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

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 + ΔΔ with CMS and VAD.
- Language-independent i-vector system.
- 2048-components UBM and T-matrix rank 400.

Evaluation:
- Performance of the binary classifier (accuracy).
- Target/non-target pairs discrimination (t-test).

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.

<table>
<thead>
<tr>
<th>Test #pairs</th>
<th>Training #pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>A 64, 800</td>
<td>B + C + D 194, 400</td>
</tr>
<tr>
<td>B 64, 800</td>
<td>A + C + D 194, 400</td>
</tr>
<tr>
<td>C 64, 800</td>
<td>A + B + D 194, 400</td>
</tr>
<tr>
<td>D 64, 800</td>
<td>A + B + C 194, 400</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th></th>
<th>2in-conc acc. tscore</th>
<th>2in-merge acc. tscore</th>
<th>siamese-net acc. tscore</th>
</tr>
</thead>
<tbody>
<tr>
<td>A (test)</td>
<td>0.49 0.67 0.52 17.66</td>
<td>0.55 32.18</td>
<td></td>
</tr>
<tr>
<td>B (test)</td>
<td>0.49 5.34 0.5 4.53</td>
<td>0.59 77.99</td>
<td></td>
</tr>
<tr>
<td>C (test)</td>
<td>0.54 7.82 0.53 18.37</td>
<td>0.62 86.17</td>
<td></td>
</tr>
<tr>
<td>D (test)</td>
<td>0.53 17.30 0.52 14.50</td>
<td>0.5 1.87</td>
<td></td>
</tr>
<tr>
<td>A (dev)</td>
<td>0.94 185.72 0.93 169.93</td>
<td>0.72 47.90</td>
<td></td>
</tr>
<tr>
<td>B (dev)</td>
<td>0.96 211.32 0.94 190.68</td>
<td>0.71 52.77</td>
<td></td>
</tr>
<tr>
<td>C (dev)</td>
<td>0.93 161.16 0.93 160.16</td>
<td>0.70 45.18</td>
<td></td>
</tr>
<tr>
<td>D (dev)</td>
<td>0.96 227.85 0.96 212.80</td>
<td>0.71 44.46</td>
<td></td>
</tr>
</tbody>
</table>

Results

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

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.

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.