FEATURE ENHANCEMENT FOR ROBUST SPEECH RECOGNITION ON SMARTPHONES WITH DUAL-MICROPHONE

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

[*] Dept. of Signal Theory, Telematics and Communications, University of Granada, Spain
[†] Dept. of Computer Science, University of Sheffield, UK
{iloes,amgg,amp}@ugr.es, j.gonzalez@sheffield.ac.uk

ABSTRACT

Latest smartphones often have more than one microphone in order to perform noise reduction. Although research on speech enhancement is already exploiting this new feature, robust speech recognition is not still benefiting from it. In this paper we propose two feature enhancement methods especially developed for the case of a smartphone with a dual-microphone operating in an adverse acoustic environment. In order to test these proposals, we have already developed a new experimental framework which includes a noisy speech database (based on AURORA2) which emulates the acquisition of dual-microphone data. Our experimental results show a clear improvement in terms of word accuracy in comparison with both using a power level difference-based speech enhancement algorithm and a single channel feature compensation approach.

Index Terms— Dual-microphone, Robust speech recognition, Feature enhancement, Smartphone, AURORA2-2C

1. INTRODUCTION

Over the last years, mobile devices such as smartphones have experienced an important growth in terms of sales and capabilities. Speech recognition technology has taken advantage of the latter, carrying to even more extended functionalities such as information search, dictation or call center interaction. We can use this kind of applications anytime, anywhere, usually taking place in noisy acoustic environments. For this reason it is crucial to tackle with the noise that contaminates the speech in order to ensure a good recognition performance [1].

Latest mobile devices include more than one microphone to perform noise reduction. This feature is employed for speech enhancement tasks, but it is not being exploited for robust speech recognition. For instance, the speech enhancement algorithms described in [2–4] are intended to work in a dual-microphone configuration. All of them are based on the power level difference (PLD) between the two microphones of the device. The PLD approach establishes that, in a conversational position (i.e. phone loudspeaker placed at the ear), the target speech power is greater at the primary microphone than at the secondary one (which is usually located at the rear or upper part of the device), while, assuming far field noise, the received noise signal powers are approximately the same at both microphones. Other papers such as [5, 6] propose speech enhancers destined to work on smartphones in a hands-free position.

In this work we present a pair of novel feature enhancement techniques intended to take advantage of the dual-microphone feature for robust speech recognition. The main idea behind these algorithms is to estimate the clean speech power spectrum in the first channel (related to the primary microphone of the device) by using the information contained in both channels. These methods are evaluated in accordance with a conversational position in two different ways: as standalone techniques and as preprocessing techniques. In this last case, the enhanced features are fed into a vector Taylor series (VTS) [7] feature compensation stage in order to further improve the recognizer performance by removing the remaining residual noise after the first enhancement. Our methods are compared with VTS along with a PLD-based speech enhancement technique in order to show the effectiveness of our proposals in terms of word accuracy.

All experiments have been carried out by using a dual-mic noisy speech database, that will be referred to as AURORA2-2C (AURORA2 - 2 Channels - Conversational Position), which is based on the well-known AURORA2 [8]. AURORA2-2C tries to emulate the noisy speech database that could have been acquired with a dual-microphone smartphone.

The rest of the paper is organized as follows: in Section 2 the dual-microphone configuration and the two proposed feature enhancement techniques are presented. In 3 the generation of the AURORA2-2C database is described. Experiments and a discussion of the results are shown in Section 4. Finally, conclusions are summarized in Section 5.

2. PROPOSED METHODS

Hereinafter we consider a noisy speech signal $y_i(m)$ which can be expressed as the sum of a clean speech signal $x_i(m)$ and a noise $n_i(m)$, i.e. $y_i(m) = x_i(m) + n_i(m)$, where

This work has been supported by the MICINN TEC2010-18009 project.

$i = 1, 2$ indicates the microphone that captures the signal (the primary one at the bottom of the device and the one at the rear, respectively). Assuming that speech and noise are independent, this additive model can be expressed in terms of power spectra as,

$$|Y_1(k,t)|^2 = |X_1(k,t)|^2 + |N_1(k,t)|^2; \quad (1)$$

$$|Y_2(k,t)|^2 = |X_2(k,t)|^2 + |N_2(k,t)|^2, \quad (2)$$

where k and t denote the frequency bin and time frame number, respectively ($k = 0, ..., \mathcal{M}-1; t = 0, ..., \mathcal{T}-1$).

In the following subsections we will develop a pair of feature enhancement techniques for the estimation of $|X_1(k,t)|^2$ from noise statistics.

2.1. Minimum mean square noise (MMSN) feature enhancer

This first algorithm is based on the minimum variance distortionless response (MVDR) beamformer [9]. According to it, the proposed linear estimator of the clean speech power spectrum bin k at time t in the first channel can be expressed as,

$$|\hat{X}_1(k,t)|^2 = \mathbf{w}_k^{\mathrm{T}} \mathbf{Y}(k,t) = \mathbf{w}_k^{\mathrm{T}} \begin{pmatrix} |Y_1(k,t)|^2 \\ |Y_2(k,t)|^2 \end{pmatrix}, \quad (3)$$

where $\mathbf{w}_k$ is a 2×1 weighting vector that must be estimated from the dual signal. We assume that the speech power in the second channel is related with the speech power in the first one through a time-invariant factor $A_{21}(k)$, i.e. $|X_2(k,t)|^2 = A_{21}(k)|X_1(k,t)|^2$. This factor can be interpreted as the target speech signal transfer function between the two microphones. This time-independent approach for $A_{21}(k)$ is appropriate if we assume a fixed relative position between the speaker and the smartphone. This seems a reasonable assumption for the case of adopting a conversational position. In this way, (2) can be rewritten as,

$$|Y_2(k,t)|^2 = A_{21}(k)|X_1(k,t)|^2 + |N_2(k,t)|^2. \quad (4)$$

In accordance with (4), the estimator in (3) can be now expressed as,

$$\begin{aligned} |\hat{X}_1(k,t)|^2 &= \mathbf{w}_k^{\mathrm{T}} |X_1(k,t)|^2 \begin{pmatrix} 1 \\ A_{21}(k) \end{pmatrix} + \\ &+ \mathbf{w}_k^{\mathrm{T}} \begin{pmatrix} |N_1(k,t)|^2 \\ |N_2(k,t)|^2 \end{pmatrix}. \end{aligned} \quad (5)$$

Our goal now is to obtain the weighting vector that minimizes the mean square noise-dependent term in (5), that is,

$$\mathbf{w}_k = \arg\min_{\mathbf{w}_k} \mathrm{E}\left[\left(\mathbf{w}_k^{\mathrm{T}} \mathbf{N}_k(t)\right)^2\right], \quad (6)$$

where $\mathbf{N}_k(t) = \left(|N_1(k,t)|^2, |N_2(k,t)|^2\right)^{\mathrm{T}}$. Rearranging terms, (6) is now rewritten as,

$$\begin{aligned} \mathbf{w}_k &= \arg\min_{\mathbf{w}_k} \mathbf{w}_k^{\mathrm{T}} \mathrm{E}\left[\mathbf{N}_k(t) \mathbf{N}_k(t)^{\mathrm{T}}\right] \mathbf{w}_k \\ &= \arg\min_{\mathbf{w}_k} \mathbf{w}_k^{\mathrm{T}} \mathbf{C}_{N,k} \mathbf{w}_k. \end{aligned} \quad (7)$$

$\mathbf{C}_{N,k}$ is the following 2×2 correlation matrix:

$$\mathbf{C}_{N,k} = \begin{pmatrix} c_{N,k}(1,1) & c_{N,k}(1,2) \\ c_{N,k}(2,1) & c_{N,k}(2,2) \end{pmatrix}, \quad (8)$$

where $c_{N,k}(i,j) = \mathrm{E}\left[|N_i(k,t)|^2 |N_j(k,t)|^2\right]$. The minimization in (7) must be subjected to the distortionless speech constraint

$$\mathbf{w}_k^{\mathrm{T}} \begin{pmatrix} 1 \\ A_{21}(k) \end{pmatrix} = 1, \quad (9)$$

therefore, the optimization problem can be solved by Lagrange multipliers using the cost function $f(\mathbf{w}_k, \lambda) = \mathbf{w}_k^{\mathrm{T}} \mathbf{C}_{N,k} \mathbf{w}_k - \lambda \left(\mathbf{w}_k^{\mathrm{T}} (1, A_{21}(k))^{\mathrm{T}} - 1\right)$. We can obtain the optimal weighting vector by solving $\nabla f(\mathbf{w}_k, \lambda) = 0$ as,

$$\mathbf{w}_k = \frac{\mathbf{C}_{N,k}^{-1} (1, A_{21}(k))^{\mathrm{T}}}{(1, A_{21}(k)) \mathbf{C}_{N,k}^{-1} (1, A_{21}(k))^{\mathrm{T}}}. \quad (10)$$

2.2. Dual-channel spectral subtraction

Along with the relationship defined in (4) we can also relate the noise power spectra in the first and second channels as $|N_1(k,t)|^2 = G_{12}(k,t)|N_2(k,t)|^2$. In such a case, (1) can be rewritten as,

$$|Y_1(k,t)|^2 = |X_1(k,t)|^2 + G_{12}(k,t)|N_2(k,t)|^2. \quad (11)$$

Similarly to $A_{21}(k)$, factor $G_{12}(k,t)$ can be understood as the frequency response of a new linear filter that relates the noise signals captured by the two microphones. Although this filter will be time-variant according to the scenario geometry (e.g. the relative position between the moving noise sources and the smartphone), we will assume that the noise sources do not meaningfully move during the speech utterance, which involves time invariance, that is, $G_{12}(k,t) \approx G_{12}(k)$. Notice that the magnitude of $G_{12}(k)$ in far field noise conditions, according to the explanation in Section 1, is around the unity.

Combining equations (11) and (4) we obtain the following dual-channel spectral subtraction (DCSS) estimator for bin k at time t:

$$|\hat{X}_1(k,t)|^2 = \frac{|Y_1(k,t)|^2 - G_{12}(k)|Y_2(k,t)|^2}{1 - G_{12}(k) A_{21}(k)}. \quad (12)$$

For every frequency bin we can estimate the noise gain factor $G_{12}(k)$ by minimizing the mean square error

$$E_k = \mathrm{E}\left[\left(|N_1(k,t)|^2 - G_{12}(k)|N_2(k,t)|^2\right)^2\right]. \quad (13)$$

Solving $\partial E_k / \partial G_{12}(k) = 0$, we can derive the desired estimate as,

$$\hat{G}_{12}(k) = \frac{\hat{c}_{N,k}(1,2)}{\hat{c}_{N,k}(2,2)}. \quad (14)$$

2.3. Implementation issues

In practice, the correlation matrix $\mathbf{C}_{N,k}$ is estimated for every k by using the M first and last frames of each utterance, since it is considered that those frames are only noise.

Factor $A_{21}(k)$ was pre-estimated for every frequency bin from stereo clean speech power spectra as the median of a set of ratios $\{|X_2(k,t)|^2/|X_1(k,t)|^2\}$ obtained from a validation dataset different from the one described in the next section. The median operator is applied in order to avoid the influence of outliers.

The reliability of the estimated $G_{12}(k)$ in (14) is critically reduced at very high signal-to-noise ratios (SNRs). In order to avoid this fact, an *a posteriori* SNR of the primary channel is calculated for the utterance being processed: the noise power is computed from the first and last 100ms while the noisy speech power is obtained from the rest of the utterance. Thus, if the SNR exceeds 30dB, $G_{12}(k)$ is set to zero, which involves that $|\hat{X}_1(k,t)|^2 = |Y_1(k,t)|^2$.

Finally, the estimates obtained in (3) and (12) are bounded in order to avoid possible negative power spectrum bins:

$$|\tilde{X}_1(k,t)|^2 = \max(|\hat{X}_1(k,t)|^2, \gamma|Y_1(k,t)|^2), \qquad (15)$$

where $\gamma \ll 1$ is a thresholding factor fixed by means of preliminary experiments.

3. THE AURORA2-2C DATABASE

Since a noisy speech database suitable to operate within the framework of this paper is not available, we have created the AURORA2-2C (AURORA2 - 2 Channels - Conversational Position) database for that purpose, which is based on the well-known AURORA2 [8]. Figure 1 depicts the block diagram of the scheme applied to generate the proposed database. The clean speech captured by the primary microphone, $x_1(m)$, is received at the secondary microphone transformed according to the acoustic path $h_{21}(m)$ between both mics. Thus, $x_2(m) = h_{21}(m) * x_1(m)$. Original (real) stereo noise $\{n'_i(m); i = 1, 2\}$ is scaled by the same gain factor G in order to get a specific SNR for $y_1(m)$. Additionally, a noise $\epsilon_2(m)$ is artificially added to model the silence (an almost negligible signal component) at the second channel. This is necessary since the original silence component in $x_1(m)$ is almost deleted by the acoustic path $h_{21}(m)$.

We assume that $h_{21}(m)$ takes into account the geometry of the problem (i.e. the fixed relative position between the speaker and the smartphone) but not the acoustic environment. This simplification is mostly needed for the next reason: the acoustics at a given environment are based on the distribution of different physical elements in space, therefore, for a perfect channel modeling it would be necessary to capture stereo clean speech into the corresponding noisy acoustic environment, which is physically impossible. Furthermore, notice that AURORA2 does not contemplate the effect of the different noisy acoustic environments for the clean speech signals either.

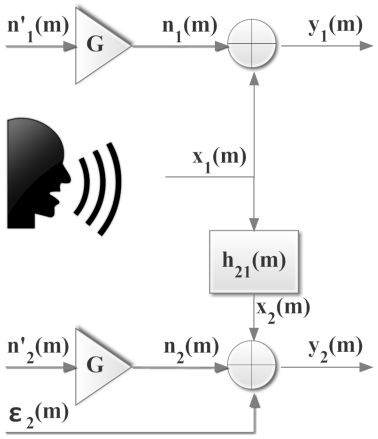

Fig. 1. The AURORA2-2C generation block diagram.

The clean speech signals included in AURORA2 are used here as the $x_1(m)$ signals. In order to get $x_2(m)$ from $x_1(m)$, the channel response $h_{21}(m)$ was modeled as a time-invariant FIR filter trained from speech recorded in a semi-anechoic environment. Stereo clean speech signals $x_1^{(tr)}(m)$ and $x_2^{(tr)}(m)$ were recorded for that purpose, in a conversational position, with a Sony Ericsson Xperia neo V (whose geometry is illustrated in Figure 2). We estimated our FIR filter through the minimization of the mean square error $\mathrm{E}[e^2(m)]$, where

$$e(m) = x_2^{(tr)}(m) - \sum_{l=0}^{p-1} \hat{h}_{21}(l) x_1^{(tr)}(m-l). \qquad (16)$$

The stereo noise signals $n'_1(m)$ and $n'_2(m)$ were recorded with the same device as for the speech, again in a conversational position, in some places where the use of a smartphone is probable: bus, babble, car, pedestrian street, cafe, street, bus station and train station. By using the application FaNT (Filtering and Noise Adding Tool) [10], the noise gain factor G was calculated as defined in [8]. Moreover, noise $\epsilon_2(m)$ was generated so that the resulting silence noise at the second channel has the same statistical distribution as the silence noise present in the first channel.

During training, the AURORA2 clean training set is used with no modification. For testing, two new sets are defined. We start from the 4 test subsets of AURORA2 with 1001 utterances in each one. Signals of each noise type are added to each subset at SNRs of 20dB, 15dB, 10dB, 5dB, 0dB and -5dB. The clean case is taken as a seventh condition too. As in [8], noise and speech are filtered with the G.712 characteristic before signal addition. In this way, we assume that speech and noise signals have been recorded with a similar equipment. Noises bus, babble, car and pedestrian street, are added to the 4 subsets in order to generate the first test

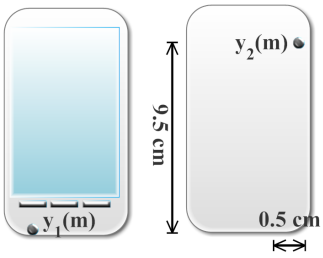

Fig. 2. Geometry of the device used for AURORA2-2C.

set, called test set *A*, which is composed of 28028 utterances (1001 utterances per subset $\times$ 4 subsets $\times$ 7 SNRs). The second test set (test set *B*) is defined similarly but using noises cafe, street, bus station and train station.

4. EXPERIMENTAL RESULTS

4.1. Experimental framework

The European Telecommunication Standards Institute front-end (ETSI FE, ES 201 108) is used to extract acoustic features from the speech signal. For feature enhancement, 129-component power spectral feature vectors are employed. Twelve Mel-frequency cepstral coefficients (MFCCs) along with the 0th order coefficient and their respective velocity and acceleration form the 39 dimensional feature vector used by the recognizer. Cepstral mean normalization (CMN) is applied to improve the robustness of the system to channel mismatches. For the recognizer, acoustic models trained only on clean speech are employed. Left to right continuous density hidden Markov models (HMMs) with 16 states and 3 Gaussians per state are used to model each digit. Silences and short pauses are modeled by HMMs with 3 and 1 states, respectively, and 6 Gaussians per state.

Different techniques are compared in terms of word recognition accuracy when they are evaluated using the AURORA2-2C database. All methods were tested both as standalone techniques and in combination with a 1st order vector Taylor series feature compensation algorithm (VTS-1). The VTS implementation was the one reported in [11], except that noise estimates are obtained by linear interpolation between the averages of the first and last 20 frames. VTS compensation is performed using a 256-component Gaussian mixture model (GMM) with diagonal covariance matrices. GMM training is performed by the expectation-maximization (EM) algorithm on the same dataset used for acoustic model training. A standalone VTS-1 was also evaluated as reference, while the baseline consists of the results obtained when the noisy speech features are employed. For both cases, the signals from the first channel, $y_1(m)$, were used.

Our proposals were compared with the speech enhancement technique in [2] (PLD). The final smoothing stage of this PLD-based algorithm was avoided since a small degradation of the recognizer performance was experimentally found.

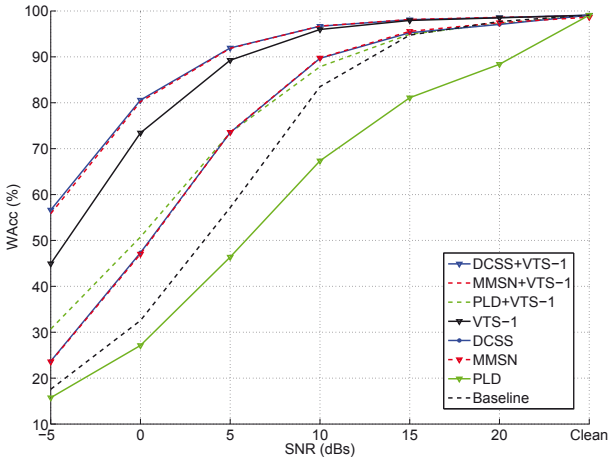

Fig. 3. Word accuracy vs. SNR for the different techniques tested.

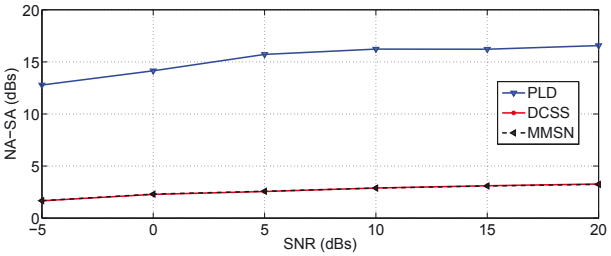

Fig. 4. NA-SA vs. SNR for the different methods evaluated.

Thresholding factor $\gamma = 0.05$ was determined by recognition experiments on an independent validation set. Finally, $M = 20$ was chosen.

4.2. Results

Word accuracy results for different SNR values, and averaged along all types of noise in test sets *A* and *B*, are shown in Table 1. A comparative graph can be seen in Figure 3. We observe that both of our proposed techniques (DCSS and MMSN) provide a similar performance: as standalone methods they produce an increase of the recognizer accuracy in comparison with the baseline case, being the best results obtained when these techniques are used as preprocessing algorithms for VTS-1. It can be shown that the expressions of MMSN tend to those of DCSS (and vice versa) when $A_{21}(k) \rightarrow 0$. For conversational position $A_{21}(k)$ factor has a relatively small magnitude, what justifies the similar performance of DCSS and MMSN. On the other hand, as can be observed, the evaluated speech enhancement method (PLD) is not adequate for speech recognition purposes. Although this kind

Method / SNR (dB)	-5	0	5	10	15	20	Clean	Avg. (-5 to 20)
Baseline	17.57	32.55	57.16	83.51	94.74	97.71	99.10	**63.87**
PLD	15.74	27.12	46.38	67.35	81.10	88.39	99.09	**54.35**
MMSN	23.58	47.04	73.58	89.77	95.56	97.45	98.64	**71.16**
DCSS	23.70	47.34	73.52	89.69	95.27	97.06	99.08	**71.10**
VTS-1	44.94	73.44	89.27	95.97	97.93	98.52	99.06	**83.35**
PLD+VTS-1	30.70	50.75	73.47	87.85	94.93	97.25	99.07	**72.49**
MMSN+VTS-1	55.83	80.23	91.85	96.72	98.15	98.61	98.94	**86.90**
DCSS+VTS-1	56.64	80.60	91.95	96.68	98.10	98.56	99.09	**87.09**

Table 1. Word accuracy (in terms of percentage and for different SNR values) obtained for the evaluated techniques. Results are averaged along all types of noise in test sets *A* and *B*.

of techniques improve the speech quality, they may not be appropriate for the recognition task. This idea has been reported in the literature, for instance in [12], where it has been experimentally demonstrated that some speech enhancement methods even produce degradation in terms of word accuracy. Figure 4 shows the noise attenuation minus speech attenuation (NA-SA) measure, used in [2] and defined in [13], for the different techniques evaluated. The bigger this measure, the better the objective speech quality. According to the comparison, as expected, the PLD-based algorithm is a good speech enhancer with respect to our proposals. This fact was confirmed through informal subjective listening opinions. In this way, the need for developing specific dual-channel feature enhancement methods for speech recognition is clearly supported.

5. CONCLUSIONS

In this paper we have proposed two feature enhancement techniques and a novel methodology for generating a noisy speech database for robust speech recognition on smartphones with dual-microphone. Our results have shown the utility of taking advantage of the signal captured by the extra microphone dedicated to noise reduction. Additionally, the AURORA2-2C database could serve for other evaluations in the future. As future work, dynamic and better parameter estimation for our methods will be investigated. Also, we aim to extend and evaluate these techniques in hands-free conditions.

REFERENCES

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

[2] M. Jeub, C. Herglotz, C.M. Nelke, C. Beaugeant, and P. Vary, "Noise reduction for dual-microphone mobile phones exploiting power level differences," in *Proc. ICASSP*, Kyoto, Japan, 2012, pp. 1693–1696.

[3] 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 *Proc. ISCSLP*, Hong Kong, 2012, pp. 206–209.

[4] N. Yousefian, A. Akbari, and M. Rahmani, "Using power level difference for near field dual-microphone speech enhancement," *Applied Acoustics*, vol. 70, pp. 1412–1421, December 2009.

[5] C.M. Nelke, C. Beaugeant, and P. Vary, "Dual microphone noise PSD estimation for mobile phones in hands-free position exploiting the coherence and speech presence probability," in *Proc. ICASSP*, Vancouver, Canada, 2013, pp. 7279–7283.

[6] H. Thumchirdchupong and N. Tangsangiumvisai, "A two-microphone noise reduction scheme for hands-free telephony in a car environment," in *Proc. ECTI-CON*, Krabi, Thailand, 2013, pp. 1–6.

[7] P.J. Moreno, B. Raj, and R.M. Stern, "A vector Taylor series approach for environment-independent speech recognition," in *Proc. ICASSP*, Atlanta, USA, 1996, pp. 733–736.

[8] D. Pearce and H.G. Hirsch, "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proc. ICSLP*, Beijing, China, 2000.

[9] Himawan I., *Speech recognition using ad-hoc microphone arrays*, Ph.D. thesis (Queensland University of Technology), 2010.

[10] G. Hirsch, "FaNT - Filtering and noise adding tool, http://dnt.kr.hsnr.de/download.html," 2005.

[11] J.C. Segura, A. de la Torre, M.C. Benítez, and A.M. Peinado, "Model-based compensation of the additive noise for continuous speech recognition. Experiments using the AURORA II database and tasks," in *Proc. Eurospeech*, Aalborg, Denmark, 2001.

[12] K.K. Paliwal, J.G. Lyons, S. So, A.P. Stark, and K.K. Wójcicki, "Comparative evaluation of speech enhancement methods for robust automatic speech recognition," in *Proc. ICSPCS*, Gold Coast, Australia, 2010, pp. 1–5.

[13] M. Pawig and P. Vary, "Quality investigations of network based acoustic noise reduction," in *Proc. ESSV*, Dresden, Germany, 2011, pp. 325–332.