6.8	(4.5, 9.6)
Tristesse	Joie	4.1	(2.3, 6.5)
Surprise	Joie	0.0051	(0, 0.03)
Tristesse	Neutre	6	(1.3, 10.01)
Surprise	Neutre	7.9	(1.5, 11.1)
Surprise	Tristesse	0.015	(0, 0.06)

B Surprise A		Moyenne	95% I.C.
Colère	Dégoût	1.65	(0.31, 3.1)
Colère	Peur	8.17	(1.05, 11.36)
Colère	Joie	0.017	(0, 0.06)
Colère	Neutre	0.054	(0.01, 0.12)
Colère	Tristesse	0.0038	(0, 0.02)
Colère	Surprise	5.4	(0.7, 8.1)
Dégoût	Peur	0.00015	(0, 0)
Dégoût	Joie	0.0046	(0, 0.02)
Dégoût	Neutre	0.016	(0, 0.05)
Dégoût	Tristesse	4.003	(2, 5.02)
Dégoût	Surprise	0.095	(0.02, 0.22)
Peur	Joie	0	(0, 0)
Peur	Neutre	0.8	(0.44, 1.3)
Peur	Tristesse	0.01	(0.04, 0.1)
Peur	Surprise	9.12	(1.36, 13.5)
Joie	Neutre	0.017	(0, 0.05)
Joie	Tristesse	0.51	(0.24, 0.89)
Joie	Surprise	0.0038	(0, 0.02)
Neutre	Tristesse	25	(20, 30)
Neutre	Surprise	14	(9, 18.6)
Tristesse	Surprise	0.015	(0, 0.05)

TABLE C.2 – *Les intervalles de chevauchements entres les clusters des vecteurs latents des différentes émotions pour la modalité visuelle avec* $\beta = 0.1$.

A sur B		Moyenne	95% I.C.
Dégoût	Colère	0.00053	(0, 0.01)
Peur	Colère	8.9	(6.1, 10.01)
Joie	Colère	0	(0, 0)
Neutre	Colère	0.012	(0, 0.04)
Tristesse	Colère	19.11	(14.01, 21.08)
Surprise	Colère	0.0014	(0, 0.01)
Peur	Dégoût	53	(59.99, 66.7)
Joie	Dégoût	0	(0, 0)
Neutre	Dégoût	29	(22.9, 32.5)
Tristesse	Dégoût	0.01	(0, 0.05)
Surprise	Dégoût	0	(0, 0)
Joie	Peur	0	(0, 0)
Neutre	Peur	41	(38, 48)
Tristesse	Peur	0.0094	(0, 0.04)
Surprise	Peur	0.0039	(0, 0.02)
Neutre	Joie	4	(3, 5)
Tristesse	Joie	0.15	(0.07, 0.26)
Surprise	Joie	9	(7, 11)
Tristesse	Neutre	25	(20.8, 30)
Surprise	Neutre	0	(0, 0)
Surprise	Tristesse	0	(0, 0)

B Surprise A		Moyenne	95% I.C.
Colère	Dégoût	3	(1.1, 7.6)
Colère	Peur	19	(12, 21)
Colère	Joie	0	(0, 0)
Colère	Neutre	0.00043	(0, 0.01)
Colère	Tristesse	2e-06	(0, 0)
Colère	Surprise	0.0002	(0, 0.01)
Dégoût	Peur	34	(29, 39)
Dégoût	Joie	0	(0, 0)
Dégoût	Neutre	1.75	(1.06, 2.29)
Dégoût	Tristesse	0.00035	(0, 0.01)
Dégoût	Surprise	5e-06	(0, 0)
Peur	Joie	0	(0, 0)
Peur	Neutre	10.17	(8.08, 11.29)
Peur	Tristesse	4.1e-05	(0, 0)
Peur	Surprise	0.15	(0.06, 0.27)
Joie	Neutre	9.5e-05	(0, 0)
Joie	Tristesse	2.69	(1.37, 3.1)
Joie	Surprise	11.0013	(9.89, 12.01)
Neutre	Tristesse	18	(12.65, 20.6)
Neutre	Surprise	3e-04	(0, 0.01)
Tristesse	Surprise	0.00024	(0, 0)

TABLE C.3 – *Les intervalles de chevauchements entres les clusters des vecteurs latents des différentes émotions pour la modalité acoustique avec* $\beta = 5 \times 10e^{-3}$.

A sur B		Moyenne	95% I.C.
Dégoût	Colère	0.0005	(0, 0.01)
Peur	Colère	48	(43, 54)
Joie	Colère	60	(55, 65)
Neutre	Colère	54	(49, 58)
Tristesse	Colère	41	(36, 47)
Surprise	Colère	33	(29, 38)
Peur	Dégoût	0.0042	(0, 0.02)
Joie	Dégoût	0.00075	(0, 0.01)
Neutre	Dégoût	0.004	(0, 0.02)
Tristesse	Dégoût	0.00076	(0, 0.01)
Surprise	Dégoût	4e-06	(0, 0)
Joie	Peur	49	(44, 54)
Neutre	Peur	57	(52, 61)
Tristesse	Peur	49	(44, 54)
Surprise	Peur	9.2	(7.3, 11)
Neutre	Joie	62	(58, 66)
Tristesse	Joie	51	(46, 56)
Surprise	Joie	21	(18, 25)
Tristesse	Neutre	46	(42, 50)
Surprise	Neutre	12	(9.8, 14)
Surprise	Tristesse	16	(13, 20)

B Surprise A		Moyenne	95% I.C.
Colère	Dégoût	8.9e-05	(0, 0)
Colère	Peur	26	(22, 30)
Colère	Joie	47	(42, 53)
Colère	Neutre	31	(28, 35)
Colère	Tristesse	39	(34, 44)
Colère	Surprise	66	(61, 70)
Dégoût	Peur	6.86	(2.65, 9.98)
Dégoût	Joie	0.0003	(0, 0.001)
Dégoût	Neutre	2	(1, 3)
Dégoût	Tristesse	2.65	(1, 5.01)
Dégoût	Surprise	0.00001	(0, 0.01)
Peur	Joie	59	(54, 64)
Peur	Neutre	59	(55, 63)
Peur	Tristesse	65	(61, 70)
Peur	Surprise	57	(51, 62)
Joie	Neutre	55	(50, 59)
Joie	Tristesse	64	(59, 69)
Joie	Surprise	67	(62, 72)
Neutre	Tristesse	67	(63, 71)
Neutre	Surprise	58	(54, 63)
Tristesse	Surprise	53	(47, 58)

TABLE C.4 – *Les intervalles de chevauchements entres les clusters des vecteurs latents des différentes émotions pour le modèle des durées avec* $\beta = 2 \times 10e^{-5}$.

A sur B		Moyenne	95% I.C.
Dégoût	Colère	6.7	(4.8, 8.9)
Peur	Colère	48	(43, 54)
Joie	Colère	60	(55, 65)
Neutre	Colère	54	(49, 58)
Tristesse	Colère	41	(36, 47)
Surprise	Colère	33	(29, 38)
Peur	Dégoût	0.004	(0, 0.02)
Joie	Dégoût	0.00074	(0, 0.01)
Neutre	Dégoût	0.004	(0, 0.02)
Tristesse	Dégoût	0.00074	(0, 0.01)
Surprise	Dégoût	0	(0, 0)
Joie	Peur	49	(44, 54)
Neutre	Peur	57	(52, 61)
Tristesse	Peur	49	(44, 54)
Surprise	Peur	9.2	(7.3, 11)
Neutre	Joie	62	(58, 66)
Tristesse	Joie	51	(46, 56)
Surprise	Joie	21	(18, 25)
Tristesse	Neutre	46	(42, 50)
Surprise	Neutre	12	(9.8, 14)
Surprise	Tristesse	16	(14, 20)

B Surprise A		Moyenne	95% I.C.
Colère	Dégoût	1e-04	(0, 0)
Colère	Peur	26	(22, 30)
Colère	Joie	47	(42, 53)
Colère	Neutre	31	(28, 35)
Colère	Tristesse	39	(34, 44)
Colère	Surprise	65	(61, 70)
Dégoût	Peur	30	(25, 35)
Dégoût	Joie	16	(13, 20)
Dégoût	Neutre	28	(23, 32)
Dégoût	Tristesse	28	(23, 33)
Dégoût	Surprise	4.1	(2.6, 5.9)
Peur	Joie	59	(54, 64)
Peur	Neutre	59	(55, 63)
Peur	Tristesse	66	(61, 71)
Peur	Surprise	57	(51, 62)
Joie	Neutre	55	(51, 59)
Joie	Tristesse	64	(59, 69)
Joie	Surprise	66	(61, 71)
Neutre	Tristesse	67	(63, 71)
Neutre	Surprise	58	(54, 63)
Tristesse	Surprise	53	(47, 58)

TABLE C.5 – *Les intervalles de chevauchements entres les clusters des vecteurs latents des différentes émotions pour le modèle audiovisuel avec* $\beta = 0$.

Bibliographie

Ossama Abdel-Hamid and Hui Jiang. Fast speaker adaptation of hybrid nn/hmm model for speech recognition based on discriminative learning of speaker code. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 7942–7946. IEEE, 2013.

Christian Abry and MT Lallouache. Audibility and stability of articulatory movements : Deciphering two experiments on anticipatory rounding in french. In *Proceedings of the XIIth International Congress of Phonetic Sciences*, volume 1, pages 220–225, 1991.

Kei Akuzawa, Yusuke Iwasawa, and Yutaka Matsuo. Expressive speech synthesis via modeling expressions with variational autoencoder. In *Proc. Interspeech 2018*, pages 3067–3071, 2018. doi : 10.21437/Interspeech.2018-1113. URL `http://dx.doi.org/10.21437/Interspeech.2018-1113`.

Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. *ICLR17*, 2017.

Shumin An, Zhenhua Ling, and Lirong Dai. Emotional statistical parametric speech synthesis using lstm-rnns. In *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1613–1616. IEEE, 2017.

Tobias S Andersen. The mcgurk illusion in the oddity task. In *Auditory-Visual Speech Processing 2010*, 2010.

Robert Anderson, Bjorn Stenger, Vincent Wan, and Roberto Cipolla. Expressive visual text-to-speech using active appearance models. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3382–3389, 2013.

Lyubomir Y Antonov. Method of and device for synthesis of speech from printed text, July 14 1981. US Patent 4,278,838.

Andrei Atanov, Alexandra Volokhova, Arsenii Ashukha, Ivan Sosnovik, and Dmitry Vetrov. Semi-conditional normalizing flows for semi-supervised learning. *arXiv preprint arXiv :1905.00505*, 2019.

G Bailly, G Gibert, and M Odisio. Evaluation of movement generation systems using the point-light technique. In *Speech Synthesis, 2002. Proceedings of 2002 IEEE Workshop on*, pages 27–30. IEEE, 2002.

Gérard Bailly, Antoine Bégault, Frédéric Elisei, and Pierre Badin. Speaking with smile or disgust : data and models. 2008.

Tanja Bänziger and Klaus R Scherer. Introducing the geneva multimodal emotion portrayal (gemep) corpus. *Blueprint for affective computing : A sourcebook*, pages 271–294, 2010.

Nelly Barbot, Olivier Boëffard, Jonathan Chevelu, and Arnaud Delhay. Large linguistic corpus reduction with scp algorithms. *Computational Linguistics*, 41(3) :355–383, 2015.

John N Bassili. Emotion recognition : the role of facial movement and the relative importance of upper and lower areas of the face. *Journal of personality and social psychology*, 37(11) :2049, 1979.

Joseph Bates et al. The role of emotion in believable agents. *Communications of the ACM*, 37 (7) :122–125, 1994.

Helen L Bear and Richard Harvey. Phoneme-to-viseme mappings : the good, the bad, and the ugly. *Speech Communication*, 95 :40–67, 2017.

Fredericka Bell-Berti and Katherine S Harris. Anticipatory coarticulation : Some implications from a study of lip rounding. *The Journal of the Acoustical Society of America*, 65(5) :1268–1270, 1979.

Fredericka Bell-Berti and Katherine S Harris. Temporal patterns of coarticulation : Lip rounding. *The Journal of the Acoustical Society of America*, 71(2) :449–454, 1982.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb) :1137–1155, 2003.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning : A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8) : 1798–1828, 2013.

Christian Benoit, Tahar Lallouache, Tayeb Mohamadi, and Christian Abry. A set of french visemes for visual speech synthesis, 1992.

Jeffrey J Berry. Accuracy of the ndi wave speech research system. *Journal of Speech, Language, and Hearing Research*, 54(5) :1295–1301, 2011.

Jonas Beskow. Rule-based visual speech synthesis. In *Fourth European Conference on Speech Communication and Technology*, 1995.

Jonas Beskow. On talking heads, social robots and what they can teach us. In *International Congress of Phonetic Sciences ICPhS 2019*, 2019.

Jonas Beskow and Mikael Nordenberg. Data-driven synthesis of expressive visual speech using an mpeg-4 talking head. In *Ninth European Conference on Speech Communication and Technology*, 2005.

Mark Beutnagel, Alistair Conkie, Juergen Schroeter, Yannis Stylianou, and Ann Syrdal. The at&t next-gen tts system. In *Joint meeting of ASA, EAA, and DAGA*, pages 18–24. Citeseer, 1999.

Elisabetta Bevacqua and Catherine Pelachaud. Expressive audio-visual speech. *Computer Animation and Virtual Worlds*, 15(3-4) :297–304, 2004.

Purnima Bholowalia and Arvind Kumar. Ebk-means : A clustering technique based on elbow method and k-means in wsn. *International Journal of Computer Applications*, 105(9), 2014.

A Black. Chatr, version 0.8, a generic speech synthesizer. *System documentation, ATR-Interpreting Telecomunications Laboratories, Kyoto, Japan*, 1996.

Alan W Black and Paul A Taylor. Automatically clustering similar units for unit selection in speech synthesis. 1997.

Alan W Black, Christina L Bennett, Benjamin C Blanchard, John Kominek, Brian Langner, Kishore Prahallad, and Arthur Toth. Cmu blizzard 2007 : A hybrid acoustic unit selection system from statistically predicted parameters. In *Blizzard Challenge Workshop, Bonn, Germany*, 2007.

Alan W Black, Simon King, and Keiichi Tokuda. The blizzard challenge 2010. 2010.

Paul Boersma et al. Praat, a system for doing phonetics by computer. *Glot international*, 5, 2002.

Dwight Bolinger. Intonation across languages. *Universals of human language*, 2 :471–524, 1978.

Dwight Bolinger and Dwight Le Merton Bolinger. *Intonation and its uses : Melody in grammar and discourse*. Stanford university press, 1989.

Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew Dai, Rafal Jozefowicz, and Samy Bengio. Generating sentences from a continuous space. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 10–21, Berlin, Germany, August 2016. Association for Computational Linguistics. doi : 10.18653/v1/K16-1002. URL `https://www.aclweb.org/anthology/K16-1002`.

Baris Bozkurt, Ozlem Ozturk, and Thierry Dutoit. Text design for tts speech corpus building using a modified greedy selection. In *Eighth European Conference on Speech Communication and Technology*, 2003.

Ronald Newbold Bracewell and Ronald N Bracewell. *The Fourier transform and its applications*, volume 31999. McGraw-Hill New York, 1986.

Matthew Brand. Voice puppetry. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 21–28. ACM Press/Addison-Wesley Publishing Co., 1999.

AP Breen, E Bowers, and W Welsh. An investigation into the generation of mouth shapes for a talking head. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*, volume 4, pages 2159–2162. IEEE, 1996.

Christoph Bregler, Michele Covell, and Malcolm Slaney. Video rewrite : driving visual speech with audio. In *Siggraph*, volume 97, pages 353–360, 1997.

NM Brooke and SD Scott. Computer graphics animations of talking faces based on stochastic models. In *Proceedings of ICSIPNN'94. International Conference on Speech, Image Processing and Neural Networks*, pages 73–76. IEEE, 1994.

The Duy Bui, Dirk Heylen, and Anton Nijholt. Combination of facial movements on a 3d talking head. In *Proceedings Computer Graphics International, 2004.*, pages 284–290. IEEE, 2004.

Felix Burkhardt and Walter F Sendlmeier. Verification of acoustical correlates of emotional speech using formant-synthesis. In *ISCA Tutorial and Research Workshop (ITRW) on speech and emotion*, 2000.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. Iemocap : Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4) :335, 2008.

Janet E Cahn. The generation of affect in synthesized speech. *Journal of the American Voice I/O Society*, 8(1) :1–1, 1990.

Hüseyin Cakmak, Jérôme Urbain, Thierry Dutoit, and Joëlle Tilmanne. The av-lasyn database : A synchronous corpus of audio and 3d facial marker data for audio-visual laughter synthesis. In *LREC*, pages 3398–3403, 2014.

La Calliope and G Fant. *La parole et son traitement automatique*. Masson Paris, 1989.

Yong Cao, Petros Faloutsos, Eddie Kohler, and Frédéric Pighin. Real-time speech motion synthesis from recorded motions. In *Proceedings of the 2004 ACM SIGGRAPH/Eurographics symposium on Computer animation*, pages 345–353. Eurographics Association, 2004.

Yong Cao, Wen C Tien, Petros Faloutsos, and Frédéric Pighin. Expressive speech-driven facial animation. *ACM Transactions on Graphics (TOG)*, 24(4) :1283–1302, 2005.

Justine Cassell, Joseph Sullivan, Elizabeth Churchill, and Scott Prevost. *Embodied conversational agents*. MIT press, 2000.

Milos Cernak, Petr Motlicek, and Philip N Garner. On the (un) importance of the contextual factors in hmm-based speech synthesis and coding. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8140–8143. IEEE, 2013.

Ff Charpentier and M Stella. Diphone synthesis using an overlap-add technique for speech waveforms concatenation. In *ICASSP'86. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 11, pages 2015–2018. IEEE, 1986.

Louise Chavarie, Kimberly L Howland, Les N Harris, Mike J Hansen, Colin P Gallagher, William J Harford, William M Tonn, Andrew M Muir, and Charles C Krueger. Habitat overlap of juvenile and adult lake trout of great bear lake : Evidence for lack of a predation gradient ? *Ecology of Freshwater Fish*, 28(3) :485–498, 2019.

Stanley F Chen and Joshua Goodman. An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*, 13(4) :359–394, 1999.

Jonathan Chevelu and Damien Lolive. Do not build your tts training corpus randomly. In *2015 23rd European Signal Processing Conference (EUSIPCO)*, pages 350–354. IEEE, 2015.

Claude C Chibelushi, John S Mason, and R Deravi. Integration of acoustic and visual speech for speaker recognition. In *Third European Conference on Speech Communication and Technology*, 1993.

Byoungwon Choe, Hanook Lee, and Hyeong-Seok Ko. Performance-driven muscle-based facial animation. *The Journal of Visualization and Computer Animation*, 12(2) :67–79, 2001.

Erika Chuang and Chris Bregler. Performance driven facial animation using blendshape interpolation. *Computer Science Technical Report, Stanford University*, 2(2) :3, 2002.

Yu-An Chung, Yuxuan Wang, Wei-Ning Hsu, Yu Zhang, and RJ Skerry-Ryan. Semi-supervised training for improving data efficiency in end-to-end speech synthesis. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6940–6944. IEEE, 2019.

Michael M Cohen and Dominic W Massaro. Modeling coarticulation in synthetic visual speech. In *Models and techniques in computer animation*, pages 139–156. Springer, 1993.

Vincent Colotte and Yves Laprie. Amélioration de la précision de la resynthèse avec td-psola. 2002.

Timothy F Cootes, Gareth J Edwards, and Christopher J Taylor. Active appearance models. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (6) :681–685, 2001.

Piero Cosi, Emanuela Magno Caldognetto, Giulio Perin, and Claudio Zmarich. Labial coarticulation modeling for realistic facial animation. In *Proceedings. Fourth IEEE International Conference on Multimodal Interfaces*, pages 505–510. IEEE, 2002.

Erica Costantini, Fabio Pianesi, and Michela Prete. Recognising emotions in human and synthetic faces : the role of the upper and lower parts of the face. In *Proceedings of the 10th international conference on Intelligent user interfaces*, pages 20–27. ACM, 2005.

Steven Davis and Paul Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE transactions on acoustics, speech, and signal processing*, 28(4) :357–366, 1980.

JR Davitz. Auditory correlates of vocal expressions of emotional meanins. u : Jr davitz (ed.)-the communication of emotional meaning, 1964.

Celso M de Melo, Peter Carnevale, and Jonathan Gratch. The effect of virtual agents' emotion displays and appraisals on people's decision making in negotiation. In *International Conference on Intelligent Virtual Agents*, pages 53–66. Springer, 2012.

Donald M Decker et al. *Handbook of the International Phonetic Association : A guide to the use of the International Phonetic Alphabet*. Cambridge University Press, 1999.

Doris M Dehn and Susanne Van Mulken. The impact of animated interface agents : a review of empirical research. *International journal of human-computer studies*, 52(1) :1–22, 2000.

Li Deng. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE Signal Processing Magazine*, 29(6) :141–142, 2012.

Emily L Denton, Soumith Chintala, Rob Fergus, et al. Deep generative image models using a laplacian pyramid of adversarial networks. In *Advances in neural information processing systems*, pages 1486–1494, 2015.

Yu Ding and Catherine Pelachaud. Lip animation synthesis : a unified framework for speaking and laughing virtual agent. In *AVSP*, pages 78–83, 2015.

Carl Doersch. Tutorial on variational autoencoders. *arXiv preprint arXiv :1606.05908*, 2016.

Gianluca Donato, Marian Stewart Bartlett, Joseph C. Hager, Paul Ekman, and Terrence J. Sejnowski. Classifying facial actions. *IEEE Transactions on pattern analysis and machine intelligence*, 21(10) :974–989, 1999.

Robert E Donovan and Phil C Woodland. Improvements in an hmm-based speech synthesiser. In *Fourth European Conference on Speech Communication and Technology*, 1995.

Homer Dudley. Remaking speech. *The Journal of the Acoustical Society of America*, 11(2) : 169–177, 1939.

Stéphane Dupont and Juergen Luettin. Audio-visual speech modeling for continuous speech recognition. *IEEE transactions on multimedia*, 2(3) :141–151, 2000.

Thierry Dutoit. *An introduction to text-to-speech synthesis*, volume 3. Springer Science & Business Media, 1997.

Thierry Dutoit, Vincent Pagel, Nicolas Pierret, François Bataille, and Olivier Van der Vrecken. The mbrola project : Towards a set of high quality speech synthesizers free of use for non commercial purposes. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*, volume 3, pages 1393–1396. IEEE, 1996.

Mark Dworkin, Apurba Chakraborty, Sangyoon Lee, Colleen Monahan, Lisa Hightow-Weidman, Robert Garofalo, Dima Qato, and Antonio Jimenez. A realistic talking human embodied agent mobile phone intervention to promote hiv medication adherence and retention in care in young hiv-positive african american men who have sex with men : qualitative study. *JMIR mHealth and uHealth*, 6(7) :e10211, 2018.

Mike Edgington. Investigating the limitations of concatenative synthesis. In *Fifth European Conference on Speech Communication and Technology*, 1997.

P Ekman, W Friesen, and J Hager. Facial action coding system : Research nexus. *Network Research Information, Salt Lake City, UT*, 1, 2002.

Paul Ekman. About brows : Emotional and conversational signals. In *Human Ethology : Claims and Limits of a New Dicipline : Contributions to the Colloquium, Cranach*. Cambridge University Press, 1979.

Paul Ekman. An argument for basic emotions. *Cognition & emotion*, 6(3-4) :169–200, 1992.

Paul Ekman and Daniel Cordaro. What is meant by calling emotions basic. *Emotion review*, 3 (4) :364–370, 2011.

Paul Ekman and Wallace V Friesen. Constants across cultures in the face and emotion. *Journal of personality and social psychology*, 17(2) :124, 1971.

Paul Ekman and Wallace V Friesen. Measuring facial movement. *Environmental psychology and nonverbal behavior*, 1(1) :56–75, 1976.

Paul Ekman and Wallace V Friesen. *Facial action coding system : Investigator's guide*. Consulting Psychologists Press, 1978.

Paul Ekman and Wallace V Friesen. A new pan-cultural facial expression of emotion. *Motivation and emotion*, 10(2) :159–168, 1986.

Kevin El Haddad, Hüseyin Cakmak, Alexis Moinet, Stéphane Dupont, and Thierry Dutoit. An hmm approach for synthesizing amused speech with a controllable intensity of smile. In *2015 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, pages 7–11. IEEE, 2015.

Saher Esmeir and Shaul Markovitch. Anytime learning of decision trees. *Journal of Machine Learning Research*, 8(May) :891–933, 2007.

Gary Faigin. *The artist's complete guide to facial expression.* Watson-Guptill, 2012.

Caroline J Falconer, E Bethan Davies, Rebecca Grist, and Paul Stallard. Innovations in practice : Avatar-based virtual reality in camhs talking therapy : two exploratory case studies. *Child and Adolescent Mental Health*, 24(3) :283–287, 2019.

Bo Fan, Lijuan Wang, Frank K Soong, and Lei Xie. Photo-real talking head with deep bidirectional lstm. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4884–4888. IEEE, 2015.

Bo Fan, Lei Xie, Shan Yang, Lijuan Wang, and Frank K Soong. A deep bidirectional lstm approach for video-realistic talking head. *Multimedia Tools and Applications*, 75(9) :5287–5309, 2016.

Yuchen Fan, Yao Qian, Feng-Long Xie, and Frank K Soong. Tts synthesis with bidirectional lstm based recurrent neural networks. In *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

Gabriele Fanelli, Juergen Gall, Harald Romsdorfer, Thibaut Weise, and Luc Van Gool. Acquisition of a 3d audio-visual corpus of affective speech. *IEEE Transactions on Multimedia*, 12(6) : 591–598, 2010.

Gunnar Fant. Speech communication research. *Royal Swedish Academy of Engineering Sciences*, 2 :331–337, 1953.

Panagiotis Paraskevas Filntisis, Athanasios Katsamanis, Pirros Tsiakoulis, and Petros Maragos. Video-realistic expressive audio-visual speech synthesis for the greek language. *Speech Communication*, 95 :137–152, 2017.

Ivan Fónagy. La mimique buccale. *Phonetica*, 33(1) :31–44, 1976.

Manuel Fonseca De Sam Bento Ribeiro. Suprasegmental representations for the modeling of fundamental frequency in statistical parametric speech synthesis. 2018.

Hélene François and Olivier Boëffard. Design of an optimal continuous speech database for text-to-speech synthesis considered as a set covering problem. In *Seventh European Conference on Speech Communication and Technology*, 2001.

Nico H Frijda and Batja Mesquita. The social roles and functions of emotions. 1994.

Nico H Frijda, Batja Mesquita, Joep Sonnemans, and Stephanie Van Goozen. The duration of affective phenomena or emotions, sentiments and passions. 1991.

Yun Fu, Renxiang Li, Thomas S Huang, and Mike Danielsen. Real-time humanoid avatar for multimodal human-machine interaction. In *2007 IEEE International Conference on Multimedia and Expo*, pages 991–994. IEEE, 2007.

Yun Fu, Renxiang Li, Thomas S Huang, and Mike Danielsen. Real-time multimodal human–avatar interaction. *IEEE Transactions on Circuits and Systems for Video Technology*, 18(4) : 467–477, 2008.

Osamu Fujimura. An approximation to voice aperiodicity. *IEEE transactions on Audio and Electroacoustics*, 16(1) :68–72, 1968.

Mark JF Gales. Cluster adaptive training of hidden markov models. *IEEE transactions on speech and audio processing*, 8(4) :417–428, 2000.

Peter Gärdenfors. How homo became sapiens : On the evolution of thinking. 2006.

Laurianne Georgeton, Nikola Paillereau, Simon Landron, Jiayin Gao, and Takeki Kamiyama. Analyse formantique des voyelles orales du français en contexte isolé : à la recherche d'une référence pour les apprenants de fle. 2012.

Andrew Gibiansky, Sercan Arik, Gregory Diamos, John Miller, Kainan Peng, Wei Ping, Jonathan Raiman, and Yanqi Zhou. Deep voice 2 : Multi-speaker neural text-to-speech. In *Advances in neural information processing systems*, pages 2962–2970, 2017.

Bertrand Le Goff. Automatic modeling of coarticulation in text-to-visual speech synthesis. In *Fifth European Conference on Speech Communication and Technology*, 1997.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

O Govokhina. *Modèles de génération de trajectoires pour l'animation de visages parlants*. PhD thesis, Thèse de l'Institut National Polytechnique de Grenoble, 2008.

Oxana Govokhina, Gérard Bailly, Gaspard Breton, and Paul Bagshaw. Evaluation de systèmes de génération de mouvements faciaux. 2006.

Björn Granström, David House, and Magnus Lundeberg. Prosodic cues in multimodal speech perception. In *Proceedings of the International Congress of Phonetic Sciences (ICPhS99)*, pages 655–658, 1999.

David Guennec. *Study of unit selection text-to-speech synthesis algorithms*. PhD thesis, 2016.

David Gunning. Explainable artificial intelligence (xai). *Defense Advanced Research Projects Agency (DARPA), nd Web*, 2 :2, 2017.

A. Hallgren and B. Lyberg. Visual speech synthesis with concatenative speech. In *AVSP*, Terrigal-Sydney, Australia, 1998.

Åsa Hällgren and Bertil Lyberg. *Facial animation using visual polyphones*. PhD thesis, ASA, 1997.

Kei Hashimoto, Keiichiro Oura, Yoshihiko Nankaku, and Keiichi Tokuda. The effect of neural networks in statistical parametric speech synthesis. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4455–4459. IEEE, 2015.

Gustav Eje Henter, Srikanth Ronanki, Oliver Watts, Mirjam Wester, Zhizheng Wu, and Simon King. Robust tts duration modelling using dnns. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5130–5134. IEEE, 2016.

Gustav Eje Henter, Jaime Lorenzo-Trueba, Xin Wang, and Junichi Yamagishi. Principles for learning controllable tts from annotated and latent variation. In *INTERSPEECH*, pages 3956–3960, 2017.

Gustav Eje Henter, Jaime Lorenzo-Trueba, Xin Wang, and Junichi Yamagishi. Deep encoder-decoder models for unsupervised learning of controllable speech synthesis. *arXiv preprint arXiv :1807.11470*, 2018.

Caroline Henton and Peter Litwinowicz. Saying and seeing it with feeling : techniques for synthesizing visible, emotional speech. In *SSW*, pages 73–76. Citeseer, 1994.

Barbara Heuft, Thomas Portele, and Monika Rauth. Emotions in time domain synthesis. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*, volume 3, pages 1974–1977. IEEE, 1996.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae : Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017.

Toshio Hirai and Seiichi Tenpaku. Using 5 ms segments in concatenative speech synthesis. In *Fifth ISCA workshop on speech synthesis*, 2004.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8) : 1735–1780, 1997.

Gregor O Hofer, Korin Richmond, and Robert AJ Clark. Informed blending of databases for emotional speech synthesis. In *Interspeech*, pages 501–504. International Speech Communication Association, 2005.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70, 1979.

John N Holmes, Ignatius G Mattingly, and John N Shearme. Speech synthesis by rule. *Language and speech*, 7(3) :127–143, 1964.

Andreas Holzinger, Chris Biemann, Constantinos S Pattichis, and Douglas B Kell. What do we need to build explainable ai systems for the medical domain ? *arXiv preprint arXiv :1712.09923*, 2017.

Pengyu Hong, Zhen Wen, and Thomas S Huang. Real-time speech-driven face animation with expressions using neural networks. *IEEE Transactions on neural networks*, 13(4) :916–927, 2002.

David House, Jonas Beskow, and Björn Granström. Timing and interaction of visual cues for prominence in audiovisual speech perception. In *Seventh European Conference on Speech Communication and Technology*, 2001.

Zhiying Huang, Jian Tang, Shaofei Xue, and Lirong Dai. Speaker adaptation of rnn-blstm for speech recognition based on speaker code. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5305–5309. IEEE, 2016.

Akemi Iida and Nick Campbell. Speech database design for a concatenative text-to-speech synthesis system for individuals with communication disorders. *International Journal of Speech Technology*, 6(4) :379–392, 2003.

Fumitada Itakura. Line spectrum representation of linear predictor coefficients of speech signals. *The Journal of the Acoustical Society of America*, 57(S1) :S35–S35, 1975.

Michelle C Jackson, Darragh J Woodford, Terence A Bellingan, Olaf LF Weyl, Michael J Potgieter, Nick A Rivers-Moore, Bruce R Ellender, Hermina E Fourie, and Christian T Chimimba. Trophic overlap between fish and riparian spiders : potential impacts of an invasive fish on terrestrial consumers. *Ecology and evolution*, 6(6) :1745–1752, 2016.

Jia Jia, Shen Zhang, Fanbo Meng, Yongxin Wang, and Lianhong Cai. Emotional audio-visual speech synthesis based on pad. *IEEE transactions on audio, speech, and language processing*, 19(3) :570–582, 2010.

W Lewis Johnson, Shrikanth Narayanan, Richard Whitney, Rajat Das, Murtaza Bulut, and Catherine LaBore. Limited domain synthesis of expressive military speech for animated characters. In *Proceedings of 2002 IEEE Workshop on Speech Synthesis, 2002.*, pages 163–166. IEEE, 2002.

Prem Kalra, Angelo Mangili, Nadia Magnenat Thalmann, and Daniel Thalmann. Simulation of facial muscle actions based on rational free form deformations. In *Computer Graphics Forum*, volume 11, pages 59–69. Wiley Online Library, 1992.

Jari Kätsyri, Vasily Klucharev, Michael Frydrych, and Mikko Sams. Identification of synthetic and natural emotional facial expressions. In *AVSP 2003-International Conference on Audio-Visual Speech Processing*, 2003.

Hideki Kawahara, Ikuyo Masuda-Katsuse, and Alain De Cheveigne. Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based f0 extraction : Possible role of a repetitive structure in sounds. *Speech communication*, 27(3-4) :187–207, 1999.

Kyosuke Kazumi, Yoshihiko Nankaku, and Keiichi Tokuda. Factor analyzed voice models for hmm-based speech synthesis. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4234–4237. IEEE, 2010.

Dacher Keltner and Jonathan Haidt. Social functions of emotions at four levels of analysis. *Cognition & Emotion*, 13(5) :505–521, 1999.

Dacher Keltner and Ann M Kring. Emotion, social function, and psychopathology. *Review of General Psychology*, 2(3) :320–342, 1998.

Farkas Kempelen. *Mechanismus der menschlichen Sprache, nebst der Beschreibung seiner sprechenden Maschine*. Degen, 1791.

John F Kihlstrom and Nancy Cantor. Social intelligence. 2000.

Simon King and Vasilis Karaiskos. The blizzard challenge 2011. In *Proc. Blizzard Challenge*, volume 2011, pages 1–10, 2011.

Simon King and Vasilis Karaiskos. The blizzard challenge 2012. In *Proc. Blizzard Challenge*, 2012.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *ICLR 2014*, 2013.

S Prahallad Kishore and Alan W Black. Unit size in unit selection speech synthesis. In *Eighth European Conference on Speech Communication and Technology*, 2003.

Dennis H Klatt. Software for a cascade/parallel formant synthesizer. *the Journal of the Acoustical Society of America*, 67(3) :971–995, 1980.

Viacheslav Klimkov, Alexis Moinet, Adam Nadolski, and Thomas Drugman. Parameter generation algorithms for text-to-speech synthesis with recurrent neural networks. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 626–631. IEEE, 2018.

Alexei Kochetov and Chris Neufeld. Examining the extent of anticipatory coronal coarticulation : A long-term average spectrum analysis. In *Proceedings of Meetings on Acoustics ICA2013*, volume 19, page 060300. ASA, 2013.

John Kominek and A Black. The blizzard challenge 2006 cmu entry introducing hybrid trajectory-selection synthesis. In *Blizzard Challenge Workshop*, 2006.

Sumedha Kshirsagar, Tom Molet, and Nadia Magnenat-Thalmann. Principal components of expressive speech animation. In *Proceedings. Computer Graphics International 2001*, pages 38–44. IEEE, 2001.

R Kubichek. Mel-cepstral distance measure for objective speech quality assessment. In *Proceedings of IEEE Pacific Rim Conference on Communications Computers and Signal Processing*, volume 1, pages 125–128. IEEE, 1993.

Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1) :79–86, 1951.

Paul Lamere, Philip Kwok, Evandro Gouvea, Bhiksha Raj, Rita Singh, William Walker, Manfred Warmuth, and Peter Wolf. The cmu sphinx-4 speech recognition system. In *IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP 2003), Hong Kong*, volume 1, pages 2–5, 2003.

RJ Laresn and E Diener. Promises and problems with the circumplex model of emotion. *Review of Personality and social psychology : Emotion*, 13 :25–29, 1992.

Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther. Autoencoding beyond pixels using a learned similarity metric. In *International conference on machine learning*, pages 1558–1566, 2016.

Bertrand Le Goff, Thierry Guiard-Marigny, Michael M Cohen, and Christian Benoît. Real-time analysis-synthesis and intelligibility of talking faces. In *SSW*, pages 53–56, 1994.

Hao Li, Jihun Yu, Yuting Ye, and Chris Bregler. Realtime facial animation with on-the-fly correctives. *ACM Trans. Graph.*, 32(4) :42–1, 2013.

Xu Li, Zhiyong Wu, Helen M Meng, Jia Jia, Xiaoyan Lou, and Lianhong Cai. Expressive speech driven talking avatar synthesis with dblstm using limited amount of emotional bimodal data. In *Interspeech*, pages 1477–1481, 2016a.

Xu Li, Zhiyong Wu, Helen M Meng, Jia Jia, Xiaoyan Lou, and Lianhong Cai. Phoneme embedding and its application to speech driven talking avatar synthesis. In *INTERSPEECH*, pages 1472–1476, 2016b.

Anders Löfqvist. Speech as audible gestures. pages 289–322, 1990.

Heng Lu, Simon King, and Oliver Watts. Combining a vector space representation of linguistic context with a deep neural network for text-to-speech synthesis. In *Eighth ISCA Workshop on Speech Synthesis*, 2013.

Patrick Lucey, Jeffrey F Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. The extended cohn-kanade dataset (ck+) : A complete dataset for action unit and emotion-specified expression. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*, pages 94–101. IEEE, 2010.

Hieu-Thi Luong, Shinji Takaki, Gustav Eje Henter, and Junichi Yamagishi. Adapting and controlling dnn-based speech synthesis using input codes. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4905–4909. IEEE, 2017.

Shuang Ma, Daniel Mcduff, and Yale Song. Neural tts stylization with adversarial and collaborative games. In *International Conference on Learning Representations*, 2019.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of machine learning research*, 9(Nov) :2579–2605, 2008.

Sébastien Le Maguer, Nelly Barbot, and Olivier Boeffard. Evaluation of contextual descriptors for hmm-based speech synthesis in french. In *Eighth ISCA Workshop on Speech Synthesis*, 2013.

Brian Mak and Etienne Barnard. Phone clustering using the bhattacharyya distance. In *Fourth International Conference on Spoken Language Processing*, 1996.

Kohki Mametani, Tsuneo Kato, and Seiichi Yamamoto. Investigating context features hidden in end-to-end tts. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6920–6924. IEEE, 2019.

Maurizio Mancini, Beatrice Biancardi, Florian Pecune, Giovanna Varni, Yu Ding, Catherine Pelachaud, Gualtiero Volpe, and Antonio Camurri. Implementing and evaluating a laughing virtual character. *ACM Transactions on Internet Technology (TOIT)*, 17(1) :1–22, 2017.

Jean-Claude Martin, Radoslaw Niewiadomski, Laurence Devillers, Stéphanie Buisine, and Catherine Pelachaud. Multimodal complex emotions : Gesture expressivity and blended facial expressions. *International Journal of Humanoid Robotics*, 3(03) :269–291, 2006.

Dominic W Massaro and Michael M Cohen. Perception of synthesized audible and visible speech. *Psychological Science*, 1(1) :55–63, 1990.

Takashi Masuko, Keiichi Tokuda, Takao Kobayashi, and Satoshi Imai. Speech synthesis using hmms with dynamic features. In *1996 ieee international conference on acoustics, speech, and signal processing conference proceedings*, volume 1, pages 389–392. IEEE, 1996.

Takashi Masuko, Takao Kobayashi, Masatsune Tamura, Jun Masubuchi, and Keiichi Tokuda. Text-to-visual speech synthesis based on parameter generation from hmm. In *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98 (Cat. No. 98CH36181)*, volume 6, pages 3745–3748. IEEE, 1998.

Takashi Masuko, Takao Kobayashi, and Keisuke Miyanaga. A style control technique for hmm-based speech synthesis. In *Eighth International Conference on Spoken Language Processing*, 2004.

Harry McGurk and John MacDonald. Hearing lips and seeing voices. *Nature*, 264(5588) :746, 1976.

Leland McInnes, John Healy, and James Melville. Umap : Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv :1802.03426*, 2018.

DG McNicholl, GK Davoren, AR Majewski, and JD Reist. Isotopic niche overlap between co-occurring capelin (mallotus villosus) and polar cod (boreogadus saida) and the effect of lipid extraction on stable isotope ratios. *Polar Biology*, 41(3) :423–432, 2018.

Albert Mehrabian. Communication without words. *Psychology today*, 2(4), 1968.

Thomas Merritt, Javier Latorre, and Simon King. Attributing modelling errors in hmm synthesis by stepping gradually from natural to modelled speech. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4220–4224. IEEE, 2015.

Thomas Merritt, Robert AJ Clark, Zhizheng Wu, Junichi Yamagishi, and Simon King. Deep neural network-guided unit selection synthesis. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5145–5149. IEEE, 2016.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

S. Minnis and A. Breen. Modeling visual coarticulation in synthetic talking heads using a lip motion unit inventory with concatenative synthesis. In *Interspeech*, Bejing, China, 2000.

Javier M Montero, J Gutiérrez-Arriola, José Colás, Emilia Enriquez, and José Manuel Pardo. Analysis and modelling of emotional speech in spanish. In *Proc. of ICPhS*, volume 2, pages 957–960, 1999.

Masahiro Mori et al. The uncanny valley. *Energy*, 7(4) :33–35, 1970.

Masanori Morise and Yusuke Watanabe. Sound quality comparison among high-quality vocoders by using re-synthesized speech. *Acoustical Science and Technology*, 39(3) :263–265, 2018.

Masanori Morise, Fumiya Yokomori, and Kenji Ozawa. World : a vocoder-based high-quality speech synthesis system for real-time applications. *IEICE TRANSACTIONS on Information and Systems*, 99(7) :1877–1884, 2016.

Michael W Morris and Dacher Keltner. *How Emotions Work : An Analysis of the Social Functions of Emotional Expression in Negotiation.* Graduate School of Business, Stanford University, 1999.

Eugene S Morton. On the occurrence and significance of motivation-structural rules in some bird and mammal sounds. *The American Naturalist*, 111(981) :855–869, 1977.

Eugene S Morton. Sound symbolism and its role in non-human vertebrate communication. *Sound symbolism*, pages 348–365, 1994.

Eric Moulines and Francis Charpentier. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech communication*, 9(5-6) :453–467, 1990.

Iain R Murray and John L Arnott. Implementation and testing of a system for producing emotion-by-rule in synthetic speech. *Speech Communication*, 16(4) :369–390, 1995.

Utpala Musti. *Acoustic-visual speech synthesis by bimodal unit selection.* PhD thesis, 2013.

Robin L Nabi. The theoretical versus the lay meaning of disgust : Implications for emotion research. *Cognition & Emotion*, 16(5) :695–703, 2002.

Radoslaw Niewiadomski, Magalie Ochs, and Catherine Pelachaud. Expressions of empathy in ecas. In *International Workshop on Intelligent Virtual Agents*, pages 37–44. Springer, 2008.

Jun-yong Noh, Douglas Fidaleo, and Ulrich Neumann. Animated deformations with radial basis functions. In *Proceedings of the ACM symposium on Virtual reality software and technology*, pages 166–174. ACM, 2000.

Magnus Nordstrand, Gunilla Svanfeldt, Björn Granström, and David House. Measurements of articulatory variation in expressive speech for a set of swedish vowels. *Speech Communication*, 44(1-4) :187–196, 2004.

Takashi Nose, Junichi Yamagishi, Takashi Masuko, and Takao Kobayashi. A style control technique for hmm-based expressive speech synthesis. *IEICE TRANSACTIONS on Information and Systems*, 90(9) :1406–1413, 2007.

Ana Margarida Belém Nunes. Cross-linguistic and cultural effects on the perception of emotions. *International Journal of Science Commerce and Humanities*, 1(8) :107–120, 2013.

Magalie Ochs and Philippe Blache. Virtual reality for training doctors to break bad news. In *European Conference on Technology Enhanced Learning*, pages 466–471. Springer, 2016.

Magalie Ochs, Yu Ding, Nesrine Fourati, Mathieu Chollet, Brian Ravenet, Florian Pecune, Nadine Glas, Ken Prepin, Chloé Clavel, and Catherine Pelachaud. Vers des agents conversationnels animés socio-affectifs. In *Proceedings of the 25th Conference on l'Interaction Homme-Machine*, pages 69–78, 2013.

Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet : A generative model for raw audio. *arXiv preprint arXiv :1609.03499*, 2016.

Charles Egerton Osgood, George J Suci, and Percy H Tannenbaum. *The measurement of meaning.* Number 47. University of Illinois press, 1957.

Douglas O'Shaughnessy, Louis Barbeau, David Bernardi, and Danièle Archambault. Diphone speech synthesis. *Speech communication*, 7(1) :55–65, 1988.

Jörn Ostermann and David Millen. Talking heads and synthetic speech : An architecture for supporting electronic commerce. In *2000 IEEE International Conference on Multimedia and Expo. ICME2000. Proceedings. Latest Advances in the Fast Changing World of Multimedia (Cat. No. 00TH8532)*, volume 1, pages 71–74. IEEE, 2000.

Slim Ouni and Sara Dahmani. Is markerless acquisition technique adequate for speech production ? *The Journal of the Acoustical Society of America*, 139(6) :EL234–EL239, 2016.

Slim Ouni and Guillaume Gris. Dynamic lip animation from a limited number of control points : Towards an effective audiovisual spoken communication. *Speech Communication*, 96 :49–57, 2018.

Slim Ouni, Loïc Mangeonjean, and Ingmar Steiner. Visartico : a visualization tool for articulatory data. In *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

Slim Ouni, Vincent Colotte, Utpala Musti, Asterios Toutios, Brigitte Wrobel-Dautcourt, Marie-Odile Berger, and Caroline Lavecchia. Acoustic-visual synthesis technique using bimodal unit-selection. *EURASIP Journal on Audio, Speech, and Music Processing*, 2013(1) :16, 2013.

Slim Ouni, Vincent Colotte, Sara Dahmani, and Soumaya Azzi. Acoustic and Visual Analysis of Expressive Speech : A Case Study of French Acted Speech. In *Interspeech 2016*, San Francisco, United States, 2016. ISCA. URL `https://hal.inria.fr/hal-01398528`.

Slim Ouni, Sara Dahmani, and Vincent Colotte. On the quality of an expressive audiovisual corpus : a case study of acted speech. In Slim Ouni, Chris Davis, Alexandra Jesse, and Jonas Beskow, editors, *The 14th International Conference on Auditory-Visual Speech Processing*, Stockholm, Sweden, 2017. URL `https://hal.inria.fr/hal-01596614`.

Astrid Paeschke, Miriam Kienast, Walter F Sendlmeier, et al. F0-contours in emotional speech. In *Proc. 14th Int. Congress of Phonetic Sciences*, volume 2, pages 929–932, 1999.

Igor S Pandzic and Robert Forchheimer. Mpeg-4 facial animation. *The standard, implementation and applications. Chichester, England : John Wiley&Sons*, 2002.

Igor S Pandzic, Jörn Ostermann, and David Millen. User evaluation : Synthetic talking faces for interactive services. *The visual computer*, 15(7-8) :330–340, 1999.

Jonathan Parker, Ranniery Maia, Yannis Stylianou, and Roberto Cipolla. Expressive visual text to speech and expression adaptation using deep neural networks. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4920–4924. IEEE, 2017.

Catherine Pelachaud and Isabella Poggi. Subtleties of facial expressions in embodied agents. *The Journal of Visualization and Computer Animation*, 13(5) :301–312, 2002.

Catherine Pelachaud, Norman I Badler, and Mark Steedman. Generating facial expressions for speech. *Cognitive science*, 20(1) :1–46, 1996.

Marc D Pell, Silke Paulmann, Chinar Dara, Areej Alasseri, and Sonja A Kotz. Factors in the recognition of vocally expressed emotions : A comparison of four languages. *Journal of Phonetics*, 37(4) :417–435, 2009.

Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove : Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

David Philippou-Hübner, Bogdan Vlasenko, Ronald Böck, and Andreas Wendemuth. The performance of the speaking rate parameter in emotion recognition from speech. In *2012 IEEE International Conference on Multimedia and Expo Workshops*, pages 296–301. IEEE, 2012.

John F Pitrelli, Raimo Bakis, Ellen M Eide, Raul Fernandez, Wael Hamza, and Michael A Picheny. The ibm expressive text-to-speech synthesis system for american english. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(4) :1099–1108, 2006.

Robert Plutchik. Emotions : A general psychoevolutionary theory. *Approaches to emotion*, 1984 : 197–219, 1984.

Jonathan Posner, James A Russell, and Bradley S Peterson. The circumplex model of affect : An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and psychopathology*, 17(3) :715–734, 2005.

Mael Pouget. *Synthèse incrémentale de la parole à partir du texte*. PhD thesis, Grenoble Alpes, 2017.

Stanisław Puppel and Stanisław Puppel. *The biology of language*. J. Benjamins, 1995.

Raheel Qader, Gwénolé Lecorvé, Damien Lolive, and Pascale Sébillot. Phonology modelling for expressive speech synthesis : a review. 2014.

Yao Qian, Yuchen Fan, Wenping Hu, and Frank K Soong. On the training aspects of deep neural network (dnn) for parametric tts synthesis. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3829–3833. IEEE, 2014.

Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv :1511.06434*, 2015.

Erhard Rank and Hannes Pirker. Generating emotional speech with a concatenative synthesizer. In *Fifth International Conference on Spoken Language Processing*, 1998.

Taehyun Rhee, Youngkyoo Hwang, James Dokyoon Kim, and Changyeong Kim. Real-time facial animation from live video tracking. In *Proceedings of the 2011 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, pages 215–224. ACM, 2011.

Manuel Sam Ribeiro, Oliver Watts, and Junichi Yamagishi. Syllable-level representations of suprasegmental features for dnn-based text-to-speech synthesis. In *INTERSPEECH*, pages 3186–3190, 2016.

Fanny Roche, Thomas Hueber, Samuel Limier, and Laurent Girin. Autoencoders for music sound synthesis : a comparison of linear, shallow, deep and variational models. *IEEE SMC 2019*, 2019.

Srikanth Ronanki. Prosody generation for text-to-speech synthesis. 2019.

Soufiane Rouibia. *Prise en compte de criteres acoustiques pour la synthese de la parole*. PhD thesis, 2006.

David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088) :533–536, 1986.

James A Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39 (6) :1161, 1980.

James A Russell and Beverly Fehr. Fuzzy concepts in a fuzzy hierarchy : Varieties of anger. *Journal of personality and social psychology*, 67(2) :186, 1994.

Zsófia Ruttkay, Claire Dormann, and Han Noot. Embodied conversational agents on a common ground : A framework for design and evaluation. In *From brows to trust : evaluating embodied conversational agents*, pages 27–66. Kluwer, 2004.

Yoshinori Sagisaka. Speech synthesis by rule using an optimal selection of non-uniform synthesis units. In *ICASSP-88., International Conference on Acoustics, Speech, and Signal Processing*, pages 679–682. IEEE, 1988.

Shinji Sako, Keiichi Tokuda, Takashi Masuko, Takao Kobayashi, and Tadashi Kitamura. Hmm-based text-to-audio-visual speech synthesis. In *Sixth International Conference on Spoken Language Processing*, 2000.

Peter Salovey, Brian T Bedell, Jerusha B Detweiler, and John D Mayer. Current directions in emotional intelligence research. 2000.

Dietmar Schabus, Michael Pucher, and Gregor Hofer. Speaker-adaptive visual speech synthesis in the hmm-framework. In *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

Dietmar Schabus, Michael Pucher, and Gregor Hofer. Joint audiovisual hidden semi-markov model-based speech synthesis. *IEEE Journal of Selected Topics in Signal Processing*, 8(2) : 336–347, 2013.

Klaus R Scherer. Vocal affect expression : A review and a model for future research. *Psychological bulletin*, 99(2) :143, 1986.

Marc Schröder. Emotional speech synthesis : A review. In *Seventh European Conference on Speech Communication and Technology*, 2001.

Marc Schröder. Experimental study of affect bursts. *Speech communication*, 40(1-2) :99–116, 2003.

Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11) :2673–2681, 1997.

Felix Shaw and Barry-John Theobald. Expressive modulation of neutral visual speech. *IEEE MultiMedia*, 23(4) :68–78, 2016.

Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4779–4783. IEEE, 2018.

Kengo Shichiri, Atsushi Sawabe, Takayoshi Yoshimura, Keiichi Tokuda, Takashi Masuko, Takao Kobayashi, and Tadashi Kitamura. Eigenvoices for hmm-based speech synthesis. In *Seventh International Conference on Spoken Language Processing*, 2002.

AD Simons. Generation of mouthshape for a synthetic talking head. *Proc. of the Institute of Acoustics*, 1990.

RJ Skerry-Ryan, Eric Battenberg, Ying Xiao, Yuxuan Wang, Daisy Stanton, Joel Shor, Ron J Weiss, Rob Clark, and Rif A Saurous. Towards end-to-end prosody transfer for expressive speech synthesis with tacotron. *arXiv preprint arXiv :1803.09047*, 2018.

Lee Sproull, Mani Subramani, Sara Kiesler, Janet H Walker, and Keith Waters. When the interface is a face. *Human-Computer Interaction*, 11(2) :97–124, 1996.

Massimo Stella, Antonio Stella, Francesco Sigona, Paolo Bernardini, Mirko Grimaldi, and Barbara Gili Fivela. Electromagnetic articulography with ag500 and ag501. In *Interspeech*, pages 1316–1320, 2013.

Heidi K Swanson, Martin Lysy, Michael Power, Ashley D Stasko, Jim D Johnson, and James D Reist. A new probabilistic method for quantifying n-dimensional ecological niches and niche overlap. *Ecology*, 96(2) :318–324, 2015.

Makoto Tachibana, Junichi Yamagishi, Koji Onishi, Takashi Masuko, and Takao Kobayashi. Hmm-based speech synthesis with various speaking styles using model interpolation. In *Speech Prosody 2004, International Conference*, 2004.

Yaniv Taigman, Lior Wolf, Adam Polyak, and Eliya Nachmani. Voiceloop : Voice fitting and synthesis via a phonological loop. *arXiv preprint arXiv :1707.06588*, 2017.

Masatsune Tamura, Takashi Masuko, Keiichi Tokuda, and Takao Kobayashi. Speaker adaptation for hmm-based speech synthesis system using mllr. In *the third ESCA/COCOSDA Workshop (ETRW) on Speech Synthesis*, 1998.

Masatsune Tamura, Takashi Masuko, Keiichi Tokuda, and Takao Kobayashi. Adaptation of pitch and spectrum for hmm-based speech synthesis using mllr. In *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221)*, volume 2, pages 805–808. IEEE, 2001.

Hao Tang, Yun Fu, Jilin Tu, Mark Hasegawa-Johnson, and Thomas S Huang. Humanoid audio–visual avatar with emotive text-to-speech synthesis. *IEEE Transactions on multimedia*, 10(6) : 969–981, 2008a.

Hao Tang, Yun Fu, Jilin Tu, Thomas S Huang, and Mark Hasegawa-Johnson. Eava : a 3d emotive audio-visual avatar. In *2008 IEEE Workshop on Applications of Computer Vision*, pages 1–6. IEEE, 2008b.

Jianhua Tao, Le Xin, and Panrong Yin. Realistic visual speech synthesis based on hybrid concatenation method. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(3) : 469–477, 2009.

Paul Taylor and Alan W Black. Speech synthesis by phonological structure matching. 1999.

Mariët Theune, Koen Meijs, Dirk Heylen, and Roeland Ordelman. Generating expressive speech for storytelling applications. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(4) :1137–1144, 2006.

Y-I Tian, Takeo Kanade, and Jeffrey F Cohn. Recognizing action units for facial expression analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 23(2) :97–115, 2001.

Noé Tits, Fengna Wang, Kevin El Haddad, Vincent Pagel, and Thierry Dutoit. Visualization and interpretation of latent spaces for controlling expressive speech synthesis through audio analysis. *Interspeech*, 2019.

Keiichi Tokuda, Takao Kobayashi, Takashi Masuko, and Satoshi Imai. Mel-generalized cepstral analysis-a unified approach to speech spectral estimation. In *Third International Conference on Spoken Language Processing*, 1994.

Keiichi Tokuda, Takao Kobayashi, and Satoshi Imai. Speech parameter generation from hmm using dynamic features. In *1995 International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 660–663. IEEE, 1995a.

Keiichi Tokuda, Takashi Masuko, Tetsuya Yamada, Takao Kobayashi, and Satoshi Imai. An algorithm for speech parameter generation from continuous mixture hmms with dynamic features. In *Fourth European Conference on Speech Communication and Technology*, 1995b.

Keiichi Tokuda, Takashi Masuko, Noboru Miyazaki, and Takao Kobayashi. Hidden markov models based on multi-space probability distribution for pitch pattern modeling. In *icassp*, volume 1, pages 229–232, 1999.

Keiichi Tokuda, Takayoshi Yoshimura, Takashi Masuko, Takao Kobayashi, and Tadashi Kitamura. Speech parameter generation algorithms for hmm-based speech synthesis. In *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 00CH37100)*, volume 3, pages 1315–1318. IEEE, 2000.

Keiichi Tokuda, Takashi Masuko, Noboru Miyazaki, and Takao Kobayashi. Multi-space probability distribution hmm. *IEICE TRANSACTIONS on Information and Systems*, 85(3) :455–464, 2002.

Michael Tomasello. *Origins of human communication*. MIT press, 2010.

Jürgen Trouvain, Jacques Koreman, Attilio Erriquez, and Bettina Zinn. Articulation rate measures and their relation to phone classification in spontaneous and read german speech. In *Isca ITR-workshop*, pages 155–158, 2001.

Susanne Van Mulken, Elisabeth André, and Jochen Müller. An empirical study on the trustworthiness of life-like interface agents. In *HCI (2)*, pages 152–156, 1999.

Philippe Verduyn, Pauline Delaveau, Jean-Yves Rotgé, Philippe Fossati, and Iven Van Mechelen. Determinants of emotion duration and underlying psychological and neural mechanisms. *Emotion Review*, 7(4) :330–335, 2015.

Jean Vroomen, René Collier, and Sylvie Mozziconacci. Duration and intonation in emotional speech. In *Third European Conference on Speech Communication and Technology*, 1993.

Janet H Walker, Lee Sproull, and R Subramani. Using a human face in an interface. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 85–91. ACM, 1994.

Vincent Wan, Robert Anderson, Art Blokland, Norbert Braunschweiler, Langzhou Chen, BalaKrishna Kolluru, Javier Latorre, Ranniery Maia, Björn Stenger, Kayoko Yanagisawa, et al. Photo-realistic expressive text to talking head synthesis. In *INTERSPEECH*, pages 2667–2669, 2013.

Angelina Wang, Nathan Blair, and Suneel Belkhale. Encouraging categorical meaning in the latent space of a vae.

Lijuan Wang, Xiaojun Qian, Wei Han, and Frank K Soong. Photo-real lips synthesis with trajectory-guided sample selection. In *Seventh ISCA Workshop on Speech Synthesis*, 2010.

Lijuan Wang, Wei Han, Frank K Soong, and Qiang Huo. Text driven 3d photo-realistic talking head. In *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al. Tacotron : Towards end-to-end speech synthesis. *arXiv preprint arXiv :1703.10135*, 2017.

Keith Waters. A muscle model for animation three-dimensional facial expression. *Acm siggraph computer graphics*, 21(4) :17–24, 1987.

Oliver Watts, Junichi Yamagishi, and Simon King. The role of higher-level linguistic features in hmm-based speech synthesis. 2010.

Oliver Watts, Adriana Stan, Yoshitaka Mamiya, Antti Suni, Jos Martn Burgos, and Juan Manuel Montero. The simple4all entry to the blizzard challenge 2013. In *Blizzard Challenge Workshop*, 2013.

Oliver Watts, Srikanth Ronanki, Zhizheng Wu, Tuomo Raitio, and Antti Suni. The nst–glotthmm entry to the blizzard challenge 2015. In *Proc. Blizzard Challenge Workshop*, 2015a.

Oliver Watts, Zhizheng Wu, and Simon King. Sentence-level control vectors for deep neural network speech synthesis. In *Sixteenth Annual Conference of the International Speech Communication Association*, 2015b.

Oliver Watts, Gustav Eje Henter, Thomas Merritt, Zhizheng Wu, and Simon King. From hmms to dnns : where do the improvements come from ? In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5505–5509. IEEE, 2016.

Michiel Wiggers. Judgments of facial expressions of emotion predicted from facial behavior. *Journal of Nonverbal Behavior*, 7(2) :101–116, 1982.

Xixin Wu, Lifa Sun, Shiyin Kang, Songxiang Liu, Zhiyong Wu, Xunying Liu, and Helen Meng. Feature based adaptation for speaking style synthesis. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5304–5308. IEEE, 2018.

Yi-Jian Wu and Ren-Hua Wang. Minimum generation error training for hmm-based speech synthesis. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, volume 1, pages I–I. IEEE, 2006.

Zhiyong Wu, Shen Zhang, Lianhong Cai, and Helen M Meng. Real-time synthesis of chinese visual speech and facial expressions using mpeg-4 fap features in a three-dimensional avatar. In *Ninth International Conference on Spoken Language Processing*, 2006.

Zhizheng Wu, Pawel Swietojanski, Christophe Veaux, Steve Renals, and Simon King. A study of speaker adaptation for dnn-based speech synthesis. In *Sixteenth Annual Conference of the International Speech Communication Association*, 2015a.

Zhizheng Wu, Cassia Valentini-Botinhao, Oliver Watts, and Simon King. Deep neural networks employing multi-task learning and stacked bottleneck features for speech synthesis. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4460–4464. IEEE, 2015b.

Zhizheng Wu, Oliver Watts, and Simon King. Merlin : An open source neural network speech synthesis system. In *SSW*, pages 202–207, 2016.

Lei Xie, Naicai Sun, and Bo Fan. A statistical parametric approach to video-realistic text-driven talking avatar. *Multimedia tools and applications*, 73(1) :377–396, 2014.

Liumeng Xue, Xiaolian Zhu, Xiaochun An, and Lei Xie. A comparison of expressive speech synthesis approaches based on neural network. In *Proceedings of the Joint Workshop of the 4th Workshop on Affective Social Multimedia Computing and first Multi-Modal Affective Computing of Large-Scale Multimedia Data*, pages 15–20, 2018a.

Yawen Xue, Yasuhiro Hamada, and Masato Akagi. Voice conversion for emotional speech : Rule-based synthesis with degree of emotion controllable in dimensional space. *Speech Communication*, 102 :54–67, 2018b.

Junichi Yamagishi, Koji Onishi, Takashi Masuko, and Takao Kobayashi. Modeling of various speaking styles and emotions for hmm-based speech synthesis. In *Eighth European Conference on Speech Communication and Technology*, 2003a.

Junichi Yamagishi, Masatsune Tamura, Takashi Masuko, Keiichi Tokuda, and Takao Kobayashi. A training method of average voice model for hmm-based speech synthesis. *IEICE transactions on fundamentals of electronics, communications and computer sciences*, 86(8) :1956–1963, 2003b.

Junichi Yamagishi, Takashi Masuko, and Takao Kobayashi. Hmm-based expressive speech synthesis-towards tts with arbitrary speaking styles and emotions. In *Proc. of Special Workshop in Maui (SWIM)*, 2004.

John Yang, Gyuejeong Lee, Simyung Chang, and Nojun Kwak. Towards governing agent's efficacy : Action-conditional beta-vae for deep transparent reinforcement learning. In *Asian Conference on Machine Learning*, pages 32–47, 2019.

Takayoshi Yoshimura, Takashi Masuko, Keiichi Tokuda, Takao Kobayashi, and Tadashi Kitamura. Speaker interpolation in hmm-based speech synthesis system. In *Fifth European Conference on Speech Communication and Technology*, 1997.

Takayoshi Yoshimura, Keiichi Tokuda, Takashi Masuko, Takao Kobayashi, and Tadashi Kitamura. Duration modeling for hmm-based speech synthesis. In *Fifth International Conference on Spoken Language Processing*, 1998.

Takayoshi Yoshimura, Keiichi Tokuda, Takashi Masuko, Takao Kobayashi, and Tadashi Kitamura. Simultaneous modeling of spectrum, pitch and duration in hmm-based speech synthesis. In *Sixth European Conference on Speech Communication and Technology*, 1999.

Takayoshi Yoshimura, Keiichi Tokuda, Takashi Masuko, Takao Kobayashi, and Tadashi Kitamura. Speaker interpolation for hmm-based speech synthesis system. *Acoustical Science and Technology*, 21(4) :199–206, 2000.

Takayoshi Yoshimura, Keiichi Tokuda, Takashi Masuko, Takao Kobayashi, and Tadashi Kitamura. Mixed excitation for hmm-based speech synthesis. In *Seventh European Conference on Speech Communication and Technology*, 2001.

Steve J Young, Julian J Odell, and Philip C Woodland. Tree-based state tying for high accuracy acoustic modelling. In *Proceedings of the workshop on Human Language Technology*, pages 307–312. Association for Computational Linguistics, 1994.

Kai Yu, François Mairesse, and Steve Young. Word-level emphasis modelling in hmm-based speech synthesis. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4238–4241. IEEE, 2010.

Kai Yu, Heiga Zen, François Mairesse, and Steve Young. Context adaptive training with factorized decision trees for hmm-based statistical parametric speech synthesis. *Speech communication*, 53(6) :914–923, 2011.

Yana Yunusova, Jordan R Green, and Antje Mefferd. Accuracy assessment for ag500, electromagnetic articulograph. *Journal of Speech, Language, and Hearing Research*, 52(2) :547–555, 2009.

Lukasz Zalewski and Shaogang Gong. Synthesis and recognition of facial expressions in virtual 3d views. In *Sixth IEEE International Conference on Automatic Face and Gesture Recognition, 2004. Proceedings.*, pages 493–498. IEEE, 2004.

Heiga Ze, Andrew Senior, and Mike Schuster. Statistical parametric speech synthesis using deep neural networks. In *2013 ieee international conference on acoustics, speech and signal processing*, pages 7962–7966. IEEE, 2013.

Heiga Zen and Tomoki Toda. An overview of nitech hmm-based speech synthesis system for blizzard challenge 2005. In *Proc. Blizzard Challenge*, 2005.

Heiga Zen, Keiichi Tokuda, and Tadashi Kitamura. An introduction of trajectory model into hmm-based speech synthesis. In *Fifth ISCA Workshop on Speech Synthesis*, 2004a.

Heiga Zen, Keiichi Tokuda, Takashi Masuko, Takao Kobayashi, and Tadashi Kitamura. Hidden semi-markov model based speech synthesis. In *Eighth International Conference on Spoken Language Processing*, 2004b.

Heiga Zen, Takashi Nose, Junichi Yamagishi, Shinji Sako, Takashi Masuko, Alan W Black, and Keiichi Tokuda. The hmm-based speech synthesis system (hts) version 2.0. In *SSW*, pages 294–299. Citeseer, 2007.

Yue Zhang, Yifan Liu, Felix Weninger, and Björn Schuller. Multi-task deep neural network with shared hidden layers : Breaking down the wall between emotion representations. In *2017*

IEEE international conference on acoustics, speech and signal processing (ICASSP), pages 4990–4994. IEEE, 2017.

Yang Zhou, Zhan Xu, Chris Landreth, Evangelos Kalogerakis, Subhransu Maji, and Karan Singh. Visemenet : Audio-driven animator-centric speech animation. *ACM Transactions on Graphics (TOG)*, 37(4) :161, 2018.

Résumé

Les travaux de cette thèse portent sur la modélisation des émotions pour la synthèse audiovisuelle expressive de la parole à partir du texte. Aujourd'hui, les résultats des systèmes de synthèse de la parole à partir du texte sont de bonne qualité, toutefois la synthèse audiovisuelle reste encore une problématique ouverte et la synthèse expressive l'est encore d'avantage. Nous proposons dans le cadre de cette thèse une méthode de modélisation des émotions malléable et flexible, permettant de mélanger les émotions comme on mélange les teintes sur une palette de couleurs.

Dans une première partie, nous présentons et étudions deux corpus expressifs que nous avons construits. La stratégie d'acquisition ainsi que le contenu expressif de ces corpus sont analysés pour valider leur utilisation à des fin de synthèse audiovisuelle de la parole.

Dans une seconde partie, nous proposons deux architectures neuronales pour la synthèse de la parole. Nous avons utilisé ces deux architectures pour modéliser trois aspects de la parole : 1) les durées des sons, 2) la modalité acoustique et 3) la modalité visuelle. Dans un premier temps, nous avons adopté une architecture entièrement connectée. Cette dernière nous a permis d'étudier le comportement des réseaux de neurones face à différents descripteurs contextuels et linguistiques. Nous avons aussi pu analyser, via des mesures objectives, la capacité du réseau à modéliser les émotions.

La deuxième architecture neuronale proposée est celle d'un auto encodeur variationnel. Cette architecture est capable d'apprendre une représentation latente des émotions sans utiliser les étiquettes des émotions. Après analyse de l'espace latent des émotions, nous avons proposé une procédure de structuration de ce dernier pour pouvoir passer d'une représentation par catégorie vers une représentation continue des émotions. Nous avons pu valider, via des expériences perceptives, la capacité de notre système à générer des émotions, des nuances d'émotions et des mélanges d'émotions, et cela pour la synthèse audiovisuelle expressive de la parole à partir du texte.

Mots-clés: Synthèse audiovisuelle expressive de la parole, auto-encodeur variationnel conditionné, tête parlante expressive, émotion, expression faciale, réseau de neurones profond récurrent à mémoire court-terme et long terme

Abstract

The work of this thesis concerns the modeling of emotions for expressive audiovisual text-to-speech synthesis. Today, the results of text-to-speech synthesis systems are of good quality, however audiovisual synthesis remains an open issue and expressive synthesis is even less studied. As part of this thesis, we present an emotions modeling method which is malleable and flexible, and allows us to mix emotions as we mix shades on a palette of colors.

In the first part, we present and study two expressive corpora that we have built. The recording strategy and the expressive content of these corpora are analyzed to validate their use for the purpose of audiovisual speech synthesis.

In the second part, we present two neural architectures for speech synthesis. We used these two architectures to model three aspects of speech : 1) the duration of sounds, 2) the acoustic modality and 3) the visual modality. First, we use a fully connected architecture. This architecture allowed us to study the behavior of neural networks when dealing with different contextual and linguistic descriptors. We were also able to analyze, with objective measures, the network's ability to model emotions.

The second neural architecture proposed is a variational auto-encoder. This architecture is able to learn a latent representation of emotions without using emotion labels. After analyzing the latent space of emotions, we presented a procedure for structuring it in order to move from a discrete representation of emotions to a continuous one. We were able to validate, through perceptual experiments, the ability of our system to generate emotions, nuances of emotions and mixtures of emotions, and this for expressive audiovisual text-to-speech synthesis.

Keywords: Expressive audiovisual speech synthesis, conditional variational auto-encoder, Expressive talking virtual agent, emotion, facial expression, deep bidirectional long short-term memory (DBLSTM)