Statistical analysis and performance of DFT domain noise reduction filters for robust speech recognition

Colin Breithaupt and Rainer Martin

Institute of Communications Acoustics (IKA)
Ruhr-University Bochum, 44780 Bochum, Germany
{colin.breithaupt,rainer.martin}@rub.de

Abstract

Noise reduction frontends have been developed independently for speech communication and speech recognition purposes with the result that one and the same algorithm does not perform well in both application domains. In this paper we show that noise reduction filters based on the discrete Fourier transform (DFT) which are used in speech communication can also perform well in robust automatic speech recognition (ASR) experiments if some form of feature smoothing is applied.

We analyse the statistics of the Mel frequency ceptral coefficients (MFCCs) that are used as speech features and describe the effects on recognition results if the mean and variance of these features change. It is shown that recognizers are more sensitive to an increase in variance of enhanced features than to errors in their mean values. We present a method that compensates for the increased variance of DFT-based noise reduction frontends by means of using prior knowledge and smoothing. We achieve high segmental SNR improvements as well as recognition results close to those of the Advanced Frontend (AFE) of the European Telecommunications Standards Institute (ETSI) for all noise types.

Index Terms: robust speech recognition, noise reduction, MFCC, feature statistics.

1. Introduction

The enhancement of noisy speech signals is a well studied topic in both speech communication [1] and robust automatic speech recognition (ASR) [2]. There exist highly specialised algorithms for each of these applications. Although both applications use noise reduction towards the common goal of enhancing speech, their measures of performance are quite different. Only few investigations have been made into algorithms suitable for both applications (see e.g. [3, 4]).

One class of noise reduction filters that are successfully employed in speech enhancement for communication applications consists of optimal estimators for the clean DFT coefficients given the noisy signal [5, 6]. If the length L of the transform is chosen large enough, these filters are able to suppress noise in spectral bands between the harmonics of the fundamental frequency of voiced speech while leaving spectral peaks untouched. This results in a relatively high improvement in the segmental SNR.

In contrast to DFT-based filtering for communication applications noise reduction frontends specialised for robust ASR use a comparably low spectral resolution. The ETSI AFE [7, 2], for instance, uses a Mel band smoothed power spectrum to compute a low resolution time domain Wiener filter. Whenever speech is present within a Mel band, noise within the same band cannot be attenuated without distorting the speech signal. Since speech distortions are to be avoided, more residual noise will remain during speech activity leading to lower segmental SNR improvements.

In this paper we present a frontend for robust ASR that uses high resolution DFT-based spectral estimators such as the log-spectral-amplitude (LSA) estimator [5] or estimators based on a Laplacian speech distribution and a Gaussian noise model (LG, [6]). These estimators are briefly discussed in Section 2. A comparison of these filters with the AFE in terms of the segmental SNR shows the superiority of filters with high spectral resolution for clean speech reconstruction. However, the MFCCs of enhanced speech have a higher variance for the DFT-based noise reduction filters than features computed using the AFE. The effects of this increase on the recognition results are analysed in Section 3. In Section 4 we present a solution that reduces the variance of the features and show that this leads to an increase in robustness of the ASR. Section 5 summarises experimental results and presents conclusions.

2. Noise reduction in the DFT domain

In this section we describe the DFT-based noise reduction. The DFT coefficients $Y(\lambda, k) = S(\lambda, k) + N(\lambda, k)$ of the noisy signal in frame k and frequency bin λ are assumed to be the sum of the clean speech spectrum $S(\lambda, k)$ and the uncorrelated noise $N(\lambda, k)$. For the sake of simplicity we will leave out the frequency bin index λ and the frame index k whenever possible.

The estimate $\widehat{S}$ of the spectral coefficient S of clean speech is a function

$$\widehat{S} = G(Y, P_n, \xi) \tag{1}$$

of the noisy coefficient Y, the noise power $P_n = E_n\left\{|N|^2\right\}$, and the *a priori* SNR $\xi = P_s/P_n$, where $P_s = E_n\left\{|S|^2\right\}$ denotes the speech power. As speech and noise are assumed to be uncorrelated we have $E_n\left\{|Y|^2\right\} = P_s + P_n$. $G(Y, P_n, \xi)$ takes small values for bins with low SNR. Thus, for time signals sampled with a sampling rate f_s speech and noise can be separated with a maximal spectral resolution of f_s/L.

In our evaluation we consider two estimators, the LSA [5] and the LG estimators [6, (27)]. Both filters need knowledge about P_n and ξ. Here we use the *decision-directed* approach [8] to obtain an estimate $\widehat{\xi}$ of the *a priori* SNR. The estimate $\widehat{P}_n$ of the noise power is calculated as the empirical mean by recursive averaging in speech pauses. The empirical mean of $|N|^2$ is an estimator that is independent of the statistical distribution of $|N|^2$ and is therefore suitable for different noise types. As we use the same speech pause detection for all experiments, it has no influence on the relative

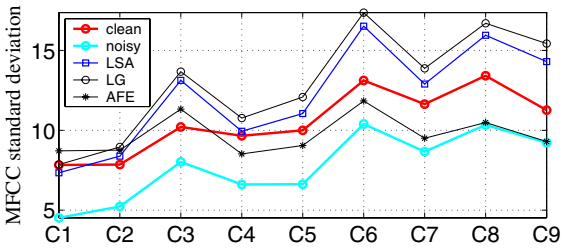

Fig. 1. Standard deviation of the first nine components c_1 to c_9 of feature vectors from state 12 of the word "oh". The noisy features are disturbed by car noise at an SNR of 10dB. The corresponding filter results for the LSA estimator, the LG estimator, and the AFE are shown.

performance and is therefore not described here.

The estimated clean speech spectrum $\widehat{S}$ is transformed back into time domain and a filtered time signal is synthesised using the overlap-add procedure. The calculation of the segmental SNR as well as the MFCC speech features is based on this time domain signal.

3. Analysis of feature variance

Figure 1 shows the standard deviation of the first 9 MFCCs $c_i, i = 1 \ldots 9$, of feature vectors calculated according to [9]. These features belong to state 12 of the HMM for the word "oh". They are extracted using the time information of a clean speech recognition experiment based on 106 utterances of male speakers. The corresponding mean values are given in Figure 2. Apart from the clean MFCCs the features are given for perturbation by car noise at 10dB SNR and for the corresponding filter results for the LSA estimator, the LG estimator, and the AFE.

In the case of the noisy, unfiltered signal the variance of the features is reduced as the dynamic range of the spectral magnitudes is limited due to the noise floor. Their mean values do not correspond to those of clean speech, though. On average, the noise reduction moves the estimated coefficients c_1 to c_8 towards the mean of the clean coefficients (cf. Fig. 2). In case of the DFT-based estimators the variance of the features rises above the level of the clean case (Fig. 1). Note that the AFE does not change the mean and the variance of the MFCCs with higher indices. Only the coarse envelope of the signal spectrum is changed which is represented by the MFCCs with low indices.

We will show in the following that speech recognizers based on HMMs are more sensitive to an increase in variance of the feature vectors than to a shift of their mean values. For this analysis, we consider a single MFCC feature vector component c_i and approximate its distribution by a normal probability density function (pdf). In the training of a HMM the mean μ_c and the variance σ_c^2 of the emission density $p_j(c_i) = \mathcal{N}(c_i; \mu_c, \sigma_c^2)$ in a state j are estimated.

During recognition the estimated clean coefficient $\widehat{c}_i$ is calculated which is also modelled by a normal random variable with mean $\mu_{\widehat{c}}$ and variance $\sigma_{\widehat{c}}^2$. We denote its pdf by $p_{\widehat{c}}(\widehat{c}_i)$. For recognition the likelihood $p_j(\widehat{c}_i)$ is evaluated. The expected normalised likelihood $E_n\{p_j(\widehat{c}_i)\}$ is given by

$$E_n\{p_j(\widehat{c}_i)\} = \frac{1}{E_c}\int_{-\infty}^{\infty} p_j(\widehat{c}_i)\, p_{\widehat{c}}(\widehat{c}_i)\, d\widehat{c}_i = \frac{\sqrt{2\sigma_c^2}}{\sqrt{\sigma_c^2+\sigma_{\widehat{c}}^2}}$$

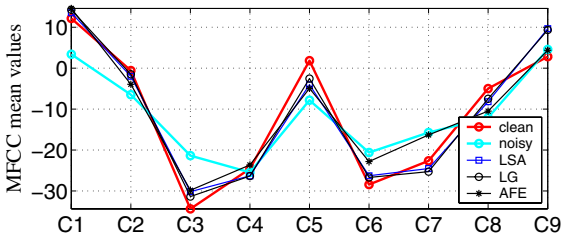

Fig. 2. Mean values of the first nine components c_1 to c_9 of feature vectors from the word "oh". The graphs shown correspond to those in Figure 1.

$$\times \exp\left[-\frac{1}{2}\left(\frac{\mu_c^2}{\sigma_c^2}+\frac{\mu_{\widehat{c}}^2}{\sigma_{\widehat{c}}^2}\right)+\frac{1}{2}\left(\frac{\mu_c}{\sigma_c^2}+\frac{\mu_{\widehat{c}}}{\sigma_{\widehat{c}}^2}\right)^2\frac{\sigma_c^2\sigma_{\widehat{c}}^2}{\sigma_c^2+\sigma_{\widehat{c}}^2}\right] \quad (2)$$

where $E_c = 1/\sqrt{4\pi\sigma_c^2}$ is the expected likelihood for clean speech, i.e. $\mu_{\widehat{c}} = \mu_c, \sigma_{\widehat{c}}^2 = \sigma_c^2$. $E_n\{p_j(\widehat{c}_i)\}$ represents the average recognition result for state j. Figure 4 shows $E_n\{p_j(\widehat{c}_i)\}$ for the case of an estimate $\widehat{c}_i$ with correct mean ($\mu_{\widehat{c}} = \mu_c$) and different variances $\sigma_{\widehat{c}}^2$. Figure 3 shows $E_n\{p_j(\widehat{c}_i)\}$ for different mean values $\mu_{\widehat{c}}$ under the condition that the variance of the estimated features equals that of clean features ($\sigma_{\widehat{c}}^2 = \sigma_c^2$). For clean speech we have $\sigma_{\widehat{c}}^2/\sigma_c^2 = 1$ and $(\mu_{\widehat{c}} - \mu_c) = 0$. Note that the slope at these points is much steeper for variance deviations than for deviations in the mean. The recognizer is more sensitive to an increase in variance than to a shift in mean values.

According to Figure 3 the shifted mean values of the noisy features in Figure 2 lowers the expected likelihood $E_n\{p_j(\widehat{c}_i)\}$ although their variance is lower than that of the clean features. The noise reduction delivers feature vectors with mean values closer to those of clean speech, but in the case of DFT-based estimators their variance rises above σ_c^2 which lowers $E_n\{p_j(\widehat{c}_i)\}$ according to Fig. 4. The advantage of the smoothed filter of the AFE is a low variance of the enhanced features (see Fig. 1).

Note that from Figs. 1 and 2 the robustness of model-based noise compensation methods can also be explained. A simple version of noise compensation shifts the mean vectors of the model densities by the mean vector of noise only features. On one side noisy features have a lower variance than the models of the recognizer suggest for clean features, that is $\sigma_{\widehat{c}}^2 < \sigma_c^2$, and on the other side the mean values of the model only need to be corrected to such an extend that $\mu_{\widehat{c}} - \mu_c < \sigma_c/2$.

4. Reduction of feature variance

As Mel coefficients are weighted sums of several DFT magnitudes, the Mel spectrum is a smoothed version of the DFT spectrum. In contrast to the Mel coefficients the magnitude squared coefficient $|Y|^2$ therefore has a comparably high variance. In consequence, the estimated coefficients $\widehat{S}$ also have a higher variance.

Additionally, the LSA and the LG estimator use the data driven *decision-directed* approach [8] to estimate the parameter $\widehat{\xi}$. It is given by

$$\widehat{\xi}(\lambda,k) = \alpha\frac{|\widehat{S}(\lambda,k-1)|^2}{\widehat{P}_n(\lambda,k-1)} + (1-\alpha)\max\left[\frac{|Y(\lambda,k)|^2}{\widehat{P}_n(\lambda,k)} - 1,\ 0\right]. \quad (3)$$

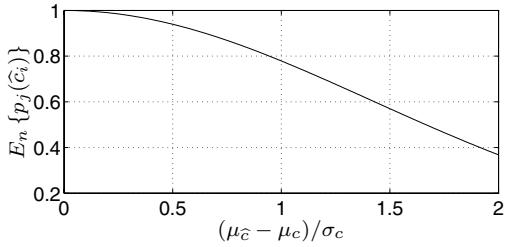

Fig. 3. Relation between expected likelihood $E_n\{p_j(\widehat{c}_i)\}$ and feature mean value $\mu_{\widehat{c}}$ for a feature variance $\sigma^2_{\widehat{c}} = \sigma^2_c$.

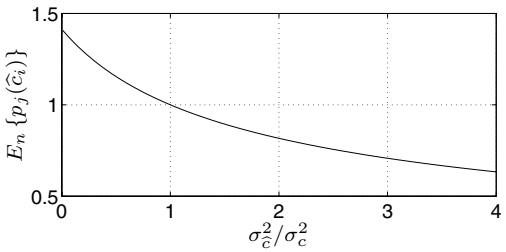

Fig. 4. Relation between expected likelihood $E_n\{p_j(\widehat{c}_i)\}$ and feature variance $\sigma^2_{\widehat{c}}$ for correct mean ($\mu_{\widehat{c}} = \mu_c$).

It is a function of $|Y(\lambda,k)|^2$ itself and therefore the variance of this estimate is especially high for frequency bins λ with low SNR.

The increase in variance can be reduced, if the data-driven *decision-directed* approach in equation (3) is replaced by a model-driven approach for frequency bands with low SNR. The following procedure is based on the rationale that we can compare a first estimate of the speech energy $\widehat{P}_s(\lambda,k)$ with prior knowledge of the spectral shape of $P_s(\lambda,k)$ created from clean speech. For the comparison we rely on those bins λ where the first estimate is relatively accurate. These are frequency bins with high SNR $\widehat{\xi}(\lambda,k)$. The result of the lookup is then used to interpolate those bins where the SNR is low and the estimate is strongly influenced by the noise.

For the description of the prior knowledge we use a Gaussian mixture model (GMM) with $N = 100$ mixture components that are trained using the clean training data of [9]. The GMM models the pdf of feature vectors composed of the $M = 23$ logarithmic Mel filter energies and the energy of the clean speech frame k. We used diagonal covariance matrices. A GMM for vectors of the spectral values of the DFT was not considered due to complexity reasons.

For each signal frame k we calculate an estimate of the clean speech power $\widehat{P}_s(\lambda) = \widehat{\xi}(\lambda)\widehat{P}_n(\lambda)$ and the power in the m-th Mel band

$$\widehat{P}_s^{mel}(m) = \sum_{\lambda=1}^{L/2+1} W_m(\lambda)\widehat{P}_s(\lambda). \tag{4}$$

The spectral weights $W_m(\lambda)$ for the calculation of the Mel filter energies are normalised to give $\sum_{\lambda=1}^{L/2+1} W_m(\lambda) = 1$. The feature vector $\widehat{\boldsymbol{P}}_s^{log}$ of the log-Mel filter energies for the lookup is calculated as $\widehat{\boldsymbol{P}}_s^{log} = (\widehat{P}_s^{log}(1)\ldots\widehat{P}_s^{log}(M), \mathcal{E}_s^l)$, with $\widehat{P}_s^{log}(m) = \log(\widehat{P}_s^{mel}(m))$ and the frame energy $\mathcal{E}_s^l = \log(\sum_{\lambda=1}^{L}|S(\lambda)|^2)$. From the GMM we calculate a new estimate

$$\widetilde{\boldsymbol{P}}_s^{mel} = \sum_{n=1}^{N} p(n|\widehat{\boldsymbol{P}}_s^{log}) \cdot \exp(\boldsymbol{\mu}_s^{mel}(n)) \tag{5}$$

$$p(n|\widehat{\boldsymbol{P}}_s^{log}) = \frac{\beta_n\, p(\widehat{\boldsymbol{P}}_s^{log}|n)}{\sum_{n=1}^{N}\beta_n\, p(\widehat{\boldsymbol{P}}_s^{log}|n)} \tag{6}$$

where β_n is the mixture weight and $\boldsymbol{\mu}_s^{mel}(n)$ is the mean vector of mixture component n. $\widetilde{\boldsymbol{P}}_s^{mel} = (\widetilde{P}_s^{mel}(1)\ldots\widetilde{P}_s^{mel}(M))$ represents the model-based Mel energy estimates. Note that $\exp(\boldsymbol{\mu}_s^{mel})$ is calculated component wise and gives a representation of $\boldsymbol{\mu}_s^{mel}$ in the linear Mel domain. In this operation we do not use $\mathcal{E}_s^l$.

We then determine low SNR bands in the Mel domain by computing

$$\widehat{\xi}^{mel}(m) = \sum_{\lambda=1}^{L/2+1} W_m(\lambda)\widehat{\xi}(\lambda). \tag{7}$$

and compare these to a threshold $\widehat{\xi}_{th}^{mel}$.

We interpolate the speech power for Mel bins with low SNR by calculating the difference between the information from the prior model and the actual value

$$\Delta\widehat{\widehat{P}}_s^{mel}(m) = \begin{cases} 0 & \widehat{\xi}^{mel}(m) \geq \widehat{\xi}_{th}^{mel} \\ \max(\widetilde{P}_s^{mel}(m) - \widehat{P}_s^{mel}(m), 0) & \text{else.} \end{cases} \tag{8}$$

As a clipping of speech has a strong negative effect on the recognition results, we use the max-function. From experiments the threshold for the local SNR was chosen as $10\log_{10}(\widehat{\xi}_{th}^{mel}) = 8$dB.

The DFT representation of $\Delta\widehat{\widehat{P}}_s^{mel}$ is then calculated as

$$\Delta\widehat{\widehat{P}}_s(\lambda) = \sum_{m=1}^{M} W_m(\lambda)\Delta\widehat{\widehat{P}}_s^{mel}(m), \tag{9}$$

Finally we obtain a reestimate of the *a priori* SNR in the DFT domain

$$\widehat{\widehat{\xi}}(\lambda) = \frac{\widehat{P}_s(\lambda) + \Delta\widehat{\widehat{P}}_s(\lambda)}{\widehat{P}_n(\lambda)}, \tag{10}$$

which is used in (1) instead of $\widehat{\xi}$. The use of $\Delta\widehat{\widehat{P}}_s(\lambda)$ has the advantage that $\widehat{\widehat{\xi}}$ takes the values of $\widehat{\xi}$ for frequency bands with high SNR. Thus the frequency resolution is not reduced in those bands.

The increase in robustness by using the prior knowledge contained in the GMM in case of frequency bins with low SNR can be seen in a comparison of the variances of the MFCCs in Fig. 5.

5. Experimental results and conclusions

In order to demonstrate the effectiveness of the DFT-based filters in terms of clean speech reconstruction we evaluated the improvement of the segmental SNR for car noise at different noise levels. Note that for the segmental SNR, noise is defined as the difference between the clean speech signal and the filtered signal in speech-active signal frames. Here, the segmental SNR for the analysis of the results considers frames with frame energy not less than -45db of the maximum frame energy of the utterance in the clean case. The segmental SNR therefore is a combined measure of speech distortion and noise suppression during speech presence. The results for the segmental SNR improvement shown in Table 1 are an average over 1001 sentences disturbed by car noise taken from the

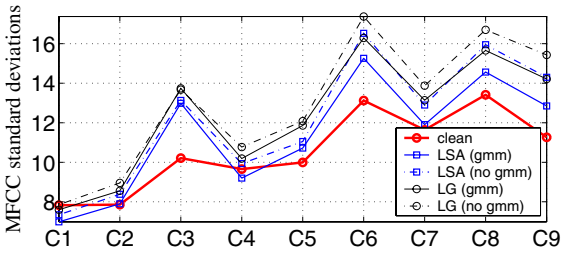

Fig. 5. Reduced standard deviation of the feature vectors by the use of the presented approach (case "gmm"). The results for the case "no gmm" are the same as in Figure 1. The feature smoothing of [10] has not been applied.

	segm. SNR impr.			word accuracy		
Filter	10dB	15dB	20dB	10dB	15dB	20dB
none	0.0	0.0	0.0	92.63%	96.51%	98.39%
AFE	5.3	2.5	0.0	95.97%	97.85%	98.78%
LSA (GMM)	6.9	5.7	4.6	95.29%	97.67%	98.57%
LG (GMM)	6.9	5.8	4.7	95.65%	97.91%	98.57%

Table 1. Improvement of the segmental SNR in dB and corresponding recognition results for car noise. The SNR of the noisy input signals was 10dB, 15dB, and 20dB as defined in [9].

test set [9]. The estimator 3 of the SNR for the DFT-based filters used a value $\alpha = 0.92$, which gave best recognition results.

The recognition results that are achieved with the AFE and our DFT-based noise reduction are also shown. The recognition scores are obtained within the AURORA2 framework [9].

The HMMs of the recognizer consisted of 16 states per model and each state had three Gaussian mixture components with diagonal correlation matrices. The features for the DFT-based frontend consisted of the thirteen MFCCs c_0 to c_{12} and their delta and delta-delta values, resulting in vectors with 39 components. The feature smoothing of [10] has been used in all recognition experiments with the DFT-based filters. The postprocessing of [10] was used in such a way that after cepstral mean and variance normalisation a smoothing of the features was performed that generates a smoothed frame k out of frames $k-2$ to $k+2$. The training was done with clean data. The AFE was not modified and the training for this frontend was done separately. The results for the three test sets given in Table 2 are the word accuracies averaged for noise levels from 0dB to 20dB as defined in [9].

From results in Tables 1 and 2 it can be seen that DFT-based noise reduction frontends perform well in terms of segmental SNR improvement and word accuracies. However, despite significantly lower segmental SNR value, Table 2 shows that the AFE is still more robust than the DFT-based methods, when the full set of experiments with all noise types is considered. We attribute this to the lower variance of the features calculated with the AFE. Furthermore, informal listening tests reveal, that the auditive quality of the AFE processed speech is lower than the quality of the DFT processed speech.

For DFT-based noise reduction filters we conclude that the gain in the segmental SNR does not result in equally improved recognition scores, since the variance of the MFCC features is also increased. Therefore improved noise reduction filters have to balance both aspects.

	test set		
Filter	A	B	C
none	84.41%	85.55%	84.29%
AFE	87.74%	87.09%	85.45%
LSA (no GMM)	85.64%	85.56%	85.97%
LSA (GMM)	85.99%	85.74%	85.73%
LG (no GMM)	85.66%	85.40%	85.54%
LG (GMM)	85.93%	85.57%	85.56%

Table 2. Word accuracies in percent for the AURORA2 recognition task and training on clean data. The results for the solution described in section 4 are marked as "GMM". The case "no GMM" describes the results for the conventional *decision-directed* approach (3).

This work is funded by the German Research Foundation DFG.

6. References

[1] R. Martin, D. Malah, R. Cox, and A. Accardi, "A noise reduction preprocessor for mobile voice communication," *EURASIP Journal of Applied Signal Processing*, pp. 1048–1058, 2004.

[2] D. Macho, L. Mauuary, B. Noé, Y. M. Cheng, D. Ealey, D. Jouvet, H. Kelleher, D. Pearce, and F. Saadoun, "Evaluation of a noise-robust DSR front-end on AURORA databases," *International Conference on Spoken Language Processing (ICSLP)*, pp. 17–20, Sept. 2002.

[3] R. Gemello, F. Mana, and R. de Mori, "Automatic speech recognition with a modified Ephraim-Malah rule," *IEEE Signal Processing Letters*, vol. 13, no. 1, pp. 56–59, Jan. 2006.

[4] P. Setiawan, S. Suhadi, T. Fingscheidt, and S. Stan, "Robust speech recognition for mobile devices in car noise," in *Interspeech – Conference on Speech Communication and Technology*, Sept. 2005, pp. 2673–2676.

[5] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 33, no. 2, pp. 443–445, Apr. 1985.

[6] R. Martin, "Speech enhancement based on minimum mean-square error estimation and supergaussian priors," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 845–856, Sept. 2005.

[7] ETSI standard document, "Speech processing, transmission and quality aspects (STQ); distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithms," *ETSI ES 202 050 V1.1.3*, Nov. 2003.

[8] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

[9] H.-G. Hirsch and D. Pearce, "The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy conditions," *ISCA Workshop on Automatic Speech Recognition*, Sept. 2000.

[10] C.-P. Chen, J. A. Bilmes, and K. Kirchhoff, "Low-resource noise-robust feature post-processing on Aurora 2.0," *International Conference on Spoken Language Processing (ICSLP)*, pp. 2445–2448, Sept. 2002.