Maximum Confidence Measure Based Interaural Phase Difference Estimation for Noise Masking in Dual-Microphone Robust Speech Recognition

Hsien-Cheng Liao[1], *Yuan-Fu Liao*[2], *Chin-Hui Lee*[3]

[1] Information & Communications Research Labs, Industrial Technology Research Institute, Taiwan
[2] Department of Electronic Engineering, National Taipei University of Technology, Taiwan
[3] School of Electrical and Computer Engineering, Georgia Institute of Technology, USA

[1]hcliao@itri.org.tw, [2]yfliao@ntut.edu.tw, [3]chl@ece.gatech.edu

Abstract

A new one-stage maximum confidence measure (MCM) based interaural phase difference estimation framework for noise masking is proposed to closely integrate the underline speech models into dual-microphone array noise filtering for robust speech recognition. The main ideas are: (1) utilizing both the speech and filler models of the recognizer to feedback confidence measures (CMs) that indicate the degree of separation between filtered speech and interference noises, and (2) automatically optimizing the parameters of the microphone array with an expectation maximization (EM) algorithm based on the proposed MCM criterion. Experimental results on a Mandarin voice command task show that the proposed approach significantly improves the final speech recognition rates. Moreover the observed performance degradation is usually graceful under low signal-to-noise ratios (SNRs) and close interference noises conditions.

Index Terms: robust speech recognition, microphone array, interaural phase difference, confidence measure

1. Introduction

Dual-microphone based noise filtering speech enhancement, which is motivated by human binaural system, is becoming a popular front-end for automatic speech recognition (ASR) recently. Two major binaural phenomena are often adopted in these methods [1-4], including: (1) interaural time difference (ITD) and (2) interaural level difference (ILD). Experiments conducted by Woodworth [5] and Feddersen [6] had shown that ITDs and ILDs indeed provide important acoustic localization cues. Between them, ITDs are especially useful when the two microphones are close to each other.

ITDs usually can be estimated reliably from interaural phase differences (IPDs) and used to produce a noise mask to filter out unwanted interference noises in the time-frequency domain. For examples, in [7-8], Kim et al. proposed an IPDs-based ITD estimation method and generate a binary noise mask by hard-decision (with a pre-determined ITD threshold). In [9], Shi et al. also adopted IPDs and proposed a phase error filtering approach using a soft-decision noise masking function.

The key to success of these approaches is how to choose proper ITD thresholds or phase error filter parameters for precise noise mask generation. For example, in [8], a minimum cross-correlation (MCC) measure between the target and interference signals and a point-by-point search algorithm was used to automatically select an optimal threshold. In [9] a maximum likelihood (ML) criterion using prior speech models and EM algorithm was adopted to choose the parameters of the phase error filter. To go further, in [3-4], probability density functions were trained in advance from a corpus of sound mixtures and then a binary mask was estimated using maximum a posteriori (MAP) classification.

Nonetheless, the MCC and the MAP classification-based binary masking generation methods in [3-4, 8], did not take into account the underlying speech recognizer and thus may only be optimal for signal separation but not necessarily for improving speech recognition accuracy. On the other hand, the ML method in [9] did consider prior speech models, but it knew nothing about the interference noises (except the assumption on phase error distributions). It may not discriminate well between the desired signal and interference noises. Besides, the prior speech model in [9] or the probability density functions in [3-4] all have to be properly trained in advance using noise-filtered training data or a mixed of sound corpus. This may raise a new problem of model parameters mismatch between training and testing conditions.

In this paper, we propose a new framework to fully integrate the dual microphone noise masking and the speech recognition system. The main idea is to utilize both prior speech and filler models to better discriminate between desired speech and unwanted interference noises. A MCM criterion and an EM algorithm can then be developed to automatically find a suitable masking threshold via enlarging the output score difference between the prior speech and filler models. In addition to improved accuracies it is also worth noting that both the prior speech and filler models can be trained using only clean data. In other words, no noise-filtered training data [9] or mixed sound corpora [3-4] are required in the proposed method.

2. Phase Error Based Noise Masking

ITDs usually could be estimated reliably using interaural phase differences. Suppose two signals, $x_L[n]$ and $x_R[n]$, are received at the left and right microphones, respectively. Following the assumption in [7], if the direction of the target speaker is known and, without loss of generality, positioned at a 90-degree angle, its ITD will be close to zero. Moreover, if the number of interference noises is I and the related ITD for the ith noise is $\delta(i)$, the two received signals can be further represented as:

$$x_L[n] = \sum_{i=0}^{I} x_i[n], \qquad x_R[n] = \sum_{i=0}^{I} x_i[n-\delta(i)] \tag{1}$$

where $x_0[n]$ represents the target signal and $x_i[n]$ ($i \neq 0$) represent noise interferences.

To obtain ITDs, phase error analysis could be performed frame-by-frame on the two received signals using short-time Fourier transform (STFT). Then the spectra of the left and right signals can be represented as

$$X_L(t,k)=\sum_{i=0}^{I}X_i(t,k),\quad X_R(t,k)=\sum_{i=0}^{I}e^{-j\omega_k d_i(t,k)}X_i(t,k) \tag{2}$$

where $X_i(t,k)$ represents the STFT of $x_i[n]$ at frame t and frequency bin k, and $d_i(t,k)$ indicates the frequency-dependent ITD of the ith source.

Assume that each time-frequency bin is dominated by a single source i^*, the two STFTs on a specific time-frequency bin (t_0,k_0) could be further simplified as:

$$X_L(t_0,k_0)\approx X_{i^*}(t_0,k_0),\quad X_R(t_0,k_0)\approx e^{-j\omega_{k_0}d(t_0,k_0)}X_{i^*}(t_0,k_0) \tag{3}$$

Accordingly, the ITD for a particular time-frequency bin (t_0,k_0) is determined by its corresponding IPD value:

$$\left|d(t_0,k_0)\right|\approx\frac{1}{\left|\omega_{k_0}\right|}\min_r\left|\angle X_R(t_0,k_0)-\angle X_L(t_0,k_0)-2\pi r\right| \tag{4}$$

By comparing the ITD of each time-frequency bin with a predefined ITD threshold τ, a binary noise mask $\mu(\tau,t,k)$ in the frequency domain can be derived:

$$\mu(\tau,t,k)=\begin{cases}1, \text{if } \left|d(t,k)\right|\le\tau\\ \varepsilon, \text{ otherwise}\end{cases} \tag{5}$$

where ε is a floor constant with a value of 0.01.

By applying the mask $\mu(\tau,t,k)$ to the averaged spectra $\overline{X}(t,k)$ of the two signals, only time-frequency bins which have smaller ITDs comparing with τ are considered to belong to the target and an enhanced spectra $\widetilde{X}(\tau,t,k)$ can be obtained.

$$\overline{X}(t,k)=\left\{X_L(t,k)+X_R(t,k)\right\}/2 \tag{6}$$

$$\widetilde{X}(\tau,t,k)=\mu(\tau,t,k)\overline{X}(t,k) \tag{7}$$

More details can be referred to [7].

It is worth noting that the quality of the noise mask, $\mu(\tau,t,k)$, and the output target signals are highly dependent on the threshold τ.

3. The Proposed MCM-Based Method

Generally speaking, if prior knowledge of the speech and interference noises is available, one can build a speech and an interference noise model to examine how good an ITD threshold is for separating speech from interference noise signals. However, it is often difficult to obtain such prior information, especially for covering all possible interference noises.

To alleviate this issue, a speech recognizer purely trained from clean speech data (which is usually available) is used to feedback CMs in order to transform the problem into a verification-based ITD threshold estimation task. The CM scores are defined as a function of the threshold τ:

$$CM(\tau)=\left[\log P\left(\tilde{\mathbf{C}}(\tau)\mid\Lambda_{\mathrm{SP}}\right)-\log P\left(\tilde{\mathbf{C}}(\tau)\mid\Lambda_{\mathrm{F}}\right)\right] \tag{8}$$

where $\tilde{\mathbf{C}}(\tau)$ is the filtered mel-frequency cepstral coefficient (MFCC) feature vector obtained from $\widetilde{X}(\tau)$, Λ_{SP} and Λ_{F} represent the speech and filler models of the speech recognizer, respectively.

By this way, if a suitable threshold τ is chosen, most of the noise interferences will be removed from a test utterance and the recognizer will report a high CM score. On the other hand, a low CM score will be observed, if an improper τ is adopted. Therefore, the estimation of best threshold, τ_{CM}, can be formulated as the following MCM optimization problem:

$$\tau_{\mathrm{CM}}=\arg\max_{\tau}\left[\log P\left(\tilde{\mathbf{C}}(\tau)\mid\Lambda_{\mathrm{SP}}\right)-\log P\left(\tilde{\mathbf{C}}(\tau)\mid\Lambda_{\mathrm{F}}\right)\right] \tag{9}$$

3.1. Preliminary Experiments on the Relationship between CM and Recognition Accuracy

To verify the feasibility of the proposed approach, a preliminary voice command experiment was first executed. The direction of the target speaker is positioned at a 90-degree angle, input SNR is 0dB and the testing utterances were interfered by a babble noise placed at 30 or 60 degrees (details will be described in the next section).

Figure 1 shows the averaged recognition rates and the CM scores after filtering by Eq. (5) at different τ. From the results, a strong correlation between recognition accuracy and CM scores could be observed. Especially, the highest recognition accuracy and CM score for 30 (or 60) degrees case all happen at the same τ value. This confirms the feasibility of the proposed approach.

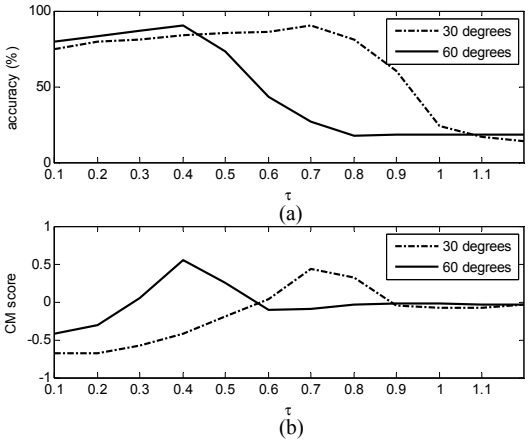

Figure 1: *The correlation between recognition accuracy and CM score at different τ: (a) Recognition accuracy, and (b) CM scores were obtained from the filtered 0dB testing utterances interfered by a babble noise placed at 30 and 60 degrees.*

3.2. MCM-Based EM Algorithm

In order to find an optimal τ to the MCM criterion automatically, we follow the ideas presented in [9, 11] and propose a method which can be done iteratively using an EM algorithm. In the E step, we formulate a Q function as follows:

$$\begin{aligned}&Q\left(\tau\mid\tau^{(k)}\right)\\&=E\left[\left(\log P\left(\tilde{\mathbf{C}}(\tau)\mid\Lambda_{\mathrm{SP}}\right)-\log P\left(\tilde{\mathbf{C}}(\tau)\mid\Lambda_{\mathrm{F}}\right)\right)\middle|\mathbf{C},\tau^{(k)},\Lambda_{\mathrm{SP}},\Lambda_{\mathrm{F}}\right]\\&=\sum_t\sum_s\sum_m\log p(\tilde{\mathbf{c}}_t(\tau),s,m\mid\Lambda_{\mathrm{SP}})p\left(s,m\mid\mathbf{c}_t,\tau^{(k)},\Lambda_{\mathrm{SP}}\right)\\&\quad-\sum_t\sum_{s'}\sum_{m'}\log p(\tilde{\mathbf{c}}_t(\tau),s',m'\mid\Lambda_{\mathrm{F}})p\left(s',m'\mid\mathbf{c}_t,\tau^{(k)},\Lambda_{\mathrm{F}}\right)\end{aligned} \tag{10}$$

here $\mathbf{C}$ is the noisy MFCC feature set, $\mathbf{c}_t$ is a feature vector in $\mathbf{C}$, $\tilde{\mathbf{c}}_t(\tau)$ is a feature vector in $\tilde{\mathbf{C}}(\tau)$, t is a frame index, (s, m) and (s', m') represent the state and mixture sequences of the speech and filler models, respectively.

If hidden Markov models (HMMs) are trained as the speech and filler models, the auxiliary function can be further simplified as:

$$Q(\tau \mid \tau^{(k)}) = \sum_{t=1}^{T}\sum_{s=1}^{S_{SP}}\sum_{m=1}^{M_{SP}} p(s,m \mid \mathbf{c}_t, \tau^{(k)}, \Lambda_{SP}) \times \left\{ \sum_{i=1}^{D} \frac{(\tilde{c}_{t,i}(\tau) - \mu_{SP,s,m,i})^2}{-2\sigma^2_{SP,s,m,i}} + Z_{SP,s,m} \right\} - \sum_{t=1}^{T}\sum_{s'=1}^{S_F}\sum_{m'=1}^{M_F} p(s',m' \mid \mathbf{c}_t, \tau^{(k)}, \Lambda_F) \times \left\{ \sum_{i=1}^{D} \frac{(\tilde{c}_{t,i}(\tau) - \mu_{F,s',m',i})^2}{-2\sigma^2_{F,s',m',i}} + Z_{F,s',m'} \right\} \quad (11)$$

where D is the dimension of the feature vector, both $Z_{SP,s,m}$ and $Z_{F,s',m'}$ are constants, $\mu_{SP,s,m,I}$, $\sigma_{SP,s,m,i}$, $\mu_{F,s',m',i}$, $\sigma_{F,s',m',i}$ are the mean and the variance of the ith component in mixture m, of state s. $\tilde{c}_{t,i}(\tau)$ is the ith cepstral coefficient of the filtered MFCC feature vector $\tilde{\mathbf{c}}_t(\tau)$ at frame t and

$$\tilde{c}_{t,i}(\tau) = \sqrt{\frac{2}{N}} \sum_{p=1}^{N} \log b_{t,p}(\tau) \cos\frac{i\pi(p-0.5)}{N} = \sqrt{\frac{2}{N}} \sum_{p=1}^{N} \log\left(\sum_{j=1}^{L_p} a_{p,j} \bar{x}_{t,p,j} \mu(\tau,t,p,j) \right) \cos\frac{i\pi(p-0.5)}{N} \quad (12)$$

here N is the number of filter banks, $b_{t,p}$ is the pth filter bank output, L_p is the number of components in the pth filter bank channel, $a_{p,j}$ is the weight of the jth component in the pth filter bank channel, and $\bar{x}_{t,p,j}$ is the jth element of the averaged spectral magnitude in the pth filter bank channel.

In order to make Eq. (11) differentiable, the binary mask in Eq. (5) was approximated by a soft "ε-1" Sigmoid function:

$$\mu(\tau,t,p,j) = (1-\varepsilon)\frac{1}{1+e^{\beta(|d(t,p,j)|-\tau)}} + \varepsilon \quad (13)$$

where β controls the steepness of the sigmoid function and $d(t,p,j)$ is the jth element of the ITD in the pth filter bank channel. The masking criterion in Eq. (13) also has smooth variation characteristics which can avoid abrupt changes in the frequency domain and provide better filtered results.

Therefore, In the M step, the derivative of Eq. (11) with respect to τ is:

$$Q'(\tau \mid \tau^{(k)}) = \sum_{t=1}^{T}\sum_{s=1}^{S_{SP}}\sum_{m=1}^{M_{SP}} p(s,m \mid \mathbf{c}_t, \tau^{(k)}, \Lambda_{SP}) \times \left\{ \sum_{i=1}^{D} \frac{(\tilde{c}_{t,i}(\tau) - \mu_{SP,s,m,i}) \frac{\partial \tilde{c}_{t,i}(\tau)}{\partial \tau}}{-\sigma^2_{SP,s,m,i}} \right\} - \sum_{t=1}^{T}\sum_{s'=1}^{S_F}\sum_{m'=1}^{M_F} p(s',m' \mid \mathbf{c}_t, \tau^{(k)}, \Lambda_F) \times \left\{ \sum_{i=1}^{D} \frac{(\tilde{c}_{t,i}(\tau) - \mu_{F,s',m',i}) \frac{\partial \tilde{c}_{t,i}(\tau)}{\partial \tau}}{-\sigma^2_{F,s',m',i}} \right\} \quad (14)$$

where

$$\frac{\partial \tilde{c}_{t,i}(\tau)}{\partial \tau} = \sqrt{\frac{2}{N}} \sum_{p=1}^{N} \cos\frac{i\pi(p-0.5)}{N} \frac{\partial}{\partial \tau}(\log b_{t,p}(\tau)) = \sqrt{\frac{2}{N}} \sum_{p=1}^{N} \cos\frac{i\pi(p-0.5)}{N} \frac{1}{b_{t,p}(\tau)} \times \sum_{j=1}^{L_p} a_{p,j} \bar{x}_{t,p,j} (1-\varepsilon) \frac{\beta e^{\beta(|d(t,p,j)|-\tau)}}{\left(1+e^{\beta(|d(t,p,j)|-\tau)}\right)^2} \quad (15)$$

Since there is no closed-form solution for τ while setting Eq. (14) to zero, a generalized M step with gradient ascent algorithm is adopted. The resulted update rule is

$$\tau^{(k+1)} = \tau^{(k)} + \eta Q'(\tau \mid \tau^{(k)})\Big|_{\tau=\tau^{(k)}} \quad (16)$$

where k is the iteration index and η is the learning rate.

4. Experimental Results

4.1. Task and Experimental Settings

The speech recognition experiments were conducted on a voice command task. Testing utterances were recorded in an anechoic room by eleven speakers (6 males and 5 females) and the sampling rate was 8 kHz. The prompt sheet has 50 short Mandarin remote control commands for toy ICs. Except some canonical commands such as 停 (/ting/, stop) and 前進 (/qian jin/, go), there are also several confusable variant sets, e.g., 左 (/zuo/, left), 向左 (/xiang zuo/, turn left), 向左轉 (/xiang zuo zhuan/, make a left turn), which certainly increase the difficulty of speech recognition.

For the setting of dual-microphone array, the speaker was positioned at 90 degrees and the distance between the center of the two microphones and the speaker was 30cm. The two microphones were placed at 5cm apart from each other. A babble noise source placed at 30 and 60 degrees was recorded separately and then digitally added to a total of 547 utterances (excluding three bad quality utterances) to form testing sets at different SNRs (0, 6, 12, and 18dB).

The Mandarin speech recognizer was trained using MAT2000 corpus (set DB4, about 2000 speakers) [12]. There are 100 right-context-dependent syllable-initial and 38 context-independent syllable-final HMMs. All 138 models have two states with a left-to-right topology and each state has two mixtures. The filler model is also trained using MAT2000. It had only one state and each state has three mixtures. Besides, all models were trained using HTK [13].

Finally, the feature vector has 13 elements (the first 8 elements of 13-dimensional MFCCs, the first 4 elements of the corresponding delta coefficients, and the delta C_0 energy) computed with a window size of 75ms and a frame shift of 25ms.

4.2. Performance Evaluation

Three approaches were evaluated and compared including (1) a single microphone baseline, (2) the MCC-based method in [8] and (3) the proposed MCM approach.

For the MCC-based approach, MATLAB source code [8] provided by the authors was adopted (some parameters were adjusted to fit the experimental settings here). Besides, for the proposed MCM-based method, the learning step size η was set to 0.01 and the initial guess of the ITD threshold τ was set to 0.2.

Table 1. *Accuracy comparison of different methods.*

(a) Babble noise located at 30 degrees

30 degrees	baseline	MCC	MCM
0dB	8.96	74.59	89.76
6dB	61.79	83.55	93.60
12dB	90.31	88.12	96.34
18dB	95.80	93.05	95.98

(b) Babble noise located at 60 degrees (closer to speech source)

60 degrees	baseline	MCC	MCM
0dB	20.29	38.57	92.32
6dB	71.85	76.23	95.80
12dB	92.14	89.76	95.61
18dB	96.53	93.97	97.26

Table 1 shows the average recognition rates under various SNR and interference noise direction conditions. From the table, it can be seen that a significant improvement was achieved by the proposed MCM-based approach. Especially,

its performance degraded more gracefully under low SNR and close interference noise cases.

It is also found that the method proposed in [8] didn't work very well here. This may due to that different experimental settings were adopted. Especially, (1) the sampling rate was changed from 16 kHz to 8 kHz, (2) the search range of τ was doubled to 3-85 degrees (original it was 3-45 degrees) and (3) the recognition task is somehow a more difficult one due to many short and confusable command sets.

4.3. Analysis

4.3.1. Convergence Behavior of the EM Iterations

Figure 2 shows two typical learning curves when the input SNR is 0dB. It could be found from the figure, the proposed approach has good convergence characteristics.

However, the convergence speed for 60 degrees was faster than that for the 30 degrees case. This may come from the setting of the initial value of the ITD threshold τ, since 0.2 is a good guess for the 60 degrees case but not for the 30 degrees (it should be around 0.5, as shown in Fig. 3). This may also explain why MCM performed better in 60 than 30 degrees case as shown in Table 1.

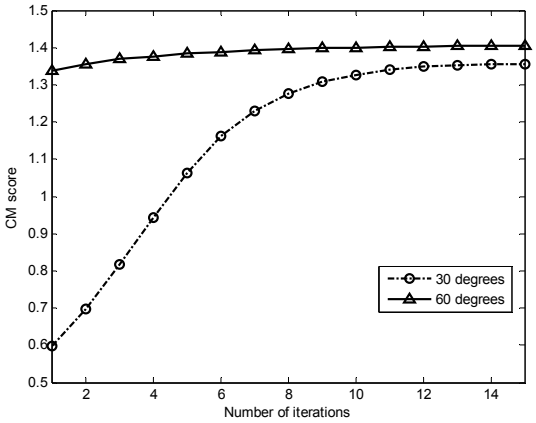

Figure 2: *Two typical convergence curves while noise interferences located at 30 and 60 degrees.*

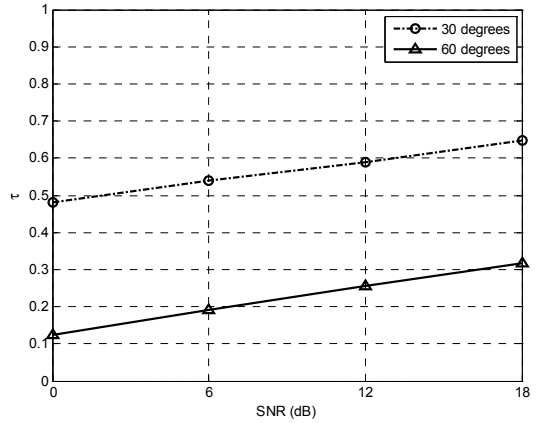

Figure 3: *The mean of the estimated τ versus different SNRs at two different noise directions.*

4.3.2. Relationship between SNR and Threshold

The means of the estimated τ at four different SNRs with noise comes from 60 or 30 degrees are shown in Fig. 3. It is interesting to find there is a strong correlation between ITD threshold τ, SNRs and interference noise direction. The trend of the curves is similar at both directions indicating that a large τ is preferred at high SNRs. Especially, a rough estimation of SNR may help to guess a proper initial value of τ in order to speed up the whole estimation procedure. On the other hand, the estimated τ could also be used to guess the SNR conditions and the directions of interference noises.

5. Conclusion

A new one-stage MCM-based interaural phase difference estimation framework for closely integrating noise masking and speech recognizer have been shown to be able to significantly improve the speech recognition rates and degrade more gracefully in worst conditions. It is also interesting to find that there is a strong correlation between ITD threshold τ and SNRs which may help to speed up the whole estimation procedure. All experimental findings confirm the feasibility of the proposed MCM approach.

6. Acknowledgment

This paper is a partial result of Project A353C41221 conducted by ITRI under sponsorship of the Ministry of Economic Affairs, Taiwan, R.O.C.

7. References

[1] Park, H. and Stern, R., "Spatial separation of speech signals using amplitude estimation based on interaural comparisons of zero-crossings", Speech Communication, 51:15-25, 2009.

[2] Cobos, M. and Lopez, J.J., "Two-microphone separation of speech mixtures based on interclass variance maximization. J. Acoust", Soc. Am., 127:1661-1672, 2010.

[3] Harding, S., Barker, J. and Brown, G., "Mask estimation for missing data speech recognition based on statistics of binaural interaction", IEEE Trans. Audio Speech Lang. Process., 14:58-67, 2006.

[4] Srinivasan, S., Roman, N. and Wang, D., "Binary and ratio time-frequency masks for robust speech recognition", Speech Communication, 48:1486-1501, 2006.

[5] Woodworth, R. S., "Experimental Psychology", New York: Holt, Rinehart, Winston, 1938.

[6] Feddersen, W.E., Sandel, T.T., Teas, D.C., and Jeffress, L.A., "Localization of high frequency tones", Journal of the Acoustical Society of America, 5:82-108, 1957.

[7] Kim, C., Kumar, K., Raj, B. and Stern, R.M., "Signal separation for robust speech recognition based on phase difference information obtained in the frequency domain", In INTERSPEECH-2009, pp. 2495-2498, 2009.

[8] Kim, C., Stern, R.M., Eom, K. and Lee, J., "Automatic selection of thresholds for signal separation algorithms based on interaural delay", In INTERSPEECH-2010, pp. 729-732, 2010.

[9] Shi, G., Aarabi, P. and Jiang, H., "Phase-Based Dual-Microphone Speech Enhancement Using A Prior Speech Model", IEEE Trans. Audio Speech Lang. Process., 15:109-118, 2007.

[10] Sukkar, R.A. and Lee, C.H., "Vocabulary independent discriminative utterance verification for nonkeyword rejection in subword based speech recognition", IEEE Trans. Speech Audio Process., 4:420-429, 1996.

[11] Chien J.T. and Liao C.P., "Maximum Confidence Hidden Markov Modeling for Face Recognition", IEEE Trans. Pattern Anal. Mach. Intell. 30:606-616, 2008.

[12] Wang, H.C., Seide, F., Tseng, C.Y., and Lee, L.S., "MAT-2000 - design, collection, and validation of a Mandarin 2000-speaker telephone speech database", In ICSLP-2000, 4:460-463, 2000.

[13] Young, S., Kershaw, D., Odell, J., Ollason, D., Valtchev, V. and Woodland, P., "The HTK Book Version 3.0", Cambridge University Press, 2000.