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
{adrien.gresse, mickael.rouvier, richard.dufour, vincent.labatut, jean-francois.bonastre}@univ-avignon.fr
LIA - University of Avignon (France)

Context & Motivation:

Context

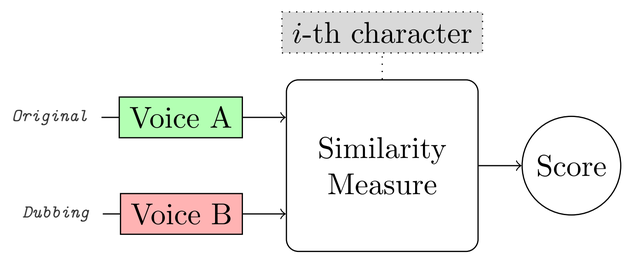
- In a multilingual context, the process of replacing original speech content (i.e. English) by the target language (i.e. French) is referred as dubbing.
- Voice casting aims at selecting an actor that mostly respects the original voice. It is performed by a human operator which raises two difficulties: 1. operator's subjectivity; 2. huge amount of available voices.

Motivation

Can we build a system that measures the similarity between a voice coming from a target language and the source one considering human perception and subjectivity ?

Approach & Protocol:

General presentation



Experimental protocol

- French and English voices segments from the Mass-Effect video game dialogs.
  - 10,000 segments per language.
  - 38 different characters.
- 3 sets: train, dev and test (70%, 10%, 20%).
- Cross-validation using 3-folds.

- MFCC: 19 coefficients + energy + Δ + $\Delta\Delta$.
- Gender-dependent GMM-UBM of 2048 components.
- T-matrix of low rank 400.

UBM, T-matrix and PLDA trained on:
- English: NIST SRE 2004, 2005 and 2006.
- French: ESTER-1/2, EPAC, ETAPE, REPERE.

An i-vector/PLDA based approach

- Voices segments represented by i-vectors (low-dimensional representation of acoustic parameters).
- Similarity between paired voices estimated with PLDA scores (likelyhood ratio).

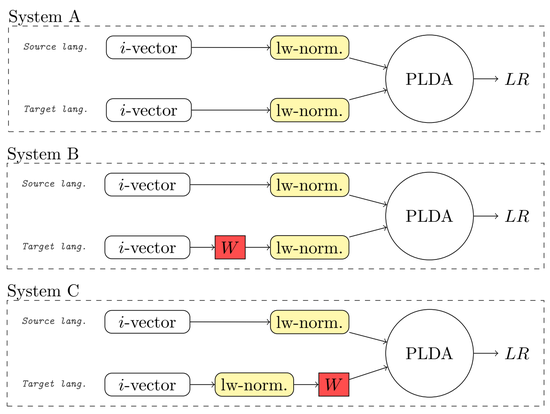


Fig. 2: Three i-vectors comparison approaches. In systems B and C, a projection matrix W trained from different languages is used for comparison.

Baseline i-vector based similarity:

Original and dubbed i-vectors extracted from the same total variability space (*System A*).

1. T-matrix trained on English corpus (EN-EN).
2. T-matrix trained on French corpus (FR-FR).

Dubbed adapted i-vector comparison:

- Use language-dependent total variability space.
- Projection of i-vectors from target language to source language.
- Matrix denoted W trained on a set of pairs $\{(x_i, y_i)\}$ where x_i is an i-vector representing a voice segment in target language and y_i its counterpart in the source language.
- Train W by minimizing $\sum_{i=1}^{n} \left\| x_i - y_i W \right\|^2$.

Three configurations for i-vectors extraction:

1. EN-EN
2. EN-FR
3. FR-FR

EN and FR refer to the language of the corpus used for T-matrix learning.

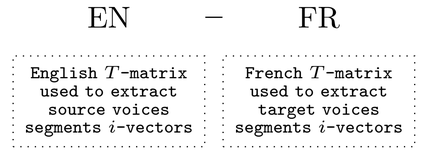


Fig. 3: Example using configuration #2

Evaluation metric

Evaluate the system capacity to detect pairs of segments from a same character:
- Computing the overall accuracy through ranked PLDA scores.
- Use a k-best approach.

$$Accuracy = \frac{\text{number of valid tests}}{\text{total number of tests}}$$

Forall character i among testing characters:
1. Score all pairs of segments where first segment belongs to character i.
2. Retrieve the k-best scored pairs.
3. Validate the test if the character of first segment is equal to the second in any of the k-best pairs.

Results & Conclusion:

Results

System	EN-EN	EN-FR	FR-FR
A	58.63	51.70	60.31
B	70.52	69.22	70.01
C	61.21	56.12	63.08

Table 1: Results of the different approaches (k=3).

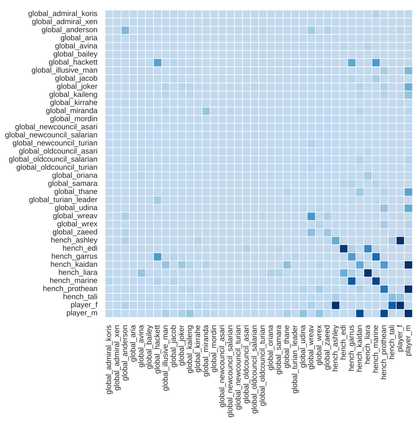


Fig. 4: Characters confusion matrix using System B with EN-EN configuration (k=1).

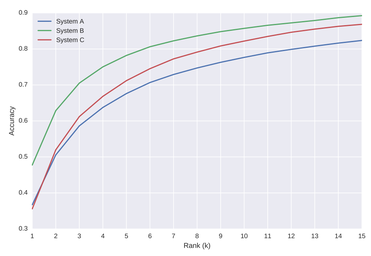


Fig. 5: Accuracy on EN-EN configuration with different values of k.

- A reasonable accuracy on *System A* (only EN-EN and FR-FR relevant).
- Highest accuracy observed on *System B* using projection with EN-EN configuration.
- Matrix shows good confusions on most representative characters (bottom right corner).
- Language-dependent T-matrix (EN-FR) do not increases accuracy surprisingly.

Conclusion

- Exploration of i-vector based systems for voices comparison in multilingual context and reduction of language variability.
- Increased accuracy using projection matrix shows the voice mapping relevance.
- A first step toward a dedicated framework for automatic voice recommendation.

Perspectives

- Explore acoustic features representations for character verification.
- Use classes of characters instead of single character in order to reduce ambiguity between speakers.
- Investigate the impact of other variability dimensions (e.g. linguistic content).
- Use a training corpus acoustically similar to targeted data.