Apprentissage automatique de représentation de voix à l'aide d'une distillation de la connaissance pour le casting vocal

Adrien Gresse, Mathias Quillot, Richard Dufour, Jean-François Bonastre

▶ **To cite this version:**

Adrien Gresse, Mathias Quillot, Richard Dufour, Jean-François Bonastre. Apprentissage automatique de représentation de voix à l'aide d'une distillation de la connaissance pour le casting vocal. 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 1 : Journées d'Études sur la Parole, 2020, Nancy, France. pp.280-288. hal-02798550v3

HAL Id: hal-02798550
https://hal.science/hal-02798550v3

Submitted on 23 Jun 2020

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.

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Apprentissage automatique de représentation de voix à l'aide d'une distillation de la connaissance pour le casting vocal

Adrien Gresse Mathias Quillot Richard Dufour Jean-François Bonastre
LIA, Avignon Université, France
prénom.nom@univ-avignon.fr

RÉSUMÉ

La recherche d'acteurs vocaux pour les productions audiovisuelles est réalisée par des directeurs artistiques (DA). Les DA sont constamment à la recherche de nouveaux talents vocaux, mais ne peuvent effectuer des auditions à grande échelle. Les outils automatiques capables de suggérer des voix présentent alors un grand intérêt pour l'industrie audiovisuelle. Dans les travaux précédents, nous avons montré l'existence d'informations acoustiques permettant de reproduire des choix du DA. Dans cet article, nous proposons une approche à base de réseaux de neurones pour construire une représentation adaptée aux personnages/rôles visés, appelée p-vecteur. Nous proposons ensuite de tirer parti de données externes pour la représentation de voix, proches de celles d'origine, au moyen de méthodes de distillation de la connaissance. Les expériences menées sur des extraits de voix de jeux vidéo montrent une amélioration significative de l'approche p-vecteur, avec distillation de la connaissance, par rapport à une représentation x-vecteur, état-de-l'art en reconnaissance du locuteur.

ABSTRACT

Learning voice representation using knowledge distillation for automatic voice casting

The search for voice actors for audiovisual productions is carried out by artistic directors (DA). DA are constantly on the lookout for new vocal talent, but are unable to conduct large-scale auditions.Automatic tools able to suggest the most suited voices are of a great interest for audiovisual industry. In previous work, we have shown the existence of acoustic information allowing us to reproduce DA choices. In this article, we propose a neural network-based approach to construct a representation adapted to targeted characters/roles, called p-vector. We then propose to take advantage of external data, close the origin one, for the representation of voices, using knowledge distillation methods. Experiments carried out on voice extracts from video games show a significant improvement in the p-vector representation, including knowledge distillation, compared to x-vectors, state-of-the-art representation in speaker recognition.

MOTS-CLÉS : distillation de la connaissance, p-vecteur, similarité perceptive, réseaux de neurones profonds.

KEYWORDS: knowledge distillation, p-vector, perceptual similarity, deep neural network.

1 Introduction

Les entreprises visant la diffusion internationale d'oeuvres audiovisuelles (films, séries, jeux vidéo...) atteignent alors un public multilingue et multiculturel. Ainsi, les producteurs de ces créations audiovisuelles accordent de plus en plus d'attention aux voix qu'ils attribuent à un personnage ou à un rôle

Volume 1 : Journées d'Études sur la Parole, pages 280–288.
HAL : hal-02798550.
Cette œuvre est mise à disposition sous licence Attribution 4.0 International.

particulier afin de renforcer le sentiment d'immersion du public. Ce processus de changement de la voix originale dans une langue par une nouvelle voix dans un autre langage est appelé *doublage vocal*. Il consiste à remplacer l'intégralité des dialogues de la création originale par de nouveaux acteurs vocaux dans le contexte linguistique et culturel ciblé. Dans ce contexte, la sélection des voix appropriées dans une langue cible en fonction à la fois de la voix d'origine et du rôle, est une tâche cruciale, appelée *casting vocal*. Habituellement, un expert humain, appelé *directeur artistique* (DA), effectue la tâche de casting de voix dans des sociétés de doublage.

Le problème majeur du doublage vocal réside dans le fait que la "similitude" recherchée entre une voix originale et une voix doublée est loin d'être une simple ressemblance acoustique. Il comprend les caractéristiques socioculturelles des langues et des pays sources et cibles. De plus, il n'y a pas de vocabulaire bien établi pour décrire les voix, les personnages et les effets immersifs. Il y a deux limites à la façon dont les DA effectuent la tâche de casting vocal : les choix des DA intègrent un certaine subjectivité, liée à leurs propres caractéristiques socioculturelles, et 2), les DA ne peuvent pas écouter et mémoriser un nombre très élevé de voix. Par conséquent, un DA travaille généralement avec une liste réduite d'acteurs qu'il a écoutés et/ou avec lesquels il a déjà travaillé.

Les outils automatiques capables de mesurer l'adéquation potentielle entre une voix originale dans une langue source et une voix doublée dans une langue et un contexte cibles, présentent un grand intérêt pour l'industrie audiovisuelle. Ils aideront les DA à remédier aux problèmes susmentionnés et à ouvrir la porte à de nouveaux talents de voix, par exemple en pré-sélectionnant un nombre raisonnable de candidats au sein d'un très large ensemble de voix.

La similarité vocale dans le contexte du doublage de voix a été étudié dans (Obin *et al.*, 2014; Obin & Roebel, 2016). Les auteurs ont montré l'importance de certaines caractéristiques para-linguistiques (*e.g.* âge, genre, état du locuteur, qualité de la voix...). Dans (Gresse *et al.*, 2017), les auteurs proposent d'estimer la proximité de "doublage" entre deux voix (une dans la langue source, et une dans la langue cible) au moyen d'une approche i-vecteur/PLDA, inspirée du domaine de la reconnaissance du locuteur. (Gresse *et al.*, 2019) supposent que des informations, ou a minima des indices, liées au casting réalisé par les DA sont présentes dans les voix de doublage choisies. L'approche proposée permet de distinguer les paires *cible* (*i.e.* une voix dans une langue source associée à la voix du personnage correspondant dans la langue cible) de *non-cible* (*i.e.* voix qui ne correspond pas au bon personnage). Une limite de ce travail est que l'utilisation de l'apprentissage supervisé binaire donne de faibles capacités de généralisation au modèle, étant donné que l'interpolation ne peut s'appuyer que sur des contre-exemples.

Des travaux récents en reconnaissance du locuteur (Variani *et al.*, 2014; Snyder *et al.*, 2016, 2017, 2018) ont montré que des représentations au moyen de réseaux de neurones profonds, et d'apprentissage bout-en-bout (end-to-end), surpassent la représentation de référence i-vecteur. Dans cet article, nous proposons d'apprendre une représentation latente originale du personnage/rôle, appelée p-vecteur, à partir d'une approche fondée sur les réseaux de neurones. Ces p-vecteurs sont conçus pour aider le système à avoir une meilleure assimilation de la dimension du personnage, et par conséquent à mieux gérer les voix inconnues. Cette approche constitue la première contribution de cet article.

Néanmoins, un frein à l'utilisation d'une telle approche s'appuyant sur les réseaux de neurones est la nécessité d'une grande quantité de données dans le domaine considéré. Dans notre contexte, la seule information que nous pouvons utiliser est la sélection vocale de l'opérateur humain (DA). Dans les travaux que nous avons initiés, seul un petit nombre de personnages est mis à notre disposition. Dans cet article, nous proposons de remédier à ce problème en appliquant des méthodes de distillation de la connaissance en utilisant des données supplémentaires, provenant d'un domaine proche, pour

extraire les informations spécifiques au personnage/rôle. Plus généralement, nous pensons que les connaissances extraites, par exemple des jeux vidéo, pourraient être transférées à d'autres contextes, tels que les personnages de voix d'émissions de télévision.

Cet article est organisé comme suit. Nous présentons d'abord l'approche p-vecteur et le cadre général de distillation de la connaissance dans la partie 2. Ensuite, nous détaillons le corpus et nous décrivons le protocole expérimental que nous avons mis en place dans la partie 3. Nous présentons nos résultats et les discutons dans la partie 4. Enfin, les conclusions et perspectives sont données dans la partie 5.

2 Approche

2.1 Représentation dédiée au personnage

Ces dernières années, des architectures à base de réseaux de neurones profonds ont été proposées pour apprendre des espaces de représentation des données (Bengio *et al.*, 2013). Nous proposons d'apprendre une représentation dédiée aux voix jouées, appelée p-vecteur. L'espace p-vecteur (p signifiant "personnage") est optimisé sur une tâche de discrimination personnage/rôle. Il permet de projeter des segments de voix d'une manière qui maximise la variabilité des personnages.

En général, la représentation des données en entrée de méthodes d'apprentissage automatique a un impact fort sur les performances des applications. Ici, nous adoptons la représentation x-vecteur, initialement introduite en reconnaissance automatique du locuteur (Snyder *et al.*, 2018). Une grande quantité de données provenant de nombreux locuteurs sont utilisées pour créer l'espace de plongement des locuteurs (*speaker embeddings*). Des segments audio sont projetés dans cet espace et caractérisés par des x-vecteurs. Les x-vecteurs sont considérés ici comme une représentation compacte et de taille fixe d'une séquence vectorielle de paramètres acoustiques de longueur variable. Nous faisons l'hypothèse que les plongements de locuteurs contiennent des informations intriquées correspondant à la dimension personnage/rôle. Nous proposons donc de construire un nouvel espace de représentation (p-vecteur) capable de discriminer les différents personnages.

2.2 Distillation de la connaissance

Dans ce travail, nous devons traiter un nombre relativement restreint de données. Nous proposons d'utiliser la distillation de la connaissance afin d'exploiter des données supplémentaires d'un domaine proche pour pallier ce problème.

La distillation (Lopez-Paz *et al.*, 2016) unifie deux techniques qui introduisent toutes deux un maître pour guider un modèle d'élève tout au long de son processus d'apprentissage. La première technique introduit le concept d'information privilégiée (*Privileged Information*) (Vapnik & Izmailov, 2015) en ajoutant un nouvel élément $x_i^\star$ à la paire caractéristique-étiquette (x_i, y_i), avec $i \in [1...N]$ où N correspond au nombre d'exemples. La deuxième technique, appelée distillation de la connaissance (*Knowledge Distillation*) (Hinton *et al.*, 2015), permet à un réseau neuronal simple de résoudre une tâche compliquée en distillant les connaissances à partir d'un modèle "lourd". Plus généralement, le maître offre au modèle étudiant la possibilité d'apprendre à partir d'une décision qui n'est pas contenue dans l'échantillon d'entraînement (Lopez-Paz *et al.*, 2016). En règle générale, un réseau neuronal utilisant une fonction d'activation *softmax* fournit une probabilité pour chaque classe obtenue avec la formule suivante :

$$q_i = \frac{\exp\left(z_i/T\right)}{\sum_j \exp\left(z_j/T\right)}$$

où T fait référence à la température et z_i désigne la sortie calculée pour chaque classe de la couche finale. Une valeur plus élevée de T donne une distribution de probabilité plus progressive sur toutes les classes. Le fait est que le vecteur de probabilité q_i contient beaucoup plus d'informations qu'un simple codage à chaud (*one-hot encoding*). La distillation consiste à élever la température jusqu'à ce que le modèle du maître produise des cibles souples (*soft-targets*) appropriées. Ces dernières correspondent à l'information privilégiée. Comme illustré dans la figure 1, nous adaptons le modèle

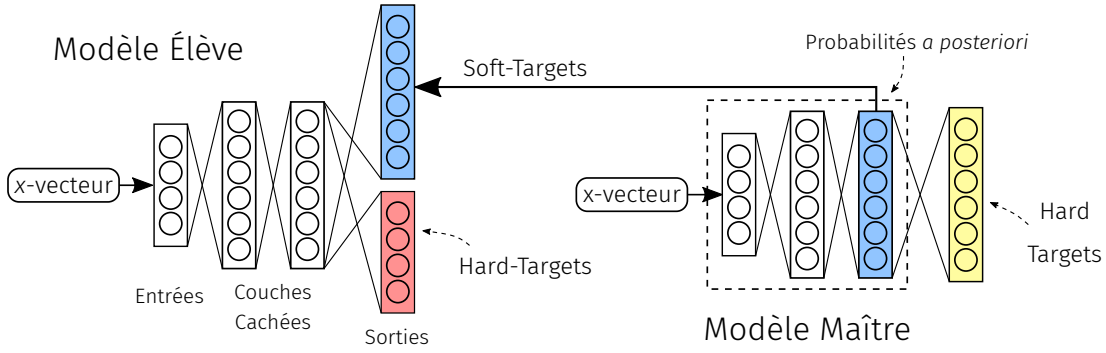

FIGURE 1 – Le modèle maître est entraîné à prédire des soft-targets afin que le modèle étudiant puisse les utiliser. Les deux modèles peuvent être entraînés sur le même corpus ou sur un corpus différent.

de l'élève aux cibles fixes (*hard-targets*, *i.e.* les étiquettes de personnages) et aux cibles souples (*soft-targets*) provenant du maître. Pour ce faire, nous utilisons un paramètre d'imitation noté λ qui contrôle la priorité entre l'imitation des probabilités *soft* et les prédictions habituelles des étiquettes *hard* pendant l'entraînement du modèle étudiant. Ceci est rendu possible en minimisant la perte :

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} [(1-\lambda) l(y_i,\ q_i) \ +\ \lambda l(s_i,\ q_i)]$$

où l désigne la perte d'entropie croisée et s_i fait référence aux *soft-targets* du modèle du maître. Le cadre maître-élève a été utilisé dans plusieurs travaux (Price *et al.*, 2016; Markov & Matsui, 2016; Li *et al.*, 2017; Watanabe *et al.*, 2017; Asami *et al.*, 2017; Joy *et al.*, 2017) pour une grande variété de tâches telles que la reconnaissance de la parole robuste au bruit, l'adaptation à un domaine, et la normalisation de locuteurs. L'approche proposée étend à l'origine ce cadre aux voix jouées et spécifiquement à la représentation des personnages/rôles.

Étant donné le nombre limité de personnages dans notre corpus, nous entraînons le modèle du maître sur un jeu de données supplémentaires contenant plus d'étiquettes de personnages. Nous supposons que cela pourrait aider le modèle étudiant à apprendre une représentation robuste et plus générale en s'adaptant aux *soft-targets* du maître.

Dans cet article, nous avons tout d'abord proposé, pour le casting vocal, un nouvel espace de représentation appelé p-vecteur obtenu à partir d'une couche d'embeddings d'un réseau de neurones profonds. Afin de pallier le problème de limitation des données, nous avons proposé une approche par distillation de la connaissance pour améliorer la représentation personnage/rôle.

3 Protocole expérimental

3.1 Corpus

Les extraits vocaux des personnages du jeu de rôle *Mass Effect 3* constituent notre corpus principal. Initialement édité en anglais, ce jeu a été traduit et doublé dans d'autres langues. Dans nos expériences, nous utilisons les versions anglaise et française des séquences audio, représentant $7,5$ d'heures de parole dans chaque langue. Les segments vocaux durent en moyenne 3 secondes, où chaque segment correspond à une interaction vocale unique. Un personnage est alors défini par un couple unique de voix jouées français-anglais. Pour éviter tout biais en termes d'identité du locuteur, nous considérons uniquement un sous-ensemble restreint de 31 personnages différents (13 féminins et 18 masculins), où nous sommes certains qu'aucun des acteurs ne joue plus d'un personnage. Chaque jeu de données anglais et français contient 10 000 segments vocaux.

Afin de remédier au nombre limité de personnages dans le corpus *Mass Effect 3*, nous utilisons des données supplémentaires d'un autre jeu vidéo multilingue appelé *Skyrim*. Nous limitons ce corpus aux dialogues anglais et français, pour un total de 120 heures de discours. Pour chaque langue, nous avons 50 000 segments annotés avec 30 étiquettes de personnages différents (7 féminins et 23 masculins). Comme nous n'avons pas suffisamment de garantie sur la correspondance français-anglais des segments et que nous ne sommes pas certains qu'un acteur joue un rôle unique, nous n'utilisons pas ce corpus dans l'étape d'évaluation. Il ne sert donc qu'à transférer les connaissances du maître au modèle étudiant dans le processus de distillation. Notons qu'il n'y a pas d'intersection entre les acteurs de *Skyrim* et *Mass-Effect 3*, ce qui évite un biais de locuteur dans l'ensemble de test. Enfin, tous les segments vocaux sont des fichiers audio enregistrés en studio de haute qualité. Tous les segments d'une durée inférieure à 1 seconde ont été supprimés.

3.2 Représentation des données

Nous réalisons une paramétrisation acoustique classique des segments audio que nous transformons en une séquence de caractéristiques de dimension 60 contenant 20 MFCCs incluant le log énergie et les dérivées de premier et second ordre ($\Delta + \Delta\Delta$). Nous utilisons une fenêtre glissante de Hamming de 20 ms (chevauchement de 10 ms), pour calculer les paramètres. Nous effectuons une normalisation des moyennes cepstrales et une détection d'activité vocale (VAD) pour supprimer les trames de faible énergie qui correspondent principalement au silence. Un système x-vecteur a été construit avec la boîte à outils Kaldi (Povey *et al.*, 2011) et entraîné sur le corpus Voxceleb (Chung *et al.*, 2018).

3.3 Protocole d'apprentissage

Le nombre de segments de voix dans le corpus *Mass Effect 3* n'est pas très bien réparti entre les différents personnages, ceux-ci n'ayant pas la même importance au sein du jeu. En conséquence, nous ne sélectionnons que 16 personnages (5 féminins, 11 masculins) qui ont tous, au moins, 90 segments vocaux dans les deux langues (anglais et français). Les segments sont tous choisis au hasard. De plus, nous créons une validation croisée en k-parties sur cet ensemble de personnages afin d'en avoir 4 dans chaque pli. Ainsi, nous avons $k = 4$ cas distincts, notés A, B, C et D qui couvrent tous les personnages, chaque cas impliquant 12 personnages pour l'apprentissage et 4 pour l'évaluation. Ces 4 personnages sont donc complètement absents du corpus d'apprentissage (ils ne partagent aucune étiquette ni aucun locuteur avec les données d'apprentissage), ce qui rend la tâche d'appariement de voix décrite dans 3.4 extrêmement difficile. Enfin, 20 % des données d'entraînement sont utilisées

pour la validation. En ce qui concerne le corpus additionnel, nous avons choisi le même nombre de segments pour chacun des 30 personnages et divisé en deux parties avec le même ratio affecté à la validation. Comme nous l'avons dit précédemment, aucune donnée de *Skyrim* n'est utilisée pour le test.

Les deux modèles maître et étudiant suivent la même architecture de réseau de neurones. Nous créons un Perceptron multicouche (MLP) en utilisant la boîte à outils Keras (Chollet *et al.*, 2015). Nous connectons une couche d'entrée avec 512 dimensions à 3 couches cachées de dimension 256 plus une couche d'embedding (*i.e.* correspondant aux p-vecteurs) de dimension 64, enfin une couche de sortie finale avec une fonction d'activation *softmax* . Les couches cachées sont combinées à une fonction d'activation tangente hyperbolique. Nous appliquons un dropout dans les 4 couches cachées avec les taux suivants : $0,25$, $0,25$, $0,25$, $0,5$. Nous utilisons une initialisation *Xavier* (Glorot & Bengio, 2010) et nous utilisons l'optimiseur *Adadelta* avec sa configuration par défaut pour résoudre la minimisation de la fonction de perte d'entropie croisée. De plus, nous utilisons une taille de batch de 12 exemples et nous entraînons le modèle sur 300 époques pendant que nous surveillons la fonction de perte sur l'ensemble de validation pour éviter le sur-apprentissage.

Le modèle du maître est entraîné sur les caractéristiques et étiquettes de l'ensemble de données supplémentaire (*Skyrim*), considéré comme une information privilégiée. Le modèle du maître peut être considéré comme un discriminateur de personnages/rôles. Ensuite, nous utilisons le maître pour calculer les *soft-targets Mass Effect 3* et former le modèle de l'élève sur les *hard-* et soft-targets de ce corpus. L'élève apprend à ajuster les 12 *hard-targets* et les 30 *soft-targets* en fonction du paramètre λ qui contrôle l'influence entre l'imitation des *soft-* et *hard-targets* pendant la phase d'apprentissage. Enfin, les p-vecteurs sont extraits de la couche d'embeddings du modèle de l'élève.

3.4 Evaluation

Pour évaluer la qualité de la représentation apprise, nous effectuons d'abord une analyse de clustering avec l'algorithme des k-moyennes sur les embeddings extraits (p-vecteurs). Nous avons expressément défini $k = 4$ pour refléter le nombre de personnages se trouvant dans l'ensemble de test. Tous les segments de voix qui sont ensuite rassemblés dans le même cluster sont affectés au personnage le plus représenté, de sorte qu'un cluster ait une seule étiquette. Ainsi, un score de F-mesure est calculé sur les segments sachant l'hypothèse de chaque cluster. Notons que plusieurs clusters peuvent être affectés au même personnage, ce qui pourrait être un problème. Mais nous considérons que cela reste un cas particulier indiquant un mauvais résultat.

De plus, nous évaluons l'approche sur une tâche d'appariement de voix avec le corpus *Mass Effect 3* en utilisant le système proposé dans (Gresse *et al.*, 2019). Ici, nous testons la capacité de faire une distinction significative entre les paires *cible* (*i.e.* paire de personnages identique en anglais-français) et *non-cible* (*i.e.* paire de personnages différente en anglais-français) lorsque nous entraînons le modèle de similarité avec les p-vecteurs.

4 Résultats

Nous utilisons différentes valeurs pour la température de distillation $T \in [1..5]$, les meilleurs résultats étant observés avec $T = 4$ en moyenne ($T = 1$ revient à apprendre sans distillation). De plus, nous vérifions les différentes valeurs dans la plage $[0, 1]$ pour le paramètre d'imitation λ : nous obtenons les meilleurs résultats en utilisant $\lambda = 0,3$ lors de la moyenne sur A, B, C et D.

4.1 Analyse par clustering

Le tableau 1 présente les résultats, en termes de F-mesure, de l'analyse par clustering utilisant avec la représentation p -vecteur. Nous observons que les p-vecteurs ont des scores de F-mesure bien meilleurs que les x-vecteurs (baseline dans ce travail), ce qui n'est pas surprenant puisque les x-vecteurs sont conçus pour se concentrer sur les identités vocales des voix des acteurs anglais et français, plus que sur leur personnage/rôle. Nous observons des résultats relativement bons, jusqu'à $0,78$ dans le meilleur des cas, ce qui indique la capacité des p-vecteurs à décrire automatiquement des personnages/rôles inconnus. En ce qui concerne le cas C, nous émettons l'hypothèse que les faibles scores de F-mesure peuvent résulter de la similitude inhérente entre les personnages – tous sont des soldats masculins – impliqués dans ce test particulier. Étonnamment, le système p-vecteur sans distillation fonctionne mieux dans ce cas spécifique.

	A	B	C	D
baseline (x-vecteur)	0,54	0,52	0,36	0,71
p-vecteur (sans distillation)	0,66	0,72	**0,59**	0,66
p-vecteur + distillation	**0,78**	**0,78**	0,40	**0,77**

TABLE 1 – F-mesures obtenues pour l'analyse par clustering sur les donnés de test.

La figure 2 illustre une projection à 2 dimensions de l'espace des p-vecteurs grâce à l'algorithme t-SNE. Sans surprise, nous voyons une distinction claire entre les personnages masculins et féminins dans les cas A, B et D (C ne contient que des soldats masculins). Les personnages de même sexe sont également correctement séparés. En considérant D, nous observons que chaque voix d'acteur des deux personnages *Hackett* (bleu) et *Illusive Man* (orange) ont une identité vocale forte, ce qui pourrait faciliter l'analyse par clustering et expliquer le score de F-mesure élevé inattendu (0.71) avec le système de référence x-vecteur.

4.2 Tâche de similarité

Nous évaluons également l'approche p-vecteur avec le système de similarité de voix dans le tableau 2. Les résultats sont présentés en termes d'exactitude et de test de Student (t-test). Le test statistique confirme la différence significative entre les scores de similitude des paires *cible* et *non-cible* puisque toutes les p-valeurs associées sont sous le seuil de rejet. En moyenne, la représentation p-vecteur surpasse le système de référence x-vecteur sur la tâche de similarité, avec une précision moyenne de $57\ \%$ et un t-score moyen de de 44,79 sur les quatre cas. De plus, nous constatons des variations plus faibles entre les différents cas de test. Compte tenu de la difficulté de cette tâche, nous pensons qu'ils constituent une preuve solide que les p-vecteurs contiennent une information personnage/rôle.

5 Conclusion

Dans cet article, nous avons tout d'abord proposé, pour le casting vocal, un nouvel espace de représentation appelé p-vecteur obtenu à partir d'une couche d'embeddings d'un réseau de neurones profond. Afin de pallier le problème de limitation des données, nous avons proposé une approche par distillation de la connaissance pour améliorer la représentation personnage/rôle. Nous avons observé une amélioration substantielle des résultats au travers de cette représentation p-vecteur, en comparaison de l'approche classique x-vecteur. Ces résultats démontrent que les p-vecteurs contiennent des informations dédiées à la dimension personnage/rôle.

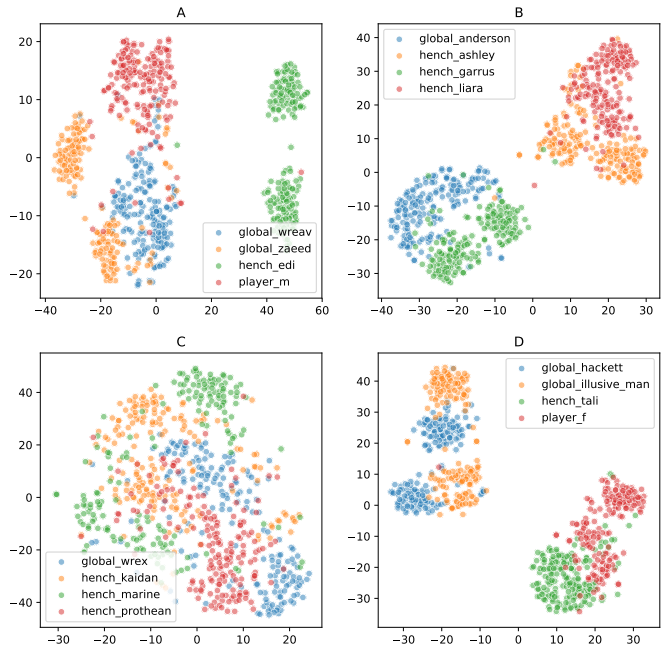

FIGURE 2 – Représentation dans l'espace des p-vecteurs pour chaque personnage dans A, B, C et D.

		exactitude	t-score
baseline x-vecteur	A	0,60	64,58
	B	0,52	20,63
	C	0,54	26,86
	D	0,49	-6,19
	moyenne	0,54	26,47
p-vecteur (sans distillation)	A	0,58	53,82
	B	0,54	20,70
	C	**0,57**	**49,86**
	D	0,54	23,34
	moyenne	0,55	36,93
p-vecteur + distillation	A	**0,63**	**80,00**
	B	**0,55**	**36,46**
	C	0,55	28,33
	D	**0,55**	**34,24**
	moyenne	0,57	44,79

TABLE 2 – Performance sur la tâche d'appariement de voix sur le corpus de test. L'exactitude sur la validation est généralemnt en dessous de 85 %.

Les paramètres T et λ, pour lesquels nous avons obtenu les meilleurs résultats en moyenne, offrent selon nous des perspectives d'analyses intéressantes. En particulier la pondération λ entre *soft-* et *hard-targets*, qui pourrait aider à mieux cerner la spécificité vocale des personnages de notre corpus par comparaison à un modèle de voix générique.

En raison des limites de notre corpus et malgré le protocole rigoureux que nous avons conçu, une certaine prudence doit être prise. Les résultats doivent être confirmés sur un corpus plus grand, avec plus de personnages, avant d'être capable de pouvoir généraliser les observations, par exemple à une autre culture. Le cadre maître-élève, pour être plus efficace, pourrait aussi être étendu sur de plus grands ensembles de données d'apprentissage, avec de nombreuses étiquettes de personnages et plusieurs acteurs par étiquette. De plus, les p-vecteurs permettent d'initier de nouvelles recherches sur les questions d'explicabilité, notamment dans le cadre des choix des directeurs artistiques. Nous souhaitons confronter les p-vecteurs à une simple décision binaire pour observer l'impact potentiel d'une caractéristique particulière sur la dimension du personnage. Les travaux futurs remplaceront le système de similitude, qui fait la distinction entre les paires de caractères identiques et différents, avec des caractéristiques explicatives (par exemple, le genre, la qualité de la voix, le timbre, la prosodie...).

Références

ASAMI T., MASUMURA R., YAMAGUCHI Y., MASATAKI H. & AONO Y. (2017). Domain adaptation of dnn acoustic models using knowledge distillation. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)* : IEEE.

BENGIO Y., COURVILLE A. & VINCENT P. (2013). Representation learning : A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, **35**(8), 1798–1828.

CHOLLET F. *et al.* (2015). Keras.

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

GLOROT X. & BENGIO Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the international conference on artificial intelligence and statistics*.

GRESSE A., QUILLOT M., DUFOUR R., LABATUT V. & BONASTRE J.-F. (2019). Similarity metric based on siamese neural networks for voice casting. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)* : IEEE.

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. In *INTERSPEECH*.

HINTON G., VINYALS O. & DEAN J. (2015). Distilling the knowledge in a neural network.

JOY N. M., KOTHINTI S. R., UMESH S. & ABRAHAM B. (2017). Generalized distillation framework for speaker normalization. In *INTERSPEECH*.

LI J., SELTZER M. L., WANG X., ZHAO R. & GONG Y. (2017). Large-scale domain adaptation via teacher-student learning.

LOPEZ-PAZ D., BOTTOU L., SCHÖLKOPF B. & VAPNIK V. (2016). Unifying distillation and privileged information. In *International Conference on Learning Representations*.

MARKOV K. & MATSUI T. (2016). Robust speech recognition using generalized distillation framework. In *INTERSPEECH*.

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

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

POVEY D., GHOSHAL A., BOULIANNE G., BURGET L., GLEMBEK O. *et al.* (2011). The kaldi speech recognition toolkit. In *IEEE 2011 ASRU*.

PRICE R., ISO K.-I. & SHINODA K. (2016). Wise teachers train better dnn acoustic models. *EURASIP Journal on Audio, Speech, and Music Processing*, **2016**.

SNYDER D., GARCIA-ROMERO D., POVEY D. & KHUDANPUR S. (2017). Deep neural network embeddings for text-independent speaker verification. In *INTERSPEECH*.

SNYDER D., GARCIA-ROMERO D., SELL G., POVEY D. & KHUDANPUR S. (2018). X-vectors : Robust dnn embeddings for speaker recognition. In *ICASSP* : IEEE.

SNYDER D., GHAHREMANI P., POVEY D., GARCIA-ROMERO D., CARMIEL Y. & KHUDANPUR S. (2016). Deep neural network-based speaker embeddings for end-to-end speaker verification. In *Spoken Language Technology Workshop (SLT)* : IEEE.

VAPNIK V. & IZMAILOV R. (2015). Learning using privileged information : similarity control and knowledge transfer. *Journal of machine learning research*, **16**, 2023–2049.

VARIANI E., LEI X., MCDERMOTT E., MORENO I. L. & GONZALEZ-DOMINGUEZ J. (2014). Deep neural networks for small footprint text-dependent speaker verification. In *ICASSP*.

WATANABE S., HORI T., LE ROUX J. & HERSHEY J. R. (2017). Student-teacher network learning with enhanced features. In *Acoustics, Speech and Signal Processing (ICASSP)* : IEEE.