icassp 2022 Singapore · imec · IDLab INTERNET & DATA LAB · GHENT UNIVERSITY

TACKLING THE SCORE SHIFT IN CROSS-LINGUAL SPEAKER VERIFICATION BY EXPLOITING LANGUAGE INFORMATION

JENTHE THIENPONDT, BRECHT DESPLANQUES, KRIS DEMUYNCK

jenthe.thienpondt@ugent.be brecht.desplanques@ugent.be

INTRODUCTION

Speaker verification
Are two utterances are spoken by the same person?

Cross-lingual challenges
Underestimation of speaker similarity in within-speaker cross-lingual trials.

Proposals
Cross-lingual fine-tuning → increase intra-speaker cross-lingual (CL) samples during fine-tuning (FT).

Language-aware calibration → incorporate language information in the logistic regression calibration stage.

BASELINE-SYSTEM: FWSE-RESNET

ResNet architecture enhanced by:

Frequency-wise Squeeze-Excitation (fwSE)
Calculates the mean descriptor across the feature maps per frequency-channel.

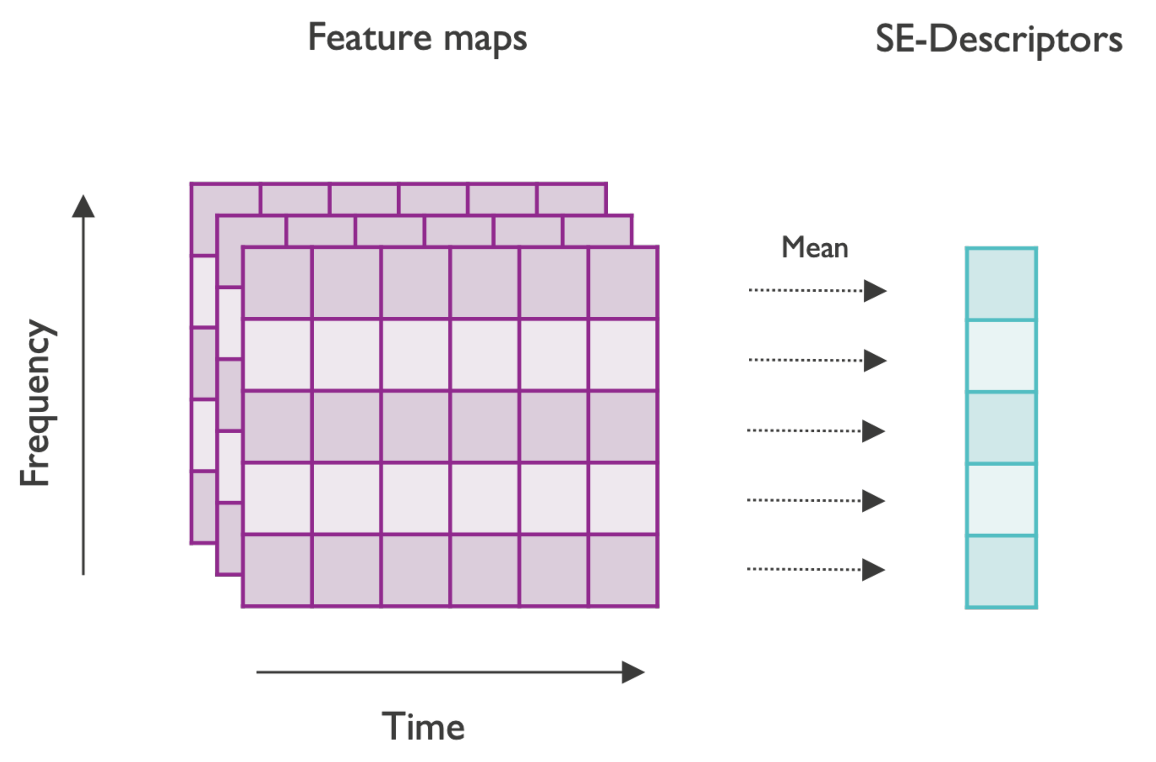


Fig. 1 *SE-descriptor calculation in the SE-block of fwSE-ResNet*

Frequency positional encodings
Enables the architecture to model frequency-dependent information.

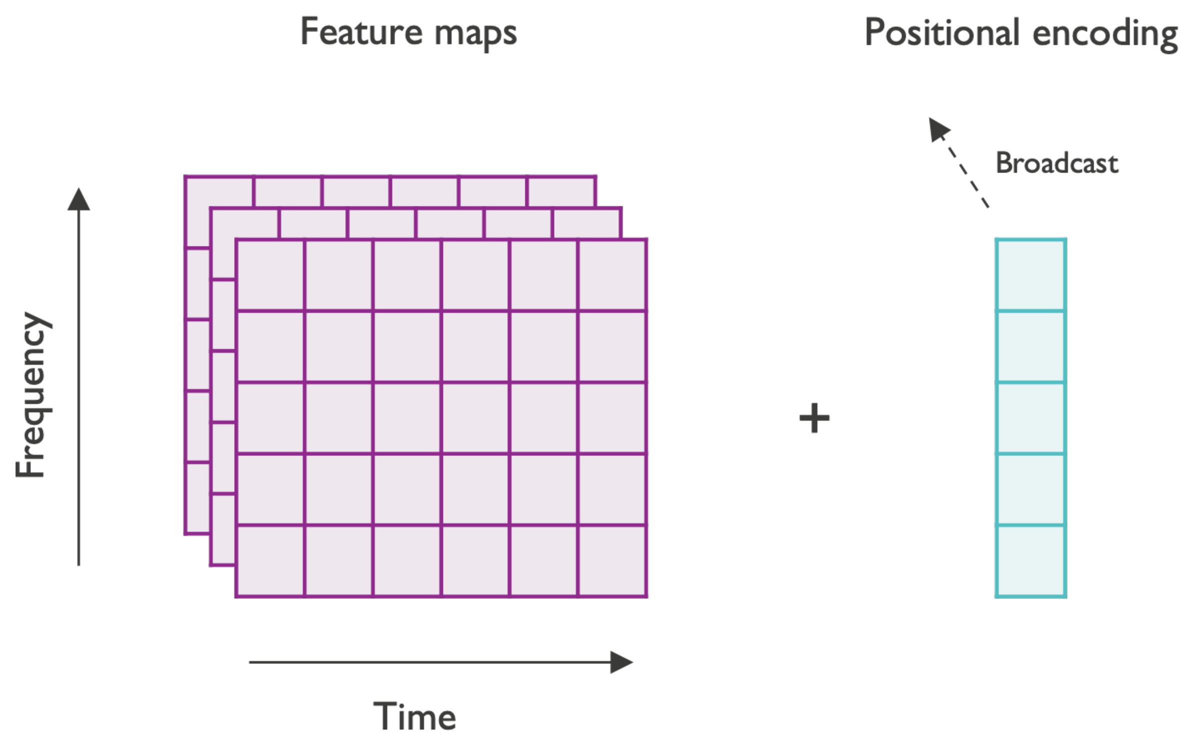


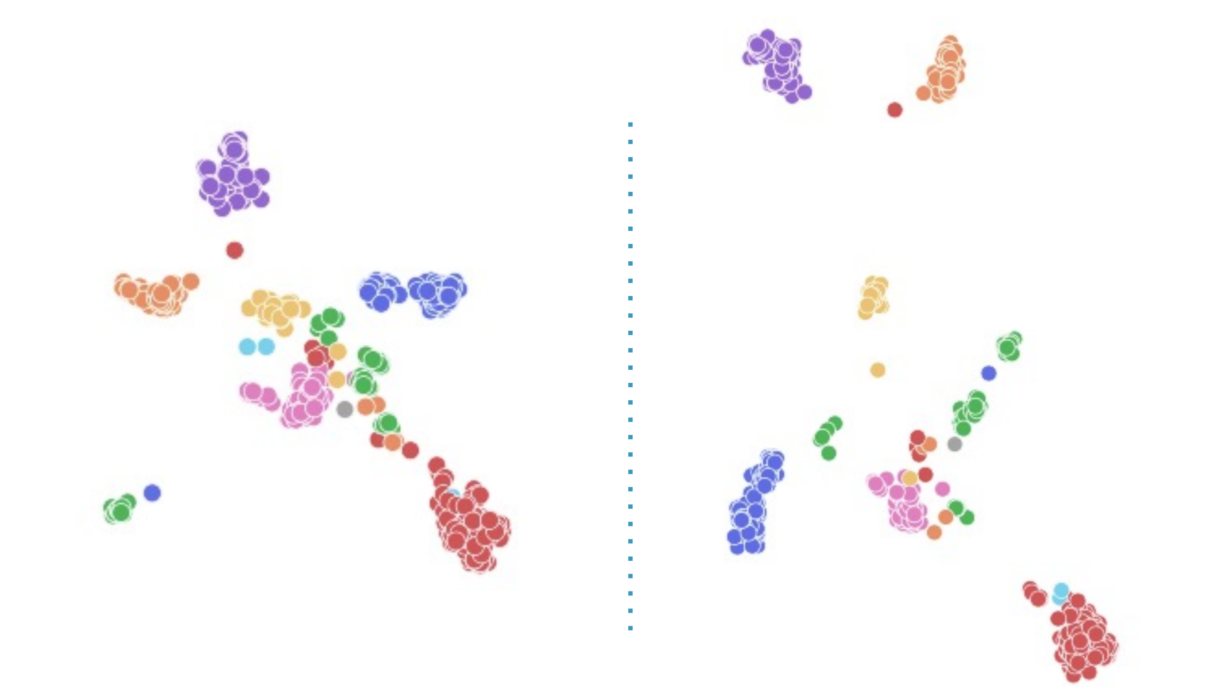
Fig. 2 *addition of frequency positional encodings in fwSE-ResNet*

CROSS-LINGUAL FINE-TUNING

Approach
1. Fine-tune model using previously proposed large-margin fine-tuning strategy.
2. Increase cross-lingual samples during FT step.

Configuration
- Select *S* random speakers from all N speakers.
- Select *U* cross-lingual utterances for each selected speaker.
- Cross-linguality determined by external language identifier.
- Resulting mini-batch size is *S x U*.

Fig. 3. *UMAP reduced embeddings of similar speakers in VoxCeleb2*



Similar male speakers before LM-FT *Similar male speakers after LM-FT*

QUALITY-AWARE SCORE CALIBRATION

- With cost of false alarms C_{fa} ,cost of a miss C_{miss} and prior target probability π:

$$\underbrace{l(s)}_{\text{Calibrated system output scores}} \geq \underbrace{log\frac{C_{fa}}{C_{miss}} - logit\,\pi}_{\text{Bayes decision threshold } \eta}$$

- Quality-aware calibration mapping function:

$$l(s) = w_s s + \boldsymbol{w}_q^T \boldsymbol{q} + b$$

with output score s, score weight w_s, bias b, learnable quality weights $\boldsymbol{w}_q$ and quality vector $\boldsymbol{q}$

→ verification decision threshold depends on the quality of the trial:

$$w_s s + b \geq \eta - \boldsymbol{w}_q^T \boldsymbol{q}$$

LANGUAGE-BASED QUALITY MEASURES (QM)

Approach
Include language information from an external language classifier in the calibration stage to compensate for score shifts due to cross-linguality.

Binary cross-linguality indicator
- Classification output of language classifier:

$$\max(\cos(s_e, s_i)) == \max(\cos(s_t, s_i))$$

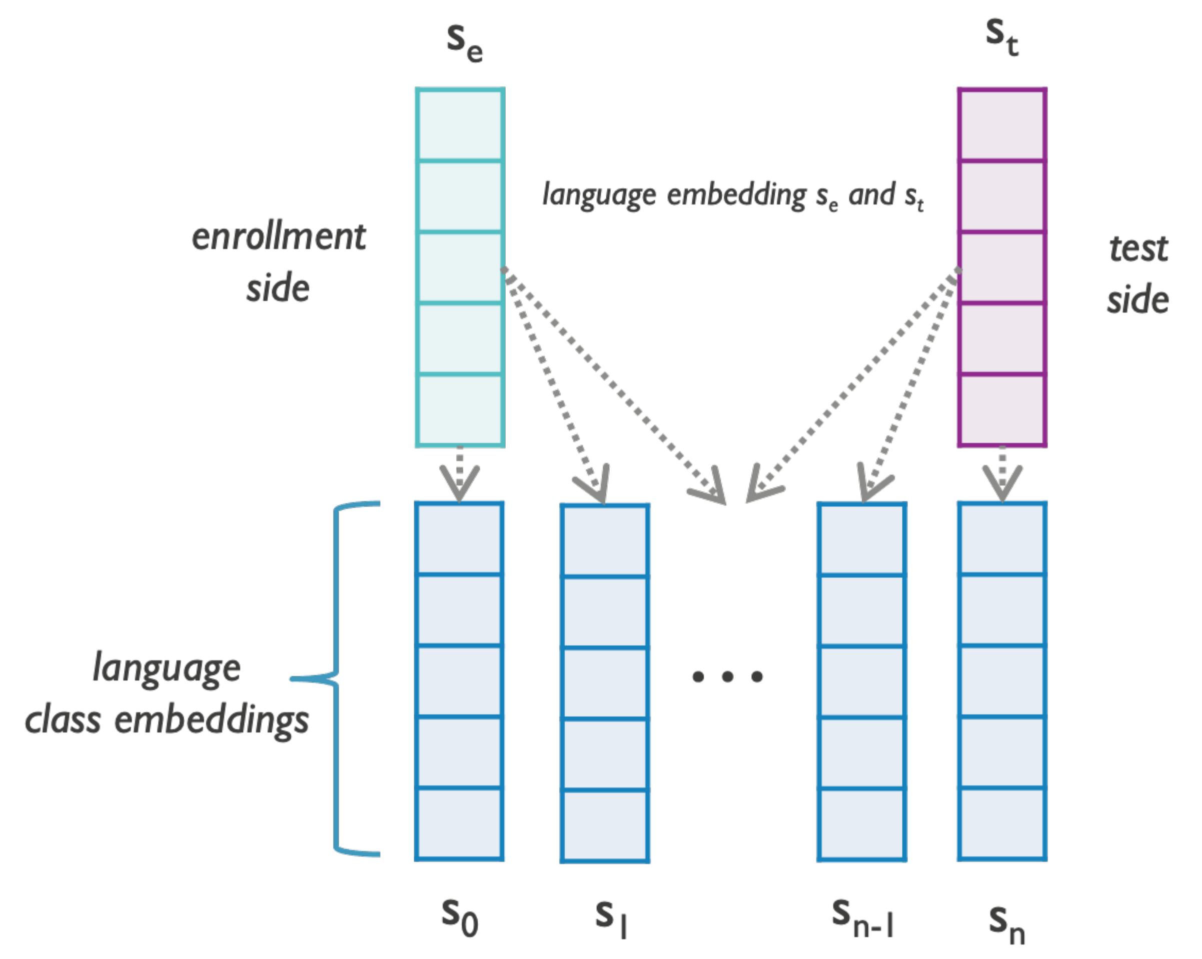


Fig. 4 *Classification output of external language classifier*

Similarity of language class probabilities
- Jensen-Shannon (JS) distance between both language classification probabilities:

$$JS\left(\left\{\frac{\cos(s_e, s_i)}{\sum_{j=1}^{N}\cos(s_e, s_j)}\right\}, \left\{\frac{\cos(s_t, s_i)}{\sum_{j=1}^{N}\cos(s_t, s_j)}\right\}\right)$$

Similarity of language embeddings
- Cosine distance of the language embeddings of the enrollment and test side of the trial:

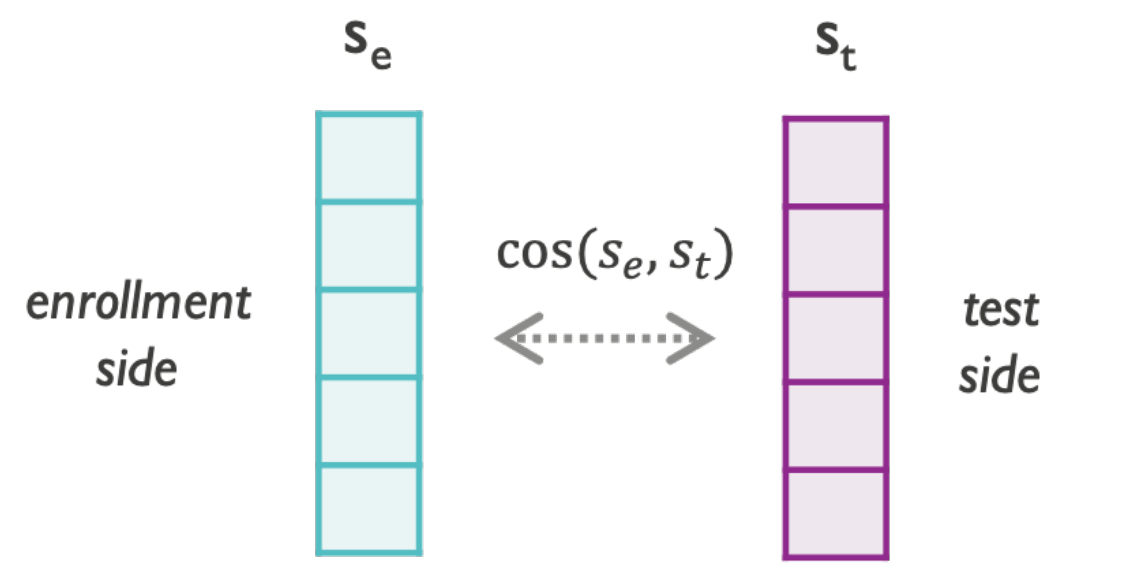
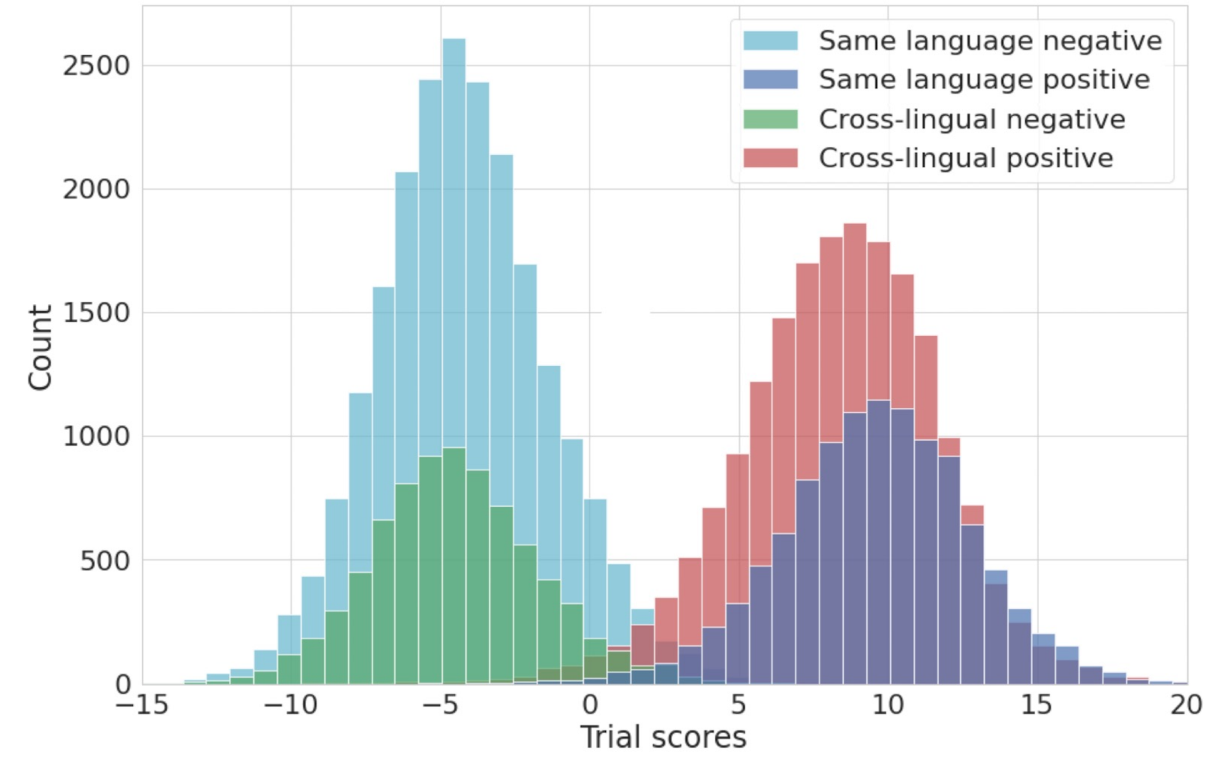


Fig. 5 *Cosine distance between language embeddings*

EXPERIMENTS & RESULTS

Fig. 6 *Histogram of the trial scores on the VoxSRC-21 validation set.*



Impact of cross-lingual sampling *on VoxSRC-21 val set*

	EER(%)	MINDCF
FWSE-RESNET	2.82	0.1538
FWSE-RESNET + LM-FT	2.41	0.1343
FWSE-RESNET + CL LM-FT	**2.25**	**0.1234**

- Cross-lingual sampling improves robustness against intra-speaker linguistic variability.

Analysis of language-aware calibration *on VoxSRC-21 val set*

	EER(%)	MINDCF
+ LOG DURATION QMF	2.11	0.1143
++ BINARY QMF	1.84	0.1038
++ JENSEN-SHANNON QMF	1.67	0.0899
++ COSINE DISTANCE QMF	**1.63**	**0.0827**

- Including language-based quality measure functions (QMF) in the calibration stage improves cross-lingual performance.
- The cosine similarity of language embeddings results in the best performance.

Results of fusion submission *on VoxSRC-21 test*

	EER(%)	MINDCF
BASELINE + LM-FT + QMF	2.78	0.1690
BASELINE + LM-FT + LANG QMF	**2.72**	**0.1492**

- Results using cosine distance language QMF.
→ 3rd place on supervised closed task of VoxSRC-21