Dual-Channel VTS Feature Compensation for Noise-Robust Speech Recognition on Mobile Devices

Iván López-Espejo[1,*], Antonio M. Peinado[1], Angel M. Gomez[1], José A. González[2]

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

[2]Dept. of Computer Science, University of Sheffield, Sheffield, UK

[*]iloes@ugr.es

Abstract: One way to improve automatic speech recognition (ASR) accuracy on the latest mobile devices, which are employed on a variety of noisy environments, consists of taking advantage of the small microphone arrays embedded in them. Since the performance of the classic beamforming techniques with small microphone arrays is rather limited, specific techniques are being developed to efficiently exploit this novel feature for noise-robust ASR purposes. In this paper, a novel dual-channel minimum mean square error (MMSE)-based feature compensation method relying on a vector Taylor series (VTS) expansion of a dual-channel speech distortion model is proposed. In contrast to the single-channel VTS approach (which can be considered as the state-of-the-art for feature compensation), our technique particularly benefits from the spatial properties of speech and noise. Our proposal is assessed on a dual-microphone smartphone (a particular case of interest) by means of the AURORA2-2C corpus. Word recognition results demonstrate the higher accuracy of our method by clearly outperforming minimum variance distortionless response (MVDR) beamforming and a single-channel VTS feature compensation approach, especially at low signal-to-noise ratios (SNRs).

1. Introduction

Automatic speech recognition (ASR) technology has recently increased its popularity due to the proliferation of speech-enabled services accessible from mobile devices (e.g. smartphones or tablets). As such devices can be used under many different acoustic scenarios, it is crucial to deal with the noise that may contaminate the speech [1, 2]. One way to improve ASR performance on noisy environments is to exploit the small microphone arrays (i.e. microphone arrays comprising a few microphones) that are being embedded in this kind of devices.

So far, most of the work done on exploiting this new feature has been for speech enhancement purposes [3–7]. In [3, 4], spectral gain masks are computed to perform speech enhancement on dual-microphone smartphones. Such masks are estimated by exploiting the power level difference (PLD) between the two microphones in the device when used in close-talk position (i.e. when the loudspeaker of the smartphone is placed at the ear of the user). In such a position, it is reasonable to assume that the clean speech power spectral density (PSD) at the primary sensor is greater than at the secondary one while approximately the same noise PSD is observed by both sensors [8]. In [5], an inter-microphone noisy speech PSD relation, similar to the PLD, is used to compute a speech presence probability (SPP) which is applied to the estimation of a spatial noise correlation matrix. Such a matrix is then used in a minimum variance distortionless response (MVDR) filter applied to enhance the noisy speech captured by a dual-microphone smartphone.

As shown in recent works [9, 10], ASR can also benefit from these small microphone arrays to further improve the recognizer performance on mobile devices. In [9], a missing-data mask estimation for dual-microphone smartphones based on deep neural network (DNN) and PLD is proposed. The resulting masks are employed to carry out spectral imputation on the noisy speech spectrogram obtained by the primary sensor of the device. In [10], two dual-channel power spectrum enhancement techniques are developed for dual-microphone smartphones as well. While one of these techniques is based on MVDR and the other on spectral subtraction (SS), both of them exploit clean speech and noise spatial correlations in order to provide more accurate estimates with respect to related single-channel approaches. Such specific solutions are necessary since a poor performance of the classic beamforming techniques can be expected in this context. The main reason for this is twofold: the small number of microphones in the device and the possible existence of sensors placed in an acoustic shadow regarding the speaker's mouth (e.g. a sensor located at the rear of a smartphone which faces backwards) [11, 12].

In this paper we propose a novel dual-channel minimum mean square error (MMSE)-based feature compensation method for noise-robust ASR on mobile devices. This method follows the well-known vector Taylor series (VTS) approach widely used to perform single-channel feature compensation (e.g., [13–15]). Our proposal is based on a stacked formulation that is able to exploit the clean speech and noise spatial correlations between the two channels of the device. As will be shown, this approach achieves more accurate clean speech estimates than a single-channel VTS scheme. Experiments are carried out on a dual-microphone smartphone in close-talk conditions. To do so, we use the AURORA2-2C (Aurora-2 - 2 Channels - Conversational Position) database [10], which is an extension to the well-known Aurora-2 corpus [16]. Word recognition results show that our dual-channel proposal greatly improves a single-channel VTS feature compensation approach as well as other state-of-the-art single- and dual-channel techniques (such as MVDR beamforming), especially at low signal-to-noise ratios (SNRs).

The rest of the paper has been organized as follows. In Section 2, the considered dual-channel distortion model is briefly introduced. Section 3 is devoted to the formulation of our dual-channel VTS feature compensation method. The experimental framework and results are shown in Section 4. Finally, in Section 5 conclusions and future work are presented.

2. Dual-Channel Distortion Model

We will consider a noisy speech signal $y_i(m)$ that consists of the sum of a clean speech signal $x_i(m)$ plus a noise signal $n_i(m)$, i.e. $y_i(m) = x_i(m) + n_i(m)$. In this additive noise distortion model i indicates the microphone that captures $y_i(m)$. Thus, $i = 1$ refers to the primary microphone often located at the bottom of the mobile device while $i = 2$ is the secondary microphone at its top or rear. Let us assume independence between speech and noise such that this additive noise distortion model can be expressed in the Mel power spectral domain as

$$|Y_i(f,t)|^2 = |X_i(f,t)|^2 + |N_i(f,t)|^2, \tag{1}$$

where $|Y_i(f,t)|^2$, $|X_i(f,t)|^2$ and $|N_i(f,t)|^2$ denote, respectively, noisy speech, clean speech and noise Mel power spectral bins from the i-th channel. Furthermore, $f = 0, 1, ..., \mathcal{M} - 1$ and $t = 0, 1, ..., T - 1$ indicate the frequency bin and time frame index, respectively. From these quantities we define the following $\mathcal{M} \times 1$ vectors:

$$\mathbf{y}_i = \left(\log |Y_i(0,t)|^2, ..., \log |Y_i(\mathcal{M}-1,t)|^2\right)^{\mathrm{T}}, \tag{2}$$

$$\mathbf{x}_i = \left(\log |X_i(0,t)|^2, ..., \log |X_i(\mathcal{M}-1,t)|^2\right)^{\mathrm{T}}, \tag{3}$$

$$\mathbf{n}_i = \left(\log |N_i(0,t)|^2, ..., \log |N_i(\mathcal{M}-1,t)|^2\right)^{\mathrm{T}}, \tag{4}$$

where the time frame index t has been omitted in the new variables $\mathbf{y}_i$, $\mathbf{x}_i$ and $\mathbf{n}_i$ for the sake of clarity. Thus, from (2)-(4), the speech distortion model in (1) is expressed in the log-Mel power spectral domain as [13, 15, 17],

$$\mathbf{y}_i = \log\left(e^{\mathbf{x}_i} + e^{\mathbf{n}_i}\right), \tag{5}$$

where the operators $\log(\cdot)$ and $e^{(\cdot)}$ are applied element-wise.

Besides the additive noise, we must also consider the acoustics involved in our problem. Thus, we assume that the clean speech signal $x_i(m)$ is the result of filtering the clean source $x(m)$ by the acoustics $h_i(m)$ that affect sensor i, that is, $x_i(m) = h_i(m) * x(m)$ or, in terms of log-Mel power spectra,

$$\begin{aligned} \mathbf{x}_i &= \mathbf{h}_i + \mathbf{x} \\ &= \mathbf{a}_{i1} + \mathbf{x}_1, \end{aligned} \tag{6}$$

where $\mathbf{h}_i$ and $\mathbf{x}$ are vectors of size $\mathcal{M}$ defined as (2)-(4) and $\mathbf{a}_{i1} = \mathbf{h}_i - \mathbf{h}_1$, $i = 1, 2$, represents the clean speech acoustic path from the source to sensor i relative to that of the primary sensor. While $\mathbf{a}_{11} = \mathbf{0}_{\mathcal{M},1}$ is an $\mathcal{M}$-dimensional zero vector by definition, $\mathbf{a}_{21}$ will be referred to as the relative acoustic path (RAP) vector.

For speech recognition purposes, we are interested in the estimation of the clean speech feature vector $\mathbf{x}_1$ derived from the signal captured by the primary microphone. This is a reasonable choice since a clear line of sight between the source (i.e. speaker's mouth) and the primary microphone can be assumed. Hence, we can expect that the primary signal $y_1(m)$ is less (or equally, in the worst case) affected by the noise than the secondary one, $y_2(m)$.

Under the described framework, we can estimate the clean speech feature vector $\mathbf{x}$ in two steps. First, $\mathbf{x}_1$ will be obtained by means of a dual-channel VTS estimation that benefits from the dual-channel noisy observation. This novel method is formulated in the next section. Then, $\mathbf{x}$ can be estimated through the application of channel deconvolution on the clean speech estimate $\hat{\mathbf{x}}_1$. For simplicity, in this work $h_1(m)$ is compensated by performing cepstral mean normalization (CMN) [18] both on training and test data. This way, we are able to cancel or mitigate the possible channel mismatch between training and test data.

3. Dual-Channel VTS Feature Compensation

In this section we develop an MMSE estimator for $\mathbf{x}_1$ that exploits the dual-channel noisy observations and relies on a VTS expansion of the dual-channel speech distortion model introduced in the previous section. Through this approach, the noisy speech statistics, needed for the MMSE estimation, are easily derived in an analytical way from clean speech, relative acoustic path (RAP) and noise statistics. Our proposal, that performs on a frame-by-frame basis, follows a stacked form that exploits clean speech and noise correlations across the two available channels.

First, we assume that the clean speech statistics at the primary channel can be accurately modeled using a $\mathcal{K}$-component Gaussian mixture model (GMM):

$$p(\mathbf{x}_1) = \sum_{k=1}^{\mathcal{K}} P(k)\mathcal{N}\left(\mathbf{x}_1 \middle| \boldsymbol{\mu}_{x_1}^{(k)}, \boldsymbol{\Sigma}_{x_1}^{(k)}\right), \tag{7}$$

where $P(k)$ is the prior probability of the k-th multivariate Gaussian component $\mathcal{N}(\cdot)$ with mean vector and covariance matrix $\boldsymbol{\mu}_{x_1}^{(k)}$ and $\boldsymbol{\Sigma}_{x_1}^{(k)}$, respectively. By considering this speech model, the log-Mel clean speech features will be estimated at every time frame t under an MMSE approach as [19],

$$\hat{\mathbf{x}}_1 = \sum_{k=1}^{\mathcal{K}} P(k|\mathbf{y})\mathrm{E}\left[\mathbf{x}_1|\mathbf{y},k\right], \tag{8}$$

where $\mathbf{y}$ is a stacked vector defined as

$$\mathbf{y} = \begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{pmatrix}, \tag{9}$$

and the k-th clean speech partial estimate $\mathrm{E}\left[\mathbf{x}_1|\mathbf{y},k\right]$ is weighted by the posterior $P(k|\mathbf{y})$ to be linearly combined. In the following subsection the estimation of the posteriors $\{P(k|\mathbf{y});\ k = 1, 2, ..., \mathcal{K}\}$ is addressed while the computation of the clean speech partial estimates is detailed in Subsection 3.2.

3.1. Estimation of the Posterior Probabilities

Let us rewrite the speech distortion model of Eq. (5) by taking into account the relationship in (6) as

$$\begin{aligned} \mathbf{y}_i \;=\; \mathbf{f}(\mathbf{x}_1, \mathbf{a}_{i1}, \mathbf{n}_i) \;&=\; \log\left(e^{\mathbf{a}_{i1}+\mathbf{x}_1} + e^{\mathbf{n}_i}\right) \\ &=\; \mathbf{x}_1 + \mathbf{a}_{i1} + \log(\mathbf{1}_{\mathcal{M},1} + e^{\mathbf{n}_i - \mathbf{x}_1 - \mathbf{a}_{i1}}), \end{aligned} \tag{10}$$

where $\mathbf{f}(\mathbf{x}_1, \mathbf{a}_{i1}, \mathbf{n}_i) : \mathbb{R}^{\mathcal{M}\times\mathcal{M}\times\mathcal{M}} \to \mathbb{R}^{\mathcal{M}}$ and $\mathbf{y}_i$, $\mathbf{x}_1$, $\mathbf{a}_{i1}$ and $\mathbf{n}_i$ are the log-Mel feature vectors at time frame t introduced in the previous section, and $\mathbf{1}_{\mathcal{M},1}$ is an $\mathcal{M}$-dimensional vector filled with ones.

From now on, let $\mathbf{a} = (\mathbf{a}_{11}^{\mathrm{T}}, \mathbf{a}_{21}^{\mathrm{T}})^{\mathrm{T}} = (\mathbf{0}_{\mathcal{M},1}^{\mathrm{T}}, \mathbf{a}_{21}^{\mathrm{T}})^{\mathrm{T}}$ and $\mathbf{n} = (\mathbf{n}_1^{\mathrm{T}}, \mathbf{n}_2^{\mathrm{T}})^{\mathrm{T}}$ be an augmented RAP vector and a stacked vector of noise, respectively, both of them $2\mathcal{M}$-dimensional. By taking into account the couple of sensors in the mobile device, the considered dual-channel distortion model is given by the following stacked vector:

$$\mathbf{y} = \begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{pmatrix} = \mathbf{F}\left(\mathbf{x}_1, \mathbf{a}, \mathbf{n}\right) = \begin{pmatrix} \mathbf{f}(\mathbf{x}_1, \mathbf{a}_{11}, \mathbf{n}_1) \\ \mathbf{f}(\mathbf{x}_1, \mathbf{a}_{21}, \mathbf{n}_2) \end{pmatrix}, \tag{11}$$

where $\mathbf{F}(\mathbf{x}_1, \mathbf{a}, \mathbf{n}) : \mathbb{R}^{\mathcal{M}\times 2\mathcal{M}\times 2\mathcal{M}} \rightarrow \mathbb{R}^{2\mathcal{M}}$. We assumed in (7) that the clean speech statistics at the primary channel are modeled by means of a $\mathcal{K}$-component GMM. To complete the generative model, we assume that the statistics for both the RAP and noise in each channel can be modeled by Gaussian distributions [13, 20], i.e. $p(\mathbf{a}_{21}) = \mathcal{N}\left(\mathbf{a}_{21} \left| \boldsymbol{\mu}_{a_{21}}, \boldsymbol{\Sigma}_{a_{21}}\right.\right)$ and $p(\mathbf{n}_i) = \mathcal{N}\left(\mathbf{n}_i \left| \boldsymbol{\mu}_{n_i}, \boldsymbol{\Sigma}_{n_i}\right.\right)$ $(i = 1, 2)$, respectively. Since any linear combination of Gaussian variables follows another Gaussian distribution [21], by linearizing the dual-channel distortion model in (11) we are able to describe the dual-channel noisy speech statistics (required to compute the posteriors $\{P(k|\mathbf{y}); k = 1, 2, ..., \mathcal{K}\}$) by means of a GMM (at every time frame t) as

$$p(\mathbf{y}) = \sum_{k=1}^{\mathcal{K}} P(k) \mathcal{N}\left(\mathbf{y} \left| \boldsymbol{\mu}_y^{(k)}, \boldsymbol{\Sigma}_y^{(k)}\right.\right). \tag{12}$$

Then, we linearize $\mathbf{y} = \mathbf{F}(\mathbf{x}_1, \mathbf{a}, \mathbf{n})$ by means of a first-order VTS expansion around the point $\left(\boldsymbol{\mu}_{x_1}^{(k)}, \boldsymbol{\mu}_a, \boldsymbol{\mu}_n\right)$, where

$$\boldsymbol{\mu}_a = \begin{pmatrix} \boldsymbol{\mu}_{a_{11}} \\ \boldsymbol{\mu}_{a_{21}} \end{pmatrix} = \begin{pmatrix} \mathbf{0}_{\mathcal{M},1} \\ \boldsymbol{\mu}_{a_{21}} \end{pmatrix} \tag{13}$$

and

$$\boldsymbol{\mu}_n = \begin{pmatrix} \boldsymbol{\mu}_{n_1} \\ \boldsymbol{\mu}_{n_2} \end{pmatrix} \tag{14}$$

are $2\mathcal{M} \times 1$ vectors of stacked means. This procedure is accomplished by accordingly linearizing the speech distortion model for each channel, $\mathbf{f}(\mathbf{x}_1, \mathbf{a}_{i1}, \mathbf{n}_i)$ $(i = 1, 2)$, around the point $\left(\boldsymbol{\mu}_{x_1}^{(k)}, \boldsymbol{\mu}_{a_{i1}}, \boldsymbol{\mu}_{n_i}\right)$, that is,

$$\begin{aligned} \mathbf{f}(\mathbf{x}_1, \mathbf{a}_{i1}, \mathbf{n}_i) \approx\ & \mathbf{f}\left(\boldsymbol{\mu}_{x_1}^{(k)}, \boldsymbol{\mu}_{a_{i1}}, \boldsymbol{\mu}_{n_i}\right) + \mathbf{J}_x^{(i,k)}\left(\mathbf{x}_1 - \boldsymbol{\mu}_{x_1}^{(k)}\right) + \\ & + \mathbf{J}_a^{(i,k)}\left(\mathbf{a}_{i1} - \boldsymbol{\mu}_{a_{i1}}\right) + \mathbf{J}_n^{(i,k)}\left(\mathbf{n}_i - \boldsymbol{\mu}_{n_i}\right), \end{aligned} \tag{15}$$

where $\mathbf{J}_x^{(i,k)}$, $\mathbf{J}_a^{(i,k)}$ and $\mathbf{J}_n^{(i,k)}$ are $\mathcal{M} \times \mathcal{M}$ Jacobian matrices, the calculation of which will be detailed later.

To finally characterize the probability density function (PDF) $p(\mathbf{y})$ we need to derive its mean vectors and covariance matrices. By taking into account (11) and (15), it is straightforward to show that the mean vectors $\left\{\boldsymbol{\mu}_y^{(k)} : k = 1, 2, ..., \mathcal{K}\right\}$ can be obtained as

$$\boldsymbol{\mu}_y^{(k)} = \begin{pmatrix} \mathrm{E}\left[\mathbf{y}_1|k\right] \\ \mathrm{E}\left[\mathbf{y}_2|k\right] \end{pmatrix} = \begin{pmatrix} \mathbf{f}\left(\boldsymbol{\mu}_{x_1}^{(k)}, \boldsymbol{\mu}_{a_{11}}, \boldsymbol{\mu}_{n_1}\right) \\ \mathbf{f}\left(\boldsymbol{\mu}_{x_1}^{(k)}, \boldsymbol{\mu}_{a_{21}}, \boldsymbol{\mu}_{n_2}\right) \end{pmatrix}. \tag{16}$$

On the other hand, the covariance matrices can be easily calculated in accordance to their definition as

$$\boldsymbol{\Sigma}_y^{(k)} = \mathrm{E}\left[\left(\mathbf{y} - \boldsymbol{\mu}_y^{(k)}\right)\left(\mathbf{y} - \boldsymbol{\mu}_y^{(k)}\right)^{\mathrm{T}}\right], \tag{17}$$

where $\mathbf{y} - \boldsymbol{\mu}_y^{(k)}$ is defined in the following manner by again considering the approximation in (15) as well as (16):

$$\mathbf{y} - \boldsymbol{\mu}_y^{(k)} = \begin{pmatrix} \mathbf{J}_x^{(1,k)}\left(\mathbf{x}_1 - \boldsymbol{\mu}_{x_1}^{(k)}\right) + \mathbf{J}_a^{(1,k)}\left(\mathbf{a}_{11} - \boldsymbol{\mu}_{a_{11}}\right) + \mathbf{J}_n^{(1,k)}\left(\mathbf{n}_1 - \boldsymbol{\mu}_{n_1}\right) \\ \mathbf{J}_x^{(2,k)}\left(\mathbf{x}_1 - \boldsymbol{\mu}_{x_1}^{(k)}\right) + \mathbf{J}_a^{(2,k)}\left(\mathbf{a}_{21} - \boldsymbol{\mu}_{a_{21}}\right) + \mathbf{J}_n^{(2,k)}\left(\mathbf{n}_2 - \boldsymbol{\mu}_{n_2}\right) \end{pmatrix}. \tag{18}$$

For notational convenience let us define the following block Jacobian matrices:

$$\mathbf{J}_x^{(k)} = \begin{pmatrix} \mathbf{J}_x^{(1,k)} \\ \mathbf{J}_x^{(2,k)} \end{pmatrix}; \tag{19}$$

$$\mathbf{J}_a^{(k)} = \begin{pmatrix} \mathbf{J}_a^{(1,k)} & \mathbf{0}_{\mathcal{M},\mathcal{M}} \\ \mathbf{0}_{\mathcal{M},\mathcal{M}} & \mathbf{J}_a^{(2,k)} \end{pmatrix}; \tag{20}$$

$$\mathbf{J}_n^{(k)} = \begin{pmatrix} \mathbf{J}_n^{(1,k)} & \mathbf{0}_{\mathcal{M},\mathcal{M}} \\ \mathbf{0}_{\mathcal{M},\mathcal{M}} & \mathbf{J}_n^{(2,k)} \end{pmatrix}, \tag{21}$$

where $\mathbf{J}_x^{(k)}$ is a $2\mathcal{M} \times \mathcal{M}$ matrix, $\mathbf{J}_a^{(k)}$ and $\mathbf{J}_n^{(k)}$ are $2\mathcal{M} \times 2\mathcal{M}$ matrices and $\mathbf{0}_{\mathcal{M},\mathcal{M}}$ is an $\mathcal{M} \times \mathcal{M}$ zero matrix. Then, (18) can be expressed in a more compact form as

$$\mathbf{y} - \boldsymbol{\mu}_y^{(k)} = \mathbf{J}_x^{(k)}\left(\mathbf{x}_1 - \boldsymbol{\mu}_{x_1}^{(k)}\right) + \mathbf{J}_a^{(k)}\left(\mathbf{a} - \boldsymbol{\mu}_a\right) + \mathbf{J}_n^{(k)}\left(\mathbf{n} - \boldsymbol{\mu}_n\right). \tag{22}$$

Finally, by combining (22) and (17), as well as considering independence between clean speech, the RAP and noise, an expression for the dual-channel noisy speech model covariance matrix can be obtained as

$$\boldsymbol{\Sigma}_y^{(k)} = \mathbf{J}_x^{(k)} \boldsymbol{\Sigma}_{x_1}^{(k)} {\mathbf{J}_x^{(k)}}^{\mathrm{T}} + \mathbf{J}_a^{(k)} \boldsymbol{\Sigma}_a {\mathbf{J}_a^{(k)}}^{\mathrm{T}} + \mathbf{J}_n^{(k)} \boldsymbol{\Sigma}_n {\mathbf{J}_n^{(k)}}^{\mathrm{T}}, \tag{23}$$

where

$$\boldsymbol{\Sigma}_a = \mathrm{E}\left[(\mathbf{a} - \boldsymbol{\mu}_a)(\mathbf{a} - \boldsymbol{\mu}_a)^{\mathrm{T}}\right] = \begin{pmatrix} \mathbf{0}_{\mathcal{M},\mathcal{M}} & \mathbf{0}_{\mathcal{M},\mathcal{M}} \\ \mathbf{0}_{\mathcal{M},\mathcal{M}} & \boldsymbol{\Sigma}_{a_{21}} \end{pmatrix} \tag{24}$$

and

$$\boldsymbol{\Sigma}_n = \mathrm{E}\left[(\mathbf{n} - \boldsymbol{\mu}_n)(\mathbf{n} - \boldsymbol{\mu}_n)^{\mathrm{T}}\right] = \begin{pmatrix} \boldsymbol{\Sigma}_{n_1} & \boldsymbol{\Sigma}_{n_{12}} \\ \boldsymbol{\Sigma}_{n_{21}} & \boldsymbol{\Sigma}_{n_2} \end{pmatrix} \tag{25}$$

are $2\mathcal{M} \times 2\mathcal{M}$ spatial covariance matrices of the RAP and noise, respectively. In addition, $\boldsymbol{\Sigma}_{n_{12}} = \boldsymbol{\Sigma}_{n_{21}}^{\mathrm{T}} = \mathrm{E}\left[(\mathbf{n}_1 - \boldsymbol{\mu}_{n_1})(\mathbf{n}_2 - \boldsymbol{\mu}_{n_2})^{\mathrm{T}}\right]$. The Jacobian matrices, which are diagonal in accordance to the speech distortion model described by Eq. (10) (independent frequency components), are easily calculated by employing the Jacobian matrix mathematical definition as,

$$\begin{aligned}
\mathbf{J}_x^{(i,k)} &= \left.\frac{\partial \mathbf{y}_i}{\partial \mathbf{x}_1}\right|_{\boldsymbol{\mu}_{x_1}^{(k)}, \boldsymbol{\mu}_{a_{i1}}, \boldsymbol{\mu}_{n_i}} = \mathrm{diag}\left(\frac{\mathbf{1}_{\mathcal{M},1}}{\mathbf{1}_{\mathcal{M},1} + e^{\boldsymbol{\mu}_{n_i} - \boldsymbol{\mu}_{x_1}^{(k)} - \boldsymbol{\mu}_{a_{i1}}}}\right); \\
\mathbf{J}_a^{(i,k)} &= \left.\frac{\partial \mathbf{y}_i}{\partial \mathbf{a}_{i1}}\right|_{\boldsymbol{\mu}_{x_1}^{(k)}, \boldsymbol{\mu}_{a_{i1}}, \boldsymbol{\mu}_{n_i}} = \begin{cases} \mathbf{0}_{\mathcal{M},\mathcal{M}} & \text{if } i = 1 \\ \mathbf{J}_x^{(2,k)} & \text{if } i = 2 \end{cases}; \\
\mathbf{J}_n^{(i,k)} &= \left.\frac{\partial \mathbf{y}_i}{\partial \mathbf{n}_i}\right|_{\boldsymbol{\mu}_{x_1}^{(k)}, \boldsymbol{\mu}_{a_{i1}}, \boldsymbol{\mu}_{n_i}} = \mathbf{I}_{\mathcal{M}} - \mathbf{J}_x^{(i,k)},
\end{aligned} \tag{26}$$

where diag$(\cdot)$ indicates a diagonal matrix whose main diagonal corresponds to its argument, division $\div$ operates element-wise and $\mathbf{I}_{\mathcal{M}}$ is an $\mathcal{M} \times \mathcal{M}$ identity matrix.

Finally, by using the Bayes' rule and the previous derivations, in the knowledge that $p(\mathbf{y}|k) = \mathcal{N}\left(\mathbf{y} \,\middle|\, \boldsymbol{\mu}_y^{(k)}, \boldsymbol{\Sigma}_y^{(k)}\right)$, the posteriors are obtained as

$$P(k|\mathbf{y}) = \frac{p(\mathbf{y}|k)P(k)}{\sum_{k'=1}^{\mathcal{K}} p(\mathbf{y}|k')P(k')}, \qquad k = 1, 2, ..., \mathcal{K}. \tag{27}$$

The computation of the parameters of the PDFs $p(\mathbf{x}_1)$, $p(\mathbf{a}_{21})$ and $p(\mathbf{n}_i)$, $i = 1, 2$, along with $\mathbf{\Sigma}_{n_{12}}$ required to perform the calculations above is detailed in Subsection 4.2.

3.2. Clean Speech Partial Estimate Computation

While the calculation of the partial expected values in (8), $\mathrm{E}\left[\mathbf{x}_1|\mathbf{y}, k\right]$ $(k = 1, 2, ..., \mathcal{K})$, is defined as

$$\mathrm{E}\left[\mathbf{x}_1|\mathbf{y}, k\right] = \int \mathbf{x}_1 p(\mathbf{x}_1|\mathbf{y}, k) d\mathbf{x}_1, \tag{28}$$

it is again necessary to linearize the non-linear speech distortion model of (10) to make inference feasible. In this regard, two different proposals for VTS feature compensation are considered in this paper.

In the first approach, that will be referred to as 2-VTS-a, we exploit the dual-channel information. If we assume that the joint PDF $p(\mathbf{x}_1, \mathbf{y}|k)$ is Gaussian then the conditional PDF $p(\mathbf{x}_1|\mathbf{y}, k)$ will also be Gaussian, so that the expected value of $p(\mathbf{x}_1|\mathbf{y}, k)$, $\mathrm{E}\left[\mathbf{x}_1|\mathbf{y}, k\right]$, can be approximated as [22]

$$\mathrm{E}\left[\mathbf{x}_1|\mathbf{y}, k\right] = \boldsymbol{\mu}_{x_1}^{(k)} + \mathbf{\Sigma}_{x_1 y}^{(k)} \mathbf{\Sigma}_{y}^{(k)^{-1}} \left(\mathbf{y} - \boldsymbol{\mu}_{y}^{(k)}\right), \tag{29}$$

where the cross-covariance matrix $\mathbf{\Sigma}_{x_1 y}^{(k)}$ is approximated by again considering a VTS approach. Thus, by using the result in (22),

$$\mathbf{\Sigma}_{x_1 y}^{(k)} = \mathrm{E}\left[\left(\mathbf{x}_1 - \boldsymbol{\mu}_{x_1}^{(k)}\right)\left(\mathbf{y} - \boldsymbol{\mu}_{y}^{(k)}\right)^{\mathrm{T}}\right] = \mathbf{\Sigma}_{x_1}^{(k)} \mathbf{J}_{x}^{(k)^{\mathrm{T}}}, \tag{30}$$

where it should be reminded that independence between clean speech, the RAP and noise was assumed.

In the second approach, only the information from the main channel is used to compute the clean speech partial estimates. For this second strategy, which is referred to as 2-VTS-b, Eq. (10) is rewritten as $\mathbf{y}_i = \mathbf{x}_1 + \mathbf{g}(\mathbf{x}_1, \mathbf{a}_{i1}, \mathbf{n}_i)$ [14,15], where $\mathbf{g}(\mathbf{x}_1, \mathbf{a}_{i1}, \mathbf{n}_i) = \mathbf{a}_{i1} + \log(\mathbf{1}_{\mathcal{M},1} + e^{\mathbf{n}_i - \mathbf{x}_1 - \mathbf{a}_{i1}})$ is a distortion vector. Then, the k-th clean speech partial estimate is calculated as

$$\mathrm{E}\left[\mathbf{x}_1|\mathbf{y}, k\right] \approx \mathrm{E}\left[\mathbf{x}_1|\mathbf{y}_1, k\right] = \mathbf{y}_1 - \mathrm{E}[\mathbf{g}(\mathbf{x}_1, \mathbf{a}_{11}, \mathbf{n}_1)|\mathbf{y}_1, k], \tag{31}$$

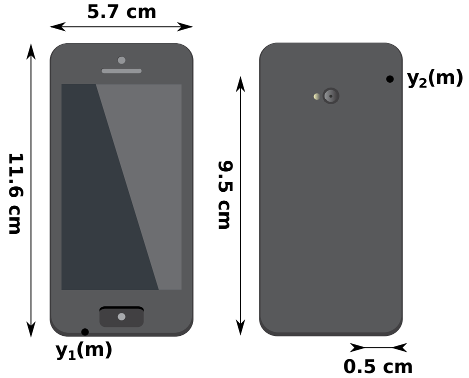

Fig. 1. *Characteristics of the mobile device used for the generation of the AURORA2-2C database.*

where it is assumed that the function $\mathbf{g}(\mathbf{x}_1, \mathbf{a}_{i1}, \mathbf{n}_i)$ is smooth for each k such that [14, 15]

$$\mathrm{E}[\mathbf{g}(\mathbf{x}_1, \mathbf{a}_{11}, \mathbf{n}_1)|\mathbf{y}_1, k] = \mathbf{g}\left(\boldsymbol{\mu}_{x_1}^{(k)}, \boldsymbol{\mu}_{a_{11}}, \boldsymbol{\mu}_{n_1}\right). \tag{32}$$

Considering 2-VTS-b instead of 2-VTS-a may be appropriate in our case, since, generally, the secondary microphone of the mobile device captures a much noisier signal than the primary one.

4. Experimental Evaluation

4.1. Experimental Settings

Our dual-channel feature compensation method is evaluated in terms of word accuracy on the AURORA2-2C (Aurora-2 - 2 Channels - Conversational Position) database [10]. This is a synthetic dual-channel noisy speech database created from the well-known Aurora-2 corpus [16]. The AURORA2-2C database emulates the recording of dual-channel noisy speech data by means of a dual-microphone smartphone in close-talk conditions (i.e. the loudspeaker of the device is placed at the ear of the user). Figure 1 depicts the geometrical characteristics of the mobile device used to generate this corpus. This database has two test sets: *A* and *B*. Following the same Aurora-2 structure, utterances in test set *A* are distorted by bus, babble, car and pedestrian street noises, while those in test set *B* are contaminated by café, street, bus station and train station noises. The SNRs considered (referred to the primary channel) for the test sets are -5 dB, 0 dB, 5 dB, 10 dB, 15 dB and 20 dB as well as the clean condition.

The European Telecommunication Standards Institute front-end (ETSI FE, ES 201 108) [23,24] is used to extract acoustic features from the speech signals. 39-dimensional feature vectors (twelve Mel-frequency cepstral coefficients (MFCCs) along with the 0th order coefficient and their respective velocity and acceleration components) are employed by the recognizer. For feature compensation, log-Mel feature vectors of $\mathcal{M} = 23$ components are computed. To obtain the cepstral coefficients for recognition, the discrete cosine transform (DCT) is applied to the enhanced log-Mel feature vectors. Finally, to improve the robustness of the system against channel mismatches, CMN is applied as previously commented.

Two sets of hidden Markov model (HMM)-based acoustic models with GMM state-output distributions are used for evaluation: clean and multi-style acoustic models. While clean models are trained on the Aurora-2 clean training dataset, multi-style models are trained with distorted speech features to strengthen the ASR system against noise. In AURORA2-2C, its multi-style training dataset is created from the 8440 training clean utterances of Aurora-2. Similarly to [16], the multi-style training dataset consists of dual-channel utterances contaminated with the types of noise in test set *A* at the SNRs (referred to the primary channel) of 5 dB, 10 dB, 15 dB and 20 dB as well as the clean condition. To train the multi-style acoustic models, training utterances are first compensated with each method tested in this work. Finally, left to right continuous density HMMs with 16 states and 3 Gaussians per state are used to model each digit for both sets of acoustic models. Silences and short pauses are modeled by HMMs with 3 and 1 states, respectively, and 6 Gaussians per state [16].

4.2. Computation of Prior Statistical Parameters

The GMM defined in (7) to describe the clean speech statistics at the primary channel is comprised of $\mathcal{K} = 256$ multivariate Gaussian components with diagonal covariance matrices. GMM training is performed by means of the expectation-maximization (EM) algorithm on the same dataset as that used for clean acoustic model training in AURORA2-2C.

The parameters of the PDF $p(\mathbf{a}_{21})$, $\boldsymbol{\mu}_{a_{21}}$ and $\mathbf{\Sigma}_{a_{21}}$, are *a priori* computed for the AURORA2-2C. In this work we assume that $p(\mathbf{a}_{21})$ follows a stationary distribution. In this way, we consider

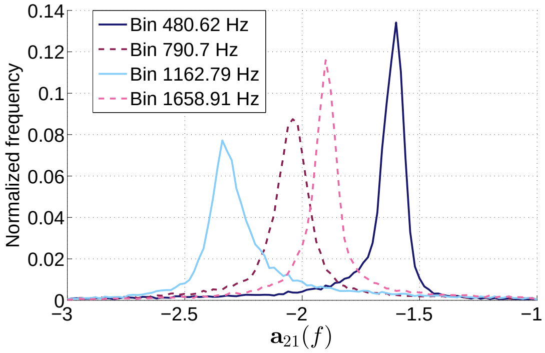

Fig. 2. *Example histograms of* $\mathbf{a}_{21}(f)$ *for four different frequency bins.*

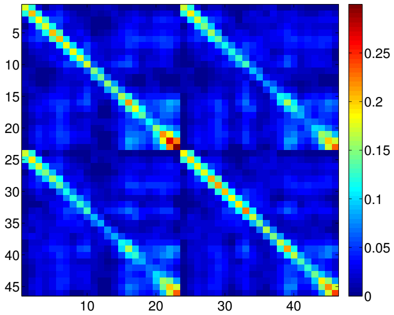

Fig. 3. *Example of noise spatial covariance matrix* $\mathbf{\Sigma}_n$ *estimated from 12 seconds of pedestrian street noise captured with a dual-microphone smartphone.*

that $\mathbf{a}_{21}$ at every time frame t is a realization of the variable. The mean vector $\boldsymbol{\mu}_{a_{21}}$ and the covariance matrix $\mathbf{\Sigma}_{a_{21}}$ are estimated as the sample mean and sample covariance, respectively, from $\mathbf{a}_{21}$ samples. Moreover, we assume independence across frequency bins for $\mathbf{a}_{21}$ and, hence, a diagonal covariance matrix $\mathbf{\Sigma}_{a_{21}}$ is used. We obtain $\mathbf{a}_{21}$ samples from the development dataset of the AURORA2-2C corpus (as $\mathbf{a}_{21} = \mathbf{x}_2 - \mathbf{x}_1$). Figure 2 plots example histograms of $\mathbf{a}_{21}(f)$ at four different frequency bins. These histograms were calculated from the aforementioned development dataset. It is worth noting that, from this figure, the Gaussian assumption seems reasonable.

Concerning noise estimation, we assume that the first and last $\nu = 20$ frames (which corresponds to 200 ms) of each utterance contain only noise energy. In this way, the mean vector of the PDF $p(\mathbf{n}_i)$, $\boldsymbol{\mu}_{n_i}$ $(i = 1, 2)$, is computed for every time frame t from a linear interpolation between the averages of the first and last ν frames in the i-th channel of each utterance in the log-Mel domain [25]. Additionally, the noise covariance matrices $\mathbf{\Sigma}_{n_i}$ $(i = 1, 2)$ and $\mathbf{\Sigma}_{n_{12}}$ are estimated per utterance as the sample covariance of the first and last ν frames as well [25]. Independence across frequency bins is also assumed for the noise so that $\boldsymbol{\Sigma}_{n_i}$ ($i = 1, 2$) and $\boldsymbol{\Sigma}_{n_{12}}$ are diagonal. Figure 3 shows an example of a noise spatial covariance matrix $\boldsymbol{\Sigma}_n$ estimated from 12 seconds of pedestrian street noise captured with a dual-microphone smartphone. This example clearly shows the suitability of the diagonal assumption.

4.3. Results

In addition to our proposals, 2-VTS-a and 2-VTS-b, the dual-channel power spectrum enhancement techniques MMSN and DCSS already proposed in [10] and a time-varying MVDR beamformer with diagonal loading [26] were also evaluated. Furthermore, the advanced front-end (AFE) [24, 27] and a single-channel VTS feature compensation algorithm [13, 14] were also tested on the primary channel as a reference. For single-channel VTS compensation, the two types of clean speech partial estimation described in Subsection 3.2 were evaluated. The corresponding experiments are labeled as 1-VTS-a (where $\mathbf{y}_1$ is considered instead of $\mathbf{y}$) and 1-VTS-b [15]. For a fair comparison, in these experiments the required clean speech GMM as well as the hyperparameters $\boldsymbol{\mu}_{n_1}$ and $\boldsymbol{\Sigma}_{n_1}$ were obtained as in Subsection 4.2. Finally, an ASR system employing noisy speech features from the primary channel after mean subtraction is used as baseline.

First, it should be noticed that the word recognition results obtained for the techniques evaluated in both this paper and [10] are slightly different. This is because the AURORA2-2C database was generated for this paper considering an anechoic chamber instead of a semi-anechoic environment for the acoustic path estimation, as was the case in [10].

Tables 1 and 2 show the word accuracy results achieved on the AURORA2-2C database when clean and multi-style acoustic models are used, respectively. Results are averaged across all types of noise in test sets *A* and *B* as well as broken down by SNR. As expected, better word recognition is generally achieved by considering multi-style instead of clean acoustic models, since the mismatch between training and test data is reduced. Likewise, the results show that the approach for clean speech partial estimate computation that only uses the information from the main channel, VTS-b, provides better accuracy than VTS-a. Additionally, in all cases, the dual-channel VTS

SNR (dB)	Baseline	MVDR	MMSN	DCSS	AFE	1-VTS-a	2-VTS-a	1-VTS-b	2-VTS-b
-5	18.15	20.94	24.16	24.37	35.81	43.06	50.74	44.25	**51.36**
0	31.85	36.66	46.31	46.69	65.46	71.86	78.14	72.75	**78.33**
5	56.11	64.41	74.78	75.06	85.16	89.05	91.74	89.69	**91.78**
10	82.78	87.84	90.66	90.65	93.80	95.23	96.32	95.44	**96.46**
15	94.72	95.99	96.14	96.03	96.96	97.69	98.03	97.71	**98.09**
20	97.76	97.82	97.87	97.64	98.33	98.50	98.54	98.49	**98.61**
Clean	99.13	98.87	98.90	99.13	**99.24**	99.09	99.04	99.09	99.04
Avg. (-5 to 20)	63.56	67.28	71.65	71.74	79.25	82.57	85.59	83.06	**85.77**
Rel. improv.	-	3.72	8.09	8.18	15.69	19.01	22.03	19.50	**22.21**

Table 1. *Word accuracy results (%) obtained on the AURORA2-2C database for different SNR values when using clean acoustic models. Results are averaged across all types of noise in test sets A and B.*

SNR (dB)	Baseline	MVDR	MMSN	DCSS	AFE	1-VTS-a	2-VTS-a	1-VTS-b	2-VTS-b
-5	36.93	46.31	46.70	47.28	48.21	45.29	53.94	47.98	**55.58**
0	66.69	77.99	78.14	78.45	78.36	74.72	81.44	76.07	**81.69**
5	88.85	92.67	92.99	93.12	92.24	91.08	**93.59**	91.22	93.45
10	95.73	96.81	96.86	**96.93**	96.54	95.97	96.86	96.00	96.85
15	97.56	98.01	98.17	98.15	98.11	97.77	**98.27**	97.91	**98.27**
20	98.31	98.53	98.53	98.45	**98.66**	98.33	98.49	98.49	98.61
Clean	98.77	98.59	98.61	98.24	**99.07**	98.84	98.73	98.79	98.82
Avg. (-5 to 20)	80.68	85.05	85.23	85.40	85.35	83.86	87.10	84.61	**87.41**
Rel. improv.	-	4.37	4.55	4.72	4.67	3.18	6.42	3.93	**6.73**

Table 2. *Word accuracy results (%) obtained on the AURORA2-2C database for different SNR values when using multi-style acoustic models. Results are averaged across all types of noise in test sets A and B.*

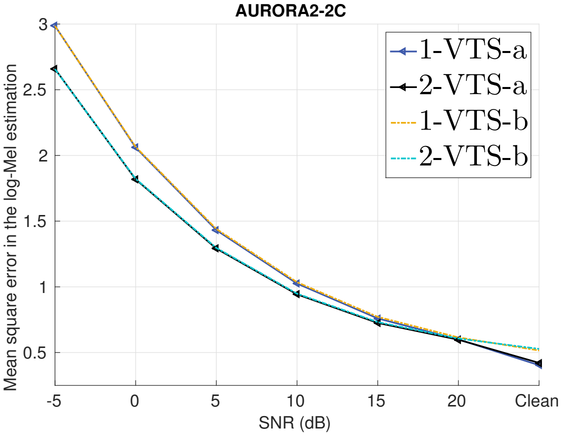

Fig. 4. *Log-Mel estimation mean square errors (for different SNR values) obtained for the VTS approaches evaluated on the AURORA2-2C database. Errors, calculated on a bin-by-bin basis, are averaged across all types of noise in test sets A and B.*

compensation approach outperforms on average the single-channel one. This is expected since the former can exploit the spatial properties of speech and noise signals by means of the RAP vector $\mathbf{a}_{21}$ and the spatial covariance matrix of noise $\mathbf{\Sigma}_n$. These parameters are directly involved in the definition of the noisy speech PDF, $p(\mathbf{y})$, determining more accurately the importance of each clean speech partial estimate in the final estimation of (8) from (27). This is confirmed by Figure 4, which depicts the mean square errors of the different estimates in the log-Mel domain as a function of the SNR.

As can be seen, 2-VTS-b is, on average, our best approach when using both clean and multi-style acoustic models. Moreover, in all cases, 2-VTS-b has the best performance at the SNRs of -5 dB and 0 dB, making it a suitable approach for challenging low-SNR environments (as those where mobile devices may be used). It is also worth mentioning that MVDR beamforming does not achieve a competitive performance in comparison with AFE and VTS when clean acoustic models are considered. This was expected since, as mentioned before, the efficiency of the classic beamforming techniques with small microphone arrays is rather limited [11, 12] (only two sensors are available in our framework and one of them is placed in an acoustic shadow regarding the speaker's mouth). That is, it is desirable to consider different noise-robust approaches to be performed on this kind of devices, as those proposed in this work.

Finally, in [10], both MMSN and DCSS were also used as preprocessing techniques for 1-

	Clean models				*Multi-style models*			
SNR (dB)	MMSN-1	MMSN-2	DCSS-1	DCSS-2	MMSN-1	MMSN-2	DCSS-1	DCSS-2
-5	56.14	56.97	56.25	**57.09**	61.58	62.83	61.61	**62.91**
0	81.05	81.58	81.04	**81.69**	84.15	85.05	84.07	**85.12**
5	92.38	**92.92**	92.37	92.86	94.10	**94.48**	94.05	94.47
10	96.55	**96.80**	96.54	96.78	97.08	**97.27**	97.14	97.26
15	98.16	**98.26**	98.13	98.21	98.28	98.36	98.24	**98.37**
20	98.61	**98.62**	98.55	98.59	**98.57**	98.56	98.54	98.51
Clean	**98.94**	98.90	98.78	98.79	**98.82**	98.76	98.73	98.70
Avg. (-5 to 20)	87.15	87.53	87.15	**87.54**	88.96	89.43	88.95	**89.44**
Rel. improv.	23.59	23.97	23.59	**23.98**	8.28	8.75	8.27	**8.76**

Table 3. *Word accuracy results (%) obtained on the AURORA2-2C database when using MMSN and DCSS as preprocessing methods for 1-VTS-b and 2-VTS-b. Clean and multi-style acoustic models are considered. Results are broken down by SNR and averaged across all types of noise in test sets A and B.*

VTS-b, outperforming 1-VTS-b when applied isolatedly. Table 3 reports the word accuracy results obtained when MMSN and DCSS are used as preprocessing techniques for 1-VTS-b (MMSN-1 and DCSS-1) and 2-VTS-b (MMSN-2 and DCSS-2). It should be remarked that, in this case, the enhanced primary spectrum from either MMSN or DCSS is used as input for 1-VTS-b. In addition, the input for 2-VTS-b consists of this enhanced primary spectrum along with the original noisy spectrum from the secondary channel. Once again, clean and multi-style acoustic modeling is considered and results are averaged across all types of noise in test sets *A* and *B* as well as broken down by SNR. While 2-VTS-b clearly outperforms 1-VTS-b when they are applied in isolation, differences become smaller when they are combined with MMSN and DCSS, as can be observed from Table 3. It must be noted that when either MMSN or DCSS is combined with 1-VTS-b, the same spatial information as in the case of 2-VTS-b is being used, since both MMSN and DCSS exploit RAP information and noise spatial correlations. Therefore, it is reasonable to expect small improvements when these are combined with 2-VTS-b. Nevertheless, as can be seen, 2-VTS-b is better able to improve MMSN and DCSS than 1-VTS-b.

5. Conclusions and Future Work

In this paper, a novel dual-channel VTS feature compensation method for noise-robust ASR on mobile devices has been proposed. It has been experimentally shown that our proposal is able to achieve high recognition accuracy when integrated in an ASR system working on challenging noisy environments, especially at low SNRs. In particular, our results have demonstrated that a dual-channel VTS approach outperforms a single-channel one by taking advantage of the spatial properties of speech and noise (modeled as the relative acoustic path between the two sensors and the dual-channel noise spatial covariance matrix, respectively).

As future work, we will evaluate the performance of our proposal on a large-vocabulary speech recognition task with a DNN-HMM-based ASR back-end. It can be expected that our proposal achieves relevant word accuracy improvements within the latter framework according to recent research, as in [28, 29], where it has been shown that the application of certain preprocessing stages (e.g. feature enhancement stages) can improve the performance of the DNN-HMM-based ASR systems trained with both clean and multi-style data.

6. Acknowledgments

This work has been supported by the MICINN and European Regional Development Fund (ERDF) TEC2013-46690-P project.

7. References

[1] J. Barker, R. Marxer, E. Vincent, and S. Watanabe. The third ‘CHiME’ speech separation and recognition challenge: Dataset, task and baselines. In *ASRU 2015 – IEEE Automatic Speech Recognition and Understanding, December 13–17, Scottsdale, USA, Proceedings*, 2015.

[2] J. M. Baker et al. Updated MINDS report on speech recognition and understanding, part 2. *IEEE Signal Processing Magazine*, 26:78–85, 2009.

[3] M. Jeub, C. Herglotz, C. M. Nelke, C. Beaugeant, and P. Vary. Noise reduction for dual-microphone mobile phones exploiting power level differences. In *ICASSP 2012 – 37th International Conference on Acoustics, Speech, and Signal Processing, March 25–30, Kyoto, Japan, Proceedings*, pages 1693–1696, 2012.

[4] J. Zhang, R. Xia, Z. Fu, J. Li, and Y. Yan. A fast two-microphone noise reduction algorithm based on power level ratio for mobile phone. In *ISCSLP 2012 – 8th International Symposium on Chinese Spoken Language Processing, December 5–8, Hong Kong, Proceedings*, pages 206–209, 2012.

[5] Z. Fu, F. Fan, and J. Huang. Dual-microphone noise reduction for mobile phone application. In *ICASSP 2013 – 38th International Conference on Acoustics, Speech, and Signal Processing, May 26–31, Vancouver, Canada, Proceedings*, pages 7239–7243, 2013.

[6] Z. Koldovsky, P. Tichavsky, and D. Botka. Noise reduction in dual-microphone mobile phones using a bank of pre-measured target-cancellation filters. In *ICASSP 2013 – 38th International Conference on Acoustics, Speech, and Signal Processing, May 26–31, Vancouver, Canada, Proceedings*, pages 679–683, 2013.

[7] A. Sugiyama and R. Miyahara. A new generalized sidelobe canceller with a compact array of microphones suitable for mobile terminals. In *ICASSP 2014 – 39th International Conference on Acoustics, Speech, and Signal Processing, May 4–9, Florence, Italy, Proceedings*, pages 820–824, 2014.

[8] N. Yousefian, A. Akbaria, and M. Rahmani. Using power level difference for near field dual-microphone speech enhancement. *Applied Acoustics*, 70:1412–1421, 2009.

[9] 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*, 8854:119–128, 2014.

[10] 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 *EUSIPCO 2014 – 22nd European Signal Processing Conference, September 1–5, Lisbon, Portugal, Proceedings*, pages 21–25, 2014.

[11] I. Tashev, S. Mihov, T. Gleghorn, and A. Acero. Sound capture system and spatial filter for small devices. In *EUROSPEECH 2008 – 9th Annual Conference of the International*

Speech Communication Association, September 22–26, Brisbane, Australia, Proceedings, pages 435–438, 2008.

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

[13] P. J. Moreno, B. Raj, and R. M. Stern. A vector Taylor series approach for environment-independent speech recognition. In *ICASSP 1996 – 21st International Conference on Acoustics, Speech, and Signal Processing, May 7–10, Atlanta, GA, Proceedings*, pages 733–736, 1996.

[14] P. Moreno. *Speech Recognition in Noisy Environments.* Ph.D. thesis (Carnegie Mellon University), 1996.

[15] 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 *EUROSPEECH 2001 – 7th European Conference on Speech Communication and Technology, September 3–7, Aalborg, Denmark, Proceedings*, 2001.

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

[17] A. Acero et al. HMM adaptation using vector Taylor series for noisy speech recognition. In *ICSLP 2000 – 6th International Conference of Spoken Language Processing, October 16–20, Beijing, China, Proceedings*, pages 229–232, 2000.

[18] B. S. Atal. Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. *J. Aco. Soc. Am.*, 55:1304–1312, 1974.

[19] J. A. González, A. M. Peinado, A. M. Gomez, and J. L. Carmona. Efficient MMSE estimation and uncertainty processing for multienvironment robust speech recognition. *IEEE Trans. on Audio, Speech, and Language Proc.*, 19, 2011.

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

[21] K. B. Petersen and M. S. Pedersen. *The Matrix Cookbook*. Technical University of Denmark, 2008.

[22] V. Stouten, H. Van Hamme, and P. Wambacq. Model-based feature enhancement with uncertainty decoding for noise robust ASR. *Speech Communication*, 48:1502–1514, 2006.

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

[24] A. M. Peinado and J. C. Segura. *Speech Recognition over Digital Channels*. Wiley, 2006.

[25] J. A. González, A. M. Peinado, N. Ma, A. M. Gomez, and J. Barker. MMSE-based missing-feature reconstruction with temporal modeling for robust speech recognition. *IEEE Trans. on Audio, Speech, and Language Proc.*, 21:624–635, 2013.

[26] X. Mestre and M. Á. Lagunas. On diagonal loading for minimum variance beamformers. In *ISSPIT 2003 – 3th International Symposium on Signal Processing and Information Technology, Darmstadt, Germany, Proceedings*, pages 459–462, 2003.

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

[28] S. Y. Chang and S. Wegmann. On the importance of modeling and robustness for deep neural network feature. In *ICASSP 2015 – 40th International Conference on Acoustics, Speech, and Signal Processing, April 19–24, Brisbane, Australia, Proceedings*, 2015.

[29] D. Baby, J. F. Gemmeke, T. Virtanen, and H. Van Hamme. Exemplar-based speech enhancement for deep neural network based automatic speech recognition. In *ICASSP 2015 – 40th International Conference on Acoustics, Speech, and Signal Processing, April 19–24, Brisbane, Australia, Proceedings*, 2015.