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

Amarnag Subramanya[‡], Zhengyou Zhang, Zicheng Liu and Alex Acero[†]

[‡]SSLI Lab, University of Washington, Seattle, WA - 98104
[†]Microsoft Research, One Microsoft Way, Redmond, WA - 98052.
asubram@ee.washington.edu,{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.

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). An efficient speech enhancement technique requires explicit and accurate statistical models for the speech signal and noise process.

Quatieri [1] provides a description of various speech enhancement techniques. Although, the above algorithms have had success in dealing with stationary noise types, they fail in the presence of non-stationary noise. Further, some of these techniques often assume, implicitly or explicitly, a single Gaussian distribution on speech signals which is a poor model as a result of the large variation in speech. Drucker [2] proposed a system using five states representing fricative, stop, vowel, glide, and nasal speech sounds. The system, however, was simulated by hand-switching between the speech states. Attempts have also been made to model state changes over time: Lim and Oppenheim [3] model the short-term speech and noise signals as an autoregressive process. Ephraim [4] 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 [5].

As mentioned earlier, while we have seen many successes in dealing with stationary noise types, enhancement in the presence of non-stationary background noise (such as interfering speech) is still an open problem. To tackle this problem, we have developed a novel hardware solution [6, 7] that makes use of an inexpensive bone-conductive microphone in addition to the regular air-conductive microphone. The bone sensor captures the sounds uttered by the speaker but transmitted via the bone and tissues in the speaker's head and is thus relatively noise-free. 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. For a detailed discussion about the bone sensor the reader is refered to [7].

In [6], we proposed an algorithm based on the SPLICE technique for speech enhancement. In the same work, a speech detector based on the energy in the bone channel was proposed. In [8], we proposed an algorithm called direct filtering (DF) based on learning mappings in a maximum likelihood framework. However, one drawback with the DF algorithm is the absence of a strong speech model, which can lead to distortion in the enhanced signal. In [9], we extended the DF algorithm to deal with the environmental noise leakage into the bone sensor, and the teethclack problem. The success of all the above algorithms, requires accurate speech activity detection to estimate noise and speech statistics. Making use of the energy in the bone sensor [6] for this task leads to two problems: 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. For a detailed decription of our previous work, the reader is refered to [11].

To address some of the above problems, in [10] we proposed an algorithm that takes into account the correlation between the two channels for speech detection and also incorporates a speech model thereby introducing robustness into the system. However, the proposed algorithm had two shortcomings: a) speech was modeled using a single Gaussian and b) the system was static, i.e., there was no information transfer across frames. In this paper, we describe some of our efforts to overcome the above problems.

2. MAGNITUDE-NORMALIZED COMPLEX SPECTRUM-BASED SPEECH MODEL

In Bayesian statistics, prior information 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. However, building accurate speech models is extremely hard on accout of the large variability of human speech due to a number of factors such as speaker change, changes due to loudness, intonation and stress. One way to deal with issues related to changes in loudness and recording device gains is to work in the mel-cepstral domain, where they only effect the first cepstral coefficient which may be neglected. However, such models have the disadvantage that they do not encode any phase information.

2.1. Model Definition

We work in the complex spectral domain as we are interested in estimating both the magnitude and phase of the clean speech signal. However, in the complex spectral domain, the variations due to loudness cannot be easily handled. 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}. We discuss an iterative approach to estimating the gain in section 4. Further, to add robustness to the model, we neglect the DC and Nyquist terms while building the model. Gain normalization has been studied in the past (for example in [13]). One important distinction between the work in [13] and the current algorithm is that here we are normalizing the clean speech signal rather than the noise.

2.2. Training

In order to train the speech model, we collected data from a large number of speakers in a clean environment. The speech frames were then extracted using a simple energy based speech detector. The resulting speech frames were then energy normalized as explained in the previous subsection. We trained 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 used $d(\tilde{X}_i, \tilde{X}_j) = \|(\log|\tilde{X}_i| - \log|\tilde{X}_j|)\|$ as the distance measure for clustering the frames, where $\log \tilde{X}$ denotes that the log operation is applied to each element of $\tilde{X}$. It should be noted here that although the above distance measure is in the log-spectral domain, the means and variances for the speech model were obtained in the normalized-complex-spectral domain.

2.3. Experimental Results

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. The above models were then used to compute the likelihoods for an utterance outside the training set but recorded using a device with similar gain setting as the training set. The aggregated likelihoods (across all frequency components) are shown in figure 1. 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. Further, the above experiment is the best case scenario for ω_2. Note 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 competing model.

Figure 2 shows the spectrogram of four clusters obtained as a result of the clustering algorithm described above. It can be seen that one cluster models fricatives, and the others model various kinds of vowels.

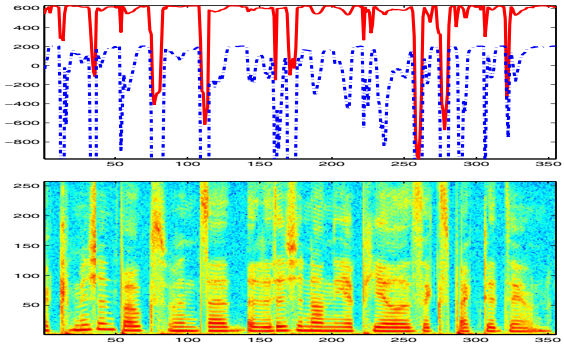

Fig. 1. Comparison of likelihoods with (solid, red lines) and without (dotted, blue lines) magnitude normalization. The second figure depicts the spectrogram.

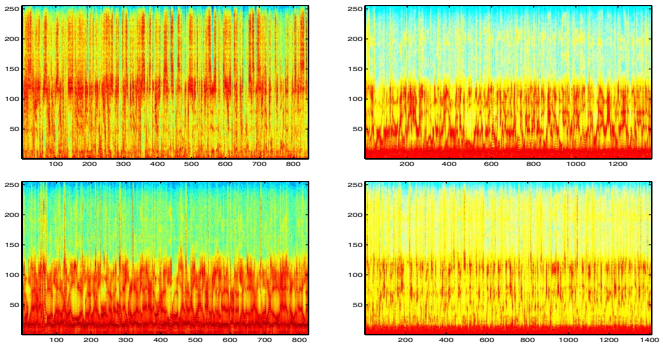

Fig. 2. Clustering results

3. MODEL FOR SPEECH ENHANCEMENT

We are now applying the speech model proposed in the last section to speech enhancement in an air- and bone-conductive integrated microphone headset [6, 7]. Due to space limitation, we only present a consice version of the inference math; for a detailed derivation, the reader is refered to [11]. 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 hamming windowed version of the signal samples. The physical process may be modeled as shown in Figure 3.

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, 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 background noise that leaks 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** for a given frame.

We make the following assumptions in the model: Background noise is modeled as $p(V_t) \sim N(0, \sigma_v^2)$; Sensor noise in the air mi-

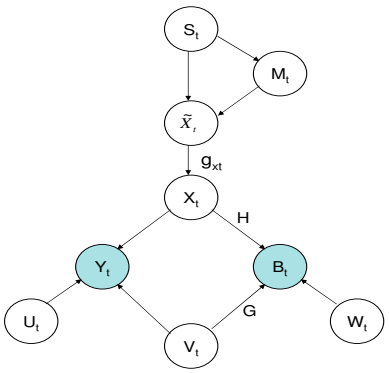

Fig. 3. The graphical model incorporating the proposed speech model.

crophone channel is modeled using $p(U_t) \sim N(0, \sigma_u^2)$; Sensor noise in the bone channel is modeled with $p(W_t) \sim N(0, \sigma_w^2)$; Speech is modeled using a mixture of Gaussians (MG),

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

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, $p(\tilde{X}_t|S_t = 1) \sim \frac{1}{M}\sum_{m=1}^{M} N(\mu_{sm}, \sigma_{sm}^2)$. For mathematical tractability we assume $p(\tilde{X}_t|X_t) \sim \delta(X_t, g_{x_t}\tilde{X}_t)$, a delta function with parameter g_{x_t}.

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, we are interested in estimating $p(\tilde{X}_t|Y_t, B_t) = \sum_{s,m} p(\tilde{X}_t, S_t = s, M_t = m|Y_t, B_t)$. Let us first consider

$$p(\tilde{X}_t, Y_t, B_t, S_t = s, M_t = m) =$$
$$\int_{V_t}\int_{U_t}\int_{W_t} p(Y_t, B_t, \tilde{X}_t, V_t, S_t, M_t, U_t, W_t)\, dU_t dW_t dV_t \quad (3)$$

After some algebra we get

$$p(\tilde{X}_t, Y_t, B_t, S_t = s, M_t = m) \sim N(\tilde{X}_t; A_1, B_1)$$
$$N(B_t; A_2, B_2)N(Y_t; g_{x_t}\mu_{sm}, \sigma_1^2)p(M_t|S_t)p(S_t) \quad (4)$$

where

$$A_1 = \frac{\sigma_{sm}^2\left(\sigma_1^2(\sigma_{uv}^2\mu_{sm} + g_{x_t}Y_t) + g_{x_t}H_m^*(B_t\sigma_{uv}^2 - G\sigma_v^2 Y_t)\right)}{\sigma_1^2\sigma_2^2 + g_{x_t}^2\sigma_{sm}^2\sigma_{uv}^2|H_m|^2},$$
$$B_1 = \frac{\sigma_1^2\sigma_{sm}^2\sigma_{uv}^2}{\sigma_1^2\sigma_2^2 + g_{x_t}\sigma_{sm}^2\sigma_{uv}^2|H_m|^2}, \sigma_1^2 = \sigma_w^2 + \frac{|G|^2\sigma_u^2\sigma_v^2}{\sigma_{uv}^2},$$
$$A_2 = g_{x_t}H_m\frac{\sigma_{uv}^2\mu_{sm} + g_{x_t}\sigma_{sm}^2 Y_t}{\sigma_2^2} + \frac{G\sigma_v^2 Y_t}{\sigma_{uv}^2}, \frac{\sigma_v^2}{\sigma_{uv}^2},$$
$$B_2 = \sigma_1^2 + g_{x_t}|H_m|^2\frac{\sigma_{sm}^2\sigma_{uv}^2}{\sigma_2^2}, \ \sigma_{uv}^2 = \sigma_u^2 + \sigma_v^2,$$
$$H_m = H - G, \ \sigma_2^2 = \sigma_{uv}^2 + g_{x_t}^2\sigma_{sm}^2. \quad (5)$$

It is not difficult to show that the posterior of $\tilde{X}_t$, $p(\tilde{X}_t|Y_t, B_t, S_t = 1, M_t = m) \propto N(\tilde{X}_t; A_1, B_1)$. In a similar vein, $p(\tilde{X}_t|Y_t, B_t, S_t = 0, M_t = 0)$ may be obtained by replacing σ_{sm}^2 by σ_{sil}^2 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:

$$p(Y_t, B_t, S_t, M_t) = \int_{\tilde{X}_t} p(\tilde{X}_t, Y_t, B_t, S_t{=}s, M_t{=}m)d\tilde{X}_t$$
$$\sim N(B_t; A_2, B_2)N(Y_t; g_{x_t}\mu_{sm}, \sigma_1^2)p(M_t|S_t)p(S_t) \quad (6)$$

Further, it can be seen that $p(M_t = m|Y_t, B_t, S_t = i) \propto p(Y_t, B_t, S_t = i, M_t = m)$ and $p(S_t = i|Y_t, B_t) \propto \sum_m p(Y_t, B_t, S_t = i, M_t = m)$. 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

$$p(Y_t^f, B_t^f, S_t, M_t) \sim L_1^f\ L_2^f\ p(M_t|S_t)p(S_t) \quad (7)$$

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

$$L(M_t = m|Y_t, B_t, S_t = i) =$$
$$p(S_t = i)p(M_t = m|S_t = i)\prod_{all\ f} L_1^f\ L_2^f\ . \quad (8)$$

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 $q(f) = p(\tilde{X}_t^f, Y_t^f, B_t^f, S_t, M_t)$ which is given by equation (4), and let the overall joint log likelihood be $F = \log\prod_{all\ f} q(f) = \sum_{all\ f}\log q(f)$. 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 $p(\tilde{X}_t|Y_t, B_t, g_{x_t})$ was obtained in the previous section. The M-step involves maximizing the objective function F w.r.t. g_{x_t} which yields

$$g_{x_t} = \frac{\sum_{all\ f}\left[\left(Y_t^*\tilde{X}_t + Y_t\tilde{X}_t^*\right)\sigma_w^2 + C\sigma_v^2\right]}{\sum_{all\ f}\left[|\tilde{X}_t|^2\sigma_w^2 + |H - G|^2|\tilde{X}_t|^2\sigma_v^2\right]}, \quad (9)$$

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 recorded utterances from a number of 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

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

the utterances are corrupted by real-world noise. 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 [9]. An estimate of the variances was obtained by using the speech detector proposed in [10]. Teethclacks in the bone channel were removed using the algorithm proposed in [9].

5.2. Propagating the prior of S_t

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

$$p(S_t = 1) = \frac{0.5 + p(S_{t-1} = 1|Y_{t-1}, B_{t-1})}{2}, \qquad (10)$$

and $p(S_t = 0) = 1 - p(S_t = 1)$. This introduces a bias towards the previous value of the state variable thereby making frame-to-frame transitions 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) [12] comparative evaluations. Table 1 shows the score criteria.

In order to gauge the sensitivity of the speech model to speakers, we trained two 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 six different speakers (three male and three female). The speaker in Ω_1 is one of the male speakers in Ω_2. The testing set consisted of ten noisy utterances (both cafeteria and office environments) recorded using speaker in Ω_1.

Each noisy utterance in the test set was processed in 3 different ways: a) SG: the algorithm described in [10] (single Gaussian for the speech model), b) MG (Ω_1): the proposed model trained with one speaker and c) MG (Ω_2): the proposed model trained with fifteen different speakers. This resulted in 3 processed utterances for each corrupted utterance. There were a total of 12 participants in the MOS evaluations. The evaluators were presented utterances in a random intra and inter set ordering. Further, the evaluators 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 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. In addition, we plan on learning the dynamics on the state variable. We also plan 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 be estimated recursively.

7. REFERENCES

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

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

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

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] A. Subramanya, Z. Zhang, Z. Liu, A Acero, "Speech Modeling with Magnitude-Normalized Complex Spectra and its Application to Multisensory Speech Enhancement," Microsoft Research Technical Report MSR-TR-2005-126, Sept., 2005.

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

[13] D. Y. Zhao and W. B. Kleijn "On Noise Gain Estimation for HMM-based Speech Enhancement." in *Proc. Eurospeech*, Sept. 2005.