Feature Denoising Using Joint Sparse Representation for In-Car Speech Recognition

Weifeng Li, Yicong Zhou, Norman Poh, Fei Zhou, and Qingmin Liao

***Abstract*—We address reducing the mismatch between training and testing conditions for hands-free in-car speech recognition. It is well known that the distortions caused by background noise, channel effects, etc., are highly nonlinear in the log-spectral or cepstral domain. This letter introduces a joint sparse representation (JSR) to estimate the underlying clean feature vector from a noisy feature vector. Performing a joint dictionary learning by sharing the same representation coefficients, the proposed method intends to capture the complex relationships (or mapping functions) between clean and noisy speech. Speech recognition experiments on realistic in-car data demonstrate that the proposed method shows excellent recognition performance with a relative improvement of 39.4% compared with the "baseline" frontends.**

***Index Terms*—Dictionary training, in-car speech recognition, log mel-filter bank (MFB) outputs, sparse representation.**

I. Introduction

THE mismatch between training and testing conditions is one of the most challenging and important problems in automatic speech recognition. This mismatch may be caused by a number of factors, such as background noise, speaker variation, a change in speaking styles, channel effects, and so on. State-of-the-art techniques for removing the mismatch usually fall into the following three categories [1]: robust features, speech enhancement, and model compensation. The first approach seeks parameterizations that are fundamentally immune to noise. The most widely used speech recognition features are the Mel-frequency cepstral coefficients (MFCCs), although they are susceptible to noise. Among the speech enhancement methods, spectral subtraction and short-time spectral attenuation based methods are commonly used, in which cases the underlying noise need to be estimated. Model compensation (e.g., maximum-likelihood linear regression [2], and Jacobian adaptation [3]) aims to adapt or transform acoustic models to match the noisy speech feature in a new testing environment. Most speech enhancement and model compensation methods are accomplished by linear functions such as simple bias removal, affine transformation, linear regression, and so on. However, it is well known that the distortion caused even by additive noise only is highly nonlinear in the log-spectral or cepstral domain.

Recently, sparse representation (SR) [4] has received growing interest in signal processing and pattern recognition. In SR a signal is approximated by a linear combination of a few atoms from a pre-defined dictionary. This line of research focuses on dictionary training, i.e., using machine learning techniques to learn an over-complete dictionary of primary signals (atoms) directly from data, so that the most relevant properties of the signals can be efficiently captured. The SR techniques have been used for compressive sensing [5], face recognition [6], and phone recognition [7]. In audio signal processing domain, SR has been used for source separation [8], speech enhancement [9], and robust speech recognition [10].

In speech enhancement and robust speech recognition domain, we have two coupled signal spaces (e.g., clean versus noisy speech spaces, or training versus testing conditions which may differ), and these two coupled spaces are usually related by some mapping functions, which could be nonlinear as mentioned above. Most existing SR methods only consider sparse modeling in an isolated signal space (such as a clean speech space only or a testing condition only), and fail to consider dictionary learning across different signal spaces. In such cases, it is often desirable to learn representations that can not only well represent each signal space individually, but also capture their relationships through the underlying SRs.

In this letter, we propose a joint SR (JSR) technique, in which a joint dictionary learning is performed across the clean and noisy feature spaces, for feature denoising (or estimating the clean features) with an application to robust speech recognition. In the proposed joint dictionary learning, sharing the same representation coefficients intends to capture possible complex relationships between the clean and noisy feature spaces. For any given testing noisy feature vector, we first find their sparse representation coefficients, and then estimate the underlying clean feature vector, which is used as input by the speech recognition system. The proposed method differs from [9] and [10] not only in dictionary training but also in clean speech/feature estimation. Furthermore, without any linear assumption of the noisy

Manuscript received October 17, 2012; revised January 19, 2013; accepted February 03, 2013. Date of publication February 08, 2013; date of current version May 23, 2013. This work was supported in part by the Macau Science and Technology Development Fund under Grant 017/2012/A1 and by the Research Committee at University of Macau under Grants MYRG113(Y1-L3)-FST12-ZYC and MRG001/ZYC/2013/FST. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Murat Saraclar.

W. Li, F. Zhou, and Q. Liao are with Shenzhen Key Laboratory of Information Science and Technology, Shenzhen Engineering Laboratory of IS&DRM, and the Department of Electronic Engineering/Graduate School at Shenzhen, Tsinghua University, Tsinghua, China (e-mail: Li.Weifeng@sz.tsinghua.edu.cn; flying.zhou@163.com; liaoqm@sz.tsinghua.edu.cn).

Y. Zhou is with Department of Computer and Information Science, University of Macau, Macau, China (e-mail: yicongzhou@umac.mo).

N. Poh is with Department of Computing, University of Surrey, Surrey, U.K. (e-mail: normanpoh@ieee.org).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LSP.2013.2245894

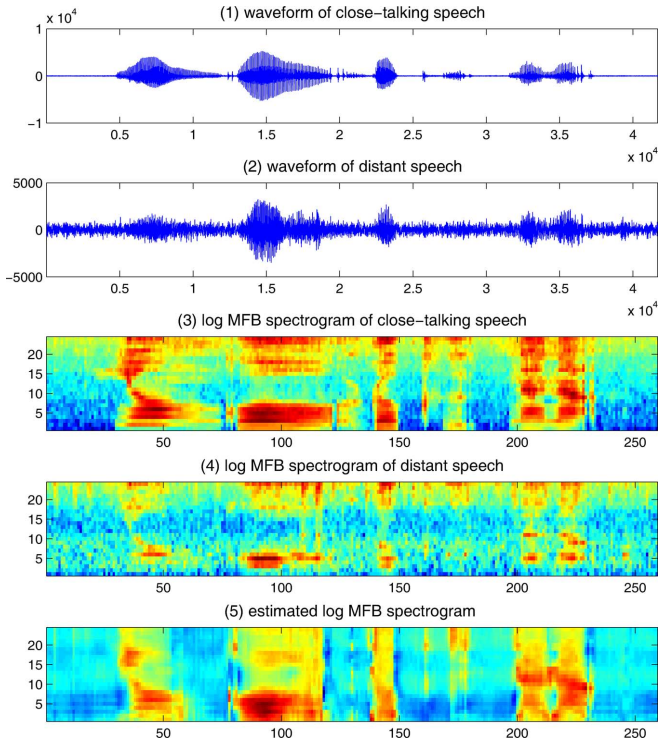

Fig. 1. Effect of car noise on log mel-filter bank (MFB) spectrogram. The speech is "4567" in Japanese from CENSREC-3 in-car database [11].

speech from clean speech required in [9] and [10], our method intends to capture the nonlinear relationship between the clean and noisy speech using a joint learning approach.

II. Proposed Approach

We show in Fig. 1(3) and Fig. 1(4) the log mel-filter bank (MFB) spectrograms of speech recorded by the close-talking microphone and distant microphone respectively (The data are from CENSREC-3 in-car database [11].). It can be observed that the background noise has contaminated the speech quality, and correspondingly degraded the performance of speech recognition. In this letter, we attempt to solve this problem by seeking a common sparse representation and such representation should be robust to the background noise. The feasibility is inspired from the property of linear object class (LOC) [12], which shows that the feature locations (or pixels) in two viewpoints of the same object could be represented by a linear combination of the bases of the two viewpoints with the same representation coefficients.

Let $\mathbf{x}^c$ and $\mathbf{x}^n$ be a log MFB output of the same speech recorded by the close-talking microphone and distant microphone, respectively. Similarly we have

$$\begin{bmatrix} \mathbf{x}^c \\ \mathbf{x}^n \end{bmatrix} = \sum_j \alpha_j \begin{bmatrix} \mathbf{d}_j^c \\ \mathbf{d}_j^n \end{bmatrix}, \tag{1}$$

where $\{\mathbf{d}_j^c\}$ and $\{\mathbf{d}_j^n\}$ are representation basis sets of clean and noisy speech, respectively. The reconstruction coefficients $\{\alpha_j\}$ are the common representations that we seek for capturing the relationships between the clean and noise speech spaces.

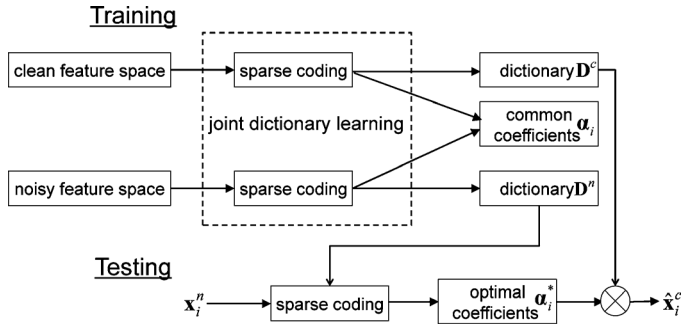

Fig. 2. Block diagram of the proposed approach.

A single sparse representation could approximate $\mathbf{x}^c$ or $\mathbf{x}^n$ by a linear combination of a few atoms from a dictionary $\mathbf{D}^c = [\mathbf{d}_1^c, \mathbf{d}_2^c, \ldots, \mathbf{d}_J^c]$ or $\mathbf{D}^n = [\mathbf{d}_1^n, \mathbf{d}_2^n, \ldots, \mathbf{d}_J^n]$ respectively, which is learned via

$$\min_{\mathbf{D}^c, \boldsymbol{\alpha}_i} \sum_{i=1}^{N} \|\mathbf{x}_i^c - \mathbf{D}^c \boldsymbol{\alpha}_i^c\|^2 + \gamma^c \|\boldsymbol{\alpha}_i^c\|_1 \tag{2}$$

or

$$\min_{\mathbf{D}^n, \boldsymbol{\alpha}_i} \sum_{i=1}^{N} \|\mathbf{x}_i^n - \mathbf{D}^n \boldsymbol{\alpha}_i^n\|^2 + \gamma^n \|\boldsymbol{\alpha}_i^n\|_1 \tag{3}$$

where i is the frame index and N is the total number of training examples. γ^c (or γ^n) is a penalty weight on sparsity. $\|\cdot\|^2$ and $\|\cdot\|_1$ denote the ℓ_2-norm and ℓ_1-norm, respectively. In order to capture the relationships between the clean and noise speech spaces and form a common representation across the two spaces, we need to learn the dictionaries $\mathbf{D}^c$ and $\mathbf{D}^n$ jointly, and then estimate the underlying $\mathbf{x}^c$ from $\mathbf{x}^n$. The flow chart is illustrated in Fig. 2.

A. Joint Dictionary Training

Given the coupled training feature sequences $\{\mathbf{x}_i^c\}_{i=1}^N$ and $\{\mathbf{x}_i^n\}_{i=1}^N$, the problem of jointly learning the dictionaries can be formulated as follows:

$$\min_{\mathbf{D}^c, \mathbf{D}^n, \boldsymbol{\alpha}_i} \sum_{i=1}^{N} (\|\mathbf{x}_i^c - \mathbf{D}^c \boldsymbol{\alpha}_i\|^2 + \|\mathbf{x}_i^n - \mathbf{D}^n \boldsymbol{\alpha}_i\|^2) + \gamma \|\boldsymbol{\alpha}_i\|_1. \tag{4}$$

Let

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}^c \\ \mathbf{x}^n \end{bmatrix}, \quad \text{and} \quad \mathbf{D} = \begin{bmatrix} \mathbf{D}^c \\ \mathbf{D}^n \end{bmatrix}. \tag{5}$$

Then (4) reduces to a standard sparse code problem:

$$\min_{\mathbf{D}, \boldsymbol{\alpha}_i} \sum_{i=1}^{N} \|\mathbf{x}_i - \mathbf{D}\boldsymbol{\alpha}_i\|^2 + \gamma \|\boldsymbol{\alpha}_i\|_1. \tag{6}$$

Equation (6) is not convex in both $\mathbf{D}$ and $\boldsymbol{\alpha}_i$, however it is convex in one of them with the other fixed[1].

[1]In our experiments, we used a Matlab package developed in [13], in which a feature-sign search algorithm is used for optimizing $\boldsymbol{\alpha}_i$ while $\mathbf{D}$ fixed, and given fixed $\boldsymbol{\alpha}_i$ $\mathbf{D}$ is learned using the Lagrange dual.

B. Clean Feature Estimation

Once we obtained the coupled dictionaries $\mathbf{D}^c$ and $\mathbf{D}^n$, for any given testing noisy feature vector $\mathbf{x}_i^n$, we first find its sparse representation in terms of $\mathbf{D}^n$

$$\boldsymbol{\alpha}_i^* = \arg\min_{\boldsymbol{\alpha}_i} \|\mathbf{x}_i^n - \mathbf{D}^n\boldsymbol{\alpha}_i\|^2 + \gamma\|\boldsymbol{\alpha}_i\|_1, \tag{7}$$

and then estimate its corresponding clean feature vector $\mathbf{x}_i^c$ in terms of $\mathbf{D}^c$ via

$$\hat{\mathbf{x}}_i^c = \mathbf{D}^c\boldsymbol{\alpha}_i^*. \tag{8}$$

Fig. 1(5) shows the estimated log MFB spectrogram of clean speech from Fig. 1(4). It can be found that by comparing Fig. 1(5) with Fig. 1(4) the interfering noise noise is reduced whereas the speech signal is enhanced. The result is that the mismatches of log MFB spectrogram between clean and noisy speech are reduced, which will be helpful for the speech recognition system.

III. Experimental Results

A. Database

The proposed approach was evaluated on a real in-car speech recognition task. Noisy speech data were taken from the CENSREC-3 (Corpus and Environments for Noisy Speech RECognition) database [11]. The "condition 3" of the CENSREC-3 was used for training and test in this letter. The training data were composed of 14,050 phonetically-balanced utterances captured by the distant microphone under two conditions: idling and driving on a city street with a normal in-car environment, and the total number of speakers for training data was 293 (202 males and 91 females). The test data were recorded by the distant microphone under 16 environmental conditions, with a total of 14,216 utterances spoken by 18 speakers (8 males and 10 females) [11].

The speech signal was sampled at 16 kHz and windowed with a 20-ms Hamming window every 10 ms. In the mel-filter bank (MFB) analysis, a cut-off was applied to frequency components lower than 250 Hz, and the total number of dimensions of the filter-bank output was 24. The acoustic models consist of triphone HMMs that have five states with three distributions. Each distribution was represented with 32-mixture Gaussians. The baseline system was trained using 39-dimensional feature vectors consisting of 12-dimensional MFCC parameters and log-energy, along with their delta and delta-delta parameters.

B. Experimental Settings

As shown in Fig. 2, we need training and testing data set. In our experiments, we randomly selected 330 utterances from CENSREC-3 training data for training the coupled dictionaries. Nine speakers (five females and four males) in the test data were used for evaluation. The baseline recognition accuracy was 78.38%. We performed JSR on log MFB features[2]. The estimated clean log MFB outputs were converted into MFCCs through Discrete Cosine Transformation (DCT), and then their delta and delta-delta parameters were calculated. The two coupled feature vectors are frame-by-frame aligned. Following [14], we also extracted noisy feature vector $\mathbf{x}_i^n$ by concatenating 11 successive frames (five before and five after) for estimating $\mathbf{x}_i^c \in \mathbb{R}^{25\times 1}$. Therefore we have the following two configurations for the JSR method.

1) JSRbs: using a single frame of log MFB outputs as $\mathbf{x}_i^n \in \mathbb{R}^{25\times 1}$;
2) JSRbm: using multi-frame (i.e., 11-frame) log MFB outputs, and concatenating them into a noisy feature vector $\mathbf{x}_i^n \in \mathbb{R}^{275\times 1}$.

For JSR based methods we empirically tuned the dictionary size and the penalty weight γ by optimizing one of them with the other fixed.

For comparisons, Generalized Spectral Subtraction (GSS) [15] and Log-Spectra Amplitude (LSA) estimator [16], and Mean and Variance Normalization (MVN) were applied. Stereo-based Piecewise LInear Compensation for Environments (SPLICE) [17], in which the compensation factor is obtained using the same stereo clean and noisy speech as that in our proposed JSR method, was performed. Exemplar-based Feature Enhancement (EFE) [10] was also compared[3]. Because the training data, recorded by the distant microphone, are noisy, for all methods we re-trained the acoustic models on their processed training data.

C. Results

Fig. 3 shows the recognition results obtained form the different methods. It can be observed that The recognition performance of "baseline" depend on the evaluation environments. When the recording environments between training and test data are not matched, the recognition correct rate can degrade into less than 60% for the in-car state of window and air-conditioner on high level (from Fig. 4). Carrying out the Mean and Variance Normalization (MVN) in cepstral domain is helpful for improving the in-car speech recognition performance, and performs comparably to the speech enhancement method the GSS but not as well as the LSA and SPLICE. The SPLICE and EFE perform better than the "baseline", and the SPLICE provides more gains than the EFE. When the joint sparse representation (JSR) is employed, the recognition performance is better than the EFE, which demonstrates the effectiveness of the proposed JSR approach. Moreover, using multiple frames is advantageous compared with using one frame only. The highest correct rate 86.9% is achieved by using "JSRbm" with a relative improvement of 39.4% compared with "baseline".

Fig. 4 compares the recognition performance of the proposed method (i.e., "JSRbm") with the LSA for the six in-car states. It is found that compared with the "baseline", the LSA can improve the recognition performance for stationary noisy conditions (e.g., air-conditioner on and window open), but it is not effective for non-stationary noisy conditions (e.g., audio CD player and hazard flasher on). Our proposed method not only

[2]Log-energy parameter is also included in our experiments.

[3]In our experiment 4000 speech exemplars and 4000 noise exemplars were randomly selected from CENSREC-3 training data, and $B = 25$ and $T = 11$ were adopted.

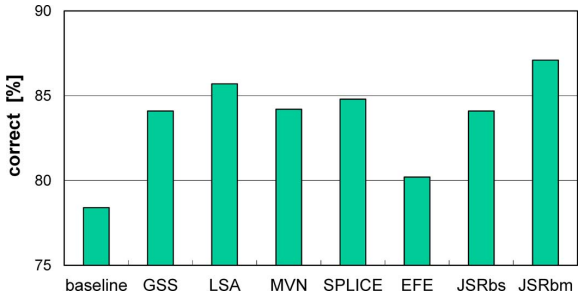

Fig. 3. Recognition performance of different methods.

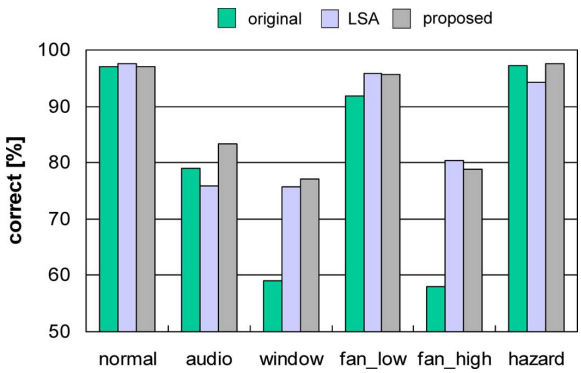

Fig. 4. Recognition performance for six in-car states. The first five in-car states are averaged over three driving speed conditions (i.e., idling, low-speed, high-speech), and the hazard flasher is on only for idling condition [11]).

TABLE I
AN EXAMPLE OF ENVIRONMENTAL SENSITIVITIES OF THE DICTIONARIES ON THE RECOGNITION RATE FOR THE CONDITION OF DRIVING IN LOW-SPEED WITH WINDOWS OPEN ("LW") BY USING "JSRBS". "LN" DENOTES THE CONDITION OF DRIVING IN LOW-SPEECH WITH NORMAL IN-CAR STATE. "IW" DENOTES THE CONDITION OF IDLING WITH WINDOWS OPEN. "HS" DENOTES THE CONDITION OF DRIVING IN HIGH-SPEECH WITH CD PLAYER ON

dictionary training condition	LW	LN	IW	HS
correct [%]	81.29	82.85	80.25	79.51

performs comparably with the LSA for the stationary noisy conditions, but also shows its robustness to non-stationary noisy conditions. Table I shows an example of environmental sensitivities of the dictionaries on the recognition rate when noise conditions have mismatch between dictionary training and testing. The testing noise condition is "LW" (driving in low-speed with windows open). It can be observed that when the dictionary training condition[4] is "LW' (matched), the recognition performance is better than those using the "IW" and "HS" data for training the dictionaries, but worse than that using the "LN" data, which may be explained by the fact that the "LN" data are included in the training data of acoustic model. In general the recognition performance is not very sensitive to the mismatch between dictionary training and testing.

IV. CONCLUSIONS

In this letter, we have proposed a joint sparse representation (JSR) technique for feature denoising in application to robust in-car speech recognition. In our proposed JSR, a joint dictionary learning is performed across the clean and noisy feature spaces in order to capture possible complex relationships between the two feature spaces. Given a noisy feature vector (e.g., log mel-filter-bank (MFB) outputs or Mel-frequency cepstral coefficients (MFCCs)), the feature vectors of clean speech are estimated using the dictionary of clean speech. Compared with the spectral subtraction and the log-spectral amplitude estimator, the proposed method shows its superiority in terms of a significant improvement in recognition performance in the speech recognition experiments conducted in 16 realistic driving conditions.

In our experiments, the data for training the dictionaries are from the training data of acoustic model. In order to develop a data-driven in-car recognition system, we need to develop an effective algorithm for automatic adapting the sparse representations to different driving conditions. Moreover, when the system encounters a new type of noise, a soft or fuzzy logic decision is desirable and will be our future work.

[4]Other nine speakers from the test data in CENSREC-3 were used for training the dictionary.

REFERENCES

[1] Y. Gong, "Speech recognition in noisy environments: A survey," *Speech Commun.*, vol. 16, no. 3, pp. 261–291, 1995.
[2] C. Leggetter and P. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models," *Comput. Speech Lang.*, vol. 9, no. 2, pp. 171–186, 2012.
[3] S. Sagayama, Y. Yamaguchi, and S. Takahashi, "Jacobian adaptation of noisy speech models," in *IEEE Wkshp on Autom. Speech Recognition and Understanding*, 1997, pp. 396–403.
[4] J. Wright, Y. Ma, J. Mairal, G. Sapiro, T. Huang, and S. Yan, "Sparse representation for computer vision and pattern recognition," *Proc. IEEE*, vol. 98, no. 6, pp. 1031–1044, 2010.
[5] E. Candes and M. Wakin, "An introduction to compressive sampling," *IEEE Signal Process. Mag.*, vol. 25, no. 2, pp. 21–30, 2008.
[6] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Patt. Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210–227, 2009.
[7] T. Sainath, A. Carmi, D. Kanevsky, and B. Ramabhadran, "Bayesian compressive sensing for phonetic classification," in *IEEE Int. Conf. on Acoust., Speech, and Signal Process.*, 2010, pp. 4370–4373.
[8] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 3, pp. 1066–1074, 2007.
[9] C. Sigg, T. Dikk, and J. Buhmann, "Speech enhancement using generative dictionary learning," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 6, pp. 1698–1712, 2012.
[10] J. Gemmeke, T. Virtanen, and A. Hurmalainen, "Exemplar-based sparse representations for noise robust automatic speech recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2067–2080, 2011.
[11] M. Fujimoto, K. Takeda, and S. Nakamura, "Censrec-3: An evaluation framework for Japanese speech recognition in real driving-car environments," *IEICE Trans. Inf. Syst.*, vol. E89-D, no. 11, pp. 2783–2793, 2006.
[12] T. Vetter and T. Poggio, "Linear object classes and image synthesis from a single example image," *IEEE Trans. Patt. Anal. Mach. Intell.*, vol. 19, no. 7, pp. 733–742, 1997.
[13] H. Lee, A. Battle, R. Raina, and A. Ng, "Efficient sparse coding algorithms," in *NIPS*, 2007, pp. 801–808.
[14] B. Zhang, S. Matsoukas, and R. Schwartz, "Long span features and minimum phoneme error heteroscedastic linear discriminant analysis," in *Proc. of EARS RT-04 Workshop*, 2004.
[15] B. Sim, Y. Tong, J. Chang, and C. Tan, "A parametric formulation of the generalized spectral subtraction method," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 4, pp. 328–337, 1998.
[16] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-32, no. 2, pp. 443–445, 1985.
[17] L. Deng, A. Acero, M. Plumpe, and X. Huang, "Large-vocabulary speech recognition under adverse acoustic environments," in *Proc. ICSLP'00*, 2000, pp. 806–809.