CLOSELY COUPLED ARRAY PROCESSING AND MODEL-BASED COMPENSATION FOR MICROPHONE ARRAY SPEECH RECOGNITION

Xianyu Zhao, Zhijian Ou, Minhua Chen, Zuoying Wang

Department of Electronic Engineering
Tsinghua University, Beijing, 100084, P.R.China
{zhaoxy, ozj}@thsp.ee.tsinghua.edu.cn

ABSTRACT

In this paper, a new microphone array speech recognition system in which the array processor and the speech recognizer are closely coupled is studied. The system includes a Generalized Sidelobe Canceller (GSC) beamformer followed by a recognizer with Vector Taylor Series (VTS) compensation. The GSC beamformer provides two outputs, allowing more information to be used in the recognizer. One is the enhanced target speech output, the other is the reference noise output. VTS is used to compensate the effect of the residual noise in the GSC speech output, utilizing the GSC reference noise output. The compensation is done in a Minimum Mean Square Error (MMSE) sense. Moreover, an iteration procedure using Expectation-Maximization (EM) algorithm is developed to refine the compensation parameters. Experimental results on MONC database showed that the new system significantly improved the speech recognition performance in the overlapping speech situations.

1. INTRODUCTION

Meetings present an important application domain for speech recognition technologies. However, when there are multiple concurrent speakers, the performance of speech recognizers is seriously degraded due to the overlapping speech. With their ability to provide directional discrimination, microphone arrays offer a potential solution to the problem of recognizing overlapping speech in the meeting environment [1][2].

Various microphone array-based speech processing methods have been studied, mainly for suppressing interfering signals to enhance the SNR of target signals [3-6]. Examples include broadband beam pattern synthesis [4], adaptive beamforming [5] and post-filtering [6], etc. Note that when used for speech recognition, these methods usually generate the enhanced target signal as a single output, which then gets treated as a single input to the recognizer. The array processor and the speech recognizer are loosely coupled. The only communication between them is through the signal output by the array processor. Other useful environmental information that can be provided by the multi-microphone array processor is ignored. In this way, we believe, the recognition performance of the system as a whole is limited.

In this paper, a new microphone array speech recognition system in which the array processor and the speech recognizer are closely coupled is studied, as shown in Fig. 1. Specifically, we consider the integration of the Generalized Sidelobe Canceller (GSC) adaptive beamforming and Vector Taylor Series (VTS) model-based noise compensation. GSC is widely used in microphone array system to suppress interferences adaptively [5]. VTS is a model-based compensation method developed to improve speech recognition performance in noisy environments for mono-microphone situations [7-9]. The proposed system includes a GSC beamformer followed by a recognizer with VTS compensation. Remarkably, here the GSC beamformer provides two outputs, allowing more information to be used in the recognizer. One is the enhanced target speech output, the other is the reference noise output. VTS is used to compensate the effect of the residual noise in the GSC speech output, utilizing the GSC reference noise output. Moreover, an iteration procedure using Expectation-Maximization (EM) algorithm was developed to refine the compensation parameters. The new system was tested on the Multi-channel Overlapping Numbers Corpus (MONC) database [10]. Experimental results showed that the new system improved the speech recognition performance in the overlapping speech situations.

2. MICROPHONE ARRAY SIGNAL PROCESSING AND RESIDUAL NOISE COMPENSATION

2.1 Microphone array signal processing with GSC

The array processor using using GSC adaptive beamforming is shown in Fig. 1. The GSC uses the configuration proposed in [5]. It includes a fixed beamformer (FBF), a multi-input adaptive filter (MIAF) and a blocking matrix (BM). The FBF enhances the desired speech signal; it can be designed as a delay-and-sum beamformer. The BM blocks the desired speech signal and passes the speech from other competing speakers. So in the BM output, the interfering signal is dominant. The MIAF uses the Normalized Least Mean Square (NLMS) algorithm to adapt the coefficients of a set of transversal FIR filterers. The desired speech output is extracted by subtracting the output of MIAF, $r(n)$, which we called the reference noise output, from the FBF output, $d(n)$.

0-7803-8874-7/05/$20.00 ©2005 IEEE

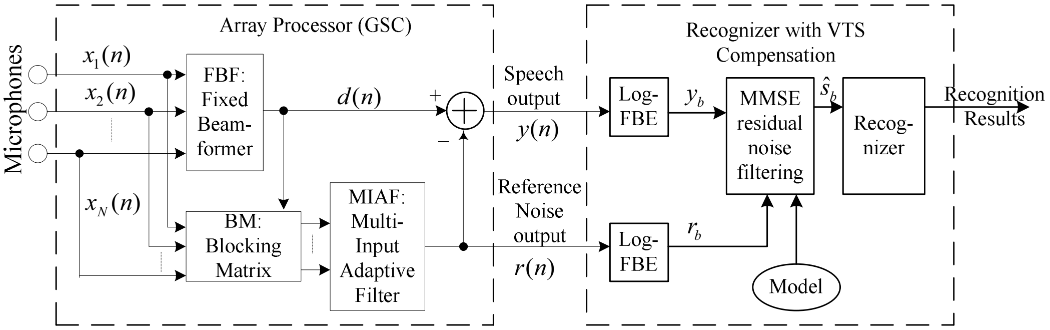

Fig. 1 Closely Coupled Array Processor and Speech Recognizer

2.2 Model-based residual noise compensation

2.2.1 A model for the GSC outputs

Although the SNR of the GSC speech output, $y(n)$, is increased comparing with the microphone input, there is still some residual noise in $y(n)$ due to the convergence error of the multi-input adaptive filter. We formulate the relation between the desired clean speech signal $s(n)$, GSC speech output $y(n)$ and the residual noise $e(n)$ as follows:

$$y(n) = s(n) + e(n). \tag{1}$$

Furthermore, we assume that the relation between the residual noise $e(n)$ and the GSC reference noise output $r(n)$ could be expressed as

$$e(n) = r(n) \otimes h(n), \tag{2}$$

where $h(n)$ stands for a linear filter which takes into account the amplitude and phase difference between $r(n)$ and $e(n)$. The relations in equations (1) and (2) can be represented using the power spectral density (PSD) as

$$\begin{aligned} |Y(\omega)| &= |S(\omega)| + |E(\omega)| \\ &= |S(\omega)| + |H(\omega)|^2 \cdot |R(\omega)|, \end{aligned} \tag{3}$$

where $|Y(\omega)|$, $|S(\omega)|$, $|E(\omega)|$, $|R(\omega)|$ and $|H(\omega)|^2$ are the PSD of $y(n)$, $s(n)$, $e(n)$, $r(n)$ and $h(n)$ respectively. Taking natural logarithms on equation (3), we get

$$\begin{aligned} \log|Y(\omega)| &= \log\left[|S(\omega)| + |E(\omega)|\right] \\ &= \log|S(\omega)| + \log\left[1 + \frac{|H(\omega)|^2 \cdot |R(\omega)|}{|S(\omega)|}\right]. \end{aligned} \tag{4}$$

For brevity, we use y_b, s_b, r_b and h_b for $\log|Y(\omega)|$, $\log|S(\omega)|$, $\log|R(\omega)|$ and $\log|H(\omega)|^2$ respectively. After some algebraic manipulation, we can get the following equation in the logarithmic filter bank energies (log-FBE) domain:

$$y_b = s_b + \log\left(1 + \exp\left(r_b + h_b - s_b\right)\right). \tag{5}$$

2.2.2 Vector Taylor Series (VTS) model-based residual noise compensation

By applying a first order Taylor series expansion around $\left(\mu_s, \mu_r, h_b^0\right)$, we can approximate equation (5) with

$$\begin{aligned} y_b \approx{} & \mu_s + \log\left(1 + \exp\left(\mu_r + h_b^0 - \mu_s\right)\right) \\ & + A\left(s_b - \mu_s\right) + \left(I - A\right)\left(r_b - \mu_r\right) + \left(I - A\right)\left(h_b - h_b^0\right), \end{aligned} \tag{6}$$

where

$$A = \frac{\partial y_b}{\partial s_b} = \frac{1}{1 + \exp\left(\mu_r + h_b^0 - \mu_s\right)}. \tag{7}$$

The clean speech s_b is modeled as a K-Gaussian mixture (we use $K = 64$):

$$p\left(s_b\right) = \sum_{k=1}^{K} P\left(\upsilon_k\right) N\left(s_b; \mu_{s,k}, \Sigma_{s,k}\right), \tag{8}$$

where υ_k is the k-th Gaussian distribution with mean $\mu_{s,k}$ and covariance matrix $\Sigma_{s,k}$, and $P\left(\upsilon_k\right)$ is the a priori probability of υ_k. The noise r_b is modeled by single Gaussian $N\left(r_b; \mu_r, \Sigma_r\right)$, and the noise is assumed to be statistically independent with the clean speech. Then through the relations in equation (6), the noisy GSC output y_b can also be modeled as a K-Gaussian mixture, and the mean and covariance of its k-th Gaussian distribution can be shown as

$$\mu_{y,k} = \mu_{s,k} + \log\left(1 + \exp\left(\mu_r + h_b^0 - \mu_{s,k}\right)\right) \tag{9}$$

and

$$\Sigma_{y,k} = A\Sigma_{s,k}A^T + \left(I - A\right)\Sigma_r\left(I - A\right)^T. \tag{10}$$

Because the residual noise is non-stationary in the case of overlapping speech, the compensation is done frame by frame. For the t-th frame, the noise mean and covariance are estimated as

$$\mu_r(t) = \frac{1}{L}\sum_{i=0}^{L-1} r_b(t - i) \tag{11}$$

and

$$\Sigma_r(t)=\frac{1}{L}\sum_{i=0}^{L-1}\left(r_b(t-i)-\mu_r(t)\right)\left(r_b(t-i)-\mu_r(t)\right)^T, \quad (12)$$

where L is the number of history frames used to make the estimation. After model compensation, the clean speech can be estimated based on the minimum mean square error (MMSE) criterion. Thus, we have

$$\begin{aligned}\hat{s}_b(t)&=E\left[s_b\,\middle|\,y_b(t),p(s_b),\mu_r(t),\Sigma_r(t),h_b^0\right]\\&\approx y_b(t)-\sum_{k=1}^{K}P\left(\upsilon_k\,|\,y_b(t)\right)\cdot\log\left(1+\exp\left(\mu_r(t)+h_b^0-\mu_{s,k}\right)\right).\end{aligned} \quad (13)$$

In the second part of the above equation, an approximation similar to that used in [9] is adopted to make the MMSE estimation efficiently. The posteriori probabilities $P\left(\upsilon_k\,|\,y_b(t)\right)$ are estimated using the compensated model for $y_b(t)$:

$$P\left(\upsilon_k\,|\,y_b(t)\right)=\frac{P(\upsilon_k)N\left(y_b(t);\mu_{y,k},\Sigma_{y,k}\right)}{\sum_{k'=1}^{K}P(\upsilon_{k'})N\left(y_b(t);\mu_{y,k'},\Sigma_{y,k'}\right)} \quad (14)$$

2.2.3 EM algorithm for h_b^0

The optimal estimation of the linear filter between the residual noise and the reference noise signal can be obtained by the maximum likelihood criterion. Since it is difficult to obtain the ML estimate directly, EM algorithm is used to iteratively update the parameter values. The auxiliary function $Q\left(h_b,h_b^0\right)$ for the EM algorithm is defined as

$$Q\left(h_b,h_b^0\right)=E\left[\log p\left(y_b,\upsilon_k\,|\,h_b\right)\,\middle|\,y_b,h_b^0\right], \quad (15)$$

where h_b^0 is the current estimate of the linear filter. The h_b that maximizes the auxiliary function can be found as:

$$\begin{aligned}\hat{h}_b^0&=h_b^0+\left[\sum_{t=1}^{T}\sum_{k=1}^{K}P\left(\upsilon_k\,|\,y_b(t),h_b^0\right)W_k^{-1}\right]^{-1}\\&\bullet\left[\sum_{t=1}^{T}\sum_{k=1}^{K}P\left(\upsilon_k\,|\,y_b(t),h_b^0\right)W_k^{-1}h_b\left(y_b(t),r_b(t),\upsilon_k\right)\right],\end{aligned} \quad (16)$$

where

$$W_k=(I-A)^{-1}A\Sigma_kA^T(I-A)^{-T} \quad (17)$$

and

$$\begin{aligned}h_b\left(y_b(t),r_b(t),\upsilon_k\right)=(I-A)^{-1}&\left(y_b(t)-\mu_{y,k}\right)-\\&\left(r_b(t)-\mu_r\right).\end{aligned} \quad (18)$$

In equation (16), T is the number of the log-FBE feature vectors used to obtain the ML estimation of h_b. Applying the simplification used in [9], we can approximate equation (16) with

$$\hat{h}_b^0=h_b^0+\frac{1}{T}\sum_{t=1}^{T}\sum_{k=1}^{K}P\left(\upsilon_k\,|\,y_b(t),h_b^0\right)h_b\left(y_b(t),r_b(t),\upsilon_k\right). \quad (19)$$

$\hat{h}_b^0$ is then used in the next EM iteration as the current estimate of h_b. This iteration proceeds until convergence.

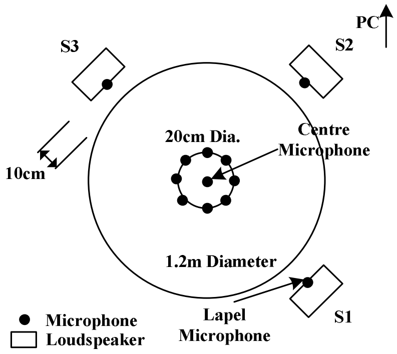

Fig. 2 Meeting room configuration for MONC [10]

2.2.4 Discussion

In [2], a post-filter after the beamforming stage is used to reduce the residual noise in the beamformer output. This post-filtering approach uses an assumed noise field coherence model to design the Wiener post-filter. However, in the assumed noise coherence model, there have parameters that are not normally known in advance. In practice, these parameters are hand-adjusted to achieve a compromise between diffuse and localized noise reduction [2]. This is not convenient for the deployment of a microphone array in different environment settings. On the other hand, the compensation parameters used in our model-based residual noise compensation algorithm can be learned automatically from data through the EM algorithm, and avoid the hand-adjustments. This feature is desirable for the use of our system in diverse environments.

3. EXPERIMENTAL RESULTS

The proposed new microphone array speech recognition system was tested on the Multi-channel Overlapping Numbers Corpus (MONC) database. The MONC is based on the Numbers Corpus (telephone quality speech, 30-word vocabulary) prepared by the Center for Spoken Language Understanding at the Oregon Graduate Institute [10]. The meeting room configuration for the MONC data acquisition is shown in Fig. 2. The loudspeakers simulate the presence of the desired speaker (S1) and the two competing speakers (S2 and S3) in a realistic meeting scenario. A circular microphone array comprising eight equally spaced microphones is placed in the middle of a round table. An additional microphone is placed at the center of the table. A lapel microphone is attached to each loudspeaker. The same type omni-directional microphones are used in all locations. The circular table is located at one end of a moderately reverberant, $8.2\text{m}\times3.6\text{m}\times2.4\text{m}$, rectangular room. The dominant non-speech noise is produced by a PC located at the opposite end of the room. Three recognition tasks based on the MONC database were carried out in our work:

- Task-S1: only the desired speaker S1 is active, no overlapping speech

- Task-S1S2: the desired speaker S1 with one competing speaker S2 active (resulting in approximately 0dB SNR at the centre table-top microphone location)
- Task-S1S2S3: the desired speaker S1 with two competing speakers S2 and S3 active (resulting in approximately -3dB SNR at the centre location)

Our speech recognition system is based on continuous-density hidden Markov models (CHMM) with 363 states and 14 Gaussians per state. The 45-dimensional feature vector is formed by 14 MFCC's, energy plus their first and second order differentials. The recognition system was trained on the clean training set from the original Numbers Corpus. The baseline system achieved a Word Error Rate (WER) of 6.19% on the clean test set from the original Numbers Corpus.

Task	S1	S1S2	S1S2S3
Lapel Microphone	8.32	47.74	50.28
Centre Microphone	13.16	67.31	83.24
GSC	13.02	22.89	32.15
GSC + Model-based Compensation	12.47	19.44	23.27
GSC + Model-based Compensation + EM	9.02	16.92	22.00

Table 1. Word Error Rate (WER) results (%)

The WER results of the speech recognition experiments are listed in Table 1. In this table, the "Lapel Microphone" row lists WER results for recordings from the desired speaker S1's lapel microphone, and the "Centre Microphone" row gives the results for the recordings from the centre table-top microphone. It is clear from these two rows of results that the speech recognition performance becomes seriously degraded when there are several concurrent competing speakers. Even when the lapel microphone is placed very near the desired speaker, the problem remains. The "GSC" row represents the experiment using the microphone array's GSC speech output, y . It is shown that microphone array signal processing, such as GSC, improves the recognition performance for overlapping speech, as expected.

The performance was further improved when model-based residual noise compensation was done. In our experiments of model-based compensation, the number of history frames (L) used to estimate the parameters of the reference noise was set to 10. Because of the non-stationary nature of overlapping speech, the length of history frames could not be too large. While 10 history frames allowed for an accurate estimation of the mean vector of the reference noise, the accuracy in the estimation of the covariance matrix was very poor. So we only applied mean compensation and left the covariance matrix unchanged. Using this kind of model-based compensation for the residual noise in the GSC speech output, we have resulted in a comparable performance with those reported by Moore and McCowan in [2], as shown in the fifth row. In this first set of compensation experiments, the linear filter between the residual noise in y and the reference noise was set a priori and remained unchanged in the subsequent compensation procedure.

Applying the EM iteration procedure proposed in section 2.2.3 for the optimal estimation of h_b^0 resulted in further performance improvements, as the last row of Table 1 shows. Because we use all the frames in one utterance for the EM parameter estimation, multiple passes of compensation are performed for this mode after the whole sentence is recorded.

4. CONCLUSIONS

In this paper, a new microphone array speech recognition system in which the array processor and the speech recognizer are closely coupled is studied. The system includes a GSC beamformer followed by a recognizer with VTS compensation. VTS compensation is performed in the log-FBE domain for the filtering of the residual noise in the GSC speech output, utilizing the GSC reference noise output. This approach does not require the a priori knowledge of the noise field coherence model, and can be used to do on-line compensation conveniently in different environmental settings. The compensation is done in a MMSE sense. Moreover, an iteration procedure using EM algorithm is developed to refine the compensation parameters. Experimental results on MONC database showed that the new system significantly improved the speech recognition performance in the overlapping speech situations.

5. REFERENCES

[1] E. Shriberg, A. Stolcke and B. Baron. Observations on overlap: findings and implications for automatic processing of multi-party conversation. In *Proceedings of Eurospeech 2000*, vol. 2, pp. 1359-1362, 2001

[2] D. Moore and I. McCowan. Microphone array speech recognition: experiments on overlapping speech in meetings. In *Proceedings of the ICASSP 2003*, 2003

[3] M. Brandstein and D. Ward. Microphone Arrays - Signal Processing Techniques and Applications. Springer-Verlag, New York, 2001

[4] R. Kennedy, P. Abhayapala, and D. Ward. Broadband nearfield beamforming using a radial beampattern transformation. In *IEEE Trans. Signal Processing*, vol. 46, no. 8, pp. 2147-2156, Aug. 1998

[5] O. Hoshuyama, A. Sugiyama and A. Hirano. A robust adaptive beamforming for microphone arrays with a blocking matrix using constrained adaptive filters. *IEEE Trans. Signal Processing*, vol. 47, no. 10, pp. 2677-2684, Oct. 1999

[6] R. Zelinski. A microphone array with adaptive post-filtering for noise reduction in reverberant rooms. In *Proceedings of the ICASSP'88*, pp. 2578-2581, 1988

[7] A. Acero, L. Deng and T. Kristjansson and J. Zhang. HMM adaptation using vector Taylor series for noisy speech recognition. In *Proceedings of ICSLP'2000*, 2000

[8] D. Y. Kim, C. K. Un, N. S. Kim. Speech recognition in noisy environments using first-order vector Taylor series. In *Speech Communication*, vol. 24, no. 1, pp. 39-49, 1998

[9] J. C. Segura, A. Torre, M. C. Benitez, A. M. Peinado. Model-based compensation of the additive noise for continuous speech recognition: experiments using the AURORA II database and tasks. In *Proceedings of the EUROSPEECH 2001*, 2001

[10] Multi Channel Overlapping Numbers Corpus (MONC) distribution. http://cslu.cse.ogi.edu/corpora/