Improved Vocal Effort Transfer Vector Estimation for Vocal Effort-Robust Speaker Verification

Iván López-Espejo[1,2], Santi Prieto[3], Alfonso Ortega[4] and Eduardo Lleida[4]

[1]Department of Electronic Systems, Aalborg University, Denmark
[2]Center for Robust Speech Systems (CRSS), The University of Texas at Dallas, USA
[3]VeriDas | das-Nano, Navarre, Spain
[4]ViVoLab, Aragón Institute for Engineering Research (I3A), University of Zaragoza, Spain

ivl@es.aau.dk, sprieto@veridas.com, {ortega,lleida}@unizar.es

AALBORG UNIVERSITY DENMARK

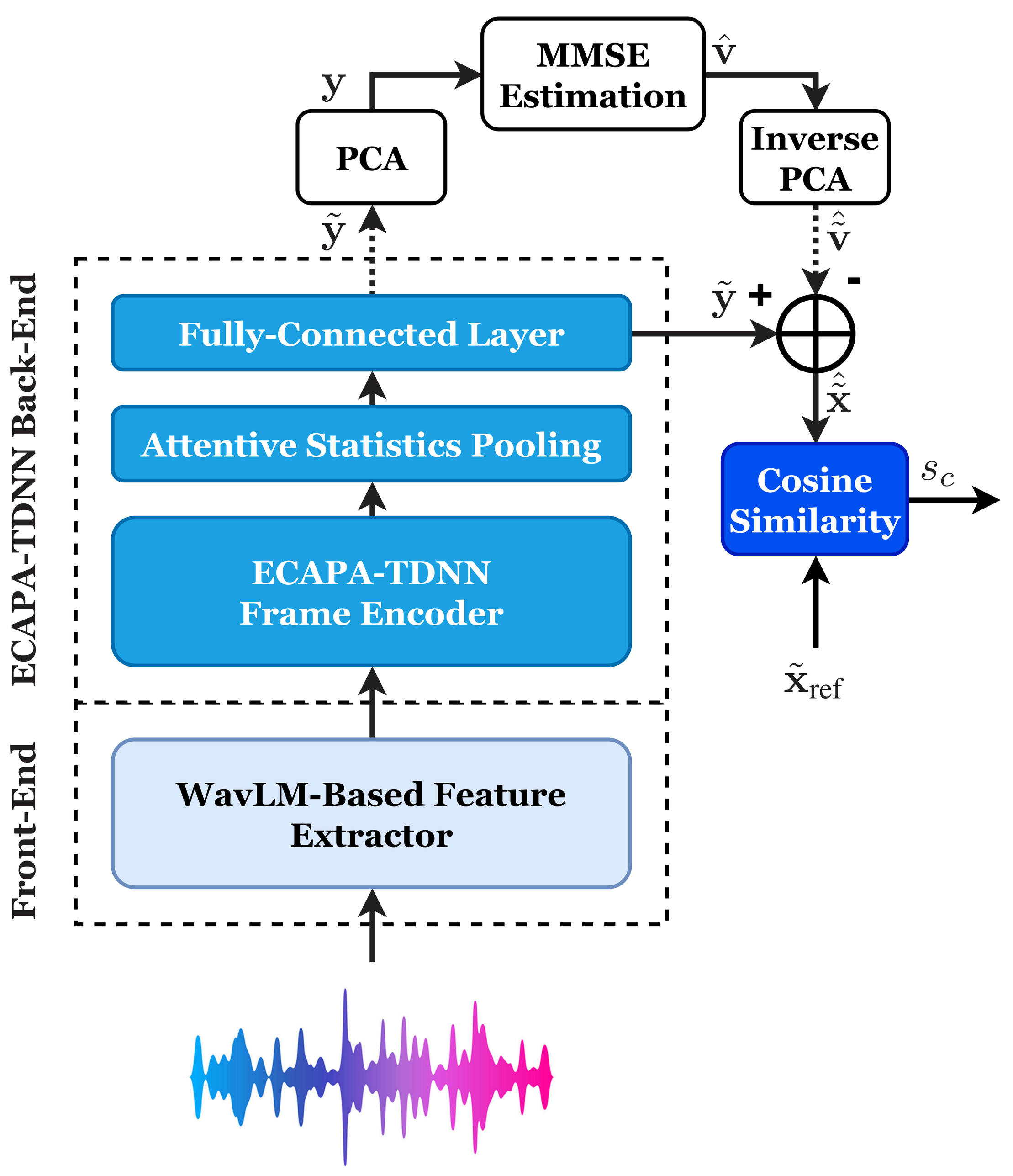
Funded by the European Union

Introduction

- Speaker verification performance tends to dramatically drop in the presence of non-neutrally-phonated (e.g., *shouted and whispered*) speech
- Previous work explored a series of *minimum mean square error* (*MMSE*) techniques estimating **normal speaker embeddings from non-neutrally-phonated ones**
- **MEMLIN** (Multi-Environment Model-based LInear Normalization) provided the best performance in terms of equal error rate (EER) when dealing with both shouted and whispered speech
- In this work **we tackle a MEMLIN's shortcoming**, which is explained in the next box 👉

System Overview

- Speaker embedding compensation is applied *only* in case that the embedding comes from non-neutrally-phonated speech
- The ECAPA-TDNN back-end is trained on an augmented version of the **VoxCeleb2** dataset



- **Shouted-normal speech corpus:** Paired utterances in Finnish from 22 speakers
- **Whispered-normal speech corpus:** Paired utterances in English from 36 speakers (*CHAINS*)
- Due to speech data scarcity, experiments are performed by following a **leave-one-speaker-out cross-validation** strategy

Normal Speaker Embedding Estimation

$\tilde{\mathbf{x}} \in \mathbb{R}^D$: Normal embedding | $\tilde{\mathbf{y}} \in \mathbb{R}^D$: Non-neutrally-phonated embedding | $\tilde{\mathbf{v}} \in \mathbb{R}^D$: *Vocal effort transfer vector*

MMSE Estimation

$$\tilde{\mathbf{y}} = \tilde{\mathbf{x}} + \tilde{\mathbf{v}} \longrightarrow \text{Assuming } \tilde{\mathbf{y}} \text{ is modeled by a } K\text{-component GMM} \longrightarrow \hat{\tilde{\mathbf{x}}} = \tilde{\mathbf{y}} - \underbrace{\sum_{k=1}^{K} P(k|\tilde{\mathbf{y}})\hat{\tilde{\mathbf{v}}}^{\{k\}}}_{\hat{\tilde{\mathbf{v}}}}$$

- **Limitation of MEMLIN:** The set of partial estimates $\left\{\hat{\tilde{\mathbf{v}}}^{\{k\}}; k = 1, ..., K\right\}$ is pre-computed (during an offline training stage) and fixed

---

- **To overcome MEMLIN's shortcoming, we propose MMSE$_\mathbf{V}$:**
  1. We jointly model $\tilde{\mathbf{v}}$ and $\tilde{\mathbf{y}}$ by a K-component GMM $p(\tilde{\mathbf{z}} = (\tilde{\mathbf{v}}, \tilde{\mathbf{y}}))$
  2. Estimation is carried out in a **principal component analysis** (**PCA**) domain **to face data scarcity**

Let $\mathbf{W}_L$ be a $D \times L$ PCA transform matrix, where $L \ll D = 256$

$$\mathbf{v} = \mathbf{W}_L^\top \tilde{\mathbf{v}}, \qquad \mathbf{y} = \mathbf{W}_L^\top \tilde{\mathbf{y}}$$

$$p\left(\mathbf{z} = (\mathbf{v}, \mathbf{y}) \in \mathbb{R}^{2L}\right) = \sum_{k=1}^{K} P(k)\mathcal{N}\left(\mathbf{z}\left|\boldsymbol{\mu}_z^{\{k\}}, \boldsymbol{\Sigma}_z^{\{k\}}\right.\right), \quad \boldsymbol{\mu}_z^{\{k\}} = \begin{pmatrix} \boldsymbol{\mu}_v^{\{k\}} \\ \boldsymbol{\mu}_y^{\{k\}} \end{pmatrix}, \quad \boldsymbol{\Sigma}_z^{\{k\}} = \begin{pmatrix} \boldsymbol{\Sigma}_{vv}^{\{k\}} & \boldsymbol{\Sigma}_{vy}^{\{k\}} \\ \boldsymbol{\Sigma}_{yv}^{\{k\}} & \boldsymbol{\Sigma}_{yy}^{\{k\}} \end{pmatrix}$$

MMSE$_\mathrm{V}$ Compensation

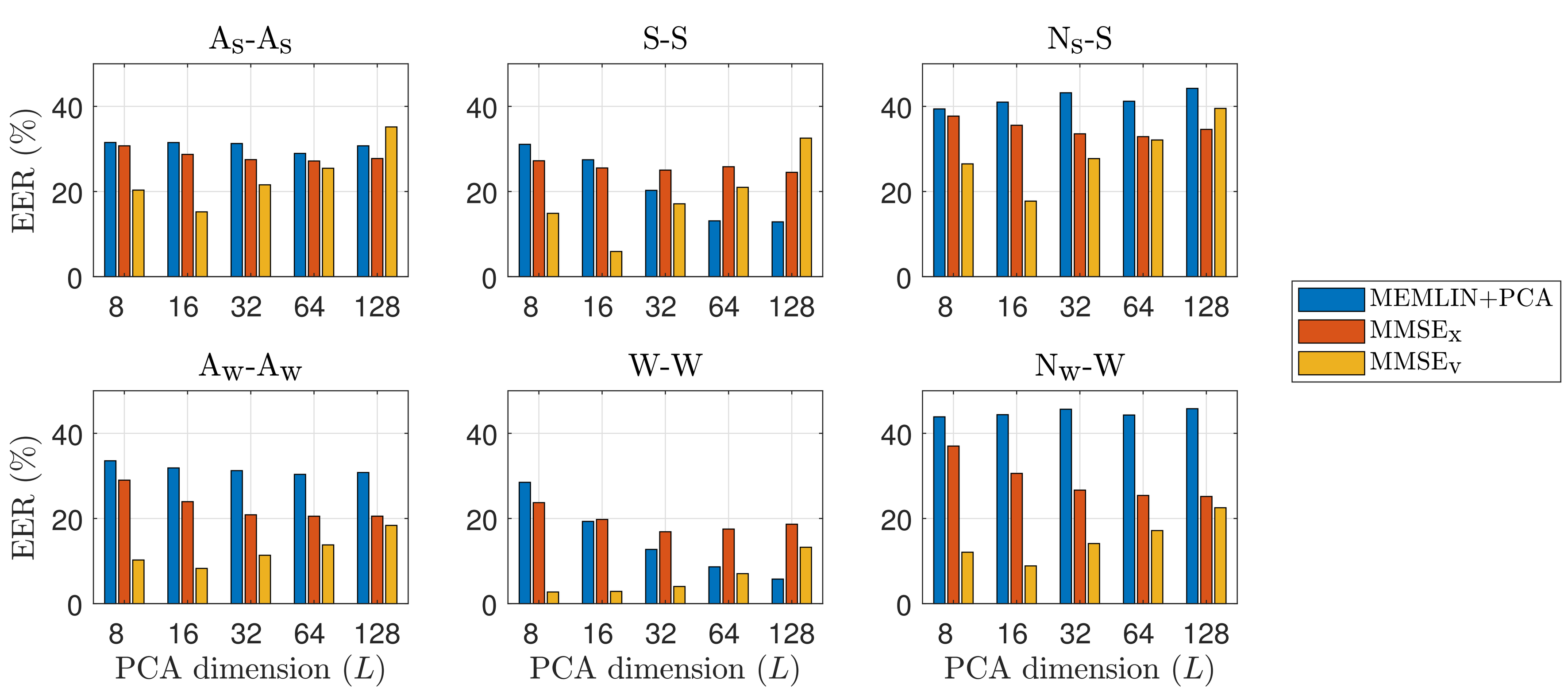
$$\hat{\mathbf{v}} = \mathbb{E}(\mathbf{v}|\mathbf{y}) = \sum_{k=1}^{K} P(k|\mathbf{y})\underbrace{\mathbb{E}(\mathbf{v}|\mathbf{y}, k)}_{\hat{\mathbf{v}}^{\{k\}}} \longrightarrow \hat{\tilde{\mathbf{x}}} = \tilde{\mathbf{y}} - \underbrace{\mathbf{W}_L\hat{\mathbf{v}}}_{\hat{\tilde{\mathbf{v}}}}$$

1. **Combination weights:** $P(k|\mathbf{y}) = \dfrac{p(\mathbf{y}|k)P(k)}{\sum_{k'=1}^{K} p(\mathbf{y}|k')P(k')}$
2. **Partial estimates:** $\mathbb{E}(\mathbf{v}|\mathbf{y}, k) = \boldsymbol{\mu}_v^{\{k\}} + \boldsymbol{\Sigma}_{vy}^{\{k\}}\left(\boldsymbol{\Sigma}_{yy}^{\{k\}}\right)^{-1}\left(\mathbf{y} - \boldsymbol{\mu}_y^{\{k\}}\right)$

- Both $\mathbf{W}_L$ and $p(\mathbf{z})$ are calculated from paired normal and non-neutrally-phonated embeddings

Experimental Results and Discussion

- **EER** (%) is the chosen speaker verification metric | Use of K = 8-component GMMs
- Embedding compensation experiments are carried out by employing *E-T+WavLM as the baseline system*
- **MMSE$_\mathbf{X}$:** MMSE estimator equivalent to MMSE$_\mathrm{V}$ that directly estimates $\hat{\tilde{\mathbf{x}}}$ from $\mathbb{E}(\mathbf{x}|\mathbf{y})$



Shouted and normal speech:

Condition	E-T+MFCC	E-T+WavLM	MEMLIN	MEMLIN+PCA	MMSE$_X$	MMSE$_V$
A_S-A_S	19.96	17.11	15.62	31.50	28.72	**15.22**
N_S-N_S	9.73	**7.25**	**7.25**	**7.25**	**7.25**	**7.25**
S-S	11.58	9.94	10.44	27.46	25.53	**5.91**
N_S-S	25.28	21.76	20.74	41.00	35.56	**17.74**

Whispered and normal speech:

Condition	E-T+MFCC	E-T+WavLM	MEMLIN	MEMLIN+PCA	MMSE$_X$	MMSE$_V$
A_W-A_W	16.54	11.24	**8.25**	31.87	23.95	8.27
N_W-N_W	1.21	**0.62**	**0.62**	**0.62**	**0.62**	**0.62**
W-W	4.38	5.26	4.00	19.31	19.77	**2.87**
N_W-W	12.81	9.81	11.47	44.38	30.59	**8.86**