Techniques de synthèse vocale neuronale à l'épreuve des données d'apprentissage non dédiées : les livres audio amateurs en français

Aghilas Sini* — **Lily Wadoux*** — **Antoine Perquin*** — **Gaëlle Vidal*** — **David Guennec*** — **Damien Lolive*** — **Pierre Alain*** — **Nelly Barbot*** — **Jonathan Chevelu*** — **Arnaud Delhay***

* *Université de Rennes, CNRS, IRISA, France*

RÉSUMÉ. *Dans cet article, nous nous intéressons à la capacité des systèmes de synthèse vocale neuronale à tirer parti des données non dédiées en langue française. En effet, ces dernières sont abondantes mais leurs conditions d'enregistrement sont hétérogènes, alors que les données dédiées à la synthèse de parole (de meilleure qualité) sont en quantité limitée et difficiles à collecter. Leur impact est mesuré sur trois systèmes : synthèse de parole monolocuteur, clonage de voix et conversion de voix. Des évaluations objectives et subjectives sur la reproduction de la voix du locuteur et sur la qualité des échantillons synthétisés ont été menées. Elles montrent qu'il est difficile de produire une synthèse vocale de qualité comparable avec l'état de l'art dans certaines conditions d'enregistrement ou pour des voix atypiques.*

MOTS-CLÉS : *Synthèse de la parole, clonage de voix, conversion de voix, synthèse vocale neuronale.*

ABSTRACT. *In this article, we consider how neural speech synthesis systems perform with non-dedicated data in French. Indeed, these are plentiful, unlike dedicated data of better quality which are limited in their availability and difficult to collect, but are recorded in heterogeneous conditions. Their impact is measured on three systems: single-speaker speech synthesis, voice cloning and voice conversion. Speaker similarity and overall quality were measured through objective and subjective evaluations. Our results outline the difficulty of producing high-quality speech synthesis under some recording conditions, or for atypical voices.*

KEYWORDS: *Speech synthesis, Voice Cloning, Voice Conversion, Neural synthesis.*

1. Introduction

Les technologies de la synthèse vocale restent fortement contraintes par les données nécessaires à leur élaboration, à la fois en termes de qualité et de quantité. Même si les premiers systèmes de synthèse de la parole en utilisaient de très faibles quantités, l'évolution des méthodes a conduit à l'utilisation de plus grandes quantités de données. La synthèse de la parole de bout en bout (*end-to-end speech synthesis*) peut, selon les applications, nécessiter des dizaines d'heures, voire des centaines d'heures dans le cas de la synthèse massivement multilocuteur. Une constante, quelle que soit la technologie, reste que les données employées sont quasi exclusivement dédiées à la tâche de synthèse et de très haute qualité, en particulier si une application industrielle est visée. Dans la suite, on utilisera le terme « données dédiées » pour référencer les jeux de données conçus spécifiquement pour la synthèse de parole.

Les méthodes actuelles de synthèse de la parole sont généralement des variantes de la synthèse monolocuteur (Tan *et al.*, 2021). Elles impliquent une importante préparation et une quantité non triviale de données pour produire une voix de synthèse. Une approche récente, le clonage de voix, ne nécessite au contraire que quelques minutes de parole pour modéliser l'identité du locuteur (Snyder *et al.*, 2017). Une autre méthode réside dans la conversion de voix visant à transformer un énoncé vocal existant, produit par un locuteur source, afin qu'il soit perçu comme produit par un locuteur cible différent (Zhao *et al.*, 2019).

Les meilleurs résultats dans les publications en synthèse de la parole sont généralement obtenus sur la base de corpus d'apprentissage dédiés et très qualitatifs. Ils tendent à respecter un fort degré d'uniformité stylistique et une homogénéité des caractéristiques des locuteurs. On a ainsi des voix très majoritairement jeunes, plus souvent féminines que masculines, employant un style calme, posé, relativement neutre émotionnellement (et ce malgré une avancée de l'expressivité générale grâce aux méthodes de bout en bout). Ce manque de diversité s'explique souvent par les attentes présumées de la population pour laquelle la voix de synthèse a été produite et aussi par la difficulté à maîtriser des données expressives. Il existe peu d'études sur l'influence du choix de la voix sur la qualité de la synthèse produite (Hinterleitner *et al.*, 2014). À notre connaissance, il n'en existe pas sur les voix atypiques en français.

Des initiatives basées sur des données de qualité variable et non dédiées à la tâche de synthèse de la parole existent cependant, notamment en anglais. On peut par exemple citer le corpus *Librispeech* (Panayotov *et al.*, 2015), construit pour la tâche de reconnaissance de la parole, et une version améliorée (*LibriTTS* (Zen *et al.*, 2019)) où les audios les moins utilisables pour les applications de synthèse ont été exclus.

À notre connaissance, cet aspect de la synthèse de parole n'est pas exploré pour le français. Dans cet article, nous tâchons de pallier ce manque en considérant la question suivante : comment les systèmes de synthèse vocale se comportent-ils lorsqu'ils ont été entraînés sur des données de qualité inférieure aux standards du domaine et, qui plus est, non dédiées à la tâche de synthèse ?

Ne pouvant traiter le problème de manière exhaustive, nous restreignons le cadre de cette étude à l'utilisation de données disparates issues de livres audio expressifs, ayant subi un ensemble minimal de prétraitements. De même, une seule technologie majeure par méthode de synthèse est prise en compte. Nous nous basons ainsi sur l'architecture Tacotron 2 (Shen *et al.*, 2018) qui représente la base architecturale de la majorité des publications depuis son apparition. Une variante du modèle est apprise pour chacune des trois applications visées : synthèse, conversion et clonage de voix. Ces modèles sont appris sur un corpus de parole obtenu à partir de livres audio de locuteurs amateurs à l'expressivité variable. Le vocodeur est commun à toutes les approches afin d'évaluer uniquement le modèle acoustique, ce dernier étant généralement l'élément de la chaîne le plus impacté par la variabilité dans les données.

Nous commençons par présenter l'état de l'art sur les techniques de synthèse vocale utilisées dans cet article et justifions nos choix. La section 3 donne ensuite une présentation détaillée des données utilisées pour nos expériences. Les choix architecturaux et les détails concernant l'entraînement des différents modèles sont décrits en section 4. Enfin, les sections 5 et 6 détaillent le protocole expérimental et les résultats obtenus. Cette dernière section se conclut par une discussion des résultats. Nos conclusions et perspectives futures sont présentées en section 7.

2. Travaux connexes

Cette partie présente les technologies de synthèse vocale et les grands défis auxquels le domaine est confronté. Nous évoquons d'abord la synthèse de bout en bout de manière globale avant de nous focaliser sur le modèle acoustique puis sur le vocodeur. Enfin, nous discutons des défis relatifs aux données en synthèse de bout en bout ainsi que des travaux s'assimilant au nôtre sur cet aspect. Une discussion de la difficulté du processus d'évaluation de la synthèse est également abordée dans ce cadre.

2.1. *Processus de synthèse vocale*

Un système de synthèse de parole à partir du texte a pour objectif de produire, à partir d'une séquence de mots, éventuellement accompagnée de consignes, un signal de parole correspondant.

Ces systèmes présentent actuellement les meilleurs résultats en termes de rendu naturel dans l'état de l'art, au contraire des systèmes précédents (sélection d'unités, synthèse par HMM (*Hidden Markov Models*) ou DNN (*Deep Neural Network*) non de bout en bout) qui apparaissent de plus en plus rarement dans les *challenges* de comparaison (Ling *et al.*, 2021). Ils présentent de nombreux avantages, l'essentiel du savoir expert étant contenu dans le modèle lui-même. Ce dernier fait essentiellement un travail de conversion entre une séquence d'unités linguistiques (texte) ou phonétiques (parfois les deux) et une séquence cible de nature acoustique (audio). S'il existe des modèles réalisant cette conversion directement (Clarinet par exemple,

Ping *et al.* (2019)), on divise généralement le processus en deux modèles distincts. Il s'agit alors de prédire une représentation acoustique intermédiaire (généralement un mel-spectrogramme) plutôt que l'audio directement. Ce modèle est alors appelé modèle acoustique. Un second modèle, appelé vocodeur, vient alors traduire le mel-spectrogramme en un signal audio. L'avantage de ce principe est que le vocodeur peut être indépendant du locuteur et appris sur de grandes quantités de données multilocuteurs. Dans cette étude, nous utilisons un unique vocodeur pour l'ensemble des approches.

2.2. *Modélisation acoustique*

Dans cette section, nous détaillons le modèle acoustique et ses variantes pour la synthèse monolocuteur, le clonage et enfin la conversion de voix. Ils sont schématisés dans la figure 1.

2.2.1. *Modèle acoustique et synthèse monolocuteur*

Le modèle acoustique le plus populaire est Tacotron 2 (Shen *et al.*, 2018). Il s'agit d'un modèle neuronal séquence à séquence autorégressif qui suit l'architecture encodeur/décodeur et inclut un module d'attention. Des extensions ont été proposées afin d'accélérer la synthèse au prix d'un apprentissage plus complexe, tel FastSpeech2 (Ren *et al.*, 2020) ou d'ajouter du contrôle de la prosodie, par exemple en modélisant la prosodie d'un signal de parole de consigne à l'aide d'un auto-encodeur variationnel (Elias *et al.*, 2021). Elles sont néanmoins moins utilisées que Tacotron 2 dans la littérature, ce qui explique notre choix.

Le cas de la synthèse neuronale de bout en bout monolocuteur est le cas le plus simple d'utilisation du modèle acoustique. Avec cette méthode, chaque voix synthétisée est produite par un modèle acoustique distinct : un modèle par locuteur.

2.2.2. *Clonage de voix*

La synthèse de bout en bout monolocuteur requiert une quantité non négligeable de données par locuteur pour entraîner les modèles acoustiques correspondant à chaque locuteur. Ce n'est pas le cas pour la synthèse neuronale multilocuteur. Son principe est d'entraîner un unique modèle acoustique sur un corpus comprenant plusieurs locuteurs, dont ceux que l'on souhaite synthétiser. La possibilité d'utiliser plusieurs locuteurs à l'entraînement offre deux avantages. Premièrement, moins de données pour chaque locuteur sont nécessaires car l'agrégation de tous les locuteurs donne le volume de données requis. Deuxièmement, le modèle a la possibilité d'apprendre des différences entre locuteurs. Lors de la synthèse, la voix souhaitée est spécifiée au modèle acoustique par un vecteur *one-hot* (Arık *et al.*, 2017 ; Ping *et al.*, 2018). Cette approche ne peut donc synthétiser que les voix de son corpus d'entraînement.

Il est cependant possible d'utiliser ce type de modèle multilocuteur pour le personnaliser avec un locuteur absent du corpus d'entraînement. Cette approche, appelée

clonage de voix, se décline en deux méthodes : l'adaptation au locuteur et l'encodage de locuteurs. La première repose sur une étape d'adaptation, ou *fine-tuning*, du modèle multilocuteur préentraîné afin qu'il ne produise plus que la voix du locuteur cible. L'encodage de locuteurs, quant à lui, ne nécessite pas d'étape de *fine-tuning*. À la place, un second modèle, appelé encodeur de locuteurs, fournit au modèle de synthèse une représentation vectorielle des caractéristiques du locuteur, appelée plongement de locuteur. Ce modèle peut être entraîné conjointement au modèle acoustique, ou séparément (Jia *et al.*, 2018). Lors de la synthèse, le plongement du locuteur cible est nécessaire pour que le modèle acoustique génère de la parole synthétique se rapprochant de sa voix réelle. En cas de changement de locuteur, il suffit de transmettre les échantillons audio de la nouvelle cible à l'encodeur de locuteurs. Ces deux méthodes n'ont besoin que d'une faible quantité de données du locuteur cible. Ainsi, dans l'étude (Chen *et al.*, 2019), elles génèrent de bons résultats à partir de dix secondes de parole, et de très bons à partir de dix minutes.

Dans cette étude, nous utilisons l'approche par encodage de locuteurs. En effet, malgré des résultats de qualité légèrement inférieure à l'approche par adaptation au locuteur (Arik *et al.*, 2018), elle ne nécessite qu'une seule phase d'entraînement et permet donc de généraliser les tests plus facilement à de nouveaux locuteurs. Les modèles choisis pour cette approche sont le modèle x-vecteurs (Snyder *et al.*, 2017) comme encodeur de locuteurs et Tacotron 2 comme modèle acoustique.

Le modèle x-vecteurs est très largement utilisé dans les travaux de vérification du locuteur, et par extension, de clonage de voix. Il est basé sur une architecture en trois blocs. Le premier est un ensemble de couches fonctionnant à l'échelle de la trame pour extraire une représentation de chaque trame et de son contexte. Le deuxième est une agrégation statistique permettant de condenser l'information apportée par chaque trame contextualisée à l'échelle du segment audio entier. Enfin, le dernier bloc est un ensemble de couches fonctionnant à l'échelle du segment et d'où est extrait le plongement de locuteur, appelé x-vecteur.

Le modèle Tacotron 2 est présenté dans la section 2.2.1. Néanmoins, le modèle utilisé ici est multilocuteurs : un plongement de locuteur est concaténé à la sortie de l'encodeur du Tacotron, avant d'être transmis à son décodeur.

2.2.3. *Conversion de voix*

L'objectif de la conversion de voix est de transformer un énoncé produit par un locuteur source, en conservant l'information linguistique, afin que celui-ci soit perçu comme ayant été prononcé par un locuteur cible.

Une étude récente reprend l'évolution des différentes techniques de conversion de voix (Sisman *et al.*, 2021). Historiquement, les premières approches se concentraient sur la modification des caractéristiques spectrales de la voix, par exemple le spectre et les formants, en utilisant des données parallèles, c'est-à-dire des données pour lesquelles les locuteurs source et cible ont prononcé le même contenu. Grâce à un alignement dynamique temporel entre séquences source et cible, il était alors possible

de calculer une fonction de transformation effectuant la correspondance entre les espaces acoustiques des locuteurs source et cible. Les premières approches proposées se fondaient sur la quantification vectorielle (Abe *et al.*, 1990), puis ont évolué vers des approches probabilistes utilisant des mélanges de lois gaussiennes (Toda *et al.*, 2007).

Les travaux de recherche dans le domaine se sont ensuite orientés vers l'apprentissage de fonctions de transformation en utilisant des données non parallèles (Erro *et al.*, 2009 ; Wang *et al.*, 2015). Plus récemment, l'introduction des PPG (*Phonetic PosteriorGrams*) constitue une nouvelle technique permettant de s'affranchir de données parallèles (Sun *et al.*, 2016). Les Phonetic PosteriorGrams (PPG) représentent l'évolution temporelle de la probabilité *a posteriori* des phonèmes (Hazen *et al.*, 2009). Ils capturent le contenu linguistique d'un énoncé tout en gardant l'information temporelle. Généralement, l'extraction des PPG est effectuée par un modèle acoustique multilocuteur permettant d'effacer les composantes liées au locuteur source. Un modèle de synthèse spécifique au locuteur cible peut ensuite être utilisé pour générer un signal de parole synthétique ayant les caractéristiques de ce locuteur. À l'heure actuelle, les techniques utilisant les PPG donnent les meilleurs résultats (Zheng *et al.*, 2020) au Voice Conversion Challenge (VCC). Notamment, lors de l'édition VCC 2020, quatre types de systèmes de conversion pouvaient être distingués : a) ceux combinant la reconnaissance automatique de la parole suivie de synthèse de parole, b) les méthodes basées sur les PPG (Tian *et al.*, 2018 ; Liu *et al.*, 2021), c) les approches de type auto-encodeur (Ho et Akagi, 2020), et d) les approches génératives antagonistes (GAN) (Tobing *et al.*, 2020).

Les méthodes qui combinent la reconnaissance et la synthèse de parole utilisent le texte comme pivot, ce qui permet également d'effacer les caractéristiques du locuteur source. Les performances d'un tel système sont, par nature, dépendantes de la qualité de la reconnaissance de parole.

Dans cet article, nous utilisons un système de conversion de l'état de l'art reposant sur les PPG (Zhao *et al.*, 2019). Le choix de cette approche est motivé par sa capacité à préserver les propriétés temporelles de l'audio source au travers du PPG. De plus, le système reprend l'architecture Tacotron 2 utilisée pour la synthèse vocale monolocuteur, ce qui permet d'avoir une base de comparaison entre les deux approches.

2.3. *Vocodeur*

La dernière étape du processus de synthèse monolocuteur est la transformation du mel-spectrogramme en un flux audio PCM (*Pulse Code Modulation*). Ce travail est opéré par un modèle distinct du modèle acoustique présenté en section 2.2 et entraîné de manière indépendante.

L'état de l'art offre un grand nombre de propositions pour cette tâche. Le modèle neuronal le plus connu est Wavenet (van den Oord *et al.*, 2018), mais des modèles comme SampleRNN (Mehri *et al.*, 2017) ou ParallelWaveGAN (Yamamoto *et al.*,

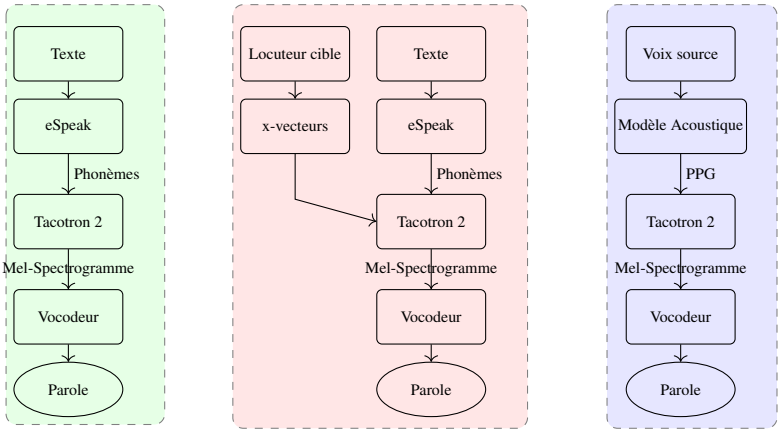

Figure 1. *Architecture des différentes techniques de synthèse vocale : monolocuteur (à gauche), le clonage (au milieu, encodeur identique au premier) et conversion de voix (à droite).*

2020) existent. Dans cette étude, nous utilisons WaveGlow (Prenger *et al.*, 2019). Notre choix s'est porté sur ce vocodeur car il n'est pas autorégressif.

2.4. *Défis propres à la synthèse vocale neuronale*

Les techniques de synthèse vocale neuronale, malgré leurs différences, sont confrontées à des problématiques semblables telles que la disponibilité des données en quantité suffisante et de haute qualité, et une évaluation adaptée à la tâche pour laquelle la technique a été mise en place.

2.4.1. *Données d'entraînement*

Malgré le progrès des techniques de synthèse vocale, la disponibilité des données en quantité suffisante et de haute qualité est souvent considérée comme un prérequis pour l'obtention d'un système de synthèse vocale de l'état de l'art. Ce défi est d'autant plus difficile à relever lorsqu'il s'agit de données issues de langues autres que l'anglais. La plupart des travaux de *benchmark* sur les différentes techniques de synthèse vocale se font ainsi sur des corpus en anglais : LJSpeech (Ito et Johnson, 2017), VCTK (Veaux *et al.*, 2017), LibriTTS (Zen *et al.*, 2019), ARCTIC (Kominek et Black, 2004). Les corpus dédiés à la synthèse vocale en langue française sont, par contraste, peu exploités et relativement peu volumineux. On peut citer FrenchSiwis (Honnet *et al.*, 2017) et SynPaFlex (Sini *et al.*, 2018), tous deux conçus pour la synthèse vocale et contenant des données d'une seule locutrice. On peut aussi évoquer Att-HACK (Le Moine et Obin, 2020) pour la conversion de voix avec plusieurs locuteurs de haute qualité, mais avec une quantité de données par locuteur relativement

limitée. BREF (Lamel *et al.*, 1991) est un autre corpus susceptible d'être utilisé pour l'apprentissage de synthèse vocale multilocuteur ou le clonage de voix, en particulier en raison d'une bonne qualité d'enregistrement. Tous ces corpus peuvent être utilisés pour avoir un système de synthèse vocale de qualité.

Cependant, les systèmes de synthèse vocale dits de bout en bout peuvent-ils utiliser toutes sortes de données, en particulier des données qui n'ont pas été collectées à des fins de synthèse, lorsqu'elles sont présentes en grande quantité ? Exposer les systèmes de synthèse vocale de l'état de l'art à des données amateurs en grande quantité sans préalablement poser de contraintes liées aux caractéristiques du locuteur ou au style semble pertinent afin d'explorer les limites des architectures actuelles sur une langue autre que l'anglais.

2.4.2. *Méthodologie d'évaluation de la parole synthétique*

L'objet de la synthèse vocale étant de produire un stimulus de parole destiné à l'humain, l'évaluation subjective reste incontournable et ce malgré ses défauts biens connus dans la communauté (Wagner *et al.*, 2019). Les raisons principales reposent sur l'absence d'une métrique restituant parfaitement l'opinion de l'humain et sur le faible nombre de travaux dans ce domaine.

Une évaluation adaptée à la tâche de synthèse est l'une des tâches les plus onéreuses du processus expérimental car elle dépend de nombreux facteurs : recrutement des testeurs et vérification de leur niveau linguistique, choix du test audio permettant de répondre à la question de recherche avec le moins de biais possible, considération des conditions d'évaluation, du matériel utilisé lors du test et du temps imparti pour éviter la fatigue des testeurs, optimisation de focalisation de l'effort cognitif des testeurs avec une interface ergonomique.

Dans les challenges VCC (Yi *et al.*, 2020 ; Lorenzo-Trueba *et al.*, 2018) et les éditions du challenge Blizzard (depuis 2005), dédiés respectivement aux techniques de conversion de voix et à la synthèse vocale à partir du texte, des tests de similarité au locuteur, de qualité et d'intelligibilité sont le plus souvent mis en place afin d'évaluer tous les participants. Les tests de qualité et d'intelligibilité sont en général fusionnés pour qualifier la qualité globale, car ils sont intrinsèquement liés.

Le principal outil utilisé pour l'évaluation reste le test MOS (*Mean Opinion Score*) qui agrège les notes attribuées en aveugle par des testeurs entre un minimum de 1 et un maximum de 5 (ITU-T, 1996). Sa déclinaison pour la dégradation par rapport à une référence (DMOS) est également fréquemment employée, tout comme les tests de préférence (A/B) et les évaluations de type MUSHRA. Cette dernière impose cependant une tâche plus astreignante au testeur, tous les stimuli évalués (en aveugle) lui étant présentés en simultané et une notation fine étant exigée pour chacun d'entre eux. En outre, MUSHRA impose l'ajout d'une ancre basse en plus de l'ancre haute (référence) généralement utilisée.

L'évaluation objective, quant à elle, repose sur des métriques utilisées dans les algorithmes pour quantifier la « qualité » du signal synthétique, à savoir : métriques

spectrales, MCD (*Mel-Cepstral Distortion*) et SNR (*Signal Noise Ratio*) ; fréquence fondamentale (F_0, rapport voisement/non-voisement, Likelihood-ratio (LL-Ratio)) ; BAP (*Band Aperiodicity Parameter*) ; durée syllabes/phonèmes. On peut même citer des initiatives reposant sur l'apprentissage profond comme MOSNet, qui visent à estimer des scores de tests subjectifs (Lo *et al.*, 2019), mais ces approches, manquant de maturité, restent marginales. Concernant l'intelligibilité, il est le plus souvent fait usage d'un système de reconnaissance automatique de la parole (Vích *et al.*, 2008). Pour la similarité au locuteur cible, un calcul de distance cosinus avec des plongements locuteurs est souvent employé. Ces méthodes d'évaluation objectives peuvent être de bons indicateurs lors de l'entraînement de systèmes de synthèse vocale, mais elles ne sont pas suffisantes. En outre, un processus d'alignement est parfois nécessaire car le calcul de ces métriques entraîne souvent le passage par des représentations intermédiaires pouvant engendrer des artefacts.

Ceci dit, les méthodes objectives reposent souvent sur une comparaison à une référence absolue. Ceci n'est cependant pas toujours adapté à la tâche d'évaluation. En effet, le fait qu'un échantillon soit différent d'un autre (même d'une référence) n'implique pas nécessairement que celui-ci lui est inférieur. Les évaluations objectives, qui s'appuient sur les systèmes de reconnaissance de la parole ou de vérification du locuteur pour le calcul de l'intelligibilité et de la similarité, ne sont pas meilleures car ces techniques comportent aussi des erreurs de prédiction.

3. Jeu de données

Pour cette étude, nous avons cherché à disposer de lectures enregistrées et accompagnées de leur texte, qui proposent du contenu expressif, et pour une multiplicité de locuteurs. La langue ciblée est le français, et une durée minimale d'environ dix heures est requise pour au moins quelques voix.

De nombreux livres audio enregistrés et partagés par des amateurs sont accessibles aujourd'hui. Ces données présentent des lectures diversifiées, par des donneurs de voix singuliers et relativement libres dans l'interprétation des œuvres de leur choix. C'est pourquoi nous avons choisi d'utiliser le corpus MUFASA (*MUltispeaker French Audiobooks corpus dedicated to expressive read Speech Analysis*) pour nos expériences.

3.1. *Le corpus MUFASA*

Le corpus MUFASA est une base évolutive d'enregistrements de livres audio réalisés par des particuliers, à partir de textes libres de droits, principalement en langue française. Les données sont issues de collectes au format MP3 128 Kbit/s pour l'audio (plus rarement 64 Kbit/s), et aux formats texte et PDF pour les transcriptions. Elles sont progressivement annotées et validées, et leur contenu en français au moment de l'étude est d'environ 600 heures de parole réparties entre vingt locuteurs, dix hommes et dix femmes.

Au cours de certaines lectures, quelques phrases ont été ajoutées au texte de référence pour présenter ou pour clore la section lue. Une partie des textes concernés ont été conformés à la parole. D'autres annotations manuelles ont relevé et documenté, pour certains fichiers audio, des dégradations de qualité ou la présence de sons exogènes à la parole. Les conditions non professionnelles des prises de son favorisent des variations d'intensité et de répartition des fréquences, et sont susceptibles d'enregistrer réverbérations et bruits de fond, ou artefacts liés à la qualité d'encodage ou à la captation. Sont signalés également, quand ils sont repérés, musique ou bruitages insérés intentionnellement par montage.

Les locuteurs ont des profils très divers. Ils sont fidèles aux textes mais assez libres dans les pauses, souvent décorrélées des ponctuations. Quatre d'entre eux ont un accent régional. Seule une voix fait quelques écarts de prononciation et a une prosodie légèrement hésitante. Trois tranches d'âge perçu sont représentées : adulte (10), senior (6) et jeune (4). Les textes sont issus de différents courants littéraires, les plus récents datent du milieu du vingtième siècle. Le genre narratif y est le plus représenté, avec une quantité importante de dialogues dans la plupart des œuvres.

Du point de vue de la prosodie, chaque lecteur a une posture de référence qui lui est propre, et que l'on retrouve dans la narration de façon générale. On peut les classer en trois groupes : onze lecteurs produisent un motif prosodique récurrent, à l'échelle de la phrase. Cinq lisent d'une parole proche du spontané, naturelle. Les quatre autres ont des stratégies plus amples, où les phrases se succèdent de façon contrastée, où le texte est plus incarné. Pour les passages au style direct, également, chaque locuteur organise sa lecture avec plus ou moins de variabilité. Le premier groupe ne marque pas nécessairement d'expressivité ou de personnification (l'abandon du motif narratif fait déjà rupture). Les lecteurs plus naturels dans les passages narratifs marquent une emphase expressive au style direct, et aussi des changements de timbre, subtils le plus souvent. Les lecteurs les plus stratégiques dans la narration sont aussi les plus théâtraux au style direct, avec une nette emphase expressive et des changements de timbre parfois radicaux.

3.2. *Annotation, sélection des données*

Pour cette étude, nous avons extrait du corpus MUFASA un sous-ensemble en langue française d'une durée globale de 222 heures, correspondant à 667 unités audio (généralement des chapitres) de durée très variable : de moins d'une minute à plus de deux heures. Les sous-corpus de chaque locuteur durent d'une centaine de minutes à près de quinze heures (dont huit font plus de dix heures). Des regroupements de chapitres d'une même œuvre ont été préférés à une pioche disparate. En moyenne, deux à trois livres différents sont représentés dans chaque sélection par locuteur, il peut y en avoir jusqu'à huit.

L'annotation automatique consiste en un alignement du texte et de la parole associée à un découpage, sur les silences, en énoncés courts. Pour notre étude, cette

fragmentation a visé à obtenir des unités audio d'une durée inférieure à 10 secondes, une partie présente donc toujours des pauses internes. Tous les textes ont été normalisés et phonétisés au format IPA avec le logiciel *eSpeak*[1]. Des règles ont été dérivées de la validation manuelle d'une partie des segments, puis appliquées aux autres. Cette opération a abouti principalement à exclure les unités correspondant aux débuts et aux fins de chapitre, les plus susceptibles de présenter de la musique. Un dernier traitement permet d'exclure automatiquement les paires signal-texte les moins vraisemblables dans le rapport entre durée audio et nombre de mots, et aussi les unités de plus de 10 secondes qui subsistent. Notre étude porte sur un ensemble correspondant à 161 heures de parole, découpées en énoncés d'une durée moyenne de 4 secondes.

Après nos expérimentations, une expertise acoustique a porté sur les segments audio utilisés, regroupés par enregistrement d'origine. Leur analyse, réalisée à l'aide des outils *open source Audacity* pour les spectres de fréquence et *FreeLCS* pour les intensités en unités LUFS (*Loudness Unit Full Scale*), montre globalement des caractéristiques constantes pour chaque locuteur. Pour six d'entre eux, le signal de parole est de qualité bonne à convenable, six autres présentent des artefacts d'acquisition (résonance médium ou ventilation). La qualité pour les locuteurs restants peut être considérée comme moyenne, elle est irrégulière pour deux d'entre eux. Quelques écarts d'intensité et de qualité sont néanmoins constatés entre les enregistrements et, pour au moins deux voix, la sélection comporte des segments où bruitages et musiques sont superposés à la parole.

3.3. *Sélection des locuteurs cibles*

Le critère principal de sélection des locuteurs cibles est l'existence de traits distinctifs saillants pour chacun d'eux. En outre, les données pour chaque locuteur doivent être disponibles en quantité suffisante, de l'ordre de dix heures de parole au minimum. Notre choix s'est ainsi porté sur *Nadine*, *Jean-Luc*, *René* et *Victoria*.

Les flux de parole de *Nadine* et de *René* sont assez constants, chacun à leur manière : ils présentent un motif rythmique et intonatif récurrent qui pour *Nadine* est emblématique de la narration, et pour *René* relève d'un style personnel très marqué, pittoresque, se déployant aussi au style direct. *Nadine* l'abandonne au style direct pour porter des expressivités plus naturelles, avec quelques changements de timbre. Les stratégies narratives de *Jean-Luc*, par effet d'énigme, et de *Victoria*, au style fantasque, sont plus sophistiquées, et au style direct ces deux lecteurs mettent en œuvre des expressivités exacerbées et des changements de timbre radicaux.

Après nos expériences, une description de l'audio utilisé pour ces locuteurs cibles a porté sur les regroupements des segments par enregistrement. Elle relève une bonne qualité du sous-corpus *Nadine*, quoique perfectible pour des écarts d'intensité (un quart des données sont en excès de 6 dB LUFS sur les autres), et la présence d'artefacts

1. https ://espeak.sourceforge.net/

sur 4 % de sa durée. La voix de *Jean-Luc* est aussi très bien enregistrée, mais 5 à 6 % de ses données sont corrompues par des musiques et bruitages forts, superposés à la parole. Musiques et bruitages apparaissent aussi superposés à la voix de *Victoria*, plus légers mais dans les mêmes proportions, et la qualité globale des enregistrements de cette locutrice est, elle, moyenne et irrégulière (5 % des durées audio sont accompagnées d'une onde parasite). La voix de *René*, quant à elle, est accompagnée d'une onde médium avec plus ou moins de résonance. Les deux dernières voix citées présentent par ailleurs des écarts d'intensité sur un cinquième de leurs données (respectivement - 10 et + 3 dB LUFS). Autres phénomènes, on note dans le sous-corpus de *Jean-Luc* la disparition du timbre pour un personnage qui s'exprime sur 3 % des segments, et l'apparition parcimonieuse d'effets dramatiques sur la voix (réverbération ou jeu d'éloignement spatial). Plus d'informations sur le corpus sont disponibles en ligne [2].

4. Entraînement

4.1. *Synthèse monolocuteur*

Dans cette étude, le modèle acoustique choisi pour la synthèse monolocuteur est Tacotron 2. Nous utilisons l'implémentation d'ESPNET (Hayashi *et al.*, 2020) qui reproduit l'architecture et les hyperparamètres[3] du Tacotron 2 introduits dans l'article originel (Shen *et al.*, 2018). La seule modification apportée à la recette d'ESPNET est le passage du facteur de réduction de 2 à 1, afin d'augmenter la précision des prédictions, au coût d'une convergence plus longue.

L'apprentissage de Tacotron 2 nécessite une grande quantité de données pour chacun des locuteurs cibles. Dans les données décrites section 3.3, nous utilisons la totalité des données disponibles pour les quatre locuteurs cibles. Pour chacun d'entre eux, 200 échantillons de parole sont conservés pour l'évaluation des systèmes, le reste est utilisé pour l'entraînement.

Les échantillons audio sont convertis en mel-spectrogrammes de dimension 80 avec une fenêtre glissante de taille 1024 trames et un décalage de 221 trames. Les silences en début et en fin des échantillons sont supprimés afin de faciliter la convergence du modèle d'attention du modèle acoustique.

Un modèle Tacotron 2 différent est appris pour chacun des locuteurs indépendamment. Cet apprentissage est effectué pendant 200 époques, avec 200 échantillons du jeu d'entraînement mis de côté pour former un jeu de validation. Un mécanisme d'arrêt prématuré est mis en place (*early stopping*) : si la fonction de coût cesse de diminuer sur le jeu de validation pendant 20 époques, l'apprentissage est interrompu pour éviter tout surapprentissage. En pratique, l'apprentissage s'est arrêté automatiquement autour de 70 époques.

2. `https://sites.google.com/view/machahu`
3. `https://shorturl.at/otK59`

4.2. *Clonage de voix*

Comme discuté en section 2.2.2, l'approche choisie pour le clonage de voix est l'encodage de locuteurs. Le système est donc composé de deux modèles : le modèle x-vecteurs en guise d'encodeur de locuteurs et Tacotron 2 pour le modèle acoustique. L'entraînement est effectué en deux étapes. Le modèle x-vecteurs est d'abord entraîné seul, puis utilisé pour l'entraînement du Tacotron 2. Chacune de ces deux étapes est effectuée sur un corpus différent.

L'implémentation pour le modèle de x-vecteurs est ici celle de l'outil Kaldi ASR[4]. Sa dimension d'entrée est de 23. La taille de ses couches intermédiaires, et donc la dimension des x-vecteurs produits, est de 512. Les autres hyperparamètres correspondent à ceux par défaut de la recette. Il nécessite un très grand nombre de locuteurs d'entraînement, mais est moins sensible à la qualité des données que le modèle acoustique. Nous utilisons ici comme corpus d'entraînement la version française du corpus CommonVoice (Mozilla, 2020 ; Ardila *et al.*, 2020), contenant 682 heures de parole (version de décembre 2020). Du fait de la diversité des moyens d'enregistrement et des environnements sonores, la qualité des échantillons est très variable. Comme nous utilisons la version du modèle x-vecteurs indépendante du texte, seuls les échantillons ont été fournis au modèle, sous forme de MFCC (*Mel Frequency Cepstral Coefficients*). L'entraînement a été réalisé en 420 étapes, sur le jeu d'entraînement par défaut, contenant 3605 locuteurs, avec une moyenne de 70 échantillons par locuteur.

Le modèle acoustique utilisé, Tacotron 2, correspond quant à lui à une implémentation d'ESPNET[5], avec les hyperparamètres par défaut de la recette. Il est entraîné sur le corpus MUFASA, présenté en section 3. Contrairement à la synthèse monolocuteur et à la conversion de voix, le principe du clonage implique que les locuteurs cibles soient absents du corpus d'entraînement. Il est donc entraîné sur les données disponibles pour tous les locuteurs, à l'exception de *Nadine*, *Jean-Luc*, *René* et *Victoria*. Le modèle converge après 30 époques d'entraînement.

4.3. *Conversion de voix*

Le système de conversion de voix utilisé ici repose sur (Zhao *et al.*, 2019) et consiste en trois principaux modèles : un modèle d'extraction de PPG, un modèle de conversion des PPG en mel-sprectrogramme (PPG-to-Mel) et un vocodeur. L'architecture globale du système de conversion de voix est présentée dans la figure 1.

Le modèle d'extraction de PPG repose sur le modèle acoustique d'un système de reconnaissance de parole. Dans cette étude, il s'agit d'un modèle TDNN-HMM (*Time Delay Neural Network - Hidden Markov Model*) (Peddinti *et al.*, 2015) préentraîné sur des données multilocuteurs[6].

4. `https://github.com/kaldi-asr/kaldi/tree/master/egs/sre16/v2`
5. `https://github.com/espnet/espnet/tree/master/egs/libritts`
6. `https://github.com/pguyot/zamia-speech`

Le modèle acoustique (*PPG-to-Mel*) est dérivé de Tacotron 2, et a pour objectif de prédire un spectrogramme à partir des PPG. Chaque locuteur cible nécessite un modèle PPG-to-Mel différent. Pendant la phase d'apprentissage, le modèle PPG-to-Mel est appris spécifiquement pour une voix cible. L'apprentissage est basé sur les PPG dérivés du signal audio de la voix cible et est indépendant de la voix source. Pendant la phase d'inférence, les PPG en entrée du modèle sont extraits du signal de parole source et un modèle PPG-to-Mel, spécifique à la voix cible, les convertit en signal audio. Pour le modèle acoustique, le jeu d'entraînement utilisé est identique à celui utilisé pour l'apprentissage du modèle acoustique de synthèse monolocuteur.

4.4. *Vocodeur*

Pour les trois systèmes de synthèse, le vocodeur WaveGLOW, reposant sur des réseaux à flux est utilisé. Sa génération de signal, à partir de mel-spectrogrammes, est rapide, tout en conservant une haute qualité. Dans cette étude, nous utilisons l'implémentation officielle (Nvidia, 2018).

Ainsi, le modèle WaveGLOW préentrainé publié sur le GitHub officiel[7] est adapté sur le corpus MUFASA. Il est entraîné sur le même corpus que le modèle acoustique de clonage, sur environ 40 époques (372 500 batches), sans les locuteurs cibles, pour éviter de corriger en partie la voix produite, et présenter un biais en faveur du clonage en termes de similarité au locuteur. Le modèle obtenu est utilisé pour toutes les expériences.

5. Protocole expérimental

Nous évaluons les performances des systèmes entraînés dans la section précédente selon deux critères : la qualité globale de la synthèse, et la capacité à reproduire fidèlement la voix des locuteurs cibles. Ces deux critères sont évalués à l'aide de mesures objectives automatiques (MCD et similarité cosinus) et de tests perceptifs (un MOS et un DMOS). Le but de ces évaluations est de commencer par identifier la qualité absolue du rendu pour chacune des trois techniques avant de qualifier leur capacité à reproduire l'identité vocale du locuteur souhaité, et donc par extension leur capacité à rester fidèle aux styles parfois très marqués des locuteurs de livres audio.

5.1. *Évaluation objective*

5.1.1. *Métriques*

Afin d'avoir une première estimation de la performance des différents systèmes, nous avons utilisé deux mesures objectives.

7. `https://github.com/NVIDIA/waveglow`

La première, la distortion mel-cepstrale (*Mel-Cepstral Distortion*, MCD) (Kominek *et al.*, 2008), permet d'estimer la qualité globale d'un système. Elle se calcule sur les coefficients mel-cepstraux généralisés (*Mel-Generalized Coefficients*, MGC) de deux signaux audio. Pour les trois paradigmes étudiés, nous appliquons une déformation temporelle dynamique (*Dynamic Time Warping*, DTW) entre le signal synthétique et le signal naturel de référence pour compenser les différences de durée. Les MGC sont extraits à l'aide des outils SPTK[8] et WORLD (Morise *et al.*, 2016).

La seconde mesure objective est la similarité cosinus entre plongements de locuteurs. On l'utilise ici pour évaluer la similarité entre locuteurs naturels et synthétiques. Elle est calculée à partir de plongements extraits d'un encodeur de locuteurs. Pour éviter de biaiser l'évaluation, nous utilisons un encodeur de locuteur différent de celui utilisé dans le modèle de clonage de voix. Dans cette étude, nous utilisons le modèle sur étagère *Resemblyzer*[9] qui implémente l'encodeur de locuteur présenté dans (Wan *et al.*, 2018). Les plongements de locuteurs à comparer sont calculés sur des signaux de parole synthétique et des signaux de parole naturelle correspondant à la même voix.

5.1.2. *Corpus de test*

Pour chacun des quatre locuteurs cibles définis dans la section 3.3, on sélectionne au moins 200 échantillons audio et leur transcription. Ces données n'ont pas été utilisées lors de l'entraînement des modèles. Les énoncés peuvent varier d'un locuteur à un autre. Ces échantillons sont utilisés pour évaluer objectivement la performance du vocodeur, ainsi que celle des systèmes proposés.

Pour évaluer le système de conversion de voix, il est nécessaire de définir la parole d'un locuteur source différent du locuteur cible. On sélectionne donc un second ensemble de couples texte/audio parmi les données n'ayant pas servi à l'entraînement des modèles acoustiques de conversion de voix. Ce choix est limité par les données parallèles à disposition dans le corpus MUFASA. Cette contrainte n'est pas nécessaire pour les évaluations subjectives qui suivront. Pour cette évaluation objective, *Nadine* et *Victoria*, sont locutrices sources l'une de l'autre. Les locuteurs sources de *Jean-Luc* et de *René* sont des femmes, *Victoria* et *Pomme* (autre locutrice de MUFASA) respectivement. *Pomme* est constante dans la narration et modérément expressive au style direct, alors que son timbre est âgé.

Pour le clonage de voix, les échantillons de référence utilisés pour extraire les x-vecteurs ne sont pas compris dans le corpus de test. Ils sont construits à partir d'échantillons du locuteur cible, sélectionnés au hasard dans le corpus d'entraînement des modèles de synthèse monolocuteur et de conversion de voix, présentés section 4. Ces échantillons sont concaténés jusqu'à obtenir un échantillon de dix minutes. Pour chacun des quatre locuteurs cibles, un seul échantillon de dix minutes est produit afin d'extraire un x-vecteur ; ce dernier est utilisé dans les tests objectifs et perceptifs.

8. `https://sp-tk.sourceforge.net/`
9. `https://github.com/resemble-ai/Resemblyzer`

5.2. *Évaluation subjective*

5.2.1. *Test de qualité (MOS)*

Le test MOS permet d'évaluer la qualité générale de la parole. La question posée au testeur est : « Merci d'écouter l'échantillon à évaluer. Comment jugez-vous la qualité générale de la parole dans cet échantillon ? ». Le testeur a comme échelle de notation : très mauvais - mauvais - moyen - bon - très bon, les notes associées allant de 1 à 5. Le test est composé de deux étapes d'introduction, non évaluées, permettant à l'évaluateur de se familiariser avec la parole produite et le processus de notation, suivies de 100 étapes de tests dont les résultats sont enregistrés. Tous les échantillons sont écoutés au casque ou aux écouteurs. Ce test a été réalisé par 19 testeurs, ce qui conduit à 1643 notes, hors étapes d'introduction.

5.2.2. *Test de similarité locuteur (DMOS)*

La question posée au testeur dans le cadre de ce test est « Merci d'écouter d'abord l'échantillon de référence, puis l'échantillon à évaluer. La voix dans l'échantillon à évaluer vous semble-t-elle proche de celle du locuteur de référence ? Il ne s'agit pas de noter la qualité de l'échantillon mais bien l'identité vocale. »Le testeur a comme échelle de notation : très éloigné - éloigné - moyennement proche - proche - très proche, les notes associées allant de 1 à 5. Les contenus textuels des échantillons de référence sont différents de ceux des échantillons à évaluer. Ce test possède le même nombre d'étapes que le précédent et a été réalisé par 13 testeurs (1204 notes, hors étapes d'introduction). Dans les deux tests d'écoute, les évaluateurs sont recrutés parmi des experts et des non-experts non rémunérés. Les conditions d'écoute ne sont pas contrôlées, mais il est demandé d'utiliser des écouteurs ou un casque audio.

5.2.3. *Corpus de test*

Le corpus est constitué pour chaque locuteur de 50 échantillons ne contenant pas de bruit ni de musique dans le naturel et dont la synthèse correspondante n'a pas échoué (synthèse vide, uniquement pour *René* et *Victoria* en clonage). Les échantillons ainsi sélectionnés sont utilisés dans le cadre des évaluations MOS et DMOS décrites précédemment comme échantillons de référence.

Au cours des deux tests, les échantillons évalués pour chaque locuteur sont des échantillons (1) vocodés à partir de mel-spectrogrammes naturels, (2) synthétisés par le modèle monolocuteur, (3) obtenus par clonage de voix en utilisant comme entrée du modèle x-vecteurs un échantillon de dix minutes de parole (voir section 5.1.2), ou (4) convertis à partir de parole du locuteur de même genre dans le jeu de test. On obtient un total de 800 stimuli (4 systèmes, 4 locuteurs, 50 échantillons). Pour le test DMOS de manière spécifique, nous évaluons aussi des faux positifs sous la forme d'échantillons auto-encodés du même genre que le locuteur dans la référence (200 stimuli supplémentaires).

6. Résultats et discussion

6.1. *Résultats*

Les résultats de l'évaluation objective (mesure de la MCD et de la similarité cosinus entre plongements de locuteurs) sont présentés dans les tableaux 1 et 2. Les résultats de l'évaluation subjective (mesure de la qualité générale et de la similarité entre locuteurs) sont présentés dans les tableaux 3 et 4. Des exemples d'échantillons synthétiques sont mis à disposition en ligne[10].

Locuteurs→ Systèmes ↓	**Nadine**	**Victoria**	**Jean-Luc**	**René**	**Moyenne**
Vocodeur	2,56 ± 0,04	2,52 ± 0,09	1,62 ± 0,08	1,67 ± 0,06	2,06 ± 0,04
Synthèse	4,69 ± 0,09	5,46 ± 0,12	4,78 ± 0,18	4,50 ± 0,13	4,79 ± 0,07
Clonage	3,86 ± 0,10	4,78 ± 0,16	4,81 ± 0,15	3,80 ± 0,14	4,20 ± 0,07
Conversion	4,74 ± 0,25	4,64 ± 0,25	3,26 ± 0,34	5,13 ± 0,22	4,53 ± 0,13
Moyenne	3,87 ± 0,08	4,33 ± 0,11	3,62 ± 0,14	3,74 ± 0,10	

Tableau 1. *Distortion mel-cepstrale (± demi-intervalle de confiance à 95 %), en dB.*

Locuteurs → Systèmes ↓	**Nadine**	**Victoria**	**Jean-Luc**	**René**	**Moyenne**
Vocodeur	0,98 ± 0,00	0,98 ± 0,00	0,98 ± 0,00	0,98 ± 0,00	0,98 ± 0,00
Synthèse	0,74 ± 0,01	0,75 ± 0,01	0,75 ± 0,01	0,78 ± 0,01	0,76 ± 0,00
Clonage	0,62 ± 0,01	0,59 ± 0,01	0,64 ± 0,01	0,59 ± 0,01	0,61 ± 0,01
Conversion	0,75 ± 0,02	0,76 ± 0,01	0,52 ± 0,01	0,48 ± 0,01	0,59 ± 0,01
Moyenne	0,77 ± 0,01	0,77 ± 0,01	0,72 ± 0,01	0,71 ± 0,01	

Tableau 2. *Similarité cosinus entre locuteurs (± demi-intervalle de confiance à 95 %).*

Locuteurs → Systèmes ↓	**Nadine**	**Victoria**	**Jean-Luc**	**René**	**Moyenne**
Vocodeur	4,7 ± 0,1	4,1 ± 0,2	4,4 ± 0,2	3,9 ± 0,2	4,2 ± 0,1
Synthèse	3,7 ± 0,2	2,7 ± 0,2	1,9 ± 0,2	2,5 ± 0,2	2,7 ± 0,1
Clonage	3,0 ± 0,2	2,6 ± 0,2	2,3 ± 0,2	1,7 ± 0,1	2,4 ± 0,1
Conversion	3,2 ± 0,2	2,9 ± 0,2	1,7 ± 0,1	2,6 ± 0,2	2,6 ± 0,1
Moyenne	3,6 ± 0,1	3,1 ± 0,1	2,6 ± 0,1	2,7 ± 0,1	

Tableau 3. *Scores MOS moyens en fonction des systèmes et des locuteurs (± demi-intervalle de confiance à 95 %)*

10. `https://sites.google.com/view/machahu`

Locuteurs → Systèmes ↓	**Nadine**	**Victoria**	**Jean-Luc**	**René**	**Moyenne**
Vocodeur	4,3 ± 0,3	4,3 ± 0,3	4,3 ± 0,3	4,9 ± 0,2	4,5 ± 0,1
Synthèse	3,9 ± 0,3	3,8 ± 0,3	3,0 ± 0,3	4,3 ± 0,3	3,7 ± 0,2
Clonage	3,1 ± 0,3	2,0 ± 0,2	2,0 ± 0,2	1,8 ± 0,3	2,2 ± 0,1
Conversion	3,6 ± 0,3	3,9 ± 0,3	2,4 ± 0,3	4,5 ± 0,2	3,6 ± 0,2
Moyenne	3,7 ± 0,2	3,5 ± 0,2	3,0 ± 0,2	3,9 ± 0,2	

Tableau 4. *Scores DMOS moyens en fonction des systèmes et des locuteurs (± demi-intervalle de confiance à* 95 %*)*

Les mesures obtenues pour le vocodeur nous renseignent sur l'impact du vocodeur sur la qualité générale lorsqu'il sera intégré à la chaîne de traitement pour la synthèse, la conversion et le clonage. Ce système représente la meilleure qualité atteignable par le vocodeur et donc la borne haute de ce qu'il est possible de reproduire à l'aide du modèle acoustique. Le système obtient une MCD moyenne de 2,06 dB sur les quatre locuteurs, ce qui est comparable à l'état de l'art (Hsu et Lee, 2020). Lors du test d'écoute, son score MOS est 4,2, ce qui peut paraître légèrement bas en comparaison avec l'état de l'art. Ceci n'est pas surprenant puisqu'il s'agit ici d'un vocodeur multilocuteur entraîné sur des données issues de livres audio non professionnels. De plus, ce score varie énormément en fonction du locuteur. Le vocodeur obtient un score MOS de 4,7 pour *Nadine*, ce qui laisse présager de très bons résultats des systèmes synthétiques pour cette voix. Il obtient un score MOS plus bas pour *Jean-Luc* (4,4) et des scores décevants pour les voix de *René* et de *Victoria* (autour de 4).

Le vocodeur obtient une similarité cosinus de 0,98 pour l'ensemble des locuteurs et un score DMOS de 4,5. Cela suggère que notre vocodeur est capable de reproduire fidèlement la voix des locuteurs de test. De manière surprenante, le vocodeur obtient un score DMOS identique pour les voix de *Nadine*, de *Victoria* et de *Jean-Luc* mais *René* obtient une note significativement plus élevée. Cela peut s'expliquer par le fait que le caractère atypique de la voix de *René* et de son élocution est si accentué qu'il favorise grandement sa reconnaissance par les testeurs.

En moyenne, le système de synthèse monolocuteur voit une augmentation significative de la MCD et une diminution significative du MOS par rapport au vocodeur. Cela correspond à la diminution de la qualité globale de la synthèse, attendue en raison des erreurs de prédictions introduites par le modèle acoustique. Bien que la MCD de 4,79 dB ne soit pas surprenante, un score MOS de 2,7 est en deçà de l'état de l'art (Weiss *et al.*, 2021). Cependant, celui-ci varie d'un locuteur à l'autre. *Nadine* obtient un score MOS de 3,7 comparable à l'état de l'art pour des données similaires en anglais (Zen *et al.*, 2019). En revanche, *Victoria*, *Jean-Luc* et *René* obtiennent des scores MOS inférieurs à 3. Cela peut s'expliquer par le caractère dégradé de leurs données. L'expressivité n'étant pas modélisée explicitement, la qualité de l'apprentissage sur les voix atypiques de *René* et de *Victoria* est probablement entravée. Il est surprenant

que la synthèse monolocuteur obtienne les meilleurs scores MOS malgré une mauvaise MCD. Comme dans Weiss *et al.* (2021), les systèmes présentant les meilleurs scores MOS n'obtiennent pas toujours les meilleures MCD, probablement à cause de l'alignement DTW.

Le système de synthèse monolocuteur subit aussi une diminution significative de la similarité cosinus et du score DMOS par rapport au vocodeur. Cela traduit une diminution globale de la fidélité de la reproduction d'une voix due au modèle acoustique. La similarité cosinus moyenne est élevée (0,76) et varie peu d'un locuteur à l'autre (celle de *René* restant légèrement supérieure). Le système de synthèse obtient un score DMOS moyen de 3,7 avec une plus grande variabilité en fonction du locuteur. Il n'y a pas de différence significative entre *Nadine*, *Victoria* et *René*. *Jean-Luc* obtient cependant une note significativement plus basse. Il est intéressant de noter que cette voix synthétique a aussi obtenu le moins bon score MOS. Il y a probablement une corrélation entre le score MOS pour la qualité globale et le DMOS pour la similarité au locuteur. En effet, il est difficile de noter la similarité d'un échantillon à un locuteur cible lorsque la qualité globale de cet échantillon est mauvaise. Ces mesures suggèrent que la synthèse monolocuteur est capable de reproduire des voix, même expressives, tant que la quantité de données pour le locuteur est suffisante. Cependant, les résultats pour *Jean-Luc* suggèrent qu'avoir des données de qualité reste important.

Comme la synthèse monolocuteur, le clonage de voix montre une diminution de la qualité globale par rapport au vocodeur qui peut être observée par une dégradation de la MCD et du score MOS. En moyenne, le clonage est aussi significativement moins bon que la synthèse monolocuteur en termes de MOS. Ceci n'est pas surprenant puisque le clonage est une tâche plus complexe que la synthèse monolocuteur. Il est cependant intéressant de noter que la tendance est inversée dans le cas de *Jean-Luc*. Le clonage pourrait ainsi s'avérer intéressant dans les cas où les données du locuteur cible sont perturbées de façon ponctuelle.

On observe aussi une diminution significative des performances du système de clonage en ce qui concerne la fidélité de la reproduction des voix des locuteurs cibles en comparaison avec le vocodeur et la synthèse monolocuteur. Le faible score de similarité cosinus (0,6 en moyenne) est confirmé par les résultats du test d'écoute DMOS (2,2 en moyenne). À part pour *Nadine*, les testeurs ne semblent pas avoir été capables de reconnaître la voix des locuteurs cibles. Bien qu'en théorie, le clonage de voix permette de reproduire fidèlement la voix de locuteurs non vus lors de l'apprentissage, sa capacité de généralisation ne semble pas bonne sur des données de livres audio amateurs français. Trois causes seraient possibles : la qualité générale des données (enregistrement amateur), l'expressivité des locuteurs (lecture de contenu non neutre), le nombre limité de locuteurs disponibles dans cette première version du corpus.

Le système de conversion de voix subit lui aussi une dégradation de la qualité générale par rapport au vocodeur, mais ne présente pas de différence significative avec le système de synthèse monolocuteur en termes de MOS moyen. Ceci n'est pas surprenant puisque les modèles acoustiques de ces deux systèmes ont été entraînés sur les mêmes quantités de données des locuteurs cibles. Il n'y a pas non plus de différence

significative entre la conversion de voix et le clonage de voix en termes de MOS. Cela suggère que la conversion de voix n'apporte pas d'amélioration ou de dégradation significative par rapport aux deux autres paradigmes pour la qualité globale.

Enfin, la conversion de voix présente une dégradation globale par rapport au vocodeur en ce qui concerne la fidélité de la reproduction de la voix des locuteurs cibles. Relativement à la similarité cosinus moyenne, la conversion de voix est inférieure à la synthèse monolocuteur et comparable au clonage de voix. Cela dit, locuteur par locuteur, les voix de *Nadine* et de *Victoria* obtiennent des scores similaires à la synthèse monolocuteur, et les voix de *Jean-Luc* et de *René* obtiennent des performances significativement dégradées par rapport à tous les autres systèmes. En moyenne et locuteur par locuteur, le score DMOS de la conversion de voix n'est pas significativement distinguable de celui pour la synthèse monolocuteur. Bien que le score DMOS moyen de la conversion de voix soit significativement supérieur à celui du clonage de voix (3,6 et 3,4 respectivement), les différences ne sont significatives que pour la moitié des locuteurs. Ces observations suggèrent que la conversion de voix n'apporte pas non plus d'amélioration ou de dégradation significative pour la similarité au locuteur par rapport aux deux autres paradigmes.

6.2. *Discussion*

Locuteur	**Jitter**	**Shimmer**	F_0 (Hz)	F_0 **min**	F_0 **max**	**HNR**
Nadine	2,54	1,10	$187{,}27 \pm 41$	90,87	503,25	8,03
Victoria	2,39	1,07	$199{,}59 \pm 51$	89,39	507,16	7,09
René	3,68	1,44	$145{,}85 \pm 34$	44,90	299,91	3,92
Jean-Luc	4,37	1,39	$116{,}91 \pm 37$	43,28	311,53	3,18

Tableau 5. *Mesures de Jitter (%), Shimmer (dB), F_0 (moyenne avec écart-type, minimum et maximum) ainsi que le rapport signal à bruit (HNR, dB) pour les 4 locuteurs de test. Jitter, Shimmer et HNR sont calculés avec OpenSmile. Les attributs liés au F_0 sont calculés avec Praat.*

L'objectif de cette étude était d'entraîner des modèles de synthèse de parole sur des données du tout-venant issues de livres audio amateurs. L'état de l'art en matière de synthèse monolocuteur montre que cette technologie est capable de produire de la parole de qualité reproduisant fidèlement la voix de locuteurs. En revanche, ces études sont souvent réalisées sur des données enregistrées spécifiquement pour entraîner un système de synthèse de parole. Dans ces travaux, nous avons montré que l'approche monolocuteur est sensible aux données d'apprentissage. Pour la voix de *Nadine*, le système appris obtient des résultats similaires à ceux de l'état de l'art alors que les systèmes appris sur les trois autres voix obtiennent des résultats bien inférieurs. Il convient alors de se poser la question de l'origine de la variabilité au sein de nos résultats. La quantité de données n'explique pas cette variabilité, puisque chaque locuteur dispose d'un nombre d'heures de parole comparable. La qualité des données

en revanche semble être un facteur important. En effet, les échantillons de *René* et de *Jean-Luc* sont plus bruités (HNR, tableau 5) que ceux de *Nadine* et mènent à des systèmes moins performants. Cependant, *Victoria* donne des résultats significativement moins bons que ceux de *Nadine* malgré un HNR proche. Il reste donc un autre facteur impactant la qualité de l'apprentissage. Nous pensons qu'il s'agit de l'expressivité. Ainsi, la prosodie de *Victoria* est moins régulière et plus stratégique dans la narration que celle de *Nadine*, elle est aussi plus expressive au style direct. Malheureusement, ces aspects de la parole sont difficilement quantifiables et mesurables.

Ces remarques à propos de l'impact de la qualité des données s'appliquent aussi au clonage et à la conversion de voix. Ainsi, les méthodes présentes dans l'état de l'art ne sont pas actuellement applicables à tous les types de données. Pour pallier ce défaut, deux pistes sont possibles. La première consiste à améliorer les modèles acoustiques pour les rendre plus robustes aux bruits et à l'expressivité présente dans les données d'apprentissage. Par exemple, la modélisation à l'aide d'un auto-encodeur variationel pourrait être une option prometteuse. La seconde piste consiste à travailler sur les données, en mettant en place des procédures de sélection automatique en fonction de leur qualité.

Le clonage de voix est soumis à un impératif supplémentaire en termes de données. Les travaux de la communauté montrent qu'un système de clonage est capable de reproduire n'importe quelle voix à partir de courts échantillons, lorsqu'ils sont entraînés sur une (ou plusieurs) centaine(s) de locuteurs. Le français ne dispose malheureusement pas, à notre connaissance, d'un jeu de données libre de droit contenant de la parole de qualité provenant d'autant de locuteurs et les résultats obtenus dans cet article montrent que la vingtaine de locuteurs présents dans le corpus MUFASA n'est pas un nombre suffisant pour généraliser à n'importe quelle voix de locuteur inconnu. Une solution naïve consisterait à ajouter plus de locuteurs au corpus MUFASA. Cependant, dans le cas de données du tout-venant, bien que de nouvelles données soient faciles à trouver, leur qualité reste un problème crucial et non trivial comme discuté précédemment. Une autre solution serait de travailler sur des modèles multilingues pour tirer profit des larges jeux de données disponibles en anglais.

Enfin, il est intéressant de noter que la conversion de voix obtient des résultats similaires à la synthèse monolocuteur malgré son cas d'usage différent. Cette observation n'est pas surprenante si l'on s'en tient au fait que les deux paradigmes ont été entraînés sur les mêmes données (mêmes locuteurs, même quantité). Cela indique cependant que, dans notre cas, l'impact de la qualité des données est plus important que celui des modalités d'entrée. Nous n'avons pas mesuré de différences significatives entre l'utilisation de séquences phonétiques issues du texte d'une part et l'utilisation de PPG issus de l'audio d'autre part, en tant qu'entrée d'un modèle acoustique. Bien que les PPG encodent des informations prosodiques en plus des informations phonétiques, ces informations supplémentaires n'ont pas eu d'impact, positif ou négatif, sur les performances du système.

7. Conclusion

La synthèse de la parole est en règle générale effectuée à partir de corpus de données de qualité construits spécifiquement pour la tâche de synthèse. La mitigation de cette contrainte sur la qualité et sur l'adéquation des données d'entraînement utilisées en synthèse vocale déverrouillerait un grand potentiel. Ceci dit, il convient avant tout de faire un état des lieux des performances actuelles avec de telles données. Dans cette étude, nous nous sommes focalisés sur le cas de la langue française.

Nous avons évalué la capacité de trois paradigmes de synthèse de parole à produire des livres audio à partir de données du même type. La synthèse monolocuteur est capable de reproduire fidèlement la voix du locuteur d'entraînement mais la qualité globale de la synthèse est grandement impactée par le niveau d'expressivité de ce même locuteur. Le clonage de voix propose une approche intéressante pour baisser le coût de la production de la synthèse vocale. Cependant, ce paradigme souffre du même défaut sur la qualité globale et présente une forte complexité pour produire un modèle acoustique capable de restituer fidèlement la voix du locuteur cible. Enfin, la conversion de voix semble une piste prometteuse car offrant des performances proches de la synthèse monolocuteur.

De futurs travaux sont à mener pour améliorer la qualité des systèmes. Tout d'abord, l'amélioration du procédé de sélection des données au sein des corpus de synthèse reste une étape importante pour une meilleure maîtrise des propriétés de la voix synthétique. On peut également noter l'intérêt de raffiner les procédures automatiques naissantes dans la communauté pour sélectionner les meilleures données parmi une large quantité disponible. Un autre axe d'étude est la modélisation et un contrôle explicite de l'expressivité. Cela peut se faire par des contraintes explicites lors de la construction du modèle acoustique. L'intégration de méthodes génératives capables de tirer parti des spécificités des données à plusieurs échelles est une solution potentielle.

Remerciements

Ces travaux ont bénéficié d'un accès aux moyens de calcul de l'IDRIS au travers de l'allocation de ressources 2023-AD011011870R2 attribuée par GENCI.

8. Bibliographie

Abe M., Nakamura S., Shikano K., Kuwabara H., « Voice conversion through vector quantization », *Journal of the Acoustical Society of Japan (E)*, 1990.

Ardila R., Branson M., Davis K., Kohler M., Meyer J., Henretty M., Morais R., Saunders L., Tyers F., Weber G., « Common Voice : A Massively-Multilingual Speech Corpus », *LREC*, 2020.

Arik S., Chen J., Peng K., Ping W., Zhou Y., « Neural voice cloning with a few samples », *Advances in Neural Information Processing Systems : Annual Conf. on Neural Information Processing Systems*, 2018.

Arık S. Ö., Chrzanowski M., Coates A., Diamos G., Gibiansky A., Kang Y., Li X., Miller J., Ng A., Raiman J. *et al.*, « Deep voice : Real-time neural text-to-speech », *Int. Conf. on Machine Learning*, 2017.

Chen Y., Assael Y., Shillingford B., Budden D., Reed S., Zen H., Wang Q., Cobo L. C., Trask A., Laurie B., Gulcehre C., van den Oord A., Vinyals O., de Freitas N., « Sample Efficient Adaptive Text-to-Speech », *Int. Conf. on Learning Representations*, 2019.

Elias I., Zen H., Shen J., Zhang Y., Jia Y., Weiss R. J., Wu Y., « Parallel tacotron : Non-autoregressive and controllable tts », *ICASSP*, 2021.

Erro D., Moreno A., Bonafonte A., « INCA algorithm for training voice conversion systems from nonparallel corpora », *IEEE Tr. on Audio, Speech, and Language Processing*, 2009.

Hayashi T., Yamamoto R., Inoue K., Yoshimura T., Watanabe S., Toda T., Takeda K., Zhang Y., Tan X., « Espnet-TTS : Unified, reproducible, and integratable open source end-to-end text-to-speech toolkit », *ICASSP*, 2020.

Hazen T. J., Shen W., White C., « Query-by-example spoken term detection using phonetic posteriorgram templates », *IEEE Workshop on Automatic Speech Recognition Understanding*, 2009.

Hinterleitner F., Manolaina C., Möller S., « Influence of a voice on the quality of synthesized speech », *2014 Sixth International Workshop on Quality of Multimedia Experience*, 2014.

Ho T. V., Akagi M., « Non-parallel Voice Conversion based on Hierarchical Latent Embedding Vector Quantized Variational Autoencoder », *Blizzard Challenge Workshop*, 2020.

Honnet P.-E., Lazaridis A., Garner P. N., Yamagishi J., The siwis french speech synthesis database ? design and recording of a high quality french database for speech synthesis, Technical report, Idiap, 2017.

Hsu P.-c., Lee H.-y., « WG-WaveNet : Real-Time High-Fidelity Speech Synthesis Without GPU », *Interspeech*, 2020.

Ito K., Johnson L., « The LJ Speech Dataset », , https ://keithito.com/LJ-Speech-Dataset/, 2017.

ITU-T, ITU-T Recommendation P.800, Technical report, International Telecommunication Union, 1996.

Jia Y., Zhang Y., Weiss R., Wang Q., Shen J., Ren F., Chen Z., Nguyen P., Pang R., Moreno I., Wu Y., « Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis », *Neural Information Processing Systems Conf.*, 2018.

Kominek J., Black A. W., « The CMU Arctic speech databases », *SSW 5*, 2004.

Kominek J., Schultz T., Black A. W., « Synthesizer voice quality of new languages calibrated with mean mel cepstral distortion. », *SLTU*, 2008.

Lamel L. F., Gauvain J.-L., Eskénazi M., « Bref, a large vocabulary spoken corpus for french », *Eurospeech*, 1991.

Le Moine C., Obin N., « Att-HACK : An Expressive Speech Database with Social Attitudes », *Speech Prosody*, 2020.

Ling Z.-H., Zhou X., King S., « The Blizzard Challenge 2021 », *Blizzard Challenge Workshop*, 2021.

Liu S., Cao Y., Wang D., Wu X., Liu X., Meng H., « Any-to-Many Voice Conversion With Location-Relative Sequence-to-Sequence Modeling », *IEEE/ACM Tr. on Audio, Speech and Language Processing*, 2021.

Lo C.-C., Fu S.-W., Huang W.-C., Wang X., Yamagishi J., Tsao Y., Wang H.-M., « MOSNet : Deep Learning based Objective Assessment for Voice Conversion », *Interspeech*, 2019.

Lorenzo-Trueba J., Yamagishi J., Toda T., Saito D., Villavicencio F., Kinnunen T., Ling Z., « The Voice Conversion Challenge 2018 : Promoting Development of Parallel and Nonparallel Methods », *Speaker Odyssey 2018*, ISCA, p. 195-202, June, 2018.

Mehri S., Kumar K., Gulrajani I., Kumar R., Jain S., Sotelo J., Courville A. C., Bengio Y., « SampleRNN : An Unconditional End-to-End Neural Audio Generation Model », *5th International Conference on Learning Representations, ICLR 2017, Toulon, France*, 2017.

Morise M., Yokomori F., Ozawa K., « WORLD : a vocoder-based high-quality speech synthesis system for real-time applications », *IEICE Tr. on Information and Systems*, 2016.

Mozilla, « CommonVoice », , `https://commonvoice.mozilla.org/`, 2020.

Nvidia, « Waveglow Github repository », , `https://github.com/NVIDIA/waveglow/`, 2018.

Panayotov V., Chen G., Povey D., Khudanpur S., « Librispeech : An ASR corpus based on public domain audio books », *ICASSP*, 2015.

Peddinti V., Povey D., Khudanpur S., « A time delay neural network architecture for efficient modeling of long temporal contexts », *Interspeech*, 2015.

Ping W., Peng K., Chen J., « ClariNet : Parallel Wave Generation in End-to-End Text-to-Speech », *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, OpenReview.net, 2019.

Ping W., Peng K., Gibiansky A., Arik S. O., Kannan A., Narang S., Raiman J., Miller J., « Deep Voice 3 : 2000-Speaker Neural Text-to-Speech », *Int. Conf. on Learning Representations*, 2018.

Prenger R., Valle R., Catanzaro B., « Waveglow : A flow-based generative network for speech synthesis », *ICASSP*, 2019.

Ren Y., Hu C., Tan X., Qin T., Zhao S., Zhao Z., Liu T.-Y., « FastSpeech 2 : Fast and High-Quality End-to-End Text to Speech », *Int. Conf. on Learning Representations*, 2020.

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., « Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions », *ICASSP*, 2018.

Sini A., Lolive D., Vidal G., Tahon M., Delais-Roussarie É., « SynPaFlex-Corpus : An Expressive French Audiobooks Corpus dedicated to expressive speech synthesis. », *LREC*, 2018.

Sisman B., Yamagishi J., King S., Li H., « An Overview of Voice Conversion and Its Challenges : From Statistical Modeling to Deep Learning », *IEEE/ACM Tr. on Audio, Speech, and Language Processing*, 2021.

Snyder D., Garcia-Romero D., Povey D., Khudanpur S., « Deep Neural Network Embeddings for Text-Independent Speaker Verification », *Interspeech*, 2017.

Sun L., Li K., Wang H., Kang S., Meng H., « Phonetic posteriorgrams for many-to-one voice conversion without parallel data training », *IEEE Int. Conf. on Multimedia and Expo*, 2016.

Tan X., Qin T., Soong F., Liu T.-Y., « A Survey on Neural Speech Synthesis », *arXiv preprint arXiv :2106.15561v3*, 2021.

Tian X., Wang J., Xu H., Chng E.-S., Li H., « Average Modeling Approach to Voice Conversion with Non-Parallel Data », *Speaker and Language Recognition Workshop (Odyssey)*, 2018.

Tobing P. L., Wu Y.-C., Toda T., « Baseline System of Voice Conversion Challenge 2020 with Cyclic Variational Autoencoder and Parallel WaveGAN », *Blizzard Challenge Workshop*, 2020.

Toda T., Black A. W., Tokuda K., « Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory », *IEEE Tr. on Audio, Speech, and Language Processing*, 2007.

van den Oord A., Li Y., Babuschkin I., Simonyan K., Vinyals O., Kavukcuoglu K., Driessche G., Lockhart E., Cobo L., Stimberg F. *et al.*, « Parallel wavenet : Fast high-fidelity speech synthesis », *Int. Conf. on Machine Learning*, 2018.

Veaux C., Yamagishi J., MacDonald K., CSTR VCTK corpus : English multi-speaker corpus for CSTR voice cloning toolkit, Technical report, University of Edinburgh. CSTR, 2017.

Vích R., Nouza J., Vondra M., « Automatic speech recognition used for intelligibility assessment of text-to-speech systems », *Verbal and Nonverbal Features of Human-Human and Human-Machine Interaction*, Springer, p. 136-148, 2008.

Wagner P., Beskow J., Betz S., Edlund J., Gustafson J., Eje Henter G., Le Maguer S., Malisz Z., Székely E., Tånnander C., Voße J., « Speech Synthesis Evaluation — State-of-the-Art Assessment and Suggestion for a Novel Research Program », *SSW 10*, 2019.

Wan L., Wang Q., Papir A., Moreno I. L., « Generalized end-to-end loss for speaker verification », *ICASSP*, 2018.

Wang H., Soong F., Meng H., « Aa spectral space warping approach to cross-lingual voice transformation in hmm-based tts », *ICASSP*, 2015.

Weiss R. J., Skerry-Ryan R., Battenberg E., Mariooryad S., Kingma D. P., « Wave-tacotron : Spectrogram-free end-to-end text-to-speech synthesis », *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, p. 5679-5683, 2021.

Yamamoto R., Song E., Kim J.-M., « Parallel WaveGAN : A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram », *ICASSP*, 2020.

Yi Z., Huang W.-C., Tian X., Yamagishi J., Das R. K., Kinnunen T., Ling Z.-H., Toda T., « Voice Conversion Challenge 2020 –- Intra-lingual semi-parallel and cross-lingual voice conversion –- », *Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020*, ISCA, oct, 2020.

Zen H., Dang V., Clark R., Zhang Y., Weiss R. J., Jia Y., Chen Z., Wu Y., « LibriTTS : A Corpus Derived from LibriSpeech for Text-to-Speech », *Interspeech*, 2019.

Zhao G., Ding S., Gutierrez-Osuna R., « Foreign Accent Conversion by Synthesizing Speech from Phonetic Posteriorgrams », *Interspeech*, 2019.

Zheng L., Tao J., Wen Z., Zhong R., « CASIA Voice Conversion System for the Voice Conversion Challenge 2020 », *Blizzard Challenge Workshop*, 2020.