HAL open science

Mesure de similarité fondée sur des réseaux de neurones siamois pour le doublage de voix

Adrien Gresse, Richard Dufour, Vincent Labatut, Mickael Rouvier, Jean-François Bonastre

To cite this version:

Adrien Gresse, Richard Dufour, Vincent Labatut, Mickael Rouvier, Jean-François Bonastre. Mesure de similarité fondée sur des réseaux de neurones siamois pour le doublage de voix. XXXIIèmes Journées d'Études sur la Parole (JEP), Jun 2018, Aix-en-Provence, France. 10.21437/JEP.2018-2 . hal-01819198

HAL Id: hal-01819198

https://hal.science/hal-01819198

Submitted on 20 Jun 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Mesure de similarité fondée sur des réseaux de neurones siamois pour le doublage de voix

Adrien Gresse Richard Dufour Vincent Labatut
Mickaël Rouvier Jean-François Bonastre
LIA - Université d'Avignon (France)
prénom.nom@univ-avignon.fr

RÉSUMÉ

Le doublage vocal d'une œuvre culturelle permet sa diffusion vers une audience plus large. Le processus de sélection de voix dans une nouvelle langue, intégralement réalisé par un opérateur humain, est appelé casting vocal. Cette sélection dépasse le simple cadre de la proximité acoustique entre deux voix, intégrant de nombreux critères plus subjectifs qui peuvent être liés notamment à des choix socioculturels, émotionnels... Dans ce papier, nous proposons une approche par réseaux de neurones siamois mesurant la proximité entre la voix originale et la voix dans la langue cible, en intégrant la notion de similarité entre les voix non pas d'un point de vue purement acoustique mais également réceptif. Les premiers résultats que nous obtenons montrent grâce à un test d'hypothèse statistique qu'il existe une information dans les paramètres acoustiques qui permet cette association.

ABSTRACT

Siamese neural networks based similarity metric for dubbing

Dubbing aims to broadcast a multimedia document to a larger audience. The process that consists in selecting a voice in a target language is referred as voice casting and it is performed by a human. This selection is not only based on acoustic similarity between two voices. Actually, it is suppported by more subjective criteria such as emotions, sociocultural choices... In this paper we propose a siamese neural networks based approach measuring proximity between the original voice and the dubbed one. The concept of similarity we want to model does not only consider the acoustic part of a voice, also it takes into account spectators receptive concerns. We perform a statistical test to evaluate our model. Our results show that there is an information in the acoustic parameters that allows a voice to be associated with another one with respect to a particular character.

MOTS-CLÉS : casting vocal, réseaux de neurones siamois, i-vecteur, similarité.

KEYWORDS: voice casting, siamese neural networks, i-vector, similarity.

1 Introduction

La voix apparaît, dans de nombreuses œuvres culturelles (films, documentaires, jeux vidéos...), comme un vecteur de stimulis émotifs pour le public qui la reçoit. Dans un contexte de diffusion internationale, un doublage vocal est souvent réalisé en remplaçant la voix originale par une nouvelle voix dans une langue cible. Le processus qui permet de sélectionner, à partir d'une voix originale, une voix parmi plusieurs voix candidates dans une autre langue est appelé *casting vocal*. Il s'agit de l'objet d'étude principal de nos travaux. À l'origine, la sélection est réalisée par un opérateur humain en fonction, d'une part, de la voix originale et de critères plus subjectifs pouvant être liés, par exemple,

au personnage, ou rôle, interprété. En sciences humaines, le terme *réception* est utilisé pour parler des effets à long terme d'une voix perçue à un moment donné. Il ne s'agit pas ici de trouver la voix la plus semblable à celle d'origine au niveau acoustique, mais de trouver celle qui aura, dans cette nouvelle langue, un effet identique à la voix d'origine, faisant intervenir des critères socioculturels, ou autres.

Un des enjeux du travail que nous menons réside dans la notion de "similarité" entre les voix. D'une manière générale, celle-ci a déjà été étudiée à maintes reprises. Nombreux sont les papiers qui ont découlé du travail de Laver (1980), qui propose un moyen de décrire la *qualité vocale*, se comprenant comme les caractéristiques auditives qui colorent la voix d'un individu. Plusieurs travaux proposent d'évaluer le degré de similarité de la voix perçue dans un groupe de voix (McDougall, 2013; Rose, 1999; Loakes, 2006; Nolan *et al.*, 2011; Baumann & Belin, 2010), bien souvent dans le cadre d'applications juridiques. Entre autres, ces travaux montrent qu'il existe des corrélations entre certaines caractéristiques acoustiques et le fait que des voix soient perçues comme étant similaires ou non. Toutefois, il n'existe pas de méthode établie pour quantifier le degré de similarité entre deux voix. Dans le vaste domaine de la reconnaissance automatique du locuteur, des systèmes permettent indirectement de mesurer la distance entre deux identités locuteurs, et par conséquent d'évaluer la "similarité" de leur voix (Kelly *et al.*, 2016; Zhang & Tan, 2008; Lindh & Eriksson, 2010). Néanmoins, cette notion de similarité n'induit principalement qu'une ressemblance acoustique. Dans notre contexte, nous souhaitons étendre cette notion de similarité en y intégrant d'autres critères de nature plus humaine, qui guident le choix de l'opérateur de casting vocal. Nous pensons aux éléments amenés notamment par le jeu de l'acteur tels que les inflexions de voix utilisées pour mieux faire ressortir les traits du personnage incarné. Récemment, certains travaux ont commencé à explorer cette dimension (Obin *et al.*, 2014; Obin & Roebel, 2016; Gresse *et al.*, 2017) qui offrent une comparaison entre l'utilisation d'un système de reconnaissance automatique du locuteur et d'un classifieur multimodal de critères para-linguistiques.

Nos travaux s'inscrivent dans le cadre de la recommandation automatique de voix pour des œuvres culturelles. Dans cette optique, nous proposons d'explorer la dimension caractéristique du personnage perçue au travers de la voix dans le cadre d'un jeu vidéo. Il s'agit d'une approche inédite pour la mesure de la similarité de la voix dans un contexte multilingue. Nous nous limitons pour le cas présent à deux langues (anglais et français). À la différence de (Gresse *et al.*, 2017), où nous avions proposé une approche i-vecteur/PLDA inspirée de la reconnaissance du locuteur, nous explorons ici l'utilisation de réseaux de neurones siamois. En effet, nous avons observé que la PLDA a tendance à se focaliser sur l'identité du locuteur. De plus, notre méthode pouvait s'apparenter à un mapping des locuteurs dans la langue originale vers les locuteurs de la langue source, plus qu'à une estimation pure de la similarité entre les deux. Notre intuition est qu'un réseaux de neurones siamois devrait être mieux adapté à la notion de similarité telle que définie dans le casting vocal. Nous nous concentrons ici sur la mesure de la similarité plus que sur la classification des voix. Nous utilisons un test statistique pour vérifier si le modèle est capable d'abstraire cette notion de similarité.

Le reste de cet article est organisé comme suit. Dans la Section 2 nous détaillons notre approche. La méthode et le protocole expérimental sont décrits dans la Section 3 avant de présenter nos résultats dans la Section 4. Enfin, nos conclusions et nos perspectives futures sont énoncées dans la Section 5.

2 Approche proposée

La Figure 1 illustre de manière simplifiée notre système automatique de casting vocal. Ce dernier accepte deux entrées qui correspond à deux extraits de voix, et une sortie qui représente le score de similarité entre celles-ci, estimant leur degré de proximité. Pour le casting vocal, le système doit

prédire dans quelle mesure la voix d'une langue *cible* peut être utilisée pour le doublage de la voix dans une langue *source*, en dépassant la simple ressemblance acoustique. Au cœur du système se trouve notre modèle de similarité appris sur un ensemble de voix.

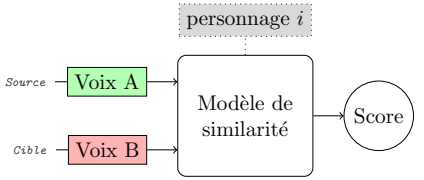

FIGURE 1 – Présentation simplifiée du système automatique de casting vocal.

Dans la Section 2.1, nous présentons les concepts et notre motivation quant à l'utilisation des réseaux de neurones siamois, constituant l'originalité de ce travail. Les données en entrée de ce réseau sont représentées au moyen de i-vecteurs, que nous présentons succinctement dans la Section 2.2.

2.1 Réseaux de neurones siamois

De manière intuitive, les architectures siamoises nous offrent un moyen d'apprendre une mesure de similarité à partir de deux entrées indépendantes qui partagent une relation abstraite de similarité. Les premiers travaux faisant utilisation de réseaux de neurones siamois font référence à (Bromley *et al.*, 1994) pour la vérification de signatures. Ce type d'architecture a la particularité de faire intervenir deux réseaux de neurones identiques qui prennent deux entrées indépendantes et qui se rejoignent finalement grâce à une fonction de pénalité (voir Figure 2). Cette fonction se base sur une métrique (ici une distance) calculée à partir des représentations de plus haut-niveau des deux réseaux. À noter que les deux réseaux qui interviennent dans ce type d'architecture partagent les mêmes paramètres.

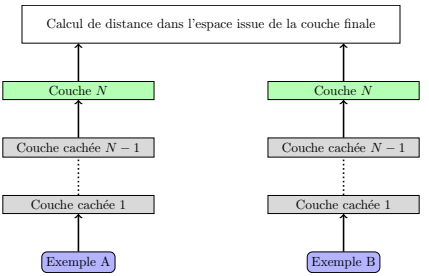

FIGURE 2 – Exemple de réseau de neurones siamois.

Dans notre travail, nous avons mis en place une architecture siamoise semblable à (Chopra *et al.*, 2005; Hadsell *et al.*, 2006). Elle nous assure deux choses (Koch *et al.*, 2015) :

— Du fait du partage des paramètres, des entrées fortement similaires ne peuvent pas être projetées à des endroits différents dans l'espace de représentation latent et inversement, une paire d'entrées différentes ne peut être projetée par les réseaux siamois à des endroits proches.

— Aucune distinction n'est faite par la fonction de pénalité quant à l'ordre des entrées constituant la paire traitée (*i.e.* la fonction de similarité est symétrique).

Dans (Chopra *et al.*, 2005) les auteurs utilisent une fonction de pénalité s'appuyant sur une mesure de l'énergie définie comme $E_W(I_1, I_2) = \|G_W(I_1) - G_W(I_2)\|_2$, où I_1 et I_2 correspondent aux entrées et G représente une fonction de projection depuis l'espace des entrées vers un nouvel espace de représentation. En jouant sur les paramètres W, il faut donc minimiser l'énergie lorsque les deux entrées I_1 et I_2 sont similaires mais aussi s'assurer que E_W est grande pour des entrées différentes. À juste titre, la fonction de pénalité est qualifiée de contrastive. Soit une variable binaire notée T telle que $T = 0$ lorsque les entrées sont similaires et $T = 1$ dans le cas contraire. On considère une constante notée m positive que l'on peut interpréter comme une marge. La fonction de pénalité est définie par l'équation suivante :

$$L(I_1, I_2, T) = (1 - T) \times (E_W(I_1, I_2))^2 + T \times max\{0, m - E_W(I_1, I_2)\}^2$$

Dans les travaux présentés, nous avons utilisé deux réseaux Perceptron multicouches (MLP) contenant 2 couches cachées de 1 000 unités plus une couche finale constituées de 500 unités, combinées à une fonction tangente hyperbolique. Les deux réseaux calculent la même fonction G_W.

2.2 Représentation des données par des i-vecteurs

La question du choix de la représentation des segments audio pour une mesure de similarité de voix de doublage apparaît comme une problématique en elle-même. De manière générale, le choix de la représentation des données en entrée (ici des segments audio) a une influence non-négligeable sur les performances finales des systèmes : c'est aussi le cas dans notre contexte de casting vocal. Ainsi, la contrainte principale concerne la variabilité de la durée des séquences audio. Afin de pouvoir représenter des séquences de durée variables par un vecteur de taille fixe, nous avons choisi de représenter ces données au moyen de i-vecteurs.

Les i-vecteurs ont été initialement présentés dans le domaine de la vérification du locuteur (Dehak *et al.*, 2011) et ont depuis montré leur robustesse. Ils contiennent, entre autres, les caractéristiques propres au locuteur mais également des informations liées au canal de transmission ou au contenu phonétique du segment audio. Cette représentation est extraite à partir de séquences pouvant être de tailles différentes. Plus généralement on dit que le i-vecteur est une représentation compacte de la séquence de paramètres acoustiques extraite à partir d'un segment de voix.

3 Protocole Expérimental

Dans cette section, nous détaillons les données utilisées pour notre problème de casting vocal (Section 3.1). Puis nous définissons la manière dont nos expériences seront menées (Section 3.3 et Section 3.2). Enfin, nous proposons un protocole d'évaluation de notre approche (Section 3.4).

3.1 Données

Nos expériences sont réalisées sur les données issues du jeu vidéo *Mass Effect 3*. Nous avons extrait les interactions vocales des différents personnages du jeu dans leurs versions originales (*i.e.* en anglais) et doublées (*i.e.* en français). Notre objectif est d'apprendre à réaliser – à l'instar de l'opérateur de casting vocal – l'appariement des voix originales et doublées de manière automatique. L'apprentissage automatique se base sur des paires de voix dont l'une d'entre elles appartient à l'ensemble des segments de voix originales (ici anglais) et la deuxième appartient quant à elle à l'ensemble des segments de voix doublées (ici français). Nous avons donc un ensemble contenant toutes les paires réalisables $\mathcal{L} = \{(x_i, y_j)\}$ avec $x_i \in X$ l'ensemble des segments originaux en

anglais et $y_j \in Y$ l'ensemble des segments en français. Nous notons au passage que les ensembles X et Y sont bijectifs du fait que chaque segment de la version originale possède exactement un équivalent dans la version française.

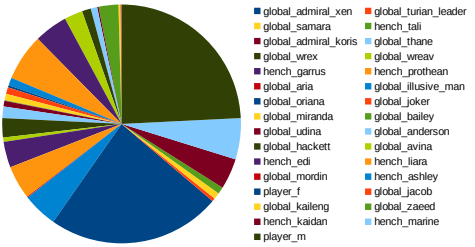

FIGURE 3 – Répartition des segments par nom de personnage.

Nous pouvons définir un personnage k comme un couple de locuteurs $\mathcal{P}_k = (S^k_{en}, S^k_{fr})$ où S^k_{en} correspond à l'ensemble des segments du locuteur correspondant au personnage k en anglais (en) et S^k_{fr} représentant les segments du locuteur doublant le même personnage k en français (fr). Il est important de noter que nous avons veillé à ce qu'aucun locuteur ne soit associé à plus d'un personnage, dans le but d'éviter tout biais à ce niveau-là. Nous soulignons également que le nombre total de segments de voix n'est pas équitablement réparti entre les différents personnages comme on peut le voir dans la Figure 3. Le protocole expérimental a été défini afin de réduire au mieux ce biais.

Chaque ensemble de segments (*en* et *fr*) compte 10 000 segments audio de haute qualité (enregistrés en studio) pour un corpus contenant un total approximatif de 7,5 heures de dialogues dans chaque langue. Les segments ont en moyenne une durée de 3,5 secondes, et comme il a été dit plus haut, le nombre de segments par personnage est relatif à son importance dans le jeu. Ainsi, nous avons en moyenne, pour les deux langues, 12 minutes de dialogue par personnage, avec un écart-type de 20 minutes. Autre élément important : nous avons un total de 31 personnages différents associés chacun à un acteur différent (tous personnages confondus et toutes langues confondues) pour le doublage, soit 62 identités locuteurs représentées dans nos données.

3.2 Traitement des séquences

Le signal audio a été transformé en vecteurs de caractéristiques de 60 dimensions, avec 20 paramètres MFCC incluant l'énergie auxquels s'ajoutent les 20 dérivées du premier ordre (Δ) et 20 du second ($\Delta\Delta$). Les paramètres sont calculés sur des fénêtres de 20 ms avec un décalage de 10 ms. Nous avons appliqué une normalisation sur la moyenne cepstrale et supprimé les trames de faible énergie, qui correspondent principalement à du silence. Nous avons ensuite entraîné un modèle du monde (UBM) de 2 048 composantes à partir des vecteurs de caractéristique et une matrice de variabilité totale T de rang 400 qui nous permet d'extraire nos i-vecteurs. Nous avons créé deux espaces i-vecteurs, un pour l'anglais et l'autre pour le français. Le modèle du monde ainsi que la matrice T ont été appris pour l'anglais sur NIST SRE 2004, 2005 et 2006. Pour le français, l'UBM et la matrice T ont été appris sur les campagnes d'évaluation ESTER-1, ESTER-2, EPAC, ETAPE et REPERE.

3.3 Apprentissage

Les paires constituées de segments de voix doublant un même personnage sont dites *target*, toutes les autres *nontarget*. Étant donné que notre ensemble de paires est issu de la combinaison des deux ensembles de segments de voix, nous avons un nombre beaucoup plus grand de paires *nontarget* que de *target*. Pour palier ce biais, nous avons réalisé un équilibrage des paires utilisées pour les tests. En effet, il nous faut avoir un parfait équilibre entre ces paires pour le corpus de test qui nous sert de corpus de contrôle. En détails, 16 personnages disposent d'un nombre de segments supérieur ou égal à 95. Nous pouvons donc utiliser 4 de ces personnages pour notre corpus de test, ceux-ci étant par conséquent retirés du corpus d'apprentissage afin d'éviter tout effet de mémorisation. A travers une validation croisée, nous obtenons 4 ensembles de tests différents ainsi que 4 corpus d'apprentissages variant également dans la mesure où les personnages non-utilisés pour les tests y sont réintégrés.

Compte tenu de nos paires, composées d'un segment anglais (source) et d'un segment français (cible), et avec 4 personnages comptant 95 segments (tirés aléatoirement pour ceux en ayant un nombre plus grand) en anglais et 95 en français, nous avons un corpus de test composé de $4 \times 95 \times 95$ paires *target* et $12 \times 95 \times 95$ paires *nontarget*, que nous ramenons donc au même nombre par tirage aléatoire. Cette procédure permet d'éliminer les biais liés aux probabilités *a priori* des 2 classes ramenées à 0,5.

Nous avons également pensé à l'impact de la langue, ainsi qu'au contenu des segments de voix comme biais possibles. Nous utilisons toujours la même configuration de langue, soit un segment en anglais combiné avec un segment en français. Étant donné la sensibilité à la durée des i-vecteurs qui est directement reliée au contenu linguistique de chaque segment, nous veillons à éviter les paires qui associent un segment anglais à son homologue en français en mélangeant aléatoirement tous les segments au préalable. Enfin, nous avons également levé le biais potentiellement introduit par la différence de genre. Pour cela, nous effectuons nos tests sur les paires de même genre uniquement. Les paires *target* respectant de par nature cette dernière contrainte, nous réduisons donc l'ensemble des paires *nontarget* aux seules paires de même genre.

3.4 Évaluation

Notre évaluation se base sur les scores obtenus pour chaque paire. Il s'agit d'une distance de Manhattan calculée dans l'espace de représentation du modèle de similarité que nous avons appris. Les scores obtenus sur les paires issues du corpus de test sont regroupés en deux groupes. Les paires *target* dans l'un, les paires *nontarget* dans l'autre. Dans le but d'évaluer la pertinence du modèle de similarité appris, nous réalisons un test statistique d'hypothèse : le t-test, ou test de *Student*. L'idée est de comparer la moyenne des scores des deux groupes. Il s'agit d'un test bilatéral où l'hypothèse nulle H_0 dit que les moyennes des deux groupes sont identiques. Pour étayer cette statistique il est impératif d'y ajouter des éléments descriptifs tels que la dispersion ou l'écart à la moyenne.

4 Résultats

Nous présentons dans la Table 1 l'ensemble des résultats obtenus au moyen du test statistique de *Student* sur les résultats des réseaux de neurones siamois. Pour chacun des 4 ensembles de tests détaillés plus haut, nous donnons la valeur du t-test avec la probabilité qui lui est associée. Ainsi nous avons obtenu des valeurs de $9,16$ et $11,22$ associées à une probabilité < 0.01 pour les ensembles de tests notés 1 et 2. Pour les ensembles notés 3 et 4, les valeurs du t-test sont de $-27,44$ et $-34,37$ aussi associées à une p-value < 0.01. Étant donné que le t-score est un ratio de la différence entre les deux groupes et de la différence à l'intérieur des groupes, nous observons que la différence entre les groupes

de paires *target* et *nontarget* est plus grande pour les tests notés 3 et 4. Cela peut signifier plusieurs choses. Le plus probable étant que, pour les personnages qui composent les ensembles de tests 1 et 2, le modèle ne soit pas parvenu à généraliser. Il est aussi envisageable que les personnages de chaque test soient déjà similaires entre eux ou au contraire plus variés. Nous entendons des personnages similaires au sens d'un même type de personnage (*e.g.* jouant le rôle de soldat). Ces résultats montrent que le système fait des confusions entre certains personnages. En effet, les deux groupes de scores étant très similaires, il y a donc confusion dès que l'on se retrouve à l'intersection des distributions des scores *target* et *nontarget*. Si l'on se réfère aux moyennes, nous observons une différence notamment sur les tests 1 et 2. En effet le score moyen des paires *target* est inférieur à celui des paires *nontarget* contrairement aux tests 3 et 4. Nos scores correspondent à une distance et devraient logiquement être en moyenne plus faibles chez les individus *target*. Hors, nous observons le phénomène inverse dans les deux premiers tests, ce qui renforce un peu plus l'idée que la configuration des personnages pour ces tests là n'est pas adéquate. Nous avons également testé l'utilisation d'une distance euclidienne au lieu d'une distance de Manhattan. Les résultats obtenus ne variant pas significativement, aussi nous n'avons pas jugé utile de les faire apparaître ici.

#	**NOMBRE DE PAIRES**	**CORPUS**		
		TRAIN	**TEST**	
1	genre identique :	19420084	72200	
	genre différent :	8218842	0	
	target / nontarget (même nb.)	13819463	36100	
		t-score / p-value :	**9,16**	**4,96E-20**
			target	*nontarget*
		moyenne :	0,65	0,63
		écart-type :	0,32	0,32
2	genre identique :	13168499	72200	
	genre différent :	2535633	0	
	target / nontarget (même nb.)	7852066	36100	
		t-score / p-value :	**11,22**	**3,25E-29**
			target	*nontarget*
		moyenne :	0,56	0,54
		écart-type :	0,23	0,24
3	genre identique :	9896643	72200	
	genre différent :	3977211	0	
	target / nontarget (même nb.)	6936927	36100	
		t-score / p-value :	**-27,44**	**6,55E-165**
			target	*nontarget*
		moyenne :	0,54	0,6
		écart-type :	0,28	0,26
4	genre identique :	18275296	72200	
	genre différent :	8117988	0	
	target / nontarget (même nb.)	13196642	36100	
		t-score / p-value :	**-34,37**	**7,53E-257**
			target	*nontarget*
		moyenne :	0,59	0,67
		écart-type :	0,29	0,32
TOTAL CUMULÉ		**t-test / p-value :**	**-21,48**	**2,71E-102**
			target	***nontarget***
		moyenne :	**0,59**	**0,61**
		écart-type :	**0,28**	**0,29**

TABLE 1 – Résultats du test statistique de *Student* sur les corpus de tests.

Nous avons aussi effectué un test de *Student* sur toutes les paires de tests cumulées. Nous observons une différence significative compte tenu du t-score de $-21,48$ associé à une probabilité < 0.01. À titre de comparaison, nous avons réalisé le même test sur les scores obtenus avec la méthode présentée dans (Gresse *et al.*, 2017). Sur l'ensemble total de paires de tests, le t-score est de $2,70$ avec une probabilité également inférieure au seuil de rejet. En définitive, ces résultats nous amènent

à la conclusion que l'approche présentée dans cet article est bien adaptée à la modélisation de la similarité de voix de doublage.

5 Conclusion

Le problème du casting vocal, et plus particulièrement de la perception des voix, est un problème complexe du fait de la multitude de facteurs pouvant influencer le choix de l'opérateur. Cette complexité se ressent dès lors que l'on essaye d'automatiser une tâche qui n'est pas clairement formalisée et qui laisse place à des critères subjectifs. Dans cet article, nous avons proposé un système automatique pour essayer de nous rapprocher des choix de l'opérateur de casting. Ce système nous permet surtout d'explorer, au moyen de méthodes statistiques, cette notion de similarité. Pour cela, nous avons utilisé une approche fondée sur les réseaux de neurones siamois. Les résultats que nous avons obtenus montrent que le modèle de similarité que nous avons appris au moyen de ces réseaux est capable de faire ressortir une différence significative entre les paires de voix doublant un même personnage et les autres paires de voix. Les résultats de nos tests montrent bien que cette différence ne peut être le fruit du hasard, il y a donc une information reliée à la dimension "personnage" sur laquelle s'appuie notre modèle.

Nous avons toutefois observé des variations dans nos 4 tests au niveau des scores moyens des groupes *target* et *nontarget*. La performance du modèle de similarité dépend donc des personnages impliqués dans nos données. Il nous faut donc étudier plus en profondeur l'impact des différents personnages pour pouvoir être capable d'expliquer les possibles confusions du système. Par exemple en mettant en place un protocole sur plusieurs itérations où l'on retire un personnage différent du corpus à chaque fois. De manière générale, la mesure de la similarité réalisée à l'aide des réseaux de neurones siamois apparaît plus pertinente qu'une approche inspirée de la reconnaissance du locuteur s'appuyant sur la PLDA (Gresse *et al.*, 2017). En effet, la PLDA a tendance à se focaliser sur le locuteur lui-même. À l'inverse, les réseaux de neurones siamois nous permettent d'apprendre un modèle sur la base de deux voix différentes. Nous considérons donc l'utilisation des réseaux de neurones siamois et l'apprentissage par paires de voix plus apte à la mesure d'une similarité. Le modèle tire à la fois parti des paires de voix doublant un même personnage et de celles doublant des personnages différents. Bien que nous observons une différence significative entre les scores *target* et *nontarget*, il nous est encore difficile d'expliquer cette différence. Nous consacrerons donc nos futurs travaux à l'approfondissement du travail présenté dans cet article. De plus, nous avons conscience que la méthode i-vecteur peut ne pas être la plus adéquate. En effet cette dernière est censée compenser une partie du bruit, mais cela peut nous amener à perdre de l'information utile pour caractériser certains aspects de la voix. La question de la représentation de l'information est donc une piste de recherche très intéressante que nous aborderons dans de futurs travaux.

Remerciements

Les travaux présentés dans cet article sont financés par la fondation de l'Université d'Avignon.

Références

BAUMANN O. & BELIN P. (2010). Perceptual scaling of voice identity : common dimensions for different vowels and speakers. *Psychological Research PRPF*, **74**(1), 110.

BROMLEY J., GUYON I., LECUN Y., SÄCKINGER E. & SHAH R. (1994). Signature verification using a" siamese" time delay neural network. In *Advances in Neural Information Processing Systems*, p. 737–744.

CHOPRA S., HADSELL R. & LECUN Y. (2005). Learning a similarity metric discriminatively, with application to face verification. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, p. 539–546 : IEEE.

DEHAK N., KENNY P. J., DEHAK R., DUMOUCHEL P. & OUELLET P. (2011). Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, **19**(4), 788–798.

GRESSE A., ROUVIER M., DUFOUR R., LABATUT V. & BONASTRE J.-F. (2017). Acoustic pairing of original and dubbed voices in the context of video game localization.

HADSELL R., CHOPRA S. & LECUN Y. (2006). Dimensionality reduction by learning an invariant mapping. In *Computer vision and pattern recognition, 2006 IEEE computer society conference on*, volume 2, p. 1735–1742 : IEEE.

KELLY F., ALEXANDER A., FORTH O., KENT S., LINDH J. & ÅKESSON J. (2016). Identifying perceptually similar voices with a speaker recognition system using auto-phonetic features. In *INTERSPEECH*, p. 1567–1568.

KOCH G., ZEMEL R. & SALAKHUTDINOV R. (2015). Siamese neural networks for one-shot image recognition. In *ICML Deep Learning Workshop*, volume 2.

LAVER J. (1980). The phonetic description of voice quality. *Cambridge Studies in Linguistics London*, **31**, 1–186.

LINDH J. & ERIKSSON A. (2010). Voice similarity-a comparison between judgements by human listeners and automatic voice comparison. In *Proceedings from FONETIK*, p. 63–69.

LOAKES D. (2006). *A forensic phonetic investigation into the speech patterns of identical and non-identical twins*. PhD thesis, University of Melbourne, School of Languages.

MCDOUGALL K. (2013). Assessing perceived voice similarity using multidimensional scaling for the construction of voice parades. *International Journal of Speech, Language & the Law*, **20**(2).

NOLAN F., FRENCH P., MCDOUGALL K., STEVENS L. & HUDSON T. (2011). The role of voice quality 'settings' in perceived voice similarity. *International Association for Forensic Phonetics and Acoustics, Vienna, Austria*.

OBIN N. & ROEBEL A. (2016). Similarity search of acted voices for automatic voice casting. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **24**(9), 1642–1651.

OBIN N., ROEBEL A. & BACHMAN G. (2014). On automatic voice casting for expressive speech : Speaker recognition vs. speech classification. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, p. 950–954 : IEEE.

ROSE P. (1999). Differences and distinguishability in the acoustic characteristics of hello in voices of similar-sounding speakers. *Australian Review of Applied Linguistics*, **22**(1), 1–42.

ZHANG C. & TAN T. (2008). Voice disguise and automatic speaker recognition. *Forensic science international*, **175**(2), 118–122.