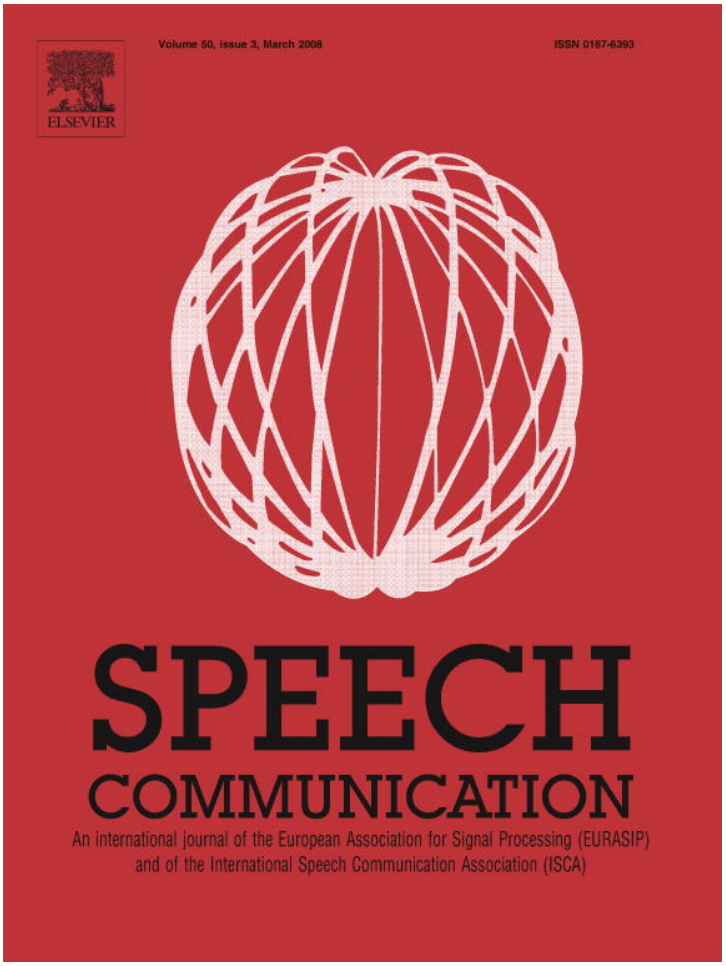

Available online at www.sciencedirect.com
ScienceDirect
Speech Communication 50 (2008) 228–243
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

Multisensory processing for speech enhancement and magnitude-normalized spectra for speech modeling ☆

Amarnag Subramanya [a,*], Zhengyou Zhang [b], Zicheng Liu [b], Alex Acero [b]

[a] *Signal, Speech and Language Interpretation (SSLI) Lab, University of Washington, Seattle, WA 98195, United States*
[b] *Microsoft Research, One Microsoft Way, Redmond, WA 98052, United States*

Received 2 February 2006; received in revised form 24 July 2007; accepted 10 September 2007

Abstract

In this paper, we tackle the problem of speech enhancement from two fronts: speech modeling and multisensory input. We present a new speech model based on statistics of magnitude-normalized complex spectra of speech signals. By performing magnitude normalization, we are able to get rid of huge intra- and inter-speaker variation in speech energy and to build a better speech model with a smaller number of Gaussian components. To deal with real-world problems with multiple noise sources, we propose to use multiple heterogeneous sensors, and in particular, we have developed microphone headsets that combine a conventional air microphone and a bone sensor. The bone sensor makes direct contact with the speaker's temple (area behind the ear), and captures the vibrations of the bones and skin during the process of vocalization. The signals captured by the bone microphone, though distorted, contain useful audio information, especially in the low frequency range, and more importantly, they are very robust to external noise sources (stationary or not). By fusing the bone channel signals with the air microphone signals, much improved speech signals have been obtained.

Keywords: Speech enhancement; Speech modeling; Multi-sensory processing; Dynamic Bayesian

1. Introduction

Speech signals captured by microphones are corrupted by various noise sources. Speech enhancement, i.e., improving the quality of degraded speech, has many applications such as speech communications and man–machine interaction. Despite more than three decades of research, speech enhancement algorithms are not robust to different operating conditions. The problems related to speech enhancement may be grouped under two broad categories, namely, noise modeling and speech modeling. While algorithms have been proposed to rid a speech signal of stationary noise sources, non-stationary noise still poses a challenge. The other dimension to the problem lies in the difficulty of learning an accurate model of human speech due to the inherent non-stationarity of speech, huge intra- and inter-speaker variability, often unpredictable environmental conditions (for example reverberation effects), and sometimes arbitrary microphone gain setting. An accurate speech enhancement technique requires explicit and accurate statistical models for the speech signal and noise processes.

Based on the domain in which removal is done, speech enhancement algorithms may be classified under two broad categories, namely, time-domain based algorithms and spectral-domain based methods. For the former category, examples include (Paliwal and Basu, 1987; Gannot et al., 1998; Lee et al., 1995). For the latter, Quatieri (2002) provides a nice description of various algorithms including spectral subtraction, Wiener filtering, model-based processing and auditory masking. The spectral domain algorithms are usually more attractive from the computational standpoint. Attempts have also been made to model state changes over time in the frequency domain. Lim and

☆ This work was done at Microsoft Research.
* Corresponding author. Tel.: +1 206 221 5216.
E-mail address: asubram@ee.washington.edu (A. Subramanya).

0167-6393/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2007.09.002

Oppenheim (1979) model the short-term speech and noise signals as an autoregressive process. Ephraim (1992) model the long-term speech and noise signals as a hidden Markov process. While autoregressive and hidden Markov models have proved extremely useful in coding and recognition, they were not found to be sufficiently refined for speech enhancement (Ephraim et al., 2005).

Based on the number of channels input, speech enhancement techniques may be classified as single channel and multi-channel algorithms. A majority of the algorithms discussed in the previous paragraph are single channel methods. Multi-channel algorithms make use of information from more than one sensor for speech enhancement (Meyer and Simmer, 1997; Lotter et al., 2003; Nandkumar and Hansen, 1995). However, the sensors are all usually of a similar type (e.g., an array of air microphones). (Jeannes et al., 2001) has a nice survey of multi-channel speech enhancement algorithms. Prominent multi-channel techniques include adaptive noise cancellation (ANC) and beamforming. ANC is based on the availability of an auxiliary channel known as the reference where a sample of the contaminating noise is assumed to be present. In practice though, it is very difficult to find a speech-free noise reference. In the case of beamforming, the gain of a microphone in an array is adjusted based on the direction of the noise source.

In general, while multi-channel methods tend to outperform single-channel algorithms, their performance in difficult environments (for example, non-stationary noise) still leaves a lot to be desired. One of the down sides of multi-channel methods is that, while there are multiple sensors at play, all the sensors are of a similar 'type', and they all capture various amounts of signal and noise simultaneously. One solution to this problem is to use different types of sensors, ideally where the sensors contain complimentary information – we refer to this as *multisensory* processing. It is important to highlight the fact that, in multisensory processing, the different sensors have different properties and thus simple modification of multi-channel techniques will not yield desired results. One area in the speech community where multisensory processing has received a lot of attention is speech recognition, i.e., audio–visual speech processing. It involves the use of a regular microphone along with a camera that captures images of the speakers mouth/face region. The visual stream is used to disambiguate between phones that are easily confused when using only the audio stream (Potamianos et al., 2004). Further, the video stream also lends noise robustness to the speech recognition engine as the visual data is immune to noise effecting the audio stream.

In the speech enhancement community though, interest in multisensory methods has been to some extent limited. Some examples of previous work in this area include, Graciarena et al. (2003), where they combined air and throat microphones for noise robustness in speech recognition. They trained a mapping from the concatenated features of both microphone signals in a noisy environment to clean speech. One down side of their approach, is that the mapping is environment dependent. This can lead to generalization problems in unseen environments. Further, their model does not produce an enhanced waveform but rather only enhances features for speech recognition. Strand et al. (2003) designed an ear plug to capture the vibrations in the ear canal, and used the signals for speech recognition with MLLR adaptation. Heracleous et al. (2003) used a stethoscope device to capture the bone vibrations of the head and used that for non-audible murmur recognition. Like in Strand et al. (2003), they only used the bone signals for adapting the recognizer.

Another work that makes use of a bone sensor for speech enhancement is that of Aliph's Jawbone headsets (http://www.jawbone.com/). This device also uses an air microphone and a bone sensor. The bone sensor is only used as a voice activity detector. When the speaker is not speaking, the air microphone signals are used to build a noise model, and when the speaker is speaking, this adaptively built noise model is used to subtract noise from the air microphone signals. In our work, the bone sensor is not only used as a voice activity detector to adaptively build a noise model but also used for speech enhancement through multisensory processing.

The broad goals of this work are, firstly, to come up with a solution that makes use of multiple sensor streams to combat highly non-stationary noise, and secondly, to develop algorithms that can take advantage of additional sensor streams to reliably estimate the clean signal. We hope to replicate the 'clean signal' as closely as possible, so as to improve the overall user experience of speech communication in highly non-stationary noisy environments.

In this paper, we attack the problem of speech enhancement in non-stationary environments on several fronts: (a) we propose to integrate alternative sensors, in particular the bone-conductive sensor, with standard air-conductive microphone to deal with highly non-stationary noise, (b) we propose a new speech model based on magnitude-normalized spectra, which alleviates the issues related to intra- and inter-speaker variations and (c) we propose algorithms/models that can take advantage of the above in order to produce good quality speech in noisy environments.

In Section 2, we present details about the new multi-sensory microphone that is robust to noise. We also discuss some of our previous work using this multi-sensory microphone and highlight its shortcomings. In Section 3 we discuss the proposed magnitude normalized speech model. Next, in Section 4 we discuss speech enhancement using the multisensory headset and the proposed speech model. Details of the experimental setup are in Section 5. Finally, Sections 6 and 7 discuss the results, conclusions and future work.

2. Air-and-bone conductive microphone

We developed an air-and-bone conductive (ABC) microphone that makes use of a bone conduction microphone in

addition to the regular air microphone (Zhang et al., 2004). Fig. 1 shows two prototypes of the ABC microphone. In the case of the first prototype, when the user wears the device, the bone sensor resides on his/her temple and the air channel is a close-talking microphone. In the case of the second prototype, the bone sensor is positioned behind the ear, while the air channel is a medium-sized boom of 45 mm. Either case, the bone conduction microphone captures the vibrations caused in the speakers' bones and skin during the process of vocalization.

Fig. 2 shows the time and frequency domain renditions of a speech signal captured using the ABC microphone in a relatively noise-free office environment. As it can be seen the bone sensor only captures frequency components up to 3 KHz. Fig. 3 shows the frequency response of the air and bone channels averaged over all the speech frames in the utterance used in Fig. 2. It can be seen that at low frequencies (<700 Hz) the bone sensor follows the air sensor closely, but tapers down for higher frequencies. Fig. 4 shows a comparison of the frequency response of the air and bone channels in a noisy environment. The utterance used to generate this figure was recorded in a cafeteria with ambient noise level 85 dBc. The first plot shows the average response in the air channel for speech and non-speech (noisy) frames. It can be seen that except in the low frequencies, the SNR for frequencies >2 KHz is about 0 dB. The second plot shows the response for signal captured

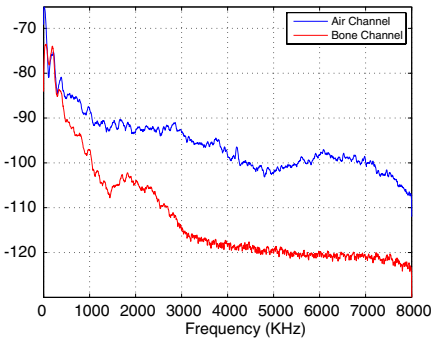

Fig. 3. Frequency Response of the two channels for the utterance "This advice sounds questionable." spoken by a female speaker. The frequency response was computed by averaging the spectra over all the speech frames in the above utterance. The x-axis represents frequency on a linear scale and the y-axis represents magnitude on a log scale.

by the bone microphone. Here it can be seen that the speech is on an average 20 dB above the noise. This illustrates the noise robustness of the bone sensor.

More information on the ABC microphone and other alternative sensors to capture speech may be obtained in (Zhang et al., 2004). In essence, with the use of the additional sensor, we have one stream with undistorted but noise corrupted speech (air channel) and another stream with distorted but fairly uncorrupted speech (bone channel). This truly exemplifies the complementary relationship

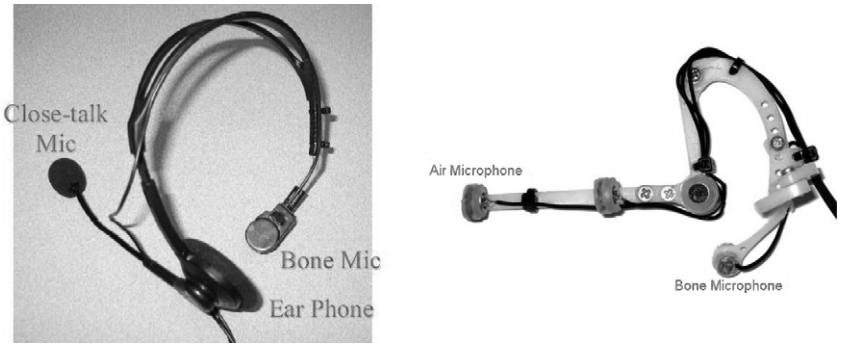

Fig. 1. Two prototypes of air-and-bone conductive (ABC) microphone headsets.

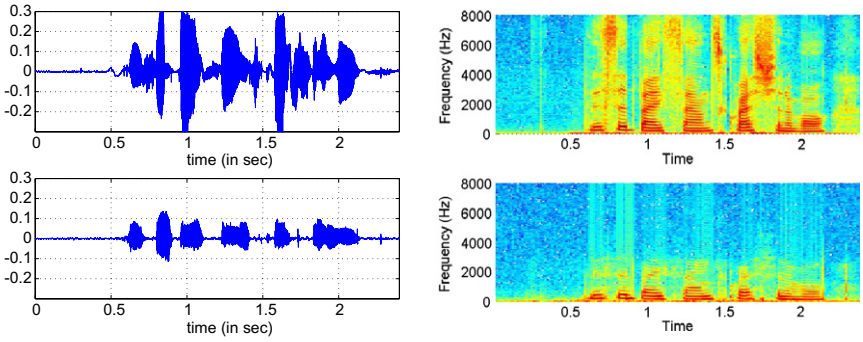

Fig. 2. Time and frequency domain renditions of the signals captured by the ABC microphone for a female speaker saying "This advice sounds questionable.". The first row shows the signal captured by the air microphone and the second row shows the signal captured by the bone microphone. Sampling rate is 16 KHz.

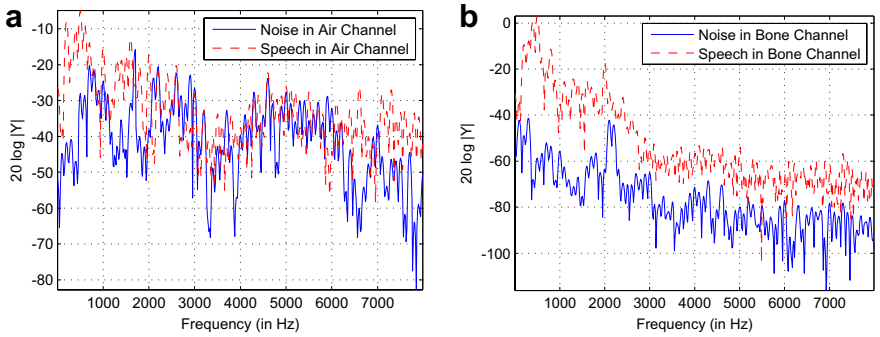

Fig. 4. Frequency response for speech and noise in the two channels (manually segmented) for the utterance "This advice sounds questionable" spoken by a female speaker. The frequency responses were computed by averaging the spectra over the speech and non-speech frames respectively in each channel. The *x*-axis represents frequency on a linear scale and the *y*-axis represents magnitude on a log scale. The response due to speech is represented using dashed lines and the response due to noise is represented using solid lines.

that exists in most multisensory problems. The challenge here is to intelligently fuse the signals from the two sensors to obtain an estimate of the 'clean' signal.

As one of the end goals of this work is to improve the overall user experience in noisy environments, it is thus imperative to reproduce both the magnitude and phase of the 'cleaned' signal. Furthermore, we are in pursuit of an algorithm that is computationally inexpensive and can be implemented using a low power digital signal processor (DSP). The above two goals constrain the possible candidates for processing domains. The time domain is ruled out because of the computational complexity. The Mel-cepstral domain, though attractive from the perceptual point of view, does not take into account the phase information. The same holds for the real cepstral domain. As a result, we are left with the spectral domain or its variants. We choose not to work in the log-spectral domain because (a) the use of the logarithm leads to non-linearities in the processing algorithms, and (b) modeling phase in the log-spectral domain is not entirely straightforward. In consequence, we choose the complex spectral domain to process our signals.

In the remainder of this section we discuss some of our previous work in this area and also highlight some of its short-comings. In (Zheng et al., 2003), we proposed an algorithm based on the SPLICE technique to learn the mapping between the two streams and the clean speech signal. One drawback of this approach is that it requires prior training and therefore can lead to generalization difficulties. In the same work, we also proposed a speech detector based on a histogram of the energy in the bone channel. In (Liu et al., 2004), we proposed an algorithm called direct filtering (DF) that does not require any prior training in order to estimate the clean speech signal. The transfer function from the close talking channel to the bone-channel is learned from the given utterance and the clean signal is estimated in a maximum likelihood framework. It was also shown that the performance of the DF algorithm is better compared with the algorithm proposed in (Zheng et al., 2003) for speech enhancement. However, one drawback with the DF algorithm is the absence of a speech model, which can lead to distortion in the enhanced signal. In (Liu et al., 2005), we extended the DF algorithm to deal with the environmental noise leakage into the bone sensor, and the teethclack problem that is caused when the users' upper and lower jaws come in contact with each other during the process of articulation. A common factor amongst all the above approaches (and for that matter any real-time speech enhancement system) is the requirement of accurate speech/voice activity detection. The technique proposed in (Zheng et al., 2003), which makes use of a function of the energy in the bone sensor, although robust to noise, has a number of problems, such as (a) some classes of phones (e.g., fricatives) have low energy in the bone sensor, thus producing false negatives; and (b) leakage of ambient noise in the bone sensor can lead to false positives. Furthermore, by using just the bone sensor for speech detection, we are not leveraging the two channels of information provided by the multisensory headset. In (Subramanya et al., 2005), we proposed an algorithm that takes into account the correlation between the two channels for speech detection and also incorporates a speech model within a graphical model (GM) framework thereby reducing the amount of distortion in the enhanced signal. However, this work made use of a simplistic speech model (single Gaussian), and thus led to distortion in the enhanced signal whenever there was a mismatch between the training and test conditions. Further, there were also issues with gain normalization as the model was built in the complex spectral domain. Another drawback of (Subramanya et al., 2005) was the inherent lack of temporal smoothness. The model had no temporal constraints and thus exhibited sudden changes in state, leading to distortions in the output signal. In this work we extend the above algorithm with a new magnitude-normalized speech model and investigate the benefit of using multiple Gaussian components. Further we also propose a dynamic Bayesian network (DBN) that uses the above speech model and takes into account temporal constraints to produce a relatively distortionless enhanced speech signal.

3. Magnitude-normalized complex spectrum-based speech model

In Bayesian statistics, prior information on hidden variables plays a crucial role in inference. If z represents the hidden variable that we wish to estimate given some observation o, using Bayes rule we have $p(z|o) = kp(o|z)p(z)$, where k is a constant independent of z. In the above, $p(z)$ is the prior information about z, and represents the knowledge about z that is known even before o is observed. In the case of speech enhancement (and in general speech processing), a speech model lends itself into such a role by providing a prior on clean speech that is hidden given noisy speech. However, modeling human speech is an extremely complex problem owing to its large variability. Some of the factors contributing to this variability include: differing spectral profiles for different speakers; changes in loudness, intonation and stress even for a single speaker due to context, emotive state, environmental conditions, etc.; variations due to gender, social background, etc.

In general, statistical models of speech have relied on using mixtures of densities to accurately model speech (Chen et al., 2002). However, even a mixture model is not suitable to deal with issues related to changes in loudness and recording device gains (as we would need infinitely many mixtures to model all possible scenarios). An approach that has been used to deal with these issues is to model speech in the cepstral domain, where such changes are reflected in the first cepstral coefficient that is then neglected for modeling purpose (Wu et al., 2003). This approach has been popular when speech enhancement is done to improve speech recognition performance. However (as discussed in the last section) working in the cepstral domain has its own disadvantages, which include issues with non-linearity and absence of phase in the reconstructed signal. In the complex spectral domain, however, it is possible to recover both the magnitude and phase, but we need some form of *gain normalization*. Gain normalization has been studied in the past and can be grouped under two categories, namely, speech gain normalization (Yoshizawa et al., 2004) and noise gain normalization (Zhao and Kleijn, 2005). One important distinction between the work in (Zhao and Kleijn, 2005) and the current algorithm is that here we are normalizing the clean speech signal rather than the noise. In the case of speech gain normalization, majority of the work has been on gain normalization in the cepstral domain (Yoshizawa et al., 2004)[1]. In our case we work in the complex spectral domain, where, the normalization problem to some extent is more challenging than the cepstral domain.

We briefly digress to explain the notation used in this paper. Lower case alphabets are used to represent signals in the time-domain; $x(k)$ represents the clean speech signal that we wish to estimate, k is the time index, $y(k)$ and $b(k)$ represent the signals captured by the air and bone sensors respectively. The signals $y(k)$ and $b(k)$ are transformed using the short-time Fourier transform (STFT) using appropriate windows yielding Y_t and B_t respectively, where t is the frame index. Note that both Y_t and B_t are vectors. Let X_t be a frame of the clean speech signal that we wish to estimate given Y_t and B_t. As we are dealing with real signals, the Fourier transform is symmetric about the zero frequency axis and thus we need to consider only one-half of the spectrum for processing. Thus if $2(\mathbf{N}-1)$ is the length of the Fourier transform (FFT), then Y_t, B_t (and thus X_t) are vectors of length $\mathbf{N}$ (including the dc and nyquist terms). We represent this as $Y_t = [Y_t^l, \ldots, Y_t^f, \ldots, Y_t^N]^{\mathrm{T}}$. Henceforth we use Y_t^f to refer to a particular component of Y_t. In other words, Y_t^f represents the value of the fth frequency component of the tth frame of $y(k)$. Note that B_t and B_t^f, X_t and X_t^f are defined in a similar fashion. Given X_t, $x(k)$ may be reconstructed using an inverse Fourier transform and the overlap-and-add procedure. If Z_t is a random variable that follows a Gaussian (normal) distribution, this is denoted using $p(Z_t) \sim N(Z_t; \mu, \Sigma)$ or $N(\mu, \Sigma)$, where μ and Σ are the mean and covariance of Z_t. Also $\sim$ is used to denote "distributed as". Unless otherwise stated, we use 16 KHz as the default sampling rate for all our experiments in this work. Other notation will be defined as required.

3.1. Model definition

We propose the use of magnitude-normalized complex spectra as features for the speech model. In order to build such a speech model, the frames of the speech signal are normalized with their energy, i.e.,

$$\widetilde{X}_t = \frac{X_t}{\|X_t\|}. \tag{1}$$

Thus all $\widetilde{X}_t$'s are unit vectors and distributed on a unit hyper-sphere. It can be easily seen that the above step has a variance reducing effect because instead of attempting to capture the variations in an n-dimensional space, we are modeling a region on a unit hyper-sphere. However, as a result of the above normalization, the model now requires a gain term g_{x_t} to match the model with the observation. We will describe an iterative approach to estimating the gain in Section 4.5.

3.2. Training

In order to train the speech model, we collected data from a number of speakers in a clean environment. We then made use of energy-based speech detector to extract all the speech frames in the above utterances. Let us denote these frames as $\{X_t\}_{t=1}^T$ (assuming T frames of speech in the training set). The resulting speech frames were then energy normalized as explained in the previous section to yield $\{\widetilde{X}_t\}_{t=1}^T$. We then ran the k-means algorithm (with random

[1] They use all the cepstral coefficients for modeling speech, i.e., the first coefficient is not neglected.

initialization) on the above energy normalized speech frames using the following distance metric:

$$d(\widetilde{X}_i, \widetilde{X}_j) = \left\| [d(\widetilde{X}_i^1, \widetilde{X}_j^1), \ldots, d(\widetilde{X}_i^f, \widetilde{X}_j^f), \ldots, d(\widetilde{X}_i^{\mathbf{N}}, \widetilde{X}_j^{\mathbf{N}})]^T \right\|, \quad 1 \leqslant i, j \leqslant T,$$
$$d(\widetilde{X}_i^f, \widetilde{X}_j^f) = \log|\widetilde{X}_i^f| - \log|\widetilde{X}_j^f|, \quad 1 \leqslant f \leqslant \mathbf{N} \tag{2}$$

to yield M clusters. We use a mixture of Gaussians to model the normalized speech. The means and variances of these Gaussians are set to the mean and variance of the clusters obtained above. The responsibility of each Gaussian in the mixtures, $\alpha_i (0 \leqslant i \leqslant M-1)$, is set to $N(i)/T$, where $N(i)$ is the number of elements in the ith cluster. Thus we have that $\sum_{i=1}^{M} \alpha_i = 1$. We discuss more about this model in Section 4.3.

4. A dynamic Bayesian network for speech enhancement in a multisensory headset

A dynamic Bayesian network (DBN) represents a family of probability distributions defined in terms of a directed graph. The nodes in the graph represent random variables, and the joint probability distribution over the variables is obtained by taking products over functions on connected subsets of nodes. By exploiting graph-theoretic representations, DBNs provide general algorithms for computing marginal and conditional probabilities of interest (Jordan and Weiss, 2002; Bilmes, 2000; Zweig et al., 2002; Bilmes, 2001). The popular Hidden Markov Models (HMMs) maybe represented as a DBN. A characteristic of DBNs that distinguishes them from other classes of graphical models is that some (or sometimes all) of the edges in the graph point in the direction of increasing time. DBNs have been applied to many tasks in the past including, speech recognition (Zweig et al., 2002), vision applications such as tracking (Beal et al., 2003), natural language processing (NLP) applications such as parsing, tagging (Klein and Manning, 2002). In this paper we propose to use a DBN for speech enhancement using a multi-sensory microphone. See (Jordan and Weiss, 2002) for more information on DBNs, their usage and inference.

4.1. Network description

Fig. 5 shows two frames of a DBN used to model the enhancement process in the complex spectral domain. Here, all observed variables are shaded. In this model,

- S_t is a discrete random variable representing the state (speech/non-speech) of the frame,
- M_t is a discrete random variable acting as an index into the mixture of distributions modeling speech/non-speech,
- $\widetilde{X}_t$ represents magnitude-normalized speech,
- g_{x_t} scales $\widetilde{X}_t$ to match the clean speech X_t,
- V_t is the background noise,

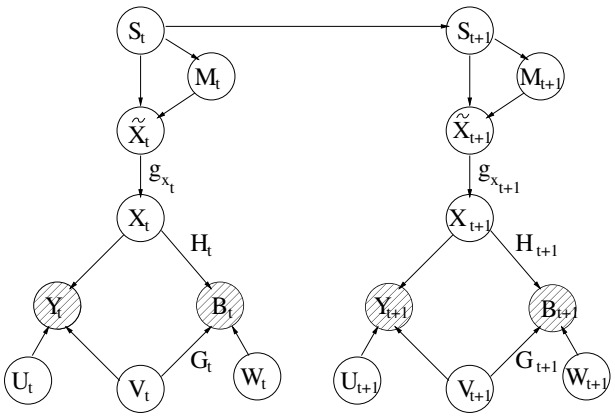

Fig. 5. A dynamic bayesian network (DBN) of the proposed enhancement framework.

- U_t and W_t represent the sensor noise in the air and bone channels respectively,
- H_t is the optimal linear mapping from X_t to B_t,
- G_t models the leakage of background noise into the bone sensor.

Recall that the variables X_t, Y_t, B_t have already been defined in Section 3. Note that the subscript t in the above variables refers to the value of that particular random variable for the frame at time t. Furthermore, the variables $\widetilde{X}_t$, X_t, Y_t, V_t, H_t, G_t and B_t are all vectors of dimension $\mathbf{N}$. The physical process that the DBN attempts to model may be expressed as

$$Y_t = X_t + V_t + U_t \quad \text{and} \quad B_t = H_t X_t + G_t V_t + W_t. \tag{3}$$

In other words, we have that the signal captured at the air microphone (Y_t) is a sum of the clean speech signal (X_t), the corrupting noise (V_t) and the sensor noise in the air microphone (U_t). The signal captured by the bone microphone (B_t) is a sum of the transformed version of the clean speech signal ($H_t X_t$), the amount of background noise that leaks into the bone sensor ($G_t B_t$) and the sensor noise in the bone microphone (W_t). The reason that $G_t V_t$ is used in the second equation is inherent to the operation of the ABC microphone, in that, we expect that not all the ambient noise that effects the air microphone is captured by the bone microphone, but a much attenuated and distorted version is seen at the bone sensor. For an ideal bone sensor which is perfectly immune to ambient noise, $G_t = 0$. Furthermore, the model also recognizes the fact that the clean speech signal X_t undergoes some form of distortion while being captured by the bone channel, and $H_t X_t$ is used to represent this phenomenon.

We make the following assumptions with regards to the model:

(1) For mathematical tractability we assume that the frequency components of $\widetilde{X}_t$, X_t, Y_t, V_t, H_t, G_t, B_t are independent. It is important to highlight the fact that we are not making any assumptions on the independence of the variables, but rather the independence of the individual frequency components, i.e.

$\widetilde{X}_t^f$ and $\widetilde{X}_t^{f+i}$ are independent $\forall f, i : -\mathbf{N} \leqslant i \leqslant \mathbf{N}$, $i \neq 0,\ 1 \leqslant f + i \leqslant \mathbf{N}$. (4)

(2) The variables S_t and M_t are scalars and are considered global over the frame at time t, i.e., the value of these variables are the same for all the frequency components for a given frame.
(3) Background noise is modeled using a zero mean Gaussian, i.e., $p(V_t) \sim N(0, \sigma_v^2)$.
(4) Sensor (device) noise in the air microphone channel is modeled with $p(U_t) \sim N(0, \sigma_u^2)$.
(5) Sensor noise in the bone channel is modeled with $p(W_t) \sim N(0, \sigma_w^2)$.
(6) For mathematical tractability we assume that $p(\widetilde{X}_t|X_t) \sim \delta(X_t, g_{x_t}\widetilde{X}_t)$, where δ is the Kronecker delta function with parameter g_{x_t}. While it is possible to put a prior on g_{x_t} and estimate it in a Bayesian setting, in this paper, we take a frequentist approach to estimating g_{x_t} (see Section 4.5).

Information on how the sensor noise variances are chosen is given in Section 5. Also, except for the variances associated with the sensor noise, all other variances are a function of time t. However to reduce notational clutter, we don't explicitly represent this.

The joint distribution over all the variables in the model factorizes as follows:

$$p(Y_t, B_t, X_t, \widetilde{X}_t, V_t, S_t, M_t, U_t, W_t) = p(Y_t|X_t, V_t, U_t)p(B_t|X_t, V_t, W_t)p(X_t|\widetilde{X}_t)p(\widetilde{X}_t|M_t, S_t) \times p(M_t|S_t)p(S_t)p(V_t)p(U_t)p(W_t). \quad (5)$$

The above factorization follows directly as a result of the semantics of DBNs (Jordan and Weiss, 2002). For now, we ignore the temporal constraint between S_t and S_{t+1} We discuss the temporal smoothness constraints in Section 4.6. As a result of the frequency independence assumption (see Eq. (4)), we can re-write the above equation as

$$p(Y_t^f, B_t^f, X_t^f, \widetilde{X}_t^f, V_t^f, S_t, M_t, U_t^f, W_t^f) = p(Y_t^f|X_t^f, V_t^f, U_t^f)p(B_t^f|X_t^f, V_t^f, W_t^f)p(X_t^f|\widetilde{X}_t^f) \times (\widetilde{X}_t^f|M_t, S_t)p(M_t|S_t)p(S_t)p(V_t^f)p(U_t^f)p(W_t^f), \quad 1 \leqslant f \leqslant N, \quad (6)$$

which is essentially $\mathbf{N}$ equations (one for each f). Recall that all terms in the above equation correspond to individual (frequency) components except for S_t and M_t (S_t, M_t are global over a given frame). In the following, for notational convenience, we drop the superscript f on all the variables, i.e. (for example) Y_t implies Y_t^f unless otherwise stated. Thus the above equation may be re-written as

$$p(Y_t, B_t, X_t, \widetilde{X}_t, V_t, S_t, M_t, U_t, W_t) = p(Y_t|X_t, V_t, U_t)p(B_t|X_t, V_t, W_t)p(X_t|\widetilde{X}_t)p(\widetilde{X}_t|M_t, S_t) \times p(M_t|S_t)p(S_t)p(V_t)p(U_t)p(W_t). \quad (7)$$

Before we discuss inference using the above model, we briefly digress to comment on how we estimate the transfer function H_t and the leakage factor G_t.

4.2. Transfer function and leakage factor

In this work, we adopt a frequentist approach to estimate H_t and G_t (similar to Liu et al., 2004). The transfer function H_t models the relationship between the air and bone channels. In other words, given a frame of speech captured by the air channel at time t, H_t is the transform it would have to undergo to be similar to the frame captured in the bone channel. In our experiments we have found that H_t is phone-dependent, i.e, H_t depends on the class of speech.

However, as a first attempt, we ignore this dependency and estimate a single H_t for all classes of speech.

The physical process of the ABC microphone is represented as in Eq. (3). As a result of frequency independence assumption, the two equations may be decomposed into $2\mathbf{N}$. This results in a set of $2\mathbf{N}$ equations in $3\mathbf{N}$ unknowns (assuming that U_t^f, W_t^f and V_t^f are given). While it might seem that we have a under determined set of equations, consider the silent regions of an utterance, i.e., regions in which speech is absent. For such frames we have $X_t^f = 0,\ \forall f$ and thus we get

$$Y_t^f = U_t^f + V_t^f,\ \ 1 \leqslant f \leqslant \mathbf{N},$$
$$B_t^f = W_t^f + G_t^f V_t^f,\ \ 1 \leqslant f \leqslant \mathbf{N}, \quad (8)$$

where $t \in N_v$ and N_v is the set of non-speech frames in the region $[0, t]$. Making use of the above formulation, it can be shown (Liu et al., 2004) that the ML estimate of G_t is given by

$$G_t = \frac{\sum_{t\in N_v}(\sigma_v^2|B_t|^2 - \sigma_w^2|Y_t|^2) + \sqrt{\left(\sum_{t\in N_v}(\sigma_v^2|B_t|^2 - \sigma_w^2|Y_t|^2)\right)^2 + 4\sigma_v^2\sigma_w^2\left|\sum_{t\in N_v}B_t^*Y_t\right|^2}}{2\sigma_v^2\sum_{t\in N_v}B_t^*Y_t}. \quad (9)$$

Note that B_t^* represents the complex conjugate (Hermitian) of B_t. As a result of our frequency independence assumptions, the above equation represents only a single frequency component and is simply repeated to estimate G_t for the

others. Once we have estimated G_t from the non-speech frames, we can estimate H_t using

$$H_t = G_t + \frac{\sum_{t\in N_s}(\sigma_v^2|B'_t|^2 - \sigma_w^2|Y_t|^2) + \sqrt{\left(\sum_{t\in N_s}(\sigma_v^2|B'_t|^2 - \sigma_w^2|Y_t|^2)\right)^2 + 4\sigma_v^2\sigma_w^2|\sum_{t\in N_s}(B'_t)^* Y_t|^2}}{2\sigma_v^2\sum_{t\in N_s}(B'_t)^* Y_t}, \tag{10}$$

where $B'_t = B_t - G_t Y_t$ and N_s is the set of speech frames in the region $[0, t]$. For more details on estimating the transfer function, the reader is refereed to (Liu et al., 2004). While a naive implementation of the above equations to estimate G_t and H_t would require storing all the relevant speech and non-speech frames, a more efficient technique is to make use of accumulators that are appended when we have new data. In our work we estimate G_t and H_t efficiently using a recursive implementation based on exponential aging (see appendix for more details).

4.3. Posterior of magnitude normalized clean speech – $p(\widetilde{X}_t|Y_t, B_t)$

Recall Eq. (7),

$$\begin{aligned} &p(Y_t, B_t, X_t, \widetilde{X}_t, V_t, S_t, M_t, U_t, W_t) \\ &\quad = p(Y_t|X_t, V_t, U_t)p(B_t|X_t, V_t, W_t)p(X_t|\widetilde{X}_t)p(\widetilde{X}_t|M_t, S_t) \\ &\qquad \times p(M_t|S_t)p(S_t)p(V_t)p(U_t)p(W_t). \end{aligned} \tag{11}$$

We have already defined $p(X_t|\widetilde{X}_t)$, $p(V_t)$, $p(U_t)$, and $p(W_t)$ (see Section 4.1). Further, we have that $p(Y_t|X_t, V_t, U_t) \sim N(Y_t; X_t + V_t, \sigma_u^2)$ and $p(B_t|X_t, V_t, W_t) \sim N(B_t; H_t X_t + G_t V_t, \sigma_w^2)$. Finally, in our system, speech is modeled using a mixture of Gaussians (MG),

$$p(\widetilde{X}_t|S_t) = \sum_{m=0}^{M-1} P(M_t = m|S_t)p(\widetilde{X}_t|S_t, M_t), \tag{12}$$

$$\text{with } p(\widetilde{X}_t|S_t = s, M_t = m) \sim N(\mu_{s,m}, \sigma_{s,m}^2). \tag{13}$$

In our model, $S_t = 0$ indicates the silence (non-speech) state and $S_t = 1$ indicates the speech state. We represent silence using a single Gaussian, and thus $P(M_t = 0|S_t = 0) = 1$ which implies $p(\widetilde{X}_t|S_t = 0) \sim N(\widetilde{X}_t; 0, \sigma_{sil}^2)$. In the case of speech we use a MG with $M = 4$ and $P(M_t = i|S_t = 1) = \alpha_i$ (see Section 3.2).

As X_t and $\widetilde{X}_t$ are related by a delta distribution, *given* g_{x_t}, estimating either one of these variables is equivalent. Thus, integrating out X_t from the joint distribution we have

$$\begin{aligned} &\int_{X_t} p(Y_t, B_t, X_t, \widetilde{X}_t, V_t, S_t, M_t, U_t, W_t)\mathrm{d}X_t \\ &\quad = p(Y_t, B_t, \widetilde{X}_t, V_t, S_t, M_t, U_t, W_t) \\ &\quad = p(Y_t|g_{x_t}\widetilde{X}_t, V_t, U_t)p(B_t|g_{x_t}\widetilde{X}_t, V_t, W_t)p(\widetilde{X}_t|M_t, S_t) \\ &\qquad \times p(M_t|S_t)p(S_t)p(V_t)p(U_t)p(W_t), \end{aligned} \tag{14}$$

where $p(Y_t|g_{x_t}\widetilde{X}_t, V_t, U_t) \sim N(g_{x_t}\widetilde{X}_t + V_t, \sigma_u^2)$ and $p(B_t|g_{x_t}\widetilde{X}_t, V_t, W_t) \sim N(g_{x_t}H\widetilde{X}_t + GV_t, \sigma_w^2)$. Now consider

$$\begin{aligned} p(\widetilde{X}_t|Y_t, B_t) &= \sum_{s,m} p(\widetilde{X}_t, S_t = s, M_t = m|Y_t, B_t) \\ &= \sum_{s,m} p(\widetilde{X}_t|Y_t, B_t, S_t = s, M_t = m) \\ &\quad \times p(M_t = m|Y_t, B_t, S_t = s)p(S_t = s|Y_t, B_t), \end{aligned} \tag{15}$$

where the above equation follows as a result of the recursive application of the chain rule of probability. Let us first consider evaluating $p(\widetilde{X}_t \mid Y_t, B_t, S_t = s, M_t = m)$. Using the definition of conditional probability, we have

$$\begin{aligned} &p(\widetilde{X}_t|Y_t, B_t, S_t = s, M_t = m) \\ &\qquad = \frac{p(\widetilde{X}_t, Y_t, B_t, S_t = s, M_t = m)}{p(Y_t, B_t, S_t = s, M_t = m)}. \end{aligned} \tag{16}$$

But

$$\begin{aligned} &p(\widetilde{X}_t, Y_t, B_t, S_t = s, M_t = m) \\ &\quad = \int_{V_t}\int_{U_t}\int_{W_t} p(Y_t, B_t, \widetilde{X}_t, V_t, S_t, M_t, U_t, W_t)\mathrm{d}U_t\,\mathrm{d}W_t\,\mathrm{d}V_t. \end{aligned} \tag{17}$$

Integrating out V_t, U_t and W_t we get

$$\begin{aligned} &p(\widetilde{X}_t, Y_t, B_t, S_t = s, M_t = m) \\ &\qquad \sim N(\widetilde{X}_t; \alpha_{s,m}, \beta_{s,m})N(B_t; \gamma_{s,m}\eta_{s,m}) \\ &\qquad\quad \times N(Y_t; g_{x_t}\mu_{s,m}, \sigma_1^2)p(M_t, |S_t)p(S_t) \end{aligned} \tag{18}$$

where

$$\begin{aligned} \alpha_{s,m} &\triangleq \frac{\sigma_{s,m}^2(\sigma_1^2(\sigma_{uv}^2\mu_{s,m} + g_{x_t}Y_t) + g_{x_t}H_m^*(B_t\sigma_{uv}^2 - G\sigma_{uv}^2 Y_t))}{\sigma_1^2\sigma_2^2 + g_{x_t}^2\sigma_{s,m}^2\sigma_{uv}^2|H_m|^2}, \\ \beta_{s,m} &\triangleq \frac{\sigma_1^2\sigma_{s,m}^2\sigma_{uv}^2}{\sigma_1^2\sigma_2^2 + g_{x_t}\sigma_{s,m}^2\sigma_{uv}^2|H_m|^2}, \\ \gamma_{s,m} &\triangleq g_{x_t}, H_m\frac{\sigma_{uv}^2\mu_{s,m} + g_{x_t}\sigma_{s,m}^2 Y_t}{\sigma_2^2} + \frac{G\sigma_v^2 Y_t}{\sigma_{uv}^2}, \\ \eta_{s,m} &\triangleq \sigma_1^2 + g_{x_t}|H_m|^2\frac{\sigma_{s,m}^2\sigma_{uv}^2}{\sigma_2^2}, \\ \sigma_{uv}^2 &\triangleq \sigma_u^2 + \sigma_v^2, \\ \sigma_1^2 &\triangleq \sigma_w^2 + \frac{|G|^2\sigma_u^2\sigma_v^2}{\sigma_{uv}^2}, \\ \sigma_2^2 &\triangleq \sigma_{uv}^2 + g_{x_t}^2\sigma_{s,m}^2, \\ H_m &\triangleq H - G\frac{\sigma_v^2}{\sigma_{uv}^2}. \end{aligned} \tag{19}$$

In the above, the parameter g_{x_t} is hidden. We propose an approach to estimating g_{x_t} in Section 4.5. Using Eq. (18) in Eq. (16) we get

$$\begin{aligned}
p(\widetilde{X}_t|Y_t,B_t,S_t=s,M_t=m) &= \frac{p(\widetilde{X}_t,Y_t,B_t,S_t=s,M_t=m)}{p(Y_t,B_t,S_t=s,M_t=m)} \\
&= \frac{p(\widetilde{X}_t,Y_t,B_t,S_t=s,M_t=m)}{\int_{\widetilde{X}_t} p(\widetilde{X}_t,Y_t,B_t,S_t=s,M_t=m)\mathrm{d}\widetilde{X}_t} \\
&\sim \frac{N(\widetilde{X}_t;\alpha_{s,m},\beta_{s,m})N(B_t;\gamma_{s,m},\eta_{s,m})N(Y_t;g_{x_t}\mu_{s,m},\sigma_1^2)p(M_t|S_t)p(S_t)}{\int_{\widetilde{X}_t} N(\widetilde{X}_t;\alpha_{s,m},\beta_{s,m})N(B_t;\gamma_{s,m},\eta_{s,m})N(Y_t;g_{x_t}\mu_{s,m},\sigma_1^2)p(M_t|S_t)p(S_t)\mathrm{d}\widetilde{X}_t} \\
&\sim \frac{N(\widetilde{X}_t;\alpha_{s,m},\beta_{s,m})N(B_t;\gamma_{s,m},\eta_{s,m})N(Y_t;g_{x_t}\mu_{s,m},\sigma_1^2)p(M_t|S_t)p(S_t)}{N(B_t;\gamma_{s,m},\eta_{s,m})N(Y_t;g_{x_t}\mu_{s,m},\sigma_1^2)p(M_t,|S_t)p(S_t)} \\
&\sim N(\widetilde{X}_t;\alpha_{s,m},\beta_{s,m}).
\end{aligned} \tag{20}$$

The above follows as $\int_{\widetilde{X}_t} N(\widetilde{X}_t;\alpha_{s,m},\beta_{s,m})\mathrm{d}\widetilde{X}_t = 1$. Thus we have the posterior

$$p(\widetilde{X}_t|Y_t,B_t,S_t=s,M_t=m) \sim N(\widetilde{X}_t;\alpha_{s,m},\beta_{s,m}), \tag{21}$$

where $\alpha_{s,m}$, $\beta_{s,m}$ are defined in Eq. (19). Note that in order to compute $p(\widetilde{X}_t|Y_t,B_t,S_t=0,M_t=0)$ we follow the same process and simply replace $\sigma^2_{s,m}$ by σ^2_{sil} in Eq. (19). In order to evaluate the posterior $p(\widetilde{X}_t|Y_t,B_t)$ in Eq. (15) we still need to compute $p(M_t=m|Y_t,B_t,St=s)$ and $p(S_t=s|Y_t,B_t)$.

First consider

$$\begin{aligned}
p(M_t=m|Y_t,B_t,S^t=s) &= \frac{p(M_t=m,Y_t,B_t,S^t=s)}{p(Y_t,B_t,S^t=s)} \\
&\propto p(Y_t,B_t,S_t=s,M_t=m).
\end{aligned} \tag{22}$$

But from Eq. (18) we have

$$\begin{aligned}
&p(Y_t,B_t,S^t=s,M_t=m) \\
&= \int_{\widetilde{X}_t} p(\widetilde{X}_t,Y_t,B_t,S_t=s,M_t=m)\mathrm{d}\widetilde{X}_t \\
&\sim N(B_t;\gamma_{s,m},\eta_{s,m})N(Y_t;g_{x_t}\mu_{s,m}\sigma_1^2) \\
&\times p(M_t=m|S_t=s)p(S_t=s).
\end{aligned} \tag{23}$$

Thus,

$$\begin{aligned}
&p(M_t=m \mid Y_t,B_t,S^t=s) \\
&\propto N(B_t;\gamma_{s,m}\eta_{s,m})N(Y_t;g_{x_t}\mu_{s,m},\sigma_1^2) \\
&\times p(M_t=m|S_t=s)p(S_t=s).
\end{aligned} \tag{24}$$

But, recall that M_t is global over a given frame. Thus let

$$\begin{aligned}
&N(B_t^f;\gamma_{s,m}^f,\eta_{s,m}^f)N(Y_t^f;g_{x_t}\mu_{s,m}^f,\sigma_1^{f2}) \\
&\times p(M_t=m|S_t=s)p(S_t=s) \triangleq k(m,f).
\end{aligned} \tag{25}$$

We can compute the value of the posterior using

$$p(M_t=m|Y_t,B_t,S_t=s) = \frac{\prod_f \kappa(m,f)}{\sum_m \prod_f \kappa(m,f)}. \tag{26}$$

The posterior of S_t may be obtained in a similar manner by observing that

$$\begin{aligned}
p(S_t=s|Y_t,B_t) &\propto \sum_m p(Y_t,B_t,S^t=s,M_t=m) \\
&\sim \sum_m N(B_t;\gamma_{s,m},\eta_{s,m})N(Y_t;g_{x_t}\mu_{s,m},\sigma_1^2) \\
&\times p(M_t=m|S_t=s)p(S_t=s).
\end{aligned} \tag{27}$$

Later (see Section 4.4), we show how the above posterior can be used to build a speech detector. Now, we return to our original problem, i.e., computing $p(\widetilde{X}_t|Y_t,B_t)$: As it can be seen, we have now computed all the terms needed to evaluate the posterior - $p(\widetilde{X}_t|Y_t,B_t,S_t=s,M_t=m)$ (Eq. (21)), $p(M_t=m|Y_t,B_t,S_t=s)$ (Eq. (26)), and $p(S_t=s|Y_t,B_t)$ (Eq. (27)). In practice, though, we are more interested in most likely value of $\widetilde{X}_t$ given Y_t and B_t. In other words, we are interested in computing $E(\widetilde{X}_t|Y_t,B_t)$. Thus, taking an expectation w.r.t $p(\widetilde{X}_t|Y_t,B_t)$ in Eq. (15) we get,

$$\begin{aligned}
\widehat{\widetilde{X}}_t &= E(\widetilde{X}_t|Y_t,B_t) = p(S_t=0|Y_t,B_t)E(\widetilde{X}_t|Y_t,B_t,S_t=0,M_t=0) \\
&+ p(S_t=1|Y_t,B_t)\sum_m p(M_t=m|Y_t,B_t,S_t=1) \\
&\times E(\widetilde{X}_t|Y_t,B_t,S_t=1,M_t=m).
\end{aligned} \tag{28}$$

As the the posterior $p(\widetilde{X}_t|Y_t,B_t,S_t=s,M_t=m)$ follows a Gaussian distribution, $E(\widetilde{X}_t|Y_t,B_t,S_t=s,M_t=m)=\alpha_{s,m}$. Once again, we would like to remind the reader that the above equation is for a single frequency component of $\widetilde{X}_t$. To generate the entire vector $\widetilde{X}_t$, we would need to use the above equation **N** times (recall **N** is the dimension of $\widetilde{X}_t$). The above equation is intuitively appealing as the expected value of $\widetilde{X}_t$ is given by a sum of its value given a particular state weighted by the probability of that state given the observations. Note that $\widehat{\widetilde{X}}_t$ is an MMSE estimator for $\widetilde{X}_t$ [Ephraim and Malah, 1984].

4.4. Speech detection

In this section, we discuss how some of the posteriors obtained in the previous section may be naturally extended

to obtain a speech detector. Recall $p(S_t = s|Y_t, B_t) \propto \sum_m p(Y_t, B_t, S^t = s, M_t = m)$ (see Section 4.3) and Eq. (23)

$$\begin{aligned} p(Y_t, B_t, S^t = s, M_t = m) &= \int_{\widetilde{X}_t} p(\widetilde{X}_t, Y_t, B_t, S_t = s, M_t = m) \mathrm{d}\widetilde{X}_t \\ &\sim N(B_t; \gamma_{s,m}, \eta_{s,m}) N(Y_t; g_{x_t}\mu_{s,m}\sigma_1^2) \\ &\times p(M_t = m|S_t = s) p(S_t = s) \end{aligned} \tag{29}$$

In the above equation, the first distribution $N(B_t; \gamma_{s,m}, \eta_{s,m})$ models the correlation between the air and bone microphone channels whereas the second term $N(Y_t; g_{x_t}\mu_{s,m}\sigma_1^2)$ makes use of the prior (along with variance and sensor noise in the air microphone channel) to explain the observation in the air microphone channel. The second term is important because we cannot rely on just the correlation for classes of phones that are weak in the bone sensor (e.g. fricatives).

As S_t is global over a given frame, we rewrite Eq. (29) as

$$p(Y_t^f, B_t^f, S_t, M_t) \sim \Psi_{s,m}^f \Delta_{s,m}^f p(M_t|S_t) p(S_t), \tag{30}$$

with $\Psi_{s,m}^f = N(B_t^f; \gamma_{s,m}^f, \eta_{s,m}^f), \Delta_{s,m}^f = N(Y_t^f; g_{x_t}\mu_{s,m}^f(\sigma_1^2)f)$, where the exponent f represents the fth frequency component. Thus, as a result of our independence of frequency components assumption, the posterior is given by

$$\begin{aligned} p(S_t = s|Y_t, B_t) &\propto \sum\nolimits_m p(S_t = s) p(M_t = m|S_t = s) \\ &\times \prod_{\mathrm{all}f} \Psi_{s,m}^f \Delta_{s,m.}^f \end{aligned} \tag{31}$$

But

$$\begin{aligned} L(S_t = s|Y_t, B_t) &= p(Y_t, B_t|S_t = s) \\ &= \frac{p(S_t = s|Y_t, B_t) p(Y_t, B_t)}{p(S_t = s)} \propto \frac{p(S_t = s|Y_t, B_t)}{p(S_t = s)}, \end{aligned} \tag{32}$$

where $L(S_t = s|Y_t, B_t)$ is the likelihood of S_t taking on the value s given the value of Y_t and B_t. Thus a frame may be classified as speech if $L(S_t = 1|Y_t, B_t) > L(S_t = 0|Y_t, B_t)$ and as non-speech otherwise. This is implemented by defining

$$D_t = \frac{\sum_m p(M_t = m|S_t = 1) \prod_{\mathrm{all}f} \Psi_{1,m}^f \Delta_{1,m}^f}{\sum_m p(M_t = m|S_t = 0) \prod_{\mathrm{all}f} \Psi_{0,m}^f \Delta_{0,m}^f} \tag{33}$$

and the frame is classified as speech if

$$D_t > \frac{p(S_t = 0)}{p(S_t = 1)} \tag{34}$$

and as non-speech other wise. Note that the above equation essentially represents a likelihood ratio test.

4.5. Estimating the gain g_{xt}

As can be noticed, gain g_{xt} is involved in a number of expressions obtained above. Since we are unable to come up with a closed-form solution, we resort to the EM algorithm to estimate g_{xt}. Let

$$q(f) = p(\widetilde{X}_t^f, Y_t^f, B_t^f, S_t, M_t), \tag{35}$$

which is given by Eq. (18). Note that $q(f)$ is the joint likelihood over some of the variables in the model. Though V_t^f is absent in the above equation, note that $\widetilde{X}_t$ was obtained after integrating out V_t^f and thus includes belief about V_t^f. The joint log likelihood over the entire frame is given by

$$F = \log \prod_{\mathrm{all}f} q(f) = \sum_{\mathrm{all}f} \log q(f), \tag{36}$$

where the above equation follows as a result of the frequency independence assumption. In order to maximize F we resort to the EM algorithm. The E-step essentially consists of estimating the most-likely value of $\widetilde{X}_t$ given the current estimate of g_{xt}, i.e., $\widehat{\widetilde{X}}_t = E(p(\widetilde{X}_t|Y_t, B_t, g_{x_t}))$, where $E(\cdot)$ is the expectation operator and $p(\widetilde{X}_t|Y_t, B_t, g_{x_t})$ is given by Eq. (15). The M-step involves maximizing the objective function F w.r.t. g_{xt}. Taking the derivative of F w.r.t g_{xt} and solving for g_{xt} yields

$$g_{x_t} = \frac{\sum_{\mathrm{all}f} [(Y_t^* \widetilde{X}_t + Y_t \widetilde{X}_t^*)\sigma_w^2 + C\sigma_v^2]}{\sum_{\mathrm{all}f} [|\widetilde{X}_t|^2 \sigma_w^2 + |H - G|^2 |\widetilde{X}_t|^2 \sigma_v^2]}, \tag{37}$$

where

$$C = (B_t - GY_t)^*(H - G)\widetilde{X}_t + (B_t - GY_t)(H - G)^* \widetilde{X}_t^*.$$

It should be noted here that we do not estimate g_{xt} for the Gaussian that models silence, and in that case, g_{xt} is set to 1. Indeed, we do not perform magnitude normalization in modeling the silence because the energy of a silence frame is essentially zero (or close to it) and this is true irrespective of device gains or changes in loudness.

4.6. Dynamics of S_t

The enhancement process starts off with both $S_t = 0$ and $S_t = 1$ being equally likely, i.e., $p\ (S_t = 0) = p(S_t = 1) = 0.5$. In order to enforce smoothness in the state estimates we use the following state dynamics:

$$p(S_t = s|S_{t-1} = s) = \frac{0.5 + p(S_{t-1} = s|Y_{t-1}, B_{t-1})}{2} \tag{38}$$

and $p(S_t = s|S_{t-1} = 1 - s) = 1 - p(S_t = s|S_{t-1} = s)$. This way, if the previous frame is a speech, i.e., $p(S_{t-1} = 1|Y_{t-1}, B_{t-1}) > 0.5$, the prior for the current frame being speech, i.e., $p(S_t 1| = 1 S_{t-1} = 1)$, is larger than 0.5. The same is true: $p(S_t = 0|S_{t-1} = 0) > 0.5$. This introduces some bias towards the state of the previous frame, making frame-to-frame transition smoother. It should be noted here that $p(S_t|S_{t-1})$ is a function of time and thus the model represents the so called non-homogeneous Markov chain.

5. Experimental setup

In this paper, we make use of Microsoft's Internal noisy speech corpus for all our experiments. Details of the corpus are as follows: a large number of utterances were recorded

from a number of speakers using the ABC microphone in various environments including cafeteria (ambient noise level 85 dBc), office with interfering speakers in the background, driving on a highway (with windows rolled down and radio running) and other real-world noisy environments. It is important to highlight that the utterances are corrupted by real-world noise, which means that we do not have the ground-truth (i.e. clean) utterances.

The speech model was learnt offline as described in Section 3. The noise model is built in the following way. Given a noisy utterance, we first remove all teethclacks in the bone channel using the algorithm proposed in (Liu et al., 2005). We then run an energy based speech detector (described in (Zheng et al., 2003)) on the first two seconds to obtain an initial estimate of σ^2 and σ_w^2. As we model noise using a single, zero-mean Gaussian, we have the unbiased estimate of the variances given by

$$\sigma_v^2 = \frac{1}{|N_v| - 1} \sum_{t \in N_v} |Y_t|^2,$$
$$\sigma_w^2 = \frac{1}{|N_v| - 1} \sum_{t \in N_v} |B_t - G_t Y_t|^2, \quad (39)$$

where $\{N_v\}$ is the set of non-speech (or noise corrupted) frames. Recall that σ_v^2 is the variance of the corrupting noise, and σ_w^2 is the variance of the sensor noise in the bone microphone. These initial estimates are then used in the framework described above. Also we set $\sigma_u^2 = 10^{-4}\sigma_w^2$. This is based on empirical studies and the observation that close-talk sensor technology is more advanced than bone-sensor technology. The transfer functions H_t and G_t are estimated using the procedure described in Section 4.2. As each of the above parameters need to be updated for each frame, this poses a huge computational overhead. Thus we use a recursive implementation as described in Appendix A. We have found that this makes the algorithm computationally efficient and speech enhancement runs in real-time. The iterative estimation for g_{x_t} usually converged within 2–3 iterations. For the first frame in the utterance, g_{x_t} was initialized to 1 at the start of the EM algorithm. In the case of subsequent frames, g_{x_t} was initialized to the last converged value (i.e. the value of g_{x_t} for the previous frame). All utterances for this work were processed using a Hamming window of size 25 ms at 100 Hz.

As explained in Section 1, for our applications, we are interested in the overall user experience in noisy environments. Thus perceptual quality of the processed utterances is very important. To measure the improvement in quality, we conducted comparative evaluations based on mean opinion score (MOS) (Deller et al., 1999). Table 1 shows the score criteria.

In order to measure the sensitivity of the speech model to speakers, we trained two speech models. The first (Ω_1) was trained on clean speech from a single speaker and the second model (Ω_2) was trained on clean speech utterances from four different speakers (two males and two females). Each model is a mixture of four Gaussians. The speaker in Ω_1 is one of the male speakers in Ω_2.

Table 1
MOS evaluation criteria

Score	Impairment
5	(Excellent) imperceptible
4	(Good) (Just) perceptible but not annoying
3	(Fair) (Perceptible and) slightly annoying
2	(Poor) annoying (but not objectionable)
1	(Bad) very annoying (objectionable)

The testing set consisted of 12 noisy utterances, with an equal male–female ratio, recorded in a number of noisy environments as explained above. The testing set included a representative utterance from all different recording conditions in the Microsoft noisy speech corpus. Each utterance in the test set was processed using five different algorithms:

1. the classical spectral subtraction algorithm (Quatieri, 2002),
2. our previously proposed direct filtering algorithm (Liu et al., 2004),
3. the algorithm described in (Subramanya et al., 2005) which uses a single Gaussian for the speech model,
4. the proposed mixture of Gaussians speech model trained with one speaker (Ω_1), and
5. the proposed mixture of Gaussians speech model trained with four speakers (Ω_2).

Therefore, each participant in the MOS study was asked to rate 72 utterances. In the case of spectral subtraction, noise profile was computed offline using the non-speech regions of the utterance. There were a total of 17 participants (subjects) in the MOS test. The evaluators were presented with a random ordering of the sets of utterances and random ordering within a set. The participants were blind to the relationship between the utterances and the processing algorithm.

6. Results

We first discuss some simulation results on the use of the magnitude normalized speech spectrum. Later we present the results of our MOS tests.

6.1. Magnitude-normalized speech model

In order to test the proposed magnitude-normalized speech model, we did two experiments: In the first experiment, we classified the frames (into speech and non-speech) in a number of utterances using the proposed model. The second experiment was designed to test the robustness of the proposed model to mismatch between training and testing conditions. Fig. 6 shows the spectrogram of four clusters obtained as a result of the clustering algorithm

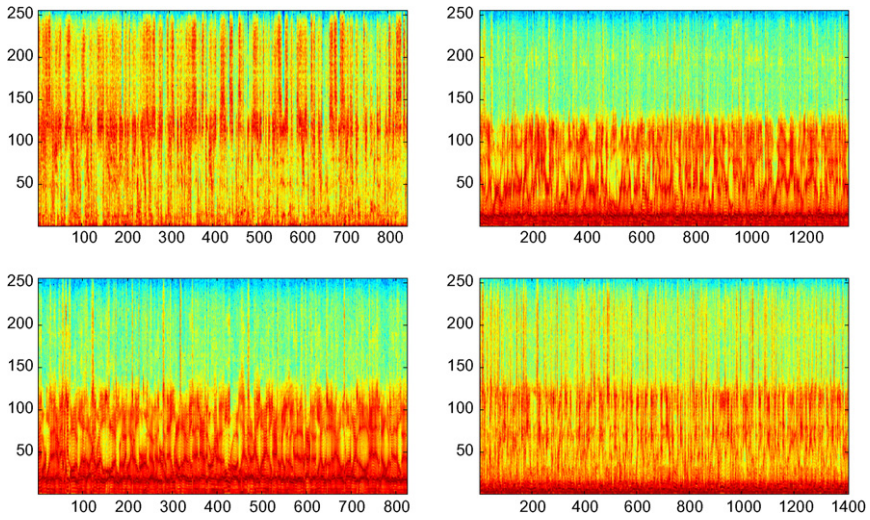

Fig. 6. Figure shows the clusters obtained from the algorithm detailed in Section 3.2. In each plot, the y-axis is the index of the frequency component (FFT index), and the x-axis presents the number of frames classified within that cluster.

described in Section 3. The figure was generated in the following manner: we first ran the algorithm proposed in Section 3 on the training set with $M = 4$ (this step yielded α_i, μ_i and $\Sigma_i, 0 \leqslant i \leqslant 3$). Next, we choose four utterances from four different speakers outside the training set. These utterances were concatenated, and then transformed into the complex spectral domain using STFT with a Hamming window of size 25 ms at l00 Hz. If $\{\widetilde{X}_t\}_{t=1}^{T_s}$ are the frames obtained from the STFT (after magnitude normalization), each $\widetilde{X}_t$ was classified using

$$\begin{aligned} k_t &= \text{argmin}_m\{d(\widetilde{X}_t, \mu_m)\}, \quad 1 \leqslant t \leqslant T_s, \\ d(\widetilde{X}_t, \mu_m) &= \|[d(\widetilde{X}_t^1, \mu_m^1 \ldots, d(\widetilde{X}_t^f, \mu_m^f), \ldots, d(\widetilde{X}_i^{\mathbf{N}}, \mu_m^{\mathbf{N}})]^T\|, \\ d(\widetilde{X}_i^f, \mu_m^f) &= \log|\widetilde{X}_i^f| - \log|\mu_m^f|, \quad 1 \leqslant f \leqslant \mathbf{N}, \end{aligned} \tag{40}$$

where k_t is the label for the frame at time t and μ_m is the mean of the mth cluster. Each of the four sub-plots in Fig. 6 show the frames that were assigned the same label according to the above distance measure. It can be seen that one cluster models fricatives (sub-figure in row 1, column 1), another cluster models purely vowels (sub-figure in row 2, column 1). This shows that the proposed distance metric leads to meaningful clusters and thus can be used as a prior for speech.

In order to test the proposed model for robustness, we built two speech models using a single Gaussian, one using the proposed energy-normalized spectra (ω_1) and the other using original spectra (ω_2) in the complex spectral domain. In other words, ω_1 used $\widetilde{X}_t$, whereas, ω_2 made use of X_t. Note that for comparison, we only use a single Gaussian in each model. The two models were then used to compute the likelihood $L(S_t = 1|Y_t, B_t)$ for all the frames in an utterance outside the training set. The speaker was outside the training set, and the gain on the recording device was set to a different level when compared to the utterances in the training set. The likelihoods in the two cases are shown in Fig. 7. It can be seen that the likelihoods resulting from ω_1 are always greater than the likelihoods resulting from ω_2 suggesting that the magnitude-normalized speech model can better explain speech signals. A similar trend was observed even when making use of an utterance recorded with a similar gain setting as the utterances used in the training set. It should be noted here that the above does not imply that a speech frame will be classified as speech in a practical setting, as this would also depend on the likelihoods from the alternate competing model (noise/silence).

6.2. Enhancement results

Table 2 shows the results of the MOS tests. There are a number of very interesting observations that can be made from the test results. First, it can be seen from the table that the subjects on average preferred the original corrupted utterances over those processed by spectral subtraction. In fact most subjects were more comfortable in listening to corrupted speech rather than to distorted speech. It is not surprising that spectral subtraction introduced distortions in the processed signal as most corrupting noise sources were non-stationary. Second, it can be seen that the system that uses a single Gaussian to model speech does worse than the proposed algorithms. Clearly a single Gaussian is unable to capture all the spectral profiles. Third, the multi-speaker model Ω_2 performs only slightly worse than the single speaker model. This suggests that our proposed magnitude-normalized speech model is able to generalize fairly well.

Fig. 8 shows the time and frequency domain renditions of a noisy utterance used in the above MOS tests[2]. This utterance was recorded in an sound proof environment

[2] The wave files used to generate this figure can be obtained at http://research.microsoft.com/~zhang/WITTY/samplewaveforms.htm.

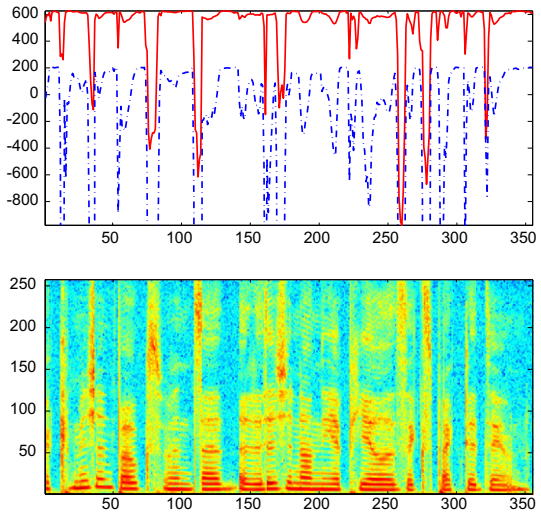

Fig. 7. Comparison of likelihoods of the two speech models. In the first plot, the y-axis represents the log-likelihood shifted by a constant c ($c = 750$) in the positive direction, the likelihood of ω_1 is represented with a solid line, and that of ω_2 is represented with a dotted line. The second plot shows the spectrogram of the utterance used to compute the likelihood. In the second plot, the y-axis is the index of the frequency component (FFT index). In both plots, the x-axis represents time in seconds.

with two interfering speakers. The ambient noise level was 75 dBc. The first two figures show the time and frequency domain information captured by the air channel, whereas the last two figures show the time and frequency domain information captured by the bone channel. It is evident that speech energy from only the air channel does not provide any indication of the speech/non-speech states at any given instant. The second figure shows the signal captured by the bone channel. It exemplifies the robustness of the bone sensor to ambient noise despite a small amount of leakage. Fig. 9 shows the resulting spectra when the above utterance was processed using the noise removal techniques listed in Section 5. The first figure shows the signal captured by the air channel (repeated from Fig. 8 for convenience). The second figure shows the results of the spectral subtraction algorithm. As it can be seen the spectrum appears distorted and thus gets a lower MOS score. The third figure shows the result of the direct filtering algorithm (Liu et al., 2004). It can be seen that algorithm does not successfully rid the the signal of all the corrupting noise. Furthermore, it can also be observed that the

Table 2
MOS results

Algorithm	User preference score	Standard deviation
Original	2.58	0.23
Spectral subtraction	2.22	0.42
Direct filtering	2.78	0.61
Single Gaussian	3.03	0.32
Mixture Gaussian (Ω_1)	3.75	0.15
Mixture Gaussian (Ω_2)	3.71	0.19

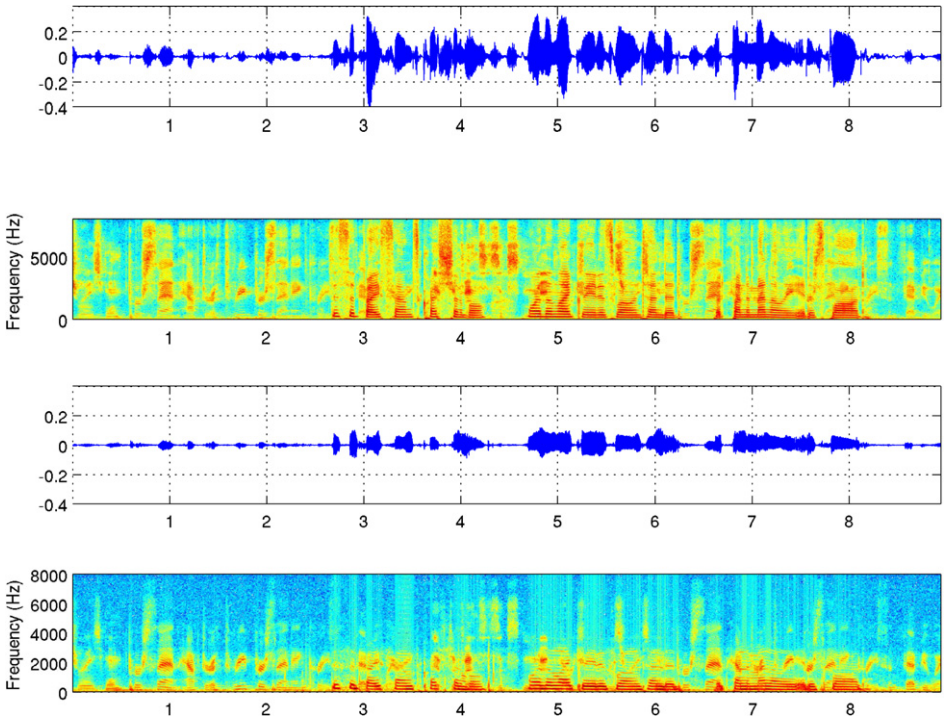

Fig. 8. Figure showing the time and frequency domain renditions of a noisy utterance from the Microsoft noisy speech corpus. The utterance consists of a female speaker saying "this advice sounds questionable most likely it is well intended but fundamentally it is flawed". All x-axes represent time in seconds. While the first two plots show the signal captured by the air microphone, the third and fourth plots show the signal captured by the bone microphone.

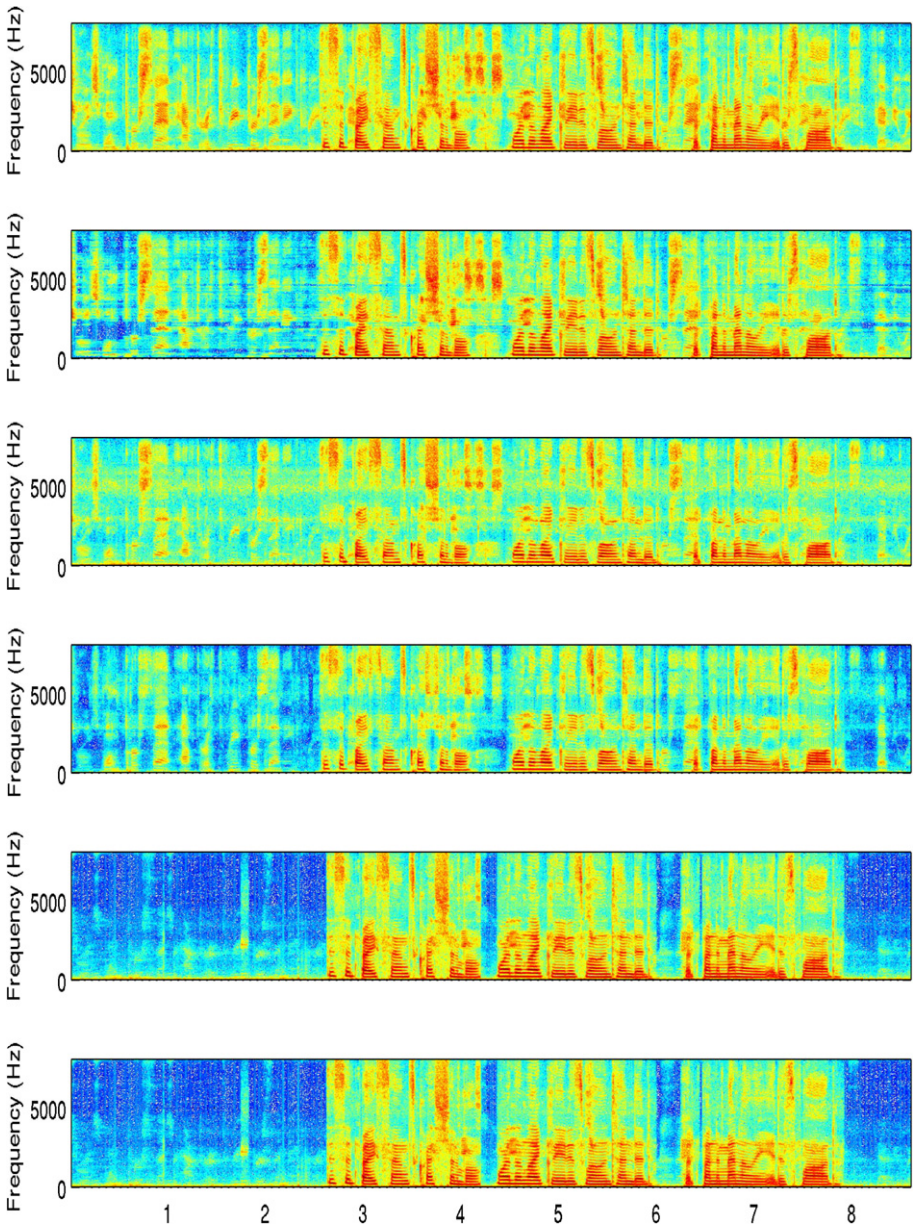

Fig. 9. Results: (a) Noisy signal, (b) enhanced by classical spectral subtraction, (c) enhanced using [Liu et al., 2004], (d) enhanced using [Subramanya et al., 2005], (e) enhanced using Ω_1, and, (f) enhanced using Ω_2.

processed signal is distorted. Next are the results obtained using the algorithm proposed in (Subramanya et al., 2005). While this does better than the (Liu et al., 2004), a single Gaussian is usually unable to explain all the variations of speech. The last two figures show the results of Ω_1 and Ω_2 respectively.

It can be seen that the proposed algorithm not only gets rid of a significant amount of noise, but does so with little/no distortion in speech. Also, there is little difference between the results of Ω_1 and Ω_2 thus, highlighting the fact that the magnitude normalized model makes it easier for even single speaker models to explain speech from other users. Therefore, it is not surprising that subjects prefer the utterances processed by the proposed algorithm.

7. Conclusion and future work

In this paper we have proposed a new multi-sensory microphone that makes use two sensor streams to enhance speech: a standard air microphone and a bone sensor. We also have proposed a mixture-of-Gaussians speech model built from magnitude-normalized complex spectra for speech enhancement. We have shown how the proposed speech model can be used in the context of speech enhancement with an air- and bone-conductive microphone. Substantial improvement has been observed in the MOS evaluation over the best of our previously developed techniques. Comparison between single-speaker trained and multi-speaker trained models suggests that the proposed

magnitude-normalized speech model is able to generalize fairly well. Furthermore, the proposed technique is computationally inexpensive and runs in real time.

For our future work, we are planning to introduce dynamics on other variables such as $\widetilde{X}_t$ and X_t. Using Kalman updates coupled with the proposed inference framework may lead to better estimates of the clean speech signal. We are also working on recursive noise update with noisy speech frames. Another area of future work includes, making use of a class dependent transfer function H_t. While the class independence assumption made in this work, performs fairly, a class independent H_t will be able to further the use of the bone sensor for enhancing the audio in the air channel.

Appendix A. Online estimation of H_t

Recall that in Section 4.2 the mapping function H_t is estimated over a number of speech frames with the following equation:

$$H = \frac{\sum(\sigma_v^2|B_t|^2 - \sigma_w^2|Y_t|^2) \pm \sqrt{(\sum(\sigma_v^2|B_t|^2 - \sigma_w^2|Y_t|^2))^2 + 4\sigma_v^2\sigma_w^2|\sum B_t^* Y_t|^2}}{2\sigma_v^2 \sum B_t^* Y_t} \tag{A.1}$$

In this appendix we show how the above equation may be implemented efficiently using a recursive procedure.

A.1. Exponential aging technique

As can be observed from the above equation, estimation of H_t requires computing several summations over the last T frames in the form of

$$S(T) = \sum_{t=1}^{T} s_t, \tag{A.2}$$

where s_t is $\sigma_v^2 \mid B_t|^2 - \sigma_w^2 \mid Y_t|^2$ or $B_t^* Y_t$.

With this formulation, the first frame ($t = 1$) is as important as the last frame ($t = T$). However, one would prefer the latest frames contribute more to the estimation of H_t than the older frames. One technique to achieve this is the so-called *exponential aging*. The idea is to replace (A.2) by

$$S'(T) = \sum_{t=1}^{T} c^{T-t} s_t, \tag{A.3}$$

where $c \leqslant 1$. If $c = 1$, then (A.3) is equivalent to (A.2). If $c < 1$, then the last frame is weighted by 1, the before-last frame is weighted by c (i.e., it contributes less than the last frame), and the first frame is weighted by c^{T-1} (i.e., it contributes significantly less than the last frame). Take an example. Let $c = 0.99$ and $T = 100$, then the weight for the first frame is only $0.99^{99} = 0.37$.

What is interesting is the fact that $S'(T)$ can be estimated recursively. Indeed

$$S'(T) = cS'(T-1) + s_T. \tag{A.4}$$

Since it automatically weighs old data less, we do not need to keep a fixed window length, and we do not need to save the data of the last T frames in the memory.

The effective length (memory of past data) is given by

$$L(T) = \sum_{t=1}^{T} c^{T-t} = \sum_{i=0}^{T-1} c^i = \frac{1-c^T}{1-c}. \tag{A.5}$$

The asymptotic effective length is given by

$$L = \lim_{T\to\infty} L(T) = \frac{1}{1-c} \tag{A.6}$$

or equivalently,

$$c = \frac{L-1}{L}. \tag{A.7}$$

Therefore, if we want to have an effective length of, say, 200 frames, we can set $c = \text{f99/200} = 0.995$. We use a recursive implementation similar to the above to estimate σ_u^2 and σ_v^2.

References

Beal, M., Jojic, N., Attias, H., 2003. A graphical model for audiovisual object tracking. IEEE Trans. Pattern Anal. Machine Intell. (PAMI) 25 (l), 828–836.

Bilmes, J., 2000. Dynamic bayesian multi-networks. Proc. Uncertainity in Artificial Intelligence. Stanford, California.

Bilmes, J., 2001. Graphical models and automatic speech recognition. University of Washington Electrical Engineering Technical Report, UWEETR-2001-0005.

Chen Tao, Huang Chao, Chang Eric, Wang Jingchun, 2002. On the use of gaussian mixture model for speaker variability analysis. In: Proc. Internat. Conf. on Spoken Language Processing (ICSLP).

Deller, J.R., Hansen, J.H.L., Proakis, J.G., 1999. Discrete-Time Processing of Speech Signals. IEEE Press.

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

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, 109–1121.

Ephraim, Y., Lev-Ari, H., Roberts, W.J.J., 2005. A brief survey of speech enhancement. CRC Electronic Engineering Handbook. CRC Press.

Zweig, G., Bilmes, J., et al., 2002. Structurally discriminative graphical models for automatic speech recognition: Results from the 2001 Johns hopkins summer workshop. In: Proc. Internat. Conf. on Acoustics, Speech, and Signal Processing (ICASSP).

Gannot, S., Burshtein, D., Weinstein, E., 1998. Iterative and sequential kalman filter-based speech enhancement algorithms. IEEE Trans. Speech Audio Process. 6, 373–385.

Graciarena, M., Franco, H., Sonmez, K., Bratt, H., 2003. Combining standard and throat microphones for robust speech recognition. IEEE Signal Process. Lett. 10, 72–74.

Heracleous, P., Nakajima, Y., Lee, A., Saruwatari, H., Shikano, K., 2003. Accurate hidden Markov models for non-audible murmur (nam) recognition based on iterative supervised adaptation. In: Proc. IEEE

Automatic Speech Recognition and Understanding Workshop (ASRU), US Virgin Islands.

Jeannes, W.L.B., Scalart, P., Faucon, G., Beaugeant, C., 2001. Combined noise and echo reduction in hands-free systems: A survey. IEEE Trans. Speech Audio Process. 9 (8).

Jordan, M.I., Weiss, Y., 2002. Graphical models: Probabilistic Inference. MIT Press.

Klein, D., Manning, C., 2002. Conditional structure versus conditional estimation in NLP models. In: Proc. Workshop on Empirical Methods in Natural Language Processing (EMNLP).

Lee, B.G., Lee, K.Y., Ann, S., 1995. An EM based approach for parameter enhancement with an application to speech signals. Signal Process. 46, 1–14.

Lim, J.S., Oppenheim, A.V., 1979. Enhancement and bandwidth compression of noisy speech. Proc. IEEE 67 (12), 1586–1604.

Liu, Z., Zhang, Z., Acero, A., Droppo, J., Huang, X., 2004. Direct filtering for air-and bone-conductive microphones. In: Proc. IEEE Internat. Workshop on Multimedia Signal Processing (MMSP), Siena, Italy, September. pp. 363–366.

Liu, Z., Subramanya, A., Zhang, Z., Droppo, J., and Acero, A., 2005. Leakage model and teeth clack removal for air- and bone-conductive integrated microphones. In: Proc. Internat Conf. on Acoustics, Speech, and Signal Processing (ICASSP), Vol. 1. Philadelphia, March, pp. 1093–1096.

Lotter, T., Benien, C., Vary, P., 2003. Multichannel speech enhancement using bayesian spectral amplitude estimation. In: Proc. Internat. Conf. on Acoustics, Speech, and Signal Processing (ICASSP).

Meyer, J., Simmer, K.U., 1997. Multi-channel speech enhancement in a car environment using wiener filtering and spectral subtraction. In: Proc. Internat. Conf. on Acoustics, Speech, and Signal Processing (ICASSP).

Nandkumar, S., Hansen, J.H.L., 1995. Dual-channel iterative speech enhancement with constraints based on an auditory spectrum. IEEE Trans. Speech Audio Process. 3 (l), 22–34.

Paliwal, K.K., Basu, A., 1987. A speech enhancement method based on kalman filtering. In: Proc. Internat. Conf. on Acoustics, Speech, and Signal Processing (ICASSP).

Potamianos, G., Neti, C., Luettin, J., Matthews, I., 2004. Audio-visual automatic speech recognition: an overview, in issues in visual and audio-visual speech processing. In: Bailly, G., Vatikiotis-Bateson, E., Perrier, P. (Eds.). MIT Press.

Quatieri, T.F., 2002. Discrete-Time Speech Signal Processing: Principles and Practice. Prentice Hall.

Strand, O.M., Holter, T., Egeberg, A., Stensby, S., 2003. On the feasibility of asr in extreme noise using the PARAT earplug communication terminal. In: Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), US Virgin Islands.

Subramanya, A., Zhang, Z., Liu, Z., Droppo, J., Acero, A., 2005. A graphical model for multi-sensory speech processing in air-and-bone conductive microphones. In: Proc. Eurospeech – Eur. Conf. on Speech Communication and Technology, Lisbon, Portugal, September.

Wu, J., Droppo, J., Deng, L., Acero, A. 2003. A noise-robust asr front-end using wiener filter constructed from mmse estimation of clean speech and noise. In Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), US Virgin Islands, December.

Yoshizawa Shingo, Hayasaka Noboru, Wada Naoya, Miyanaga Yoshikazu, 2004. Cesptral gain normalization for noise robust speech recognition. In: Proc. Internat. Conf. on Acoustics, Speech, and Signal Processing (ICASSP).

Zhang, Z., Liu, Z., Sinclair, M., Acero, A., Deng, L., Droppo, J., Huang, X., Zheng, Y., 2004. Multi-sensory microphones for robust speech detection, enhancement and recognition. In: Proc. Internat. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), Vol. 3. Quebec Canada, Montreal, pp. 781–784.

Zhao, D.Y., Kleijn, W.B., 2005. On noise gain estimation for hmm-based speech enhancement. In: Proc. Eurospeech – Eur. Conf. on Speech Communication and Technology.

Zheng, Y., Liu, Z., Zhang, Z., Sinclair, M., Droppo, J., Deng, L., Acero, A., Huang, X., 2003. Air- and bone-conductive integrated microphones for robust speech detection and enhancement. In: Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), US Virgin Islands, December, pp. 249–254.