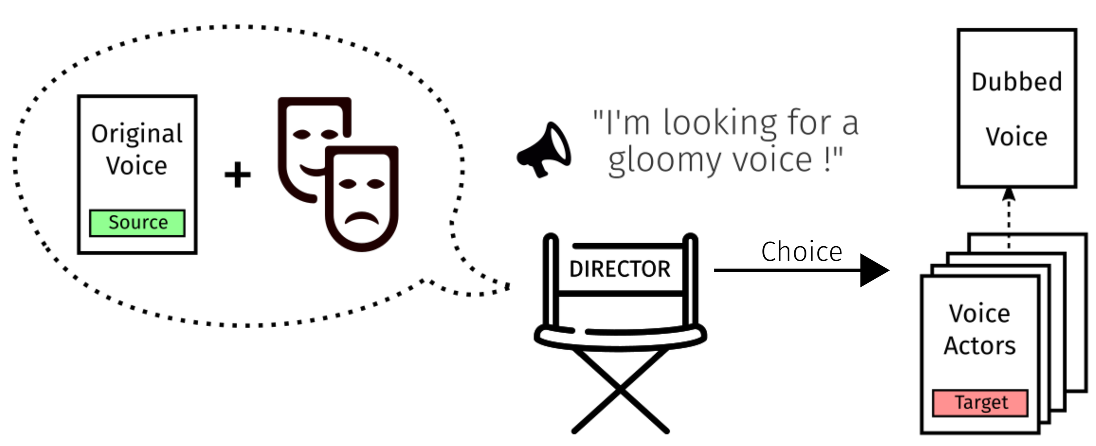
Similarity Metric Based on Siamese Neural Networks for Voice Casting

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

LIA, Avignon University

Context of application



- Replacing the original voice by a new voice in a target language is referred as *dubbing*.
- Voice casting is performed by a human operator and aims to find the most suited voice for the role.

Difficulties:

1. There is no formal description of voices.
2. Operator has too many voice-actors to cast.
3. The subjectivity of the choice.

Motivations

- **Approximate automatically the operator's choice to help him in future decisions.**
- Learn a multilingual similarity metric beyond the simple acoustic resemblance.
- Build a character/role dedicated representational space.

Proposed approach

We use pairwise relationship between two voices (original, dubbed) that share an abstract notion of similarity.

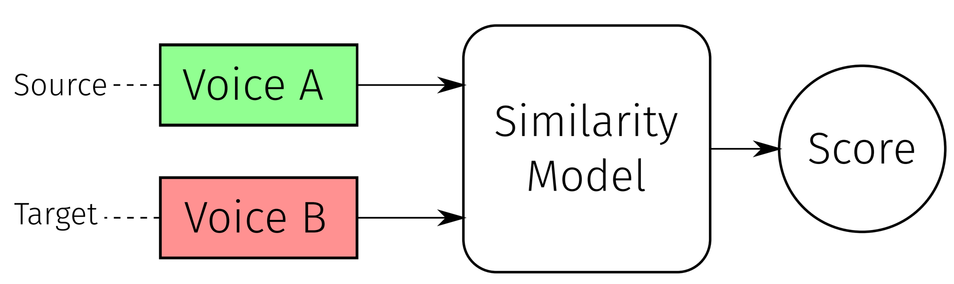


Figure 1: Voice A is in the source language, voice B in the target language. The score reflects the operator's similarity perception.

We train a binary-classifier using *Siamese Neural Networks* that learn to discriminate between *target* pairs (same character) and *non-target* pairs (different characters).

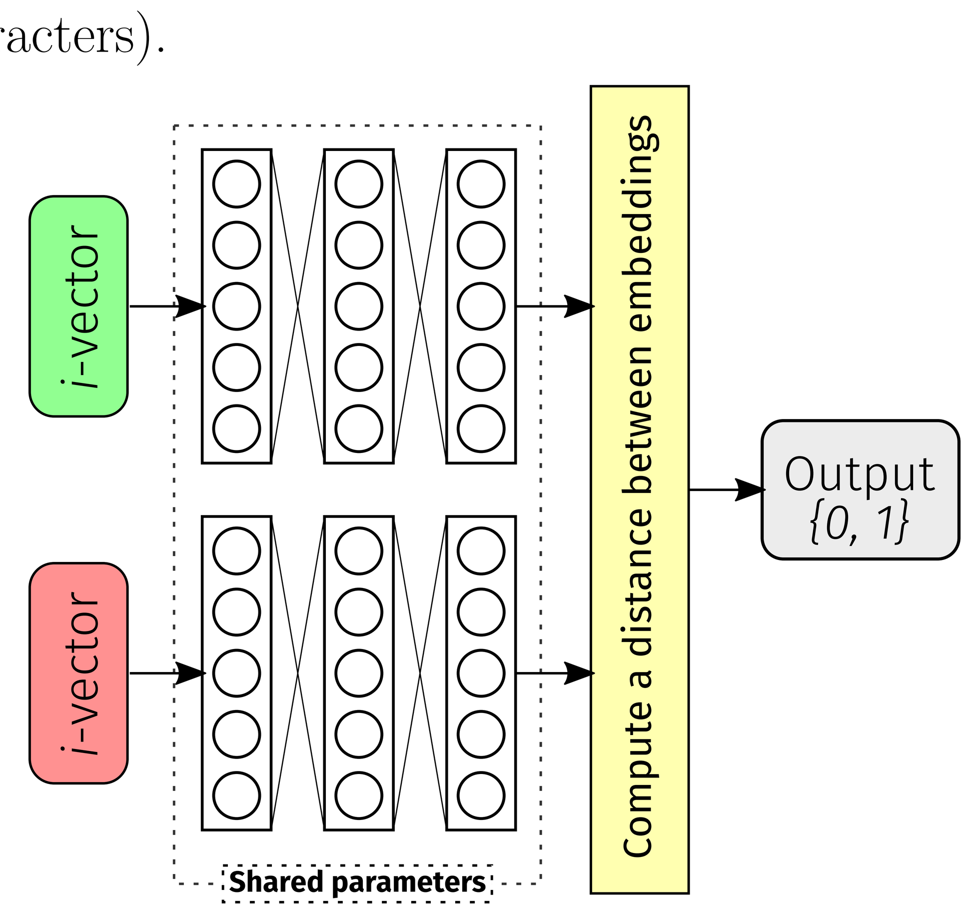


Figure 2: Siamese Neural Networks (SNN) involves two networks sharing same parameters allowing a comparison between independent inputs.

Experimental protocol

Corpus

- 16 characters from *Mass Effect* video-game.
- 180 voice segments per character.
- 2 different languages: *English* and *French*.

Sequences extraction (i-vector):

- 19 MFCCs + energy + Δ + $\Delta\Delta$ with CMS and VAD.
- Language-independent i-vector system.
- 2048-components UBM and T-matrix rank 400.

Table 1: We perform a 4-fold cross-validation (jackknifing) over the 16 characters of *Mass Effect* in order to tackle the dataset limitation. Each case contains 4 distinct characters.

Test	#pairs	Training	#pairs
A	$64,800$	$B+C+D$	$194,400$
B	$64,800$	$A+C+D$	$194,400$
C	$64,800$	$A+B+D$	$194,400$
D	$64,800$	$A+B+C$	$194,400$

Evaluation:

1. Performance of the binary classifier (accuracy).
2. *Target*/*non-target* pairs discrimination (t-test).

Results

Table 2: We compare accuracy and t-score of SNN with classic architectures.

	2in-conc		2in-merge		**siamese-net**	
	acc.	*tscore*	*acc.*	*tscore*	*acc.*	*tscore*
A (test)	0.49	~~0.71~~	0.52	17.66	**0.55**	52.18
B (test)	0.49	5.34	0.50	4.53	**0.59**	77.99
C (test)	0.51	7.82	0.53	18.37	**0.62**	86.17
D (test)	**0.53**	17.30	0.52	14.50	0.50	~~1.87~~
A (dev)	0.94	185.72	0.93	169.93	0.72	47.90
B (dev)	0.96	211.32	0.94	190.68	0.71	52.77
C (dev)	0.93	161.16	0.93	160.16	0.70	45.18
D (dev)	0.96	227.85	0.96	212.80	0.71	44.46

SNN generalize better on 3 out of 4 test cases while standard architectures seem to memorize couple of speakers.

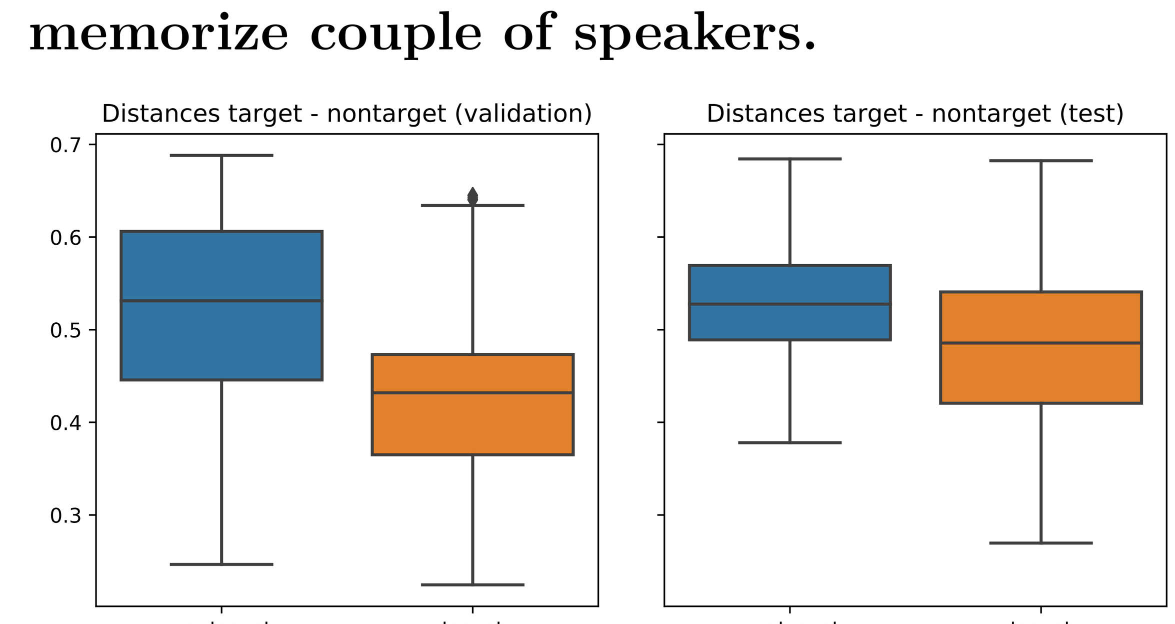


Figure 3: Target (blue) and non-target (orange) distances on case C for development (left) and test (right) with SNN.

Conclusion

- Results show that we are able to discriminate *target* and *non-target* pairs on unknown voices using siamese networks.
- We built a latent representational space emphasizing the information that reflects an abstract notion of similarity.

Limits:

- The dataset limitation.
- We do not discriminate the character himself.
- We suppose the existence of other bias.

Perspectives

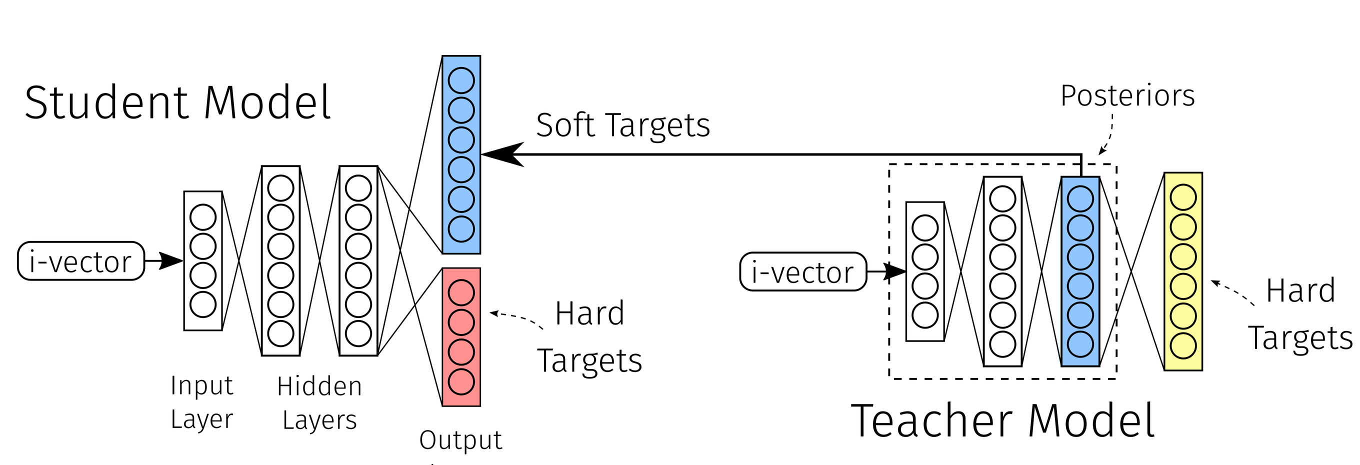


Figure 4: The teacher-student framework allows student model to learn from an intelligent teacher. We train the teacher on additional data (yellow). We use the soft-targets (blue) produced by teacher conjointly with data from *Mass Effect* (red) to train the student model.

Knowledge distillation:

- Teacher model is a character/role classifier trained on additional data with extra labels.
- We raise the temperature in *softmax* activation layer to smooth the class probabilities distribution.
- The knowledge coming from the soft-labels help the student model to discriminate on the *Mass Effect* characters.