Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 54 (2012) 108–118

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Speech enhancement based on Sparse Code Shrinkage employing multiple speech models

Peter Jančovič [*], Xin Zou, Münevver Köküer

School of Electronic, Electrical & Computer Engineering, University of Birmingham, Pritchatts Road, B15 2TT Birmingham, UK

Received 17 September 2010; received in revised form 11 July 2011; accepted 12 July 2011
Available online 23 July 2011

Abstract

This paper presents a single-channel speech enhancement system based on the Sparse Code Shrinkage (SCS) algorithm and employment of multiple speech models. The enhancement system consists of two stages: training and enhancement. In the training stage, the Gaussian mixture modelling (GMM) is employed to cluster speech signals in ICA-based transform domain into several categories, and for each category a super-Gaussian model is estimated that is used during the enhancement stage. In the enhancement stage, the estimate of each signal frame is obtained as a weighted average of estimates obtained by using each speech category model. The weights are calculated according to the probability of each category, given the signal enhanced using the conventional SCS algorithm. During the enhancement, the individual speech category models are further adapted at each signal frame. Experimental evaluations are performed on speech signals from the TIMIT database, corrupted by Gaussian noise and three real-world noises, Subway, Street, and Railway noise, from the NOISEX-92 database. Evaluations are performed in terms of segmental SNR, spectral distortion and PESQ measure. Experimental results show that the proposed multi-model SCS enhancement algorithm significantly outperforms the conventional WF, SCS and multi-model WF algorithms.

Keywords: Speech enhancement; Sparse Code Shrinkage; Independent Component Analysis; Multiple models; Clustering; Super-Gaussian distribution; Gaussian mixture model (GMM)

1. Introduction

In recent years there has been a wide increase in the use of devices and applications capturing and processing speech signals, such as mobile phones, hearing aids and voice-automated services. The speech signal in real-world applications is often corrupted due to communication channel or background additive noise. The aim of speech enhancement is to estimate the clean speech signal from the observed noisy signal. The enhancement may be performed by employing signals recorded by a single microphone or by multiple microphones. This paper is focused on enhancement of speech corrupted by additive noise when using single microphone recordings.

Speech enhancement is typically performed in a short-time manner in a transform domain, i.e., the speech signal is split into short-time segments, referred to as frames, which are usually transformed by a (linear) transformation. Currently the most widely-used is the short-time spectral domain obtained by employing the discrete Fourier transform (DFT). Spectral subtraction (Boll, 1979) and Wiener filtering (WF) (Lim and Oppenheim, 1979; McAulay and Malpass, 1980) were among the first introduced speech enhancement techniques. They remain popular today due to their reasonable performance and low computational complexity. Spectral subtraction performs subtraction of an estimated noise magnitude spectrum from a noisy speech magnitude spectrum. The WF algorithm performs filtering of a noisy speech signal by using a filter derived

[*] Corresponding author. Tel.: +44 121 4144316; fax: +44 121 4144291.
E-mail addresses: p.jancovic@bham.ac.uk (P. Jančovič), x.zou@ulster.ac.uk (X. Zou), m.kokuer@bham.ac.uk, m.kokuer@qmul.ac.uk (M. Köküer).

0167-6393/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2011.07.005

based on the minimum mean-square error (MSE) criterion. Another widely-used speech enhancement technique based on minimising the MSE of the speech spectral amplitude, or its logarithm, was proposed in (Ephraim and Malah, 1984, 1985).

Some enhancement techniques attempt to incorporate masking properties of the human auditory system, which state that an additive noise of amplitude falling below the masking threshold is not audible to the human ear. This was investigated within a generalised spectral subtraction framework in (Virag, 1999), and within an adaptive β-order minimum MSE speech enhancement algorithm in (You et al., 2006). The incorporation of perceptual weighting presented in (Hu and Loizou, 2003) was shown to have the form of a modified version of the Wiener filter.

The speech enhancement approaches based on minimising MSE rely on a Gaussian model for both noise and speech. While for many applications, the spectral components of noise can be modelled by a Gaussian random variable, it has been shown that clean speech spectral components can be more effectively modelled by non-Gaussian distributions, e.g. (Lotter and Vary, 2005; Gazor and Zhang, 2005). This has led to recent speech enhancement techniques based on the maximum-a-posteriori (MAP) framework that incorporate a non-Gaussian model of speech. An MAP-based signal enhancement algorithm, referred to as Sparse Code Shrinkage (SCS), was proposed in (Hyvarinen, 1999, 2001). The SCS algorithm performs the enhancement in the transform domain obtained by employing the Independent Component Analysis (ICA). This algorithm was originally proposed for image enhancement and was later successfully applied for speech signal enhancement (Lee et al., 2000; Potamitis et al., 2001; Choi et al., 2001). Recently, other MAP-based speech enhancement algorithms have been proposed in (Martin, 2002; Wolfe and Godsill, 2003; Lotter and Vary, 2005; Gazor and Zhang, 2005; Zou et al., 2008a).

Currently, most of the speech enhancement algorithms employ a single PDF model of speech and noise. The employment of more sophisticated statistical models of speech and noise, such as hidden Markov model (HMM), was proposed some time ago in (Ephraim, 1992a,b), but investigated only more recently to a greater extent. For instance, the employment of HMMs in (Sameti et al., 1998; Zhao and Kleijn, 2007), Gaussian mixture models and codebooks in (Srinivasan et al., 2006; Vaseghi, 2005; Kundu et al., 2008; Ming et al., 2011). The authors in (Srinivasan et al., 2006) used codebooks for modelling autoregressive (AR) parameters of the clean speech and several types of noises. The enhancement was performed by searching for the best combination of the speech and noise AR parameters from the codebooks that maximise the likelihood of the noisy speech. A similar idea but incorporated within the HMM framework was presented in (Zhao and Kleijn, 2007). The authors in (Kundu et al., 2008) derived an MMSE estimator under the consideration of a GMM model for speech signal. The authors in (Ming et al., 2011) used a GMM, constructed from a speech corpus corrupted by various noise types, to identify the longest matching segments between the noisy sentence and the corpus sentences. The identified corpus segments of the corresponding clean speech were then employed in the Wiener filter framework to enhance the noisy speech. The paper, however, was oriented towards the separation of speech from a mixture signal containing other speech or music. We have recently investigated the employment of multiple super-Gaussian PDF models of speech in the context of the SCS algorithm in (Zou et al., 2008b; Jančovič et al., 2011), where speech frames were categorised into two classes based on voicing information estimated using the method presented in (Jančovič and Köküer, 2007).

In this paper, we present our further research on speech enhancement employing multiple speech models within the Sparse Code Shrinkage algorithm. Unlike our initial works, as mentioned above, the speech signals here are clustered into a number of categories by employing the Gaussian mixture modelling (GMM), the enhancement scheme is improved, and a more in-depth performance analysis is carried out. The GMM clustering is performed on the ICA-transformed signal frames and each component of the GMM represents a speech category. For each speech category, a super-Gaussian model of speech, which is used in the enhancement process, is then estimated. The clustering and individual speech model parameter estimation is performed on the training data which is different to data used during the enhancement process. The estimate of the enhanced signal frame is calculated as a weighted average of a set of estimates obtained by using each individual speech category model. The weights are calculated based on the probability of each speech category, given the signal frame enhanced using the conventional SCS algorithm that employs a single speech model (referred to as the primary enhanced signal). A comparison to the idealised results obtained using the clean speech signal frames for weights calculation is also presented. Further, individual speech category models are adapted at each signal frame during the enhancement process based on the primary enhanced signal. The experimental evaluation is performed on speech signals from the TIMIT database, corrupted by Gaussian noise and three real-world non-stationary noises from the NOISEX-92 database, specifically, Subway, Street, and Railway noise. The enhancement performance is evaluated in terms of the segmental signal-to-noise ratio (SNR), spectral distortion (SD) and ITU-Perceptual Evaluation of Speech Quality (PESQ) measure. Evaluations of the multi-model SCS algorithm are provided as a function of the number of speech clusters used and using various ways of weights calculation. A comparison to the conventional WF and SCS algorithms employing a single speech model is provided. In addition, for comparison, we also developed a multi-model WF algorithm that follows the proposed scheme. Experimental results demonstrate large improvement gains of the proposed multi-model SCS algorithm in comparison to the above mentioned algorithms. The paper is organised as follows. Section 2 presents a background on the conventional WF and SCS enhancement algorithms. Section 3 then describes the proposed multi-model SCS enhancement algorithm. Section 4 presents the experimental results and Section 5 gives conclusion.

2. Speech enhancement in additive noise

2.1. Problem formulation

Let us consider the problem of the enhancement of a speech signal contaminated by an independent additive noise. Let $x(n)$ and $v(n)$ denote the sampled clean speech signal and noise signal, respectively. The observed noisy speech signal $y(n)$ is

$$y(n) = x(n) + v(n). \tag{1}$$

We consider speech enhancement performed in a transform domain, i.e., speech signal is split into short-time segments (referred to as frames), which are usually weighted by a windowing function, and then transformed by a (linear) transformation. Let us denote by vector $\boldsymbol{y}$ and $\boldsymbol{x}$ a block of K samples of the noisy signal and the clean signal, respectively, and by $\boldsymbol{z}$ and $\boldsymbol{s}$ their corresponding transformed versions. Throughout the paper, a vector of a signal frame is denoted by bold type and its individual components by normal type, with the component indices being omitted for simplicity. The variances of the signal and noise in transform domain are denoted by σ_s^2 and σ^2, respectively. Note that the variances of the signal and noise and the parameters of the models are calculated for each individual component.

2.2. Wiener filtering

Wiener filtering (WF) performs the enhancement of the noisy speech signal in the DFT-based transform domain by using a filter derived based on the minimum mean-square error (MMSE) criterion. For each frequency bin, the estimate of the clean signal, denoted by $\hat{s}$, is obtained by multiplying the noisy signal z with a filter H, i.e., $\hat{s} = Hz$, where the filter H is defined as

$$H = \frac{\sigma_s^2}{\sigma_s^2 + \sigma^2}. \tag{2}$$

The WF enhancement rule is derived based on the assumption that both signal and noise are Gaussian distributed.

2.3. Sparse Code Shrinkage

The Sparse Code Shrinkage (SCS) algorithm performs the enhancement of the noisy speech in the ICA-based transform domain, where the ICA transformation matrix is obtained based on clean speech data before the enhancement process. The following sections describe the process of obtaining the ICA transformation matrix and the SCS enhancement rule.

2.3.1. Independent Component Analysis

The Independent Component Analysis (ICA) can be used to transform data to be as statistically independent from each other as possible. Considering linear ICA, the data model in our context can be written as $\boldsymbol{x} = \boldsymbol{A}\boldsymbol{s}$, where $\boldsymbol{x}$ is the frame vector of the K observed clean speech samples, $\boldsymbol{s}$ is the vector of the underlying (independent) sources, and $\boldsymbol{A}$ is a mixing matrix. The goal of ICA is to estimate, from the observed signal $\boldsymbol{x}$, both the $\boldsymbol{A}$ and $\boldsymbol{s}$ such that the estimated sources $\boldsymbol{s}$ are statistically independent.

The ICA algorithms usually consist of two stages: pre-whitening and unmixing. The pre-whitening stage decorrelates and normalises the input mixture $\boldsymbol{x}$ by applying PCA-based transformation matrix, denoted by $\boldsymbol{U}$. The unmixing stage then further transforms the pre-whitened signal, denoted by $\boldsymbol{x}'$, to be independent using an orthonormal transformation matrix $\boldsymbol{B}$. The orthonormal matrix $\boldsymbol{B}$ can be estimated based on various criteria. One of the most widely used ICA algorithms is the negentropy-based fastICA (Hyvarinen et al., 2001). The updating rule of $\boldsymbol{B}$ can be written as

$$\boldsymbol{B} \leftarrow E\{\boldsymbol{x}'g(\boldsymbol{B}^T\boldsymbol{x}')\} - E\{g'(\boldsymbol{B}^T\boldsymbol{x}')\}\boldsymbol{B}, \tag{3}$$

where g is a non-linear function and g' is the first derivative of g.

For a given signal frame $\boldsymbol{x}$, the independent components of $\boldsymbol{x}$ can be obtained by

$$\boldsymbol{s} = \boldsymbol{W}\boldsymbol{x}, \tag{4}$$

where $\boldsymbol{W}$ is the estimated ICA unmixing matrix (i.e., the inverse of the mixing matrix $\boldsymbol{A}$), which is obtained as $\boldsymbol{W} = \boldsymbol{B}\boldsymbol{U}$.

2.3.2. Enhancement rule

The SCS algorithm is performed in the orthogonolised ICA transform domain. Each signal frame $\boldsymbol{x}$ is transformed using Eq. 4 with $\boldsymbol{W}$ being replaced by its orthogonalised version calculated as

$$\boldsymbol{W} = \boldsymbol{W}(\boldsymbol{W}^T\boldsymbol{W})^{-1/2}. \tag{5}$$

In the ICA domain, the components of the transformed speech frame $\boldsymbol{s}$ are independent from each other and thus the enhancement of individual components can be performed separately.

We employ the PDF model of independent components s as used in (Hyvarinen, 1999; Potamitis et al., 2001)

$$p(s) = \frac{1}{2d}\,\frac{(\alpha+2)[\alpha(\alpha+1)/2]^{(\alpha/2+1)}}{\left[\sqrt{\alpha(\alpha+1)/2} + |s/d|\right]^{(\alpha+3)}}, \tag{6}$$

where a, α and d are distribution model parameters, which can be calculated using the following formulas (Choi et al., 2001)

$$p(s=0)=\frac{v\eta}{2\Gamma(1/v)},\quad v=F\left(\frac{\mu_{|s|}}{\sqrt{\sigma_s^2}}\right), \tag{7}$$

$$F(u)=\frac{\Gamma(2/u)}{\sqrt{\Gamma(1/u)\Gamma(3/u)}},\quad \eta=\frac{1}{\sqrt{\sigma_s^2}}\left(\frac{\Gamma(3/v)}{\Gamma(1/v)}\right)^{1/2}, \tag{8}$$

$$d=\sqrt{\sigma_s^2},\quad k=d^2p(s=0)^2, \tag{9}$$

$$\alpha=\frac{2-k+\sqrt{k(k+4)}}{2k-1},\quad a=\sqrt{\frac{\alpha(\alpha+1)}{2}}, \tag{10}$$

where $\mu_{|s|}$ denotes the mean of $|s|$ and $\Gamma(\cdot)$ is the gamma function.

The estimate $\hat{s}$ of the clean signal in the ICA transform domain can be obtained from the noisy signal z by using the SCS estimation rule (Hyvarinen, 1999)

$$\hat{s}=\operatorname{sign}(z)\times\max\left(0,\frac{|z|-ad}{2}+\frac{1}{2}\sqrt{(|z|+ad)^2-4\sigma^2(\alpha+3)}\right). \tag{11}$$

The enhanced speech signal frame in the time-domain $\hat{\boldsymbol{x}}$ can then be obtained by $\hat{\boldsymbol{x}}=\boldsymbol{W}^T\hat{\boldsymbol{s}}$.

The SCS enhancement rule is derived based on the assumption that the signal is non-Gaussian and noise is Gaussian.

3. Speech enhancement using Sparse Code Shrinkage and multiple speech models

Speech enhancement using the SCS algorithm has previously been performed by employing a single model of speech signal. This section presents the proposed enhancement scheme based on employing multiple speech models and the SCS enhancement algorithm. Our initial research in (Zou et al., 2008b; Jančovič et al., 2011) categorised the speech signals at the frame level into only two categories, based on the voicing information estimated using the method presented in (Jančovič and Köküer, 2007). Here, we present an improved enhancement scheme and employ GMM for clustering of speech into a larger number of clusters.

3.1. Motivation

The denoising capability of speech signal enhancement algorithms is influenced by the accuracy of the PDF model of the signal. Furthermore, it has been shown in (Hyvarinen, 1999) that the denoising capability of the SCS algorithm improves with increasing the non-gaussianity of the signal $\boldsymbol{s}$. The employment of a single PDF model of the entire speech, as in the conventional SCS speech enhancement, captures only the global statistical characteristics. Due to the variety of speech sounds, the employment of multiple PDF models of speech can better account for different statistical properties of various speech sounds and also increase the non-gaussianity and thus should provide improved enhancement performance.

3.2. The proposed scheme

The overall schematic diagram of the proposed enhancement algorithm employing multiple PDF models is depicted in Fig. 1. The proposed enhancement scheme consists of three main parts: the GMM-based data clustering, the primary enhancement, and the advanced enhancement. The GMM-based data clustering and both of the enhancement stages are performed in the same ICA-based transform domain. The aim of the GMM-based data clustering, which is performed using the training data before the enhancement process itself, is to cluster the entire speech data into several categories. With each category, we then associate and construct a PDF model that will be used in the advanced enhancement stage.

The enhancement is performed at a frame level. In the primary enhancement stage, a conventional enhancement algorithm employing a global speech PDF model is used to provide the primary signal estimate, denoted by $\hat{\boldsymbol{s}}^{pri}$. The obtained primary signal estimate is used in the advanced enhancement stage in two ways: (i) for adaptation of the PDF model of each speech category; (ii) for calculation of the weights with which each of the signal estimate obtained in the advanced enhancement will contribute to the final signal estimate, denoted by $\hat{\boldsymbol{s}}$. In the advanced enhancement stage, a set of signal estimates, denoted by $\hat{\boldsymbol{s}}^{adv}$, are obtained based on the adapted PDF models of each category. The final signal estimate of a

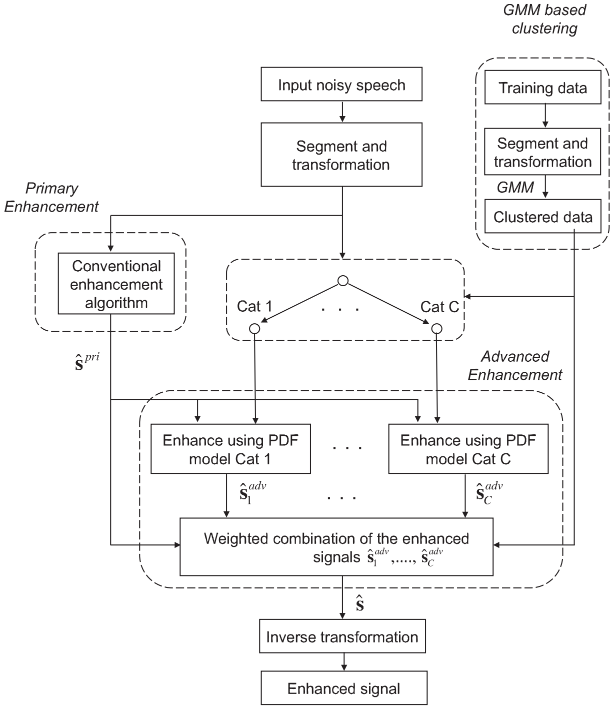
GMM based clustering
Input noisy speech
Training data
Segment and transformation
Segment and transformation
GMM
Clustered data
Primary Enhancement
Conventional enhancement algorithm
Cat 1
. . .
Cat C
$\hat{\mathbf{s}}^{pri}$
Advanced Enhancement
Enhance using PDF model Cat 1
. . .
Enhance using PDF model Cat C
$\hat{\mathbf{s}}_1^{adv}$
. . .
$\hat{\mathbf{s}}_C^{adv}$
Weighted combination of the enhanced signals $\hat{\mathbf{s}}_1^{adv}$,...., $\hat{\mathbf{s}}_C^{adv}$
$\hat{\mathbf{s}}$
Inverse transformation
Enhanced signal

Fig. 1. Schema of the proposed speech enhancement algorithm employing multiple speech models.

given frame is obtained as a weighted sum of the advanced estimates from each category.

The following sections describe the GMM-based clustering of speech signals and the advanced enhancement stage. The primary enhancement stage corresponds to conventional WF or SCS algorithms, presented earlier in Sections 2.2 and 2.3, for multi-model WF and multi-model SCS algorithms respectively.

3.3. Clustering of speech based on GMM

The clustering of speech signal frames into several categories and the calculation of the PDF model for each category, which is then used in the advanced enhancement stage, are performed before the enhancement process itself. The clustering is performed using a large collection of clean speech data, which are not used in the evaluation of the enhancement performance, and these are referred to as the training speech data.

The clustering could be performed using various techniques. Here we employed the Gaussian mixture modelling (GMM). The GMM models the distribution of the ICA transformed signal frames $\boldsymbol{s}$ of the clean training speech data by using a linear combination of C Gaussian density functions as

$$P(\boldsymbol{s}|\lambda) = \sum_{c=1}^{C} w_c \mathcal{N}\left(\boldsymbol{s}; \mu_c, \sigma_c^2\right), \tag{12}$$

where $\mathcal{N}(\mu_c, \sigma_c^2)$ denotes the cth Gaussian mixture component with mean μ_c and variance σ_c^2, and w_c is the weight. We refer to each Gaussian mixture component as a category and denote its parameters by $\lambda_c = \{w_c, \mu_c, \sigma_c^2\}$.

After the parameters λ_c of the individual speech categories are obtained, each transformed signal frame $\boldsymbol{s}$ from the training data is assigned to a category c^* based on the maximum likelihood criterion as $c^* = \text{argmax}_c P(\boldsymbol{s}|\lambda_c)$. The parameters of the super-Gaussian PDF model of each category to be used in the enhancement process are then calculated based on the set of signal frames associated with that category. The parameters of the PDF model of each category c used for enhancement are $\mu_{|s_c|}$ and $\sigma_{s_c}^2$ (for the model as in Eq. 6) in the case of the multi-model SCS enhancement, and $\sigma_{s_c}^2$ in the case of multi-model WF enhancement. Examples of PDF models of two speech categories, in the case of speech being clustered into sixteen categories, are depicted in Fig. 2. It can be seen that there are large differences between models in each category.

3.4. Advanced enhancement

The proposed advanced enhancement stage employs multiple PDF models, with each PDF model corresponding to a particular speech category. In order to better model the dynamics of the speech, the category PDF model parameters are adapted for each component using the primary estimate of the frame $\hat{\boldsymbol{s}}^{pri}$ as

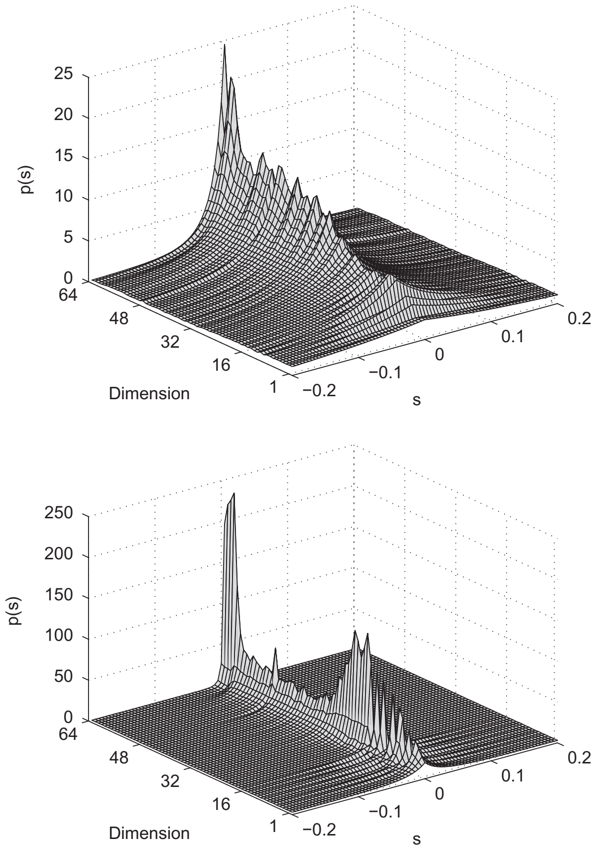

Fig. 2. Examples of PDF models of two speech categories.

$$\mu_{|s_c|} = \eta\mu_{|s_c|} + (1-\eta)|\hat{s}^{pri}|, \tag{13}$$

$$\sigma_{s_c}^2 = \eta\sigma_{s_c}^2 + (1-\eta)(\hat{s}^{pri})^2, \tag{14}$$

where the parameter η controls the level of adaptation from the current model and the value of 0.95 was found to be suitable based on experiments similar to those performed in Section 4.2. We have observed that it is advantageous to adapt the parameters of only the few models which achieve highest likelihood values of the primary enhanced signal, i.e., $P(\hat{\boldsymbol{s}}_{pri}|\lambda_c)$. This is reasonable, because models which are far from the current signal frame should not be adapted. Adapting many (or all) of the models would effectively cause the models to converge to the same single model. In our experiments, similar performance was observed when adapting between 1 to 4 best models.

The advanced estimate of the signal $\hat{\boldsymbol{s}}_c^{adv}$ is obtained using such adapted PDF model parameters of each category c. The final estimate is then obtained as a weighted summation of all the advanced estimates as

$$\hat{\boldsymbol{s}} = \sum_{c=1}^{C} a_c \hat{\boldsymbol{s}}_c^{adv}, \tag{15}$$

where a_c denotes the weight with which each estimate contributes to the final estimate. The weights a_c correspond to the probability of each category c, given the primary signal frame estimate $\hat{\boldsymbol{s}}^{pri}$, and are calculated as

$$a_c = \frac{P(\hat{\boldsymbol{s}}^{pri}|\lambda_c)}{\sum_{c=1}^{C} P(\hat{\boldsymbol{s}}^{pri}|\lambda_c)}, \tag{16}$$

where $P(\hat{\boldsymbol{s}}^{pri}|\lambda_c) = w_c\mathcal{N}(\hat{\boldsymbol{s}}^{pri}; \mu_c, \sigma_c^2)$. We have observed that the direct use of the likelihood values in Eq. 16 provides a set of weights in which a single weight often largely dominates, i.e., a single weight has a value close to 1 while all other weights are close to 0. Such close to binary weights may negatively affect the enhancement performance because they are calculated based on the primary enhanced signal $\hat{\boldsymbol{s}}^{pri}$ which inevitably contains some errors. In order to provide a softer set of weights, we have explored the use of the sigmoid and logarithm functions to compress the range of the likelihood values $P(\hat{\boldsymbol{s}}^{pri}|\lambda_c)$ prior to their use in Eq. 16 and found that this can improve the performance. In the case of using the logarithm function, the log-likelihood values were floored at 0.1. The flooring will cause the weight values to be approximately equal in the case of the likelihood values being low for all the categories. After applying the compression function, the weight values were rescaled in order to sum to 1. A summary of these experimental evaluations is presented in Section 4.2.

4. Experimental evaluations

4.1. Experimental set-up and evaluation measures

Experiments were performed using the TIMIT database, downsampled to 8 kHz. Ten utterances of the first 100 speakers from the test set were used. Eight utterances (‘si’ and ‘sx’) of each speaker were collected over all the speakers to form the training data set – this was used for estimating the ICA transformation matrix and training the speech PDF models. The remaining two utterances (‘sa’) of each speaker were used for enhancement evaluation. Clean speech signal was corrupted at input global signal-to-noise ratios (SNRs) of 10, 15, and 20 dB by Gaussian noise and real-world Subway, Street and Railway noises from the NOISEX-92 database. The noise corrupting each speech utterance was taken from the entire noise signal at a random position. The utterances were segmented into frames of 64 samples with an overlap of 32 samples between adjacent frames. Noise variance was estimated for each element in the transform domain from the noisy signal based on the first 80 ms of the signal which was assumed to contain no speech. This provided an initial noise estimate. A speech activity detector, as described in (Zou et al., 2008b), was employed to detect speech absent signal frames. The noise variance was updated over time, based on the detected speech absent frames as $\sigma^2 = \rho\sigma^2 + (1-\rho)z^2$, where ρ was set to 0.98. All the enhancement algorithms presented used the same procedure. The proposed multi-model SCS enhancement scheme, denoted by MultiSCS, employed orthogonalised ICA transformation matrix obtained based on applying the fastICA algorithm (Hyvarinen et al., 2001) on the training data set. The GMM-based speech clustering was performed using HTK software (Young et al., 2000). The performance of the proposed multi-model SCS algorithm is compared to the conventional Wiener filtering (WF) and conventional SCS algorithms, which both employ a single speech model. In addition, we also developed a multi-model WF algorithm, denoted by MultiWF, which employs multiple speech models according to the proposed scheme of the MultiSCS algorithm, but uses the WF enhancement rule. The WF and MultiWF are performed in the Fourier domain. All the enhancement algorithms employed dynamic speech model(s) – the parameters of the models were adapted over time based on the detected speech frames using a similar formula as used above for noise variance adaptation. The adaptation coefficient was set for each algorithm to a value that provided the best performance.

Experimental evaluations are performed in terms of segmental SNR, spectral distortion (SD) and ITU-Perceptual Evaluation of Speech Quality (PESQ) measure. In the calculation of the segmental SNR and SD, the silence parts at the beginning and ending of each utterance were excluded. Let us denote the error between the original speech signal $x(n)$ and the estimated speech signal $\hat{x}(n)$ by $e(n) = x(n) - \hat{x}(n)$. The segmental SNR is then defined as

$$segSNR = \frac{1}{J}\sum_{j=0}^{J-1} 10\log_{10}\frac{\sum_{i=0}^{L-1} x^2(L\cdot j+i)}{\sum_{i=0}^{L-1} e^2(L\cdot j+i)}, \tag{17}$$

where J is the number of frames, j is the frame-index, and L is the frame length, which was set to 80 samples. The SD between the original clean signal and the estimate is defined as the average difference between the logarithm spectrum of each frame of the clean speech spectrum $X(k,j)$ and the enhanced speech signal $\widehat{X}(k,j)$ as

$$SD = \frac{1}{4J}\sum_{j=0}^{J-1}\sum_{k=0}^{255} 20|\log_{10}|X(k,j)| - \log_{10}|\widehat{X}(k,j)||, \tag{18}$$

where k and j are the frequency-index and frame-index, respectively.

4.2. Evaluation of the effect of clustering and cluster weight calculation on the enhancement performance

This section evaluates the enhancement performance of the proposed multi-model SCS-based enhancement algorithm with respect to the number of speech clusters used and the way of calculating the weights in Eq. 16 with which each cluster estimate contributes to the overall enhanced signal estimate. Experiments presented in this section were performed using a subset of 20 utterances from 10 randomly-selected speakers and the speech signal was corrupted by Gaussian noise at 10 dB global SNR.

First, we evaluate the effect of the number of speech clusters employed. The results are presented in Fig. 3. In figures, the MultiSCS-Ideal and MultiSCS denotes the idealised and practical case, respectively. The weights in Eq. 16 are calculated using the clean signal $\boldsymbol{s}$ in the case of MultiSCS-Ideal, while the primary enhanced signal $\hat{\boldsymbol{s}}^{pri}$ is used in MultiSCS case. In both cases, the weights are calculated using the log-likelihood values. It can be seen that the enhancement performance improves significantly by

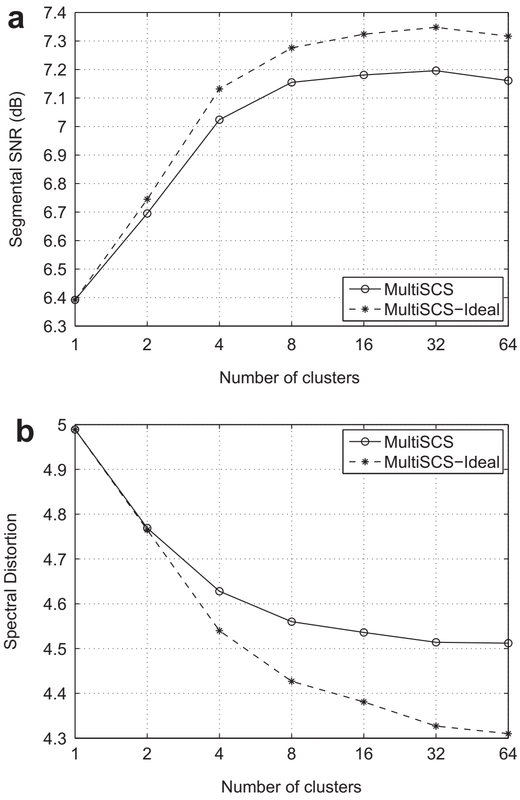

Fig. 3. Experimental results in terms of segmental SNR (a) and spectral distortion (b) on Gaussian noise at the SNR of 10 dB obtained by using various number of speech clusters.

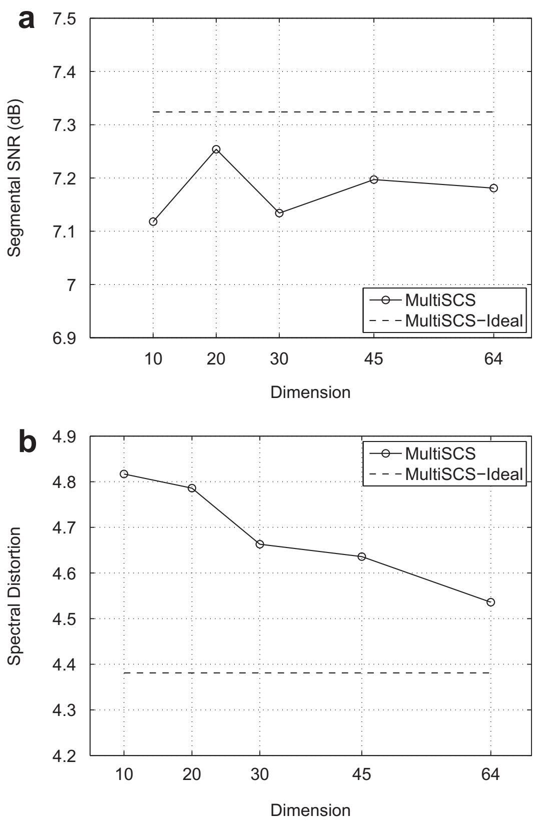

Fig. 4. Experimental results in terms of segmental SNR (a) and spectral distortion (b) on Gaussian noise at the SNR of 10 dB obtained by using the clustering and weight calculation in Eq. 16 based on the ICA-transformed signal frame vector $\boldsymbol{s}$ of various (reduced) dimensions.

employing multiple speech models instead of a single speech model that is used in the conventional SCS algorithm. For instance, the MultiSCS-Ideal case, using 32 speech models, obtained (compared to the use of a single speech model) an increase in the segmental SNR from 6.39 dB to 7.35 dB, and at the same time a decrease in the SD from 4.99 to 4.33. In terms of the segmental SNR, both the idealised and practical cases obtained the best performance using 32 speech clusters. When using 64 clusters, the segmental SNR started to decrease, which may be caused by the fact that although the increase of the number of clusters can provide better accuracy in modelling the training speech data, it then does not perform well on unseen speech data used for evaluation of the enhancement performance. In terms of the SD, the performance improves slightly further when using 64 clusters in the idealised case, but in the practical case is the same as in the case of 32 clusters. Similar performance trends as in the Gaussian noise were also observed in other noisy conditions. Since the performance in the practical case differs only slightly when using 16 and 32 speech models, the remaining experiments in this section were performed using 16 speech models to reduce the computation time.

Next, we explore the weight calculation in Eq. 16 using the log-likelihoods computed based on employing a modified feature vector. The speech clustering and weight calculation above were performed using the entire ICA-transformed signal frame vector $\boldsymbol{s}$ (or its estimate $\hat{\boldsymbol{s}}^{pri}$). However, some elements of the vector may express a low variance and thus carry only small amount of information about the speech signal. In noisy conditions, these elements may contain noise and thus negatively affect the weight calculation – and consequently the enhancement performance. Thus, we explored the effect of performing the speech clustering and weight calculation using a feature vector containing only a given number of high-variance elements. The variance of each element was estimated using the training data set. The enhancement performance as a function of the vector dimension is depicted for the MultiSCS practical case in

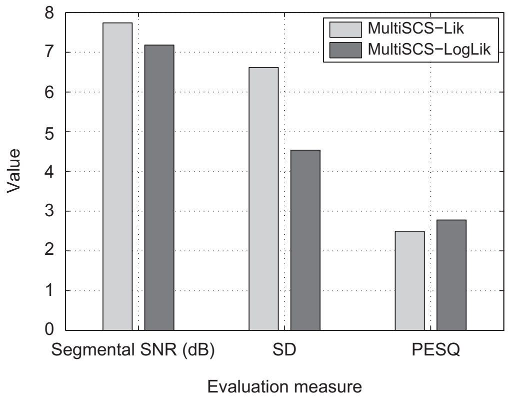

Fig. 5. Experimental results on Gaussian noise at the SNR of 10 dB obtained by calculating the weights a_c using the likelihood and log-likelihood values.

Fig. 4. For comparison, results are also presented for the MultiSCS-Ideal case, in which the weights are calculated using the full 64-dimensional clean speech signal vector. It can be seen that the segmental SNR shows some smaller fluctuations but no distinct trend as a function of the feature vector dimension. In the case of the SD, the lowest SD value is achieved by using the entire 64-dimensional feature vector, and the SD increases considerably as the dimension decreases. Results of a similar trend were observed also in other noisy conditions. Thus, the entire 64-dimensional feature vector is used throughout the rest of the paper.

Finally, we explore the calculation of the weights in Eq. 16 by employing directly the likelihood values and logarithm of the likelihood values. The obtained results are presented for the practical case in terms of the segmental SNR, SD and PESQ measures in Fig. 5. It can be seen that the direct use of the likelihood values provides little improvement of the segmental SNR than when using the log-likelihood values, however, this is at the expense of a large increase of the SD. Further, the PESQ measure is considerably better when using the log-likelihood values. As such, the weights are calculated using the log-likelihood values throughout the experimental evaluations presented in the paper.

4.3. Experimental results of the proposed algorithm and comparison to other methods

This section presents the experimental results of the proposed multi-model SCS algorithm on 200 speech utterances from 100 speakers in various noisy conditions and compares its performance to the conventional WF and SCS algorithms and the multi-model WF algorithm. The results presented for MultiWF ad MultiSCS algorithms are obtained using 32 speech models and practical weight calculation using log-likelihoods.

The results obtained in Gaussian, Subway, Street and Railway noisy conditions are presented in Tables 1–4,

Table 1
Speech quality evaluation in terms of the segmental SNR, SD and PESQ for speech corrupted by Gaussian noise at various input SNRs.

Algorithm	10 dB			15 dB			20 dB		
	SNR	SD	PESQ	SNR	SD	PESQ	SNR	SD	PESQ
Baseline	1.84	9.26	2.10	6.84	6.73	2.47	11.84	4.70	2.83
WF	5.77	5.35	2.59	9.64	4.01	2.94	13.44	3.01	3.23
SCS	6.72	4.56	2.75	10.09	3.53	3.09	13.73	2.73	3.37
MultiWF	6.17	5.12	2.60	10.08	3.75	2.96	13.90	2.78	3.27
MultiSCS	7.15	4.47	2.81	10.62	3.38	3.14	14.42	2.53	3.43

Table 2
Speech quality evaluation in terms of the segmental SNR, SD and PESQ for speech corrupted by Subway noise at various input SNRs.

Algorithm	10 dB			15 dB			20 dB		
	SNR	SD	PESQ	SNR	SD	PESQ	SNR	SD	PESQ
Baseline	2.75	7.26	2.22	7.75	5.18	2.58	12.75	3.58	2.94
WF	4.85	5.28	2.48	9.08	3.86	2.84	13.28	2.81	3.16
SCS	5.13	5.27	2.48	9.12	3.84	2.87	13.42	2.73	3.22
MultiWF	5.22	5.14	2.48	9.44	3.71	2.85	13.67	2.65	3.19
MultiSCS	5.50	5.02	2.53	9.59	3.58	2.91	13.96	2.51	3.26

Table 3
Speech quality evaluation in terms of the segmental SNR, SD and PESQ for speech corrupted by Street noise at various input SNRs.

Algorithm	10 dB			15 dB			20 dB		
	SNR	SD	PESQ	SNR	SD	PESQ	SNR	SD	PESQ
Baseline	3.18	6.04	2.41	8.18	4.27	2.77	13.18	2.92	3.12
WF	5.33	4.56	2.65	9.35	3.36	2.98	13.45	2.44	3.28
SCS	5.26	4.75	2.59	9.21	3.42	2.97	13.58	2.42	3.32
MultiWF	5.73	4.42	2.65	9.72	3.20	3.01	13.83	2.29	3.32
MultiSCS	5.72	4.42	2.67	9.76	3.13	3.04	14.14	2.17	3.38

Table 4
Speech quality evaluation in terms of the segmental SNR, SD and PESQ for speech corrupted by Railway noise at various input SNRs.

Algorithm	10 dB			15 dB			20 dB		
	SNR	SD	PESQ	SNR	SD	PESQ	SNR	SD	PESQ
Baseline	2.57	4.47	2.83	7.57	3.12	3.17	12.57	2.10	3.49
WF	6.75	3.29	3.02	10.74	2.41	3.31	14.70	1.73	3.56
SCS	7.38	3.20	2.95	11.67	2.23	3.31	16.05	1.56	3.62
MultiWF	7.15	3.13	3.03	11.10	2.25	3.34	15.04	1.60	3.62
MultiSCS	8.11	2.87	3.04	12.29	1.99	3.39	16.46	1.37	3.69

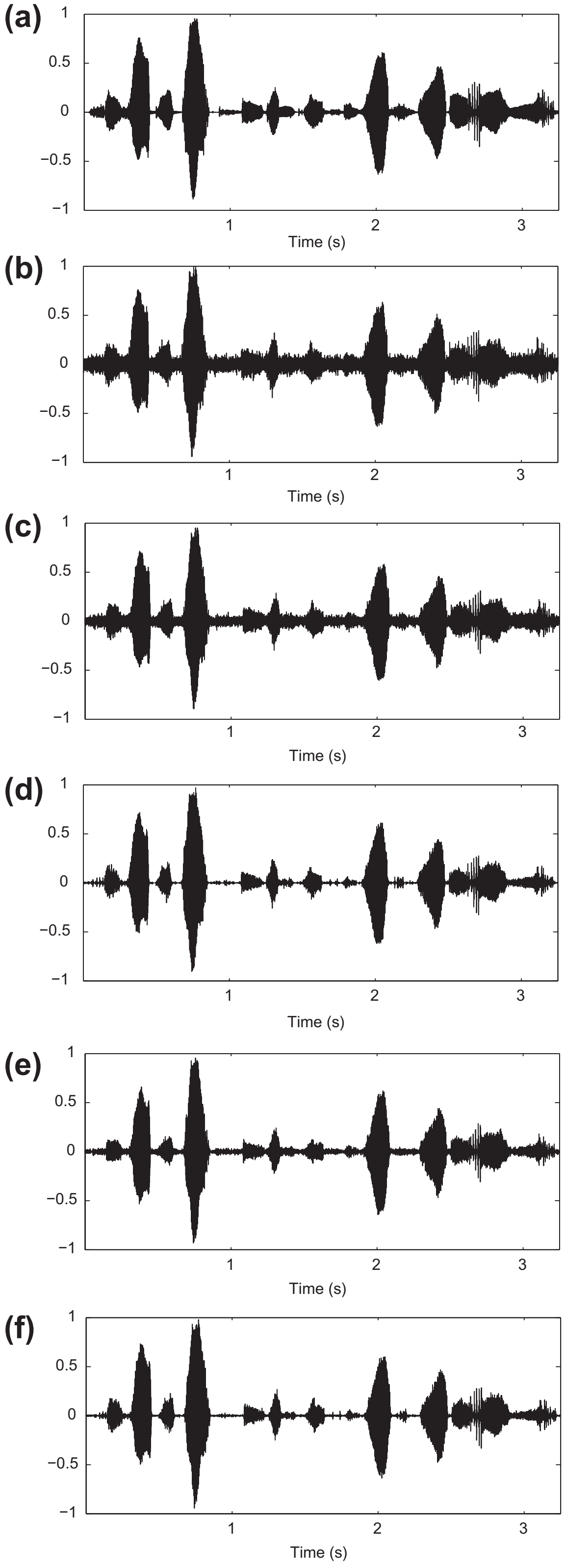

Fig. 6. Waveforms of a speech utterance: clean speech (a), 10 dB Gaussian noisy speech (b), speech enhanced by the conventional WF (c), conventional SCS (d), multi-model WF (e), and multi-model SCS (f).

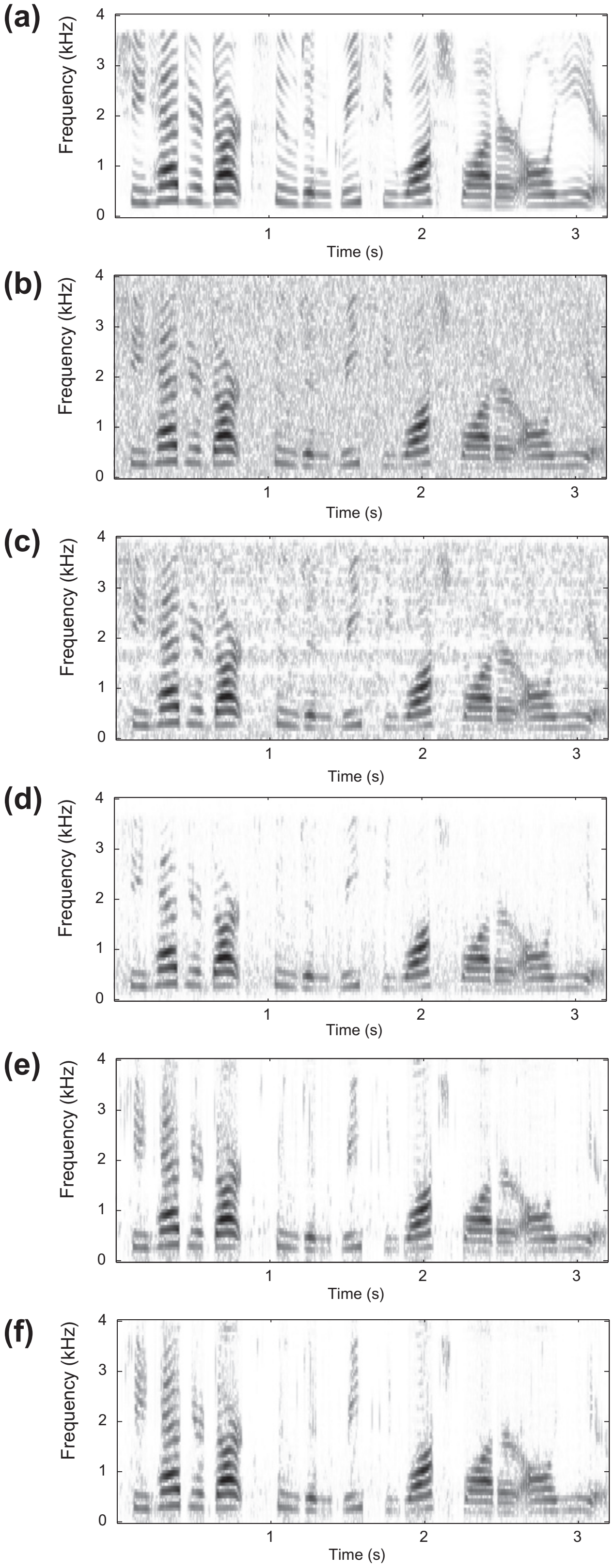

Fig. 7. Spectrograms a speech utterance: clean speech (a), 10 dB Gaussian noisy speech (b), speech enhanced by the conventional WF (c), conventional SCS (d), multi-model WF (e), and multi-model SCS (f).

respectively. In the tables, the ‘Baseline’ denotes the use of the noisy signal itself. Let us first compare the conventional WF and SCS algorithms. It can be seen that the SCS performed significantly better than WF in Gaussian noisy conditions, which demonstrates that the use of non-Gaussian model of speech as in the case of SCS provides improved performance. In other more non-stationary noisy conditions, the SCS obtained better results in most of the Subway and Railway noise conditions, while the WF worked better in particular for Street noise at low SNRs. Let us now discuss the performance of the proposed MultiSCS algorithm. It can be seen that the MultiSCS algorithm obtained consistent and significant performance improvements over both the conventional WF and SCS algorithms in all noisy conditions and all three measures. It also substantially outperformed the MultiWF algorithm, with some very large improvements in Gaussian and Railway noise and moderate improvements in Subway and Street noise conditions. The relative performance improvements (averaged over all noisy conditions), with reference to the baseline, achieved by the proposed MultiSCS algorithm over the conventional SCS algorithm is 37.5%, 22.7%, and 30.9% and over the MultiWF algorithm is 23.2%, 16.0%, and 27.0%, in terms of the segmental SNR, SD, and PESQ, respectively.

An example of spectrograms of the noisy speech and enhanced speech using various algorithms is depicted in Figs. 6 and 7, respectively. It can be seen that the employment of multiple models, i.e., MultiWF and MultiSCS algorithms, can provide reduced level of noise in comparison to the conventional single-model WF and SCS algorithms. The signal enhanced by the MultiSCS algorithm retains more faithfully the structure of the speech signal (i.e., clearer harmonics) than the conventional SCS algorithm and has lower amount of the residual noise than the MultiWF algorithm.

5. Conclusion

This paper presented a single-channel speech enhancement scheme based on Sparse Code Shrinkage (SCS) and the employment of multiple speech models. The motivation for the use of multiple speech models was to better account for the variety of the speech sounds, and we have shown on an example that different speech categories resulted in significantly different non-Gaussian speech models. The presented algorithm employed Gaussian mixture modelling (GMM) for the clustering of speech signals into several categories, in which each component of the trained GMM was considered as one speech category. For each speech category, a non-Gaussian speech model was estimated based on the signal frames assigned to the category. Clustering and individual model estimation was performed on the training data, which were not seen during the enhancement process. In the enhancement stage, the estimate of each signal frame was obtained as a weighted average of a set of estimates provided by using each speech category model. The weights were calculated according to the probability of the primary enhanced signal belonging to each category. We have demonstrated that the use of the logarithm to compress the range of the likelihood values can provide improved performance. During the enhancement, the individual speech category models were further adapted for each signal frame based on the primary enhanced signal. Experimental evaluations were performed on speech signals from the TIMIT database, corrupted by Gaussian noise and three real-world noises, Subway, Street, and Railway noise, from the NOISEX-92 database. The performance was evaluated in terms of the segmental SNR, spectral distortion and PESQ measures. The proposed multi-model SCS algorithm was compared to the conventional Wiener filtering (WF) and SCS algorithms employing a single speech model and also to the WF algorithm employing multiple speech models. It was demonstrated that the proposed multi-model SCS algorithm consistently substantially outperformed all the other algorithms in all noisy conditions.

Acknowledgement

This work was supported by UK EPSRC Grant EP/F036132/1.

References

Boll, S., 1979. Suppression of acoustic noise in speech using spectral subtraction. IEEE Trans. Acoust. Speech Signal Process. 27 (2), 113–120.

Choi, C., Choi, S., Kim, S., 2001. Speech enhancement using sparse code shrinkage and global soft decision. In: Proc. Int. Conf. on Independent Component Analysis and Blind Signal Separation.

Ephraim, Y., 1992a. A Bayesian estimation approach for speech enhancement using hidden Markov models. IEEE Trans. Signal Process. 40 (4), 725–735.

Ephraim, Y., 1992b. Statistical-model-based speech enhancement systems. Proc. IEEE 80 (10), 1524–1555.

Ephraim, Y., Malah, D., 1984. Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator. IEEE Trans. Acoust. Speech Signal Process. 32 (6), 1109–1121.

Ephraim, Y., Malah, D., 1985. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. IEEE Trans. Acoust. Speech Signal Process. 33 (2), 251–266.

Gazor, S., Zhang, W., 2005. Speech enhancement employing Laplacian-Gaussian mixture. IEEE Trans. Speech Audio Process. 13 (5), 896–904.

Hu, Y., Loizou, P., 2003. A perceptually motivated approach for speech enhancement. IEEE Trans. Speech Audio Process. 11 (5), 457–465.

Hyvarinen, A., 1999. Sparse code shrinkage: denoising of nonGaussian data by maximum likelihood estimation. Neural Comput. 11 (7), 1739–1768.

Hyvarinen, A., Karhunen, J., Oja, E., 2001. Independent Component Analysis. John Wiley & Sons, Inc.

Jančovič, P., Köküer, M., 2007. Estimation of voicing-character of speech spectra based on spectral shape. IEEE Signal Process. Lett. 14 (1), 66–69.

Jančovič, P., Zou, X., Köküer, M., 2011. Speech enhancement and representation employing the Independent Component Analysis. In: Ramirez, J., Gorriz, J. (Eds.), Recent Advances in Robust Speech Recognition Technology. Bentham Science Publishers, pp. 103–113.

Kundu, A., Chatterjee, S., Murthy, A., Sreenivas, T., 2008. GMM based Bayesian approach to speech enhancement in signal/transform domain. ICASSP, Las Vegas, USA, pp. 4893–4896.

Lee, J.-H., Jung, H.-Y., Lee, T.-W., Lee, S.-Y., 2000. Speech enhancement with MAP estimation and ICA-based speech features. Electronic Lett. 36, 1506–1507.

Lim, J., Oppenheim, A., 1979. Enhancement and bandwidth compression of noisy speech. Proc. IEEE 67, 1586–1604.

Lotter, T., Vary, P., 2005. Speech enhancement by MAP spectral amplitude estimation using a super-Gaussian speech model. EURASIP J. Appl. Signal Process., 1110–1126.

Martin, R., 2002. Speech enhancement using MMSE short time spectral estimation with Gamma distributed speech priors. ICASSP, Orlando, Florida 1, pp. 253–256.

McAulay, R., Malpass, K., 1980. Speech enhancement using a minimum mean-square error short-time spectral amplitude enhancement. IEEE Trans. Acoust. Speech Signal Process. 28 (2), 137–145.

Ming, J., Srinivasan, R., Crookes, D., 2011. A corpus-based approach to speech enhancement from nonstationary noise. IEEE Trans. Audio Speech Language Process. 19 (4), 822–836.

Potamitis, I., Fakotakis, N., Kokkinakis, G., 2001. Speech enhancement using the sparse code shrinkage technique. ICASSP, Salt Lake City, Utah, pp. 621–624.

Sameti, H., Sheikhzadeh, H., Deng, L., 1998. HMM-based strategies for enhancement of speech signals embedded in nonstationary noise. IEEE Trans. Speech Audio Process. 6 (5), 445–455.

Srinivasan, S., Samuelsson, J., Kleijn, W., 2006. Codebook driven short-term predictor parameter estimation for speech enhancement. IEEE Trans. Audio Speech Language Process. 14 (1), 163–176.

Vaseghi, S., 2005. Advanced Digital Signal Processing and Noise Reduction. John Wiley & Sons.

Virag, N., 1999. Single channel speech enhancement based on masking properties of the human auditory system. IEEE Trans. Speech Audio Process. 7 (2), 126–137.

Wolfe, P., Godsill, S., 2003. Efficient alternatives to the Ephraim and Malah suppression rule for audio signal enhancement. EURASIP J. Appl. Signal Process. 10, 1043–1051.

You, C., Koh, S., Rahardja, S., 2006. Masking-based β-order MMSE speech enhancement. Speech Commun. 48, 50–70.

Young, S., Kershaw, D., Odell, J., Ollason, D., Valtchev, V., Woodland, P., 2000. The HTK Book V3.1.

Zhao, D., Kleijn, W., 2007. HMM-based gain modeling for enhancement of speech in noise. IEEE Trans Audio Speech Language Process. 15 (3), 882–892.

Zou, X., Jančovič, P., Liu, J., Köküer, M., 2008a. Speech signal enhancement based on MAP algorithm in the ICA space. IEEE Trans. Signal Process. 56 (5), 1812–1820.

Zou, X., Jančovič, P., Köküer, M., Russell, M., 2008b. ICA-based MAP speech enhancement with multiple variable speech distribution models. Interspeech, Brisbane, Australia, pp. 415–418.