SORBONNE UNIVERSITÉ FACULTÉ DE SCIENCES
ANNÉE UNIVERSITAIRE 2019-2020 M2 ISSI CMI
PERCEPTION ET MODÉLISATION DE L'INTERACTION - MU5SPI01

Projet modélisation de l'interaction

Projet Mimos (μῖμος)

Application à la *Trumpologie*

Étudiants :

Jeremy JASPAR - 3401421
Mathias ROESLER - 3523660
Camilo SARMIENTO - 3523391

Résumé

L'objectif de ce travail a été la création d'un système capable d'identifier les *Deep Fakes*, dès lors qu'ils utilisent des imitateurs pour réaliser la voix. Pour cela, un système de reconnaissance vocale bio-inspiré, i.e. basé sur l'acoustique, la phonétique et indépendant de la sémantique a été développé. Ce dernier caractérise les signaux vocaux par neuf coefficients MFCC, ainsi que par la fréquence fondamentale et l'énergie des différents segments obtenus par segmentation du signal. Puis, grâce à une distance de Bhattacharyya, la mesure de similarité avec les échantillons correspondant à la personne concernée est réalisée, permettant ainsi de déterminer s'il s'agit ou pas d'un *Deep Fake*.

Introduction

Dans le cadre de cette étude, nous avons souhaité développer un système de reconnaissance d'identité à partir de la voix. Plus précisément, nous avons voulu mettre en place un système permettant de distinguer l'actuel président des États-Unis, Donald J. Trump, d'autres personnes, notamment d'imitateurs.
Cette application très spécifique a pour but d'être généralisée et de s'inscrire dans l'actualité, une actualité particulière où la notion de vérité est confrontée aux *Deep Fakes*. En effet, le système que nous avons développé pourrait, à terme, permettre d'identifier ces vidéos créées par des réseaux de neurones de plus en plus performants à partir de leur point faible, la voix. Dans le cadre de cette étude, nous voulons évaluer notre algorithme en utilisant des imitateurs performant dans un registre comique. Afin de rester sur une analyse de la voix et de ne pas biaiser nos résultats en nous concentrant sur le contexte, nous avons mis en place un système indépendant du contenu sémantique du signal étudié.

La reconnaissance de la voix indépendemment du contenu sémantique de celui-ci, est un mécanisme très développé chez l'être humain et ceci dès le plus jeune âge. En effet, comme l'explique l'article Effects of Experience on Fetal Voice Recognition [1], cette reconnaissance se fait dès la période prénatale, bien avant l'acquisition du langage. Cette observation fait donc penser que cette capacité de reconnaissance se fait donc principalement sur des aspects non verbaux.
Dans l'article Multiple levels of linguistic and paralinguistic features contribute to voice recognition [2], les chercheurs s'intéressent à l'importance de chaque niveau de décomposition de la parole dans la reconnaissance d'une personne grâce à sa voix. Pour cela, ils demandent à cinq individus de produire cinq types de signaux vocaux : un signal exclusivement non voisé, on ne dispose donc que de l'information acoustique ; un signal en mandarin, on enrichi le signal en apportant de la phonétique ; un signal en allemand, la phonétique est plus riche car plus proche de l'anglais ; un signal en pseudo anglais, l'information contenue dans la phonétique est maximale sans pour autant ajouter de l'information sémantique ; un signal en anglais, nous avons à présent accès à la sémantique. Une fois ces signaux obtenus, ils les font écouter à un échantillon d'individus qui doivent associer le signal vocal à un des cinq échantillons. Les résultats montrent que l'information acoustique est suffisante pour avoir une reconnaissance meilleur que le hasard et que, plus le signal est enrichi, plus la reconnaissance à partir de la voix est précise. De plus, la comparaison des performances montre aussi que l'apport de l'information phonétique est ce qui permet d'améliorer le plus les performances. Pour rappel, lorsque l'on étudie la phonétique on s'intéresse au conduit vocal et plus précisément les différentes configurations de celui-ci. Ces résultat confirment bien que la capacité de reconnaissance chez l'être humain se fait principalement sur des aspects non verbaux, observation qui vient réconforter notre choix de ne pas prendre en compte la sémantique dans notre système.

La première partie de ce document présente l'étape de construction des bases de don-

nées utilisées pour l'apprentissage, la validation et le test de notre algorithme. Puis, nous passerons à une des étapes les plus importantes dans le processus, l'extraction des caractéristiques de la voix qui nous intéressent. Puis, nous présenterons la méthode de comparaison que nous avons utilisé pour finir avec une présentation des résultats obtenus.

1 Bases de données

La création des bases de données peut-être décomposée en deux étapes : sélection d'extraits pertinents et divers; segmentation des extraits en échantillons. Ces derniers ont une durée comprise entre huit et douze secondes. Il est également important de préciser que lors de la sélection, nous avons fait en sorte de prendre des passages ayant le moins de bruit possible, évitant les cas avec des interventions de tiers ou des applaudissements.

Trois bases différentes ont été construites. Comme nous l'avons mentionné précédemment, nous voulons uniquement apprendre à reconnaître le vrai Trump, la base d'apprentissage aura donc uniquement des exemples le concernant. Pour cette base, nous avons pris des interventions de Trump dans trois contextes différents : un discours de campagne lors des élections de mi-mandat, un discours face à l'organisation des Nations Unies et une conférence de presse. Puis, pour chaque intervention, nous avons sélectionné quatre extraits, obtenant ainsi une base composée de douze exemples.
Les deux autres bases sont celle de validation et de test. La première catégorie que nous avons créée est celle des imitateurs de Donald Trump (Alec Baldwin, John di Domenico, Jimmy Fallon et Trevor Noah). La deuxième est composée d'hommes politiques américains (George W. Bush, Barack Obama, Bernie Sanders et Adam Schiff) et la troisième de femmes politiques américaines (Hillary Clinton, Kamala Harris, Nancy Pelosi et Elizabeth Warren). La quatrième regroupe à nouveau des hommes politiques, qui cette fois-ci sont français (Jacques Chirac, Nicolas Sarkozy, François Hollande et Emmanuel Macron). La cinquième catégorie joue le rôle de catégorie témoin, car elle est composée de personnes de divers horizons dans différents contextes, lointains à celui qui nous intéresse : présentateurs (Stephen Colbert et Seth Meyers), comédien (James Veitch) et *youtubeur* (SuperCarlinBrothers). Finalement, la dernière catégorie correspond à d'autres interventions de Donald Trump. Une fois le choix des différentes catégories réalisé, nous récupérons six extraits pour chaque individu dans la base. Nous avons finalement pris trois de ces extraits pour chaque individu afin de former la base de validation, le reste des extraits forme alors notre base de test.

2 Extraction des caractéristiques

Comme expliqué dans l'introduction, la phonétique est le niveau de langage le plus exploité dans la reconnaissance à partir de la voix. Nous nous sommes donc focalisés sur des caractéristiques acoustiques et des caractéristiques en relation avec le conduit vocal et plus précisément les différentes configurations de celui-ci.

2.1 Phonétique, extraction MFCC

Nous souhaitons obtenir une représentation de notre signal qui soit moins volumineuse, qui s'intéresse à la phonétique du signal et qui, compte tenu du fait que notre système sera appliqué dans des contextes divers, soit invariante aux légers changements. Comme expliqué dans l'article Speaker identification using Mel Cepstral coefficients [3], la représentation paramétrique par extraction des MFCC est une des plus répandues et elle est moins sensible aux variations que l'étude de la forme du signal.
La représentation par MFCC est basée sur les capacités auditives de l'être humain. En effet, on retrouve cette idée de bancs de filtres linéairement espacés en dessous de 1kHZ, puis logarithmiquement espacés au dessus, au lieu

d'une identification isolée des fréquences. Les filtres appliqués sont généralement triangulaires. L'unité de mesure utilisée est le Mel qui correspond a la tonalité perçue par l'oreille humaine d'une certaine fréquence. La correspondance est donnée par l'équation suivante [4] :

$$F_{Mel} = \frac{1 \times 10^3}{\log(2)} \left[1 + \frac{F_{Hz}}{1 \times 10^3}\right]. \quad (1)$$

Nous segmentons notre signal grâce à un fenêtrage ... et nous obtenons de segments de ...ms. Nous appliquons alors la représentation en MFCC à notre signal en prenant k coefficients, ce qui nous donne un vecteur acoustique [3] de taille k qui caractérise notre signal.

2.2 Énergie et fréquence

Comme nous le verrons dans la section où l'on présente les résultats, le vecteur acoustique obtenu après extraction des coefficients MFCC n'est pas suffisant pour obtenir les performances souhaitées. C'est pour cette raison que nous allons extraire, en plus, à chaque segment l'énergie et la fréquence fondamentale. Ces deux caractéristiques viennent s'ajouter aux k coefficients de la MFCC de sorte à former un vecteur qui caractérise chaque segment, de dimension $k + 2$.

3 Algorithme de comparaison

L'approche la plus répandue dans la littérature pour la reconnaissance vocale est la combinaison de la MFCC avec le DTW (*Dynamic Time Warping*) [5]. Cette technique de calcul de distances est utile lorsque l'on compare le même mot où la même phrase, mais que l'on veut être invariants aux différents temps de production de ce signal. Malheureusement, dans notre cas, nous n'avons pas le même contenu sémantique dans les signaux que nous comparons, c'est pourquoi il nous a fallu trouver d'autres méthodes.

Comme nous voulons comparer des imitateurs et d'autres personnes à Trump, nous utilisons une mesure de distance pour trouver la personne qui se rapproche le plus de Trump. Nous avons donc modélisé Trump par les caractéristiques extraites des exemples de la base d'apprentissage. Nous avons utilisé plusieurs calculs de distance différents pour comparer Trump et les imitateurs. D'abord nous avons utilisé une distance euclidienne. Cette distance ne prend pas en compte la variance des différents exemples. À cause de cela, nous avons cherché à utiliser une distance qui prend en compte la variance des exemples.

Nous avons utilisé en premier lieu la distance de Mahalanobis qui prend en compte la variance et la corrélation des données. Elle donne une mesure de la similarité entre deux séries de données. Nous avons également utilisé la distance de Bhattacharyya qui est souvent utilisée dans le domaine de la vision par ordinateur. Elle estime également la similarité entre deux séries de données. Ayant des résultats légèrement meilleurs avec la distance de Bhattacharyya nous avons décidé d'utiliser cette dernière plutôt que la distance de Mahalanobis.

4 Résultats

L'obtention des résultats obtenus a été possible grâce à un processus progressif d'amélioration, à travers lequel nous avons découvert des éléments fort intéressants. C'est pour cette raison que nous avons décidé de ne pas nous contenter uniquement des résultats finaux, mais de détailler la logique suivie.

4.1 Une approche uniquement phonétique

Dans un premier temps, étant partis sur une prise en compte principalement de la phonétique, nous avons pris uniquement les coefficients MFCC. La première étape de notre analyse consiste alors à évaluer la cohérence globale de notre méthode. Pour cela, nous vérifions que la distance obtenue lorsque nous testons un extrait de Trump est bien minimale par rapport aux autres. Il suffit alors de

comparer les extraits de Trump avec les possibles voix derrière les *Deep Fakes*, i.e. celles des imitateurs qui, logiquement, devraient être les plus proches.
Grâce à cette première expérience, nous souhaitons également vérifier cette première hypothèse consistant à dire que les extraits des imitateurs vont forcément être plus proches que tous les autres. Pour cela, nous avons inclus également des extraits d'hommes politiques américains autres que Trump. Les résultats sont présentés sur la figure 1.

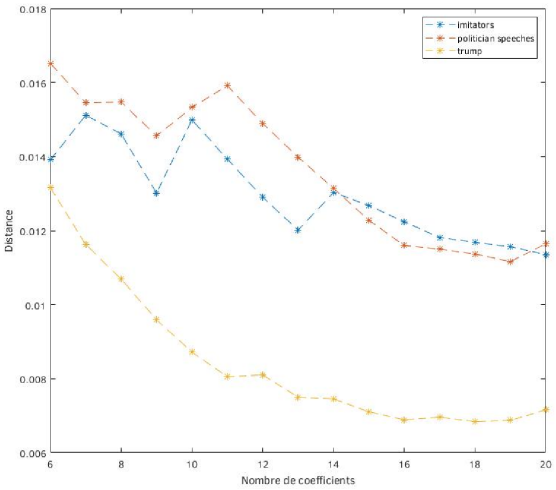

FIGURE 1 – Évolution de la distance de Bhattacharyya en fonction du nombre de coefficients pris en compte pour la MFCC, ceci pour les extraits de Trump, des imitateurs de Trump et de discours d'hommes politiques américains dans la base de validation.

Comme le montrent les résultats ci-dessus, la méthode adoptée est cohérente, nous avons bien une distance qui est minimale lorsque l'on compare des extraits de Trump avec les imitateurs. Néanmoins, notre hypothèse supposant que les extraits des imitateurs vont forcément être plus proches que tous les autres n'est pas vérifiée. En effet, comme on peut le voir sur la figure 1, pour un nombre de coefficients supérieur à 14, les extraits d'hommes politiques américains sont plus proches que les imitateurs.
La réfutation de cette hypothèse nous pousse alors à inclure les autres types d'extraits que nous avons dans notre base afin d'êtres sûrs que la cohérence que nous cherchons est bien respectée.

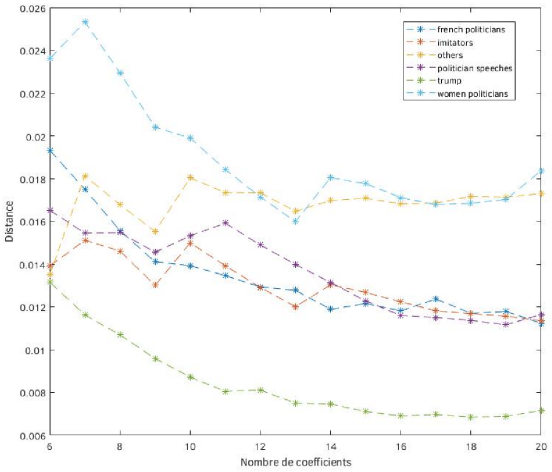

FIGURE 2 – Évolution de la distance de Bhattacharyya en fonction du nombre de coefficients pris en compte pour la MFCC, ceci pour tous les extraits dans la base de validation.

Les résultats obtenus sur la figure 2 nous permettent de conclure sur la cohérence de notre méthode. En effet, comme attendu, pour un nombre de coefficients MFCC supérieur à six, nous avons toujours une distance plus petite lorsque nous comparons Trump avec lui même.
Mais ces résultats révèlent un élément très intéressant, en prenant un nombre suffisamment élevé de coefficients MFCC, nous voyons apparaître trois groupes distincts : un premier composé uniquement d'extraits de Trump, un deuxième composé d'extraits d'hommes politiques (anglophones et francophones) et des imitateurs, puis un troisième composé d'extraits de femmes politiques et des personnes autres.
La première conclusion que nous pouvons tirer de cette tendance est que nous réussissons bien à ne prendre en compte que les dimensions phonétiques et inférieures. En effet, nous obtenons des distances très similaires entre les hommes politiques anglophones et francophones, deux langues proches phonéti-

quement, alors que le contenu sémantique est très différent. La deuxième observation que nous pouvons faire est que la similarité obtenue entre hommes politiques peut-être expliquée par le fait que les discours politiques respectent un certain nombre de codes, notamment sur le rythme et les intonations, qui sont indépendants du pays, du moins entre les États-Unis et la France. Finalement, nous observons que, même si dans le groupe où l'on a des voix de femmes nous prenons des discours politiques, les résultats restent éloignés de ceux des hommes politiques. Cette différence nous indique que, même si notre système est sensible au contexte (distance faible entre discours politiques), les caractéristiques du conduit vocale sont également prises en compte, permettant de séparer les femmes et les hommes dans un même contexte.
Comme nous venons de le voir, les tendances observées sur la figure 2 sont intéressantes car très révélatrices. Néanmoins, afin de construire un système plus robuste, nous allons enrichir notre système de sorte à creuser la différence entre les différents groupes.

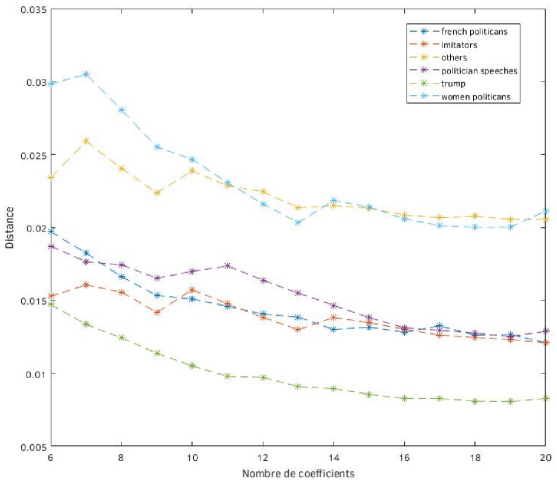

FIGURE 3 – Évolution de la distance de Bhattacharyya en fonction du nombre de coefficients pris en compte pour la MFCC et en prenant en compte la fréquence fondamentale et l'énergie, ceci pour tous les extraits dans la base de validation.

4.2 Enrichissement et classement

Nous allons à présent essayer de creuser l'écart entre les groupes distincts que nous avons vu apparaître dans les résultats précédents. Pour cela, nous allons enrichir la caractérisation de nos exemples en ajoutant l'information la plus bas niveau possible, i.e. l'acoustique. Comme mentionné dans la section dédiée à l'extraction des caractéristiques, nous prenons en compte la fréquence fondamentale et l'énergie du signal de façon brute, sans traiter cette information statistiquement. La figure 3 présente les résultats obtenus avec cette nouvelle caractérisation.

Comme nous pouvons le voir, l'ajout de ces paramètres acoustiques nous permet bien de séparer les trois groupes plus distinctement. Dans notre cas particulier, l'utilisation de sept coefficients pour la MFCC est la configuration idéale, les deuxièmes extraits les plus proches sont les imitateurs et l'on vient nos trois groupes distinctement.
Maintenant que nous avons choisi les paramètres les plus adéquats pour notre contexte, nous allons appliquer notre algorithme sur notre base test de sorte à établir un classement du meilleur imitateur de Trump.

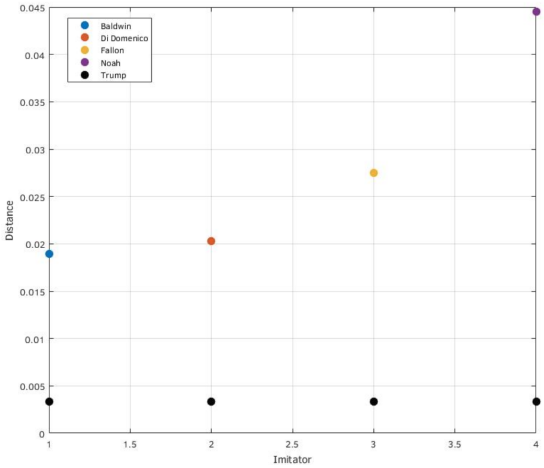

FIGURE 4 – Classement des imitateurs de Trump obtenu en prenant en compte la fréquence fondamentale, l'énergie et sept paramètres pour la MFCC.

Comme le montre la figure 4, nous retrouvons en dernière position le présentateur Trevor Noah juste après Jimmy Fallon. Puis, dans la première position, nous retrouvons Alec Baldwin suivi de Marc di Domenico.
Nous avons mis en place un système nous permettant d'établir un classement des meilleurs imitateurs vocaux de Trump. Néanmoins, avant que ce système soit utile pour la reconnaissance de *Deep Fakes*, il est nécessaire de mettre en place une métrique permettant de convertir notre distance en une décision indiquant si il s'agit de Trump ou non. Pour cela, nous avons mis en place un protocole basée sur les performances humaines. Celui-ci consiste à retirer l'information sémantique des extraits en inversant l'ordre du signal lors de l'écoute, puis en demandant à différents individus d'indiquer si il s'agit de Trump ou pas.

Références

[1] B. S. Kisilevsky, S. M. Hains, K. Lee, X. Xie, H. Huang, H. H. Ye, K. Zhang, and Z. Wang, "Effects of Experience on Fetal Voice Recognition," Psychological Science, vol. 14, no. 3, pp. 220–224, May 2003. [Online]. Available : https://doi.org/10.1111/1467-9280.02435

[2] J. M. Zarate, X. Tian, K. J. P. Woods, and D. Poeppel, "Multiple levels of linguistic and paralinguistic features contribute to voice recognition," Scientific Reports, vol. 5, no. 1, pp. 1–9, Jun. 2015. [Online]. Available : https://www.nature.com/articles/srep11475

[3] R. Hasan, M. Jamil, G. Rabbani, and S. Rahman, "Speaker identification using MEL frequency cepstral coefficients," in ICECE. Dhaka, Bangladesh : International Conference on Electrical & Computer Engineering, Dec. 2004.

[4] J. R. D. Jr, J. H. L. Hansen, and J. G. Proakis, Discrete–Time Processing of Speech Signals. New York : Wiley-Blackwell, Sep. 1999.

[5] L. Muda, M. Begam, and I. Elamvazuthi, "Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques," arXiv :1003.4083 [cs], Mar. 2010, arXiv : 1003.4083. [Online]. Available : http://arxiv.org/abs/1003.4083