Conversion d'Identité de la Voix Chantée par Sélection et Concaténation d'Unités Spectrales

Nicolas Obin, Pascal Pham, Axel Roebel

Conversion d'Identité de la Voix Chantée par Sélection et Concaténation d'Unités Spectrales

Nicolas Obin, Pascal Pham, Axel Roebel
IRCAM, Sorbonne Université, CNRS, Paris, France
{nobin, pham, roebel}@ircam.fr

RÉSUMÉ

Cet article présente un algorithme de sélection d'unités spectrales pour la conversion de l'identité de la voix chantée à partir de bases de données non parallèles. Les algorithmes de conversion basés sur des unités de parole présentent des avantages importants pour la conversion de l'identité vocale : la conversion vocale par sélection d'unités permet la préservation des caractéristiques originales de la voix cible, en utilisant des unités réelles ; et la segmentation en unités linguistiques permet d'apprendre la conversion à partir d'enregistrements de la voix cible non nécessairement alignés avec ceux de la voix source. La contribution principale de cet article est de réaliser la sélection des unités spectrales de la voix cible en fonction de plusieurs facteurs : acoustique, linguistique (phonèmes) et musicaux (hauteur, intensité et durée). Pour ce faire, la sélection de la séquence d'unités d'enveloppe spectrale est établie comme un problème d'optimisation à partir d'une fonction de coût multiple qui comprend la distorsion spectrale des chanteurs source et cible ainsi que les différences de hauteur, d'intensité et de durée des unités spectrales correspondantes. L'objectif est de guider la sélection vers des enveloppes spectrales du chanteur cible partageant un contexte musical similaire avec celles du chanteur source. Il est montré lors d'une expérience perceptive que l'algorithme proposé améliore le naturel de la conversion et la similarité avec la voix cible.

ABSTRACT

This paper presents a unit-selection algorithm for non-parallel singing voice conversion. Unit-based algorithms presents important advantages for voice conversion : the speech segmentation into linguistic units allows the possibility to learn the conversion from on-the-fly databases of the target voice not necessarily aligned to the source voice, and unit-selection voice conversion allows the preservation of the original characteristics of the target voice, by using real units. The main idea of this paper is that the spectral envelopes of a speaker vary according to multiple factors : linguistics (phonemes), and musical (pitch, intensity, and duration). Accordingly, the selection of the sequence of spectral envelope units is established as a multi-target optimization problem, including the spectral distortion of the source and target singers, and the pitch, intensity, and duration differences of the corresponding spectral envelopes. The objective is to guide the selection towards spectral envelopes of the source and target singers sharing a similar musical context. It is shown that the proposed algorithm improves conversion naturalness and target similarity.

MOTS-CLÉS : conversion de l'identité vocale, voix chantée, conversion non-parallèle, sélection d'unités, optimisation multi-cible.

KEYWORDS: voice conversion, singing voice, non-parallel conversion, unit-selection, multi-target

optimization.

1 Introduction

La conversion d'identité vocale consiste à modifier la voix d'un locuteur source afin d'être perçue comme celle d'un locuteur cible. Grâce aux avancées récentes, la conversion vocale a considérablement gagné en popularité et en qualité au cours des dernières années, menant notamment aux premières compétitions internationales sur la conversion d'identité vocale (Toda *et al.*, 2016; Lorenzo-Trueba *et al.*, 2018), et avec son extension à la conversion vocale entre des langues différentes (Sündermann *et al.*, 2006; Nakashika *et al.*, 2016; Kinnunen *et al.*, 2017) et la conversion de voix chantée (Villavicencio & Bonada, 2010; Villavicencio & Kenmochi, 2011; Doi *et al.*, 2012; Kobayashi & Toda, 2014). La conversion vocale a un large éventail d'applications : du divertissement (parler avec la voix d'une autre personne, par exemple via des applications mobiles), créative (reconstruire la voix de personnalités), et médicale (≪réparation vocale≫ pour les personnes présentant un handicap vocal). Considérant la conversion de voix chantée, les applications créatives sont importantes dans l'industrie musicale : du karaoké à la production musicale jusqu'aux chanteurs virtuels, en contrôlant l'identité d'un chanteur réel ou artificiel (Villavicencio & Bonada, 2010; Kenmochi, 2010).
Les algorithmes de conversion d'identité vocale reposent principalement sur la conversion spectrale : la conversion du signal vocal est limitée à la conversion du timbre représenté au moyen d'enveloppes spectrales. La conversion de la voix consiste alors à apprendre une fonction de conversion entre l'espace acoustique d'une voix source et d'une voix cible. La fonction de conversion est généralement modélisée par des modèles statistiques, historiquement avec les modèles de mélange Gaussiens (GMM, (Stylianou *et al.*, 1998)) et plus récemment avec des réseaux de neurones (Desai *et al.*, 2009; Sun *et al.*, 2015). La fonction de conversion est alors apprise à partir d'une base de données pré-alignée (dite ≪parallèle≫) dans laquelle les voix source et cible ont prononcé le même ensemble de phrases, de sorte qu'une correspondance directe entre les trames des voix source et cible puisse être établie pour l'apprentissage.
Ce paradigme de conversion de la voix présente cependant des limites importantes et bien connues : les effets de sur-apprentissage et de moyennage relatifs à la modélisation statistique (Toda *et al.*, 2007) qui conduit à une dégradation de la voix convertie, et la nécessité de construire des bases de données parallèles extrêmement restrictives et non souhaitées pour des applications réelles. Pour répondre à ces limitations, des algorithmes de conversion à partir d'unités - ou d'exemples réels - ont été récemment proposés (Sündermann *et al.*, 2006; Wu *et al.*, 2013; Aihara *et al.*, 2014; Jin *et al.*, 2016). Tout d'abord, la conversion vocale à partir de sélection et de concaténation d'unités spectrales présente l'avantage de préserver les caractéristiques et la dynamique d'origine de la voix cible dans la mesure où elle repose uniquement sur l'utilisation d'unités vocales réelles. Deuxièmement, la segmentation de la parole en unités linguistiques telles que les phonèmes (voir, par exemple, la conversion vocale entre des langues différentes (Sündermann *et al.*, 2006; Nakashika *et al.*, 2016; Kinnunen *et al.*, 2017)) offre la possibilité d'utiliser des bases de données ≪à la volée≫ (dite ≪non-parallèles≫) des voix source et cible. Néanmoins, le choix de la fonction de coût utilisée pour la sélection d'unités constitue un défi important de la conversion par sélection d'unités : la mesure de la distorsion spectrale entre les voix source et cible (Sündermann *et al.*, 2006) peut être efficace dans une certaine mesure, notamment lorsque peu de données de la voix cible sont disponibles. En revanche, elle n'en demeure pas moins limitée par définition (Sündermann *et al.*, 2007) : en effet, la voix la plus proche spectralement de la voix source ne serait, à la limite, qu'elle-même. Cette limitation démontre la nécessité de considérer d'autres facteurs pour la conversion par sélection d'unités.

Les algorithmes de conversion de la voix chantée reposent sur le même paradigme que ceux utilisés pour la voix parlée (Villavicencio & Bonada, 2010; Villavicencio & Kenmochi, 2011; Doi *et al.*, 2012; Kobayashi & Toda, 2014), sans vraiment tenir compte des spécificités de la voix chantée telles que le registre étendu de hauteur, d'intensité et de durée de la voix chantée par comparaison à la voix parlée. Cet article propose un algorithme de conversion de la voix par sélection et concaténation d'unités spectrales avec un focus sur la conversion de la voix chantée. La principale contribution de l'article repose sur l'observation que les enveloppes spectrales d'un chanteur varient en fonction de plusieurs facteurs : linguistique (phonèmes), et musicale (hauteur, intensité et durée) (voir par exemple (Joliveau *et al.*, 2005) sur la voix chantée). En conséquence, la correspondance des unités spectrales d'un chanteur source et d'un chanteur cible doit être recherchée autour d'un contexte musical similaire (par exemple, mêmes hauteurs, intensités, et durées). L'algorithme de conversion vocale proposé est basé sur la sélection d'unités de phonèmes, capitalisant les avantages de la conversion par sélection d'unités et de la conversion de voix non-parallèle. Pour intégrer les spécificités de la voix chantée dans l'algorithme de sélection d'unités, une fonction de coût multiple est établie à partir de : la distorsion spectrale entre les chanteurs source et cible, et les informations musicales telles que les différences de hauteur, d'intensité et de durée entre les chanteurs source et cible. Cette fonction de coût multiple est définie afin de guider la sélection vers des enveloppes spectrales provenant de contextes musicaux similaires, et ainsi d'augmenter la correspondance entre les espaces acoustiques des chanteurs source et cible relativement à ces contextes. En d'autres termes et par analogie, les enveloppes spectrales sélectionnées de la voix cible doivent représenter pour la voix cible ce que les enveloppes spectrales de la voix source représentent pour la voix source.
L'article est organisé de la manière suivante : la conversion de voix par sélection d'unité est présentée dans la Section 2 et l'algorithme de sélection d'unités proposé avec la fonction de coût multi-cible est détaillé dans la Section 2.1. Une expérience perceptive est rapportée dans la Section 3 pour évaluer la conversion multi-cible et quelques variantes sur une tâche de conversion appliquée sur la sortie d'un synthétiseur de voix chantée.

2 Conversion de l'Identité de la Voix Chantée

Cette section présente un rapide aperçu des algorithmes de conversion vocale basés sur la sélection d'unités d'enveloppes spectrales, suivi d'une description détaillée de l'algorithme de conversion vocale concaténative non parallèle (coVoC) (voir (Lorenzo-Trueba *et al.*, 2018)). L'algorithme est basé sur une bases de données non-parallèle des chanteurs source et cible. Les bases de données comprennent des analyses des signaux vocaux tels que la fréquence fondamentale F0, l'intensité, les enveloppes spectrales calculées en échelle de fréquence Mel, et des transcriptions phonétiques alignées sur le signal vocal. Les principales caractéristiques de l'algorithme coVoC sont : l'exploitation des unités de phonèmes, la normalisation des différences spectrales des voix source et cible, et la nouvelle fonction de coût multi-cible qui intègre les connaissances musicales sur les chanteurs source et cible.

2.1 Conversion de l'identité par sélection d'unités spectrales

Le problème général de la conversion de l'identité vocale consiste à déterminer la séquence la plus vraisemblable des enveloppes spectrales de la voix cible $\mathbf{x}^{tgt} = [\mathbf{x}_1^{tgt}, \ldots, \mathbf{x}_T^{tgt}]$ à partir de la séquence

observée des enveloppes spectrales de la voix source $\mathbf{x}^{src} = [\mathbf{x}_1^{src}, \ldots, \mathbf{x}_T^{src}]$:

$$\widehat{\mathbf{x}}^{tgt} = \underset{\mathbf{x}^{tgt}}{\operatorname{argmax}}\ p(\mathbf{x}^{tgt}|\mathbf{x}^{src}) \tag{1}$$

La solution pour la conversion par sélection d'unités (Sündermann *et al.*, 2006) est obtenue classiquement par minimisation d'une fonction de coût comprenant un coût cible défini pour chaque trame et un coût de concaténation défini entre des trames successives :

$$p(\mathbf{x}^{tgt}|\mathbf{x}^{src}) = \sum_{t=1}^{T} \mathcal{C}^t(\mathbf{x}_t^{tgt}, \mathbf{x}_t^{src}) + \mathcal{C}^c(\mathbf{x}_t^{tgt}, \mathbf{x}_{t-1}^{tgt}) \tag{2}$$

où $\mathcal{C}^t(\mathbf{x}_t^{tgt}, \mathbf{x}_t^{src})$ est la distorsion spectrale entre les trames source et cible, et $\mathcal{C}^c(\mathbf{x}_t^{tgt}, et\mathbf{x}_{t-1}^{tgt})$ est la distance euclidienne entre les trames cibles sélectionnées. La séquence $\mathbf{x}^{tgt}$ la plus vraisemblable est alors déterminée en utilisant un algorithme de Viterbi.

2.2 Conversion à partir d'unités phonétiques

L'algorithme de conversion présenté ci-dessus est tout d'abord étendu pour pouvoir exploiter des unités longues commes les phonèmes. L'avantage est d'une part de contraindre la conversion vocale à partir des informations linguistiques disponibles, et d'autre part de préserver la dynamique naturelle de la voix sur l'échelle des phonèmes. Pour ce faire, la sélection d'unités est reformulée en termes d'unités phonétiques :

$$\widehat{\mathbf{u}}^{tgt} = \underset{\mathbf{u}^{tgt}}{\operatorname{argmax}}\ p(\mathbf{u}^{tgt}|\mathbf{u}^{src}) \tag{3}$$

où $\mathbf{u} = [\mathbf{u}_1, \ldots, \mathbf{u}_N]$ est une partition de la séquence de trames $\mathbf{x}$ en N unités phonétiques, et $\mathbf{u}_n = [\mathbf{x}_{n_1}, \ldots, \mathbf{x}_{L_n}]$ est la séquence de trames de longueur L_n correspondant à l'unité n et au label phonétique l_n.
La sélection est réalisée sous la contrainte que la séquence des étiquettes des phonèmes cibles est la même que la séquence des étiquettes de phonèmes sources, de sorte que :

$$l(\mathbf{u}_n^{tgt}) = l(\mathbf{u}_n^{src}) = l_n, \qquad \forall n \in [1, N] \tag{4}$$

où $l(\mathbf{u}_n) = l_n$ est le label du n-ième phonème de la séquence. En d'autres termes, chaque unité cible $\mathbf{u}_n^{tgt}$ doit être sélectionnée parmi l'ensemble des unités cibles candidates correspondant au label phonétique l_n, et référencée à partir de maintenant par $\mathcal{U}_{l_n}^{tgt}$.
L'optimisation est similaire à celle décrite dans la section précédente, à l'exception que la fonction de coût est définie sur les N unités au lieu des T trames. La fonction de coût est alors définie par comparaison d'unités phonétiques en place de la comparaison de trames.

2.3 Fonction de coût spectral

La distorsion spectrale $\mathcal{C}_{spec}^t$ entre les unités source et cible $\mathbf{u}_n^{src}$ et $\mathbf{u}_j^{tgt} \in \mathcal{U}_{l_n}^{tgt}$ de longeurs différentes est calculée en utilisant la déformation temporelle dynamique (DTW) comme :

$$\mathcal{C}_{spec}^t \quad = \quad D(A(\mathbf{u}_j^{tgt}, \mathbf{u}_n^{src}), \mathbf{u}_n^{src}) \tag{5}$$

où $D(.,.)$ est la distance euclidienne et $A(\mathbf{u}_j^{tgt}, \mathbf{u}_n^{src})$ la séquence de l'unité cible déformée temporellement $\mathbf{u}_j^{tgt}$ et alignée avec la séquence de l'unité source $\mathbf{u}_n^{src}$.

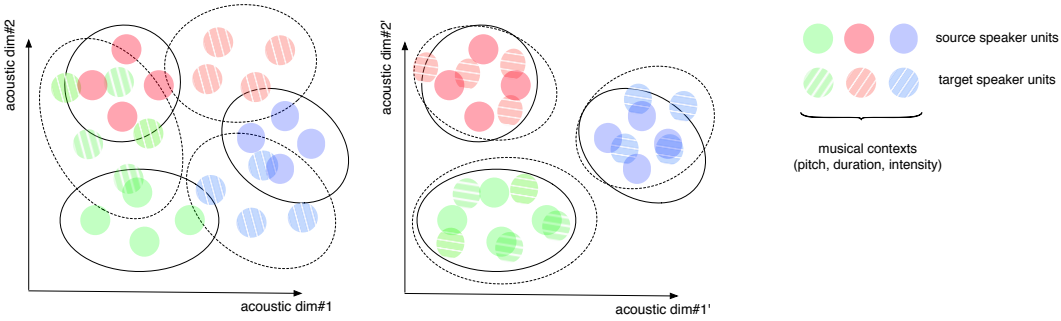

FIGURE 1 – Illustration de la fonction de côut multi-cible pour aligner les unités des voix source et cible dans l'espace musical. Chaque point représente une unité phonétique et sa séquence d'enveloppe spectrale correspondante.

Bien que la distorsion spectrale se soit montrée efficace en première approximation pour la conversion de la voix (Sündermann *et al.*, 2006), elle reste extrêmement limitée, en particulier en termes de similarité entre le locuteur source converti et le locuteur cible (Sündermann *et al.*, 2007).

En effet, la distorsion spectrale ne mesure que la similarité absolue entre les unités source et cible. Or, pour la conversion d'identité vocale, le coût entre une unité source et une unité cible doit être définie de manière à mesurer la similarité des unités *relativement* à chaque locuteur : la sélection doit être opérée de sorte à ce que l'enveloppe cible sélectionnée doit représenter pour la voix cible ce que l'enveloppe source représente pour la voix source. Pour ce faire, il est donc nécessaire d'aligner au préalable les distributions des unités source et cible pour pouvoir mesurer leur similarité de manière cohérente et relative à chaque chanteur. Ceci peut être obtenu en transformant l'espace acoustique des voix source et cible : soit en utilisant un pré-traitement de normalisation, soit en définissant une fonction de coût cible qui tient compte des multiples facteurs affectant la variabilité acoustique des chanteurs. Nous présentons ci-dessous des solutions pour ces deux types de transformation.
Une première source de différence entre l'espace acoustique des chanteurs source et cible réside dans la différence intrinsèque entre les chanteurs et entre les conditions d'enregistrement. Pour supprimer ces différences, un filtre moyen F_l est créé pour chaque phonème l, défini comme le rapport entre les moyennes des enveloppes source et cible observées pour ce phonème. La version normalisée de la distorsion spectrale s'écrit alors :

$$\mathcal{C}^t_{spec} \quad = \quad D(A(\mathbf{u}^{tgt}_j, F\mathbf{u}^{src}_n), F\mathbf{u}^{src}_n) \tag{6}$$

Afin d'augmenter la pertinence perceptive de la mesure de distorsion, la représentation de l'enveloppe spectrale utilisée pour l'algorithme DTW et pour le calcul de la distorsion spectrale utilise une échelle de fréquence Mel.

Le coût de concaténation correspondant $\mathcal{C}^c_{spec}$ entre les unités déformées temporellement du chanteur cible est alors défini comme suit :

$$\mathcal{C}^c_{spec} \quad = \quad D(\mathbf{u}^{tgt}_{i,r}, \mathbf{u}^{tgt}_{j,l}) \tag{7}$$

où $\mathbf{u}^{tgt}_{i,r}$ et $\mathbf{u}^{tgt}_{j,l}$ représentent les trames droite (de fin) de l'unité $\mathbf{u}^{tgt}_i$ et gauche (de début) de l'unité $\mathbf{u}^{tgt}_j$.

2.4 Fonction de coût multi-cible

Une deuxième source de différence réside dans le fait que l'enveloppe spectrale d'un locuteur varie en fonction de multiples facteurs, tels que la hauteur, l'intensité et la durée (voir par exemple (Joliveau *et al.*, 2005) sur la voix chantée). Pour compenser ces différences, la comparaison des unités spectrales d'un chanteur source et d'un chanteur cible doit être mesurée de manière relative au contexte musical dans lequel elles sont observées. Pour ce faire, une fonction de coût multi-cibles est proposée afin de prendre en compte ces facteurs, avec comme motivation principale de guider la sélection des unités spectrales vers des unités provenant d'un contexte musical similaire. En conséquence, la fonction de coût multi-cible proposée est écrite de la manière suivante :

$$\mathcal{C}^t = \sum_{factor} w_{factor} \mathcal{C}^t_{factor} \tag{8}$$

où $\mathcal{C}^t_{factor}$ représentent les fonctions de coûts partiels mesurant la similarité entre la source et la cible en fonction de l'un des facteurs, et w_{factor} les poids correspondant attribués aux facteurs. Comme mentionné dans (Taylor, 2006), la définition d'une fonction de coût cible multiple peut être interprétée comme une projection des unités d'origine dans un espace où chaque facteur est représenté par sa propre dimension dont la métrique est définie par le coût partiel $\mathcal{C}^t_{factor}$, avec le facteur de mise à l'échelle w_{factor}. Dans cet article, les facteurs considérés sont les facteurs spectraux, la hauteur, l'intensité et la durée. La fonction de coût multiple peut alors être interprétée comme une projection des unités spectrales source et cible afin de les aligner *relativement* aux facteurs musicaux, comme le montre la Figure 1. Concrètement, l'effet de cette projection est de rapprocher les enveloppes spectrales des chanteurs source et cible issues d'un contexte musical similaire, définies par la hauteur, l'intensité et la durée des unités correspondantes.
La fonction des coûts partiels est définie en utilisant la distance euclidienne entre les valeurs moyennes entre les caractéristiques normalisées mesurées sur les unités source et cible. Par exemple :

$$\mathcal{C}^t_{F0} = D(\overline{\mathbf{u}^{src}_i}(\mathbf{F0}^{src}_{norm}), \overline{\mathbf{u}^{tgt}_j}(\mathbf{F0}^{tgt}_{norm})) \tag{9}$$

avec $\mathbf{F0}_{norm}$ la séquence de F0 normalisée du chanteur, et $\overline{\mathbf{u}_i}(\mathbf{F0}_{norm})$ la F0 normalisée moyenne de la i-ème unité. Les autres fonctions de coût partiel $\mathcal{C}^t_{int}$ et $\mathcal{C}^t_{dur}$ sont définies de la même manière. La conversion vocale multi-cible a été utilisée avec les poids suivants : $w_{spec} = 15$, $w_{F0} = 10$ et $w_{dur} = 5$, qui ont été choisi empiriquement à partir d'essais informels. Le facteur d'intensité n'a pas été utilisé puisque la base de données utilisées pour la synthèse de chant avait une dynamique presque constante sur chaque unité de diphone (Ardaillon, 2017).

3 Expérience

3.1 Matériel

L'algorithme de sélection d'unités multi-cibles proposé a été évalué dans une expérience de perception sur la conversion de la voix chantée. Le chanteur cible est un chanteur français qui a enregistré huit chansons du chanteur français Jacques Brel. Les enregistrements ont été réalisés dans des conditions professionnelles (studio d'enregistrement, ingénieur du son), et numérisés avec une fréquence d'échantillonnage de 48.000 Hz et avec un encodage de 16 bits par échantillon. La voix chantée de la source a été créée en utilisant un synthétiseur de voix chantée (Ardaillon, 2017) à partir

des partitions musicales des chansons. Le but de l'utilisation d'un synthétiseur vocal est d'évaluer la conversion de la voix dans des conditions contrôlées, en garantissant le respect de la partition musicale et la connaissance précise des labels et des limites des phonèmes. En outre, l'utilisation d'une voix de synthèse constitue une preuve de concept que l'identité d'une voix de synthèse peut être contrôlée efficacement par conversion de l'identité vocale (Villavicencio & Bonada, 2010).

3.2 Banc d'essai des algorithmes de conversion

Un banc d'essai de systèmes de conversion à partir de sélection d'unités a été élaboré pour comparaison : d'une part, une fonction de coût basée sur la distorsion spectrale seule, et d'autre part la fonction de coût multi-cibles proposée. L'un des avantages de l'algorithme de conversion vocale proposé réside dans la possibilité de restreindre la conversion à un sous-ensemble de phonèmes et de garder le reste des phonèmes inchangés. En conséquence, cet article évalue également la conversion vocale chantée obtenue en convertissant tous les phonèmes, ou en convertissant seulement les voyelles du chanteur source. Ceci est basé sur l'hypothèse que l'identité est majoritairement portée par les voyelles, qui sont en outre plus stables et beaucoup plus longues que les consonnes dans la voix chantée. Pour résumer, les échantillons utilisés pour l'expérience sont : chant synthétisé source (S), chanteur cible (T), conversion de tous les phonèmes par distorsion spectrale (SD), conversion de tous les phonèmes avec la fonction de coût multi-cible (MO), et la même conversion vocale en convertissant seulement les voyelles (respectivement, SD_VOW et MO_VOW).

3.3 Configurations expérimentales

L'expérience a consisté dans le jugement par des auditeurs d'échantillons vocaux chantés, basés sur la similarité avec le chanteur cible et le caractère naturel du chanteur, comme utilisé pour la compétition de conversion d'identité vocale de 2016 (Toda *et al.*, 2016). Pour ce faire, quatre chansons ont été sélectionnées pour l'expérience parmi les huit disponibles, et les deux premières phrases de ces quatre chansons ont été utilisées pour la conversion. Pour une chanson donnée, la conversion de la voix a été effectuée en utilisant les morceaux disponibles restants, soit les sept chansons restantes. Pendant l'expérience, le participant devait choisir entre l'une des quatre chansons à évaluer. Ensuite, les échantillons vocaux chantés (originaux, synthétisés et convertis) étaient présentés dans un ordre aléatoire, avec toujours la possibilité d'écouter le vrai chanteur cible, et le participant devait évaluer le naturel de l'échantillon et la similarité avec le chanteur cible. L'expérience a été menée en ligne, encourageant l'utilisation d'écouteurs et de casque audio dans un environnement silencieux. Vingt personnes ont participé à l'expérience. Chaque personne a évalué les conversions d'une seule chanson, soit 2 phrases fois 6 versions (incluant le chant synthétisé S et la voix cible T).

3.4 Résultats et Discussion

La Figure 2 présente les scores obtenus par la voix chantée synthétisée, le chanteur cible et les algorithmes de conversion vocale. La voix chantée synthétisée a un naturel acceptable (3,3) mais la plus faible similarité avec la voix cible (2,1). Tous les algorithmes de conversion ont une similarité significativement plus élevée, ce qui vient malheureusement avec une dégradation du signal de chant converti. Par comparaison des algorithmes VC : l'algorithme multi-cible proposé a une similarité significativement plus élevée avec le chanteur cible (3,2) que la conversion basée sur la distorsion spectrale (2,7), et avec un naturel comparable (respectivement, 2,0 et 1,9). Cela montre que la prise en compte du contexte musical dans la recherche d'enveloppes spectrales améliore la

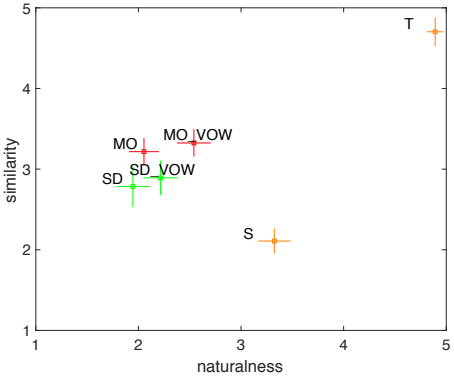

FIGURE 2 – Scores de conversion : score moyen et intervalles de confiance à 95%.

sélection d'enveloppes spectrales adéquates, conduisant à un gain de naturel et de similarité. Enfin, la préservation des consonnes conduit à un naturel significativement plus élevé (respectivement, 2,5 et 2,2) qui vient également avec une augmentation de la similarité (respectivement, 3,3 et 2,9). Bien que le gain de naturel soit clairement attendu, le gain de similarité suggère que tous les phonèmes n'ont pas la même importance dans la conversion de la voix. En particulier, les voyelles peuvent transmettre plus d'informations sur l'identité du chanteur que les consonnes. D'autre part, ce résultat suggère que les deux dimensions de jugement de la conversion ne sont clairement pas orthogonales : la dégradation de la conversion affecte directement le jugement de la similarité à la voix cible.

4 Conclusion

Cet article a présenté un algorithme de sélection d'unités multi-cibles pour la conversion non-parallèle de la voix chantée. L'idée principale est que la sélection des enveloppes spectrales doit être faite pour que les enveloppes spectrales sélectionnées du chanteur cible soient non seulement similaires mais aussi issues d'un même contexte linguistique (phonèmes) et musical (hauteur, intensité, durée) que celles du chanteur source. Une expérience perceptive a été menée pour convertir un synthétiseur vocal en un célèbre chanteur français. La conversion vocale multi-cible proposée a été jugée sensiblement plus similaire au chanteur cible par rapport à un algorithme classique de sélection d'unités. Les recherches futures porteront sur l'apprentissage des enveloppes spectrales en fonction des facteurs musicaux et leur exploitation en conversion de la voix chantée par sélection d'unités.

Références

AIHARA R., NAKASHIKA T., TAKIGUCHI T. & ARIKI Y. (2014). Voice conversion based on non-negative matrix factorization using phoneme-categorized dictionary. In *International Conference on Audio, Speech, and Signal Processing (ICASSP)*, p. 7944–7948.

ARDAILLON L. (2017). *Synthesis and expressive transformation of singing voice*. PhD thesis, Ircam-Upmc, Paris, France.

DESAI S., RAGHAVENDRA E. V., YEGNANARAYANA B., BLACK A. W. & PRAHALLAD K. (2009). Voice conversion using artificial neural networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

DOI H., TODA T., NAKANO T., GOTO M. & NAKAMURA S. (2012). Singing voice conversion method based on many-to-many eigenvoice conversion and training data generation using a singing-to-singing synthesis system. In *Asia-Pacific Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC)*.

JIN Z., FINKELSTEIN A., DIVERDI S., LU J. & MYSORE G. J. (2016). CUTE : A concatenative method for voice conversion using exemplar-based unit selection. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

JOLIVEAU E., SMITH J. & WOLFE J. (2005). Vocal tract resonances in singing : The soprano voice. *Journal of the Acoustical Society of America*, **116**(4), 2434–2439.

KENMOCHI H. (2010). VOCALOID and Hatsune Miku phenomenon in Japan. In *Intersinging, Interdisciplinary Workshop on Singing Voice*.

KINNUNEN T., JUVELA L., ALKU P. & YAMAGISHI J. (2017). Non-parallel voice conversion using i-vector plda : towards unifying speaker verification and transformation. In *International Conference on Audio, Speech, and Signal Processing (ICASSP)*.

KOBAYASHI K. & TODA T. (2014). Statistical singing voice conversion with direct waveform modification based on the spectrum differential. In *Interspeech*, p. 2514–2518.

LORENZO-TRUEBA J., YAMAGISHI J., TODA T., SAITO D., VILLAVICENCIO F., KINNUNEN T. & LING Z. (2018). The voice conversion challenge 2018 : Promoting development of parallel and nonparallel methods. In *Speaker Odyssey*.

NAKASHIKA T., TAKIGUCHI T. & MINAMI Y. (2016). Non-parallel training in voice conversion using an adaptive restricted boltzmann machine. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **24**(11), 2032 – 2045.

STYLIANOU Y., CAPPÉ O. & MOULINES E. (1998). Continuous probabilistic transform for voice conversion. *IEEE Transactions on Speech and Audio Processing*, **6**(2), 131–142.

SUN L., KANG S., LI K. & MENG H. (2015). Voice conversion using deep bidirectional long short-term memory based recurrent neural networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

SÜNDERMANN D., HÖGE H., BONAFONTE A., NEY H., BLACK A. & NARAYANAN S. (2006). Text-independent voice conversion based on unit selection. In *International Conference on Audio, Speech, and Signal Processing (ICASSP)*, p. 1173–1176.

SÜNDERMANN D., SMREKAR J., HÖGE H., BONAFONTE A. & NEY H. (2007). The speech alignment paradox. In *International Workshop on Advances in Speech Technology (AST)*.

TAYLOR P. (2006). The target cost formulation in unit selection speech synthesis. In *Interspeech*.

TODA T., BLACK A. W. & TOKUDA K. (2007). Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory. *IEEE Transactions on Audio, Speech, and Language Processing*, **15**(8), 2222–2235.

TODA T., CHEN L.-H., SAITO D., VILLAVICENCIO F., WESTER M., WU Z. & YAMAGISHI J. (2016). The voice conversion challenge 2016. In *Interspeech*.

VILLAVICENCIO F. & BONADA J. (2010). Applying voice conversion to concatenative singing-voice synthesis. In *Interspeech*, p. 803–806.

VILLAVICENCIO F. & KENMOCHI H. (2011). Non-parallel singing-voice conversion by phoneme-based mapping and covariance approximation. In *DAFx*, p. 241–244.

WU Z., VIRTANEN T., KINNUNEN T., CHNG E. S. & LI H. (2013). Exemplar-based unit selection for voice conversion utilizing temporal information. In *Interspeech*.