MICROSOFT RESEARCH TECHNICAL REPORT MSR-TR-2005-126

SPEECH MODELING WITH MAGNITUDE-NORMALIZED COMPLEX SPECTRA AND ITS APPLICATION TO MULTISENSORY SPEECH ENHANCEMENT

Amarnag Subramanya[*]

SSLI Lab,
University of Washington,
Seattle, WA 98195.
asubram@ee.washington.edu

Zhengyou Zhang, Zicheng Liu, Alex Acero

Microsoft Research,
One Microsoft Way,
Redmond, WA 98052.
{zhang,zliu,alexac}@microsoft.com

ABSTRACT

A good speech model is essential for speech enhancement, but it is very difficult to build because of huge intra- and extra-speaker variation. We present a new speech model for speech enhancement, which is based on statistical models of magnitude-normalized complex spectra of speech signals. Most popular speech enhancement techniques work in the spectrum space, but the large variation of speech strength, even from the same speaker, makes accurate speech modeling very difficult because the magnitude is correlated across all frequency bins. By performing magnitude normalization for each speech frame, we are able to get rid of the magnitude variation and to build a much better speech model with only a small number of Gaussian components. This new speech model is applied to speech enhancement for our previously developed microphone headsets that combine a conventional air microphone with a bone sensor. Much improved results have been obtained.

Keywords: Speech modeling, speech enhancement, audio processing, multisensory processing

1. INTRODUCTION

Speech enhancement in a noisy environment has many applications including communications and speech recognition. Despite more than three decades of research, it remains unsolved. The difficulty is due to non-stationarity of speech and noise, huge intra- and extra-speaker variability, often unpredictable environmental conditions (noise and reverberation), and sometimes arbitrary microphone gain setting. An efficient speech enhancement technique requires explicit and accurate statistical models for the speech signal and noise process.

Much work has been done on speech enhancement using traditional signal processing techniques. Quatieri [1] provides a description of various techniques including spectral subtraction, Wiener filtering, model-based processing and auditory masking. They have many well-known successes, such as removing fan noise, because the reconstructed clean spectra are close to optimal if the true noise spectra are given and the correlation between speech and noise is known. One shortcoming of these techniques is that they often assume, implicitly or explicitly, a single Gaussian distribution on speech signals. As mentioned earlier, a single Gaussian is a poor model to account for huge speech variation.

McAulay and Malpass in their pioneering work [2] used a speech presence uncertainty model and developed a soft-decision noise suppression filter for speech enhancement. Drucker [3] first proposed a system using five states representing fricative, stop, vowel, glide, and nasal speech sounds. Based on a series of listening tests for confusion, the number of states was reduced to three (fricative, stop, and other sounds). The system, however, was simulated by hand-switching between the sound state (the test monitor knows the speech stream).

Attempts have also been made to model state changes over time. Lim and Oppenheim [4] model the short-term speech and noise signals as an autoregressive process. Ephraim [5] models 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 [6]. Instead of working in the spectral domain, Wu et al. [7] trained a HMM or a GMM in the cepstral or MFCC domain to estimate the clean speech and noise spectra as a front-end to a speech recognition system. Experiments with the Aurora2 database show that models trained using MFCC work better (28% of relative WER reduction over baseline systems).

In this paper, we propose a novel speech model. Instead of training the models on the complex spectra of speech, we normalize the spectral components of each frame by its energy, and then build a Gaussian Mixture Model from the magnitude-normalized complex spectra. The above step normalizes for variations in the loudness of the speaker's voice; in addition it makes the models robust to other factors such as microphone gain (assuming the signals are not saturated or floored.). One way of getting around the above

[*]The work was done at Microsoft Research.

problem, which has been extensively used in the past, is to use MFCCs, where the first term always represents energy in the frame and is simply ignored for normalization. However working in the mel-cepstral domain has its own disadvantages, which include issues with non-linearity and absence of phase in the reconstructed signal.

As mentioned earlier, while we have seen many successes to deal with stationary noise such as fan, enhancement in the presence of non-stationary background noise is still an open problem. To tackle this problem, we have developed a novel hardware solution to combat against highly non stationary acoustic noise such as background interfering speech [8, 9]. The device makes use of an inexpensive bone-conductive microphone in addition to the regular air-conductive microphone. The signal captured by the latter is corrupted by environmental conditions, whereas the signal in the former is relatively noise-free. The bone sensor captures the sounds uttered by the speaker but transmitted via the bone and tissues in the speaker's head. High frequency components ($>$ 3Khz) are absent in the bone sensor signal. Thus, the challenge here is to enhance the signal in the air-channel by fusing the two streams of information.

In [8], 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 problems. In the same work, we also proposed a speech detector based on a histogram of the energy in the bone channel. In [10], we proposed an algorithm called direct filtering (DF) that does not require any prior training in order to estimate the clean speech signal, i.e. 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 in comparison to the [8] 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 [11], 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. All these approaches require accurate speech/voice activity detection, and the technique proposed in [8] makes use of a function of the energy in the bone sensor. This has two problems associated with it: A) some classes of phones (e.g., fricatives) have low energy in the bone sensor causing false negatives; and B) leakage in the bone sensor can lead to false positives. Further, by using just the bone sensor for speech detection, we are not leveraging the two channels of information provided by the multisensory headset. In [12], we proposed an algorithm that takes into account the correlation between the two channels for speech detection and also incorporates a speech model within the graphical model framework thereby reducing the amount of distortion in the enhanced signal. However, the speech model is only a single Gaussian. In this paper, we describe how the proposed speech model is used to achieve even better speech enhancement.

The paper is organized as follows. Section 2 describes our proposed speech model. Sections 3 and 4 show how its application to speech enhancement with an air- and bone-conductive microphone headset. Section 5 provides the experiment results. Section 6 concludes the paper.

2. MAGNITUDE-NORMALIZED COMPLEX SPECTRUM-BASED SPEECH MODEL

In Bayesian statistics, prior information on hidden variables plays a crucial role in inference. A speech model lends itself into such a role by providing a prior on clean speech that is hidden given noisy speech. Human speech is very difficult to model due to its large variability. Some of the factors contributing to this variability include: differing speech profiles for different speakers; changes in loudness, intonation and stress for a single speaker; variations due to gender. One way to deal with issues related to changes in loudness and changes in recording device gains is to build the model in the mel-cepstral domain, where such changes are reflected in the first cepstral coefficient that is then neglected for modeling purpose. However, such models have the disadvantage that they do not encode any information about the phase of the speech signals.

2.1. Model Definition

In our case, we are interested in estimating both the magnitude and phase of the clean speech signal which explains why we work in the complex spectral domain. However, in the complex spectral domain, the variations due to loudness cannot be accounted for by simply getting rid of some components. Thus, 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.,

$$\tilde{X}_t = \frac{X_t}{\|X_t\|}. \quad (1)$$

Thus all $\tilde{X}_t$'s are unit vectors and distribute 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} for inference to match a particular speech frame. We discuss an iterative approach to estimating the gain in

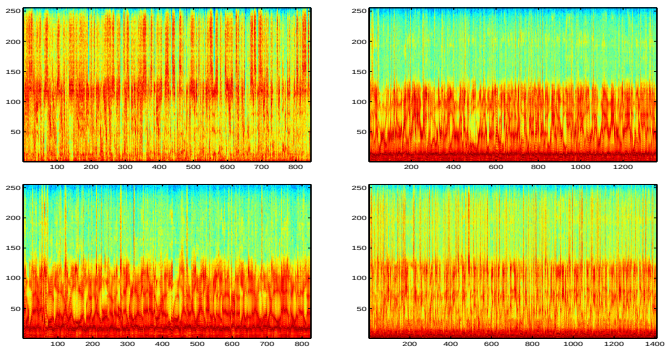

Fig. 1. Clustering results

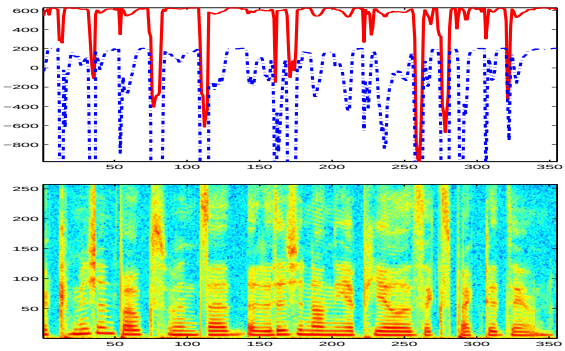

Fig. 2. Comparison of likelihoods of two speech models with and without magnitude normalization.

section 4. Furthermore to add robustness to the model, we neglect the DC and Nyquist terms while building the model. Thus our models are of $\frac{N}{2} - 1$ dimensions, where N is the length of the FFT.

2.2. Training

In order to train the speech model, we collected data from a number of speakers in a clean environment. The speech frames from the utterances were extracted using a simple energy based speech detector. The resulting speech frames were then energy normalized as explained in the previous subsection. We use a mixture of Gaussians to model the normalized speech frames using the k-means algorithm with random initialization. Since it is well known that human are perceptually more sensitive to log magnitude, we use the following quantity to measure the difference between two vectors $\tilde{X}_i$ and $\tilde{X}_j$:

$$d(\tilde{X}_i, \tilde{X}_j) = \|(\log|\tilde{X}_i| - \log|\tilde{X}_j|)\| \,, \qquad (2)$$

where $\log \tilde{X}$ denotes that the log operation is applied to each element of $\tilde{X}$.

2.3. Experimental Results

Figure 1 shows the spectrogram of four clusters obtained as a result of the clustering algorithm described above. Here, we have concatenated all speech frames in the same cluster and shown them in a separate picture. It can be seen that one cluster models fricatives, and the others model various forms of vowels.

In order to test the model robustness, we built two speech models using a single Gaussian, one using energy normalized spectra (ω_1) and the other using original spectra (ω_2) in the complex spectral domain. Note that for comparison, we only use a single Gaussian in each model. The two built models were then used to compute the likelihoods in an utterance which is not in the training set but is recorded using a device with similar gain setting as that in the training set. The aggregated likelihoods (across all frequency components) are shown in figure 2. 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. 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).

3. GRAPHICAL MODEL FOR SPEECH ENHANCEMENT IN A MULTISENSORY HEADSET

We are now applying the speech model proposed in the last section to speech enhancement in an air- and bone-conductive integrated microphone headset [8, 9]. Since we work in the complex spectral domain, we transform the time domain signals from the air microphone and the bone sensor into complex spectra by applying the fast-Fourier transform (FFT) to the windowed version of signal samples. The physical process is then modeled in the complex spectral domain as shown in Figure 3. The variables used in the model are described below.

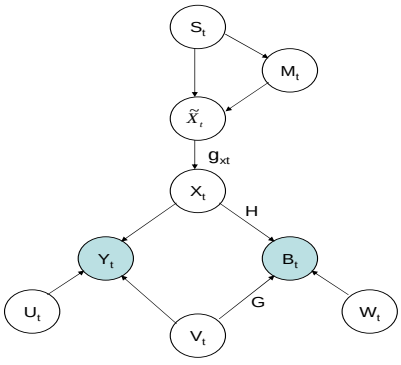

Fig. 3. The graphical model incorporating the proposed speech model.

In the above model, S_t is a discrete random variable representing the state (speech / silent) of the frame at time t, M_t is a discrete random variable acting as an index into the mixture of the speech model, $\tilde{X}_t$ represents the *scaled* version of clean speech signal that needs to be estimated, X_t represents the clean speech signal that needs to be estimated, g_{x_t} scales $\tilde{X}_t$ to match the clean speech X_t from the air conductive microphone, Y_t is the signal captured by the air microphone, B_t is the signal captured by the bone sensor, V_t is the background noise, H is the optimal linear mapping between clean speech and bone signal, G models the leaking of background noise into the bone sensor. The variables $\tilde{X}_t, X_t, Y_t, V_t, B_t$ are all in the complex spectral domain and have $\frac{N}{2}-1$ dimensions, where N is the FFT length. For mathematical tractability we assume that the different components of the above variables, except for S_t and M_t, are all independent. S_t and M_t are global over a given frame.

We make the following assumptions with regard to some of the dependencies between variables and priors in the model,

1. Background noise is modeled with a zero mean Gaussian, i.e., $p(V_t) \sim N(0, \sigma_v^2)$,
2. Sensor noise in the air microphone channel is modeled with $p(U_t) \sim N(0, \sigma_u^2)$
3. Sensor noise in the bone channel is modeled with $p(W_t) \sim N(0, \sigma_w^2)$
4. Speech is modeled using a mixture of Gaussians (MG),

$$p(\tilde{X}_t|S_t) = \sum_{m=1}^{M} P(M_t = m|S_t) p(\tilde{X}_t|S_t, M_t), \quad (3)$$

$$\text{with } p(\tilde{X}_t|S_t, M_t) \sim N(\mu_{sm}, \sigma_{sm}^2) \quad (4)$$

   We assume that $S_t = \{0, 1\}$, where 0 and 1 indicate silence and speech respectively. We model silence using a single Gaussian, and thus $P(M_t = 1|S_t = 0) = 1$ and $p(\tilde{X}_t|S_t = 0) \sim N(0, \sigma_{sil}^2)$. In the case of speech we use a MG with $M = 4$. For simplicity we assume that all the Gaussians in the mixture are equally likely and thus, $P(M_t = i|S_t = 1) = \frac{1}{M}$ for $i = 1, \ldots, M$ and thus,

$$p(\tilde{X}_t|S_t = 1) \sim \frac{1}{M} \sum_{m=1}^{M} N(\mu_{sm}, \sigma_{sm}^2) \quad (5)$$

5. For mathematical tractability we model $p(\tilde{X}_t|X_t) \sim \delta(X_t, g_{x_t}\tilde{X}_t)$, a delta function with parameter g_{x_t}.

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

$$\begin{aligned} p(Y_t, B_t, X_t, \tilde{X}_t, V_t, S_t, M_t, U_t, W_t) &= p(Y_t|X_t, V_t, U_t) \\ & p(B_t|X_t, V_t, W_t) p(X_t|\tilde{X}_t) p(\tilde{X}_t|M_t, S_t) p(M_t|S_t) \\ & p(S_t) p(V_t) p(U_t) p(W_t). \end{aligned} \quad (6)$$

It is very important to note here that in the above equation all variables except S_t, M_t are individual components, whereas S_t, M_t are global over a given frame. As X_t and $\tilde{X}_t$ are related by a delta distribution, given g_{x_t}, estimating either one of these variables is equivalent to estimating the other. Thus, integrating out X_t from the joint distribution gives

$$\begin{aligned} p(Y_t, B_t, \tilde{X}_t, V_t, S_t, M_t, U_t, W_t) &= p(Y_t|g_{x_t}\tilde{X}_t, V_t, U_t) \\ & p(B_t|g_{x_t}\tilde{X}_t, V_t, W_t) p(\tilde{X}_t|M_t, S_t) p(M_t|S_t) \\ & p(S_t) p(V_t) p(U_t) p(W_t)\ . \end{aligned} \quad (7)$$

We are interested in estimating

$$\begin{aligned} p(\tilde{X}_t|Y_t, B_t) &= \sum_{s,m} p(\tilde{X}_t, S_t = s, M_t = m|Y_t, B_t), \\ &= \sum_{s,m} p(\tilde{X}_t|Y_t, B_t, S_t = s, M_t = m) \\ & \qquad p(M_t = m|Y_t, B_t, S_t = s) p(S_t = s|Y_t, B_t) \\ &= p(S_t{=}0|Y_t, B_t) p(\tilde{X}_t|Y_t, B_t, S_t{=}0, M_t{=}0) + \\ & \quad p(S_t{=}1|Y_t, B_t) \sum_m p(M_t{=}m|Y_t, B_t, S_t{=}1) \\ & \qquad p(\tilde{X}_t|Y_t, B_t, S_t{=}1, M_t{=}m) \end{aligned} \quad (8)$$

Let us first consider

$$\begin{aligned} & p(\tilde{X}_t, Y_t, B_t, S_t = s, M_t = m) = \\ & \int_{V_t} \int_{U_t} \int_{W_t} p(Y_t, B_t, \tilde{X}_t, V_t, S_t, M_t, U_t, W_t)\ dU_t dW_t dV_t \end{aligned} \quad (9)$$

After some algebra we get

$$\begin{aligned} & p(\tilde{X}_t, Y_t, B_t, S_t = s, M_t = m) \sim N(\tilde{X}_t; A_1, B_1) \\ & \quad N(B_t; A_2, B_2) N(Y_t; g_{x_t}\mu_{sm}, \sigma_1^2) p(M_t|S_t) p(S_t) \end{aligned} \quad (10)$$

where

$$\begin{aligned} A_1 &= \frac{\sigma_{sm}^2 \left(\sigma_1^2(\sigma_{uv}^2 \mu_{sm} + g_{x_t} Y_t) + g_{x_t} H_m^* (B_t \sigma_{uv}^2 - G\sigma_v^2 Y_t)\right)}{\sigma_1^2 \sigma_2^2 + g_{x_t}^2 \sigma_{sm}^2 \sigma_{uv}^2 |H_m|^2}, \\ B_1 &= \frac{\sigma_1^2 \sigma_{sm}^2 \sigma_{uv}^2}{\sigma_1^2 \sigma_2^2 + g_{x_t} \sigma_{sm}^2 \sigma_{uv}^2 |H_m|^2}, \\ A_2 &= g_{x_t} H_m \frac{\sigma_{uv}^2 \mu_{sm} + g_{x_t} \sigma_{sm}^2 Y_t}{\sigma_2^2} + \frac{G\sigma_v^2 Y_t}{\sigma_{uv}^2}, \\ B_2 &= \sigma_1^2 + g_{x_t} |H_m|^2 \frac{\sigma_{sm}^2 \sigma_{uv}^2}{\sigma_2^2}, \\ \sigma_{uv}^2 &= \sigma_u^2 + \sigma_v^2, \qquad \sigma_1^2 = \sigma_w^2 + \frac{|G|^2 \sigma_u^2 \sigma_v^2}{\sigma_{uv}^2}, \\ \sigma_2^2 &= \sigma_{uv}^2 + g_{x_t}^2 \sigma_{sm}^2, \qquad H_m = H - G\frac{\sigma_v^2}{\sigma_{uv}^2}. \end{aligned} \quad (11)$$

Now we can calculate the posterior of $\tilde{X}_t$ as

$$p(\tilde{X}_t|Y_t, B_t, S_t = 1, M_t = m) = \frac{p(\tilde{X}_t, Y_t, B_t, S_t{=}1, M_t{=}m)}{\int_{\tilde{X}_t} p(\tilde{X}_t, Y_t, B_t, S_t{=}1, M_t{=}m) d\tilde{X}_t} \sim N(\tilde{X}_t; A_1, B_1). \quad (12)$$

Furthermore, $p(\tilde{X}_t|Y_t, B_t, S_t = 0, M_t = 0)$ may be obtained by replacing σ^2_{sm} by σ^2_{sil} in the above equation.

3.1. Posteriors of S_t and M_t

To calculate the posteriors of S_t and M_t, we first compute the following joint distribution:

$$p(Y_t, B_t, S_t, M_t) = \int_{\tilde{X}_t} p(\tilde{X}_t, Y_t, B_t, S_t{=}s, M_t{=}m) d\tilde{X}_t \sim N(B_t; A_2, B_2) N(Y_t; g_{x_t}\mu_{sm}, \sigma_1^2) p(M_t|S_t) p(S_t) \quad (13)$$

Now the posteriors can be obtained as

$$p(M_t{=}m|Y_t, B_t, S_t{=}i) \propto p(Y_t, B_t, S_t{=}i, M_t{=}m),$$
$$p(S_t{=}i|Y_t, B_t) \propto \sum_m p(Y_t, B_t, S_t{=}i, M_t{=}m). \quad (14)$$

Up to now, we treat each of the frequency components independently. As explained previously, both S_t and M_t are defined over each frame across all frequency bins. Therefore, we should aggregate the likelihoods due to individual components to obtain a single most likely estimate for S_t and M_t. Thus the above equation may be rewritten as

$$p(Y_t^f, B_t^f, S_t, M_t) \sim L_1^f \; L_2^f \; p(M_t|S_t) p(S_t) \quad (15)$$

with $L_1^f = N(B_t^f; A_2^f, B_2^f)$, $L_2^f = N\big(Y_t^f; g_{x_t}\mu_{sm}^f, (\sigma_1^2)^f\big)$, where the exponent f represents the f^{th} frequency component. Finally, the likelihoods for a state are given by

$$L(M_t = m|Y_t, B_t, S_t = i) = p(S_t = i) p(M_t = m|S_t = i) \prod_{all\ f} L_1^f \; L_2^f \; . \quad (16)$$

4. ESTIMATING THE GAIN g_{x_t}

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

$$q(f) = p(\tilde{X}_t^f, Y_t^f, B_t^f, S_t, M_t) \quad (17)$$

which is given by equation (10), and let the overall joint log likelihood be

$$F = \log \prod_{all\ f} q(f) = \sum_{all\ f} \log q(f) \; . \quad (18)$$

The E-step essentially consists in estimating the most-likely value of $\tilde{X}_t$ given the current estimate of g_{x_t}, i.e., $\hat{\tilde{X}}_t = E\big(p(\tilde{X}_t|Y_t, B_t, g_{x_t})\big)$, where $E(.)$ is the expectation operator and $p(\tilde{X}_t|Y_t, B_t, g_{x_t})$ is given by equation (8). The M-step involves maximizing the objective function F w.r.t. g_{x_t}. This yields

$$g_{x_t} = \frac{\sum_{all\ f} \big[\big(Y_t^* \tilde{X}_t + Y_t \tilde{X}_t^*\big)\sigma_w^2 + C\sigma_v^2\big]}{\sum_{all\ f} \big[|\tilde{X}_t|^2\sigma_w^2 + |H - G|^2|\tilde{X}_t|^2\sigma_v^2\big]}, \quad (19)$$

where

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

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

5. EXPERIMENTAL RESULTS

5.1. Setups

We have recorded a number of utterances by four different speakers using the air-and-bone conductive microphone in various environments including cafeteria (ambient noise level 85 dBc) and office with an interfering speaker in the background. It is important to note that the utterances are corrupted by real-world noise, which implies that we do not have the ground-truth utterances. Each of the utterances were processed using the above framework to obtain an estimate of the clean speech signals.

The transfer functions H and G were estimated as explained in [11]. An estimate of the variances was obtained by using the speech detector proposed in [12]. Teethclacks in the bone channel were removed using the algorithm proposed in [11].

5.2. Propagating the prior of S_t

The enhancement process starts off with both $S_t = 0$ and $S_t = 1$ being equally likely. In order to enforce smoothness in the state estimates we use the following state dynamics:

$$p(S_t = 1) = \frac{0.5 + p(S_{t-1} = 1|Y_{t-1}, B_{t-1})}{2}, \quad (20)$$

and $p(S_t = 0) = 1 - p(S_t = 1)$. This introduces some bias towards the state of the previous frame, making frame-to-frame transition smoother.

5.3. Results

For our applications, we are more interested in perceptual quality than speech recognition. To measure the quality, we conducted mean opinion score (MOS) [13] comparative evaluations. Table 1 shows the score criteria.

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)

Table 2. MOS Results.

Original	SG	MG (Ω_1)	MG (Ω_2)
2.5833	3.0361	3.7583	3.6194

In order to gauge 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. The testing set consists of five noisy utterances recorded in a cafeteria with 85 dBc noise using the male speaker in both models.

Each noisy utterance in the test set was process in 3 different ways: a) SG: the algorithm described in [12] which uses a single Gaussian for the speech model, b) MG (Ω_1): the proposed mixture Gaussian model trained with one speaker and c) MG (Ω_2): the proposed mixture Gaussian model trained with four speakers. Therefore, together with the original utterance, there is a set of 4 utterances for each noisy utterance. There were a total of 17 participants 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. Table 2 shows the results of the MOS tests.

It can be seen that all the processed utterances outperform the original noisy ones. In addition, the proposed speech model outperforms our previously proposed algorithm, and it is not surprising that the model built using the same (single) speaker in both training and testing sets performs the best. However, the multi-speaker model Ω_2 only performs slightly worse than the single speaker model. This suggests that our proposed magnitude-normalized speech model is able to generalize fairly well.

6. CONCLUSION AND FUTURE WORK

In this paper we have proposed to use a mixture Gaussian speech model built from magnitude-normalized complex spectra for speech enhancement. We have also shown how the proposed mixture Gaussian model can be used in the context of speech enhancement with an air-and-bone conductive microphone. Substantial improvement have 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.

For our future work, we plan to collect a large amount of data with more speakers in order to build better speech models. we are also planning to introduce dynamics on other variables such as $\tilde{X}_t$ and X_t which may lead to better estimates of the clean speech signal. Finally, we are working on a system where the noise can estimated recursively.

7. REFERENCES

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

[2] R. J. McAulay and M. L. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Trans. Acoustics, Speech, Signal Processing*, vol. 28, no. 2, pp. 137–145, 1980.

[3] H. Drucker, "Speech processing in a high ambient noise environment," *IEEE Trans. Audio Electroacoust*, vol. 16, no. 2, pp. 165–168, 1968.

[4] J.S. Lim and A.V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proc. IEEE*, vol. 67, no. 12, pp. 1586–1604, 1979.

[5] Y. Ephraim, "A bayesian estimation approach for speech enhancement using hidden markov models," *IEEE Trans. Signal Processing*, vol. 40, no. 4, pp. 725–735, 1992.

[6] Y. Ephraim, H. Lev-Ari, and W. J. J. Roberts, "A brief survey of speech enhancement," in *CRC Electronic Engineering Handbook*. CRC Press, Feb. 2005.

[7] J. Wu, J. Droppo, L. Deng, and A. Acero, "A noise-robust asr front-end using wiener filter constructed from mmse estimation of clean speech and noise,," in *Proc. ASRU*, Dec. 2003.

[8] Y. Zheng, Z. Liu, Z. Zhang, M. Sinclair, J. Droppo, L. Deng, A. Acero, and X. Huang, "Air- and bone-conductive integrated microphones for robust speech detection and enhancement," in *Proc. ASRU*, Dec. 2003, pp. 249–254.

[9] Z. Zhang, Z. Liu, M. Sinclair, A. Acero, L. Deng, J. Droppo, X. Huang, and Y. Zheng, "Multi-sensory microphones for robust speech detection, enhancement and recognition," in *Proc. ICASSP*, May 2004, vol. 3, pp. 781–784.

[10] Z. Liu, Z. Zhang, A. Acero, J. Droppo, and X. Huang, "Direct filtering for air-and bone-conductive microphones," in *Proc. MMSP*, Sept. 2004, pp. 363–366.

[11] Z. Liu, A. Subramanya, Z. Zhang, J. Droppo, and A. Acero, "Leakage model and teeth clack removal for air- and bone-conductive integrated microphones," in *Proc. ICASSP*, Mar. 2005, vol. 1, pp. 1093–1096.

[12] A. Subramanya, Z. Zhang, Z. Liu, J. Droppo, and A. Acero, "A graphical model for multi-sensory speech processing in air-and-bone conductive microphones," in *Proc. Eurospeech*, Sept. 2005.

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