Vérification automatique de la voix de locuteurs après resynthèse à l'aide de PPG

Thibault Gaudier, Marie Tahon, Anthony Larcher, Yannick Estève

Vérification automatique de la voix de locuteurs après conversion à l'aide de PPGs

Thibault Gaudier[1, 2] Marie Tahon[1] Anthony Larcher[1] Yannick Estève[2]
(1) Laboratoire d'Informatique de l'Université du Mans (LIUM), 72100 Le Mans, France
(2) Laboratoire d'Informatique d'Avignon (LIA), 83000 Avignon, France
{prenom}.{nom}@univ-lemans.fr

RÉSUMÉ

La création de contenu journalistique peut être assistée par des outils technologiques comme la synthèse de parole. Cependant l'éditeur doit avoir la possibilité de contrôler la génération du contenu audio comme la prosodie, la prononciation ou le contenu linguistique. Dans ces travaux, un système de conversion de voix génère un signal de locuteur cible à partir d'une représentation temporelle de type Phonetic PosteriorGrams (PPGs) extraite d'un audio source. Les PPGs dêmelent le contenu phonétique du contenu rythmique, et sont généralement considérés indépendants du locuteur. Cet article présente un système de conversion utilisant les PPGs, et son évaluation en qualité audio avec un test perceptif. Nous montrons également qu'un système de vérification du locuteur ne parvient pas à identifier le locuteur source après la conversion, même si le modèle a été entraîné sur des données synthétiques.

ABSTRACT

Automatic Speaker's Voice Verification after Speech Conversion using PPGs

The creation of journalistic content can be assisted by technologies such as speech synthesis. In any cas, the editor needs the possibility to control the audio content generation such as prosody, pronunciation or linguistics. In the present work, a voice conversion system generates a target speaker signal from a temporal representation, using Phonetic PosteriorGrams (PPGs) extracted in the source audio. This representation disentangles rhythmic and phonetic information, and is usually considered speaker-independant. This paper presents a PPGs-based speech conversion system, and its evaluation in terms of general quality. We also demonstrate that a speaker verification model is not able to recover the source speaker after conversion with PPGs, even when the model is trained on synthetic data.

MOTS-CLÉS : synthèse de parole, représentation interprétable de la parole, reconnaissance du locuteur.

KEYWORDS: speech synthesis, interpretable speech representation, speaker recognition.

1 Introduction

Les journalistes et médias ont désormais accès à des flux importants de contenu, venant de différents endroits du monde et dans différentes langues. Dans ce contexte, le projet européen SELMA [1] vise à développer des outils permettant de réaliser du doublage automatique, en générant un signal de parole d'une voix cible à partir d'un texte éventuellement traduit. Une manière de faire serait d'utiliser un

1. https://selma-project.eu/

système de synthèse à partir de texte (TTS) afin de générer le signal correspondant au texte traduit. Cependant, malgré les avancées récentes du domaine, le signal ainsi synthétisé ne correspond pas nécessairement aux besoins des utilisateurs. Il y a donc la nécessité d'avoir des systèmes de conversion de parole permettant un contrôle plus fin de la parole générée selon différents aspects, en utilisant des représentations interprétables. Par exemple, EdiTTS (Tae *et al.*, 2022) utilise le texte comme représentation afin de modifier le contenu linguistique, et (Zhao *et al.*, 2019) utilise les Phonetic PosteriorGrams (PPG) comme représentation permettant de contrôler les contenus rythmique et phonétique.

Les PPGs sont une représentation temporelle des probabilités de présences de différentes unités phonétiques. Ainsi, il est possible, techniquement, de modifier les durées, sans changer les phonèmes (et réciproquement) donnant ainsi du contrôle aux utilisateurs (Zhao *et al.*, 2019) and (Yeh *et al.*, 2018). Evidemment, cette modification est artificielle et ne pourra pas fournir de parole audible si le modèle utilisé pour la conversion ne compense pas certaines modifications (par exemple allonger la durée d'une plosive). Cet aspect sera étudié dans un futur proche.

Les PPGs ont déjà été utilisés pour la tâche de conversion de voix (Levy-Leshem & Giryes, 2021). Cette représentation contient également des informations relatives à l'accent du locuteur, permettant de réaliser de la conversion d'accent (Zhao *et al.*, 2019) (modifier l'accent sans changer la voix). Cependant, cela signifie que certaines informations relatives au locuteur, présentes dans les PPGs, pourraient passer dans le signal audio généré à partir de cette représentation. À long terme, notre objectif est de synthétiser un signal de parole (source) à partir de texte, puis de convertir ce signal vers une voix cible à partir de PPGs corrigés manuellement. Nous cherchons donc à vérifier que le locuteur de l'audio cible ne peut pas être identifié à partir de l'audio source, que cette source soit naturelle (comme présenté ici) ou bien synthétique (comme dans notre objectif à long terme).

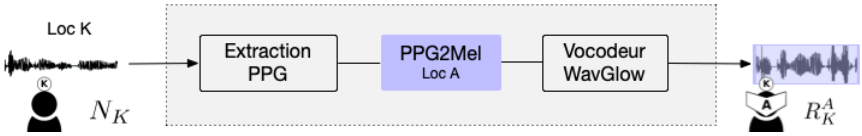


FIGURE 1 – Approche de synthèse utilisant les PPG (PPG2Mel). Les éléments en bleu sont spécifiques à un locuteur. N est un échantillon naturel, R un échantillon resynthétisé (voir Section 4.1)

Pour cela, nous entraînons différents modèles de conversion à partir de PPGs, de manière similaire à (Zhao *et al.*, 2019) and (Levy-Leshem & Giryes, 2021). Plus précisément, comme illustré Figure 1, un système PPG2Mel est entraîné à générer une voix A à partir de PPGs. Ainsi, à l'inférence, lorsqu'un signal de parole N_K provenant du locuteur source K passe dans le système, le signal synthétisé R_K^A est converti vers la voix cible A. Nous réalisons une évaluation subjective de la qualité de la parole générée, comparant notre signal généré à celui obtenu par un système TTS, un signal obtenu par un vocodeur et le signal naturel. L'objectif de cette évaluation, décrite dans la Section 3, est de s'assurer de générer de la parole de qualité convenable à partir de PPGs. Comme notre but n'est pas de proposer un nouveau système de synthèse mais d'en utiliser un existant pour une autre tâche, nous ne cherchons pas à comparer ce système à d'autres à partir d'un score de qualité.

À notre connaissance, aucune étude n'a tenté d'identifier une éventuelle information relative au locuteur original après conversion à partir de PPGs. Notre contribution principale est l'étude de la capacité d'un système de vérification du locuteur (SV) à identifier le locuteur original d'un audio synthétisé. Ceci est réalisé en synthétisant des audios de différents locuteurs source (base de données

VoxCeleb (Nagrani *et al.*, 2017; Chung *et al.*, 2018)) en utilisant deux modèles PPG2Mel entraînés sur deux voix cibles différentes. La Section 4.1 détaille le contenu des bases de données utilisées, leur usage et les notations utilisées. À partir de cette base de données synthétique annotée avec les locuteurs sources, nous pouvons entraîner des systèmes de SV de plusieurs manières. La Section 4.2 détaille notre protocole et la Section 4.3 montre les résultats obtenus. Contrairement au protocole d'évaluation de conversion de voix, nous ne cherchons pas à identifier la similarité entre les audios synthétiques et les locuteurs cibles, mais à vérifier si un système de vérification du locuteur peut identifier les locuteurs sources à partir d'audio synthétique, malgré la présence de la voix cible.

2 Synthèse de parole

Les tâches de synthèse de parole sont souvent divisées en deux étapes : la prédiction d'une représentation fréquentielle (mel-spectrogramme par exemple) à partir de l'entrée, puis l'utilisation d'un vocodeur pour obtenir le signal audio correspondant.

De nos jours, les systèmes de synthèse à partir de texte sont principalement des systèmes neuronaux autorégressifs comme Tacotron2 (Shen *et al.*, 2018), ou des systèmes séquence-vers-séquence, par exemple utilisant des Transformers comme FastSpeech (Ren *et al.*, 2019). L'introduction de contrôle dans ces systèmes est généralement réalisée en conditionnant la génération de parole à un locuteur spécifique (Cooper *et al.*, 2020) and (Valle *et al.*, 2020). Cependant, le contrôle pour d'autres aspects comme la prosodie (Sini *et al.*, 2020), l'intonation (Łańcucki, 2021), le style (Wang *et al.*, 2018) ou l'émotion (Diatlova & Shutov, 2023) ont été introduits dans les systèmes de synthèse.

Le but des système de conversion de voix est généralement de préserver certains aspects provenant d'un audio source (par exemple le contenu linguistique) et d'autres provenant d'un audio cible (les indices acoustiques relatifs à un locuteur). L'utilisation de représentation démêlées (principalement entre locuteur, contenu linguistique et/ou prosodie) extraites du signal cible ont été utilisées par (Qian *et al.*, 2019, 2020), (Polyak *et al.*, 2021). Les PPGs ont également été utilisés pour la conversion de voix (Levy-Leshem & Giryes, 2021), d'accent (Zhao *et al.*, 2019) ou de rythme (Yeh *et al.*, 2018).

Comme beaucoup de systèmes de synthèse génèrent des mel-spectrogrammes à partir de la consigne, le rôle du vocodeur est alors de produire le signal audio correspondant dans le domaine temporel. Aujourd'hui les vocodeurs mainstream sont basés sur des réseaux génératifs comme HifiGan (Kong *et al.*, 2020), ou WaveGlow (Prenger *et al.*, 2019).

3 Synthèse à partir de Phonetic PosteriorGrams (PPG)

3.1 Les Phonetic PosteriorGrams (PPG)

Les Phonetic PosteriorGrams (PPG) (exemple Figure 2) sont une représentation temporelle et probabiliste des phonèmes prononcés dans un audio. Ainsi, pour chaque trame de 30 ms, on obtient la probabilité de présence de chacun des phonèmes. Cette représentation présente certains avantages : elle permet à un utilisateur de contrôler finement l'audio représenté selon certaines caractéristiques, comme le contenu phonétique ou le rythme de parole. Cette représentation contient également une information plus riche qu'une séquence de phonème, car la confusion entre plusieurs classes de phonèmes, qui s'observe par une probabilité non nulle pour différentes classes dans une même trame, peut être interprétée comme une différence de réalisation de ces phonèmes. Ainsi, les PPG contiennent des informations relatives à la phonétique de la phrase et au rythme de celle-ci, mais d'autres informations pourraient être également présentes de manière cachée. L'expérience présentée dans cet article cherche à identifier une éventuelle présence d'information relative au locuteur dans un PPG.

Les PPG sont extraits à partir du modèle présenté dans (Zhao *et al.*, 2019), qui est un Generalized Maxout Network fourni par Kaldi (Zhang *et al.*, 2014), entraîné à imiter un GMM-HMM représentant 5816 unités acoustiques, qui sont ensuite regroupées en 40 phonèmes pour l'anglais. 100 trames de PPG sont extraites chaque seconde.

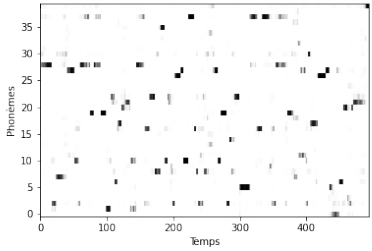


FIGURE 2 – Exemple de PPG pour la phrase : "Such risks can be lessened when the President recognizes the security problem"

3.2 Synthèse de parole à partir de PPG

Les PPG sont une représentation présentant des similitudes avec un encodage one-hot des phonèmes. Ainsi, il est possible de générer un audio correspondant à un PPG de la même manière que pour d'autres représentations de la parole telle une séquence de phonème. Nous avons utilisé une approche similaire à celle présentée dans (Zhao *et al.*, 2019) et (Levy-Leshem & Giryes, 2021). Nous avons donc entraîné Tacotron2 (Shen *et al.*, 2018) en utilisant les PPG extraits des audios en entrée pour prédire les mel-spectrogrammes extraits de ces mêmes audios. Ce système sera désigné par PPG2Mel. Le modèle converge plus rapidement en utilisant des PPG qu'en utilisant du texte, car les PPG sont déjà alignés temporellement avec l'audio. La seule modification que nous avons faite à l'architecture de Tacotron2 est le remplacement de la couche d'embedding de caractères par une couche linéaire transformant les 40 probabilités de présence des phonèmes en une représentation interne de dimension 512.

Comme le système PPG2Mel produit des mel-spectrogrammes, nous devons ensuite les convertir dans le domaine audio. Nous avons utilisé WaveGlow, un vocodeur neuronal décrit dans (Prenger *et al.*, 2019). Notre vocodeur est entraîné sur le dataset LJSpeech (Ito & Johnson, 2017) en utilisant la configuration par défaut à l'exception de la fréquence d'échantillonage, que nous avons passée de 22,05kHz à 16kHz. Nous avons utilisé l'implémentation proposée par Nvidia, disponible sur GitHub [2]

3.3 Évaluation perceptive de la parole

Notre objectif avec cette évaluation est de s'assurer que le système présenté dans la Section 3.2 génère des échantillons audios de qualité suffisante pour étudier la présence d'information relative au locuteur dans la synthèse. Nous avons donc comparé la qualité du système PPG2mel avec une version obtenue par TTS, une version obtenue en utilisant uniquement le vocodeur (analyse-synthèse), ainsi que l'échantillon naturel original. Les systèmes TTS et PPG2Mel ont été entraînés sur la base de données LJSpeech (Ito & Johnson, 2017). Notre baseline TTS utilise l'implémentation de Tacotron2 par Nvidia [3], en changeant uniquement la fréquence d'échantillonage à 16kHz afin de rester consistant avec les autres échantillons. Nous avons utilisé les ensembles d'entraînement, validation et test provenant du même dépôt GitHub. Nous avons ensuite divisé l'ensemble de test en

2. https://github.com/nvidia/waveglow
3. https://github.com/nvidia/tacotron2

trois sous-ensembles en fonction de la durée des audios. 20 segments sont sélectionné dans chaque sous-ensemble afin d'avoir une représentation des audios courts, moyens et longs. Un test Mean Opinion Score (MOS) (1 : mauvais, 5 excellent) a été mis en place avec la plateforme FlexEval (Fayet *et al.*, 2020) à partir de la question suivante : "jugez la qualité de cet échantillon audio". 36 participants sur 44, majoritairement non natifs, ont évalué l'ensemble des 20 segments audios qui leur était présentés. Les 4 versions des 60 segments ont été évalués en moyenne 12 fois.

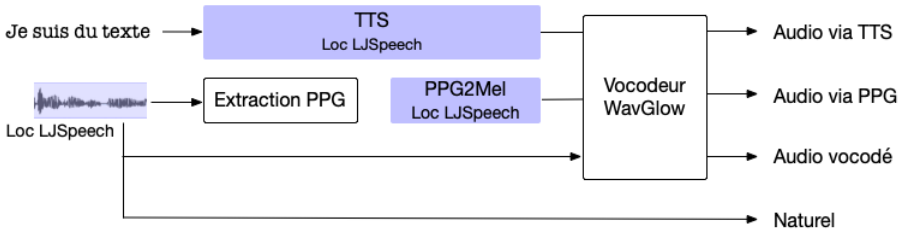


FIGURE 3 – Représentation des 4 versions de l'audio présentées lors du test perceptif. Les blocs bleus sont spécifiques à un locuteur

Les résultats sont décrits dans la Table 1. Nous avons retiré les étapes d'introduction et utilisé toutes les autres réponses, dont celles venant de participants n'ayant pas complété toutes les étapes. À partir de ces résultats, nous pouvons conclure que l'utilisation de PPG pour la synthèse de parole ne dégrade pas la qualité par rapport à la synthèse à partir de texte. Nous observons également qu'une partie importante de la dégradation en qualité provient du vocodeur. Cela peut s'expliquer par le fait que le vocodeur est biaisé par le locuteur de LJspeech. Pour améliorer ce point, il pourrait être intéressant de fine-tuné le vocodeur sur nos données. Nous sommes conscients que nos résultats de MOS sont inférieurs à ceux de la littérature, peut-être parce qu'ils sont non-natifs. Cependant, nous observons également que l'audio naturel n'est pas non plus évalué avec d'aussi bons scores.

TABLE 1 – MOS obtenus lors de l'évaluation. Intervalles de confiance à 95%

Système	Audio naturel	Audio vocodeur	Audio TTS	Audio PPG
MOS	4.35 ± 0.07	3.47 ± 0.07	3.11 ± 0.07	3.24 ± 0.07

4 Identification du locuteur source

Dans cette section, notre objectif est de déterminer si un système de vérification naïf, entraîné sur de la parole naturelle, peut identifier le locuteur source après conversion (Q1). Ensuite, nous utilisons des données synthétiques pour entraîner un modèle informé à identifier ce locuteur source. Nous souhaitons savoir à quel point ce modèle informé parvient à identifier le **locuteur source** dans des échantillons convertis vers la voix cible, mais aussi à partir d'échantillons naturels afin d'identifier le décalage entre parole naturelle et synthétique (Q2). Le modèle informé est supposé apprendre à différencier les locuteurs dans un espace adapté à la voix cible. Si la conversion cache complètement le locuteur source, on s'attend à une forte dégradation des résultats avec les deux modèles (naïf et informé). En revanche, si elle ne cache que partiellement le locuteur source, le modèle naïf devrait obtenir de mauvais résultats, mais le modèle informé devrait parvenir à identifier le locuteur source, et donc obtenir de meilleurs scores. Enfin, nous étudions à quel point les modèles parviennent à lier l'identité du **locuteur cible** provenant des échantillons naturels avec les échantillons convertis vers cette même voix cible (Q3).

4.1 Données et notations

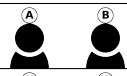
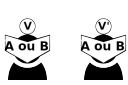
Cette expérience utilise 3 bases de données. La première est la section anglaise de M-AILABS(Solak, 2019), un corpus basé sur LibriVox. Nous avons utilisé 2 locuteurs, E. Klett, notée A, et E. Miller, noté B. Pour chaque locuteur A et B, nous avons 30 à 45 heures de parole, que nous avons divisé en entraînement, validation et test. Au cours de cette expérience, ces deux voix ont servi comme **locuteurs cibles**. Ceci signifie que tous les échantillons synthétiques ont été générés avec l'une de ces voix. Nous avons entraîné deux modèles mono-locuteurs notés PPG2Mel$_A$ et PPG2Mel$_B$, pour les locuteurs A et B (voir Figure 1).

La base de données LibriSpeech-test-clean (Panayotov *et al.*, 2015) est notre base d'enrôlement et de test pour l'expérience de vérification du locuteur. Elle contient 40 locuteurs, équilibrés en terme de genre ($\simeq$ 8 min. de parole par locuteur), notés 1 à 40. Ces locuteurs sont les **locuteurs source** que nous voulons identifier avant et après synthèse. Nous créons deux versions synthétiques de cette base en utilisant les modèles PPG2Mel$_A$ et PPG2Mel$_B$.

Enfin, les bases de données VoxCeleb1&2 (Nagrani *et al.*, 2017; Chung *et al.*, 2018) sont utilisées pour entraîner les systèmes de vérification du locuteur. Les modèles PPG2Mel$_A$ et PPG2Mel$_B$ sont utilisés pour synthétiser tous les échantillons de VoxCeleb vers les **locuteurs cibles**, choisis aléatoirement entre A et B pour chaque échantillon. Les labels de locuteur pour l'apprentissage des modèles sont conservés à l'identique, même si la voix perçue est maintenant différente.

Les échantillons naturels sont notés N_{source}, où $source$ est dans $\{A, B, 1-40\}$. Les échantillons synthétiques sont notés R^{cible}_{source}, où $source$ est identique à précédemment et $cible$ est A ou B selon le modèle PPG2Mel utilisé. Nous ne mentionnerons pas les locuteurs de VoxCeleb. Une illustration des différents cas est présente à la Table 2.

TABLE 2 – Description des notations des différents locuteurs et échantillons synthétiques. Cercle : locuteurs sources ; masques : signal synthétisé en utilisant le modèle lié à la voix indiquée sur le masque. K et K' sont considérés différents.

Données	Locuteurs	Notation	Détails
M-AILABS	4		N_A et N_B. A est E. Klett, B est E. Miller
LibriSpeech-test naturel	40		N_K, $N_{K'}$. $K, K' \in [\![1, 40]\!]$, $K \neq K'$
LibriSpeech-test synthétique	40		$R^A_{K'}$, R^B_K A, B, K, K' décrits précédemment
VoxCeleb1&2	7363		Utilisé pour entraîner le modèle naïf
VoxCeleb1&2 synthétique	7363		Utilisé pour entraîner le modèle informé Synthétisé en utilisant les locuteurs A et B de M-AILABS.

4.2 Modèles de vérification du locuteur

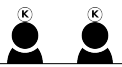
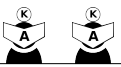
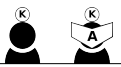
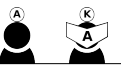
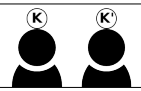
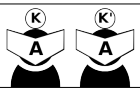
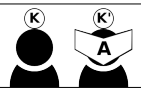
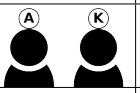
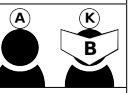
Les modèles de vérification naïf et informé utilisent l'architecture ECAPA-TDNN (Desplanques *et al.*, 2020), avec comme entrée des représentations de l'audio extraites par WavLM-Large [4] et en utilisant l'Additive Angular Margin comme fonction de coût. Nous utilisons des x-vecteurs de dimension 256. Le modèle naïf est entraîné sur les données naturelles de VoxCeleb1&2. Ce modèle obtient un

4. https://github.com/microsoft/unilm/tree/master/wavlm

EER de 1.57% sur VoxCeleb-o après 4 jours d'entraînement sur une carte GPU RTX8000, ce qui est légèrement inférieur à l'état de l'art actuel sur ces données. Le modèle informé est entraîné sur la version synthétique (locuteurs cibles A et B) de VoxCeleb1&2 décrite plus haut. La meilleure version de ce modèle est obtenue après un jour d'entraînement sur la même carte et obtient un EER de seulement 20% sur la version synthétique de VoxCeleb-o.

4.3 Expériences et résultats

TABLE 3 – Définition des expériences par leurs cibles et imposteurs, et taux d'égale erreur (EER) des modèles naïf et informé. Pour chaque expérience, les tests sont définis comme *enrôlement/test*.

Expérience	(1)	(2)	(3)	(4)	(5)
Définition de la cible	N_{1-40}/N_{1-40}	R^A_{1-40}/R^A_{1-40}	N_{1-40}/R^A_{1-40}	N_A/R^A_{1-40}	N_A/R^A_{1-40}
Définition de la cible					
Définition des imposteurs	$N_{1-40}/N_{\overline{1-40}}$	$R^A_{1-40}/R^A_{\overline{1-40}}$	$N_{1-40}/R^A_{\overline{1-40}}$	N_A/N_{1-40}	N_A/R^B_{1-40}
Définition des imposteurs					
Modèle naïf	1.98 %	49.46 %	48.02 %	45.13 %	49.81%
Modèle informé	29.44 %	49.80 %	49.00 %	33.58 %	45.52%

La Table 3 résume les différentes cibles, imposteurs et les résultats obtenus pour 5 expériences de vérification. Les expériences sont décrites par leurs paires enrôlement/test. Un test N_{1-40}/N_{1-40} compare les couples d'échantillons naturels d'un même locuteur parmi 1 à 40, par exemple N_1 et N_1. Un test $N_{1-40}/R^A_{\overline{1-40}}$ compare les échantillons naturels de chaque locuteur 1 à 40 avec les échantillons synthétisés des autres locuteurs, par exemple N_1 avec R^A_2. Chaque expérience donne un EER pour les modèles naïf et informé. Pour les expériences (1), (2) et (3), les labels relatifs au locuteur correspondent au locuteur source parmi $1-40$, tandis que pour les expériences (4) et (5) les labels sont ceux des locuteurs cibles A et B.

La première expérience (1) permet de s'assurer que notre modèle naïf obtient des résultats corrects. Pour cela, nous voulons identifier le locuteur à partir d'audio naturel. Comme attendu, le modèle naïf obtient un bon résultat (EER=1.98%) puisque il s'agit de la tâche d'entraînement de ce modèle. Le modèle informé induit une forte dégradation (EER=29.44%), ce qui indique une différence de domaine entre les données d'entraînement de ce modèle et ce test.

Dans l'expérience (2), on compare les versions converties des audios venant des locuteurs sources $1-40$ avec le modèle PPG2Mel$_A$ (R^A_{1-40}) entre eux, afin de voir si les modèles parviennent à lier les échantillons provenant d'un même locuteur source. Les résultats montrent qu'aucun des modèles ne parvient à réaliser cette tâche (EER $> 49\%$). Cela permet de répondre à la question Q1 : le modèle naïf ne parvient pas à reconnaître des échantillons provenant d'un même locuteur source après synthèse. Une hypothèse est que l'identité du locuteur source a été cachée après conversion. On observe que le modèle informé reconnaît mieux les locuteurs dans l'espace naturel (EER$= 29.44\%$, exp (1)) que dans l'espace synthétique (EER$= 49.80\%$, exp (2)). Durant son entraînement, le modèle informé a

peu convergé, mais il semble que le peu d'éléments discriminants appris permettent uniquement de distinguer les locuteurs dans l'espace naturel, cette tâche étant plus facile. Ceci permet de répondre à la question Q2 : même un modèle informé ne parvient pas à reconnaître le locuteur source après conversion.

L'expérience (3) évalue la capacité des deux modèles à lier un même locuteur dans l'espace naturel et dans l'espace synthétique. Pour cela, nous utilisons les données de LibriSpeech décrites précédemment comme enrôlement, et les versions converties de cette même base utilisant PPG2Mel$_A$ comme données de test. On observe qu'aucun de nos modèles ne parvient à identifier les 40 locuteurs entre ces deux espaces. On peut en conclure que l'approche utilisant les PPGs pour la conversion permet bien de cacher le locuteur source à des modèles de vérification du locuteur, même appris sur des données synthétiques. Les éventuels indices acoustiques permettant d'identifier le locuteur source ne sont pas détectés après la resynthèse.

Les expériences (4), respectivement (5), mesurent la proximité entre les locuteurs source $1-40$ convertis vers la voix A et leur version naturelle (resp. et leur version convertie vers la voix B) par rapport à la proximité avec les échantillons naturels du locuteur A. On conclut de l'expérience (4) que l'identité des échantillons des locuteurs source $1-40$ covnertis avec le modèle PPG2Mel$_A$ ne correspondent pas à l'identité de A, ce qui confirme le fait que le système de conversion ne parvient par à rapprocher le locuteur source du locuteur A. Cependant, les résultats montrent que les échantillons convertis sont plus proches des échantillons naturels de A selon le modèle informé que selon le modèle naïf. On peut donc confirmer que pour le modèle informé, la conversion rapproche les identités naturelle et synthétique. L'expérience (5) montre que les échantillons synthétiques générés avec les modèles PPG2Mel$_A$ et PPG2Mel$_B$ ne sont pas distinguables par le modèle naïf, et sont tous deux éloignés des échantillons naturels de A. Le modèle informé fait une légère distinction entre les échantillons synthétiques générés par PPG2Mel$_A$ et PPG2Mel$_B$. Le modèle de conversion ne permet pas d'atteindre le locuteur cible (ici A ou B) d'un point de vue de ces modèles de vérification. On peut donc répondre à la question Q3 : le modèle informé est légèrement meilleur pour identifier le lien entre la voix cible synthétisée et la voix cible naturelle. Cependant, ce résultat doit être manié avec précaution car réalisé avec deux voix cibles uniquement, et un unique système de vérification.

5 Conclusion

La première expérience présentée vise à s'assurer que notre système de conversion à base de PPGs produit de l'audio de qualité correcte. Le test perceptif réalisé montre que nous obtenons une qualité similaire à celle d'un système TTS habituel, et que le vocodeur utilisé est la source d'une grande partie de la dégradation. L'utilisation d'un meilleur vocodeur ainsi que la réalisation d'un test perceptif de similarité locuteur pourraient être des pistes de recherche pour continuer ces travaux.

Nous avons ensuite entraîné deux systèmes de vérification du locuteur sur de l'audio naturel et sur des données synthétiques afin d'identifier l'information relative au locuteur source qui aurait été cachée par la conversion. Nos expériences montrent que même si le système naïf obtient des résultats comparables avec l'état de l'art sur de la parole naturelle, ni ce système ni le système informé ne parviennent à identifier les locuteurs originaux une fois l'étape de conversion passée. De même, aucun de nos systèmes n'a été capable de lier les versions naturelles et synthétiques d'un même locuteur cible. Nous concluons donc que l'ensemble de la chaine de conversion depuis l'extraction des PPGs à la génération de l'audio permet de cacher les indices acoustiques à un système de vérification du locuteur. Ainsi notre modèle de conversion à partir de PPGs semble pertinent pour controller la génération de parole, tout en cachant les locuteurs sources.

Références

CHUNG J. S., NAGRANI A. & ZISSERMAN A. (2018). Voxceleb2 : Deep speaker recognition. In *INTERSPEECH*.

COOPER E., LAI C.-I., YASUDA Y., FANG F., WANG X., CHEN N. & YAMAGISHI J. (2020). Zero-shot multi-speaker text-to-speech with state-of-the-art neural speaker embeddings. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 6184–6188 : IEEE.

DESPLANQUES B., THIENPONDT J. & DEMUYNCK K. (2020). ECAPA-TDNN : Emphasized Channel Attention, Propagation and Aggregation in TDNN based speaker verification. In *INTERSPEECH 2020*, p. 3830–3834 : International Speech Communication Association (ISCA).

DIATLOVA D. & SHUTOV V. (2023). EmoSpeech : guiding FastSpeech2 towards Emotional Text to Speech. In *Proc. 12th ISCA Speech Synthesis Workshop*, p. 106–112. DOI : 10.21437/SSW.2023-17.

FAYET C., BLOND A., COULOMBEL G., SIMON C., LOLIVE D., LECORVÉ G., CHEVELU J. & LE MAGUER S. (2020). FlexEval, création de sites web légers pour des campagnes de tests perceptifs multimédias. In *Journées d'Études sur la Parole*, p. 22–25, Nancy, France.

ITO K. & JOHNSON L. (2017). The lj speech dataset. https://keithito.com/LJ-Speech-Dataset/.

KONG J., KIM J. & BAE J. (2020). Hifi-gan : Generative adversarial networks for efficient and high fidelity speech synthesis. In H. LAROCHELLE, M. RANZATO, R. HADSELL, M. BALCAN & H. LIN, Éds., *Advances in Neural Information Processing Systems*, volume 33, p. 17022–17033 : Curran Associates, Inc.

ŁAŃCUCKI A. (2021). Fastpitch : Parallel text-to-speech with pitch prediction. In *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 6588–6592 : IEEE.

LEVY-LESHEM R. & GIRYES R. (2021). Taco-vc : A single speaker tacotron based voice conversion with limited data. In *2020 28th European Signal Processing Conference (EUSIPCO)*, p. 391–395. DOI : 10.23919/Eusipco47968.2020.9287448.

NAGRANI A., CHUNG J. S. & ZISSERMAN A. (2017). Voxceleb : a large-scale speaker identification dataset. *Telephony*, **3**, 33–039.

PANAYOTOV V., CHEN G., POVEY D. & KHUDANPUR S. (2015). Librispeech : An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 5206–5210. DOI : 10.1109/ICASSP.2015.7178964.

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. Interspeech 2021*, p. 3615–3619. DOI : 10.21437/Interspeech.2021-475.

PRENGER R., VALLE R. & CATANZARO B. (2019). Waveglow : A flow-based generative network for speech synthesis. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 3617–3621 : IEEE.

QIAN K., ZHANG Y., CHANG S., HASEGAWA-JOHNSON M. & COX D. (2020). Unsupervised speech decomposition via triple information bottleneck. In H. D. III & A. SINGH, Éds., *Proceedings of the 37th International Conference on Machine Learning*, volume 119 de *Proceedings of Machine Learning Research*, p. 7836–7846 : PMLR.

QIAN K., ZHANG Y., CHANG S., YANG X. & HASEGAWA-JOHNSON M. (2019). AutoVC : Zero-shot voice style transfer with only autoencoder loss. In K. CHAUDHURI & R. SALAKHUTDINOV,

Éds., *Proceedings of the 36th International Conference on Machine Learning*, volume 97 de *Proceedings of Machine Learning Research*, p. 5210–5219 : PMLR.

REN Y., RUAN Y., TAN X., QIN T., ZHAO S., ZHAO Z. & LIU T.-Y. (2019). Fastspeech : Fast, robust and controllable text to speech. In H. WALLACH, H. LAROCHELLE, A. BEYGELZIMER, F. D'ALCHÉ-BUC, E. FOX & R. GARNETT, Éds., *Advances in Neural Information Processing Systems*, volume 32 : Curran Associates, Inc.

SHEN J., PANG R., WEISS R. J., SCHUSTER M., JAITLY N., YANG Z., CHEN Z., ZHANG Y., WANG Y., SKERRV-RYAN R., SAUROUS R. A., AGIOMVRGIANNAKIS Y. & WU Y. (2018). Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 4779–4783. DOI : 10.1109/ICASSP.2018.8461368.

SINI A., MAGUER S. L., LOLIVE D. & DELAIS-ROUSSARIE E. (2020). Introducing Prosodic Speaker Identity for a Better Expressive Speech Synthesis Control. In *Speech Prosody 2020*, p. 935–939 : ISCA.

SOLAK I. (2019). The m-ailabs speech dataset. https://www.caito.de/2019/01/03/the-m-ailabs-speech-dataset/.

TAE J., KIM H. & KIM T. (2022). EdiTTS : Score-based Editing for Controllable Text-to-Speech. In *Proc. Interspeech 2022*, p. 421–425. DOI : 10.21437/Interspeech.2022-6.

VALLE R., LI J., PRENGER R. & CATANZARO B. (2020). Mellotron : Multispeaker Expressive Voice Synthesis by Conditioning on Rhythm, Pitch and Global Style Tokens. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 6189–6193.

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 *International Conference on Machine Learning*, p. 5180–5189 : PMLR.

YEH C.-C., HSU P.-C., CHOU J.-C., LEE H.-Y. & LEE L.-S. (2018). Rhythm-flexible voice conversion without parallel data using cycle-gan over phoneme posteriorgram sequences. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, p. 274–281. DOI : 10.1109/SLT.2018.8639647.

ZHANG X., TRMAL J., POVEY D. & KHUDANPUR S. (2014). Improving deep neural network acoustic models using generalized maxout networks. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 215–219. DOI : 10.1109/ICASSP.2014.6853589.

ZHAO G., DING S. & GUTIERREZ-OSUNA R. (2019). Foreign Accent Conversion by Synthesizing Speech from Phonetic Posteriorgrams. In *Proc. Interspeech 2019*, p. 2843–2847. DOI : 10.21437/Interspeech.2019-1778.