Acoustic Pairing of Original and Dubbed Voices in the Context of Video Game Localization

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

Acoustic Pairing of Original and Dubbed Voices in the Context of Video Game Localization

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

LIA - University of Avignon (France)

firstname.lastname@univ-avignon.fr

Abstract

The aim of this research work is the development of an automatic voice recommendation system for assisted voice casting. In this article, we propose preliminary work on acoustic pairing of original and dubbed voices. The voice segments are taken from a video game released in two different languages. The paired voice segments come from different languages but belong to the same video game character. Our wish is to exploit the relationship between a set of paired segments in order to model the perceptual aspects of a given character depending on the target language. We use a state-of-the-art approach in speaker recognition (*i.e.* based on the paradigm i-vector/PLDA). We first evaluate pairs of i-vectors using two different acoustic spaces, one for each of the targeted languages. Secondly, we perform a transformation in order to project the source-language i-vector into the target language. The results showed that this latest approach is able to improve significantly the accuracy. Finally, we challenge the system ability to model the latent information that holds the video-game character independently of the speaker, the linguistic content and the language.

Index Terms: voice casting, voice similarity, speaker recognition, i-vector, video game.

1. Introduction

Voice casting aims at selecting the most appropriate voice to dub a character [1]. In a multilingual context, this voice selection process does not simply rely on the character but also on the voice of the original actor. Voice casting is mainly performed by a human operator and raises two main problems. First, the process is very sensitive to the operator's subjectivity. Second, the operator cannot take a decision based on all available voices, since their number is usually too large. Above all, because of his expertise and his past, he is unconsciously encouraged to consider a small subset of voices with which he is used to work.

We focus our work on the domain of voice dubbing from a source language, the dubbing being the targeted language (for example, dubbing in French from an English voice), in the context of video games. Indeed, video games are now not simply a playful experience but also a medium offering a close immersion, akin to movies [2]. One of the critical aspects for these scripted games is the voice attribution process for the different characters. Since it directly impacts the gaming experience by influencing the player's immersion, video game producers want their characters to have a voice that mostly respects its specificities, like role, character's traits, attitudes, etc. This importance translates to game localization, the process of dubbing the audio tracks from a source to a target language. The dubbed voice is not just a simple transposition of the original voice to the target one, but more based on the speaker's ability to transpose his voice in order to act like the original one. In addition, the human experts have to take into account more sophisticated aspects, like cultural or social values related to the considered language. They also have to take care of the mental representations of voices, stereotypes for short. As an illustration, a soldier could be stereotypically described as having a slightly hoarse and screamed voice.

Our final goal is to approximate the human operator expertise through the definition of appropriate statistical models. In such an automatic recommendation system for assisted voice casting, different sources of variability should be considered. The first concerns the speaker, which can be easily confused with the second: the character. One of our objectives is to build a system where these two dimensions can be observed independently. In other words, when considering a single speaker, we want to be able to identify all the different characters he can play. Note that a one-to-one association between actors and characters would just lead to a voice similarity system –close to a speaker recognition system– working on multiple languages. This brings us to the third source of variability, the language, which we want to decrease together with the linguistic content of voice segments. We propose a preliminary work to automatically pair original and dubbed voices using character voice segments. We extend an i-vector based automatic speaker recognition approach to the characterization of higher-level informations representing acted voices similarity. Given the fact that speakers use different languages, our proposed contribution for video game voice localization aims at reducing the language variability in order to better extract acoustic information relying on characters. To do so, we propose to perform a transformation in order to project the dubbed segments to the original ones.

The rest of this paper is organized as follows. Firstly, we present related work about voice casting and perceptual voice similarity in Section 2. We then describe the proposed acoustic pairing approach in Section 3. The experimental protocol as well as the experiments and discussions about our system performance are presented in Sections 4 and 5 respectively. Finally, Section 6 concludes and gives some perspectives.

2. Related work

Due to its subjective nature, automatic voice perception is a very difficult problem. The first step of this process is to handle how humans perceive voices, and the literature abounds of such work, especially in the sociology and psychology domains [3, 4, 5]. They show different factors that impact judgment and perception, not directly for voice perceptual similarity but for perception of speaker, state, traits, or even emotions. These factors are voice indicators (e.g. pitch, intensity, ...) and speech cues (e.g. non-fluency, speech rate, ...). In addition, these work present particular acoustic features that could clearly help distinguishing personality traits. [6] showed that percep-

tual voice similarity is strongly correlated with particular voice quality settings. However, this work only considers speakers with the same accents. Others work have also been conducted on the automatic attribution of acoustic features to personality traits [7]. The research on voice perception has focused much more on the automation of personality trait attribution than on voice similarity. Work on personality traits are strongly related to our own work, because video game characters also have personality traits that must be supported by their voice. In addition, these aspects have to be preserved in the dubbed version.

There are few papers relating to voice casting, and the problem has recently started gaining interest. In [8], the authors increase the users satisfaction of computer-generated characters through their instant casting movie system, called *Future Cast*, by using perceptual similar voices selection. They estimate the voice similarity by performing a weighted sum of acoustic distances between cepstral features computed with DTW. More recent work [1], [9] dive deeply into the automatic voice casting problem, by exploring different models for perceptual voice similarity search. The authors compare GMM-based acoustic models and multi-label classification of paralinguistic features, with an objective experiment in order to measure the perceptual similarity of acted voices. In addition, they carry out a subjective experimentation by presenting to different human evaluators the 3 most similar voice samples according to both GMM-based acoustic models and multi-label recognition systems. Their results show an increased performance with respect to paralinguistic features recognition instead of classical speaker recognition systems. However, they finally conclude by considering the use of an acoustic model that would be more specific to voice casting than the one used for speaker recognition.

To our knowledge, there is no other scientific work that propose to look for acoustic representations of acted voices and even less between different languages. We propose an original approach in the form of a preliminary study that seeks to extend speaker recognition systems to the purpose of voice casting.

3. Proposed approach

In this paper, we focus on video game localization, because of the strongly typed voices present in this kind of multimedia products. Video games editors largely make use of stereotyped voices for their characters. In this approach, we explore the relation between original and dubbed voices by performing character verification given two different voice excerpts.

3.1. General system presentation

The desired system is coarsely represented in Figure 1, where two audio files represent the main inputs. The first one is referred to as the *original* audio segment, and the second one as the *dubbed* audio segment. As we want to compare voice similarity, our desired system outputs scores corresponding to the proximity of the two voice segments. The voice similarity system is voluntarily presented as a black box, as its composition will be discussed here. In this preliminary approach, we chose to measure the voice similarity using off-the-shelf i-vector/PLDA components. We build three variants of our system, as illustrated in Figure 2. The different approaches are denoted *System A* (details in Section 3.3), and *Systems B* and *C* (details in Section 3.4). All of them consider two voice segments referred to by *Src* and *Tgt*, which correspond to the source (*i.e.* original) and target (*i.e.* dubbed) voice segments, respectively. In addition, all our systems consider different language configurations. An i-vector is extracted from an audio segment thanks to a total variability matrix (the T matrix). We use different T-matrix, adapted to the two considered languages (English and French). We thus have three language configurations for i-vector extraction: EN-EN, EN-FR and FR-FR. The first part of the configuration name refers to the language used to represent the source voice segments, and the second part to the dubbed voice segments.

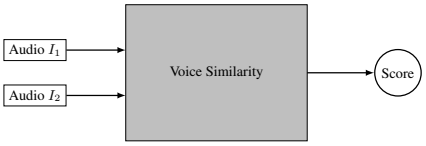

Figure 1: *Similarity measure between two audio files is performed by our system.*

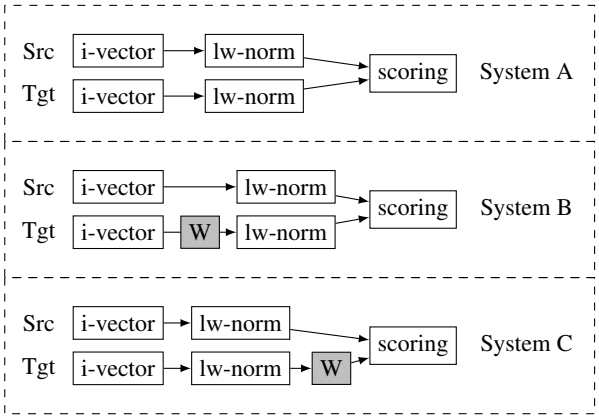

Figure 2: *Three proposed approaches for the comparison of i-vectors. In systems B and C, a projection matrix W trained from different languages is used for comparing i-vectors.*

3.2. An i-vector/PLDA based approach

The voice segments are represented by i-vectors and the similarity between two i-vectors is estimated by a PLDA approach (a probabilistic version of Linear Discriminant Analysis [10]), thus the obtained score is a likelihood ratio. An i-vector is a low-dimensional feature that characterizes speakers [11]. It is extracted from a variable length sequence of acoustic parameters representing a speech segment. In the i-vectors space, no distinction is made between the variability sources.

3.3. Baseline i-vector based similarity

Our approach considers a baseline i-vector system using a language-dependent T-matrix to extract the i-vector. So, each T-matrix has been trained on a corpus corresponding to a single language. Since our problem involves two languages, we are distinguishing two cases. In the first case, paired voice segments are represented by i-vectors extracted using a T-matrix trained for English (*i.e.* T matrix learned on an English corpus). In the second case, both original and dubbed voice segments are represented by i-vectors extracted for the French language (*i.e.* T matrix learned on a French corpus). This approach is referred to as *System A* in Figure 2.

3.4. Dubbed adapted i-vector comparison system

Since this work is dedicated to source/target-language paired audio segments, we have to deal with two languages in each voice similarity measure. We need to extract the i-vector for each audio file using the corresponding language-dependent T-matrix. So original voice segments are represented by i-vectors with a total variability matrix learned on English data, and dubbed segments are represented by i-vectors learned on French data. Thus, an adaptation is needed in order to allow a comparison between these two i-vectors using PLDA.

The heart of our contribution is dedicated to this language adaptation. We consider projecting i-vectors from a language to an other in order to perform comparison and scoring between them. The approach is inspired from [12], where authors carry out word translation, and has been adapted to our problem. The projection is learned on a set of segment pairs that are represented by $\{x_m,\ y_m\}$, where x is a voice segment i-vector in the source language, and y is its counterpart in the target language. Here, both segment x and y convey the same message m. We want to find a transformation function such that $x_m \approx L(y_m)$. We use the least squares method [13] to solve this linear equation, by computing the transformation matrix W that minimizes $\|x_m - y_m W\|^2$. This can be seen as a linear mapping between two different spaces: source and target languages i-vector spaces. Here, the i-vectors are projected into the target language, but the inverse is also possible, depending on how we see the task. We perform the projection in *System B* and *System C* and it refers to W in Figure 2. In the same way as for *System A*, we use i-vectors extracted from different languages and we use the three configurations that have been described previously: EN-EN, EN-FR and FR-FR.

4. Experimental Protocol

4.1. Corpus Description

Our corpus has been built by extracting dialogs from the role playing video game *Mass Effect 3*. It was originally released in English, but it has also been translated into several others languages such as French. We extracted the audio files composing both the English and French versions of the dialogs. Each language dialog subsets contains 10,000 audio segments that are all hight-quality studio recorded audio files. A segment generally refers to a sentence. Our corpus includes 38 different characters dubbed by at least 30 different actors. We consider a set of characters all belonging to the video game, denoted $\mathcal{P}$. Since two different actors act as the voice for the character k, we represent the character as a pair $\mathcal{P}_k = (X_k, Y_k)$ where X_k is the set of its original voice segments and Y_k its dubbed ones. We also dispose of all the different pairs of possible combinations of utterances as a set $E = \{(x_{l,m,i,k}, y_{l,m,i,k})\}$, which constitutes our corpus, where x is the English segment i-vector and y the French one. Here, variable m refers to the linguistic content (*i.e.* the message), i to the speaker, and k to the character. Finally, variable l denotes the language used to extract the i-vector. We finally split the corpus into 3 subsets with 70%, 10% and 20% for train, dev, and test respectively.

The corpus approximately contains a total of 7.5 hours of speech, for each language. The different segments are 3.5 seconds long, in average. The amount of speech by character is related to the importance of the character in the game, and it is thus not uniformly distributed: its mean is 12 minutes and its standard deviation is 20 minutes.

4.2. Voice similarity estimation

Our experiments operate on 20 MFCC parameters (including log-energy) augmented with 20 first (Δ) and 20 second ($\Delta\Delta$) derivatives, providing 60 dimensional feature vectors. We apply a cepstral mean normalization using a sliding window size of 3 seconds. We remove the low-energy frame, which corresponds mainly to silence. The low-energy algorithm is based on thresholding the log-energy and taking the consensus of threshold decisions within a window of 11 frames centered on the current frame. Finally, we train a Gender-dependent 2048 full component UBM and total variability matrix of low rank 400.

We set up two speaker recognition systems: English and French. For the English system, we train the UBM, total variability matrix and PLDA on NIST SRE 2004, 2005 and 2006. For the French one, we train them on ESTER-1, ESTER-2, EPAC, ETAPE and REPERE.

4.3. Evaluation Metric

Our metric evaluates the system capacity to detect paired segments belonging to a same character. We chose to measure this capacity by computing the overall accuracy of the system given all the pairs of segments being scored. The accuracy is computed as a ratio between the number of correctly classified pairs and the total number of classified pairs. We chose to measure the correctness of a test by considering the k-best scores of the system among all possible pairs $\{(x_{i,n}, y_{j,m})\}$. We put 2 constraints for the test validation. The first is that $i = j,\ \forall n, m$, which means both paired segments belong to the same character. The second is that the considered pairs must have been ranked among the k-best scores of the system for the set of pairs $\{(x_{i,n}, Y)\}$, where $x_{i,n}$ is the segment n of character i, and Y refers to every segment from every character of the target language.

5. Results & Discussion

The results of voice similarity measures are presented in Table 1. The evaluation has been performed using a 3-fold cross-validation on the overall corpus. To better analyze the systems performance, accuracies are computed with respect to the 3-best scores (*i.e.* pairing is considered as correct if the correct dubbed voice has been retrieved in the 3-best hypothesis).

Table 1: *Results of the different approaches (3-best accuracy).*

System	EN-EN	EN-FR	FR-FR
A	60.9	53.9	60.7
B	71.7	69.6	69.4
C	70.7	68.9	69.6

First of all, the baseline i-vector system (*System A*) is mainly evaluated given 2 different language configurations: EN-EN and FR-FR. Indeed, the EN-FR configuration, which obtained 53.9% for information, has no real interest in this approach since the comparison of i-vectors extracted from different languages is not relevant (no transformation is performed). As we can see, the first results show a reasonable accuracy for voice similarity detection, with an accuracy of 60.9% for EN-EN and 60.7% for FR-FR.

We now compare the results of *System A* with those of *Systems B* and *C*. Our first observation is that the performance is higher than the baseline system *A* for both *Systems B* and *C*.

The best results are obtained on the language configuration EN-EN, with accuracies of 71.7% and 70.7% for *System B* and *C*, respectively; whereas the language configuration EN-FR leads to a 69.6% accuracy for *System B* and 68.9% for *System C*. For the configuration FR-FR, the accuracy is slightly lower, with 69.4% for *System B* and 69.6% for *System C*. To get a global idea of the measured performance, Figure 3 presents the accuracies obtained by each system when varying the k-best number using the best configuration (EN-EN). In comparison to Table 1, similar tendencies can be observed no matter the k value chosen.

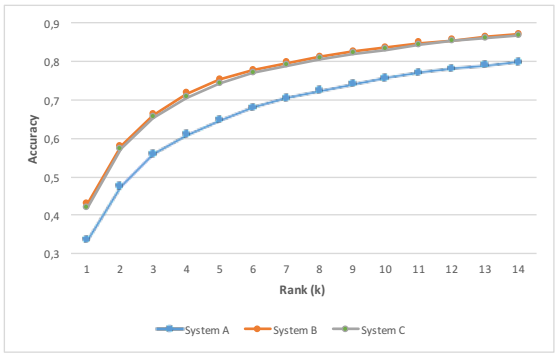

Figure 3: *Accuracy on EN-EN configuration using different values of k.*

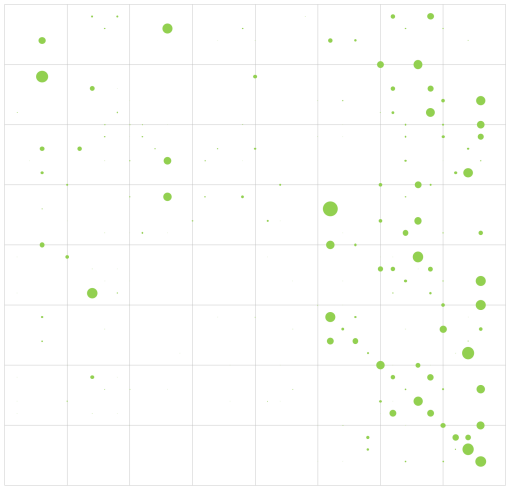

Figure 4: *Characters confusion matrix using System B with EN-EN configuration (k=1).*

We observe that in *Systems B* and *C*, there is no increased representational capacity when using English i-vectors to represent original voice segments and French i-vectors for dubbed ones. However, according to our results, the i-vector projection significantly increases the system performance. It seems that the projection operates like a speaker mapping system. The W matrix has learned to map speakers in the source language to ones in the target language. A possible explanation for the better results obtained with the EN-EN configuration is that 1) the comparison is performed in the same language i-vector space, and 2) it might benefit from the speaker mapping learned by the projection matrix. Since this projection is just an approximation of $x_m = L(y_m)$, it could explain why the system performances with the FR-FR configuration are not as good as with the EN-EN configuration. Finally, Figure 4 shows the overall confusion matrix obtained for the characters using *System B* and the EN-EN configuration for $k = 1$ (global accuracy of 43%). We observe good performances on the diagonal in the bottom-right corner, that corresponds to the most represented characters. More important, this figure displays some errors which can be explained by the different numbers between actors and characters in the corpus. Indeed, we know some actor can lend his voice to several characters, and the matrix shows that some of the errors are associated to the different characters acted by a same voice actor. For example, the same actor plays one character at two different ages (new and old), and is considered as two different characters by the system. This could also be the reason why good results are observed in Table 1 (correct character being retrieved in the 3-best hypothesis). Finally, it reveals that this proposed preliminary work on voice similarity does not allow to completely answer the global problematic, namely make the different dimensions independent, but is a first step in the automatic voice casting problem.

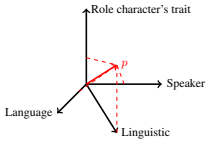

Figure 5: *Voice casting variability dimensions. Here p represents a character in the variability space.*

6. Conclusion & Perspectives

In this paper, we present an automatic pairing system of original and dubbed voice segments for the detection of acted voice similarity. The proposed approach explores i-vectors-based systems for automatic voices comparison between different languages and use projection matrix to reduce the language variability. Experiment have showed that a mapping between speakers can be learned from a source language to speakers from a target language, since the projection approach has significantly increased the accuracy of the system. These preliminary results for acted voice similarity is a first step toward a dedicated framework for automatic voice recommendation.

These initial encouraging results, however, highlighted some weaknesses in the proposed approach. First, the system could be improved by using similar acoustic data for training, the corpus used by now being far from the targeted data. In further work, we will explore acoustic features representation for character verification. In order to do so, we will investigate the use of other speech feature representations. As an illustration, Figure 5 shows different variability dimensions of our problematic. According to our experimental results, the approach presented here suffers from an ambiguity between the speaker and the character axes. Further researches will perform a contrastive experiment to observe how the system reacts. Finally, we will consider the use of classes of characters in order to reduce the variability of this dimension.

7. Acknowledgements

This work is supported by the Foundation of the University of Avignon.

8. References

[1] N. Obin, A. Roebel, and G. Bachman, "On automatic voice casting for expressive speech: Speaker recognition vs. speech classification," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, pp. 950–954, 2014.

[2] C. Bateman, *Game Writing: Narrative Skills for Videogames*, ser. Applied English Series. Charles River Media, 2007. [Online]. Available: https://books.google.fr/books?id=TJgdAQAAIAAJ

[3] M. Zuckerman and R. E. Driver, "What sounds beautiful is good: The vocal attractiveness stereotype," *Journal of Nonverbal Behavior*, vol. 13, no. 2, pp. 67–82, 1989.

[4] K. Scherer, "Vocal communication of emotion: A review of research paradigms," *Speech Communication*, vol. 40, no. 1-2, pp. 227–256, 2003.

[5] R. M. Krauss, R. Freyberg, and E. Morsella, "Inferring speakers' physical attributes from their voices," *Journal of Experimental Social Psychology*, vol. 38, pp. 618–625, 2002.

[6] F. Nolan, P. French, K. Mcdougall, L. Stevens, and T. Hudson, "The role of voice quality 'settings' in perceived voice similarity," 2011.

[7] G. Mohammadi and A. Vinciarelli, "Automatic Attribution of Personality Traits Based on Prosodic Features," *ACII 2015 Affective Computing and Intelligent Interaction*, vol. 3, pp. 29–32, 2015.

[8] Yoshihiro Adachi, Shinichi Kawamoto, Shigeo Morishima, and Satoshi Nakamura, "Perceptual similarity measurement of speech by combination of acoustic features," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2008, pp. 4861–4864.

[9] N. Obin and A. Roebel, "Similarity search of acted voices for automatic voice casting," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 9, pp. 1642–1651, 2016.

[10] S. J. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*. IEEE, 2007, pp. 1–8.

[11] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[12] T. Mikolov, Q. V. Le, and I. Sutskever, "Exploiting Similarities among Languages for Machine Translation," *arXiv preprint arXiv:1309.4168v1*, pp. 1–10, 2013.

[13] A. M. Legendre, *Nouvelles méthodes pour la détermination des orbites des comètes*. F. Didot, 1805, no. 1.