Dual-Channel Spectral Weighting for Robust Speech Recognition in Mobile Devices

Iván López-Espejo[a,*], Antonio M. Peinado[a], Angel M. Gomez[a], Jose A. Gonzalez[b]

[a] *Dept. of Signal Theory, Telematics and Communications, University of Granada, Granada, Spain*
[b] *Dept. of Computer Science, University of Sheffield, Sheffield, UK*

Abstract

Many mobile devices now include an extra microphone, frequently placed at their rear, intended to obtain information about the environmental noise for speech de-noising purposes. Although this secondary sensor can be regarded as just another element in a microphone array when performing beamforming, in this paper we show that it can be considered differently in order to better exploit the information about the acoustic environment. In particular, we propose a novel spectral weighting based on Wiener filtering that takes benefit from this secondary microphone to perform noise-robust automatic speech recognition (ASR) in mobile devices. At first it is assumed that the secondary microphone only captures noise while a reference sensor in the array (primary microphone) observes the same noise spectrum (homogeneous noise field). Since both assumptions are not always accurate, the Wiener filter (WF) weighting is modified through *1)* a bias correction term (to rectify the resulting spectral weights when a non-negligible speech component is present at the secondary channel) and *2)* a novel noise equalization to be applied on the secondary channel before spectral weight computation. Speech recognition experiments are performed on a dual-microphone smartphone (AURORA2-2C-CT/FT corpora) and a tablet with six microphones (CHiME-3/4). Our results show the high

*Corresponding author
Email address: iloes@ugr.es (Iván López-Espejo)

performance of our approach as well as its great versatility regardless of the analyzed mobile device and usage scenario.

Keywords: Power spectrum enhancement, spectral weighting, robust speech recognition, dual-channel, mobile device

1. Introduction

Automatic speech recognition (ASR) systems suffer from accuracy issues when they are deployed in noisy environments [1, 2, 3]. Many techniques have been proposed to improve ASR robustness against environmental noise. However, they may likely fail when they are exposed to rapidly changing and low-SNR (Signal-to-Noise Ratio) environments [4], such as those that can be expected when using ASR on a mobile device. Mobile devices can be employed anywhere at any time, in such a way that coping with a wide variety of noisy environments is mandatory to provide a good user experience. With the aim of enhancing noisy speech, these devices have begun to integrate small microphone arrays [5, 6, 7, 8], i.e., microphone arrays comprised of a few sensors close each other. This technology can also be exploited for noise-robust ASR purposes as is shown in [9, 10, 11, 12, 13]. In the first two references, the output of a filter-and-sum beamformer is enhanced by post-filtering. Thus, while the output of a Wiener post-filter is employed for the estimation of soft missing-data masks in [9], the post-filter in [10] is implemented through a deep neural network (DNN) and its output is used to feed the recognition engine. Similarly, post-filtering with multi-channel noise reduction is applied after minimum variance distortionless response (MVDR) beamforming in [11], where the speech gains of the steering vector are computed by eigenvalue decomposition of the clean speech spatial covariance matrix. In [12], a binary mask is estimated from the dual noisy observation by means of a DNN to perform feature compensation on smartphones with a dual-microphone. Mask estimation is based on the power level difference (PLD) [5] between the two microphones of the device: in a close-talk configuration (i.e., the loudspeaker of the smartphone is placed

at the ear) greater speech power is received at the primary microphone than at the secondary one while almost the same noise power is assumed to be received at both of them [14]. This way, a missing-data mask for the primary channel can be inferred by comparing the noisy speech power at both channels and applied to missing-data imputation [15]. Another alternative approach for dual-microphone smartphones is that proposed in [13] in which two power spectrum enhancement techniques are developed, one based on MVDR and another on spectral subtraction (SS).

Mobile devices that embed small microphone arrays have one or more microphones facing toward the speaker intended to capture her/his voice. In addition, many of these mobile devices have at least one extra microphone that is placed at their rear (the so-called secondary microphone in this work). In this case, valuable information about the acoustic environment can be captured since the microphone is placed in an acoustic shadow with respect to the target speech signal. While beamforming has proven to be useful in providing robustness in general microphone array scenarios [16, 17, 18], the secondary microphone is barely able to improve beamforming performance since its main task is to obtain information about the acoustic environment rather than directing the array towards the speaker [19, 20].

In this paper we propose a novel spectral weighting technique based on Wiener filtering intended for noise-robust ASR in mobile devices that integrate this kind of secondary sensor. We assume that only a single primary (front) microphone is available in the device. If the device integrates more than one front sensor, a virtual primary channel will be computed through the application of beamforming. In such a case, our Wiener filter (WF)-based weighting behaves as a post-filter, the performance of which will be proven in this paper to be superior than that of other well-known beamforming post-filters [21, 11]. To estimate the *a priori* signal-to-noise ratio (SNR) required by the WF, we initially assume that the secondary microphone captures no speech and that noise acquired by this microphone matches that one captured by the primary microphone. Although these assumptions can be acceptable in some situations (e.g.,

when the mobile device is used in close-talk position and into a homogeneous noise field [5, 14]), in general, they will not be satisfied. Hence, we introduce two modifications to the basic WF weighting oriented to overcome the lack of realism of these two initial assumptions. First, a bias correction term is introduced to rectify the resulting spectral weights when a non-negligible speech component is present at the secondary channel. Second, we propose a novel noise equalization procedure to be applied on the secondary channel before spectral weight computation to make the noise power spectral densities (PSDs) at both channels similar.

Our proposals are evaluated on a smartphone with a dual-microphone as well as on a tablet with six microphones (five of them facing forward and one facing backwards). In the case of the smartphone, the primary microphone is located at the bottom while the secondary one is located at the rear. This device is analyzed in two different scenarios: close-talk and far-talk. The latter implies that the device is held at some distance (from a few centimeters to less than one meter) from the face of the user. Thus, while for interactive voice response (IVR) applications such as telephone banking it is common to make use of the device in close-talk position, the far-talk configuration is especially interesting for some other ASR applications where viewing the screen is required such as search-by-voice. To experiment with both scenarios two synthetic dual-channel noisy speech databases, based on the well-known Aurora-2 corpus [22], are used: the AURORA2-2C-CT (Aurora-2 - 2 Channels - Close-Talk) database (reported in [13]) and the similarly defined AURORA2-2C-FT (Aurora-2 - 2 Channels - Far-Talk) corpus. In the case of the tablet, a virtual primary channel is obtained by means of beamforming while the secondary microphone corresponds to the one facing backwards. This type of device is only analyzed in far-talk conditions by using the CHiME-3 [1] and CHiME-4 [2] frameworks in order to test the performance of our proposals under real data conditions. Our experiments show the great versatility of our approach according to its high performance regardless of the analyzed device and usage scenario.

The rest of the paper is organized as follows. In Section 2, the calculation of

the spectral weights based on Wiener filtering for power spectrum enhancement is briefly revisited and extended to a dual-channel framework. The biased and unbiased approaches to compute the *a priori* SNR in order to derive the spectral weights along with the noise equalization procedure are presented in Section 3. The full enhancement system that integrates the stages described in the previous sections as well as the implementation issues are explained in Section 4. In Section 5, the experimental framework is described. The experimental results are shown in Section 6. Finally, in Section 7 conclusions and future work are summarized.

2. Distortion Model and WF Filtering

Let us consider a dual-channel additive noise distortion model $y_k(m) = x_k(m) + n_k(m)$, where $y_k(m)$, $x_k(m)$ and $n_k(m)$ represent the noisy speech, clean speech and noise signals, respectively, from the k-th channel ($k \in \{1, 2\}$). In the following, sensors 1 and 2 will be referred to as the primary and secondary microphones[2]. Assuming that clean speech and noise are uncorrelated, the above distortion model can be expressed in the power spectral domain as

$$|Y_k(f,t)|^2 = |X_k(f,t)|^2 + |N_k(f,t)|^2, \tag{1}$$

where $f = 0, 1, ..., \mathcal{M} - 1$ denotes the frequency bin index and $t = 0, 1, ..., T - 1$ refers to the time frame index.

As the primary microphone is faced towards the speaker, we will assume that the signal captured by this microphone has a higher SNR than the signal captured by the secondary sensor and, hence, our objective is to provide an estimate of the clean speech power spectrum at the primary channel, $|\hat{X}_1(f,t)|^2$, by taking advantage of the dual-channel information. To this end, a Wiener filtering approach, widely used by many speech enhancement methods [23, 24, 25], is adopted. As it is well-known, the Wiener filter (WF) is optimal in the

[2]In the case of multiple front sensors, $k = 1$ refers to the virtual primary channel obtained by means of beamforming.

sense of minimizing the mean square error between the target signal and the estimated one given the input corrupted signal. Under our framework, the desired optimal non-causal filter in the frequency domain is given by [26]

$$H_1(f,t) = \frac{\mathcal{S}_{x_1}(f,t)}{\mathcal{S}_{x_1}(f,t) + \mathcal{S}_{n_1}(f,t)}, \tag{2}$$

where $\mathcal{S}_{x_1}(f,t)$ and $\mathcal{S}_{n_1}(f,t)$ are the PSDs of the clean speech and the noise, respectively, at the primary channel. Thus, the clean speech power spectrum bin $|X_1(f,t)|^2$ can be estimated as

$$|\hat{X}_1(f,t)|^2 = H_1^2(f,t)|Y_1(f,t)|^2. \tag{3}$$

It should be noted that $H_1^2(f,t) \in [0\ 1]$ may be seen as a spectral weight such that $H_1^2(f,t) \to 1$ ($H_1^2(f,t) \to 0$) if speech (noise) dominates.

The WF can be alternatively expressed as

$$H_1(f,t) = \frac{\xi_1(f,t)}{\xi_1(f,t) + 1}, \tag{4}$$

where

$$\xi_1(f,t) = \frac{\mathcal{S}_{x_1}(f,t)}{\mathcal{S}_{n_1}(f,t)} \tag{5}$$

is the *a priori* SNR of the primary channel. Under our additive noise distortion model, that is,

$$\mathcal{S}_{y_k}(f,t) = \mathcal{S}_{x_k}(f,t) + \mathcal{S}_{n_k}(f,t) \quad (k = 1,2), \tag{6}$$

Eq. (5) can be alternatively expressed as

$$\xi_1(f,t) = \frac{\mathcal{S}_{y_1}(f,t) - \mathcal{S}_{n_1}(f,t)}{\mathcal{S}_{n_1}(f,t)}. \tag{7}$$

A straightforward single-channel approach for obtaining $\xi_1(f,t)$ consists of directly estimating the noise PSD $\mathcal{S}_{n_1}(f,t)$ from signal $y_1(m)$ [27]. This is often a difficult task since speech and noise overlap. In the next section we will alternatively exploit the spatial characteristics of speech and noise under the considered dual-microphone configuration to obtain estimates of $\xi_1(f,t)$ from the available signals $y_1(m)$ and $y_2(m)$.

3. Dual-Channel Spectral Weight Estimation

3.1. Biased Spectral Weight Estimation

Previous work on dual-channel noise reduction has shown that, when a mobile device is used in a close-talk position, the clean speech PSD is considerably greater at the primary sensor than at the secondary one while a similar noise PSD is observed by both sensors (i.e., $\mathcal{S}_{n_1}(f,t) \approx \mathcal{S}_{n_2}(f,t) \gg \mathcal{S}_{x_2}(f,t)$) [5, 14]. This is due to the geometry of the speaker-device acoustic system (the secondary microphone is purposely placed in an acoustic shadow with respect to the speech source) and the typical existence of a homogeneous noise field. Under ideal conditions, we can consider that $\mathcal{S}_{n_1}(f,t) = \mathcal{S}_{n_2}(f,t)$ and $\mathcal{S}_{x_2}(f,t) = 0$. Therefore, $\mathcal{S}_{n_1}(f,t) = \mathcal{S}_{y_2}(f,t)$ and the *a priori* SNR of Eq. (7) can be expressed as

$$\xi_{1,b}(f,t) \quad = \quad \frac{\mathcal{S}_{y_1}(f,t) - \mathcal{S}_{y_2}(f,t)}{\mathcal{S}_{y_2}(f,t)}. \tag{8}$$

Hence, the corresponding WF is

$$H_{1,b}(f,t) = \frac{\mathcal{S}_{y_1}(f,t) - \mathcal{S}_{y_2}(f,t)}{\mathcal{S}_{y_1}(f,t)}. \tag{9}$$

Its practical computation is described in Subsection 4.3.

The WF estimation described above strongly depends on the accuracy of the two assumptions made, that is, a negligible speech component at the secondary sensor and similar noise PSDs at both sensors. These assumptions can be acceptable in some specific cases but, in general, they will not be accurate. In the next subsections we will present two novel procedures that will allow us the application of the WF-based spectral weighting in a wider range of situations.

3.2. Unbiased Spectral Weight Estimation

The assumption of a negligible speech component at the secondary channel may be appropriate, for instance, when a dual-microphone smartphone is employed in close-talk position [5], but it will clearly fail when the device is used in far-talk conditions [6]. In fact, the rear-side microphone also captures a component of speech mainly because of diffraction at the borders of the device and

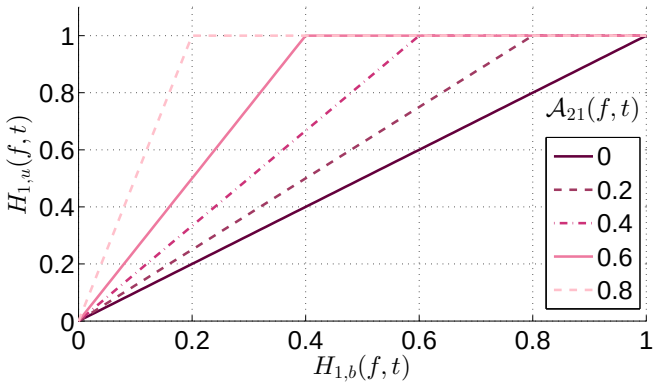

Figure 1: Result of applying the bias correction term on $H_{1,b}(f,t)$ for several $\mathcal{A}_{21}(f,t)$ values.

reflections from the acoustic environment. In this case, $\mathcal{S}_{y_2}(f,t) > \mathcal{S}_{n_2}(f,t)$ so that $\xi_{1,b}(f,t)$ and, therefore, $H_{1,b}(f,t)$, will be underestimated, i.e., biased (indicated by subscript b in the above variables). To address this problem, a bias correction term is introduced in the following.

Let us assume that the clean speech PSD at the secondary channel can be related to that at the first one by means of the gain factor $\mathcal{A}_{21}(f,t)$. This factor can be seen as the relative speech gain (RSG) between both microphones, that is, $\mathcal{S}_{x_2}(f,t) = \mathcal{A}_{21}(f,t)\mathcal{S}_{x_1}(f,t)$. Assuming, again, a homogeneous noise field ($S_{n_1}(f,t) = S_{n_2}(f,t)$), Eq. (6) for $k=2$ can be written as

$$\begin{aligned} \mathcal{S}_{y_2}(f,t) &= \mathcal{A}_{21}(f,t)\mathcal{S}_{x_1}(f,t) + \mathcal{S}_{n_1}(f,t) \\ &= \mathcal{A}_{21}(f,t)\left(\mathcal{S}_{y_1}(f,t) - \mathcal{S}_{n_1}(f,t)\right) + \mathcal{S}_{n_1}(f,t). \end{aligned} \tag{10}$$

This equation allows us to express the noise PSD at the primary channel in terms of the PSDs of the available noisy signals, that is,

$$\mathcal{S}_{n_1}(f,t) = \frac{\mathcal{S}_{y_2}(f,t) - \mathcal{A}_{21}(f,t)\mathcal{S}_{y_1}(f,t)}{1 - \mathcal{A}_{21}(f,t)}. \tag{11}$$

By substituting this noise PSD into (7) we obtain the following expression for the *a priori* SNR:

$$\xi_{1,u}(f,t) = \frac{\mathcal{S}_{y_1}(f,t) - \mathcal{S}_{y_2}(f,t)}{\mathcal{S}_{y_2}(f,t) - \mathcal{A}_{21}(f,t)\mathcal{S}_{y_1}(f,t)}, \tag{12}$$

where subscript u indicates an unbiased approach. The SNR expression in (12) yields the following WF:

$$H_{1,u}(f,t) = \frac{\mathcal{S}_{y_1}(f,t) - \mathcal{S}_{y_2}(f,t)}{\mathcal{S}_{y_1}(f,t)(1 - \mathcal{A}_{21}(f,t))}. \tag{13}$$

By comparing this expression with that of Eq. (9), we observe that the WF bias can be corrected by dividing (9) by $B(f,t) = (1 - \mathcal{A}_{21}(f,t))$. In other words, the new WF can be obtained from the one in the previous subsection by applying the bias correction term $B^{-1}(f,t)$ as

$$H_{1,u}(f,t) = B^{-1}(f,t) H_{1,b}(f,t). \tag{14}$$

Figure 1 shows the effect of the bias correction term on $H_{1,b}(f,t)$ for different values of $\mathcal{A}_{21}(f,t)$. As can be observed, if $\mathcal{A}_{21}(f,t) = 0$ (i.e., no speech is captured by the secondary microphone), the assumption made when calculating $\xi_{1,b}(f,t)$ according to (8) holds true, so that the WF is not modified. On the other hand, as $\mathcal{A}_{21}(f,t)$ increases, the initial underestimation of $H_{1,b}(f,t)$ due to a non-negligible speech component at the secondary channel is rectified. It must be noted that, since $H_{1,u}(f,t) > 1$ has no physical sense, $B^{-1}(f,t)H_{1,b}(f,t)$ has been bounded by 1 in Figure 1.

The way the RSG $\mathcal{A}_{21}(f,t)$ is computed in this work is explained in Subsection 4.1.

3.3. Noise Equalization

The assumption $\mathcal{S}_{n_1}(f,t) \approx \mathcal{S}_{n_2}(f,t)$ used for deriving the WF-based weighting in Subsections 3.1 and 3.2 can be acceptable when the mobile device is employed within a homogeneous noise field (e.g., in a diffuse noise field as in interior spaces, urban streets with high-rise buildings, etc. [28]). However, this assumption may not be satisfied in several scenarios even in the presence of a homogeneous noise field (e.g., in the case of two microphones with different characteristics). In this subsection we propose a novel noise equalization procedure to be performed before spectral weight computation. This procedure transforms the signal at channel 2 so that the noise component is forced

to follow the one at the primary channel while keeping the speech component untouched (this is a distortionless constraint similar to that of MVDR beamforming). Hence, we aim to obtain a signal $|\bar{Y}_2(f,t)|^2 = |\bar{X}_2(f,t)|^2 + |\bar{N}_2(f,t)|^2$, where $|\bar{X}_2(f,t)|^2 \approx |X_2(f,t)|^2$ and $|\bar{N}_2(f,t)|^2 \approx |N_1(f,t)|^2$, to replace $|Y_2(f,t)|^2$ in the estimation of the PSD $\mathcal{S}_{y_2}(f,t)$ required in Eqs. (9) and (13) for WF computation.

To make our equalization procedure more effective, we will additionally introduce an overestimation of the noise power spectrum. This overestimation is inspired by the oversubtraction typically applied in spectral subtraction (SS) which helps to reduce the "musical" artifacts that SS tends to introduce, thus yielding a better recognition performance as shown in [29, 30]. In our case, since the PSD of the secondary channel carries the information about the noise required in (9) and (13), we will consider an overestimation factor $\beta(f,t) \geq 1$ so that

$$|\bar{Y}_2(f,t)|^2 \approx |X_2(f,t)|^2 + \beta(f,t)|N_1(f,t)|^2. \tag{15}$$

For computing $\beta(f,t)$, we follow the same approach reported in [31], where

$$\beta(f,t) = \left(1 + \frac{std\left(|N_1(f,t)|^2\right)}{|N_1(f,t)|^2}\right) \tag{16}$$

in which $std\left(|N_1(f,t)|^2\right)$ is the standard deviation of $|N_1(f,t)|^2$ at frequency bin f and time frame t. As a result, we aim to approximate

$$|\bar{Y}_2(f,t)|^2 \approx |X_2(f,t)|^2 + |N_1(f,t)|^2 + std\left(|N_1(f,t)|^2\right). \tag{17}$$

In particular, we will obtain $|\bar{Y}_2(f,t)|^2$ from the following linear combination of the dual-channel noisy observation:

$$\begin{aligned} |\bar{Y}_2(f,t)|^2 &= \mathbf{g}_{f,t}^{\mathrm{T}} \begin{pmatrix} |Y_2(f,t)|^2 \\ |Y_1(f,t)|^2 \end{pmatrix} \\ &= \mathbf{g}_{f,t}^{\mathrm{T}} \begin{pmatrix} |X_2(f,t)|^2 \\ |X_1(f,t)|^2 \end{pmatrix} + \mathbf{g}_{f,t}^{\mathrm{T}} \begin{pmatrix} |N_2(f,t)|^2 \\ |N_1(f,t)|^2 \end{pmatrix}, \end{aligned} \tag{18}$$

where $\mathbf{g}_{f,t}$ is the weight vector to be estimated. By using the RSG $\mathcal{A}_{21}(f,t)$ as in $|X_2(f,t)|^2 = \mathcal{A}_{21}(f,t)|X_1(f,t)|^2$, and $\boldsymbol{\nu}(f,t) = \left(|N_2(f,t)|^2, |N_1(f,t)|^2\right)^{\mathrm{T}}$,

(18) can be expressed as

$$|\bar{Y}_2(f,t)|^2 = \mathbf{g}_{f,t}^{\mathrm{T}} \begin{pmatrix} 1 \\ \mathcal{A}_{21}^{-1}(f,t) \end{pmatrix} |X_2(f,t)|^2 + \mathbf{g}_{f,t}^{\mathrm{T}} \boldsymbol{\nu}(f,t). \tag{19}$$

Then, by comparing (17) and (19) we can see that our goal is to estimate the weight vector $\hat{\mathbf{g}}_{f,t}$ that transforms $\mathbf{g}_{f,t}^{\mathrm{T}} \boldsymbol{\nu}(f,t)$ into $|N_1(f,t)|^2 + std\left(|N_1(f,t)|^2\right)$ under a minimum mean square error (MMSE) criterion plus a distortionless constraint for $|X_2(f,t)|^2$. In other words, if we define $\boldsymbol{\alpha}(f,t) = \left(1, \mathcal{A}_{21}^{-1}(f,t)\right)^{\mathrm{T}}$ and

$$\varepsilon_{f,t} = \left(|N_1(f,t)|^2 + std\left(|N_1(f,t)|^2\right)\right) - \mathbf{g}_{f,t}^{\mathrm{T}} \boldsymbol{\nu}(f,t), \tag{20}$$

we want to obtain

$$\begin{aligned} &\hat{\mathbf{g}}_{f,t} = \arg\min_{\mathbf{g}_{f,t}} \mathrm{E}\left[\varepsilon_{f,t}^2\right]; \\ &\text{subject to } \ \mathbf{g}_{f,t}^{\mathrm{T}} \boldsymbol{\alpha}(f,t) = 1. \end{aligned} \tag{21}$$

The optimization problem above is solved by the Lagrange multipliers method, yielding the weight vector estimate

$$\hat{\mathbf{g}}_{f,t} = \boldsymbol{\Phi}_N^{-1}(f,t) \left[\boldsymbol{\gamma}_N(f,t) - \frac{\boldsymbol{\alpha}^{\mathrm{T}}(f,t) \boldsymbol{\Phi}_N^{-1}(f,t) \boldsymbol{\gamma}_N(f,t) - 1}{\boldsymbol{\alpha}^{\mathrm{T}}(f,t) \boldsymbol{\Phi}_N^{-1}(f,t) \boldsymbol{\alpha}(f,t)} \boldsymbol{\alpha}(f,t) \right], \tag{22}$$

which, as can be seen, depends on the noise spatial correlation matrix $\boldsymbol{\Phi}_N(f,t)$ and the overestimated noise spatial correlation vector $\boldsymbol{\gamma}_N(f,t)$. First, $\boldsymbol{\Phi}_N(f,t)$ is defined as

$$\boldsymbol{\Phi}_N(f,t) = \begin{pmatrix} \phi_{N,f,t}(2,2) & \phi_{N,f,t}(2,1) \\ \phi_{N,f,t}(1,2) & \phi_{N,f,t}(1,1) \end{pmatrix}, \tag{23}$$

where $\phi_{N,f,t}(k,l) = \mathrm{E}\left[|N_k(f,t)|^2 |N_l(f,t)|^2\right]$ $(k,l = 1,2)$. Second, the overestimated noise spatial correlation vector is

$$\boldsymbol{\gamma}_N(f,t) = \boldsymbol{\phi}_N^{(1)}(f,t) + std\left(|N_1(f,t)|^2\right) \boldsymbol{\mu}_N(f,t), \tag{24}$$

where

$$\boldsymbol{\phi}_N^{(1)}(f,t) = \begin{pmatrix} \phi_{N,f,t}(2,1) \\ \phi_{N,f,t}(1,1) \end{pmatrix} \tag{25}$$

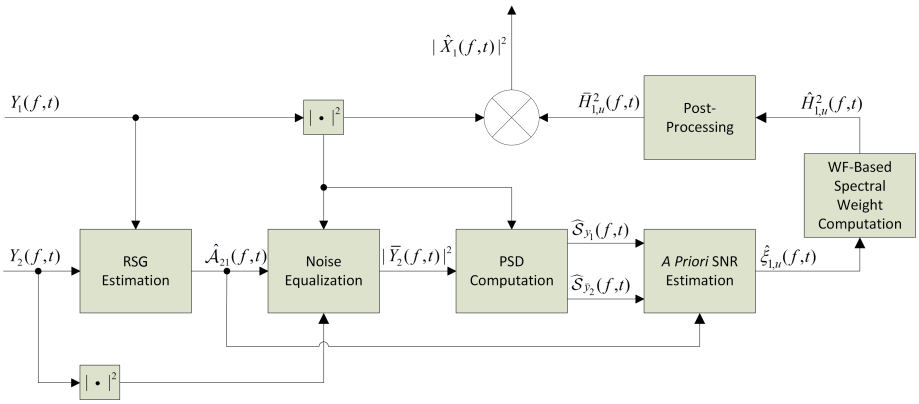

Figure 2: Block diagram of the full proposed enhancement system.

is a noise spatial correlation vector, and the noise mean vector is given by

$$\boldsymbol{\mu}_N(f,t) = \begin{pmatrix} \mathrm{E}\left[|N_2(f,t)|^2\right] \\ \mathrm{E}\left[|N_1(f,t)|^2\right] \end{pmatrix}. \tag{26}$$

In practice, the required noise statistical parameters $\boldsymbol{\Phi}_N(f,t)$ and $\boldsymbol{\gamma}_N(f,t)$ may be estimated during noise-only periods identified by means of a voice activity detector (VAD). In Subsection 4.2 we detail how such parameters are finally obtained in this work.

4. Implementation Issues

A block diagram of the full proposed enhancement system is depicted in Figure 2. This system integrates the different stages described above plus two additional steps: RSG estimation and post-processing of the spectral weights. Both the implementation issues and the additional blocks are explained in the subsections below. It should be remarked that the parameter values shown below were chosen by means of preliminary speech recognition experiments over development datasets.

Figure 3 shows an example of applying the estimated spectral weights by means of the full enhancement system in Figure 2 to the primary noisy power spectrum of an utterance captured by a dual-microphone smartphone used in

close-talk position. In this example, the noise reduction capability of our proposal can be visually inspected.

4.1. Relative Speech Gain Estimation

As seen above, the relative speech gain (RSG) between two microphones relates either the clean speech PSD or the instantaneous clean speech power at both sensors. Let us consider the latter case for the rest of this subsection, i.e., $|X_2(f,t)|^2 = \mathcal{A}_{21}(f,t)|X_1(f,t)|^2$. The RSG $\mathcal{A}_{21}(f,t)$ may be pre-computed or calculated online in various ways. Following the former approach makes sense when dual-channel clean speech data are available in advance and $\mathcal{A}_{21}(f,t)$ is virtually fixed. For example, this might be the case of using a dual-microphone smartphone in close-talk conditions. On the other hand, if the above requirements are not fulfilled, $\mathcal{A}_{21}(f,t)$ may be derived from an online estimation of the steering vector as explained in the following.

In a dual-channel configuration, we can state that

$$\begin{pmatrix} X_1(f,t) \\ X_2(f,t) \end{pmatrix} = \mathbf{d}(f,t)X(f,t), \tag{27}$$

where $X_1(f,t)$ and $X_2(f,t)$ are the clean speech primary and secondary signals in the short-time Fourier transform (STFT) domain, respectively. Moreover, $X(f,t)$ is the source speech signal (i.e., as emitted by the speaker) in the STFT domain, and $\mathbf{d}(f,t)$ is the steering vector, which is defined as

$$\begin{aligned} \mathbf{d}(f,t) &= (d_1(f,t),\ d_2(f,t))^{\mathrm{T}} \\ &= \begin{pmatrix} a_1(f,t)e^{-j2\pi f\tau_1(t)} \\ a_2(f,t)e^{-j2\pi f\tau_2(t)} \end{pmatrix}. \end{aligned} \tag{28}$$

In the above equation, $a_k(f,t)$ and $\tau_k(t)$ are, respectively, the gain and traveling time of the signal from the source to the k-th sensor, $k = 1, 2$. From (27) and (28), it is clear that

$$|X_2(f,t)|^2 = \underbrace{\left|\frac{d_2(f,t)}{d_1(f,t)}\right|^2}_{\mathcal{A}_{21}(f,t)} |X_1(f,t)|^2, \tag{29}$$

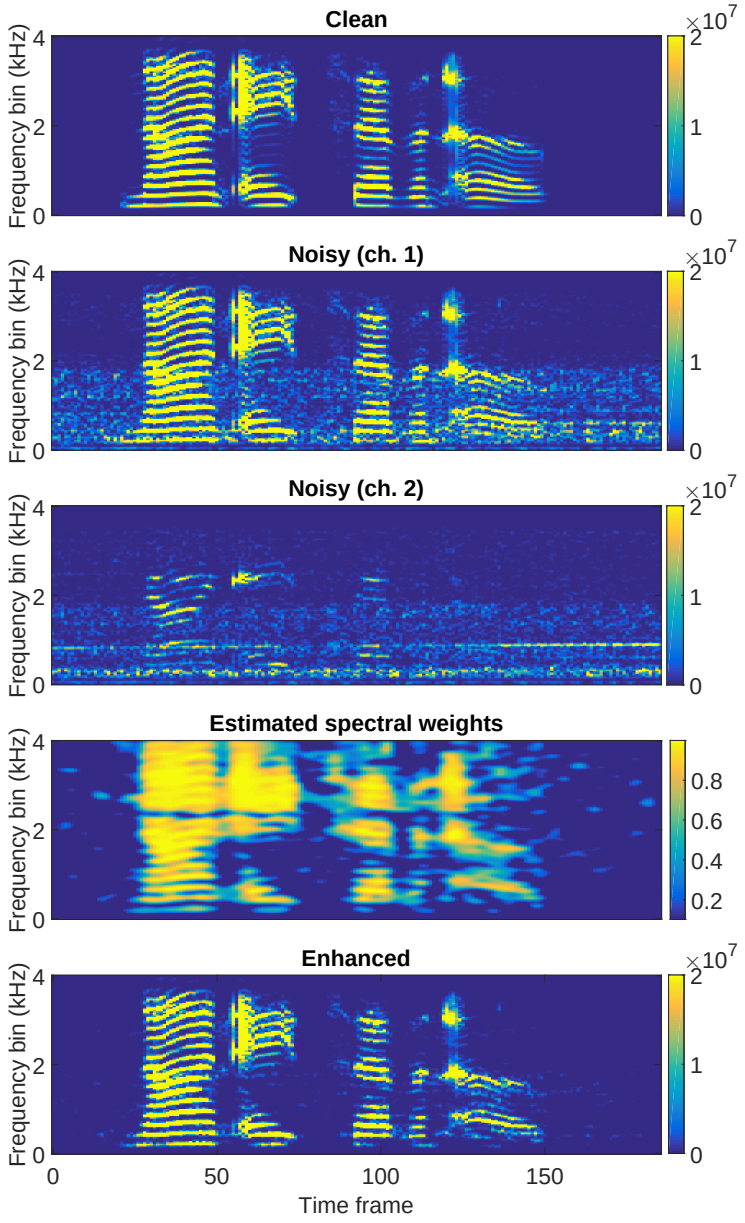

Figure 3: Example of spectral weighting by using the enhancement system of Fig. 2 for the utterance "nine eight seven oh" obtained from a dual-microphone smartphone in close-talk position. The utterance is contaminated with car noise at 0 dB in the primary channel. From top to bottom: clean speech power spectrum in the primary channel, noisy versions at the 1st and 2nd channels, estimated spectral weights and enhanced power spectrum.

where $\mathcal{A}_{21}(f,t)$ can be alternatively expressed as

$$\mathcal{A}_{21}(f,t) = \frac{a_2^2(f,t)}{a_1^2(f,t)}. \tag{30}$$

The particular algorithm applied in this work for steering vector estimation is specified in Section 6.

4.2. Noise Equalization Implementation

In practice, both the noise spatial correlation matrix $\mathbf{\Phi}_N(f,t)$ of Eq. (23) and the overestimated noise spatial correlation vector $\boldsymbol{\gamma}_N(f,t)$ of (24) are calculated as follows. First, two initial noise spatial correlation matrices as well as two initial overestimated noise spatial correlation vectors are computed per utterance and frequency bin f, one from the first M frames ($\mathbf{\Phi}_{N,f}^{(0)}$ and $\boldsymbol{\gamma}_{N,f}^{(0)}$, respectively) and another from the last M frames ($\mathbf{\Phi}_{N,f}^{(e)}$ and $\boldsymbol{\gamma}_{N,f}^{(e)}$, respectively). As for the noise estimates, $\mathbf{\Phi}_N(f,t)$ ($\boldsymbol{\gamma}_N(f,t)$) is calculated by means of linear interpolation between $\mathbf{\Phi}_{N,f}^{(0)}$ ($\boldsymbol{\gamma}_{N,f}^{(0)}$) and $\mathbf{\Phi}_{N,f}^{(e)}$ ($\boldsymbol{\gamma}_{N,f}^{(e)}$).

To avoid any possible negative power spectrum bin in (18), $|\bar{Y}_2(f,t)|^2$ is bounded below by $\eta|Y_2(f,t)|^2$, where $0 < \eta \ll 1$ is a thresholding factor set to 0.1.

An example of application of the proposed noise equalization procedure is depicted in Figure 4. The figure shows the noise spectra obtained from a dual-microphone smartphone in close-talk position averaged across time over the whole utterance. It can be observed that the equalized noise $|\bar{N}_2(f,t)|^2$ is much more similar to $|N_1(f,t)|^2$ than the original $|N_2(f,t)|^2$. The effect of the noise overestimation factor $\beta(f,t)$ can also be assessed.

4.3. Spectral Weight Computation

The PSDs of the two available noisy signals required by the *a priori* SNR computation in (8) and (12) are obtained by applying a two-dimensional 3×3 mean smoothing filter over the spectrogram $|Y_k(f,t)|^2$ ($k = 1, 2$), that is,

$$\hat{\mathcal{S}}_{y_k}(f,t) = \frac{1}{\mathcal{K}} \sum_{\nu=-1}^{1} \sum_{\tau=-1}^{1} |Y_k(f+\nu, t+\tau)|^2, \tag{31}$$

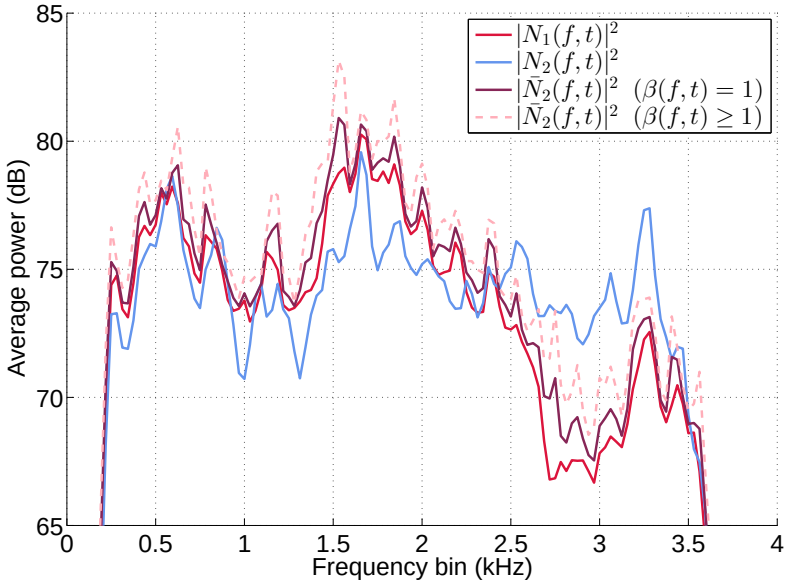

Figure 4: Example of the proposed noise equalization when applied on an utterance from a dual-microphone smartphone used in close-talk position. Both the estimated noise average power $|\bar{N}_2(f,t)|^2$ $(\beta(f,t) = 1)$ and its overestimated version, $|\bar{N}_2(f,t)|^2$ $(\beta(f,t) \geq 1$ as in Eq. (16)), are represented by frequency bin along with the actual noise average power from the two channels.

where we also assume $|Y_k(f,t)|^2 = 0$ for $f, t < 0$, $f \geq \mathcal{M}$ and $t \geq T$, and $\mathcal{K}$ is a normalizing factor equal to 9, 6 or 4 depending on the number of available spectrogram points. From these PSDs, the *a priori* SNRs $\xi_{1,b}(f,t)$ and $\xi_{1,u}(f,t)$ are similarly estimated in practice as

$$\hat{\xi}_{1,b}(f,t) = \max\left(\frac{\hat{\mathcal{S}}_{y_1}(f,t) - \hat{\mathcal{S}}_{y_2}(f,t)}{\hat{\mathcal{S}}_{y_2}(f,t)}, \eta_\xi\right), \tag{32}$$

$$\hat{\xi}_{1,u}(f,t) = \max\left(\frac{\hat{\mathcal{S}}_{y_1}(f,t) - \hat{\mathcal{S}}_{n_1}(f,t)}{\hat{\mathcal{S}}_{n_1}(f,t)}, \eta_\xi\right), \tag{33}$$

where both expressions are floored at η_ξ in order to avoid negative values (discussed in the following subsection). Moreover, the noise PSD $\mathcal{S}_{n_1}(f,t)$ calculated through (11) is also thresholded by $\eta_n = 10^3$ (which roughly corresponds to an SNR of 40 dB) to avoid negative PSD bins, i.e.,

$$\hat{\mathcal{S}}_{n_1}(f,t) = \max\left(\frac{\hat{\mathcal{S}}_{y_2}(f,t) - \hat{\mathcal{A}}_{21}(f,t)\hat{\mathcal{S}}_{y_1}(f,t)}{1 - \hat{\mathcal{A}}_{21}(f,t)}, \eta_n\right). \tag{34}$$

Finally, the estimated WFs $\hat{H}_{1,b}(f,t)$ and $\hat{H}_{1,u}(f,t)$ are obtained by substituting (32) and (33) into (4), respectively.

4.4. Post-Processing Block

Once either the WF-based spectral weights $\hat{H}^2_{1,b}(f,t)$ or $\hat{H}^2_{1,u}(f,t)$ are obtained, some post-processing operations are performed on them. For the sake of clarity let us consider only $\hat{H}^2_{1,u}(f,t)$ in the rest of this subsection. First, for speech recognition purposes, previous works have shown that better speech recognition accuracy can be achieved by leaving a small fraction of noise energy in the enhanced signal [29, 32]. Hence, $\hat{H}^2_{1,u}(f,t)$ is bounded below in accordance with

$$\bar{H}^2_{1,u}(f,t) = \max\left(\hat{H}^2_{1,u}(f,t), \eta\right), \tag{35}$$

where $\eta = 0.1$ is the same thresholding factor as in Subsection 4.2. Indeed, thresholding $\hat{H}^2_{1,u}(f,t)$ by η is equivalent to consider

$$\eta_\xi = \frac{\sqrt{\eta}}{1 - \sqrt{\eta}} \tag{36}$$

in (33) in accordance with the WF definition of (4). Therefore, this thresholding enhancement is directly accomplished substituting (36) into (33). It should be noticed that $\eta = 0.1$ implies that $\eta_\xi \approx -3.35$ dB.

Second, as in [33], we exploit the spectro-temporal correlation of speech in order to refine $\bar{H}^2_{1,u}(f,t)$ by applying a couple of two-dimensional filters in the time-frequency domain. The first one consists of a median filter of size 3×5, that tries to remove high-valued $\bar{H}^2_{1,u}(f,t)$ bins surrounded by low values of $\bar{H}^2_{1,u}(f,t)$. This procedure is justified by the fact that it is more likely that those bins constitute artifacts rather than real isolated clean speech spectral bins. Indeed, this kind of artifact often appears when the assumption $\mathcal{S}_{n_1}(f,t) \approx \mathcal{S}_{n_2}(f,t)$ does not hold but instead $\mathcal{S}_{n_1}(f,t)$ is significantly greater than $\mathcal{S}_{n_2}(f,t)$. Finally, in order to further increase the spectro-temporal coherence, the spectral weights resulting from median filtering are smoothed by convolving them with a Gaussian kernel of standard deviation $\sigma = 1$ and size 5×5.

5. Experimental Framework

The techniques presented above are assessed in terms of word accuracy and/or word error rate for two types of mobile device: a smartphone with a dual-microphone, used both in close-talk and far-talk conditions, and a tablet with six microphones only employed in a far-talk position. In this section we briefly describe the different corpora used during our experimental evaluation (AURORA2-2C-CT/FT, and CHiME-3 and CHiME-4 databases) along with their related setup particularities (i.e., the feature extraction process and the baseline back-end configuration).

5.1. The AURORA2-2C-CT/FT Databases and Settings

The AURORA2-2C-CT (Aurora-2 - 2 Channels - Close-Talk) and the AURORA2-2C-FT (Aurora-2 - 2 Channels - Far-Talk) databases are two synthetic dual-channel noisy speech databases generated from the well-known Aurora-2 corpus [22]. On the one hand, the AURORA2-2C-CT database, described in detail

in [13], tries to emulate the acquisition of dual-channel noisy speech data by using a dual-microphone smartphone in close-talk conditions (i.e., when the loudspeaker of the smartphone is placed at the ear of the user). On the other hand, the AURORA2-2C-FT database is generated in a similar way but emulating a far-talk scenario (i.e., when the user holds the device in one hand at some distance from her/his face). To the best of our knowledge, no similar real data corpora are available.

Two test sets, A and B (with the same structure as in Aurora-2), are defined in AURORA2-2C-CT/FT with different types of noise in each one. The types of noise used in test set A are bus, babble, car and pedestrian street, while test set B comprises the noises café, street, bus station and train station. These noises were recorded with a smartphone equipped with a dual-microphone (one at the rear).

To extract acoustic features from the speech signals, the European Telecommunications Standards Institute front-end (ETSI FE, ES 201 108) is used [34, 3]. A 39-dimensional feature vector, comprising twelve Mel-frequency cepstral coefficients (MFCCs) along with the 0th order coefficient and their respective velocity and acceleration, is employed by the recognizer. The power spectra considered by our techniques are represented by $\mathcal{M} = 129$ frequency bins as the sampling frequency of the AURORA2-2C-CT/FT speech data is 8 kHz [34]. To obtain the cepstral coefficients for recognition, the discrete cosine transform (DCT) is applied to the enhanced 23-component log-Mel feature vectors computed from the power spectra. Finally, to improve the robustness of the system against channel mismatches, cepstral mean normalization (CMN) is applied.

Two sets of Gaussian mixture model (GMM)-based acoustic models are used for evaluation by employing the HTK toolkit: clean acoustic models trained on the Aurora-2 clean speech training dataset and multi-style models trained with distorted speech features to strengthen the ASR system against noisy conditions. In AURORA2-2C-CT/FT, their respective training datasets for multi-style acoustic modeling are built from the 8440 clean training utterances of Aurora-2. Similarly to [22], the multi-style training datasets consist of dual-

channel utterances contaminated with the same types of noise as in test set A, at the SNRs of 5 dB, 10 dB, 15 dB and 20 dB as well as the clean condition. To train the multi-style acoustic models, the multi-style training datasets are first compensated with the technique under evaluation. To model each digit in both sets of acoustic models, left-to-right continuous density hidden Markov models (HMMs) with 16 states and 3 Gaussians per state are employed. Silences and short pauses are modeled by HMMs with 3 and 1 states, respectively, and 6 Gaussians per state [22].

5.2. CHiME-3 and CHiME-4 Databases and Settings

CHiME-3 [1] and CHiME-4 [2] are novel frameworks especially intended for researching on multi-channel noise-robust speech recognition that include ASR baseline software which uses the Kaldi ASR toolkit [35]. The CHiME-3 and CHiME-4 databases are comprised of both simulated and real speech data. Real data were recorded in noisy environments by using a tablet with six microphones, five of them facing forward and one facing backwards. Similarly, simulated data were created by mixing clean speech utterances with background noise recordings. In particular, the speech data correspond to utterances from the well-known speaker-independent medium (5k) vocabulary subset of the Wall Street Journal (WSJ0) corpus [36]. The main difference between CHiME-3 and CHiME-4 is that the latter increases the level of difficulty of the former by constraining the number of microphones available for evaluation. Therefore, the rest of the description in this subsection is common to both frameworks and their differences are explicitly remarked.

Training data are composed by 8738 noisy utterances (1600 real plus 7138 simulated from the standard WSJ0 training dataset) in the four different noisy environments considered: public transport (BUS), café (CAF), pedestrian area (PED) and street junction (STR). Development and evaluation datasets are also defined separately for the simulated and real cases. Each development dataset contains 1640 utterances (410 from each noisy environment) while each evaluation dataset is comprised of 1320 utterances (330 per noisy environment).

Speech recognition tests are performed by using not only GMMs but also DNNs for multi-style acoustic modeling. Again, by using the ETSI FE front-end we compute a 13-dimensional feature vector that consists of twelve MFCCs along with the 0th order coefficient. In this case $\mathcal{M} = 257$ as the sampling frequency of the CHiME-3 and CHiME-4 speech data is 16 kHz [34].

For evaluation, the ASR engines provided by the CHiME Challenge organizers are used with no modifications. While these CHiME-3 and CHiME-4 baseline systems are briefly described down below, the reader is referred to [1] and [2] for further details.

For GMM-based acoustic modeling, three frames from the left and right temporal context are appended to each frame, which defines an augmented 91-dimensional MFCC feature vector. Then, a linear discriminant analysis (LDA) procedure (to reduce the number of components of the augmented feature vector to only 40) as well as maximum likelihood linear transformation (MLLT) and feature-space maximum likelihood linear regression (fMLLR) with speaker adaptive training (SAT) are applied. The resulting feature vectors are then used to train 2500 different tied triphone HMM states, which are modeled by a total of 15000 Gaussians [1].

On the other hand, for DNN-based acoustic modeling, the Kaldi recipe for Track 2 of the 2nd CHiME Challenge is considered [37]. A DNN with 7 hidden layers and 2048 neurons per layer is employed. In this case, five frames from the left and right temporal context are appended to each frame, which generates an augmented 143-dimensional MFCC feature vector that is used as input to the DNN. A generative pre-training using restricted Boltzmann machines (RBMs) as well as cross-entropy training and sequence-discriminative training employing the state-level minimum Bayes risk (sMBR) criterion [38] are performed on the DNN [1]. In first instance, the DNN is trained from the alignments generated by the above GMM-based ASR system. Then, once the DNN is trained, realignments are done and the DNN is re-trained from these new alignments. This procedure is repeated until completing four iterations [37].

In CHiME-3, a 3-gram language model is employed. On the other hand,

the ASR performance is improved in CHiME-4 (only when considering DNNs for acoustic modeling) by means of a 5-gram language model with Kneser-Ney smoothing [39] (5-gram KN), and a recurrent neural network language model (RNNLM)-based rescoring [40].

6. Experimental Results

In this section, our power spectrum enhancement proposals are evaluated on the different databases described in the previous section. We should recall here that the primary channel in the AURORA2-2C-CT/FT databases is identified with the primary microphone of the smartphone. On the other hand, in CHiME-3 and CHiME-4, a virtual primary channel is obtained by means of MVDR beamforming from all the six microphones in the tablet. With this arrangement, our WF-based weighting acts as a beamformer post-filter which helps to mitigate the beamformer weak points such as its poor performance at low frequencies or the effect of noise sources placed along the steering direction [21]. The use of MVDR in particular will be justified in the CHiME results subsection.

In addition to our proposals, other single-channel and multi-channel noise-robust techniques are evaluated for comparison purposes. First, the following three single-channel noise-robust methods are tested on the primary channel: a soft-mask weighting (SMW) technique in the log-Mel domain [33], the ETSI advanced front-end (AFE) [41] and a classic Wiener filtering with the same post-processing as in Subsection 4.4 (Wiener+Int). For Wiener+Int, $\mathcal{S}_{n_1}(f,t)$ in Eq. (7) is approximated from noise estimates obtained by means of a linear interpolation in the log-power spectral domain that uses the averages of the first and last $M = 20$ frames in each utterance. Noise estimation by linear interpolation is selected for a fair comparison, as the noise statistical parameters required by our noise equalizer are obtained by following this same approach (as explained in Subsection 4.2).

Our dual-channel power spectrum enhancement methods MMSN and DCSS (which also use, for CHiME, the virtual primary channel described above), pre-

viously proposed in [13], are also evaluated as a reference. In addition, two beamforming techniques are also tested for comparison: delay-and-sum (D&S) [42] and MVDR [43]. On the one hand, time difference of arrival (TDOA) estimation for D&S is performed as explained in [1] and [44]. On the other hand, the tested MVDR employs a state-of-the-art technique to estimate the steering vector based on eigenvalue decomposition (ED) where the clean speech spatial covariance matrix is derived from complex GMM-based time-frequency masks [43]. In addition, two post-filtering methods are evaluated in CHiME when applied after MVDR beamforming. The first one consists of a multi-channel Wiener post-filter (Lefkimmiatis) [21] using a noise coherence matrix estimated per utterance and frequency bin from the noise spatial correlation matrix as computed for MVDR beamforming [43]. The second post-filter considered is a multi-channel noise reduction post-filter as in [11] (MCNR), which also employs the steering vector from [43] for consistency. The baseline system uses noisy speech features from the primary channel in the case of the AURORA2-2C-CT/FT databases and from the fifth microphone in the case of CHiME-3 and CHiME-4 (as in [1]). Finally, our biased WF-based spectral weighting approach (Prop-B) is evaluated along with its unbiased version with and without noise equalization (Prop-U+Eq and Prop-U, respectively). Again for a fair comparison, the required RSG term $\mathcal{A}_{21}(f,t)$ is computed using (30) from the same ED-based steering vector estimator as the one used for MVDR beamforming [43] in the case of CHiME-3 and CHiME-4. However, fixed $\mathcal{A}_{21}(f)$ factors were *a priori* determined in AURORA2-2C-CT/FT from their corresponding development datasets to be used with these corpora as better performance is achieved by following this simpler approach. This might be related with the fact that estimating a steering vector from a few microphones (only two in our scenario) is prone to severe inaccuracies.

6.1. AURORA2-2C-CT/FT Results

Tables 1 and 2 summarize the word accuracy results obtained for the AURORA2-2C-CT database (close-talk) when clean and multi-style acoustic models are

SNR (dB)	Baseline	SMW	Wiener+Int	MMSN	DCSS	D&S	MVDR	Prop-B	Prop-U	Prop-U+Eq	AFE
-5	18.15	26.23	26.69	23.91	24.15	12.26	13.86	29.30	29.45	32.03	**35.81**
0	31.85	51.76	49.32	45.57	46.12	21.68	27.59	53.94	54.40	62.53	**65.46**
5	56.11	77.03	75.47	73.96	74.42	39.49	57.15	79.63	79.71	83.96	**85.16**
10	82.78	89.49	90.15	90.08	90.23	67.73	85.79	92.05	91.98	**94.73**	93.80
15	94.72	94.19	95.76	95.88	95.86	90.00	95.49	96.49	96.42	**97.36**	96.96
20	97.76	96.09	97.71	97.75	97.63	96.85	97.98	97.95	97.89	**98.42**	98.33
Clean	99.13	98.40	99.08	98.98	98.63	99.04	99.05	99.05	98.99	99.03	**99.24**
Average (-5 to 20)	63.56	72.47	72.52	71.19	71.40	54.67	62.98	74.89	74.98	78.17	**79.25**

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

SNR (dB)	Baseline	SMW	Wiener+Int	MMSN	DCSS	D&S	MVDR	Prop-B	Prop-U	Prop-U+Eq	AFE
-5	36.93	37.36	47.25	45.93	46.67	22.80	25.01	52.12	52.29	**54.26**	48.21
0	66.69	68.37	76.28	77.10	77.73	46.27	58.23	80.49	80.75	**82.81**	78.36
5	88.85	87.02	91.68	92.70	92.99	77.65	87.13	93.19	93.15	**94.12**	92.24
10	95.73	94.47	96.28	96.87	97.01	92.21	95.88	96.90	96.93	**97.27**	96.54
15	97.56	96.91	97.67	98.18	98.23	96.44	97.79	98.12	98.07	**98.27**	98.11
20	98.31	97.95	98.29	98.61	**98.68**	97.69	98.60	98.51	98.51	98.64	98.66
Clean	98.77	98.77	98.90	98.71	98.49	98.44	98.79	98.87	98.59	98.62	**99.07**
Average (-5 to 20)	80.68	80.35	84.58	84.90	85.22	72.18	77.11	86.57	86.62	**87.56**	85.35

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

SNR (dB)	Baseline	SMW	Wiener+Int	MMSN	DCSS	D&S	MVDR	Prop-B	Prop-U	Prop-U+Eq	AFE
-5	21.13	29.06	30.72	25.01	26.22	17.30	19.23	28.94	29.84	30.33	**39.37**
0	35.03	53.33	53.65	44.19	46.08	31.09	38.05	49.22	49.20	52.15	**68.06**
5	58.96	77.77	77.89	69.81	70.64	54.86	71.68	74.50	73.32	76.05	**86.66**
10	84.74	89.52	91.66	88.02	87.33	82.33	91.43	90.43	89.09	90.70	**94.59**
15	95.33	94.15	96.47	95.12	94.01	95.13	97.16	96.01	95.36	95.93	**97.30**
20	98.00	96.10	98.13	97.54	96.75	97.99	**98.50**	97.86	97.58	97.84	98.34
Clean	99.10	98.40	99.08	98.88	97.92	99.07	99.17	99.07	98.99	99.02	**99.24**
Average (-5 to 20)	65.53	73.32	74.75	69.95	70.17	63.12	69.34	72.83	72.40	73.83	**80.72**

Table 3: Word accuracy results (in terms of percentage and for different SNR values) obtained for the techniques evaluated on the AURORA2-2C-FT database when using clean acoustic models. Results are averaged across all types of noise in test sets *A* and *B*.

employed, respectively. Results are broken down by SNR and averaged across all types of noise in test sets *A* and *B*. As can be observed, the best result is obtained with multi-style acoustic models when our unbiased proposal with noise equalization (Prop-U+Eq) is applied, yielding a relative average improvement over the baseline of 6.88%. Moreover, this approach also presents the best behavior in the most adverse acoustic condition tested (-5 dB) with an absolute word accuracy of 54.26% and a relative improvement of 17.33% with respect to the baseline. As expected, since the speech component at the secondary channel of a dual-microphone smartphone can be safely neglected in close-talk conditions, Prop-U and Prop-B perform virtually the same.

The word accuracy results obtained for the AURORA2-2C-FT database (far-talk), when clean and multi-style acoustic models are employed, are shown in Tables 3 and 4, respectively. With a relative average improvement of 4.94% with respect to the baseline system using multi-style models, Prop-U+Eq is again the best approach according to the results. In addition, with an absolute word accuracy of 52.67% and a relative improvement of 14.25% with respect to the baseline, Prop-U+Eq with multi-style acoustic models is the best option at -5 dB as well. In this case, the speech component at the secondary channel is not negligible. Along with this, the relative position between the speaker and the

SNR (dB)	Baseline	SMW	Wiener+Int	MMSN	DCSS	D&S	MVDR	Prop-B	Prop-U	Prop-U+Eq	AFE
-5	38.42	36.40	48.83	46.91	47.45	29.64	31.74	48.20	49.65	**52.67**	50.64
0	67.81	66.77	76.88	76.37	76.69	57.82	68.20	75.61	76.22	79.27	**79.46**
5	89.81	87.69	91.96	91.87	92.05	84.67	91.58	91.74	91.66	**92.79**	92.54
10	96.20	94.72	96.56	96.57	96.56	95.47	**96.97**	96.52	96.43	96.93	96.85
15	97.80	97.15	97.79	97.74	97.80	97.79	**98.28**	97.87	97.84	98.02	98.22
20	98.46	98.06	98.32	98.36	98.40	98.54	**98.67**	98.59	98.46	98.48	98.65
Clean	98.76	97.23	98.81	98.54	98.45	98.60	98.85	98.79	98.64	98.63	**99.06**
Average (-5 to 20)	81.42	80.13	85.06	84.64	84.83	77.32	80.91	84.76	85.04	**86.36**	86.06

Table 4: Word accuracy results (in terms of percentage and for different SNR values) obtained for the techniques evaluated on the AURORA2-2C-FT database when using multi-style acoustic models. Results are averaged across all types of noise in test sets *A* and *B*.

device is more variable in far- than in close-talk conditions, and, therefore, the RSG is also more variable. This could explain the slight degradation of Prop-U regarding Prop-B when using clean acoustic models, as we are considering a fixed RSG factor. Recall that employing $\mathcal{A}_{21}(f,t)$ computed as in (30) from the ED-based steering vector estimator of [43] is even more harmful in this scenario than using a fixed RSG factor. As aforementioned, this might be related with the fact that estimating a steering vector from only two microphones (one of them placed in an acoustic shadow with respect to the source to be localized) is prone to errors.

While Prop-U+Eq with multi-style acoustic models achieves on average the highest results, AFE performs the best on AURORA2-2C-CT/FT when clean acoustic models are employed. In this case, recall that AFE involves multiple state-of-the-art strategies (e.g., a sophisticated two-stage Mel-warped Wiener filter approach that uses a VAD, and waveform processing and blind equalization stages [41, 3]), which are not incompatible with our proposals.

It is worth noticing that beamforming techniques do not provide a successful performance. D&S yields a drop in performance since it only aligns the target signals from each channel so the primary channel is eventually combined with a much noisier secondary one. On the other hand, while MVDR beamforming

additionally manages both the speech gains (through the steering vector) and the noise signals, it is only able to achieve an improvement under clean acoustic modeling in far-talk conditions with respect to the baseline. These results are coherent with the fact that a poor performance of the classic beamforming techniques can be expected with only two microphones very close each other, one of them (i.e., the secondary sensor) placed in an acoustic shadow with respect to the target signal [19, 20].

On average, Prop-U+Eq always outperforms both MMSN and DCSS, which can also be interpreted as spectral weighting techniques. In particular, MMSN follows an approach similar to MVDR beamforming but in the power spectral domain (i.e., it neglects the phase information). As can be observed, on average, MMSN outperforms MVDR in all conditions. In this sense we can consider MMSN as an *ad-hoc* MVDR to be used with small microphone arrays, where acoustic shadows are more important than (not accurately estimated) time delays.

6.2. CHiME-3 and CHiME-4 Results

Tables 5 and 6 report, for all the methods evaluated in this work, the word error rates (WERs) obtained on the CHiME-3 real data evaluation set when using multi-style GMM- and DNN-based acoustic models, respectively. In all cases, WERs are broken down by type of noise. Beamforming methods were tested by using either the signals from only the five microphones facing forward (5 ch.) or all the six microphones (6 ch.) in the tablet (i.e., by also including the sensor that faces backwards). Similarly to what happened with the AURORA2-2C-CT/FT corpora, considering the secondary sensor for D&S yields a drop in performance while MVDR modestly improves with respect to using only the five sensors facing forward. In this way, an MVDR with all the six microphones in the tablet is selected as the best beamforming choice and we apply it to obtain our virtual primary channel for CHiME, as aforementioned. Nevertheless, even better results may be obtained using a more sophisticated beamforming technique (e.g., a generalized sidelobe canceller (GSC) [45, 46] or a post-filtered

beamformer [47, 48, 49]) for the virtual primary channel.

As expected, similar trends are obtained by employing GMMs and DNNs for acoustic modeling. Moreover, the baseline WER from GMM-based acoustic modeling is 1.33% lower than that from DNN-based acoustic modeling as a result of the more sophisticated front-end used in the former case (as explained in Subsection 5.2). Nevertheless, all the tested techniques perform better by using DNN-based acoustic models than the GMM-based ones. As can be seen, the best result is achieved by Prop-U+Eq under DNN-based acoustic modeling, with an absolute WER of 16.77% and a relative average improvement of 17.23% and 1.87% with respect to the baseline and MVDR (6 ch.), respectively. Though the secondary channel has already been used to define the virtual primary one, these results reveal the convenience of treating the secondary signal in a differentiated manner since it can be further exploited to provide useful information about the acoustic environment. This is confirmed by the results in both absolute and relative terms since, on average and considering DNN-based acoustic models, the percentage change between MVDR (5 ch.) and MVDR (6 ch.) is 0.36% while the percentage change between MVDR (6 ch.) and Prop-U+Eq is 2.30%.

As we can see, using the ED-based steering vector estimator of [43] to derive the RSG in (30) leads Prop-U to improve Prop-B for both types of acoustic models. However, it is true that while Prop-U+Eq meaningfully improves Prop-U, this latter approach slightly enhances the results of Prop-B (which in turn worsens MVDR (6 ch.)). This is because MVDR beamforming yields a strong dehomogenization of the noise at the virtual primary and secondary channels. Under this circumstance, the homogeneity assumption underlying Prop-U is not accomplished, so the substantial improvement only comes when bias correction and noise equalization, which also relies on $\mathcal{A}_{21}(f,t)$, are applied together.

Wiener+Int, which can also be considered a single-channel post-filter and unlike both multi-channel post-filters (Lefkimmiatis and MCNR), is able to improve MVDR (6 ch.). Prop-U+Eq, the parameters of which are obtained in the same conditions as Wiener+Int, yields a relative average improvement of 1.83% under DNN-based acoustic modeling with respect to this technique.

	GMM-based acoustic modeling				
	BUS	CAF	PED	STR	Average
Baseline	49.64	32.72	27.30	21.03	32.67
SMW	33.09	21.39	22.33	18.27	23.77
Wiener+Int	32.04	14.92	15.64	14.01	19.15
MMSN	33.18	15.58	16.78	14.76	20.08
DCSS	34.19	14.90	16.95	15.58	20.41
D&S (5 ch.)	32.08	22.60	25.82	15.13	23.91
D&S (6 ch.)	35.13	25.03	27.50	16.25	25.98
MVDR (5 ch.)	34.08	17.35	18.16	15.02	21.15
MVDR (6 ch.)	32.90	16.83	17.40	14.72	20.46
Lefkimmiatis	42.31	14.79	19.04	17.48	23.41
MCNR	33.60	15.92	17.31	15.14	20.49
Prop-B	34.51	17.24	17.69	14.85	21.07
Prop-U	33.95	16.14	17.17	15.13	20.60
Prop-U+Eq	**29.48**	**13.02**	**14.82**	**13.86**	**17.80**
AFE	35.91	16.96	17.84	15.75	21.62

Table 5: Word error rate results (in terms of percentage and per type of noise) for the different techniques evaluated with CHiME-3 when multi-style acoustic models are employed. Results are from the real data evaluation set when considering GMMs for acoustic modeling.

	DNN-based acoustic modeling				
	BUS	CAF	PED	STR	Average
Baseline	51.13	35.06	28.31	21.48	34.00
SMW	31.13	17.99	19.08	17.09	21.32
Wiener+Int	31.20	13.13	15.88	14.19	18.60
MMSN	29.50	13.09	16.12	**13.39**	18.03
DCSS	29.52	13.02	15.70	13.99	18.06
D&S (5 ch.)	30.90	21.16	25.52	14.61	23.05
D&S (6 ch.)	33.82	22.81	26.49	15.09	24.55
MVDR (5 ch.)	29.89	14.66	16.54	14.62	18.93
MVDR (6 ch.)	29.50	14.79	16.37	13.88	18.64
Lefkimmiatis	35.02	15.26	17.73	16.29	21.08
MCNR	29.40	14.82	16.95	14.96	19.03
Prop-B	30.57	13.95	16.52	13.78	18.71
Prop-U	29.73	13.78	16.27	14.27	18.51
Prop-U+Eq	**26.94**	**11.80**	**14.63**	13.69	**16.77**
AFE	31.00	14.77	16.69	14.10	19.14

Table 6: Word error rate results (in terms of percentage and per type of noise) for the different techniques evaluated with CHiME-3 when multi-style acoustic models are employed. Results are from the real data evaluation set when considering DNNs for acoustic modeling.

	5-gram KN					*RNNLM*				
	BUS	CAF	PED	STR	Average	BUS	CAF	PED	STR	Average
Baseline	34.32	20.02	15.21	13.41	20.74	32.85	17.97	13.55	12.14	19.12
MVDR (6 ch.)	20.17	8.09	9.73	9.68	11.92	18.86	7.19	8.45	8.24	10.68
Wiener+Int	20.34	8.20	**9.42**	9.53	11.87	18.71	7.10	**8.00**	8.57	10.59
Prop-U+Eq	**18.08**	**7.38**	9.60	**8.93**	**11.00**	**16.41**	**6.00**	8.03	**7.88**	**9.58**

Table 7: Word error rate results (in terms of percentage and per type of noise) for different techniques evaluated with CHiME-4 when multi-style acoustic models are employed. Results are from the real data evaluation set when considering DNNs for acoustic modeling as well as a 5-gram language model with Kneser-Ney smoothing (5-gram KN) and a recurrent neural network language model (RNNLM)-based rescoring.

This confirms that the secondary microphone is providing additional valuable information about the ambient noise. Furthermore, unlike for the synthetic AURORA2-2C-CT/FT corpora, it should be noticed that AFE does not present a competitive performance on the CHiME-3 real data. In addition, Prop-U+Eq is again clearly superior to both MMSN and DCSS. Finally, the result of Table 6 for Prop-U+Eq with respect to those of the CHiME-3 Challenge provided in [1] (Table 3) demonstrates that our proposal is quite competitive if we take into account that it only applies multi-channel enhancement.

The more challenging CHiME-4 1-channel and 2-channel tracks are not addressed in this paper. First, the 1-channel track is not evaluated as we require two channels. Second, the 2-channel track is not considered because our dual-channel proposals require a secondary channel and such a track randomly selects signals only from the five front sensors of the tablet. Therefore, the most relevant methods according to the results achieved on the CHiME-3 corpus are evaluated with CHiME-4 again using all the microphones in the tablet (6-channel track). Table 7 reports the WERs obtained on the CHiME-4 real data evaluation set when using multi-style DNN-based acoustic models as well as a 5-gram language model with Kneser-Ney smoothing [39] (5-gram KN) and a recurrent neural network language model (RNNLM)-based rescoring [40]. Once again,

WERs are broken down by type of noise. We recall that both Wiener+Int and Prop-U+Eq behave as beamformer post-filters since they are applied after MVDR (6 ch.). As expected, the same result trends as in CHiME-3 have been obtained on CHiME-4. Thus, the best result is again achieved by Prop-U+Eq with an absolute average WER of 9.58% with RNNLM-based rescoring and a relative average improvement of 9.54% and 1.10% with respect to the baseline and MVDR (6 ch.), respectively. The benefit of using advanced language models can also be assessed, as both 5-gram KN and RNNLM significantly enhanced the 3-gram language model performance (see Table 6).

6.3. Summary of Results

The dual-channel proposals in this paper and comparison methods were evaluated in terms of speech recognition accuracy when applied on a dual-microphone smartphone (AURORA2-2C-CT/FT) and a tablet with six microphones (CHiME-3 and CHiME-4). Both devices have a rear microphone the main purpose of which is to get information about the acoustic environment. In the case of the tablet, a virtual primary channel was generated by application of MVDR beamforming, so our WF-based weighting behaved as a beamformer post-filter.

Results when correcting bias on the AURORA2-2C-CT/FT corpora by means of Prop-U were consistent with the fact that the speech channel between the primary and secondary sensors is more variable and yields less attenuation in far- than in close-talk conditions. In the case of CHiME, we observed that MVDR beamforming yielded a strong dehomogenization of the noise at the virtual primary and secondary channels, so the homogeneity assumption underlying Prop-U was not accomplished and the substantial improvement came when bias correction and noise equalization were jointly applied.

In general, our results proved the convenience of treating the secondary signal in a differentiated manner as such a signal is rather useful to provide information about the acoustic environment. Thus, beamforming exhibited a poor performance when integrating the secondary signal, as was expected. Our best results

were consistently achieved by our WF-based weighting with bias correction and noise equalization, Prop-U+Eq, which showed a remarkable performance at low SNRs.

7. Conclusions and Future Work

In this paper we have developed a novel spectral weighting approach for noise-robust ASR in mobile devices that is capable of exploiting the information provided by a secondary microphone placed in an acoustic shadow with respect to the speaker. A twofold contribution has been presented in this work: two new ways of estimating the *a priori* SNR in a dual-channel context, and a novel noise equalizer. According to our experimental results, we conclude that although beamforming is able to provide robustness in ASR with general microphone arrays [16, 17, 18], a proper integration of a secondary microphone can be achieved when that is treated in a suitable and differentiated manner, as it is proposed in this paper. Moreover, our results have shown the great versatility of our approach according to its high performance regardless of the analyzed mobile device and usage position. Future work will investigate the performance of our approach on other mobile devices with different small microphone array configurations and other beamformers to be used for virtual primary channel computation.

Acknowledgements

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

References

[1] J. Barker, R. Marxer, E. Vincent, S. Watanabe, The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines, in: Proc. of IEEE Automatic Speech Recognition and Understanding, December 13–17, Scottsdale, USA, 2015.

[2] E. Vincent, S. Watanabe, A. A. Nugraha, J. Barker, R. Marxer, An analysis of environment, microphone and data simulation mismatches in robust speech recognition, Computer Speech & Language 46 (2017) 535–557.

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

[4] T. Yoshioka, T. Nakatani, Noise model transfer: Novel approach to robustness against nonstationary noise, IEEE Transactions on Audio, Speech, and Language Processing 21 (2013) 2182–2192.

[5] M. Jeub, C. Herglotz, C. M. Nelke, C. Beaugeant, P. Vary, Noise reduction for dual-microphone mobile phones exploiting power level differences, in: Proc. of 37$^{\text{th}}$ International Conference on Acoustics, Speech, and Signal Processing, March 25–30, Kyoto, Japan, 2012, pp. 1693–1696.

[6] C. M. Nelke, C. Beaugeant, P. Vary, Dual microphone noise PSD estimation for mobile phones in hands-free position exploiting the coherence and speech presence probability, in: Proc. of 38$^{\text{th}}$ International Conference on Acoustics, Speech, and Signal Processing, May 26–31, Vancouver, Canada, 2013, pp. 7279–7283.

[7] J. Zhang, R. Xia, Z. Fu, J. Li, Y. Yan, A fast two-microphone noise reduction algorithm based on power level ratio for mobile phone, in: Proc. of 8$^{\text{th}}$ International Symposium on Chinese Spoken Language Processing, December 5–8, Hong Kong, 2012, pp. 206–209.

[8] Z. Fu, F. Fan, J. Huang, Dual-microphone noise reduction for mobile phone application, in: Proc. of 38$^{\text{th}}$ International Conference on Acoustics, Speech, and Signal Processing, May 26–31, Vancouver, Canada, 2013, pp. 7239–7243.

[9] I. A. McCowan, A. Morris, H. Bourlard, Improving speech recognition performance of small microphone arrays using missing data techniques,

in: Proc. of 7th International Conference of Spoken Language Processing, September 16–20, Denver, USA, 2002, pp. 2181–2184.

[10] N. Ma, R. Marxer, J. Barker, G. J. Brown, Exploiting synchrony spectra and deep neural networks for noise-robust automatic speech recognition, in: Proc. of IEEE Automatic Speech Recognition and Understanding, December 13–17, Scottsdale, USA, 2015.

[11] S. Zhao, X. Xiao, Z. Zhang, T. N. Nguyen, X. Zhong, B. Ren, L. Wang, D. L. Jones, E. S. Chng, H. Li, Robust speech recognition using beamforming with adaptive microphone gains and multichannel noise reduction, in: Proc. of IEEE Automatic Speech Recognition and Understanding, December 13–17, Scottsdale, USA, 2015.

[12] I. López-Espejo, J. A. González, A. M. Gomez, 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 (2014) 119–128.

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

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

[15] J. A. González, A. M. Peinado, N. Ma, A. M. Gomez, J. Barker, MMSE-based missing-feature reconstruction with temporal modeling for robust speech recognition, IEEE Transactions on Audio, Speech, and Language Processing 21 (2013) 624–635.

[16] M. Omologo, M. Matassoni, P. Svaizer, D. Giuliani, Experiments of hands-free connected digit recognition using a microphone array, in: Proc. of

IEEE Automatic Speech Recognition and Understanding, Santa Barbara, USA, 1997, pp. 490–497.

[17] T. B. Hughes, K. Hong-Seok, J. H. DiBiase, H. F. Silverman, Performance of an HMM speech recognizer using a real-time tracking microphone array as input, IEEE Transactions on Speech and Audio Processing 7 (1999) 346–349.

[18] I. A. McCowan, C. Marro, L. Mauuary, Robust speech recognition using near-field superdirective beamforming with post-filtering, in: Proc. of 25^{th} International Conference on Acoustics, Speech, and Signal Processing, June 5–9, Istanbul, Turkey, 2000, pp. 1723–1726.

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

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

[21] S. Lefkimmiatis, P. Maragos, A generalized estimation approach for linear and nonlinear microphone array post-filters, Speech Communication 49 (2007) 657–666.

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

[23] T. V. Sreenivas, P. Kirnapure, Codebook constrained Wiener filtering for speech enhancement, IEEE Transactions on Speech and Audio Processing 4 (1996) 383–389.

[24] L. Lin, W. H. Holmes, E. Ambikairajah, Subband noise estimation for speech enhancement using a perceptual Wiener filter, in: Proc. of 28$^{\text{th}}$ International Conference on Acoustics, Speech, and Signal Processing, April 6–10, Hong Kong, 2003.

[25] X. Bingyin, B. Changchun, Wiener filtering based speech enhancement with weighted denoising auto-encoder and noise classification, Speech Communication 60 (2014) 13–29.

[26] J. S. Lim, A. V. Oppenheim, Enhancement and bandwidth compression of noisy speech, Proceedings of the IEEE 67 (1979) 1586–1604.

[27] J. S. Erkelens, R. Heusdens, Fast noise tracking based on recursive smoothing of MMSE noise power estimates, in: Proc. of 33$^{\text{rd}}$ International Conference on Acoustics, Speech, and Signal Processing, March 30–April 4, Las Vegas, USA, 2008.

[28] B. Truax, Handbook for Acoustic Ecology, Cambridge St. Pub., 1999.

[29] M. Berouti, R. Schwartz, J. Makhoul, Enhancement of speech corrupted by acoustic noise, in: Proc. of 4$^{\text{th}}$ International Conference on Acoustics, Speech, and Signal Processing, April 2–4, Washington D.C., USA, 1979, pp. 208–211.

[30] S. Kamath, P. Loizou, A multi-band spectral subtraction method for enhancing speech corrupted by colored noise, in: Proc. of 27$^{\text{th}}$ International Conference on Acoustics, Speech, and Signal Processing, May 13–17, Orlando, USA, 2002, pp. IV–4164.

[31] S. V. Vaseghi, Advanced Digital Signal Processing and Noise Reduction (4th Edition), John Wiley & Sons, 2008.

[32] R. Martin, Spectral subtraction based on minimum statistics, in: Proc. of 7$^{\text{th}}$ European Signal Processing Conference, September, Edinburgh, Scotland, 1994, pp. 1182–1185.

[33] J. Hout, A. Alwan, A novel approach to soft-mask estimation and log-spectral enhancement for robust speech recognition, in: Proc. of 37$^{\text{th}}$ International Conference on Acoustics, Speech, and Signal Processing, March 25–30, Kyoto, Japan, 2012, pp. 4105–4108.

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

[35] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, The Kaldi speech recognition toolkit, in: Proc. of IEEE Automatic Speech Recognition and Understanding, December 11–15, Waikoloa, USA, 2011.

[36] D. Paul, J. Baker, The design of Wall Street Journal-based CSR corpus, in: Proc. of 2$^{\text{nd}}$ International Conference of Spoken Language Processing, October, Alberta, Canada, 1992, pp. 899–902.

[37] C. Weng, D. Yu, S. Watanabe, B. H. Juang, Recurrent deep neural networks for robust speech recognition, in: Proc. of 39$^{\text{th}}$ International Conference on Acoustics, Speech, and Signal Processing, May 4–9, Florence, Italy, 2014, pp. 5532–5536.

[38] K. Veselý, A. Ghoshal, L. Burget, D. Povey, Sequence-discriminative training of deep neural networks, in: Proc. of 14$^{\text{th}}$ Annual Conference of the International Speech Communication Association, August 25–29, Lyon, France, 2013, pp. 2345–2349.

[39] R. Kneser, H. Ney, Improved backing-off for M-gram language modeling, in: Proc. of 20$^{\text{th}}$ International Conference on Acoustics, Speech, and Signal Processing, May 9–12, Detroit, USA, 1995, pp. 181–184.

[40] T. Mikolov, M. Karafiát, L. Burget, J. H. Černocký, S. Khudanpur, Recurrent neural network based language model, in: Proc. of 11$^{\text{th}}$ Annual Conference of the International Speech Communication Association, September 26–30, Chiba, Japan, 2010, pp. 1045–1048.

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

[42] L. Ziomek, Fundamentals of Acoustic Field Theory and Space-Time Signal Processing, CRC Press, 1994.

[43] T. Higuchi, N. Ito, T. Yoshioka, T. Nakatani, Robust MVDR beamforming using time-frequency masks for online/offline ASR in noise, in: Proc. of 41st International Conference on Acoustics, Speech, and Signal Processing, March 20–25, Shanghai, China, 2016.

[44] C. Blandin, E. Vincent, A. Ozerov, Multi-source TDOA estimation in reverberant audio using angular spectra and clustering, Signal Processing 92 (2012) 1950–1960.

[45] L. Pfeifenberger, T. Schrank, M. Zohrer, M. Hagmüller, F. Pernkopf, Multi-channel speech processing architectures for noise robust speech recognition: 3rd CHiME challenge results, in: Proc. of IEEE Automatic Speech Recognition and Understanding, December 13–17, Scottsdale, USA, 2015.

[46] L. Griffiths, C. Jim, An alternative approach to linearly constrained adaptive beamforming, IEEE Transactions on Antennas and Propagation 30 (1982) 27–34.

[47] T. Menne, J. Heymann, A. Alexandridis, K. Irie, A. Zeyer, M. Kitza, P. Golik, I. Kulikov, L. Drude, R. Schlüter, H. Ney, R. Haeb-Umbach, A. Mouchtaris, The RWTH/UPB/FORTH system combination for the 4th CHiME challenge evaluation, in: Proc. of 17th Annual Conference of the International Speech Communication Association, September 8–12, San Francisco, USA, 2016.

[48] C. Marro, Y. Mahieux, K. U. Simmer, Analysis of noise reduction and dereverberation techniques based on microphone arrays with postfiltering, IEEE Transactions on Speech and Audio Processing 6 (1998) 240–259.

[49] X. Xiao, S. Watanabe, H. Erdogan, L. Lu, J. Hershey, M. L. Seltzer, G. Chen, Y. Zhang, M. Mandel, D. Yu, Deep beamforming networks for multi-channel speech recognition, in: Proc. of 41st International Conference on Acoustics, Speech, and Signal Processing, March 20–25, Shanghai, China, 2016.

Iván López-Espejo received the M.Sc. degree in Telecommunications Engineering and the M.Sc. degree in Electronics Engineering both from the University of Granada (UGR), Granada, Spain, in 2011 and 2013, respectively. He is currently working towards the Ph.D. degree in robust speech recognition on multi-microphone mobile devices at the UGR.

Since 2013, he has been with the Signal Processing, Multimedia Transmission and Speech/Audio Technologies (SigMAT) Group, Department of Signal Theory, Telematics and Communications, UGR. His research interests are on robust speech recognition, noise estimation and signal processing.

Antonio M. Peinado received the M.S. and Ph.D. degrees in Physics from the University of Granada, Granada, Spain, in 1987 and 1994, respectively. He has held the positions of Associate Professor from 1996 to 2010 and has been Full Professor since 2010 with the Department of Signal Theory, Networking and Communications, University of Granada, where he is currently the Head of the research group on Signal Processing, Multimedia Transmission and Speech/Audio Technologies (SigMAT). His current research interests are focused on robust speech recognition and transmission, robust image/video transmission, and ultrasound and proteomic signal processing.

Angel M. Gomez received the M.Sc. and Ph.D. degrees in Computer Science from the University of Granada, Spain, in 2001 and 2006, respectively.

In 2002 he joined the Signal Theory, Networking, and Communications Department, University of Granada, where he is a member of the research group on Signal Processing, Multimedia Transmission and Speech/Audio Technologies (SigMAT). Currently he is an associate professor at the University of Granada. His research interests are on robust speech recognition, speech and audio coding, and multimedia transmission.

Jose A. Gonzalez received the B.Sc. and Ph.D. degrees in Computer Science both from the University of Granada, Spain, in 2006 and 2013, respectively.

Since 2013, he is a Research Associate in the Department of Computer Science, University of Sheffield, U.K., working in clinical applications of speech technology. His research interests include human speech processing, robust automatic speech recognition and machine learning.