MULTICHANNEL SPEECH ENHANCEMENT BASED ON SPEECH SPECTRAL MAGNITUDE ESTIMATION USING GENERALIZED GAMMA PRIOR DISTRIBUTION

Tran Huy Dat[*]

Institute for Infocomm Research, Singapore

Kazuya Takeda, Fumitada Itakura

Nagoya University, Japan

ABSTRACT

We present multichannel speech enhancement method based on MAP speech spectral magnitude estimation using a generalized gamma model of speech prior distribution, where the model parameters are adapted from actual noisy speech in a frame-by-frame manner. The utilization of a more general prior distribution with its online estimation is shown to be effective for speech spectral estimation. We tested the proposed algorithm in an in-car speech database and obtained significant improvements on the speech recognition performance, particularly under nonstationary noise conditions such as music, air-conditioner and open window.

1. INTRODUCTION

Multichannel speech enhancement systems are widely used for hands-free communication and speech recognition. The most fundamental method is beamforming, where the positions of the target and noise sources are assumed to be apart or known in advance. Recently, statistical speech enhancement methods, extended from single-channel approaches, have been investigated [1]-[2] and have shown better results than beamforming. These methods model and estimate the distributions of speech and noise spectra, assuming the Gaussian model and then statistical estimators (MMSE or MAP) are employed. However, the Gaussian model, yielding an independence between magnitude and phase, is unnatural for speech signal [3]. Moreover, the noise field was assumed to be incoherent [1] which is also untypical in real conditions. Previously, we presented a generalized gamma model of speech prior distribution [3], where the distribution parameters are adapted from actual noisy speech. We have shown that, this method provides more accurate modeling of speech prior distribution and it improved the performance of single channel speech enhancement in terms of both sound quality and speech recognition. In this study, we extend this method for a multichannel approach. Furthermore, we develop a method to adapt the model parameters in a frame-by-frame manner. The motivation for this extension is that the multichannel systems, which can better distinguish the target signal from interfering noises, should improve the performances of the parameter estimation and speech enhancement system. The paper is organized as follows. In the next section, we introduce the multichannel model and assumptions used in this study. In Sec.3, we derive the multichannel MAP speech spectral magnitude estimation using the proposed model. In Sec.4, we describe algorithms to estimate the model parameters in a frame-by-frame manner. In Sec.5, we report the experimental evaluation.

2. MULTICHANNEL MODEL AND ASSUMPTIONS

Consider the additive model of multichannel signals in the STDFT domain

$$\mathbf{X}_l\left(n,k\right) = \mathbf{H}_l\left(n,k\right) S\left(n,k\right) + \mathbf{N}_l\left(n,k\right), \qquad (1)$$

where $\mathbf{X} = \left[X_1, ..., X_D\right]^T$, $\mathbf{N} = \left[N_1, ..., N_D\right]^T$ are the vectors of the complex spectra of noisy and noise signals, $\mathbf{H} = \left[H_1, ..., H_D\right]^T$ is the vector of transfer functions, S is clean speech spectrum and $l = 1 : D$ is the microphone index. Couple (n,k) denotes the time-frequency index and will be omitted in Secs.2 and 3. The differences between our model and those proposed in [1]-[2] are as follows.

-Due to a possible change in the positions of speakers, we do not assume transfer functions to be constant in each frequency bin.

-The noise is assumed to be spatially coherent and the noise spectral components follow the zero-mean independent identical multi-variable Gaussian distribution with a full covariance matrix

$$\begin{array}{l} p\left(\mathbf{N}_R\right) = \frac{1}{2\pi^{D/2}\det(2\mathbf{C_n})^{1/2}} \exp\left\{-\mathbf{N}_R^T\mathbf{C_n}^{-1}\mathbf{N}_R\right\}, \\ p\left(\mathbf{N}_I\right) = \frac{1}{2\pi^{D/2}\det(2\mathbf{C_n})^{1/2}} \exp\left\{-\mathbf{N}_I^T\mathbf{C_n}^{-1}\mathbf{N}_I\right\}. \end{array} \qquad (2)$$

$$p\left(\mathbf{N}_R, \mathbf{N}_I\right) = p\left(\mathbf{N}_R\right) p\left(\mathbf{N}_I\right), \qquad (3)$$

where $(.)_R$, and $(.)_I$ denote the real and imaginary parts of the complex spectrum, respectively and $\mathbf{C_n}$ is half of the covariance matrix (real and symmetrical).

-Speech spectral magnitude $|S|$ follows a generalized gamma distribution given as

$$p\left(|S|\right) = \frac{b^a}{\Gamma\left(a\right)\sigma_S}\left(\frac{|S|}{\sigma_S}\right)^{La-1} \exp\left[-b\left(\frac{|S|}{\sigma_S}\right)^L\right], \quad (4)$$

[*] The author was with Nagoya University, Japan.

1-4244-0469-X/06/$20.00 ©2006 IEEE

where σ_S^2 denotes the variance of speech spectrum. (a, b, L) are distribution parameters but the system has two remaining free parameters due to the normalization $\left\langle |S|^2 \right\rangle = \sigma_S^2$, where $\langle . \rangle$ denotes the expectation. This distribution is a superset of conventional models, including Gaussian, generalized Gaussian and gamma distributions and was used in our previous study for single-channel approaches [3].

3. MULTICHANNEL MAP SPEECH SPECTRAL MAGNITUDE ESTIMATION USING GENERALIZED GAMMA MODEL

For the proposed generalized gamma model, the MAP speech spectral magnitude estimation is preferable to use due to its simplicity in implementation and the effectiveness [3]. Unlike Lotter et al [1], we derive the estimation for the case of a spatially coherent noise using generalized gamma model of speech prior distribution. The multichannel MAP estimation equation $\left|\hat{S}\right| = \arg\max\limits_{|S|} \left[p\left(|S| \,|\mathbf{X}\right)\right]$ can be expressed as

$$\frac{\partial}{\partial |S|} \left\{\log\left[p\left(\mathbf{X}|\,|S|\right)\right] + \log\left[p\left(|S|\right)\right]\right\} = 0. \tag{5}$$

3.1. Derivation of multichannel Rician distribution

From (3), $p\left(\mathbf{X}|\,|S|\right)$ can be factorized as

$$p\left(\mathbf{X}|S\right) = p\left(\mathbf{X}_R|S_R, S_I\right) p\left(\mathbf{X}_I|S_R, S_I\right). \tag{6}$$

Note that here, we consider the transfer functions as deterministic variables, which will be estimated from observations. Two terms in right side of (6) can be denoted using (2) and (3)

$$\begin{aligned} p\left(\mathbf{X}_R|S_R, S_I\right) &= \tfrac{1}{2\pi^{D/2}\det(\mathbf{C_n})^{1/2}} \exp\left\{-\mathbf{Y}_R^T\mathbf{C_n^{-1}}\mathbf{Y}_R\right\}, \\ p\left(\mathbf{X}_I|S_R, S_I\right) &= \tfrac{1}{2\pi^{D/2}\det(\mathbf{C_n})^{1/2}} \exp\left\{-\mathbf{Y}_I^T\mathbf{C_n^{-1}}\mathbf{Y}_I\right\}, \end{aligned} \tag{7}$$

where

$$\begin{aligned} \mathbf{Y}_R &= \mathbf{X}_R - \mathbf{H}_R S_R + \mathbf{H}_I S_I, \\ \mathbf{Y}_I &= \mathbf{X}_I - \mathbf{H}_R S_I - \mathbf{H}_I S_R. \end{aligned} \tag{8}$$

Substituting (7) and (8) into (6) yields the conditional distribution of the complex spectrum of noisy speech expressed as

$$p\left(\mathbf{X}|S\right) = Q\left(\mathbf{X}\right)\exp\left\{-\left[\begin{array}{l} \bar{\mathbf{H}}^T\mathbf{C_n^{-1}}\mathbf{H}\left(S_R^2 + S_I^2\right) - \\ -2Re\left(\bar{\mathbf{H}}^T\mathbf{C_n^{-1}}\mathbf{X}\right)S_R - \\ -2Im\left(\bar{\mathbf{H}}^T\mathbf{C_n^{-1}}\mathbf{X}\right)S_I \end{array}\right]\right\}, \tag{9}$$

where $Q\left(\mathbf{X}\right)$ is independent of S and this term will be reduced with further estimation. Since $\mathbf{C_n}$ is real and symmetrical the term $\bar{\mathbf{H}}^T\mathbf{C_n^{-1}}\mathbf{H}$ is real. The conditional distribution (9) can be transformed into magnitude and phase, using Jacobian transform, yielding

$$p\left(\mathbf{X}|\varphi_{\mathrm{S}}, |S|\right) \triangleq \exp\left\{\begin{array}{l} -\left(\bar{\mathbf{H}}^T\mathbf{C_n^{-1}}\mathbf{H}\right)|\mathbf{S}|^2 - \\ -2\,|\mathbf{S}|\left|\bar{\mathbf{H}}^T\mathbf{C_n^{-1}}\mathbf{X}\right|^2\cos\left(\varphi_{\mathbf{X}} - \varphi_S\right) \end{array}\right\} \tag{10}$$

where φ_S denotes the phase of the clean speech spectrum and φ_X- the phase of $\bar{\mathbf{H}}^T\mathbf{C_n^{-1}}\mathbf{X}$. Integrating (10) over φ_S, we obtain the conditional distribution of noisy speech magnitude as a multichannel version of the Rician distribution [3].

$$p\left(\mathbf{X}|\,|S|\right) \triangleq \exp\left(-\bar{\mathbf{H}}^T\mathbf{C_n^{-1}}\mathbf{H}\,|\mathrm{S}|^2\right) I_0\left(2\,|\mathrm{S}|\left|\bar{\mathbf{H}}^T\mathbf{C_n^{-1}}\mathbf{X}\right|\right). \tag{11}$$

Here, I_0 is the Bessel function of the first kind, which can be approximated by

$$I_0(x) \approx \frac{1}{\sqrt{2\pi x}}e^x, x > 0. \tag{12}$$

The first term in (5) is then derived as

$$\frac{\partial}{\partial |S|}\left\{\log\left[p\left(\mathbf{X}|\,|S|\right)\right]\right\} = -\frac{2\,|S|}{U_n^2} - \frac{1}{2S} + \frac{\frac{\left|\bar{\mathbf{H}}^T\mathbf{C_n^{-1}}\mathbf{X}\right|}{\bar{\mathbf{H}}^T\mathbf{C_n^{-1}}\mathbf{H}}}{U_n^2}, \tag{13}$$

where

$$U_n^2 = \frac{1}{\bar{\mathbf{H}}^T\mathbf{C_n^{-1}}\mathbf{H}}. \tag{14}$$

3.2. Gain function

From (4), the second term in (5) is given by

$$\frac{\partial}{\partial |S|}\left\{\log\left[p\left(|S|\right)\right]\right\} = \frac{(La-1)}{|S|} - Lb\frac{|S|}{\sigma_S^{L-1}}. \tag{15}$$

Substituting (13) and (15) into (5) yields the estimation equation, which generally can be solved by the Newton-Raphson method. However, in this study, we consider the solution for the case of $L = 2$, which yields a closed-form solution and therefore is suitable in the implementation. The estimation equation is then derived as

$$-G^2 + \frac{G}{\left(1 + \frac{b}{\xi}\right)} + \frac{4a-3}{G} = 0. \tag{16}$$

Here the gain function is

$$G = \frac{\bar{\mathbf{H}}^T\mathbf{C_n^{-1}}\mathbf{H}}{\left|\bar{\mathbf{H}}^T\mathbf{C_n^{-1}}\mathbf{X}\right|}|S| \tag{17}$$

and the generalized priori and posteriori SNR are

$$\xi = \frac{\sigma_S^2}{U_n^2}, \gamma = \left|\bar{\mathbf{H}}^T\mathbf{C_n^{-1}}\mathbf{X}\right|^2. \tag{18}$$

4. ONLINE PARAMETER ESTIMATION

4.1. Noise covariance and transfer functions

The noise covariance is initially estimated in each frequency bin using the first 250ms of observations. Then it is recursively updated, using voice activity detection (VAD),

$$\begin{aligned} &\mathbf{C_n}(n,k) = \\ &\left\{\begin{array}{ll} \alpha\mathbf{C_n}(n-1,k) + (1-\alpha)\,Re\left[\bar{\mathbf{X}}^T(n,k)\,\mathbf{X}(n,k)\right] & \mathrm{D1} \\ \mathbf{C_n}(n-1,k) & \mathrm{D0} \end{array}\right. \end{aligned} \tag{19}$$

where $D1$ and $D0$ are hypotheses of speech presence and absence, α is a smoothing coefficient. Differently from those methods proposed in [1]-[2], we estimate the transfer function via priori SNR (i.e., short-term statistics). Not losing the generality, we assume

$$\mathbf{H}_1 = 1. \tag{20}$$

The magnitude of the transfer function at microphone i is determined and smoothed as

$$|\mathbf{H}_i \widehat{(n,k)}| = \\ \begin{cases} \chi\left|\mathbf{H}_i\left(n-1,k\right)\right| + \left(1-\chi\right)\sqrt{\frac{\xi_i(n,k)}{\xi_1(n,k)}\frac{\sigma_{n_1}^2(n,k)}{\sigma_{n_i}^2(n,k)}} & \text{D1} \\ \left|\mathbf{H}_i\left(n-1,k\right)\right|, & \text{D0} \end{cases} \tag{21}$$

where χ is a smoothing coefficient. The priori SNR in the i-channel $\xi_i = \frac{|H_i|^2\sigma_S^2}{\sigma_{n_i}^2}$ is estimated by the decision-directed method [4]. The phase of the transfer function is estimated using the covariance matrix relationship

$$\mathbf{C}_x = \bar{\mathbf{H}}^T\mathbf{H}\sigma_S^2 + \left(\mathbf{C}_n + j\mathbf{C}_n\right), \tag{22}$$

where $\mathbf{C}_x$ is the noisy speech covariance matrix, which is estimated as in (21) without the VAD.

$$\mathbf{C}_x\left(n,k\right) = \alpha\mathbf{C}_x\left(n-1,k\right) + \left(1-\alpha\right)\bar{\mathbf{X}}^T\left(n,k\right)\mathbf{X}\left(n,k\right) \tag{23}$$

Taking into account (20), the phase of the transfer function can be estimated using the first column of $\mathbf{C}_x$.

4.2. Online estimation of prior distribution parameter

The main point of the proposed model is that the modeled distribution parameters are determined from actual noisy speech. Unlike the single channel system [3], here we use cross-channel statistics to perform the estimation in a frame-by-frame manner. Denote the noisy speech power as

$$\left|X_i\right|^2 = \left|H_i\right|^2\left|S\right|^2 + \left|N_i\right|^2 + 2\left|H_i\right|\left|S\right|\left|N_i\right|\cos\left(\Delta\phi_i\right), \tag{24}$$

where $\Delta\phi_i$ is the phase difference, which is assumed to follow a uniform distribution [3]. Taking into account the independence of the phase and the magnitude of the noise spectrum, the cross-channel correlations of noisy speech power are expressed as

$$\left\langle\left|X_1\right|^2\left|X_i\right|^2\right\rangle = \left|H_i\right|^2\left\langle\left|S\right|^4\right\rangle + \left\langle\left|N_1\right|^2\left|N_i\right|^2\right\rangle + \\ + \left\langle\left|S\right|^2\right\rangle\left(\left|H_i\right|^2\left\langle\left|N_1\right|^2\right\rangle + \left\langle\left|N_i\right|^2\right\rangle + 4\left|H_i\right|\left|\left\langle N_1 N_i\right\rangle\right|\right), \tag{25}$$

$$\left\langle\left|X_1\right|^2\right\rangle\left\langle\left|X_i\right|^2\right\rangle = \left|H_i\right|^2\left(\left\langle\left|S\right|^2\right\rangle\right)^2 + \\ + \left\langle\left|N_1\right|^2\right\rangle\left\langle\left|N_i\right|^2\right\rangle + \left\langle\left|S\right|^2\right\rangle\left(\left|H_i\right|^2\left\langle\left|N_1\right|^2\right\rangle + \left\langle\left|N_i\right|^2\right\rangle\right). \tag{26}$$

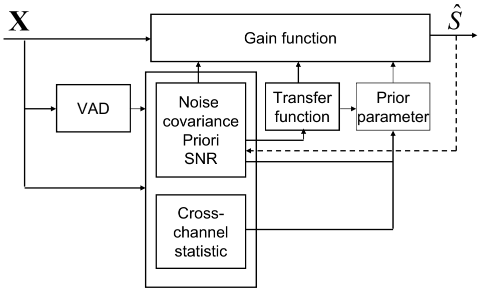

Fig. 1. Block-diagram processing

Since all components in the lower term of (25) and (26) are given from transfer function, noise covariance matrix and priori SNR estimations, the ratio of the fourth to the second-moments of $|S|$ can be determined (at each time-frequency index). This ratio is used to "match" the generalized gamma distribution parameters. For $L = 2$, yields [3]

$$\left\langle\left|S\right|^4\right\rangle = \frac{a\left(a+1\right)}{b^2}\left(\left\langle\left|S\right|^2\right\rangle\right)^2 = \left(1+\frac{1}{a}\right)\left(\left\langle\left|S\right|^2\right\rangle\right)^2. \tag{27}$$

Note that the normalization $\left\langle\left|S\right|^2\right\rangle = \sigma_S^2$ implies the relationship $a = b$. The terms $\left\langle\left|X_1\right|^2\left|X_i\right|^2\right\rangle$ and $\left\langle\left|N_1\right|^2\left|N_i\right|^2\right\rangle$ are estimated using recursive moving averages as in (19) and (23). Finally, we smooth the estimated parameter as

$$a\left(n,k\right) = \mu a\left(n-1,k\right) + \left(1-\mu\right)a\left(n,k\right), \tag{28}$$

where μ is a smoothing coefficient.

4.3. Voice activity detection

For VAD, we assume that the less noisy channel is known in advance. In addition to the conventional energy feature given from this channel, we use the cross-channel correlation coefficient ρ calculated from each frame to determine VAD.

$$\rho\left(n\right) = \frac{1}{D-1}\sum_{i=2}^{D}\sqrt{\frac{\left|\left\langle X_1\left(n\right)X_i\left(n\right)\right\rangle\right|}{\left\langle\left|X_1^2\left(n\right)\right|\right\rangle\left\langle\left|X_i^2\left(n\right)\right|\right\rangle}} \tag{29}$$

Here we assume that the channels have a higher correlation in the target signal durations. The resulting VAD is expressed as

$$VAD\left(n\right) = \begin{cases} 1 & if \quad \left|\Delta\left(n\right)\right| > \gamma_1 \text{ \& } \rho\left(\mathrm{n}\right) > \gamma_2 \\ 0 & otherwise \end{cases} \tag{30}$$

where γ_1 and γ_2 are constant boosting factors. Currently $\gamma_1 = 3$ and $\gamma_2 = 0.3$ are used. The energy distance Δ is the ratio of current frame energy $E_1\left(n\right)$ to the stored-in-memory noise energy and is calculated in decibels. Unlike VAD, the noise energy is updated when $\left|\Delta\left(n\right)\right| > \gamma_1$.

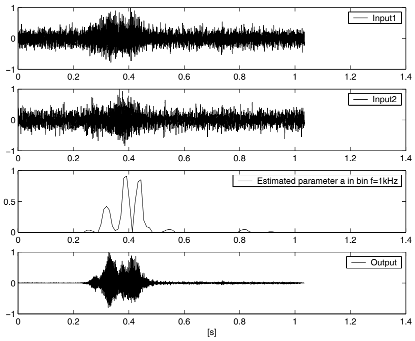

Fig. 2. Input 2-channel noisy signals (top two plots), example of estimated prior distribution parameter in bin f=1kHz (third plot) and output waveform (bottom)

5. EXPERIMENT

We evaluate the proposed algorithm in CIAIR in-car speech corpus [5]. In the implementation, 16-kHz sampling signals from two microphones, attached to the ceiling [5] were used. The Hamming window of length of 20ms with 50% overlap was applied in FFT. Figure 1 shows the block diagram of processing. Given the multichannel noisy signals, VAD is performed using (30). Then noise covariance, priori SNR and cross-channel statistics are estimated using recursive averages. The transfer function is estimated using (21) and (22). The prior parameters are estimated and updated using (25)-(28). Then the gain function is given by (16)-(18). Finally, the phase adding and overlap and add are used to resynthesize the enhanced sounds. Examples of waveforms is plotted in Figure 2. Note that, the smoothing coefficients are chosen by hearing the output sounds. Currently, $\alpha = \beta = \chi = 0.9$, and $\mu = 0.6$ are used. For reference, the Ephraim-Malah method (LSA), the single-channel version using generalized gamma modeling (GG) [3], and the multi-channel method based on psychoacoustic motivation (PA) [2] are also implemented. The overall results of speech recognition are shown in Figure 3. The proposed multichannel method is superior to others with approximately 16% improvement compared to the performance in the nearest microphone. The proposed method is particularly effective under nonstationary noise conditions. Table 1 shows the results for the cases of driving along an express way with a CD playing, high air-conditioner (AC) and open window. The superiority of proposed methods can be explained by as follows. Firstly, using the more general distribution with its online estimation improves the accuracy of prior distribution modeling which is realized in the MAP estimation. Secondly, multi-channel systems improve the performance of VAD and parameter estimation. The online parameter estimation can also be considered as an optimization of the gain function, which controls the trade-off between noise reduction and distortion, resulting in the best speech recognition performance and the quality of enhanced sound which is confirmed by the listening test.

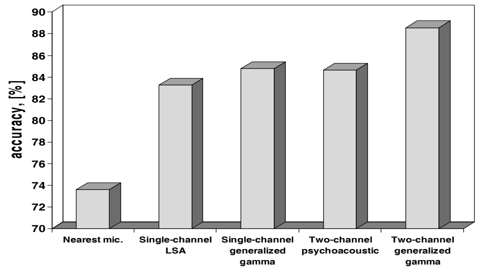

Fig. 3. Speech recognition overall performances on CIAIR in-car database

Table 1. Speech recognition rate on expressway under several driving conditions [%]

Method	Nearest mic	1-ch LSA	1-ch GG	2-ch PA	2-ch GG
CD playing	82.27	82.61	85.62	83.96	92.98
High AC	51.00	87.33	89.00	88.67	91.67
Window open	42.67	76.67	78.33	77.33	86.33

6. CONCLUSIONS

This study shows the effectiveness of using a more general prior distribution with online adaptation for multichannel speech enhancement. The accuracy of prior distribution modeling using multichannel observations is key point, which is realized in MAP speech spectral magnitude estimation. The experimental results show the superiority of the proposed method under nonstationary noise conditions.

7. REFERENCES

[1] T. Lotter, C. Benien, and P. Vary, “Multichannel Direction-Independent Speech Enhancement Using Spectral Amplitude Estimation,” *EURASIP Journal on Applied Signal Processing*, vol.11, pp.1147-1156, 2003.

[2] J. Rosca, R. Balan, and C. Beaugeant, “Multi-channel psychoacoustically motivated speech enhancement,” *Proc. ICASSP*, Hong Kong, 2003.

[3] T.H. Dat, K. Takeda, and F. Itakura, “Generalized gamma modeling of speech and its online estimation for speech enhancement,” *Proc. ICASSP*, Philadelphia, USA, 2005.

[4] Y. Ephraim and D. Malah, “Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator, ” *IEEE Trans. ASSP*, vol.ASSP-32, pp.1109-1121, 1984.

[5] K. Takeda, H. Fujimura, K. Itou, N. Kawaguchi, and S. Matsubara, “Construction and Evaluation of a Large In-Car Speech Corpus,” *IEICE Trans.*,vol.E88-D , pp. 553-561, 2005.