La voix actée : pratiques, enjeux, applications

Mathias Quillot, Lauriane Guillou, Adrien Gresse, Rafaël Ferro, Raphaël Röth, Damien Malinas, Richard Dufour, Axel Roebel, Nicolas Obin, Jean-François Bonastre, et al.

▶ **To cite this version:**

Mathias Quillot, Lauriane Guillou, Adrien Gresse, Rafaël Ferro, Raphaël Röth, et al.. La voix actée : pratiques, enjeux, applications. 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.525-533. hal-02798582v3

HAL Id: hal-02798582
https://hal.science/hal-02798582v3

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

La voix actée : pratiques, enjeux, applications

Mathias Quillot[3] Lauriane Guillou[2] Adrien Gresse[3] Rafaël Ferro[1]
Raphaël Roth[2] Damien Malinas[2] Richard Dufour[3], Axel Roebel[1]
Nicolas Obin[1], Jean-François Bonastre[3], Emmanuel Ethis[2]
(1) STMS Lab - Ircam, CNRS, Sorbonne Université, Paris, France
(2) Laboratoire Culture et Communication, Avignon Université, Avignon, France
(3) Laboratoire d'Informatique d'Avignon, Avignon Université, Avignon, France

Résumé

La voix actée représente un défi majeur pour les futures interfaces vocales avec un potentiel d'application extrêmement important pour la transformation numérique des secteurs de la culture et de la communication, comme la production ou la post-production de voix pour les séries ou le cinéma. Un aspect central de la voix actée repose sur la notion d'interprétation, un aspect peu étudié dans la communauté scientifique de la parole. Cet article propose un état des lieux et une réflexion sur les défis scientifiques et les applications technologiques de la voix actée : à la croisée de l'acoustique, de la linguistique, de la culture, et de l'apprentissage machine. Une analyse préliminaire des pratiques permet de rendre compte de la diversité de l'écosystème des "*métiers de la voix*" et de pointer les fonctions et les conventions qui s'y rattachent. Nous nous intéresserons ensuite à la pratique particulière du doublage de voix, en faisant ressortir ses enjeux et problématiques spécifiques puis en présentant des solutions proposées pour modéliser les codes expressifs de la voix d'un acteur ou les choix d'un opérateur pour le doublage.

Abstract

Acted voice : practices, challenges, applications
The acted voice represents a major challenge for the next generation of voice interfaces with an extremely important application potential for the digital transformation of cultural and communication sectors, such as production or post-production of voice for series or films. A central aspect of the acted voice relies on the notion of interpretation, an aspect that has been little studied in the speech community. This article offers an overview and a reflection on the scientific challenges and technological applications of acted voice: at the crossroads of acoustics, linguistics, culture, and machine learning. A preliminary analysis of practices is presented to account for the diversity of the "voice professions" ecosystem and to point out the functions and conventions associated with it. The remaining of the paper focuses on the specific practice of voice dubbing, highlighting its specific issues and presenting some solutions proposed to model the expressive codes of an actor or the choices of an operator for dubbing.

Mots-clés : voix actée, interprétation, analyse des pratiques, doublage
Keywords: acted speech, interpretation, practice analyse, dubbing

Volume 1 : Journées d'Études sur la Parole, pages 525–533.
HAL : hal-02798582.
Cette œuvre est mise à disposition sous licence Attribution 4.0 International.

1 Introduction

Les avancées spectaculaires réalisées cette décennie en traitement automatique de la parole, pour des tâches comme la reconnaissance ou la synthèse de parole, ont permis la réalisation de nombreuses applications devenues d'usage quotidien, comme l'interaction vocale avec un smartphone. Tour à tour et successivement voix de "laboratoire", "pathologique", "ordinaire", "spontanée", "expressive" ou "conversationnelle", la recherche sur la parole s'attaque à des domaines variés et s'affranchit peu à peu des défis posés par des facteurs de variabilité de plus en plus nombreux et complexes. Par contraste, l'étude de la parole "actée" demeure aujourd'hui marginale dans la communauté de la parole au point où un article sur le doublage était encore catalogué il y a quelques années dans la catégorie "parole anormale" ("abnormal speech"). Contrairement à la parole "ordinaire" ou "spontanée", la voix "actée" est un artifice, une construction fruit d'une interprétation maîtrisée et planifiée de sa voix par un acteur afin de produire un effet désiré chez un spectateur, par exemple rendre manifeste le comportement d'un personnage fictif, et faciliter la crédibilité et l'immersion du public dans une situation ou une trame. La voix est souvent dénaturée, grossie, pour rendre audible et sensible les effets expressifs produits par le comédien. Ainsi, la voix actée élargit le champs de recherche des fonctions expressives de la voix humaine et ouvre sur le domaine encore peu étudié de l'*interprétation*. Par ailleurs, une brève revue des "*métiers de la voix*" (voix-off pour le commentaire, publicité, voix doublée, voix au théâtre, voix au cinéma, etc…) montre la diversité des pratiques et les usages spécifiques de la voix principalement dans le secteur de la culture et de la communication, et des industries créatives. Pour comprendre les fonctions de la voix dans ces usages, la voix ne peut être considérée seule, mais dans une relation à un être humain dont elle ne constitue que l'une des modalités d'expression : l'incarnation de la voix par un corps (les voix-off sont ainsi totalement désincarnées) ou l'adéquation d'une voix et de son corps (le doublage offre un exemple de composition du corps d'un acteur réel ou fictif et de la voix d'un autre acteur) (Le Breton, 2011).

L'appréhension de la voix actée soulève à la fois un ensemble de questions inédites pour la recherche et de nouvelles possibilités technologiques et applicatives, en ouvrant sur la dimension interprétative de la voix humaine. Premièrement, l'étude de la voix actée — qui est par essence expressive — nécessite d'être en mesure de se confronter à des modalités de productions inusuelles et des registres extrêmes de variations acoustiques. Par ailleurs, elle pose la vaste question de la notion d'interprétation appliquée au domaine de la parole. La compréhension de cette notion nécessite d'étudier les pratiques des métiers de la voix : depuis la diversité de ses pratiques et de ses fonctions jusqu'aux conventions et aux choix de l'interprétation. On peut citer, à titre d'exemple, la production des voix dans les industries créatives qui suit un protocole extrêmement codifié, faisant intervenir un ensemble de conventions implicites depuis le choix d'un acteur capable d'incarner un physique ou un personnage, la supervision de son interprétation par un directeur artistique lors de séances d'enregistrements, jusqu'à la diffusion finale à un public cible. L'appréhension de la voix actée ne se limite manifestement pas à des facteurs acoustique ou linguistique et fait intervenir des facteurs sociologiques et culturels nécessaires pour sa compréhension et à sa modélisation. Cet article tente de présenter un état des lieux et une réflexion d'ensemble sur les défis et les applications de la voix actée, à la croisée de l'acoustique, de la linguistique, de la culture, et de l'apprentissage automatique. Il propose en outre un cadre méthodologique pour mieux définir le champs de la voix actée et en particulier les critères utilisés par un opérateur humain expert pour qualifier la voix et l'interprétation vocale d'un acteur. La compréhension des processus cognitifs mis en œuvre par un opérateur humain dans ses choix d'un acteur ou d'une interprétation devrait permettre, à terme, d'apprendre à la machine à reproduire ces choix.

L'article est organisé de la manière suivante : une première partie présente une enquête sociologique menée sur les pratiques des métiers de la voix, une seconde partie esquisse les défis et problématiques spécifiques au traitement de la voix actée, et une dernière partie présente une application dans le cadre du doublage au cinéma.

2 Enquête sociologique sur le rapport à la voix dans le travail d'interprétation

Dans l'optique de produire des connaissances sur les qualités de la voix actée, nous avons mis en place une enquête auprès de professionnels du doublage vocal. Artistes interprètes, directeurs artistiques et adaptateurs ont été interrogés dans le cadre de 7 entretiens semi-directifs conduits auprès de 9 enquêtés dans trois contextes en 2019 (Festival d'Avignon, festival international des voix du cinéma d'animation Voix d'Étoiles, échanges téléphoniques). L'analyse de ce matériau a permis de mettre en exergue un ensemble de caractéristiques propres à la voix dans le contexte du jeu d'acteur et du doublage, celui des fictions audiovisuelles (prise de vue réelle et animation) et des documentaires ; une diversité assurant une représentativité au sein de ce secteur professionnel.

Dérivé de l'anglais *dubbing*, le doublage est aujourd'hui un terme intégré dans le langage courant. Ce sont les Cahiers du cinéma qui, pour la première fois en mai 1930, ont présenté cette technique au public francophone : « Le *dubbing* ou doublage, est un système qui consiste à tourner des paroles en n'importe quelle langue et à l'adapter sur une bande tournée primitivement en parlant américain » (Cornu, 2014). Le doublage s'est institutionnalisé en un secteur à part entière l'industrie audiovisuelle. Il est particulièrement sollicité au regard de l'expansion du nombre de productions en circulation nécessitant une réponse technique et artistique de qualité pour leur diffusion internationale. Le doublage ne concerne pas tous les pays de la même manière, certains comme les pays scandinaves privilégient les versions originales sous-titrées, à l'inverse de la France qui est l'un des principaux *doubleurs* (en découle aussi une grande qualité des doublages français). D'autres pays en restent éloignés pour des raisons socio-culturelles : la Chine réserve ainsi le doublage aux films d'animation pour le jeune public. Un doublage dont le synchronisme labial n'est pas suffisant, voire non adapté au genre cinématographique ou audiovisuel, peut générer une suspension volontaire d'incrédulité pour les spectateurs (une notion entendue au sens de Samuel Taylor Coleridge dans Biographia (1817), relevant d'abord du champ littéraire, qui renvoie à l'acceptation, par le récepteur d'une œuvre, d'un univers fictionnel).

Le doublage tel que nous le connaissons aujourd'hui est un aboutissement d'une histoire culturelle et technologique. Avant « d'être parlant, le cinéma était muet, mais pas silencieux. Certes, les films muets n'offraient pas une parfaite *synchronisation*, mais ils n'étaient pas dénués de sons et d'illustrations musicales » (Roth, 2017). Le doublage se caractérise par un impératif de synchronisation, à commencer par un synchronisme labial (Bosseaux, 2015). La généralisation du doublage va entraîner la création de métiers qui lui sont spécifiques à l'instar des traducteurs-adaptateurs, des repéreurs et détecteurs, des coordinateurs linguistiques ou encore des monteurs en synchronisation. Elle va également générer une spécialisation de certains comédiens dont la carrière peut se concentrer presque exclusivement sur le doublage (post-synchronisation, narration, *voice over*, habillage d'antenne, publicité, bandes annonces, etc.). D'ailleurs, les équipes artistiques tentent de garantir une certaine continuité dans la relation au spectateur en gardant le même doubleur pour un comédien très reconnu (i.e. Patrick Poivey pour Bruce Willis ou Céline Monsarrat pour Julia Roberts). Si « L'arrivée du parlant va influencer la façon de produire le cinéma, de le voir et de l'entendre » (Roth, 2017 : 26), l'expansion du doublage va générer une spécialisation au sein de ce secteur industriel. Nous en distinguons quatre grands

champs : la postsynchronisation (domaine de la fiction), la narration (une voix *off* qui décrit le développement d'une trame, le plus souvent un documentaire), le *voice over* (traduire la personne qui parle à l'écran sans synchronisme labial), la publicité et les autres activités comme l'habillage d'antenne, les voix institutionnelles ou les bandes annonces.

L'enquête conduite auprès des comédiens a permis de mieux qualifier le travail sur la voix dans un contexte d'interprétation, notamment du doublage qui s'y consacre pleinement, soit le passage de la voix naturelle à la voix actée, et quelquefois chantée. Les artistes apprennent ainsi à *fixer leur voix*, à stabiliser leur *identité*, voire leur *signature de voix*. Certains acteurs admettent le fait de *changer leur voix* pour s'adapter au personnage ou à l'acteur en version originale. Pour d'autres, il s'agit plutôt de *prendre un accent*, de *moduler la voix*, de la *modifier* ou de l'*adapter*, mais non pas de la changer ou de la *transformer*. D'autres ont un langage plus artistique et parlent de *modeler la voix*, d'imaginer une sculpture, voire d'être *artisan de la voix*, de travailler à partir de leurs *tessitures* et de l'*amplitude de la palette vocale* qu'ils maîtrisent. Le travail sur la voix s'accompagne de deux autres démarches : une attention portée à la *respiration* et au *rythme du personnage* ou de l'acteur à doubler. Celui-ci peut relever de la *prosodie* de l'acteur ou de la langue (les acteurs doublant des productions audiovisuelles sont le plus souvent tributaires des traducteurs et des adaptateurs du scénario ou du texte de la version originale vers la version française). Plus encore, le travail sur la voix doit être conforme « au genre que l'on perçoit » qui se construit dans « l'immédiateté d'un objet concret avant de rejoindre les catégories de discours et vient fermer à sa façon une des fonctions inachevées telle qu'elle est directement reçue depuis le film projeté » (Ethis, 2004 : 166).

Les enquêtés ont également été interrogés sur une part de mimétisme vis-à-vis de la version originale. Il est intéressant de constater que la plupart des comédiens s'appuient sur la version à doubler, tout en s'attachant à conserver leur identité vocale dans la mesure où c'est à partir de celle-ci qu'ils ont été choisis. La voix actée se caractérise par des qualificatifs qui renvoient en particulier à l'idée d'*intention* : elle est le reflet d'une *énergie* qui appartient à l'artiste en version originale ou à la situation présentée à l'écran. La voix actée est ainsi une adresse qui ne doit ni trahir une œuvre, ni le jeu en version originale. Elle se situe dans un système de contraintes communicationnelles et symboliques (genre audiovisuel, registre de jeu, registre de langue). Les artistes devant être en capacité d'incarner une multitude de personnages durant leur carrière, et donc s'inscrire dans divers registres de jeu, au-delà de la voix elle-même, c'est leur palette, autrement dit leur amplitude vocale, qui importe. Enfin, la question de la ressemblance physique entre un artiste en version originale et en version française a été posée : si elle est parfois un repère, c'est souvent la personnalité de l'interprète qui va importer dans un casting.

Pour conclure, il convient d'insister sur le positionnement du doublage dans l'industrie audiovisuelle contemporaine : il est à la fois un espace technique, voire technologique, économique, artistique et culturel. Le travail sur la voix actée doit s'inscrire dans des registres filmiques ou audiovisuels pour être au service d'un projet artistique et des publics auxquels il s'adresse.

3 Problématiques et verrous scientifiques pour le traitement de la voix actée

La suite de cet article s'intéresse à la voix au cinéma et plus particulièrement les voix pour le doublage, dont le *voice over*. Plusieurs protagonistes interviennent dans la chaîne de traitement du *voice over*. Tout d'abord, les voix des doubleurs doivent à la fois rester fidèles à l'oeuvre

d'origine, mais aussi correspondre aux attentes du public cible. Le public est l'auditeur final du travail et doit être immergé dans l'oeuvre à son visionnage. Ensuite, la qualité d'un doublage ressort de la responsabilité des Directeurs Artistiques (DA). Ce métier nécessite une expérience en comédie et une culture cinématographique développée afin de sélectionner des voix, via le casting vocal, et de guider les acteurs lors d'enregistrements. Pour finir, le client a des objectifs commerciaux ou artistiques et contrôle les prises de décision. Il intervient lors de la sélection des acteurs et valide la qualité du doublage. Cette chaîne implique de nombreux choix humains, ainsi que des conventions culturelles et esthétiques qui sont implicites ou explicites. Identifier les codes explicitement ou modéliser ces choix implicitement imposent des verrous scientifiques que nos travaux cherchent à briser. Pour cela, nous avons concentré nos efforts sur trois points principaux : la taxonomie de la voix, la modélisation d'un processus artistique et la dépendance de réception de la voix à la langue et la culture.

Dans le but de comprendre et modéliser les mécanismes du doublage, il est nécessaire de savoir représenter une voix actée. Cette représentation peut être construite à partir de codes explicites donnant définition au terme palette vocale. Il n'existe cependant aucune taxonomie partagée par les DA et autres protagonistes de la chaîne de production du doublage. Ces faits rendent difficile la création d'un protocole d'annotation. Quand bien même une taxonomie existerait, les DA ont hélas très peu de temps à accorder à l'annotation. Sans ces annotations, il est difficile à la fois de construire des modèles d'apprentissage supervisé pour détecter automatiquement des facteurs de variabilité dans la voix, d'isoler et d'analyser les impacts de certaines caractéristiques dans des prises de décisions. Evaluer des systèmes automatiques de classification de la voix actée, et expliquer de manière intelligible à l'utilisateur les critères qui ont guidé les décisions de nos systèmes sont aussi des limites auxquelles nous nous confrontons. Autant dans une problématique sociologique qu'informatique, se pose alors la question suivante : comment définir une taxonomie de la voix actée sans l'accès à une connaissance de ses codes explicites ?

Le jugement de l'adéquation entre une voix et son personnage conduit par les DA implique de nombreux facteurs humains et perceptifs. Pour modéliser un tel processus, il nous faut comprendre quels critères ou filtres y interviennent. Nous supposons que cette adéquation porte autant sur la nature de la voix d'un comédien que sur l'interprétation qu'il y apporte en jouant le personnage cible. Nous étudions dans nos travaux comment inférer ces critères et filtres depuis des enregistrements de voix et des décisions de DA. Les critères peuvent être différents en fonction du DA. (chacun présente un profil différent, avec une expérience et des préférences artistiques qui lui sont propres) A l'instar de la reconnaissance de la parole où certains modèles sont adaptés au locuteur, il nous est nécessaire d'étudier l'adaptation de nos modèles au DA cible. Différentes pistes sont à explorer. La construction d'un modèle par DA. La construction d'un modèle universel. L'adaptation d'un modèle universel à chaque DA. Ces modèles doivent aussi s'adapter dans le temps. Le DA, à force d'expérience, évolue et ses décisions aussi.

La voix actée est intrinsèquement liée à la perception humaine. Partant de ce postulat, nous recherchons quels codes sont universels et lesquels varient en fonction des individus. Ces codes sont sujets à un processus cognitif qui rend leur modélisation difficile. Les facteurs de variation de ces derniers peuvent être d'ordre culturel ou liés à la structure de la langue parlée. Cette variabilité peut intervenir dans différentes dimensions. Comme, par exemple, la dimension de l'émotion. Certaines émotions sont universelles et ne varient pas d'une langue à l'autre, là où d'autres sont dépendantes de l'individu. Ohala (Ohala, 1996) unifie des théories guidant les recherches sur l'expression d'émotion par la voix et tend à montrer que certaines émotions sont partagées par différentes cultures. De la même manière, nous supposons que le jeu des personnages est construit pour mettre en avant des stéréotypes qui diffèrent d'une culture à l'autre.

Pour finir, tout comme notre société évolue, ces codes évoluent, et nous devons chercher à prendre en compte dans nos modèles l'évolution temporelle de la réception de la voix.

4 Application : l'aide au doublage de voix d'acteurs

Un cadre applicatif possible pour la voix actée pourrait être un système de recommandation de voix pour les DA. Un tel système consisterait à proposer automatiquement un ensemble de voix selon une requête formulée. Dans notre contexte d'étude, nous proposons de nous intéresser au casting vocal et plus particulièrement au doublage : considérant une voix dans une langue source, quelles voix d'acteurs dans une langue cible peuvent le mieux correspondre aux attentes d'un DA ? En réponse à ces problématiques, des travaux ont proposé d'étudier la similarité de voix, qui est une notion centrale dans l'élaboration de systèmes de recommandation. Elle consiste à définir une mesure de similarité entre deux voix. Comment considère-t-on que deux voix sont similaires ? A partir de quelles données, vérité, peut-on construire une telle mesure ? L'apparition de la similarité de la voix au sens du personnage a donné des réponses à ces questions. Celle-ci part de l'hypothèse suivante : il existe, dans le signal acoustique de parole, des signes acoustiques caractéristiques du personnage joué. La similarité de la voix au sens du personnage consiste alors à définir une mesure qui, si les deux acteurs jouent le même personnage, doit informer que les voix sont similaires. Dans le cas contraire, la mesure doit indiquer que les deux voix paraissent dissimilaires. Deux approches différentes ont été proposées pour cette mesure de similarité.

Dans la première approche (Gresse, 2017), deux modèles *i*-vecteurs (Dehak, 2010) sont appris, l'espace source (anglais), l'espace cible (français). Le modèle *i*-vecteur est un modèle de représentation de voix plus communément utilisé pour des tâches de reconnaissance du locuteur. Le système proposé consiste à appliquer une matrice de passage *W* à l'espace cible pour le projeter dans un nouvel espace, que nous nommerons l'espace modifié, comparable avec l'espace source. La matrice *W* est apprise sur un ensemble de données d'entraînement. Un score de similarité est ensuite calculé entre les vecteurs de l'espace source et de l'espace modifié dans le cas du système B et C. Pour le système *baseline* A, la mesure est calculée entre l'espace source et l'espace cible.

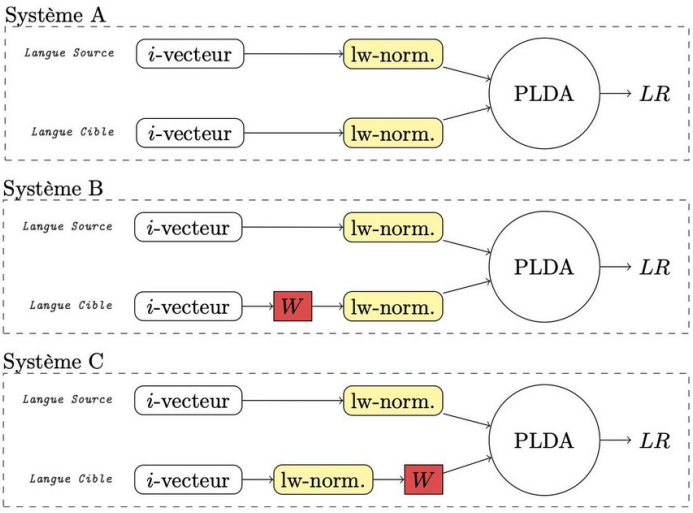

Figure 1: Systèmes de similarité proposés dans (Gresse, 2017)

Dans la seconde approche (Gresse, 2018), un réseau de neurones siamois est entraîné à projeter les deux voix fournies en entrée, une en anglais et une en français, dans un espace commun où celles-ci sont comparables au moyen d'un score de similarité. L'objectif de cet espace est de représenter la dimension personnage des voix qui y sont projetées indépendamment du locuteur, de la langue et du contenu linguistique. Dans la continuité de ces travaux, nous nous sommes intéressés à la création d'un espace personnage en s'inspirant des modèles x-vecteurs (Snyder, 2018). Nous entraînons un réseau de neurones profond à reconnaître le personnage joué quelque soit sa langue. Le réseau fournit en sortie des probabilités indiquant l'appartenance de l'enregistrement à un personnage spécifique. Une fois le réseau entraîné, sa dernière couche sert d'espace de représentation, le p-vecteur. La figure 2 donne un exemple de cet espace de représentation orienté "personnage". Des enregistrements de 4 paires de voix, chacune composée d'une voix anglaise et d'une voix française jouant le même personnage, sont projetée dans l'espace p-vecteur dans lequel nous avons appliqué un algorithme de regroupement pour tenter de retrouver les classes personnage. Chaque paire est représentée avec une couleur donnée. Notons que le système n'a jamais vu durant la phase d'apprentissage, ni les personnages, ni les locuteurs concernés.

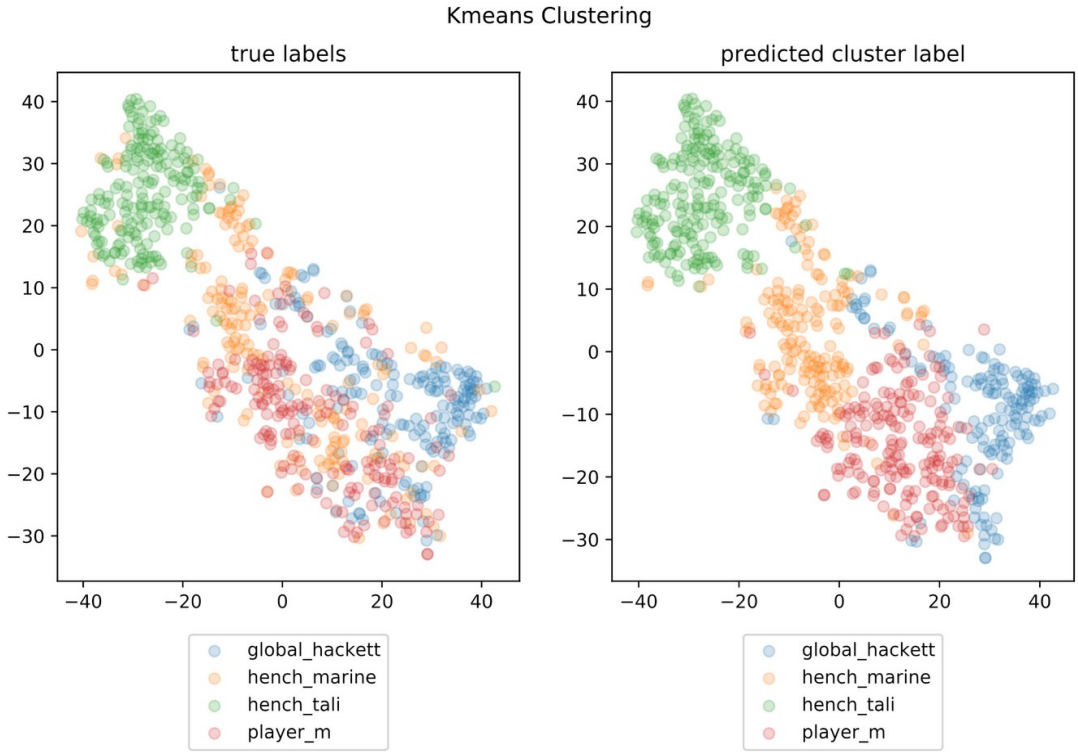

Figure 2: Comparaison des vrais labels de personnage avec les labels associés à chaque enregistrement par un algorithme de clustering sur un espace de représentation de la dimension personnage de la voix actée.

D'autres travaux ont proposé une mesure de similarité perceptuelle (Obin, 2014). Cette contribution se démarque des précédentes en proposant d'abord de représenter un enregistrement par des classes au lieu de calculer un score directement sur un espace acoustique. Ces classes sont liées à la physiologie, la phonétique, au timbre, à l'articulation, à la prosodie ou au jeu d'acteur. Un système de classification multi-labels est alors entraîné sur des enregistrements étiquetés par un expert. Pour un enregistrement donné, le système calcule un vecteur de probabilités où chaque valeur représente l'affinité de l'enregistrement pour le label. Chaque enregistrement est donc représenté par un vecteur d'affinité, référé comme la signature vocale de l'enregistrement. Appliquer une distance entre deux signatures revient à calculer un score de similarité entre deux

enregistrements. L'article confronte deux signatures vocales, l'une issue de vecteurs d'affinités appris depuis les *i*-vecteurs, l'autre des *i*-vecteurs + *PLDA*, et montre au travers d'une expérience perceptive que l'apprentissage du classifieur a permis de mettre significativement en lumière, dans l'espace i-vecteur, des informations caractéristiques de la perception de la voix actée.

D'un point de vue applicatif, ces travaux sur la similarité et la représentation de voix ouvrent des perspectives pour construire des systèmes de recommandation pour le casting vocal. Ces espaces de représentations vont permettre d'initialiser les systèmes de recommandation pour proposer des voix qui seront pertinentes ou inattendues pour les DA. Un tel système pourra prendre la forme d'une application informatique où le DA pourra demander des recommandations et réagir aux résultats. Dans une dynamique itérative, cette application fournira de nouvelles données à la recherche qui alimenteront nos travaux pour améliorer encore la finesse des recommandations proposées aux DA.

5 Conclusion

Nous avons proposé dans cet article un état des lieux et une réflexion sur les défis scientifiques et les applications technologiques de la voix actée. L'appréhension de la voix actée soulève un ensemble de questions inédites de recherche pour la communauté de la parole. Tout d'abord, une analyse des pratiques montre qu'il existe une grande diversité de métiers de la voix, chacun avec ses fonctionnalités et ses objectifs propres. Les pratiques de ces métiers ne sont cependant pas homogènes et relèvent d'une dimension artistique. En particulier, l'étude des voix de doublage montre l'existence d'un cadre partagé pour la sélection et la direction de voix d'acteurs dans le but de produire des effets définis, mais les codes sous-jacents restent implicites et difficiles à formaliser. Ainsi, l'appréhension de la voix actée ne se limite manifestement pas à des facteurs purement acoustiques ou linguistiques, mais fait intervenir des facteurs sociologiques et culturels reflétés par les choix d'opérateurs humains. La modélisation des choix des opérateurs humains nécessite donc de modéliser explicitement ou implicitement les filtres cognitifs ayant amené à ces choix. La mise en évidence de codes explicites par une taxonomie partagée et standardisée et/ou la modélisation implicite des choix représente clairement un défi pour la recherche dans ce domaine. Une étude est en cours pour essayer de faire apparaître une taxonomie partagée par les professionnels pour qualifier une voix dans les métiers en production ou post-production de voix. Par ailleurs, nous avons présenté dans cet article une solution pour modéliser implicitement les choix d'un opérateur autour de la notion de "personnage", avec les résultats préliminaires obtenus en doublage de voix. Les études à venir visent à préciser les méthodologies présentées pour permettre de mieux comprendre et modéliser les nombreux facteurs de la voix actée et, à terme, les exploiter pour créer des applications technologiques pour les industries créatives.

Remerciements

Cette recherche a été menée avec le projet TheVoice (ANR-17-CE23-0025) financé par l'Agence Nationale de la Recherche.

Références

Bosseaux C. (2015). *Dubbing, Film and Performance. Uncanny Encounters.* Bern : Peter Lage.

Cornu, J.-F. (2014). *Le doublage et le sous-titrage: Histoire et esthétique*. Presses universitaires de Rennes.

Ethis E., Malinas D. (2012). *Films de Campus. L'Université au cinéma*. Paris : Armand Colin.

DIAS G., Éd. (2015). *Actes de TALN 2015 (Traitement automatique des langues naturelles)*, Caen. ATALA, HULTECH.

Ethis E. (2004). *Pour une po(ï)étique du questionnaire en sociologie de la culture. Le spectateur imaginé*. Paris : L'Harmattan.

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*. INTERSPEECH.

Gresse A., Quillot M., Dufour R., Labatut V., Bonastre J.-F. (2018). Similarity Metrics Based on Siamese Neural Networks for Voice Casting. ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)

N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouelle (2010). "Front-End Factor Analysis For Speaker Verification," IEEE Transactions on Audio, Speech and Language Processing, vol. 19, no. 4, pp. 788 – 798

Launier J.-J. (2016). *L'Art des studios d'animation Walt Disney. Le mouvement par nature*. Art Ludique - Le Musée.

Le Breton, D. (2011). *Eclats de voix.. Une anthropologie des voix*. Métailié.

Obin N., Roebel A., (2016). Similarity search of acted voices for automatic voice casting, IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 24 , Issue. 9, p. 1642 – 1651.

Obin N., Roebel A., Bachman G. (2014). On Automatic Voice Casting for Expressive Speech: Speaker Recognition vs. Speech Classification

Ohala J., (1996). *Ethological Theory and the Expression of Emotion in the Voice*. Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96

Roth R. (2017). À l'écoute de Disney. *Une sociologie de la réception de la musique au cinéma*. Paris : L'Harmattan.

Snyder D., Garcia-Romero D., Sell G., Povey D., Khudanpur S. (2018). *X-Vectors: Robust DNN Embeddings for Speaker Recognition*. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)