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
{ilopez,sprieto}@das-nano.com, {amp,amgg}@ugr.es, jgonzalez@lcc.uma.es

***Abstract*—The use of dual-microphones is a powerful tool for noise-robust automatic speech recognition (ASR). In particular, it allows the reformulation of classical techniques like vector Taylor series (VTS) feature compensation. In this work, we consider a critical issue of VTS compensation such as posterior computation and propose an alternative way to estimate more accurately these probabilities when VTS is applied to enhance noisy speech captured by dual-microphone mobile devices. Our proposal models the conditional dependence of a noisy secondary channel given a primary one not only to outperform single-channel VTS feature compensation, but also a previous dual-channel VTS approach based on a stacked formulation. This is confirmed by recognition experiments on two different dual-channel extensions of the Aurora-2 corpus. Such extensions emulate the use of a dual-microphone smartphone in close- and far-talk conditions, obtaining our proposal relevant improvements in the latter case.**

***Index Terms*—VTS feature compensation, Posterior probability, Robust ASR, Dual-channel, Mobile device**

I. Introduction

Achieving robustness against noise in automatic speech recognition (ASR) is of utmost importance nowadays due to the wide use of mobile devices [1]. These devices frequently incorporate several microphones for speech enhancement purposes. Additionally, the microphones can be exploited to improve ASR performance in noisy conditions [2]–[5]. In this regard, in our previous work [5], a vector Taylor series (VTS) feature compensation approach was extended to be performed on dual-microphone mobile devices. This method consists of a minimum mean square error (MMSE)-based estimator of the log-Mel clean speech features relying on a VTS expansion of a dual-channel speech distortion model. For posterior computation, it follows a stacked formulation in which the two-channel joint information is indirectly exploited by means of the spatial covariance matrix of noise and a term modeling the clean speech relative acoustic path (RAP) between the two sensors of the device. This dual-channel VTS method proved to be quite effective when applied to dual-microphone recordings from a smartphone employed in close-talk conditions (i.e., when the phone loudspeaker is placed at the ear of the user). However, as shown in this work, in far-talk conditions (i.e., when the user holds the smartphone in one hand at a particular distance from her/his face), the improvement provided by this dual-channel VTS method over the single-channel one, in terms of recognition performance, is limited.

Thus, in this paper we propose a novel alternative to compute the posteriors required for dual-channel VTS feature compensation. Unlike the previous stacked formulation, the strategy followed in this work explicitly models the conditional dependence of the noisy secondary channel given the primary one. This leads to a new derivation where the correlations between the two channels are better exploited and not in an indirect way as for the stacked case discussed above. This is confirmed by our speech recognition results not only by outperforming our previous dual-channel VTS approach in close-talk conditions, but also achieving meaningful improvements in far-talk conditions.

The rest of the paper is organized as follows. In Section II, the dual-channel VTS feature compensation is briefly revisited and the problem addressed in this work is stated. The novel approach for computing the posteriors required in Section II is developed in Section III. Both the experimental framework and results are shown in Section IV. Finally, in Section V, conclusions and future work are outlined.

II. Dual-Channel VTS Feature Compensation

First of all, for convenience reasons, the same framework and mathematical notation as in [5] are adopted in this paper. Thus, let us again consider the well-known speech distortion model for additive noise in the log-Mel power spectral domain [6], [7]:

$$\mathbf{y}_i = \log\left(e^{\mathbf{x}_i} + e^{\mathbf{n}_i}\right), \tag{1}$$

where $\mathbf{y}_i$, $\mathbf{x}_i$ and $\mathbf{n}_i$ represent noisy speech, clean speech and noise log-Mel feature vectors from the i-th channel of the mobile device at a particular time frame. Specifically, $i = 1$ corresponds to the primary microphone and $i = 2$ to the secondary one. Since the primary microphone is often located at the bottom of the device and the secondary one at its top or rear, it is expected that the primary signal is not more affected by the ambient noise than the secondary one. Under the assumption that the log-Mel clean speech features at the primary channel can be modeled by a $\mathcal{K}$-component Gaussian

This work has been supported by the Spanish MINECO/FEDER Project TEC2016-80141-P.

ISBN 978-90-827970-1-5 © EURASIP 2018

mixture model (GMM), the log-Mel clean speech features are estimated in [5] at every time frame as

$$\hat{\mathbf{x}}_1 = \sum_{k=1}^{\mathcal{K}} P(k|\mathbf{y})\hat{\mathbf{x}}_1^{(k)}, \tag{2}$$

where $\mathbf{y} = \left(\mathbf{y}_1^\top, \mathbf{y}_2^\top\right)^\top$ is a stacked vector and $P(k|\mathbf{y})$ is the k-th posterior probability which weights the corresponding clean speech partial estimate, $\hat{\mathbf{x}}_1^{(k)}$.

A main feature of the stacked formulation described above is that the secondary channel is treated in a parallel manner to the primary one, using similar distortion models. However, as suggested above, it is likely that noise dominates at the secondary channel, and, therefore, we can expect that the relation between the noisy speech captured by the secondary microphone and the clean speech is more uncertain (because of the speech masking effect) than that of the primary channel. We have found it more robust to condition this distortion model at the secondary channel to the actual noisy observation from the primary channel since both channels are greatly correlated. This can be accomplished by replacing $P(k|\mathbf{y})$ in (2) by $P\left(k|\mathbf{y}_1, \mathbf{y}_2\right)$, which is further decomposed as the product of an *a priori* and a conditional probability density function (PDF) as shown in Eqs. (4) and (5). In addition, since the secondary signal is usually noisier than the primary one, obtaining $\hat{\mathbf{x}}_1^{(k)}$ by only taking into account the primary channel instead of the dual-channel information was shown to perform better [5]. Hence, that same clean speech partial estimate computation approach is followed in this work, namely

$$\hat{\mathbf{x}}_1^{(k)} = \mathbf{y}_1 - \log\left(\mathbf{1} + e^{\boldsymbol{\mu}_{n_1} - \boldsymbol{\mu}_{x_1}^{(k)}}\right), \tag{3}$$

where $\boldsymbol{\mu}_{n_i}$ is a noise mean vector from the i-th channel ($i = 1, 2$), $\boldsymbol{\mu}_{x_1}^{(k)}$ is the mean vector of the k-th component of the clean speech GMM, and $\mathbf{1}$ is an $\mathcal{M}$-dimensional vector filled with ones, where $\mathcal{M}$ is the number of filterbank channels.

III. Improved Posterior Probability Computation

The posteriors $\{P\left(k|\mathbf{y}_1, \mathbf{y}_2\right); k = 1, 2, ..., \mathcal{K}\}$ can be calculated by employing the Bayes' theorem as

$$P\left(k|\mathbf{y}_1, \mathbf{y}_2\right) = \frac{p\left(\mathbf{y}_1, \mathbf{y}_2|k\right) P(k)}{\sum_{k'=1}^{\mathcal{K}} p\left(\mathbf{y}_1, \mathbf{y}_2|k'\right) P(k')}, \tag{4}$$

where $P(k)$ is the prior probability of the k-th component of the clean speech GMM, and the PDF $p\left(\mathbf{y}_1, \mathbf{y}_2|k\right)$ can be factored as

$$p\left(\mathbf{y}_1, \mathbf{y}_2|k\right) = p(\mathbf{y}_1|k)p(\mathbf{y}_2|\mathbf{y}_1, k). \tag{5}$$

Then, by using a VTS approach [5], [8], both $p(\mathbf{y}_1|k)$ and $p(\mathbf{y}_2|\mathbf{y}_1, k)$ will be modeled as Gaussian PDFs and their parameters are obtained as described in the following.

First, the speech distortion model of (1) is adapted to the primary and secondary channels, respectively, as

$$\mathbf{y}_1 = \mathbf{x}_1 + \log\left(\mathbf{1} + e^{\mathbf{n}_1 - \mathbf{x}_1}\right), \tag{6}$$

$$\mathbf{y}_2 = \mathbf{x}_1 + \mathbf{a}_{21} + \log\left(\mathbf{1} + e^{\mathbf{n}_2 - \mathbf{x}_1 - \mathbf{a}_{21}}\right), \tag{7}$$

where $\mathbf{a}_{21}$ is the relative acoustic path (RAP) between the two sensors so that $\mathbf{x}_2 = \mathbf{a}_{21} + \mathbf{x}_1$. Eqs. (6) and (7) can be combined to define an alternative speech distortion model for the secondary channel given $\mathbf{y}_1$, as,

$$\mathbf{y}_2(\mathbf{y}_1) = \mathbf{y}_1 + \mathbf{a}_{21} + \log\left[\frac{\mathbf{1} + e^{\mathbf{n}_2 - \mathbf{x}_1 - \mathbf{a}_{21}}}{\mathbf{1} + e^{\mathbf{n}_1 - \mathbf{x}_1}}\right]. \tag{8}$$

Assuming that both $\mathbf{a}_{21}$ and $\mathbf{n}_i$ ($i = 1, 2$) can be modeled by Gaussian distributions [5], [8], [9], Eqs. (6) and (8) are linearized by means of a first-order VTS expansion to obtain the parameters (i.e., mean vectors and covariance matrices) of $p(\mathbf{y}_1|k) = \mathcal{N}\left(\mathbf{y}_1 \left| \boldsymbol{\mu}_{y_1}^{(k)}, \boldsymbol{\Sigma}_{y_1}^{(k)}\right.\right)$ and $p(\mathbf{y}_2|\mathbf{y}_1, k) = \mathcal{N}\left(\mathbf{y}_2 \left| \boldsymbol{\mu}_{y_2|y_1}^{(k)}, \boldsymbol{\Sigma}_{y_2|y_1}^{(k)}\right.\right)$, respectively. By following this procedure, it is straightforward to demonstrate that the mean vectors are given by

$$\begin{aligned}
\boldsymbol{\mu}_{y_1}^{(k)} &= \boldsymbol{\mu}_{x_1}^{(k)} + \log\left(\mathbf{1} + e^{\boldsymbol{\mu}_{n_1} - \boldsymbol{\mu}_{x_1}^{(k)}}\right), \\
\boldsymbol{\mu}_{y_2|y_1}^{(k)} &= \mathbf{y}_1 + \boldsymbol{\mu}_{a_{21}} + \log\left[\frac{\mathbf{1} + e^{\boldsymbol{\mu}_{n_2} - \boldsymbol{\mu}_{x_1}^{(k)} - \boldsymbol{\mu}_{a_{21}}}}{\mathbf{1} + e^{\boldsymbol{\mu}_{n_1} - \boldsymbol{\mu}_{x_1}^{(k)}}}\right],
\end{aligned} \tag{9}$$

where $\boldsymbol{\mu}_{a_{21}}$ is the mean vector of the RAP term. By proceeding analogously, it is easy to show that the covariance matrix of $p(\mathbf{y}_1|k)$ can be approximated as

$$\boldsymbol{\Sigma}_{y_1}^{(k)} = \mathbf{J}_{x_1}^{(1,k)} \boldsymbol{\Sigma}_{x_1}^{(k)} \mathbf{J}_{x_1}^{(1,k)^\top} + \mathbf{J}_{n_1}^{(1,k)} \boldsymbol{\Sigma}_{n_1} \mathbf{J}_{n_1}^{(1,k)^\top}, \tag{10}$$

where $\boldsymbol{\Sigma}_{x_1}^{(k)}$ and $\boldsymbol{\Sigma}_{n_i}$ are the covariance matrices of the k-th component of the clean speech GMM and the noise at the i-th channel ($i = 1, 2$), respectively, and the Jacobian matrices have the following definitions:

$$\begin{aligned}
\mathbf{J}_{x_1}^{(1,k)} &= \left.\frac{\partial \mathbf{y}_1}{\partial \mathbf{x}_1}\right|_{\boldsymbol{\mu}_{x_1}^{(k)}, \boldsymbol{\mu}_{n_1}} = \operatorname{diag}\left(\frac{\mathbf{1}}{\mathbf{1} + e^{\boldsymbol{\mu}_{n_1} - \boldsymbol{\mu}_{x_1}^{(k)}}}\right), \\
\mathbf{J}_{n_1}^{(1,k)} &= \left.\frac{\partial \mathbf{y}_1}{\partial \mathbf{n}_1}\right|_{\boldsymbol{\mu}_{x_1}^{(k)}, \boldsymbol{\mu}_{n_1}} = \mathbf{I}_{\mathcal{M}} - \mathbf{J}_{x_1}^{(1,k)},
\end{aligned} \tag{11}$$

where $\operatorname{diag}(\cdot)$ is the diagonal matrix operator and $\mathbf{I}_{\mathcal{M}}$ is an $\mathcal{M} \times \mathcal{M}$ identity matrix. Similarly, the covariance matrix of the conditional PDF $p\left(\mathbf{y}_2|\mathbf{y}_1, k\right)$ is estimated as

$$\begin{aligned}
\boldsymbol{\Sigma}_{y_2|y_1}^{(k)} &= \mathbf{J}_{x_1}^{(2,k)} \boldsymbol{\Sigma}_{x_1}^{(k)} \mathbf{J}_{x_1}^{(2,k)^\top} + \mathbf{J}_{a_{21}}^{(2,k)} \boldsymbol{\Sigma}_{a_{21}} \mathbf{J}_{a_{21}}^{(2,k)^\top} \\
&\quad + \mathbf{J}_{n_1}^{(2,k)} \boldsymbol{\Sigma}_{n_1} \mathbf{J}_{n_1}^{(2,k)^\top} + \mathbf{J}_{n_2}^{(2,k)} \boldsymbol{\Sigma}_{n_2} \mathbf{J}_{n_2}^{(2,k)^\top} \\
&\quad + \mathbf{J}_{n_1}^{(2,k)} \boldsymbol{\Sigma}_{n_1 n_2} \mathbf{J}_{n_2}^{(2,k)^\top} + \mathbf{J}_{n_2}^{(2,k)} \boldsymbol{\Sigma}_{n_2 n_1} \mathbf{J}_{n_1}^{(2,k)^\top},
\end{aligned} \tag{12}$$

where $\boldsymbol{\Sigma}_{n_1 n_2} = \boldsymbol{\Sigma}_{n_2 n_1}^\top$ is a cross-covariance matrix of noise, $\boldsymbol{\Sigma}_{a_{21}}$ is the covariance matrix of the RAP factor, and the

ISBN 978-90-827970-1-5

corresponding Jacobian matrices are calculated in a similar way as in (11):

$$\begin{aligned}
\mathbf{J}_{x_1}^{(2,k)} &= \text{diag}\left(\frac{e^{\boldsymbol{\mu}_{n_1}-\boldsymbol{\mu}_{x_1}^{(k)}}-e^{\boldsymbol{\mu}_{n_2}-\boldsymbol{\mu}_{x_1}^{(k)}-\boldsymbol{\mu}_{a_{21}}}}{\left(\mathbf{1}+e^{\boldsymbol{\mu}_{n_1}-\boldsymbol{\mu}_{x_1}^{(k)}}\right)\left(\mathbf{1}+e^{\boldsymbol{\mu}_{n_2}-\boldsymbol{\mu}_{x_1}^{(k)}-\boldsymbol{\mu}_{a_{21}}}\right)}\right), \\
\mathbf{J}_{a_{21}}^{(2,k)} &= \text{diag}\left(\frac{\mathbf{1}}{\mathbf{1}+e^{\boldsymbol{\mu}_{n_2}-\boldsymbol{\mu}_{x_1}^{(k)}-\boldsymbol{\mu}_{a_{21}}}}\right), \\
\mathbf{J}_{n_1}^{(2,k)} &= -\mathbf{J}_{n_1}^{(1,k)}, \\
\mathbf{J}_{n_2}^{(2,k)} &= \mathbf{I}_{\mathcal{M}}-\mathbf{J}_{a_{21}}^{(2,k)}.
\end{aligned} \tag{13}$$

To perform the above calculations, the parameters of the different PDFs (i.e., $p(\mathbf{x}_1)$, $p(\mathbf{a}_{21})$ and $p(\mathbf{n}_i)$, $i=1,2$) plus $\boldsymbol{\Sigma}_{n_1n_2}$ are obtained as suggested in [5]. In summary, a 256-component clean speech GMM is employed, both $\boldsymbol{\mu}_{a_{21}}$ and $\boldsymbol{\Sigma}_{a_{21}}$ are determined *a priori* from a development dataset, and the parameters of $p(\mathbf{n}_i)$, $i=1,2$, and $\boldsymbol{\Sigma}_{n_1n_2}$ are estimated on an utterance-by-utterance basis by considering that no speech is present in the first and last 20 frames of every utterance (i.e., noise only).

IV. Experiments and Results

The method developed in this work is evaluated in terms of word recognition accuracy when applied to a dual-microphone smartphone employed in noisy environments in both close- and far-talk conditions. Subsection IV-A briefly describes the experimental framework, while the results are set out in Subsection IV-B.

A. Experimental Framework

In this paper the AURORA2-2C-CT (Aurora-2 - 2 Channels - Close-Talk) and the AURORA2-2C-FT (Aurora-2 - 2 Channels - Far-Talk) corpora, generated as extensions to the well-known Aurora-2 database [10], are used. The AURORA2-2C-CT, described in detail in [3], emulates the capturing of noisy speech by using a dual-microphone smartphone in close-talk conditions. By following an analogous procedure as in [3], the AURORA2-2C-FT was created for far-talk condition experiments. Since both corpora follow the Aurora-2 structure, two test sets, A and B, are defined for each database from dual-channel utterances contaminated with different types of noise. The signal-to-noise ratios (SNRs) considered are referred to the primary channel and they are the same as in [10]: from -5 dB to 20 dB with a step of 5 dB (plus the clean case).

The European Telecommunications Standards Institute front-end (ETSI FE, ES 201 108) [11] is used for speech feature extraction. Once the cepstral coefficients are obtained for recognition, cepstral mean and variance normalization (CMVN) is applied to strengthen the ASR system.

DNN-HMM-based acoustic models are used. First, clean models are trained on the Aurora-2 clean training dataset comprising 8440 utterances. Additionally, multi-style acoustic models are also evaluated. These models are obtained from distorted speech features to further strengthen the ASR system against noise. In AURORA2-2C-CT/FT, the corresponding multi-style training datasets are also composed of 8440 utterances and generated from the clean training dataset of Aurora-2. These datasets consist of dual-channel utterances contaminated with the types of noise in test set A (i.e., bus, babble, car and pedestrian street noises) at the SNRs (referred again to the primary channel) of 5 dB, 10 dB, 15 dB and 20 dB plus the clean condition. To train the multi-style acoustic models, training utterances are first processed with each method evaluated in this paper.

In first instance, different GMM-HMM-based acoustic models are trained for the AURORA2-2C-CT and the AURORA2-2C-FT databases. For each set of models, HMMs with 16 states are used to model each of the 11 digits. Additionally, silence is modeled by an HMM with 3 states [10]. The training speech feature vectors are then used to train the resulting 179 different HMM states, which are modeled by a total of 3000 Gaussians. Next, DNNs with 5 hidden layers and 2048 neurons per layer are trained from the alignments resulting from the above GMM-HMM-based ASR systems.

Besides dual-channel VTS feature compensation integrating our novel posterior computation (2-VTS-C), two additional VTS approaches are evaluated for comparison. One of them is the single-channel VTS feature compensation (1-VTS) from [7] applied on the primary channel. The other one is the dual-channel VTS approach of [5] based on a stacked formulation (2-VTS-S). For a fair comparison, all the parameters required by these techniques are computed in the same way, as well as all of them consider the clean speech partial estimate computation of Eq. (3). Furthermore, MVDR (Minimum Variance Distortionless Response) beamforming [12] and the ETSI advanced front-end (AFE) [13] applied on the primary channel are tested as a reference along with the baseline (i.e., when using the noisy speech features from the primary channel). Finally, as in [3], [5], the dual-channel power spectrum enhancement methods MMSN and DCSS are evaluated when used as pre-processing techniques for the different VTS approaches considered. In particular, those techniques are applied to generate an enhanced primary channel to be used instead of the original one for VTS compensation.

B. Experimental Results

The word accuracy results achieved for the different VTS feature compensation approaches along with those obtained for the additional techniques tested are shown in Table I (when using both clean and multi-style acoustic models under both close- and far-talk conditions). These word accuracies are averaged across all types of noise in each test set and SNRs from -5 dB to 20 dB. As can be seen, 2-VTS-C not only outperforms 1-VTS, but also 2-VTS-S under both close- and far-talk conditions as well as by employing either clean or multi-style acoustic models. Besides this, it is noticeable that MVDR beamforming achieves quite poor results when clean acoustic models are employed (while it obtains competitive results under multi-style acoustic modeling due to the minor mismatch between training and test data). This can be ex-

ISBN 978-90-827970-1-5

TABLE I
WORD ACCURACY RESULTS IN TERMS OF PERCENTAGE OBTAINED FOR THE ASSESSED METHODS WHEN USING BOTH CLEAN AND MULTI-STYLE ACOUSTIC MODELS UNDER BOTH CLOSE- AND FAR-TALK CONDITIONS. RESULTS ARE AVERAGED ACROSS ALL TYPES OF NOISE IN EACH TEST SET AND SNRS FROM -5 DB TO 20 DB.

	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**

TABLE II
WORD ACCURACY RESULTS IN TERMS OF PERCENTAGE OBTAINED WHEN USING MMSN AND DCSS AS PRE-PROCESSING TECHNIQUES FOR VTS FEATURE COMPENSATION. BOTH CLEAN AND MULTI-STYLE ACOUSTIC MODELS ARE EMPLOYED UNDER BOTH CLOSE- AND FAR-TALK CONDITIONS. RESULTS ARE AVERAGED ACROSS ALL TYPES OF NOISE IN EACH TEST SET AND SNRS FROM -5 DB TO 20 DB.

	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**

TABLE III
DETAILED WORD ACCURACY RESULTS (IN TERMS OF PERCENTAGE AND FOR DIFFERENT SNR VALUES) OBTAINED FOR THE ASSESSED TECHNIQUES WHEN EMPLOYING MULTI-STYLE ACOUSTIC MODELS UNDER FAR-TALK CONDITIONS. RESULTS ARE AVERAGED ACROSS ALL TYPES OF NOISE IN TEST SETS *A* AND *B*.

SNR (dB)	**Baseline**	**AFE**	**MVDR**	**1-VTS**	**2-VTS-S**	**2-VTS-C**	**MMSN-1**	**MMSN-2S**	**MMSN-2C**	**DCSS-1**	**DCSS-2S**	**DCSS-2C**
-5	43.70	51.40	59.84	55.38	59.29	61.30	63.84	63.68	66.54	64.85	64.66	**67.00**
0	74.41	80.88	85.55	81.62	83.43	84.85	85.91	86.03	87.26	86.47	86.28	**87.41**
5	89.41	93.41	95.14	93.39	94.07	94.42	94.81	94.72	95.28	95.02	94.80	**95.39**
10	95.28	97.43	97.80	97.43	97.56	97.68	97.74	97.78	**97.99**	97.89	97.66	97.96
15	97.22	**98.82**	98.57	98.57	98.55	98.65	98.62	98.59	98.68	98.62	98.49	98.67
20	98.41	**99.33**	98.94	99.01	99.08	99.08	99.02	98.98	99.00	98.97	98.77	98.88
Clean	99.42	**99.58**	96.84	99.38	99.35	99.31	99.22	99.20	99.21	98.99	98.92	98.95
Avg. (-5 to 20)	83.07	86.88	89.30	87.57	88.66	89.33	89.99	89.96	90.79	90.30	90.11	**90.88**

plained because of the limitations of the classical beamforming methods in this scenario (i.e., only two microphones are available and one of them is placed in an acoustic shadow with respect to the speaker's mouth) as mentioned in [14], [15].

Table II shows the results obtained when MMSN and DCSS are used as pre-processing techniques for VTS feature compensation. In combination with these pre-processing techniques, 2-VTS-C outperforms 1-VTS and 2-VTS-S with either clean or multi-style acoustic models under both close- and far-talk conditions. In this respect, it is interesting to note that the new way of computing the posteriors has allowed to overcome the constraints of the stacked formulation when combined with MMSN and DCSS in far-talk conditions (where 1-VTS provides slightly better performance than 2-VTS-S). This is of particular importance since the far-talk scenario is of special interest for ASR with mobile devices. For this reason, Table III details the word accuracy results achieved under far-talk conditions for the different methods tested when multi-style acoustic models are employed. These results are

ISBN 978-90-827970-1-5

broken down by SNR and averaged across all types of noise in test sets *A* and *B*. From Table III we can observe that while 2-VTS-C exhibits a very good performance over the whole SNR range considered (when applied either isolatedly or jointly with MMSN/DCSS), it particularly stands out at low SNRs. This is a remarkable result, since mobile devices are often used in highly noisy environments such as crowded streets or other public venues. Finally, it is worth noting that AFE achieves the best results at higher SNRs at the expense of a modest performance at lower SNRs.

V. Conclusions

In this work we have proposed a novel posterior computation approach for improved dual-channel VTS feature compensation for mobile devices. Accurate posteriors have been obtained by explicitly modeling the conditional dependence of the noisy secondary channel given the primary one. Thus, especially under far-talk conditions, noticeable improvements in terms of recognition accuracy have been achieved with respect to using a dual-channel VTS approach based on a stacked formulation. As future work, we will research on how to exploit the dual-channel information for better clean speech partial estimate computation.

References

[1] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, "An overview of noise-robust automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, pp. 745–777, 2014.

[2] I. A. McCowan, A. Morris, and H. Bourlard, "Improving speech recognition performance of small microphone arrays using missing data techniques," in *Proc. of ICSLP 2002 – 7th International Conference of Spoken Language Processing, September 16–20, Denver, USA*, 2002, pp. 2181–2184.

[3] I. López-Espejo, A. M. Gomez, J. A. González, and A. M. Peinado, "Feature enhancement for robust speech recognition on smartphones with dual-microphone," in *Proc. of EUSIPCO 2014 – 22nd European Signal Processing Conference, September 1–5, Lisbon, Portugal*, 2014, pp. 21–25.

[4] I. López-Espejo, J. A. González, A. M. Gomez, and A. M. Peinado, "A deep neural network approach for missing-data mask estimation on dual-microphone smartphones: Application to noise-robust speech recognition," *Lecture Notes in Computer Science*, vol. 8854, pp. 119–128, 2014.

[5] I. López-Espejo, A. M. Peinado, A. M. Gomez, and J. A. Gonzalez, "Dual-channel VTS feature compensation for noise-robust speech recognition on mobile devices," *IET Signal Processing*, vol. 11, pp. 17–25, 2017.

[6] A. Acero, L. Deng, T. Kristjansson, and J. Zhang, "HMM adaptation using vector Taylor series for noisy speech recognition," in *Proc. of ICSLP 2000 – 6th International Conference of Spoken Language Processing, October 16–20, Beijing, China*, 2000, pp. 229–232.

[7] J. C. Segura, A. Torre, M. C. Benitez, and A. M. Peinado, "Model-based compensation of the additive noise for continuous speech recognition. Experiments using the AURORA II database and tasks," in *Proc. of EUROSPEECH 2001 – 7th European Conference on Speech Communication and Technology, September 3–7, Aalborg, Denmark*, 2001.

[8] P. J. Moreno, B. Raj, and R. M. Stern, "A vector Taylor series approach for environment-independent speech recognition," in *Proc. of ICASSP 1996 – 21st International Conference on Acoustics, Speech, and Signal Processing, May 7–10, Atlanta, USA*, 1996, pp. 733–736.

[9] F. Faubel, J. McDonough, and D. Klakow, "On expectation maximization based channel and noise estimation beyond the vector Taylor series expansion," in *Proc. of ICASSP 2010 – 35th International Conference on Acoustics, Speech, and Signal Processing, March 14–19, Dallas, USA*, 2010.

[10] D. Pearce and H. G. Hirsch, "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proc. of ICSLP 2000 – 6th International Conference of Spoken Language Processing, October 16–20, Beijing, China*, 2000, pp. 29–32.

[11] "ETSI ES 201 108 - Distributed speech recognition; Front-end feature extraction algorithm; Compression algorithms," .

[12] X. Mestre and M. Á. Lagunas, "On diagonal loading for minimum variance beamformers," in *Proc. of ISSPIT 2003 – 3rd International Symposium on Signal Processing and Information Technology, Darmstadt, Germany*, 2003, pp. 459–462.

[13] "ETSI ES 202 050 - Distributed speech recognition; Advanced front-end feature extraction algorithm; Compression algorithms," .

[14] I. Tashev, S. Mihov, T. Gleghorn, and A. Acero, "Sound capture system and spatial filter for small devices," in *Proc. of EUROSPEECH 2008 – 9th Annual Conference of the International Speech Communication Association, September 22–26, Brisbane, Australia*, 2008, pp. 435–438.

[15] I. Tashev, M. Seltzer, and A. Acero, "Microphone array for headset with spatial noise suppressor," in *Proc. of IWAENC 2005 – 9th International Workshop on Acoustic, Echo and Noise Control*, 2005.

ISBN 978-90-827970-1-5