Dual-Channel VTS Feature Compensation with Improved Posterior Estimation

Iván López-Espejo[1], *Antonio M. Peinado*[2], *Angel M. Gomez*[2], *José A. González*[3], *and Santiago Prieto-Calero*[1]

[1]VeriDas | das-Nano, Spain

[2]Dept. of Signal Theory, Telematics and Communications, University of Granada, Spain

[3]Dept. of Languages and Computer Science, University of Málaga, Spain

Introduction

Motivation

- **Noise-robust ASR** is of utmost importance in **mobile devices**.
- Current mobile devices embed **several microphones**.

Objective

- To improve **posterior** computation for **dual-channel VTS feature compensation**.

Dual-Channel VTS Feature Compensation

- It is expected that the 1st ch. is less affected by noise than the 2nd one.
- Log-Mel clean speech features are estimated as ($\mathcal{K}$-component GMM)

$$\hat{\mathbf{x}}_1 = \sum_{k=1}^{\mathcal{K}} P(k|\mathbf{y})\hat{\mathbf{x}}_1^{(k)}, \text{ where } \mathbf{y} = \begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{pmatrix}$$

Clean speech partial estimates

- $\hat{\mathbf{x}}_1^{(k)} = \mathbf{y}_1 - \log\left(\mathbf{1} + e^{\boldsymbol{\mu}_{n_1} - \boldsymbol{\mu}_{x_1}^{(k)}}\right)$

Posterior probabilities

- **Stacked formulation:** The 2nd ch. is treated in a parallel manner to the 1st one.
- **The relation between the noisy speech at the 2nd ch. and the clean speech is more uncertain** (because of the speech masking effect) than that of the 1st ch.
- **More robust:** To condition the distortion model of the 2nd ch. to the 1st ch. observation.
- **New derivation** from replacing $P(k|\mathbf{y})$ by $P(k|\mathbf{y}_1, \mathbf{y}_2)$.

Improved Posterior Probability Computation

Posterior computation

- $P(k|\mathbf{y}_1, \mathbf{y}_2) = \dfrac{p(\mathbf{y}_1, \mathbf{y}_2|k)P(k)}{\sum_{k'=1}^{\mathcal{K}} p(\mathbf{y}_1, \mathbf{y}_2|k')P(k')}$
- $p(\mathbf{y}_1, \mathbf{y}_2|k) = p(\mathbf{y}_1|k)p(\mathbf{y}_2|\mathbf{y}_1, k)$
- Using a **VTS** approach both $p(\mathbf{y}_1|k)$ and $p(\mathbf{y}_2|\mathbf{y}_1, k)$ are modeled as **Gaussian PDFs**.

Dual-channel speech distortion model

- $\mathbf{y}_1 = \mathbf{x}_1 + \log\left(\mathbf{1} + e^{\mathbf{n}_1 - \mathbf{x}_1}\right)$
- $\mathbf{y}_2 = \mathbf{x}_1 + \mathbf{a}_{21} + \log\left(\mathbf{1} + e^{\mathbf{n}_2 - \mathbf{x}_1 - \mathbf{a}_{21}}\right)$
- **Relative acoustic path:** $\mathbf{x}_2 = \mathbf{a}_{21} + \mathbf{x}_1$

$$p(\mathbf{y}_1|k) = \mathcal{N}\left(\mathbf{y}_1 \left| \boldsymbol{\mu}_{y_1}^{(k)}, \boldsymbol{\Sigma}_{y_1}^{(k)} \right.\right)$$

- **Speech distortion model**

$$\mathbf{y}_1 = \mathbf{x}_1 + \log\left(\mathbf{1} + e^{\mathbf{n}_1 - \mathbf{x}_1}\right)$$

- **Mean vector**

$$\boldsymbol{\mu}_{y_1}^{(k)} = \boldsymbol{\mu}_{x_1}^{(k)} + \log\left(\mathbf{1} + e^{\boldsymbol{\mu}_{n_1} - \boldsymbol{\mu}_{x_1}^{(k)}}\right)$$

- **Covariance matrix**

$$\boldsymbol{\Sigma}_{y_1}^{(k)} = \mathbf{J}_{x_1}^{(1,k)} \boldsymbol{\Sigma}_{x_1}^{(k)} \mathbf{J}_{x_1}^{(1,k)\top} + \mathbf{J}_{n_1}^{(1,k)} \boldsymbol{\Sigma}_{n_1} \mathbf{J}_{n_1}^{(1,k)\top}$$

$$p(\mathbf{y}_2|\mathbf{y}_1, k) = \mathcal{N}\left(\mathbf{y}_2 \left| \boldsymbol{\mu}_{y_2|y_1}^{(k)}, \boldsymbol{\Sigma}_{y_2|y_1}^{(k)} \right.\right)$$

- **Speech distortion model**

$$\mathbf{y}_2(\mathbf{y}_1) = \mathbf{y}_1 + \mathbf{a}_{21} + \log\left[\frac{\mathbf{1} + e^{\mathbf{n}_2 - \mathbf{x}_1 - \mathbf{a}_{21}}}{\mathbf{1} + e^{\mathbf{n}_1 - \mathbf{x}_1}}\right]$$

- **Mean vector**

$$\boldsymbol{\mu}_{y_2|y_1}^{(k)} = \mathbf{y}_1 + \boldsymbol{\mu}_{a_{21}} + \log\left[\frac{\mathbf{1} + e^{\boldsymbol{\mu}_{n_2} - \boldsymbol{\mu}_{x_1}^{(k)} - \boldsymbol{\mu}_{a_{21}}}}{\mathbf{1} + e^{\boldsymbol{\mu}_{n_1} - \boldsymbol{\mu}_{x_1}^{(k)}}}\right]$$

- **Covariance matrix**

$$\begin{aligned}\boldsymbol{\Sigma}_{y_2|y_1}^{(k)} = {} & \mathbf{J}_{x_1}^{(2,k)} \boldsymbol{\Sigma}_{x_1}^{(k)} \mathbf{J}_{x_1}^{(2,k)\top} + \mathbf{J}_{a_{21}}^{(2,k)} \boldsymbol{\Sigma}_{a_{21}} \mathbf{J}_{a_{21}}^{(2,k)\top} \\ & + \mathbf{J}_{n_1}^{(2,k)} \boldsymbol{\Sigma}_{n_1} \mathbf{J}_{n_1}^{(2,k)\top} + \mathbf{J}_{n_2}^{(2,k)} \boldsymbol{\Sigma}_{n_2} \mathbf{J}_{n_2}^{(2,k)\top} \\ & + \mathbf{J}_{n_1}^{(2,k)} \boldsymbol{\Sigma}_{n_1 n_2} \mathbf{J}_{n_2}^{(2,k)\top} + \mathbf{J}_{n_2}^{(2,k)} \boldsymbol{\Sigma}_{n_2 n_1} \mathbf{J}_{n_1}^{(2,k)\top}\end{aligned}$$

Experimental Framework

Speech recognition experiments on a **dual-microphone smartphone** used in both

- **close-talk** (AURORA2-2C-CT) and
- **far-talk** (AURORA2-2C-FT) conditions.

Front-end

- (13 MFCCs with CMVN)+Δ + $\Delta\Delta$

DNN-HMM back-end

- Clean acoustic models
- Multi-style acoustic models

Conclusions

Conclusions

- **Accurate posteriors have been obtained** by modeling the conditional dependence of the noisy 2nd ch. given the 1st one.
- **The new way of computing the posteriors has overcome the constraints of the stacked formulation** when combined with MMSN and DCSS in far-talk conditions.

Future work

- We will research on how to exploit the dual-channel information for better clean speech partial estimate computation.

Results

- **Word accuracy results (%)**.
- Our proposal (**2-VTS-C**) against dual-channel VTS based on a stacked formulation (**2-VTS-S**) and single-channel VTS applied on the 1st ch. (**1-VTS**).

	CLOSE-TALK						*FAR-TALK*					
	Clean models			*Multi-style models*			*Clean models*			*Multi-style models*		
	Test *A*	**Test *B***	**Average**	**Test *A***	**Test *B***	**Average**	**Test *A***	**Test *B***	**Average**	**Test *A***	**Test *B***	**Average**
Baseline	36.76	31.52	34.14	90.97	77.27	84.12	40.96	30.53	35.74	91.46	74.69	83.07
AFE	74.32	69.00	71.66	89.84	83.89	86.86	74.33	69.20	71.77	90.38	83.37	86.88
MVDR	46.72	38.98	42.85	91.34	83.57	87.45	52.71	39.71	46.21	**93.90**	84.71	89.30
1-VTS	84.37	78.05	81.21	89.76	84.20	86.98	84.39	79.06	81.72	90.01	85.12	87.57
2-VTS-S	88.23	83.23	85.73	91.50	87.36	89.43	86.57	81.23	83.90	91.01	86.32	88.66
2-VTS-C	**88.70**	**83.44**	**86.07**	**91.87**	**87.66**	**89.77**	**87.82**	**82.46**	**85.14**	91.61	**87.05**	**89.33**

- Dual-channel power spectrum enhancement (**MMSN** and **DCSS**) as pre-processing for VTS.

	CLOSE-TALK						*FAR-TALK*					
	Clean models			*Multi-style models*			*Clean models*			*Multi-style models*		
	Test *A*	**Test *B***	**Average**	**Test *A***	**Test *B***	**Average**	**Test *A***	**Test *B***	**Average**	**Test *A***	**Test *B***	**Average**
MMSN-1	89.55	84.62	87.08	92.74	88.84	90.79	88.29	82.67	85.48	92.47	87.51	89.99
MMSN-2S	90.02	85.49	87.75	92.99	89.44	91.22	88.07	82.70	85.39	92.21	87.72	89.96
MMSN-2C	**91.03**	86.26	88.64	**93.56**	**90.03**	**91.80**	89.60	**84.04**	86.82	93.00	**88.58**	90.79
DCSS-1	89.65	84.72	87.19	92.92	88.99	90.95	88.77	83.10	85.93	92.66	87.95	90.30
DCSS-2S	90.06	85.57	87.82	92.84	89.46	91.15	88.37	83.14	85.76	92.33	87.90	90.11
DCSS-2C	91.02	**86.31**	**88.67**	93.47	89.93	91.70	**89.70**	**84.04**	**86.87**	**93.24**	88.53	**90.88**

Contact Information

Iván López-Espejo

VeriDas | das-Nano (www.veri-das.com)

Navarre, Spain

E-mail: ilopez@das-nano.com