Int. J. Electron. Commun. (AEÜ) 65 (2011) 99–106

Contents lists available at ScienceDirect

Int. J. Electron. Commun. (AEÜ)

journal homepage: www.elsevier.de/aeue

A model distance maximizing framework for speech recognizer-based speech enhancement

Bagher BabaAli [a,*], Hossein Sameti [a], Tiago H. Falk [b]

[a] Department of Computer Engineering, Sharif University of Technology, Tehran, Iran
[b] Bloorview Research Institute, Bloorview Kids Rehab, University of Toronto, Toronto, Canada

ARTICLE INFO

Article history:
Received 20 August 2009
Accepted 11 January 2010

Keywords:
Robust speech recognition
Speech recognizer-based speech enhancement
Model distance maximizing
Spectral subtraction

ABSTRACT

This paper has presented a novel discriminative parameter calibration approach based on the model distance maximizing (MDM) framework to improve the performance of our previously-proposed method based on spectral subtraction (SS) in a likelihood-maximizing framework. In the previous work, spectral over-subtraction factors were adjusted based on the conventional maximum-likelihood (ML) approach that utilized only the true model and did not consider other confused models, thus likely reached suboptimal solutions. While in the proposed MDM framework, improved speech recognition performance is obtained by maximizing the dissimilarities among models. Experimental results based on FARSDAT, TIMIT and real distant-talking databases have demonstrated that the MDM framework outperformed ML in terms of recognition accuracy.

1. Introduction

Despite the advances in automatic speech recognition (ASR), system performance has shown to be higher when training and test conditions are matched. As a consequence, existing systems lack robustness when used in real environments outside laboratory settings. Mismatch between training and operating environments is still a major cause of degradation in recognition performance. In the past two decades, a large number of compensation techniques have been proposed to improve ASR robustness in adverse conditions. Model compensation techniques such as maximum-likelihood linear regression (MLLR) [1], maximum a posteriori (MAP) [2] and parallel model combination (PMC) [3] often aim at modifying the means and variances for each state in the hidden Markov Model (HMMs) trained with clean data so that they match noisy data observed during testing. Feature compensation approaches, in turn, attempt to either extract noise-invariant features or to increase the noise-robustness of conventional features using techniques such as codeword-dependent cepstral normalization (CDCN) [4], vector Taylor series (VTS) [5], cepstral mean subtraction (CMS) [6], relative spectral (RASTA) [7] and perceptual linear prediction (PLP) [8].

Signal compensation methods, on the other hand, aim at alleviating the detrimental noise effects by performing speech enhancement prior to the signal being fed to the recognizer. The goal is to improve the quality of the noisy speech signal and to make it sound as close as possible to its clean counterpart. Representative signal compensation strategies commonly used for speech recognition include spectral subtraction (SS) [9], Wiener filter [10], signal subspace decomposition [11], and model-based speech enhancement [12]. Among these techniques, spectral subtraction stands out as a simple yet effective method for suppressing slowly-varying additive noise.

It must be emphasized, however, that conventional speech enhancement techniques have been primarily designed to improve the intelligibility or quality of the speech signal without taking into consideration any possible detrimental effects that processing may have on succeeding systems (e.g., an ASR engine). In conventional methods (e.g., [11,13–16]) there is no feedback from the recognition stage to the enhancement stage based on the implicit assumption that an "enhanced" speech signal will result in improved recognition performance. In essence, the waveform-level criteria used during the enhancement process (e.g., maximizing signal to noise ratio or minimizing mean square error) do not guarantee a decrease in speech recognition error rate. More recently, it has been shown that by incorporating feedback information from the speech recognition system into the enhancement process (e.g., via parameter calibration), recognition rates can be further improved compared to just blindly applying the enhancement algorithm [17,18].

In previous work [19], we proposed a speech recognizer-based approach to optimize multi-band spectral subtraction parameters using a likelihood-maximizing framework. The method was based on using speech recognition likelihoods as the optimization criteria for noise suppression, as opposed to conventional methods

* Corresponding author.
E-mail addresses: babaali@ce.sharif.edu (B. BabaAli), sameti@sharif.edu (H. Sameti), tiago.falk@utoronto.ca (T.H. Falk).

1434-8411/$ - see front matter © 2010 Elsevier GmbH. All rights reserved.
doi:10.1016/j.aeue.2010.02.002

based on signal level criterion. More specifically, an utterance for which the transcription was available was used to formulate the relation between SS filter parameters and the likelihood of the true model. In a likelihood-maximizing framework that only utilizes the true model without considering other confused models, however, it is probable that only a suboptimal solution is reached. To this end, this paper presents a discriminative approach for speech recognizer-based spectral subtraction based on the framework of model distance maximization (MDM), as described in [20]. In this framework, it has been shown that by maximizing the dissimilarities between the true model and other competing models, the performance of speech recognizer-based spectral subtraction could be further improved. The proposed method has two phases: adaptation and decoding. In the adaptation phase, spectral subtraction parameters are calibrated based on maximizing acoustic likelihood distances between the true model and other competing models. In the decoding phase, these optimized parameters are applied to all incoming speech.

In Fig. 1, a block diagram of the newly-proposed framework to speech recognizer-based SS is presented in which the multiband SS enhances the received signals in such a manner so as to maximize the probability that the recognition system would estimate the correct hypothesis. This is achieved by choosing the SS parameters (in this case, spectral over-subtraction vector with coefficients) which generate a sequence of feature vectors for which the likelihood distances between the true model and other competing models ($D(\boldsymbol{a})$) is maximum. This approach results in noticeably improved recognition performance over the previously-proposed method based on the maximum-likelihood (ML) based framework. Since this framework makes use of not only the true models, but also all the competitive models to estimate the SS parameters, its adaptation phase thus has higher computational complexity, which is about M (number of competing models) times of the previous framework. We apply a strategy so that complexity is reduced to about two times of the previously-proposed method without any noticeable performance degradation. However, this gain must be considered in the context of the overall complexity of the ASR process.

The remainder of this paper is organized as follows. Sections 2 and 3 provide background on the spectral subtraction technique and the maximum likelihood-based SS (MLBSS) framework, respectively. The MDM framework for spectral subtraction is later derived in Section 4. In Section 5, we describe the proposed model distance maximization based SS (MDMBSS) paradigm. Experiments to verify the effectiveness of the proposed method are presented in Section 6. Lastly, Section 7 presents the conclusions.

2. Multi-band spectral subtraction

Among the available speech enhancement techniques, spectral subtraction is one of the most established and well-known enhancement methods in removing additive and uncorrelated noise from noisy speech. Its popularity is largely due to its simplicity, ease of implementation and low computational load. This method has been extensively studied for almost 30 years and many different variations have been proposed (e.g., [21–25]) with the majority being variants of the method proposed by Berouti et al. [21].

With the method described in [21] the speech utterance is divided into speech and nonspeech regions. It first estimates the noise spectrum from nonspeech regions and then subtracts this estimated noise spectrum from the noisy speech to obtain an estimate of the clean speech spectrum. More specifically:

$$|S_n(k)|^2 = \begin{cases} |Y_n(k)|^2 - \alpha|N_n(k)|^2 & \text{if } |Y_n(k)|^2 - \alpha|N_n(k)|^2 > \gamma|Y_n(k)|^2, \\ \gamma|Y_n(k)|^2 & \text{Otherwise,} \end{cases} \tag{1}$$

where $Y_n(k)$ represents the noisy speech short-time spectrum, $S_n(k)$ the estimated clean speech short-time spectrum, and $N_n(k)$ the noise power spectrum estimate, respectively. Parameter γ is the spectral floor factor which is a small positive number. Parameter α is the spectral over-subtraction factor and is used to compensate for errors in noise spectrum estimation. In order to obtain improved performance, these two parameters should be

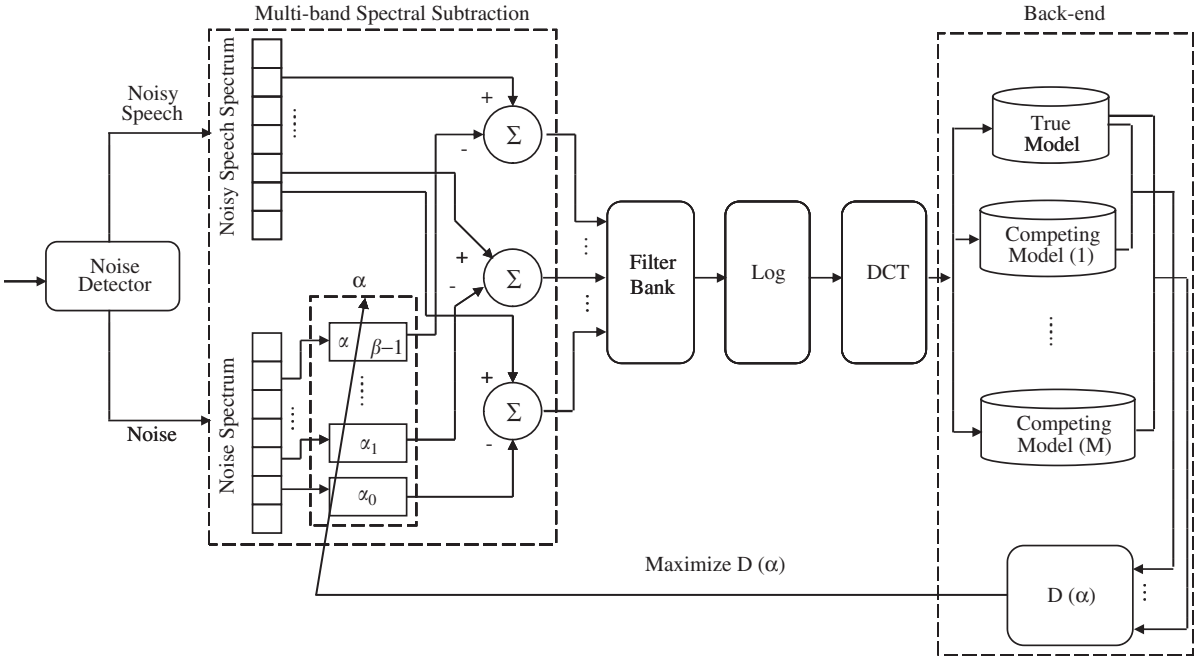

Fig. 1. Block diagram of the proposed framework.

adaptively estimated. Berouti's method assumed that the noise influenced the speech spectrum uniformly over all frequencies. Since real-world noise sources are colored and do not affect the speech signal uniformly, frequency adaptive subtraction factor based approaches have been proposed [25,26]. Lockwood and Boudy [26], for example, proposed a nonlinear spectral subtraction (NSS) method where the parameter α was frequency dependent in each frame of speech. Kamath and Loizou [25], in turn, extended this concept and proposed a multi-band spectral subtraction method that divided the speech spectrum into N bands and estimated the over subtraction factor independently for each band to take into account the variation of signal-to-noise ratios across the speech spectrum.

Previous results have shown that the multi-band approach achieves superior noise reduction and yields improved recognition results relative to full-band spectral subtraction [13,19]. Hence, this paper applies a Mel scale frequency spacing multi-band spectral subtraction approach which takes into account the fact that colored noises affect the speech spectrum differently at various frequencies. We divide the speech spectrum into mel frequency spacing bands and apply Berouti's spectral subtraction method in each band. Since the spectral over-subtraction factor is the most important parameter driving the spectral subtraction paradigm, we expect that adjusting this parameter in a model distance maximization manner for each Mel frequency band, improvement in speech recognition performance will be achieved.

3. Previous MLBSS framework

In this section, we investigate the problem of applying the likelihood maximizing framework to select the spectral over-subtraction vector so as to maximize the acoustic likelihood of the true model. Hence, the relationship between the spectral over-subtraction vector in the pre-processing stage with the acoustic likelihood of the true hypothesis in the decoding stage is formulated. These formulas depend on the feature extraction method and on the acoustic unit model. In this work, mel-frequency cepstral coefficients (MFCC) and hidden Markov model with Gaussian mixtures in each state are used as features and for modeling of the acoustic unit respectively. Speech recognition systems based on the statistical model find the acoustic unit sequence most likely to generate observed feature vectors $Z=\{\boldsymbol{z}_1,\ldots\boldsymbol{z}_2\}$ extracted from the enhanced speech signal. These observed features are a function of both the incoming speech signal and the spectral over-subtraction vector. The speech recognizer selects the most likely hypothesis based on the optimal Bayes classification criterion:

$$\hat{w}=\underset{w}{\operatorname{argmax}} P(Z(\boldsymbol{\alpha})|w)P(w), \tag{2}$$

where the observed feature vectors Z is a function of spectral over-subtraction vector $\boldsymbol{\alpha}$. Our goal is to find the vector $\boldsymbol{\alpha}$ that achieves the best recognition performance. Similar to either speaker or environmental adaptation methods, to adjust $\boldsymbol{\alpha}$, adaptation data with known transcriptions are needed. We assume that in the adaptation phase the true model of the utterance is known w_C. Hence, we can maximize Eq. (2) with respect to $\boldsymbol{\alpha}$ as:

$$\hat{\boldsymbol{\alpha}}=\underset{\boldsymbol{\alpha}}{\operatorname{argmax}} P(Z(\boldsymbol{\alpha})|w_C). \tag{3}$$

Eq. (3) shows that $\hat{\boldsymbol{\alpha}}$ is estimated based on the likelihood-maximizing framework. The likelihood-maximizing framework first generates an initial state sequence using the speech recognizer. This sequence is used to optimize the vector $\boldsymbol{\alpha}$ using a gradient-descent algorithm ensuring optimal parameters for the proposed state sequence. The utterance is decoded again using the new parameters to generate a new state sequence. This joint optimization of both the vector $\boldsymbol{\alpha}$ and state sequence continues until the recognition likelihood converges. The reader is referred to [19] for more details regarding the MLBSS framework. In the following section, the proposed MDMBSS framework is described in detail.

4. Derivation of the MDMBSS framework

As mentioned previously, in this study the model distance maximization based criterion has been used to replace the traditional likelihood based criterion with the aim to further improve speech recognition accuracy. The likelihood-maximizing framework is widely used for training and adaptation because of its simplicity and mathematical tractability. However, such framework only considers the likelihood for the true model. When there are confusable models or the amount of training or adaptation data is limited, it is very likely that only a local optimization solution is reached and suboptimal recognition accuracy is obtained.

In [20], it was shown that a significant reduction in recognition error rates could be achieved with discriminative training or an adaptation framework relative to the ML framework. Here, we propose a novel discriminative framework based on the model distance maximization criterion; a block diagram of the proposed MDMBSS paradigm is shown in Fig. 1. This approach differs from the ML approach in that ML only considers the likelihood for a single model, while the MDM framework compares the likelihood against other competing models and maximizes their likelihood differences. In MDM framework $\hat{\boldsymbol{\alpha}}$ is estimated as below:

$$\hat{\boldsymbol{\alpha}}=\underset{\boldsymbol{\alpha}}{\operatorname{argmax}}\left(P(Z(\boldsymbol{\alpha})|w_C)-\sum_{m=1}^{M}P(Z(\boldsymbol{\alpha})|w_{C_m})\right), \tag{4}$$

where C is the true model and C_m is the m th competing model and M is the number of all competing models. With HMM-based speech recognition, the acoustic likelihood is estimated by the single most likely state sequence. If S^C and S^{C_m} represent all state sequences in the true and m th competing models, respectively, and s^C and s^{C_m} represent their respective single most likely state sequences, then the maximum likelihood estimation of $\boldsymbol{\alpha}$ is given by:

$$\hat{\boldsymbol{\alpha}}=\underset{\alpha,s^C\in S^C,s^{C_m}\in S^{C_m}}{\operatorname{argmax}}\left\{\left(\begin{array}{l}\sum_i\log(P(\boldsymbol{z}_i(\boldsymbol{\alpha})|s_i^C))\\ \quad+\sum_i\log(P(s_i^C|s_{i-1}^C,w_C))\end{array}\right)-\sum_{m=1}^{M}\left(\begin{array}{l}\sum_i\log(P(\boldsymbol{z}_i(\boldsymbol{\alpha})|s_i^{C_m}))\\ \quad+\sum_i\log(P(s_i^{C_m}|s_{i-1}^{C_m},w_{C_m}))\end{array}\right)\right\}, \tag{5}$$

where $\boldsymbol{z}_i$ and s_i represent i th feature vector and state, respectively. In order to obtain $\hat{\boldsymbol{\alpha}}$ using Eq. (5), the acoustic likelihood of the true transcription (model) should be jointly maximized with respect to the state sequences of true and competing models and $\boldsymbol{\alpha}$ parameters. This joint optimization should be performed in an iterative manner.

In the adaptation phase, noisy speech is passed through the spectral subtraction filter and feature vectors $Z(\boldsymbol{\alpha})$ are extracted for known $\boldsymbol{\alpha}$ parameter. Then optimal state sequences for all models are computed using the Viterbi algorithm [27]. Given the known state sequences $\hat{s}^C$ and $\hat{s}^{C_m}$ for true and m th competing models, we want to find $\hat{\boldsymbol{\alpha}}$ such that:

$$\hat{\boldsymbol{\alpha}}=\underset{\boldsymbol{\alpha}}{\operatorname{argmax}}\left(\sum_i\log(P(\boldsymbol{z}_i(\boldsymbol{\alpha})|\hat{s}_i^C))-\sum_{m=1}^{M}\sum_i\log(P(\boldsymbol{z}_i(\boldsymbol{\alpha})|\hat{s}_i^{C_m}))\right). \tag{6}$$

Obtaining a closed-form solution for computing the optimal $\boldsymbol{\alpha}$ given a state sequence is not possible; hence, Eq. (6) can not be directly optimized with respect to $\hat{\boldsymbol{\alpha}}$ and non-linear optimization is used; more specifically, we use the gradient descent technique. We define $D(\boldsymbol{\alpha})$ to be the difference between the log likelihood of the true model and the sum of competing models. Thus

$$D(\boldsymbol{\alpha}) = \sum_i \log(P(\boldsymbol{z}_i(\boldsymbol{\alpha})|\hat{s}_i^C)) - \sum_{m=1}^{M} \sum_i \log(P(\boldsymbol{z}_i(\boldsymbol{\alpha})|\hat{s}_i^{C_m})). \tag{7}$$

The gradient vector $\nabla_{\boldsymbol{\alpha}} D(\boldsymbol{\alpha})$ is computed as:

$$\nabla_{\boldsymbol{\alpha}} D(\boldsymbol{\alpha}) = \left[\frac{\partial D(\boldsymbol{\alpha})}{\partial \alpha_0}, \frac{\partial D(\boldsymbol{\alpha})}{\partial \alpha_1}, \ldots, \frac{\partial D(\boldsymbol{\alpha})}{\partial \alpha_{B-1}}\right], \tag{8}$$

where B is the number of the mel-scaled frequency bands. Clearly, computing the gradient vector depends on both the statistical distributions in each state and the feature extraction algorithm. We derive $\nabla_{\boldsymbol{\alpha}} D(\boldsymbol{\alpha})$ assuming that each state is modeled by K mixtures of multi-dimensional Gaussians with diagonal covariance matrices. Let $\boldsymbol{\mu}_{ik}$ and $\boldsymbol{\Sigma}_{ik}$ be the mean vector and covariance matrix of the k th Gaussian density function in state s_i, respectively. We can then rewrite Eq. (7) given the optimal state sequences of true and most competing models as:

$$D(\boldsymbol{\alpha}) = \sum_i \log\left(\sum_{k=1}^{K} \exp(G_{ik}^{C}(\boldsymbol{\alpha}))\right) - \sum_{m=1}^{M} \left(\sum_i \log\left(\sum_{k=1}^{K} \exp(G_{ik}^{C_m}(\boldsymbol{\alpha}))\right)\right), \tag{9}$$

where $G_{ik}(\boldsymbol{\alpha})$ is generally defined as:

$$G_{ik}(\boldsymbol{\alpha}) = -\tfrac{1}{2}(\boldsymbol{z}_i(\boldsymbol{\alpha}) - \boldsymbol{\mu}_{ik})^T \boldsymbol{\Sigma}_{ik}^{-1} (\boldsymbol{z}_i(\boldsymbol{\alpha}) - \boldsymbol{\mu}_{ik}) + \log(\tau_{ik}\kappa_{ik}). \tag{10}$$

In Eq. (10), τ_{ik} is the weight of the k th mixture in the i th state and κ_{ik} is a normalizing constant. Using the chain rule, we have:

$$\nabla_{\boldsymbol{\alpha}} D(\boldsymbol{\alpha}) = \sum_i \sum_{k=1}^{K} \left(\gamma_{ik}^{C}(\boldsymbol{\alpha}) \frac{\partial G_{ik}^{C}(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}}\right) - \sum_{m=1}^{M} \left(\sum_i \sum_{k=1}^{K} \left(\gamma_{ik}^{C_m}(\boldsymbol{\alpha}) \frac{\partial G_{ik}^{C_m}(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}}\right)\right), \tag{11}$$

where γ_{ik} is generally defined as:

$$\gamma_{ik} = \frac{\exp(G_{ik}(\boldsymbol{\alpha}))}{\sum_{j=1}^{K} \exp(G_{ij}(\boldsymbol{\alpha}))}. \tag{12}$$

$\partial G_{ik}(\boldsymbol{\alpha})/\partial \boldsymbol{\alpha}$ is derived as:

$$\frac{\partial G_{ik}(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}} = -\frac{\partial \boldsymbol{z}_i(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}} \boldsymbol{\Sigma}_{ik}^{-1} (\boldsymbol{z}_i(\boldsymbol{\alpha}) - \boldsymbol{\mu}_{ik}). \tag{13}$$

By substituting Eq. (13) into Eq. (11), we get:

$$\nabla_{\boldsymbol{\alpha}} D(\boldsymbol{\alpha}) = -\sum_i \sum_{k=1}^{K} \left(\gamma_{ik}^{C}(\boldsymbol{\alpha}) \frac{\partial \boldsymbol{z}_i(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}} (\boldsymbol{\Sigma}_{ik}^{C})^{-1} (\boldsymbol{z}_i(\boldsymbol{\alpha}) - \boldsymbol{\mu}_{ik}^{C})\right) + \sum_{m=1}^{M} \sum_i \sum_{k=1}^{K} \left(\gamma_{ik}^{C_m}(\boldsymbol{\alpha}) \frac{\partial \boldsymbol{z}_i(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}} (\boldsymbol{\Sigma}_{ik}^{C_m})^{-1} (\boldsymbol{z}_i(\boldsymbol{\alpha}) - \boldsymbol{\mu}_{ik}^{C_m})\right). \tag{14}$$

In Eq. (14), $\partial \boldsymbol{z}_i(\boldsymbol{\alpha})/\partial \boldsymbol{\alpha}$ is the Jacobian matrix comprised of partial derivatives of each element of the i th frame feature vector with respect to each component of the over-subtraction vector $\boldsymbol{\alpha}$ and given by

$$\boldsymbol{J}_i = \frac{\partial \boldsymbol{z}_i}{\partial \boldsymbol{\alpha}} = \begin{bmatrix} \frac{\partial z_i^0}{\partial \alpha_0} & \frac{\partial z_i^1}{\partial \alpha_0} & \cdots & \frac{\partial z_i^{F-1}}{\partial \alpha_0} \\ \frac{\partial z_i^0}{\partial \alpha_1} & \frac{\partial z_i^1}{\partial \alpha_1} & \cdots & \frac{\partial z_i^{F-1}}{\partial \alpha_1} \\ \vdots & \vdots & \cdots & \vdots \\ \vdots & \vdots & \cdots & \vdots \\ \vdots & \vdots & \cdots & \vdots \\ \frac{\partial z_i^0}{\partial \alpha_{B-1}} & \frac{\partial z_i^1}{\partial \alpha_{B-1}} & \cdots & \frac{\partial z_i^{F-1}}{\partial \alpha_{B-1}} \end{bmatrix}. \tag{15}$$

The dimensionality of the Jacobian matrix is $B \times F$, where B is the number of elements in vector $\boldsymbol{\alpha}$ and F is the dimension of the feature vector. The full derivation of the Jacobian matrix when the feature vectors are MFCC is given in [19].

Since Eq. (6) makes use of not only the true model, but also all the competing models to estimate the vector $\hat{\boldsymbol{\alpha}}$, the model distance maximization procedure thus has higher computational complexity than ML. The amount of computation for the proposed framework depends primarily on the number of competing models to be computed. Consequently, the computational expense increases in proportion to the number of competing models employed. In an effort to reduce the computational complexity, in this section, we describe a technique of using a threshold to select the highest competing model amongst all models. In the proposed method, we propose to select only the highest competing model by calculating the log probability shown as follows:

$$\hat{\boldsymbol{\alpha}} = \underset{\boldsymbol{\alpha}}{\operatorname{argmax}} \left(\sum_i \log(P(\boldsymbol{z}_i(\boldsymbol{\alpha})|\hat{s}_i^{C})) - \sum_i \log(P(\boldsymbol{z}_i(\boldsymbol{\alpha})|\hat{s}_i^{C_1})) \right), \tag{16}$$

where C_1 being the identified model that attains the highest probability score among all competing models. In the adaptation procedure, only the statistical accumulators of this top competitor will be calculated and attributed to the discriminative adaptation of the vector $\boldsymbol{\alpha}$. In other words, this selection strategy indicates that only two individual models will contribute to the $\boldsymbol{\alpha}$ estimation as follows:

$$\nabla_{\boldsymbol{\alpha}} D(\boldsymbol{\alpha}) = -\sum_i \sum_{k=1}^{K} \left(\gamma_{ik}^{C}(\boldsymbol{\alpha}) \frac{\partial \boldsymbol{z}_i(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}} (\boldsymbol{\Sigma}_{ik}^{C})^{-1} (\boldsymbol{z}_i(\boldsymbol{\alpha}) - \boldsymbol{\mu}_{ik}^{C})\right) + \sum_i \sum_{k=1}^{K} \left(\gamma_{ik}^{C_1}(\boldsymbol{\alpha}) \frac{\partial \boldsymbol{z}_i(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}} (\boldsymbol{\Sigma}_{ik}^{C_1})^{-1} (\boldsymbol{z}_i(\boldsymbol{\alpha}) - \boldsymbol{\mu}_{ik}^{C_1})\right). \tag{17}$$

5. MDMBSS algorithm

Using the aforementioned framework, we propose two algorithms that are applicable to optimize parameter vector $\boldsymbol{\alpha}$. In the first method, called supervised MDMBSS, an enrollment utterance with a known transcription is used to optimize the vector $\boldsymbol{\alpha}$. This optimized vector $\boldsymbol{\alpha}$ is then used to process subsequent utterances. This algorithm is appropriate for situations in which the environment does not vary significantly over time, such as in front of a desktop computer in an office. For time-varying environments, we propose an algorithm for optimizing the vector $\boldsymbol{\alpha}$ in an unsupervised manner. In unsupervised MDMBSS, the optimization is performed on each utterance using a hypothesized transcription obtained from the recognition process. Though both of these methods are able to obtain improvements in recognition accuracy, by using the supervised method an upper bound on the performance of the proposed framework can be studied. We now describe the supervised MDMBSS algorithm in more detail.

The most likely state sequence corresponding to each utterance with a known transcription is estimated for all HMM models using the Viterbi algorithm. The features used to estimate the state sequence are generated using the vector $\boldsymbol{\alpha}$ that is derived from a previous iteration or an initial value in the first iteration. Using the estimated state sequence and applying gradient-descent-based optimization algorithm, the vector $\boldsymbol{\alpha}$ is optimized iteratively. This forms one iteration of the calibration process. Using the optimized vector $\boldsymbol{\alpha}$, a second iteration can be performed. An improved set of features for the utterance is generated and used to re-estimate the state sequence for true model and most competing model. The vector $\boldsymbol{\alpha}$ optimization process can then be repeated using the updated state sequence. The calibration process continues in an iterative manner until the

convergence condition is satisfied. Once convergence occurs, the calibration process is complete. The resulting vector $\boldsymbol{\alpha}$ is now used to process future incoming speech.

In the next section, we show that performing speech recognizer-based SS according to the model distance maximization criterion results in noticeable improvement in speech recognition accuracy over the previously-proposed MLBSS algorithm that operates according to the maximum likelihood criterion.

6. Experiments and results

In this section, the proposed MDMBSS algorithm is evaluated and is also compared with the previously-proposed MLBSS algorithm under a variety of noise conditions. In order to assess the effectiveness of the proposed algorithm, speech recognition experiments are conducted on three speech databases: FARSDAT [28], TIMIT [29], and a recorded database in a real office environment. The first and second test sets are obtained by artificially adding seven noise types (alarm, brown, multitalker, pink, restaurant, volvo, and white noise) from the NOISEX-92 database [30] to the FARSDAT and TIMIT speech databases, respectively.

Speech recognition experiments are conducted using Nevisa [31], a large-vocabulary, speaker-independent, continuous HMM-based speech recognition system developed in the speech processing lab of the Computer Engineering Department of Sharif University of Technology. Also, it is the first system to demonstrate the feasibility of accurate, speaker-independent, large-vocabulary continuous speech recognition in Persian language. In our experiments, we use clean speech for training the recognizer and the noisy speech at different SNRs to evaluate its performance. A feature vector set consisting of 36 coefficients, 12 MFCC and their first and second-order derivatives is used for all experiments.

Experiments are done in two different operational modes of the Nevisa system: phoneme recognition on FARSDAT and TIMIT databases and isolated command recognition on a distant talking database recorded in a real noisy environment. In each test, one sentence of the test set is used in the optimization phase of the MDMBSS algorithm. After vector $\boldsymbol{\alpha}$ is extracted, speech recognition is performed on the remaining test set sentences using the obtained optimized vector $\boldsymbol{\alpha}$. For each noise type, the optimization phase is done separately.

In the MDMBSS algorithm, the number of competing models (M) is one essential parameter. While a smaller value usually decreases the recognition performance of the algorithm, a greater stack size leads to increased computational complexity. Thus it is very important to find the best parameter value, which suggests a trade-off between accuracy and complexity. It has been observed that above a specific parameter threshold, no significant gains in recognition performance are observed; such threshold is often regarded as the "optimal" parameter value. Mean recognition accuracy of the MDMBSS algorithm on the TIMIT database as a function of the number of competing models M is shown in Fig. 2.

From the graph in Fig. 2, we see that the performance of MDMBSS has no noticeable improvement (about 0.3%) when the number of competing models is larger than one. Therefore, to obtain the optimal trade-off between calculation complexity and recognition performance, only the top-1 competing model is considered in the experiments described below.

6.1. Evaluation on artificially-corrupted noisy speech databases

In this section, we evaluate the recognition performance of the proposed MDMBSS method for speech artificially corrupted by noises at different SNRs using the FARSDAT and the TIMIT databases and compare it with the previously-proposed MLBSS method.

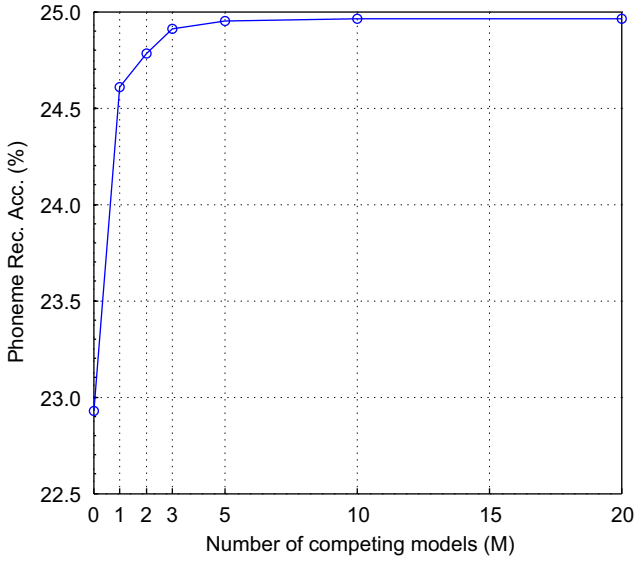

Fig. 2. Mean recognition accuracy of the MDMBSS algorithm on the TIMIT database as a function of the number of competing models in SNR value of 0 dB.

FARSDAT database consists of 6080 Persian utterances, uttered by 304 speakers. Male to female population ratio is two to one. There are a total of 405 sentences in the database and 20 utterances per speaker. The sentences are formed by using over 1000 Persian words. The database is recorded in a low noise environment featuring 31 dB signal to noise ratio in average. In all, it could be stated that FARSDAT is the counterpart of TIMIT in the Persian language. Our clean test set is selected from this database which contains 140 sentences from 7 speakers. All of the other sentences are used as the training set. To simulate a noisy environment, the testing data are contaminated by seven different NOISES at several SNRs ranging from 0 to 20 dB, at 5 dB increments, to produce various noisy test sets. To add a noise at a desired SNR, noise samples are multiplied by an attenuation factor before adding them to the speech samples. This factor depends on the speech and noise rms values calculated over the corresponding whole file, so it is computed for each speech file. The other database employed in our experiments is the well-known TIMIT database which is also corrupted using the aforementioned method.

Experiments are done in phoneme recognition mode on the hand-segmented FARSDAT and TIMIT databases. The reason for reporting phoneme recognition accuracy results instead of word recognition accuracy is that in the former case the recognition performance lies primarily on the acoustic model. For word recognition, the performance becomes sensitive to various factors such as the language model type.

The phoneme recognition accuracy results are listed in Tables 1 and 2, for MLBSS and MDMBSS methods respectively, for the noisy FARSDAT database. The corresponding results for the TIMIT database are provided in Tables 3 and 4, respectively. In the tables, the first column shows the SNRs of noisy speech and the other columns show the results for different noise types. The mean recognition accuracies are obtained by averaging the corresponding phoneme recognition accuracies for all noise types and are shown in the last column. In order to make the comparison between the MDMBSS and MLBSS methods easier, we plot the mean recognition accuracy on the FARSDAT and the TIMIT database as a function of SNR in Fig. 3.

Table 1
MLBSS method: phoneme recognition accuracy (%) on the FARSDAT database.

SNR	Noise type							Mean
	Alarm	Brown	Multitalker	Pink	Restaurant	Volvo	White	
Clean	77.36	77.36	77.36	77.36	77.36	77.36	77.36	77.36
20 dB	68.32	76.18	70.45	74.20	73.45	76.82	70.86	72.90
15 dB	61.80	75.84	64.69	62.92	66.59	76.39	59.50	66.82
10 dB	55.06	75.26	56.56	49.98	56.21	72.01	48.02	59.01
5 dB	46.64	72.26	46.23	37.06	46.12	68.78	36.78	50.55
0 dB	35.01	63.8	33.79	23.24	34.14	63.61	22.84	39.49

Table 2
MDMBSS method: phoneme recognition accuracy (%) on the FARSDAT database.

SNR	Noise type							Mean
	Alarm	Brown	Multitalker	Pink	Restaurant	Volvo	White	
Clean	77.41	77.41	77.41	77.41	77.41	77.41	77.41	77.41
20 dB	69.36	77.25	70.96	75.33	74.32	77.06	71.91	73.74
15 dB	63.10	77.19	65.38	64.21	67.92	77.65	60.80	68.04
10 dB	56.58	76.93	57.30	51.54	57.84	73.63	49.33	60.45
5 dB	48.26	73.96	46.96	38.65	47.65	70.47	38.34	52.04
0 dB	36.78	65.89	35.05	25.02	36.03	65.56	24.43	41.25

Table 3
MLBSS method: phoneme recognition accuracy (%) on the TIMIT database.

SNR	Noise type							Mean
	Alarm	Brown	Multitalker	Pink	Restaurant	Volvo	White	
Clean	66.79	66.79	66.79	66.79	66.79	66.79	66.79	66.79
20 dB	54.97	65.05	54.49	43.14	50.09	63.04	36.29	52.44
15 dB	48.25	64.58	47.17	33.77	42.22	59.38	28.31	46.24
10 dB	40.38	60.56	37.44	24.84	34.43	55.53	22.38	39.37
5 dB	31.97	54.31	26.17	17.21	25.76	51.51	11.91	31.26
0 dB	21.42	43.12	15.98	11.17	16.71	46.25	5.86	22.93

Table 4
MDMBSS method: phoneme recognition accuracy (%) on the TIMIT database.

SNR	Noise type							Mean
	Alarm	Brown	Multitalker	Pink	Restaurant	Volvo	White	
Clean	66.83	66.83	66.83	66.83	66.83	66.83	66.83	66.83
20 dB	56.01	66.38	55.12	44.23	51.25	64.40	37.25	53.52
15 dB	49.41	66.03	47.75	35.22	43.48	60.93	29.74	47.51
10 dB	41.60	62.13	38.21	26.60	35.91	57.12	24.08	40.81
5 dB	33.18	55.96	27.05	18.91	27.42	53.38	13.64	32.79
0 dB	22.86	45.10	17.09	12.98	18.40	48.20	7.64	24.61

We can make the following observations from Tables 1–4 and Fig. 3: (1) In terms of its recognition performance on the noisy speech test sets, the MDMBSS method does better than the MLBSS method. In all cases, MLBSS method achieves lower performance than the MDMBSS method. This is due to spectral distortions caused by suboptimal adjustment of the spectral subtraction factors. Moreover, (2) the higher SNR difference between the training and testing speech causes a higher degree of mismatch, thus resulting in greater degradation in recognition performance and therefore the MDMBSS method is more effective. Lastly, (3) for the clean speech test set, the recognition performance of the MDMBSS method is nearly same that of the MLBSS method.

It can be concluded from the aforementioned experiments that the MDMBSS algorithm has the capability to improve the robustness of the recognition system on artificially noise-added data. However, a direct comparison is still missing as the desired performance is needed for real environments. Therefore, a third set of experiments is performed and will be described below.

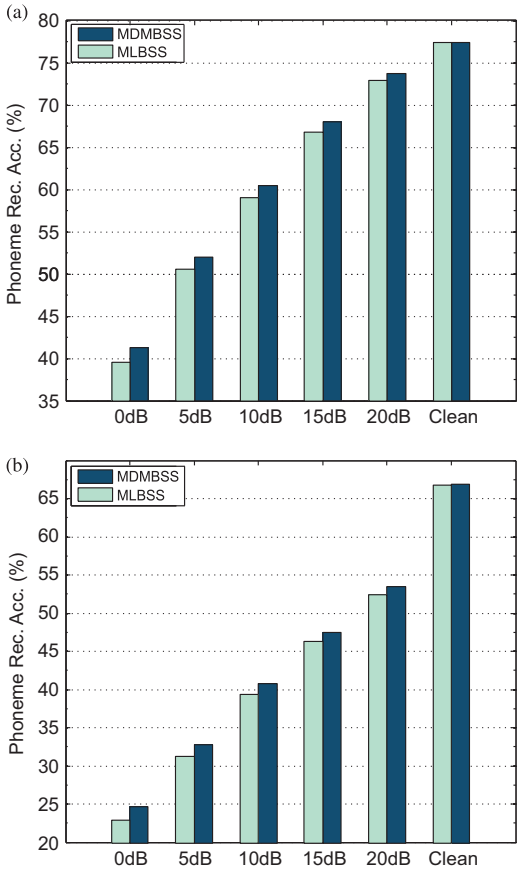

Fig. 3. Mean recognition accuracy of the MDMBSS and the MLBSS methods as a function of SNR. (a) FARSDAT database speech tested with seven added noises. (b) TIMIT database speech tested with the same seven noises.

6.2. Evaluation on real distant-talking speech in noisy environments

In order to measure the performance of the proposed algorithm in comparison with the MLBSS method, speech recognition experiments are carried out on speech data recorded in a real noisy office environment. In this experiment, we use an isolated command recognition task trained with clean isolated commands and test with noisy data captured from a microphone positioned 2 m away from the speaker. We collect the training dataset using a close-talking microphone in a quiet office using 16 female and 32 male talkers; each uttered 30 commands such as turn on/off or open/close different devices in an office. We gather the test data in a noisy office environment. For the test set, 22 male and 11 female talkers, different from those used to produce the training dataset, utter commands at a 2 m distance from the microphone. Room is 4.5 m $\times$ 3.5 m wide and the ceiling height is 3.5 which resulted in a reverberation time of approximately 300 ms ($T_{60} \cong 0.3$ s). There are some sources of noise such as three computers and a loudspeaker propagating office noise from the NOISEX database at a 40° angle with the wall. The average SNR of the test set is 15 dB. The utterances are sampled at a 16-kHz sampling rate and stored with 16-bit resolution. One utterance of each speaker in the test set is used in the optimization phase of the MDMBSS algorithm separately. Speech recognition is performed using the Nevisa system in isolated command recognition

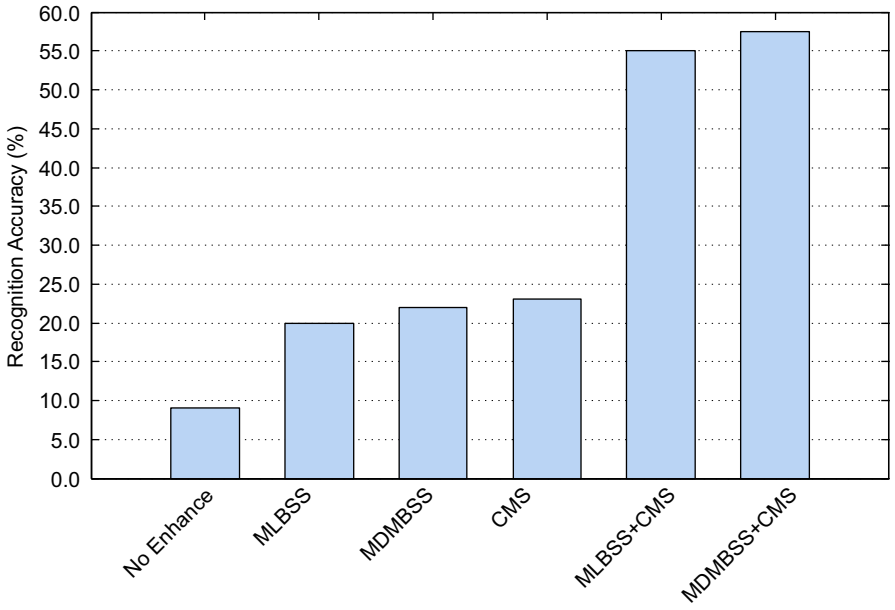

Fig. 4. Recognition accuracy (%) in isolated command recognition operational mode on data recorded in real environment versus different combinations of the proposed MDMBSS method, MLBSS and CMS.

mode. A 16-state left-to-right HMM model without skips over states is trained for each of the 30 commands. The output densities are two Gaussians with diagonal covariance matrices.

In the distant-talking environment, not only the desired speech but also sound from interfering sources is picked up. Additionally, the received signals are corrupted by echoes introduced by the acoustic environment. Therefore, these channel distortions can drastically degrade the speech recognition performance. Cepstral Mean Subtraction (CMS) [6] is one of the most popular methods employed to reduce the effects of channel distortion and variability. It reduces errors caused by the channel difference between test and training conditions caused by different recording devices and communication channels, and it is also very simple to implement. Thus, it has been adopted in many current systems. Due to the presence of the natural logarithm in the feature extraction process, linear filtering usually results in a constant offset in the filter bank or cepstral domains and hence can be subtracted from the signal. The conventional CMS estimates the sample mean vector of the cepstral vectors of an utterance and then subtracts this mean vector from every cepstral vector of the utterance. We combine the CMS with the proposed MDMBSS by mean normalization of the Jacobian matrix similar to what was done for MLBSS in [19].

The results of the different experiments are shown in Fig. 4 for MDMBSS, MLBSS, CMS, MDMBSS+CMS and MLBSS+CMS methods. Results show that adding CMS to the enhancement techniques compensates for the channel effect. This figure also indicates that: (1) each approach is able to improve the robustness of the system, (2) MDMBSS combined with CMS is more effective than all other combinations and reduces the error rate by up to 35 percent relative to MDMBSS alone and up to 48 percent relative to the no-enhancement baseline, and (3) The MDMBSS and MLBSS approaches yield better performance when combined by CMS.

7. Conclusion

We presented a speech recognizer-based approach for optimizing spectral subtraction parameters in a discriminative framework and used it as a speech enhancement method applied to the front-end of speech recognition systems. A discriminative adaptation strategy for the SS parameters based on the model distance maximizing criterion was proposed and experimental results with noisy speech demonstrated that the proposed framework improved ASR performance by an average 2% (in low SNR) relative to a previously-proposed method based on a maximum-likelihood based framework. Additionally, we showed that further improvement can be obtained in real environments by combining the proposed approach with cepstral mean subtraction to compensate for channel effects. Another line of research not in the scope of this paper but which should be explored in the future is the use of the missing data theory in the proposed framework to further improve ASR performance.

Acknowledgment

This research was in part supported by a grant from Iran Telecommunication Research Center (ITRC).

References

[1] Leggetter CJ, Woodland PC. Speaker adaptation of continuous density hmms using multivariate linear regression. In: Proceedings of the third international conference on spoken language processing (ICSLP), Japan, 1994, p. 451–4.
[2] Gauvain JL, Lee CH. Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains. IEEE Transactions on Speech and Audio Processing 1994;2:291–8.
[3] Gales MJF, Young SJ. Robust continuous speech recognition using parallel model combination. IEEE Transactions on Speech and Audio Processing 1996;4:352–9.
[4] Acero A. Acoustical and environmental robustness in automatic speech recognition. Norwell, Mass, USA: Kluwer Academic Publishers; 1993.
[5] Moreno PJ, Raj B, Stern RM. Data-driven environmental compensation for speech recognition: a unified approach. Speech Communication 1998;24: 267–85.
[6] Furui S. Cepstral analysis technique for automatic speaker verification. IEEE Transactions on Acoustics, Speech, and Signal Processing 1981;29:254–72.
[7] Hermansky H, Morgan N. Rasta processing of speech. IEEE Transactions on Speech and Audio Processing 1994;2:578–89.
[8] Hermansky H. Perceptual linear predictive (plp) analysis of speech. Journal of the Acoustical Society of America 1990;87:1738–52.
[9] Boll S. Suppression of acoustic noise in speech using spectral subtraction. IEEE Transactions on Acoustics, Speech, and Signal Processing 1979;27: 113–20.

[10] Stahl V, Fischer A, Bippus R. Quantile based noise estimation for spectral subtraction and wiener filtering. In: Proceedings of the IEEE international conference on acoustics speech and signal processing, Turkey, 2000. p. 1875–8.
[11] Hermus K, Wambacq P, Hamme HV. A review of signal subspace speech enhancement and its application to noise robust speech recognition. EURASIP Journal on Advances in Signal Processing 2007;2007.
[12] Stouten V, Hamme HV, Demuynck K, Wambacq P. Robust speech recognition using model-based feature enhancement. In: Proceedings of the eighth European conference on speech communication and technology (EUROSPEECH), Switzerland, 2003. p. 17–20.
[13] Chen J, Paliwal KK, Nakamura S. Sub-band based additive noise removal for robust speech recognition. In: Proceedings of the seventh European conference on speech communication and technology (EUROSPEECH), Denmark, 2001. p. 571–4.
[14] Fujimoto M, Ogata J, Ariki Y. Large vocabulary continuous speech recognition under real environments using adaptive sub-band spectral subtraction. In: Proceedings of the sixth international conference on spoken language processing (ICSLP), China, 2000. p. 305–8.
[15] Yamamoto H, Yamada M, Komiri Y, Ohora Y. Estimated segmental snr based adaptive spectral subtraction approach for speech recognition Technical report SP94-50, IEICE, Tokyo, Japan; 1994.
[16] Vaseghi SV, Milner BP. Noise compensation methods for hidden markov model speech recognition in adverse environments. IEEE Transactions on Speech and Audio Processing 1997;5:11–21.
[17] Seltzer ML, Raj B, Stern RM. Likelihood-maximizing beamforming for robust hands-free speech recognition. IEEE Transactions on Speech and Audio Processing 2004;12:489–98.
[18] BabaAli B, Sameti H, Safayani M. Spectral subtraction in likelihood-maximizing framework for robust speech recognition. In: Proceedings of the interspeech, Australia, 2008.
[19] BabaAli B, Sameti H, Safayani M. Likelihood-maximizing-based multiband spectral subtraction for robust speech recognition. EURASIP Journal on Advances in Signal Processing 2009;2009.
[20] Kwong S, He QH, Man KF, Tang KS. Maximum model distance approach for hmm-based speech recognition. Pattern Recognition 1998;31:219–29.
[21] Berouti M, Schwartz R, Makhoul J. Enhancement of speech corrupted by acoustic noise. In: Proceedings of the IEEE international conference on acoustics, speech, and signal processing (ICASSP), USA, 1979. p. 208–11.
[22] Sovka P, Pollak P, Kybic J. Extended spectral subtraction. In: Proceedings of European signal processing conference (EUSIPCO), Italy, 1996. p. 963–6.
[23] Sim BL, Tong YC, Chang JS, Tan CT. A parametric formulation of the generalized spectral subtraction method. IEEE Transactions on Speech and Audio Processing 1998;6:328–37.
[24] Virag N. Single channel speech enhancement based on masking properties of the human auditory system. IEEE Transactions on Speech and Audio Processing 1999;7:126–37.
[25] Kamath S, Loizou P. A multi-band spectral subtraction method for enhancing speech corrupted by colored noise. In: Proceedings of the IEEE international conference on acoustics, speech, and signal processing (ICASSP), USA, 2002. p. 4160–4.
[26] Lockwood P, Boudy J. Experiments with a nonlinear spectral subtractor (nss), hidden markov models and the projection, for robust speech recognition in cars. Speech Communication 1992;11:215–28.
[27] Viterbi A. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. IEEE Transactions on Information Theory 1967;13:260–9.
[28] Bijankhan M, Sheikhzadegan MJ. Farsdat—the speech database of farsi spoken language. In: Proceedings of the fifth Australian international conference on speech science and technology, Australia, 1994. p. 826–9.
[29] Zue V, Seneff S, Glass J. Speech database development at mit: timit and beyond. Speech Communication 1990;9:351–6.
[30] Varga AP, Steeneken HJM, Tomlinson M, Jones D. The noisex-92 study on the effect of additive noise on automatic speech recognition. Technical report, Defense Evaluation and Research Agency (DERA), Speech Research Unit, Malvern, UK; 1992.
[31] Sameti H, Veisi H, Bahrani M, BabaAli B, Hosseinzadeh K. Nevisa, a persian continuous speech recognition system. In: Proceedings of the 13th international CSI computer conference, Iran, 2008. p. 485–92.

Bagher BabaAli received his B.S. in Computer Engineering from Shiraz University, Shiraz-Iran in 2001, and his M.S. in artificial intelligence from Sharif University of Technology, Tehran, Iran in 2004. He is currently working on speech recognizer based speech enhancement method as his Ph.D. thesis in department of computer engineering, Sharif University of Technology. His research interests include speech recognition, statistical pattern recognition, speech enhancement, soft computing.

Hossein Sameti received his M.S. from Sharif University of Technology, Tehran-Iran in 1989, and his Ph.D. in Electrical Engineering from University of Waterloo, Waterloo, Canada in 1994. He is a faculty member of department of Computer Engineering at Sharif University of Technology, Tehran, Iran since 1995. His research interests include speech recognition, statistical modeling, speech enhancement and statistical language processing.

Tiago H. Falk received the B.Sc. degree from the Federal University of Pernambuco, Brazil, in 2002, and the M.Sc. (Eng.) and Ph.D. degrees from Queen's University, Canada, in 2005 and 2008, respectively, all in electrical engineering. He is currently a Postdoctoral Fellow at Bloorview Kids Rehab, affiliated with the University of Toronto, Canada. His research interests include multimedia and biomedical signal processing. Dr. Falk is recipient of the IEEE Kingston Section Ph.D. Research Excellence Award (2008), the Best Student Paper Awards at ICASSP (2005) and IWAENC (2008), and the Newton Maia Young Scientist Award (2001).