Exploration de la représentation multidimensionnelle de paramètres acoustiques unidimensionnels de la parole extraits par des modèles profonds non supervisés.

Maxime Jacquelin, Maëva Garnier, Laurent Girin, Rémy Vincent, Olivier Perrotin

Exploration de la représentation multidimensionnelle de paramètres acoustiques unidimensionnels de la parole extraits par des modèles profonds non supervisés.*

Maxime Jacquelin[1, 2] Maëva Garnier[1] Laurent Girin[1] Rémy Vincent[2]
Olivier Perrotin[1]
(1) Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, F-38000 Grenoble, France
(2) Vogo, F-38190 Bernin, France
maxime.jacquelin, maeva.garnier, laurent.girin,
olivier.perrotin@grenoble-inp.fr, r.vincent@vogo-group.com

RÉSUMÉ

Cet article propose une méthodologie pour interpréter les dimensions de variation de la parole conversationnelle, extraites de façon non-supervisée, et sur des données multilocuteurs, par un algorithme d'apprentisage profond (Auto-Encodeur Variationnel). Par des analyses de corrélation et de similarité cosinus, nous montrons que la distribution de la fréquence fondamentale et de la fréquence centrale des trois premiers formants de l'ensemble d'apprentissage est encodée par une direction dédiée de l'espace latent. Lorsque la distribution est multimodale, les différents modes du paramètre acoustique sont encodés dans des dimensions distinctes. De plus, nous avons identifié les directions expliquant la variation des paramètres au sein de chaque mode, et entre eux.

ABSTRACT

Exploring the multidimensional representation of unidimensional speech acoustic parameters extracted by deep unsupervised models

This paper proposes a methodology for interpreting the dimensions of variations of conversational speech, extracted in an unsupervised manner, and on multi-speaker data, by a deep learning algorithm (Variational Auto-Encoder). Using correlation and cosine similarity analyses, we show that the distribution of the fundamental frequency and the central frequencies of the first three formants of the training set is encoded by one dedicated latent space direction. When the distribution is multimodal, different modes of the acoustic feature are encoded in separate dimensions. In addition, we also have identified the directions that explain the variation of the feature within and across modes.

MOTS-CLÉS : apprentissage de représentations, codage de la parole, auto-encodeur variationnel, modèle source-filtre.

KEYWORDS: representation learning, speech encoding, variational autoencoder, source-filter model.

*. note aux relecteurs : cette soumission est une traduction d'un article publié à XAI-SA IEEE ICASSP Workshop, Explainable Machine Learning for Speech and Audio, 2024 (Jacquelin *et al.*, 2024).

1 Introduction

Depuis les modèles physiques jusqu'aux approches d'apprentissage profond sur données massives, la modélisation de la parole trouve des applications dans la reconnaissance automatique de la parole, le codage de la parole, la synthèse vocale expressive ou la conversion de voix. Pour chaque cas, l'objectif de la modélisation de la parole est de comprendre comment les signaux sont générés et comment leurs caractéristiques acoustiques peuvent être modulées à partir d'un nombre limité de dimensions de contrôle, aussi indépendantes que possible les unes des autres.

En ce sens, les premiers modèles acoustiques tels que le modèle source-filtre de Fant (Fant, 1971) établissent une distinction claire entre les variations acoustiques liées à la source glottique (variations de la fréquence fondamentale f_0, d'apériodicité, d'inharmonicité ou de la pente spectrale) et celles, supposées indépendantes, liées à l'articulation du conduit vocal (formants ou pics spectraux dans les bruits turbulents). Bien qu'un tel ensemble de paramètres acoustiques présente l'avantage d'être facilement interprétable en termes de physiologie et de contrôle gestuel sous-jacent, ils sont largement interdépendants, avec des contraintes anatomiques et physiques qui sous-tendent les covariations de ces paramètres au sein d'un même individu et d'un individu à l'autre (Coleman, 1971).

Bien que les modèles génératifs profonds non- et auto-supervisés soient des outils puissants pour modéliser toute la complexité des signaux de parole (Kingma & Welling, 2014; Van Der Oord *et al.*, 2017; Baevski *et al.*, 2020; Hsu *et al.*, 2021; Lakhotia *et al.*, 2021), peu de recherches ont été menées jusqu'à présent sur l'interprétation de leurs représentations latentes. Plusieurs études ont déjà utilisé des auto-encodeurs variationnels (VAE) (Kingma & Welling, 2014), VQ-VAE (Van Der Oord *et al.*, 2017), ou des modèles auto-supervisés tels que HuBERT (Hsu *et al.*, 2021), pour trouver des espaces de représentation des variations discrètes de la parole, avec des dimensions qui distinguent les informations phonémiques (de bas niveau) de celles liées à la langue (Williams *et al.*, 2021), à l'identité du locuteur (Chou *et al.*, 2018) et/ou au "style" de parole (Williams & King, 2019) (de haut niveau). D'autres études récentes ont été en mesure de trouver un espace latent de faible dimension pour représenter et contrôler les variations acoustiques continues de la parole expressive (en termes d'intonation, de contenu spectral ou de rythme) (Blaauw & Bonada, 2016; Hsu *et al.*, 2017; Wang *et al.*, 2018; Zhang *et al.*, 2019; Tits *et al.*, 2019; Bous & Roebel, 2022; Lenglet *et al.*, 2022b; Vaidya *et al.*, 2022; Sadok *et al.*, 2023). Elles ont montré que, même avec une approche d'apprentissage entièrement non-supervisée, les paramètres acoustiques étaient encodés selon différents sous-espaces quasi orthogonaux de la représentation apprise. Cela leur a permis de contrôler certains aspects de l'intonation liés à la source glottique, de manière presque indépendante des variations de l'enveloppe spectrale, liées à l'articulation du conduit vocal.

Un phénomène qui reste cependant inexpliqué est que chaque paramètre acoustique est souvent encodé par plusieurs dimensions latentes (Sadok *et al.*, 2023), et la question de savoir quel type d'information est capturé par chacune de ces dimensions reste peu explorée. Parmi les multiples interactions possibles entre les paramètres acoustiques, nous faisons l'hypothèse dans cette étude que *la multiplicité des dimensions latentes observées peut refléter l'encodage des différentes sources de variabilité inter- et intra-individuelle de chaque paramètre acoustique*. À notre connaissance, il s'agit de l'une des premières tentatives visant à étudier l'interaction entre les représentations vocales de bas niveau (paramètres acoustiques) et de haut niveau (liées au locuteur, telles que le genre) dans les espaces latents. Pour cela, le choix d'un VAE offre plusieurs avantages par rapport à d'autres méthodes non- ou auto-supervisées. D'abord, sa nature stochastique et la régularisation de son espace latent fournit une représentation latente désenchevêtrée, ce qui permet de modéliser les variations des

caractéristiques de la parole dépendantes et indépendantes du locuteur dans des directions distinctes de l'espace latent. Ensuite, les VAEs permettent une forte réduction de dimension, ce qui favorise l'interprétabilité de l'espace latent. Enfin, les modèles auto-supervisés tels que wav2vec 2.0 ou HuBERT (Baevski *et al.*, 2020; Hsu *et al.*, 2021) sont entraînés en utilisant une tâche de clustering de phonèmes, donnant ainsi la priorité à l'encodage des informations phonétiques au détriment de la prosodie ou de la paralinguistique, qui sont souvent modélisées séparément (Polyak *et al.*, 2021). Dans notre étude, il est crucial de garantir l'encodage à la fois des informations vocales de bas niveau et de haut niveau dans l'espace latent du modèle.

Nous optons donc pour une approche basée sur les VAEs et introduisons une méthodologie reposant sur l'analyse en composantes principales (PCA), la régression linéaire (LR) et l'analyse discriminante linéaire (LDA), afin d'analyser et d'interpréter l'aspect multidimensionnel de la représentation des paramètres acoustiques individuels, avant de tester notre hypothèse.

2 Méthodologie

2.1 Entraînement du VAE

L'architecture VAE utilisée dans cette étude est similaire à celle utilisée par Sadok *et al.* (2023). Elle prend en entrée des trames de spectrogrammes d'amplitude de la transformée de Fourier à court terme (TFCT) de taille 513. La dimension du vecteur latent $\mathbf{z}$ est fixée à 16. L'encodeur se compose de trois couches cachées entièrement connectées de 256, 64 et 2×16 unités (pour les vecteurs de moyenne et de variance de $\mathbf{z}$), toutes avec une activation tangente hyperbolique. Le décodeur est construit symétriquement à l'encodeur.

Deux modèles VAEs indépendants ont été entraînés, l'un avec la base de données VCTK (Yamagishi *et al.*, 2019), l'autre avec la base de données Att-HACK (Le Moine & Obin, 2020). VCTK comprend 109 locuteurs anglais lisant les mêmes 400 énoncés. Att-HACK comprend 25 locuteurs français qui ont acté les mêmes 100 énoncés dans quatre attitudes sociales (amicale, distante, dominante et séductrice) avec 3 à 5 répétitions. Nous appelons VAE-VCTK le modèle entraîné sur VCTK et VAE-AH le modèle entraîné sur Att-HACK. L'ensemble d'entraînement VAE-VCTK contient 25 heures de parole provenant de 29 locuteurs féminins et 29 locuteurs masculins, sélectionnés de manière aléatoire. L'ensemble de validation contient 3 heures provenant de 10 locuteurs féminins et 10 locuteurs masculins qui n'ont pas été utilisés pour l'entraînement. L'ensemble d'entraînement VAE-AH contient 20 heures de parole de 7 locuteurs féminins et 7 locuteurs masculins, sélectionnés de manière aléatoire. L'ensemble de validation contient 3 heures provenant de 2 locuteurs féminins et 2 locuteurs masculins non utilisés pour l'entraînement. Tous les signaux ont été sous-échantillonnés de 48 pour VCTK et 44.1 pour Att-HACK à $16\,\mathrm{kHz}$. La TFCT a été définie avec une fenêtre de Hanning de $64\,\mathrm{ms}$ et un chevauchement de $50\,\%$.

Les modèles ont été entraînés avec l'optimiseur Adam (Kingma & Ba, 2015) sur 500 époques, avec une taille de batch de 128 et un taux d'apprentissage de 10^{-4}. Nous avons utilisé le VAE d'Itakura-Saito (IS), c'est-à-dire que la fonction de coût est la somme pondérée d'une divergence IS pour le terme de reconstruction et de la divergence de Kullback-Leibler (KL) pour le terme de régularisation (Girin *et al.*, 2019) : $\mathcal{L}_{total} = \mathcal{L}_{\mathsf{IS}} + \beta\mathcal{L}_{\mathsf{KL}}$. Pour éviter le problème de disparition du gradient rencontré dans les VAEs, le critère de régularisation β a été utilisé (Higgins *et al.*, 2017).

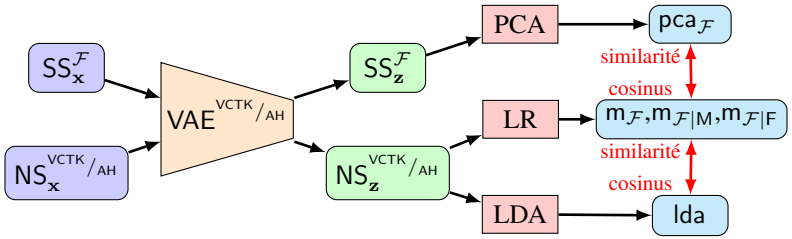


FIGURE 1 – Résumé de la méthodologie de test proposée.

2.2 Ensembles de test

Nous avons conçu plusieurs ensembles de tests pour évaluer les capacités de représentation de nos deux VAE. Tout d'abord, quatre ensembles de données ont été conçus pour démontrer l'aspect multidimensionnel de la représentation latente de chaque paramètre acoustique. Suivant Sadok *et al.* (2023), nous avons utilisé le logiciel Soundgen (Anikin, 2019) pour générer quatre signaux de 5 s avec une variation respective de f_0 et de la fréquence des trois premiers formants F_1, F_2, ou F_3. Quand l'un des paramètres varie, les trois autres paramètres sont laissés constants, fixés à la médiane de leur distribution dans les deux ensembles de données VCTK et Att-HACK (c'est-à-dire 140 Hz, 450 Hz, 1600 Hz et 2800 Hz, pour f_0, F_1, F_2 et F_3, respectivement). Pour chaque signal, la plage de variation du paramètre correspondant correspond à sa distribution dans les ensembles de données VCTK et Att-HACK : 85–310 Hz pour f_0 ; 290–890 Hz pour F_1 ; 960–2360 Hz pour F_2 ; et 2000–3430 Hz pour F_3. Nous avons ensuite calculé le spectrogramme d'amplitude de chaque signal, chacun constituant une base de données de test appelé $\mathsf{SS}_{\mathbf{x}}^{\mathcal{F}}$, $\mathcal{F} \in \{f_0, F_1, F_2, F_3\}$, avec SS pour "synthesis speech " (parole synthétique). $\mathsf{SS}_{\mathbf{z}}^{\mathcal{F}}$ sont les ensembles de test correspondants dans l'espace latent, c'est-à-dire $\mathsf{SS}_{\mathbf{x}}^{\mathcal{F}}$ passés par l'encodeur VAE.

Ensuite, pour analyser la capacité des VAEs à représenter les covariations naturelles entre les paramètres acoustiques, nous avons généré deux ensembles de test supplémentaires appelés $\mathsf{NS}_{\mathbf{x}}^{\mathsf{VCTK}}$ pour VAE-VCTK et $\mathsf{NS}_{\mathbf{x}}^{\mathsf{AH}}$ pour VAE-AH (avec $\mathsf{NS}_{\mathbf{z}}^{\mathsf{VCTK}}$ et $\mathsf{NS}_{\mathbf{z}}^{\mathsf{AH}}$ les ensembles de test correspondants dans l'espace latent des VAE), NS signifiant "natural speech" (parole naturelle). Ces ensembles de données sont constitués de 3 heures de signaux de parole naturelle provenant de 9 femmes et 9 hommes pour VCTK, et de 3 femmes et 3 hommes pour Att-HACK, qui ne font partie ni de l'ensemble d'apprentissage ni de l'ensemble de validation. Pour chaque signal de cet ensemble de test, une analyse acoustique par trame a été effectuée avec Praat (Boersma & Weenink, 2001) pour extraire f_0 et $F_{1,2,3}$.

2.3 Analyse PCA, LR et LDA

Notre analyse vise à identifier les directions dans l'espace latent qui capturent la plus grande variabilité expliquée par chaque paramètre acoustique de parole considéré. Notre première étape a été d'*étudier l'encodage de chaque paramètre acoustique séparément*, en utilisant les ensembles de données de test *de parole synthétique* $\mathsf{SS}_{\mathbf{x}}^{\mathcal{F}}$ spécifiquement conçus à cette fin. Indépendamment pour chaque paramètre acoustique $\mathcal{F} \in \{f_0, F_1, F_2, F_3\}$, nous avons appliqué une PCA sur les bases de données encodées $\mathsf{SS}_{\mathbf{z}}^{\mathcal{F}}$, afin d'extraire par le biais des composantes principales notées $\mathsf{pca}_{\mathcal{F}}$ les multiples directions dans l'espace latent qui expliquent la variation du paramètre acoustique.

Notre deuxième étape consiste à *identifier le rôle de ces dimensions multiples*, grâce à l'analyse de *la parole naturelle*. Pour ce faire, nous avons recherché la direction de variation (DV) de chaque paramètre acoustique dans nos bases de données de test de parole naturelle encodée $\mathsf{NS}_{\mathbf{z}}^{\mathsf{VCTK}}$ et $\mathsf{NS}_{\mathbf{z}}^{\mathsf{AH}}$. Indépendamment pour chaque paramètre $\mathcal{F}$, nous avons calculé une régression linéaire des valeurs $\mathcal{F}$, extraites de $\mathsf{NS}_{\mathbf{x}}$ avec Praat, sur les valeurs correspondantes de $\mathbf{z}$ dans $\mathsf{NS}_{\mathbf{z}}$. La DV de $\mathcal{F}$, notée $\mathsf{m}_{\mathcal{F}} \in \mathbb{R}^{16}$, est le vecteur des coefficients de la LR, c'est-à-dire :

$$\hat{\mathcal{F}} = \mathsf{m}_{\mathcal{F}}^{\top}\mathbf{z} + b_{\mathcal{F}} \approx \mathcal{F}, \tag{1}$$

au sens des moindres carrés ($b_{\mathcal{F}}$ est l'ordonnée à l'origine et $^{\top}$ désigne l'opérateur de transposition). Pour identifier le rôle de chaque dimension de la PCA ($\mathsf{pca}_{\mathcal{F}}$), nous avons ensuite analysé leur colinéarité avec les DV des paramètres acoustiques extraits dans des conditions spécifiquement choisies. En particulier, pour vérifier notre hypothèse, à savoir si les différentes dimensions latentes reflètent des sources de variabilité inter- et intra-individuelle de chaque paramètre acoustique, nous avons mesuré la DV sur l'ensemble du jeu de test, ainsi que la DV pour chaque genre de locuteurs. Nous désignons la DV résultante par $\mathsf{m}_{\mathcal{F}|\mathsf{M}}$ et $\mathsf{m}_{\mathcal{F}|\mathsf{F}}$ pour les locuteurs masculins et féminins, respectivement.

La représentation possible des paramètres acoustiques liés au genre dans des directions distinctes nous amène à faire un pas de plus pour *identifier une représentation indépendante de la variabilité inter- et intra-genre* dans l'espace latent. Le genre étant l'une des caractéristiques les plus discriminantes entre individus dans la parole, nous émettons l'hypothèse qu'une LDA calculée sur les locuteurs d'une base de données encodées $\mathsf{NS}_{\mathbf{z}}$ (noté lda), qui trouve la combinaison linéaire des dimensions latentes qui discrimine le mieux les locuteurs, devrait afficher une direction inter-genre sur sa première composante, et donc une direction intra-genre sur les composantes restantes. Pour vérifier que cette LDA met en évidence une représentation démêlée de la variabilité inter- et intra-genre, nous avons analysé la colinéarité entre les composantes de lda et les DV calculées sur des valeurs de f_0 propres à chaque genre et indépendantes du genre. La figure 1 résume notre analyse, réalisée indépendamment pour chaque paramètre acoustique $\mathcal{F}$. La colinéarité des directions extraites par PCA, LR ou LDA a été évaluée à l'aide de la similarité cosinus (CS) entre $\mathsf{pca}_{\mathcal{F}}$, $\mathsf{m}_{\mathcal{F}}$, $\mathsf{m}_{\mathcal{F}|\mathsf{M}}$, $\mathsf{m}_{\mathcal{F}|\mathsf{F}}$, and lda.

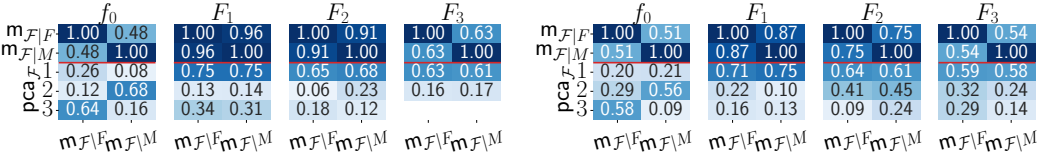
3 Résultats

3.1 Représentation multidimensionnelle des paramètres acoustiques

La première étape de notre analyse consiste à étudier séparément la représentation de chaque paramètre acoustique par les VAEs. Pour VAE-VCTK, les quatre PCA distinctes appliquées à $\mathsf{SS}_{\mathbf{z}}^{\mathcal{F}}$ ont montré que trois composantes principales (PC) sont nécessaires pour expliquer au moins $80\,\%$ de la variance de chaque base de données encodée, à l'exception de F_3 (deux PC). Dans le cas de VAE-AH, chaque $\mathcal{F}$ a besoin de cinq PCs pour expliquer $80\,\%$ de la variance. En particulier, les variances minimales expliquées par les premières PCs sont d'environ $43\,\%$ pour VAE-VCTK sur f_0 et $36\,\%$ pour VAE-AH sur f_0. Nous avons également observé pour les deux VAEs que toutes les premières PCs de paramètres acoustiques différents sont relativement orthogonaux entre eux, avec une valeur maximale de CS de 0.33 entre pca_{f_0} et pca_{F_1}. Bien que ces résultats soient similaires à ceux de Sadok *et al.* (2023), nous n'avons pas obtenu le même nombre de PCs pour expliquer $80\,\%$ de la variance et une plus grande orthogonalité entre les PCs était rapportée. Ces différences suggèrent une dépendance possible de ces analyses aux données de parole utilisées pour l'entraînement et le test.

	m_{f_0}	$m_{f_0\|F}$	$m_{f_0\|M}$	m_{F_1}	$m_{F_1\|F}$	$m_{F_1\|M}$	m_{F_2}	$m_{F_2\|F}$	$m_{F_2\|M}$	m_{F_3}	$m_{F_3\|F}$	$m_{F_3\|M}$
VAE-VCTK	**0.65**	**0.64**	**0.58**	0.37	**0.73**	**0.75**	0.40	**0.71**	**0.74**	0.32	**0.64**	**0.58**
VAE-AH	**0.61**	**0.58**	**0.53**	0.31	**0.58**	**0.61**	0.36	**0.65**	**0.67**	0.27	**0.56**	**0.52**

TABLE 1 – Score de régression R^2 pour toutes les LRs testées



(a) VAE-VCTK (b) VAE-AH

FIGURE 2 – Similarité cosinus entre $m_{\mathcal{F}|M}$, $m_{\mathcal{F}|F}$ et $pca_{\mathcal{F}}$. Pour plus de clarté, seules les trois premières composantes principales sont présentées.

Cependant, l'analyse des colinéarités entre les PCs extraites pour chaque paramètre a montré que notre VAE crée un équilibre entre les représentations latentes. D'une part, une pseudo-indépendance des paramètres de source (f_0) et de filtre ($F_{1,2,3}$) sur la première PC qui permet une modélisation séparée des variations d'intonation et d'articulation, et d'autre part une colinéarité plus forte sur les autres PCs qui laisse envisager la modélisation des covariations bien connues entre les paramètres acoustiques (Titze, 2004; Sundberg & Nordenberg, 2006). Ces résultats sont observés pour nos deux VAEs entraînés sur des données différentes, et ont également été rapportés par Sadok *et al.* (2023). Cela montre une propriété générale de la représentation multidimensionnelle des paramètres acoustiques unidimensionnelles considérées dans l'espace latent du VAE, qui sera examinée plus en détail dans la section suivante.

3.2 Interprétation des dimensions apprises

Nous avons observé que les variations de chaque paramètre acoustique, lorsqu'elles sont isolées dans l'ensemble de test de la parole synthétique, sont encodées par au moins deux directions dans l'espace latent du VAE. Notre hypothèse est que ces multiples dimensions sont nécessaires pour modéliser les variations acoustiques inter- et intra-individuelles de la parole naturelle. Ainsi, pour vérifier la fonction de chaque dimension, nous étudions maintenant les DV de chaque paramètre en contexte, c'est-à-dire sur les ensembles de tests de parole naturelle encodée $\mathsf{NS}_{\mathbf{z}}^{\mathsf{VCTK}}$ et $\mathsf{NS}_{\mathbf{z}}^{\mathsf{AH}}$ et données par $m_{\mathcal{F}}$, $m_{\mathcal{F}|M}$, et $m_{\mathcal{F}|F}$. Nous essayons alors de corréler ces directions avec celles observées sur chaque ensemble de tests de parole synthétique $\mathsf{SS}_{\mathbf{z}}^{\mathcal{F}}$ ($pca_{\mathcal{F}}$).

Le score de régression R^2 pour chaque DV est donné dans le Tab. 1, et la Fig. 2 affiche les CS entre $m_{\mathcal{F}|M}$ et $m_{\mathcal{F}|F}$ (obtenues sur la parole naturelle) et $pca_{\mathcal{F}}$ (obtenues sur la parole synthétique). Pour les deux modèles et les trois formants, nous pouvons observer un contraste entre les petites valeurs de R^2 obtenues sur les DV globales (m_{F_i}, $i \in \{1,2,3\}$) et les valeurs élevées de R^2 obtenues sur les DV en fonction du genre. Parallèlement, nous observons une CS élevée entre $m_{F_i|M}$ et $m_{F_i|F}$, pour $i \in \{1,2,3\}$ (Fig. 2). Tout cela montre que, pour les deux modèles, les valeurs de fréquence des formants sont encodées linéairement dans l'espace latent lorsque l'on considère les deux genres séparément, avec des DV assez similaires, mais des ordonnées à l'origine différentes. En outre, la première PC de pca_{F_i} est la plus corrélée avec les DV des deux genres pour les deux modèles.

En ce qui concerne f_0, le Tab. 1 montre des scores de régression élevés pour les deux modèles dans toutes les conditions (par genre et globalement). En outre, la Figure 2 montre que $\mathsf{m}_{f_0|\mathsf{M}}$ et $\mathsf{m}_{f_0|\mathsf{F}}$ sont les DVs par genre les moins corrélées entre elles, mais qu'elles sont les plus corrélées respectivement avec la deuxième et la troisième PC de pca_{f_0}. Rappelons que pca_{f_0} est calculé sur l'ensemble de test de parole synthétique $\mathsf{SS}_{\mathbf{z}}^{f_0}$, pour lequel aucun paramètre autre que f_0 ne varie dans le signal d'entrée, c'est-à-dire qu'aucune autre information sur le genre n'est disponible. Pourtant, sur cette base de données, les deux VAEs sont capables de distinguer les valeurs de f_0 qui sont plus susceptibles d'appartenir à des locuteurs masculins ou féminins. Nous supposons que les modèles ont appris la distribution bimodale des valeurs f_0 rencontrées dans leurs ensembles d'apprentissage respectifs et qu'ils sont capables de distinguer les trames synthétiques sur la base de ces distributions.

Pour tester cette hypothèse, nous avons calculé les corrélations entre la distribution de f_0, mesurée sur chaque base de données utilisée pour entraîner les VAEs, et la projection des vecteurs latents $\mathbf{z}$ dans $\mathsf{SS}_{\mathbf{z}}^{f_0}$ sur les PCs de pca_{f_0} (en résumé, les coefficients de la PCA pour f_0). La corrélation la plus élevée a été obtenue avec la première PC de pca_{f_0} pour les deux VAEs (VAE-VCTK : 0.48, VAE-AH : 0.53). Comme on peut le voir sur la Fig. 3, les deux principaux pics du profil du premier coefficient pca_{f_0} sont proches des médianes des deux modes la distribution f_0 pour les deux modèles. De plus, les valeurs de PC sont élevées pour les deux modes de la distribution de f_0, alors qu'elles sont proches de 0 ou négatives lorsque les deux modes se confondent, modélisant ainsi l'incertitude de la classification entre les trames de parole masculines ou féminines. Nous avons mené la même expérience et observé un comportement similaire sur les trois formants, pour nos deux modèles VAE. Pour chaque formant, la deuxième PC est la plus corrélée avec la distribution de fréquence des formants (valeur absolue de corrélation supérieure à 0.8 pour les deux VAEs). Dans chaque cas, la distribution est unimodale, ce qui explique de manière cohérente la corrélation d'une seule autre PC avec le DV de la valeur du formant, comme démontré précédemment (Fig. 2).

Dans l'ensemble, nous avons montré que la représentation multidimensionnelle d'un paramètre acoustique unique est étroitement liée à la multimodalité de la distribution du paramètre. Pour chaque paramètre, nous avons constaté qu'une PC encode la distribution du paramètre qui est apprise à partir de l'ensemble d'apprentissage, et que les variations de paramètres acoustiques spécifiques à chaque mode sont encodées par quelques autres PCs distinctes.

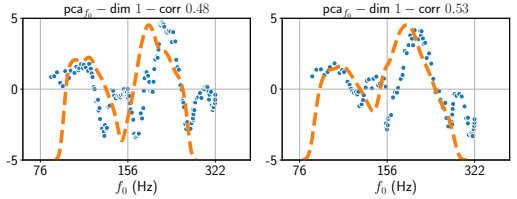


FIGURE 3 – Distribution des valeurs f_0 sur les données d'entraînement (en orange) et projection de $\mathsf{SS}_{\mathbf{z}}^{\mathsf{VCTK}}$ (gauche) et $\mathsf{SS}_{\mathbf{z}}^{\mathsf{AH}}$ (droite) sur la composante PCA la plus corrélée (en bleu).

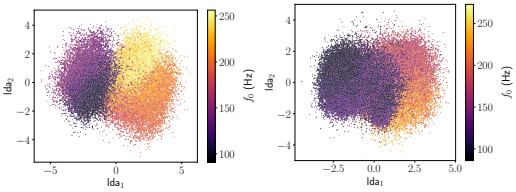


FIGURE 4 – Projection de $\mathsf{NS}_{\mathbf{z}}^{\mathsf{VCTK}}$ (gauche) et $\mathsf{NS}_{\mathbf{z}}^{\mathsf{AH}}$ (droite) sur les deux premières composantes de leurs lda respectifs, colorés en fonction de leurs valeurs f_0.

3.3 Modélisation des variations inter- vs. intra-genre

Nous avons observé que, pour nos deux modèles VAE, les valeurs f_0 de la parole des hommes et des femmes sont encodées dans des directions distinctes, mais non orthogonales de l'espace latent (voir Fig. 2). En revanche, pouvons-nous identifier des directions orthogonales qui modélisent

les variations inter- et intra-genre, afin de fournir une représentation latente mieux démêlée de la variabilité liée au genre ? Comme indiqué dans la Section 2.3, les candidats appropriés pour de telles directions sont la première et deuxième composante de lda, une LDA sur les locuteurs calculée sur l'espace latent de chaque VAE ($\mathsf{NS}_{\mathbf{z}}^{\mathsf{VCTK}}$ et $\mathsf{NS}_{\mathbf{z}}^{\mathsf{AH}}$ pour VAE-VCTK et VAE-AH, respectivement). Pour valider cette hypothèse, nous avons mesuré la colinéarité des composantes de lda avec les DV des représentations de f_0 globale (m_{f_0}) et selon le genre ($\mathsf{m}_{f_0|\mathsf{M}}$ et $\mathsf{m}_{f_0|\mathsf{F}}$). Les CS mettent en évidence une forte corrélation (supérieure à 0.85 pour les deux VAE) entre la première composante lda et m_{f_0}, qui comprend des informations sur le genre. Par ailleurs, la deuxième composante lda est bien corrélée avec les DV calculées par genre, respectivement 0.68 avec $\mathsf{m}_{f_0|\mathsf{M}}$ et 0.61 avec $\mathsf{m}_{f_0|\mathsf{F}}$.

Ces résultats sont cohérents avec notre hypothèse selon laquelle les informations intra- et inter-genres sont encodées suivant des directions LDA distinctes. Pour illustrer cette analyse, la Fig. 4 représente la position des trames des signaux encodés $\mathsf{NS}_{\mathbf{z}}^{\mathsf{VCTK}}$ (à gauche) et de $\mathsf{NS}_{\mathbf{z}}^{\mathsf{AH}}$ (à droite) selon leurs deux premières composantes lda respectives. Nous observons que les trames sont regroupées en deux groupes le long de la première composante, les trames des locuteurs masculins (violet, f_0 faible) et féminins (orange-jaune, f_0 élevé) étant associées à des valeurs négatives et positives, respectivement. La seconde composante lda modélise la variation intra-genre de f_0. Dans l'ensemble, ces résultats mettent en évidence la capacité du VAE à démêler les variations inter- et intra-genre le long de deux directions distinctes de l'espace latent que nous avons identifiées grâce à l'analyse LDA, et ceci est observé pour deux ensembles de données de parole différents. Les faibles CS entre ces deux directions (< 0.35) sont prometteuses pour imaginer un contrôle indépendant de f_0 entre les classes de genre et à l'intérieur de celles-ci.

4 Conclusion

Nous avons introduit une méthodologie pour analyser l'espace latent du VAE entraîné sur une base de données multilocuteurs en combinant l'utilisation d'ensembles de données de test synthétiques et naturelles, et l'extraction et la comparaison des directions qui expliquent le mieux la variation des paramètres acoustiques sélectionnés. Après avoir montré que la variation de chaque paramètre est encodée par de multiples dimensions dans l'espace latent, nous avons démontré que l'une de ces dimensions encode la forme globale de la distribution des paramètres sur l'ensemble d'apprentissage. Dans le cas de f_0, la distribution est bimodale et les valeurs de f_0 appartenant à différents modes sont encodés sur des dimensions distinctes supplémentaires. Dans ce cas, nous avons identifié des directions dans l'espace latent du VAE qui expliquent les variations inter- et intra-genre de f_0. Pour valider notre approche, nous avons mené nos expériences sur deux ensembles de données (VCTK et Att-HACK) qui diffèrent en termes de langue (anglais vs. français) et de style (lecture/narration vs. jeu d'acteur/expression).

Alors que plusieurs études ont utilisé la réduction de la dimension de l'espace latent (Tits *et al.*, 2019; Dieck *et al.*, 2022), ont abordé l'orthogonalité des différentes directions qui expliquent un paramètre donné (Hsu *et al.*, 2017; Sadok *et al.*, 2023), ou identifié la variation de paramètres acoustiques dans l'espace latent par régression linéaire (Sadok *et al.*, 2023; Vaidya *et al.*, 2022; Lenglet *et al.*, 2022a), ce travail est l'un des rares à tenter d'interpréter la représentation multidimensionnelle de chaque paramètre acoustique unidimensionnel. Dans de futurs travaux, nous visons à augmenter le nombre de paramètres d'intérêt et à utiliser notre méthode pour contrôler la variation des paramètres acoustiques dans l'espace latent des modèles non- ou auto-supervisés.

Références

ANIKIN A. (2019). Soundgen : an open-source tool for synthesizing nonverbal vocalizations. *Behavior Research Methods*, **51**, 778–792.

BAEVSKI A., ZHOU Y., MOHAMED A. & AULI M. (2020). wav2vec 2.0 : A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, p. 12449–12460, virtual conf.

BLAAUW M. & BONADA J. (2016). Modeling and transforming speech using variational autoencoders. In *Proc. of Interspeech*, p. 1770–1774, San Francisco, CA, USA.

BOERSMA P. & WEENINK D. (2001). Praat, a system for doing phonetics by computer. *Glot International*, **5**(9-10), 341–347.

BOUS F. & ROEBEL A. (2022). A bottleneck auto-encoder for F0 transformations on speech and singing voice. *Information*, **13**(3), 102–121.

CHOU J.-C., YEH C.-C., LEE H.-Y. & LEE L.-S. (2018). Multi-target voice conversion without parallel data by adversarially learning disentangled audio representations. In *Proc. of Interspeech*, p. 501–505, Hyderabad, India.

COLEMAN R. O. (1971). Male and female voice quality and its relationship to vowel formant frequencies. *J. of speech and hearing research*, **14**(3), 565–577.

DIECK T. T., PÉREZ-TORO P. A., ARIAS T., NOETH E. & KLUMPP P. (2022). Wav2vec behind the scenes : How end2end models learn phonetics. In *Proc. of Interspeech*, p. 5130–5134, Incheon, Korea.

FANT G. (1971). *Acoustic theory of speech production*. Mouton.

GIRIN L., ROCHE F., HUEBER T. & LEGLAIVE S. (2019). Notes on the use of variational autoencoders for speech and audio spectrogram modeling. In *Int. Conf. on Digital Audio Effects*, p. 1–8, Birmingham, UK.

HIGGINS I., MATTHEY L., PAL A., BURGESS C., GLOROT X., BOTVINICK M., MOHAMED S. & LERCHNER A. (2017). β-VAE : Learning basic visual concepts with a constrained variational framework. In *Int. Conf. on Learning Representations*, Toulon, France.

HSU W.-N., BOLTE B., TSAI Y.-H. H., LAKHOTIA K., SALAKHUTDINOV R. & MOHAMED A. (2021). Hubert : Self-supervised speech representation learning by masked prediction of hidden units. *Trans. on Audio, Speech, and Language Processing*, **29**, 3451–3460.

HSU W.-N., ZHANG Y. & GLASS J. (2017). Learning latent representations for speech generation and transformation. In *Proc. of Interspeech*, p. 1273–1277, Stockholm, Sweden.

JACQUELIN M., GARNIER M., GIRIN L., VINCENT R. & PERROTIN O. (2024). Exploring the multidimensional representation of unidimensional speech acoustic parameters extracted by deep unsupervised models. In *ICASSP Workshop XAI-SA*, Seoul, Korea.

KINGMA D. P. & BA J. (2015). Adam : A method for stochastic optimization. In *Int. Conf. on Learning Representations*, San Diego, USA.

KINGMA D. P. & WELLING M. (2014). Auto-encoding variational bayes. In *Int. Conf. on Learning Representations*, Banff, Canada.

LAKHOTIA K., KHARITONOV E., HSU W.-N., ADI Y., POLYAK A., BOLTE B., NGUYEN T.-A., COPET J., BAEVSKI A., MOHAMED A. & DUPOUX E. (2021). On generative spoken language modeling from raw audio. *Trans. of the Association for Computational Linguistics*, **9**, 1336–1354.

LE MOINE C. & OBIN N. (2020). Att-hack : An expressive speech database with social attitudes. In *Speech Prosody*.

LENGLET M., PERROTIN O. & BAILLY G. (2022a). Modélisation de la parole avec tacotron2 : Analyse acoustique et phonétique des plongements de caractere. In *Journées d'Études sur la Parole*, p. 788–796, Noirmoutier, France.

LENGLET M., PERROTIN O. & BAILLY G. (2022b). Speaking rate control of end-to-end TTS models by direct manipulation of the encoder's output embeddings. In *Proc. of Interspeech*, p. 11–15, Incheon, Korea.

POLYAK A., ADI Y., COPET J., KHARITONOV E., LAKHOTIA K., HSU W.-N., MOHAMED A. & DUPOUX E. (2021). Speech resynthesis from discrete disentangled self-supervised representations. In *Proc. of Interspeech*, p. 3615–3619, Brno, Czechia.

SADOK S., LEGLAIVE S., GIRIN L., ALAMEDA-PINEDA X. & SÉGUIER R. (2023). Learning and controlling the source-filter representation of speech with a variational autoencoder. *Speech Comm.*, **148**, 53–65.

SUNDBERG J. & NORDENBERG M. (2006). Effects of vocal loudness variation on spectrum balance as reflected by the alpha measure of long-term-average spectra of speech. *The J. of the Acoust. Soc. of Am.*, **120**(1), 453–457.

TITS N., WANG F., EL HADDAD K., PAGEL V. & DUTOIT T. (2019). Visualization and interpretation of latent spaces for controlling expressive speech synthesis through audio analysis. In *Proc. of Interspeech*, p. 4475–4479, Graz, Austria.

TITZE I. R. (2004). A theoretical study of F0-F1 interaction with application to resonant speaking and singing voice. *J. of Voice*, **18**(3), 292–298.

VAIDYA A. R., JAIN S. & HUTH A. (2022). Self-supervised models of audio effectively explain human cortical responses to speech. In *Int. Conf. on Machine Learning*, p. 21927–21944, Baltimore, USA.

VAN DER OORD A., VINYALS O. *et al.* (2017). Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, p. 6306–6315, Long Beach, USA.

WANG Y., STANTON D., ZHANG Y., RYAN R.-S., BATTENBERG E., SHOR J., XIAO Y., JIA Y., REN F. & SAUROUS R. A. (2018). Style tokens : Unsupervised style modeling, control and transfer in end-to-end speech synthesis. In *Int. Conf. on Machine Learning*, p. 5180–5189, Stockholm, Sweden.

WILLIAMS J., FONG J., COOPER E. & YAMAGISHI J. (2021). Exploring disentanglement with multilingual and monolingual VQ-VAE. In *ISCA Speech Synthesis Workshop*, p. 124–129, Budapest, Hungary.

WILLIAMS J. & KING S. (2019). Disentangling style factors from speaker representations. In *Proc. of Interspeech*, p. 3945–3949, Graz, Austria.

YAMAGISHI J., VEAUX C. & MACDONALD K. (2019). CSTR VCTK corpus : English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92).

ZHANG Y.-J., PAN S., HE L. & LING Z.-H. (2019). Learning latent representations for style control and transfer in end-to-end speech synthesis. In *Int. Conf. on Acoustics, Speech and Signal Processing*, p. 6945–6949, Brighton, UK.