논문 2013-08-23

Recognition Performance Improvement of Unsupervised Limabeam Algorithm using Post Filtering Technique

Dinh Cuong Nguyen, Suk-Nam Choi, Hyun-Yeol Chung*

Abstract : Abstract- In distant-talking environments, speech recognition performance degrades significantly due to noise and reverberation. Recent work of Michael L. Selzer shows that in microphone array speech recognition, the word error rate can be significantly reduced by adapting the beamformer weights to generate a sequence of features which maximizes the likelihood of the correct hypothesis. In this approach, called Likelihood Maximizing Beamforming algorithm (Limabeam), one of the method to implement this Limabeam is an UnSupervised Limabeam(USL) that can improve recognition performance in any situation of environment. From our investigation for this USL, we could see that because the performance of optimization depends strongly on the transcription output of the first recognition step, the output become unstable and this may lead lower performance. In order to improve recognition performance of USL, some post-filter techniques can be employed to obtain more correct transcription output of the first step. In this work, as a post-filtering technique for first recognition step of USL, we propose to add a Wiener-Filter combined with Feature Weighted Malahanobis Distance to improve recognition performance. We also suggest an alternative way to implement Limabeam algorithm for Hidden Markov Network (HM-Net) speech recognizer for efficient implementation. Speech recognition experiments performed in real distant-talking environment confirm the efficacy of Limabeam algorithm in HM-Net speech recognition system and also confirm the improved performance by the proposed method.

Keywords : Robust speech recognition, Microphone array processing, Beamforming, Limabeam, hidden markov model, HM-Net, Post filtering, Feature Weighted Malahanobis Distance.

I. INTRODUCTION

In State-Of-The-Art Automatic Speech Recognition (ASR), the performance can be attained well when the speech signals are captured by close-talking microphone. However, in distant-talking environments, such as meeting room, inside car, information kiosk, etc., the speech recognition accuracy degrades significantly due to the effects of additive noise and reverberation. In these scenarios, microphone array can be used to compensate the distortions of the speech signals by spatial filtering to the sound field.

Many methods were proposed for microphone array speech recognition. The simplest approach is that of the delay-and-sum beamformer, in which delays were inserted in each channel to compensate for differences in travel time between the desired sound source

* Corresponding Author (hychung@yu.ac.kr)
Received: 19 Feb. 2013, Revised: 26 Mar. 2013, Accepted: 18 Apr. 2013.
※ This work was supported by the 2013 Yeungnam University research grant.

ISSN : 1975-5066

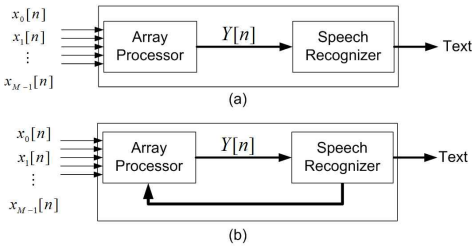

Fig. 1 (a) Conventional architecture and (b) Limabeam architecture for speech recognition with microphone array

and the various sensors before summing them to give a single enhanced output channel. This process ensures constructive in-phase addition of the desired signal during the summation. As the noise components in the signal are combined out of phase, the procedure leads to a relative increase in the signal level with respect to the noise level [1-4].

Other more sophisticated beamforming techniques which calculate the channel filters to optimize a particular criterion such as gain with respect to an isotropic noise field or a set of particular noise locations. The typical one is Generalized Sidelobe Canceller (GSC) [5], in which the array parameters are updated on a sample-by-sample or frame-by-frame basis. This method assumes that the target and jammer signals are uncorrelated. When this assumption is violated, as is the case for speech in a reverberation environment, the methods suffer from signal cancellation because reflected copies of the target signal appear as unwanted jammer signals.

All of microphone array processing approaches described above, shown in Figure 1(a), were designed for signal enhancement (improve Signal to Noise Ratio - SNR). However, a speech system does not interpret waveform-level information directly. It is statistical pattern classifier that operates on a sequence of features derived from the waveform. Therefore, above techniques do not decrease the Word Error Rate (WER) significantly.

Recent work of Michael L. Seltzer shows that in microphone array speech recognition, the WER can be dramatically reduced by adapting the beamformer weights to generate a sequence of features which maximizes the likelihood of the correct hypothesis [6]. In this approach, called Likelihood Maximizing Beamforming (Limabeam), shown in Figure 1(b), information from the recognition system is used to optimize the beamformer weights. Limabeam has several advantages over classical methods such as enhancing those signal components which is important for accuracy recognition; no assumption about the interfering signals; requiring no prior knowledge of the microphone array geometry, room configuration, speaker-to -receiver impulse responses, etc. Limabeam is mainly a data-driven approach.

Originally, Limabeam was investigated with Sphinx3, an HMM-based large-vocabulary speech recognition system [7], and English database. In this paper, we present the results of investigation the performance of speech recognition using microphone array and the way to implement Limabeam algorithm with Hidden Markov Network (HM-Net) speech recognition system and Korean database.

In practice, usually post-filter is applied to the output of beamformer for improving the system performance. Most approaches use the input channel auto- and cross-spectral densities to estimate a Wiener post-filter. These post-filters have been used successfully in a number of speech enhancement and robust speech recognition applications [13]. Feature Weighted Mahalanobis Distance (FWMD) is also known to be effective for improving the performance of Limabeam algorithm [16] and if we apply a Wiener-Filter combined with FWMD to the system, we can expect that the recognition performance of the system is improved.

The remainder of this paper is organized as follow. In Section II, HM-Net speech recognition system is briefly described. Section

III describes about Delay-and-Sum (D&S) beamforming, Limabeam algorithm, M. Sheltzer's and our proposed implementation method, postfiltering technique, and FWMD. The implementations of our proposed Unsupervised Limabeam algorithm in HM-Net speech recognition system, which combines FWMD with post filtering technique, is presented in Section IV. Section V evaluates the performance of these algorithms through a series of experiments performed using microphone array with 4 elements. Finally, we present some conclusions in Section VI.

II. Hm-net Speech Recognition System

HM-Net is a large-vocabulary, speaker -independent, HMM-based continuous speech recognition system [8]. The core structure of HM-Net is using Successive State Splitting algorithm to handle the problem of limited training data and unseen contexts in training step. HM-Net, shown in Figure 2, is a network of Hidden Markov Model which represents context dependent left-to-right HMMs in an efficient way. HMMs in HM-Net have various state lengths and share their state each other. Each node of the network is corresponding to an HMM state and has following information

- State number,
- Acceptable context class,
- Lists of preceding states and succeeding states,
- Parameters of the output probability density distribution,
-State transition probabilities.

In the HM-Net, if a phoneme context of a sample is given, the model corresponding to the context can be determined by concatenating several states, each of which can accept the context, using the restriction of the preceding state list and the succeeding state list. Since this model and an HMM are equivalent, we can treat the HM-Net as well

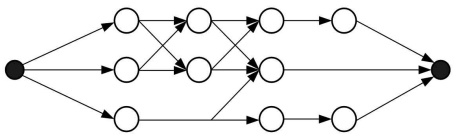

Fig. 2 HM-Net Architecture.

as common HMMs and thus so Baum-Welch re-estimation and Viterbi decoding can be applied to them [9,10,11].

III. The Limabeam Algorithm

3.1 Delay-and-Sum Beamforming

Delay-and-Sum (D&S) is the most popular and simplest method in microphone array processing. To process, one channel is chosen as a reference and the Time-Difference Of Arrival (TDOA) for the rest of the channels is estimated using Generalized Cross-Correlation (GCC) Phase Transform (PHAT) [12] or any other Time Delay Estimation (TDE) techniques. Next the time aligned speech signals are summed up as

$$y[n] = \sum_{m=0}^{M-1} \alpha_m x_m[n-\tau_m] \quad (1)$$

where M is the number of microphones, α_m is a weight applied to the signal received by microphone m. Usually, all . Usually, all $\alpha_m = 1/M$.

3.2 Limabeam Algorithm

Limabeam uses the filter-and-sum approach for beamforming. The signal captured by each microphone is time-delay-compensated and then convoluted with its associated filter and the filtered outputs are then summed up together

$$y[n] = \sum_{m=0}^{M-1} \sum_{p=0}^{P-1} h_m[p] x_m[n-p] \quad (2)$$

where $x_m[n]$ is time-delayed signals, $y[n]$ is the output signal, $h_m[n]$ is the filter associated with channel m. All filter coefficients for all

microphones are represented by

$$\zeta = [h_0[0],\ h_0[1],\ \ldots,\ h_{M-1}[P-2],\ h_{M-1}[P-1]]^T \quad (3)$$

The feature vector of the filter-and-sum output $y[n]$, $Z = \{z_1, z_2, \ldots, z_T\}$ is the function of both the incoming speech and the array processing parameters. There cognizer chooses a hypothesis $\hat{\omega}$ according to

$$\hat{\omega} = \arg\max_{\omega} P(Z(\zeta)|\omega)P(\omega) \quad (4)$$

Limabeam tries to optimize the recognition performance by finding the parameter vector ζ that maximizes the likelihood of the correct transcription of the utterance that was spoken. Assuming that the correct transcription ω_c is known, the optimal parameter vector $\hat{\zeta}$ can be selected by maximizing the acoustic log-likelihood of the given transcription ω_c

$$\hat{\zeta} = \arg\max_{\zeta} \log(P(Z(\zeta)|\omega_c)) \quad (5)$$

Let S_c represents the set of all possible state sequences and S represents one such state sequence, then the maximum likelihood estimate of ζ can written as

$$\arg\max_{\zeta, s \in S_c} \left\{ \sum_i \log(P(Z(\zeta)|s_i)) + \sum_i \log(P(s_i|s_{i-1}, \omega_c)) \right\} \quad (6)$$

According to (6), in order to find $\hat{\zeta}$ is a weight applied to the signal received by microphone m. Usually, all

3.3 M. Seltzer's Implementation Method

In the previous section, the theory of Limabeam algorithm was presented in which prior knowledge of correct transcription is required in order to maximize its likelihood, but if the prior knowledge of correct transcription is known, we do not need to do any recognition. To solve this problem, M. Seltzer proposed two different implementations of Limabeam in practice. The first method, called Calibrated Limabeam, is appropriate for situations in which the environment and the user's position do not vary significantly over time while the second method, called

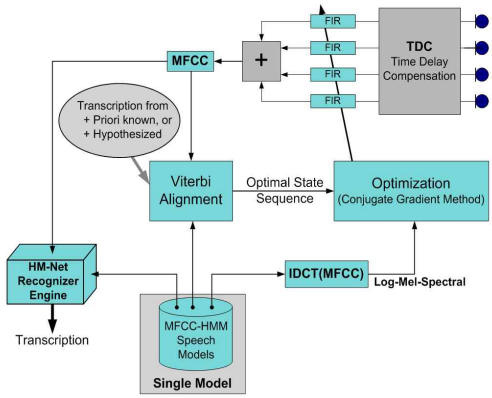

Fig. 3 Flowchart of proposed implementation of Limabeam algorithm in HM-Net Speech Recognition System.

Unsupervised Limabeam, is more appropriate for time-varying environments. So, we choose an unsupervised Limabeam that can improve recognition performance in any situation of environment.

In Limabeam algorithm, array parameter optimization is performed in the LogMelSpec domain. For this reason, M. Seltzer employed a parallel set of HMMs trained on log mel spectra, rather than cepstral by using the Statistical Re-estimation (STAR) algorithm [12], which ensures that the two sets of models have identical frame-to-state alignments.

3.4 Proposed Implementation Method

LogMelSpec can be derived from MFCC feature by taking Inverse Discrete Cosine Transform (IDCT) of MFCC

$$Z_{LogMelSpec} = IDCT(Z_{MFCC}) \quad (7)$$

Based on (14), Limabeam algorithm can be implemented in another way, in which the optimal state sequence is estimated in cepstral domain, and do not need to employ the parallel model. LogMelSpec HMM model will be calculated directly from MFCC HMM model through IDCT and thus it is easy and efficient to implement. Figure 3 illustrates the flowchart

of this alternative implementation.

3.5 Post-filtering technique

Zenlinski, McCowan and Leukimmiatis [13, 14,15] proposed a Wiener post-filter to the output of beamformer for improving the system performance. By using the auto-spectral and cross-spectral densities between channels, this technique could enhance the output signal. It can be used in both spatial and frequency domain. The formula in frequency domain is presented as following

$$v_i(f) = s(f) + n_i(f) \tag{8}$$

where s is the desired signal and n_i is the noise on microphone i. The general Wiener filter expression for a microphone array is given as

$$h_{opt}(f) = \frac{\Phi_{ss}(f)}{\Phi_{ss}(f) + \Phi_{nn}(f)} \tag{9}$$

where $\Phi_{ss}(f)$ and $\Phi_{nn}(f)$ are the auto-spectral density of the desired $s(f)$ and the noise at the output of the beamformer $n(f)$ respectively.

Practically, $\Phi_{ss}(f)$ is calculated as the average of auto - spectral densities and $\Phi_{ss}(f) + \Phi_{nn}(f)$ is the average of cross - spectral density between channels.

3.6 Feature Weighted Mahalanobis Distance

In our previous work [16, 17], the proposed Feature Weighted Mahalanobis Distance (FWMD) is effective to improve the performance of Limabeam algorithm. The FWMD basically gives less weight to noise features and higher weight to noise free features and written as

$$\forall i,j : D_{i,j}^{weighted} = \sum_{c=1}^{N} w_{i,j} D_{i,j}[c] \tag{10}$$

To improve speech recognition performance of the system, we can estimate the recognition performance using this distance after post filtering.

IV. Proposed Unsupervised Limabeam Algorithm

Limabeam algorithm is originaly developed by Michael L. Seltzer at CMU such as a framework for speech recognition system, another extension is that using N-best parallel model to maximum likelihood beamformers [18]. This system applies N-best parallel beamformers to multi-channel signals. In [19] K. Bahram et al proposed to use phone-based filter and sum parameter optimization, the parameters of array processing are adjusted in calibration phase based on phones used in language and maximum likelihood method.

In our works, to improve recognition performance of Limabeam algorithm we propose to combine FWMD and Wiener filter in Unsupervised Limabeam. We take the following steps:

1. Time-align the signals from the M microphones
2. Initialize ζ as $h_i[0] = 1/M,\ h_i[k] = 0,\ k \neq 0$
3. Process the signals using ζ to generate an output signal
4. Apply Wiener post-filtering to enhance the output signal
5. Perform speech recognition on the array output to obtain a word sequence in HM-Net
6. Determine optimal state sequence through the hypothesized transcription, array output signal, and HMM from speech recognition
7. Use optimal state sequence to estimate ζ using FWMD
8. If $L(\zeta)$ has not converged, go tos tep 3.

The flowchart of Unsupervised Limabeam combined with post filtering schema in HMNet speech recognition system is shown in Figure 4. From Figure 4, filter parameters were optimized afresh for each utterance in the following manner. Delay and Sum beamforming was used to process the array signals in order

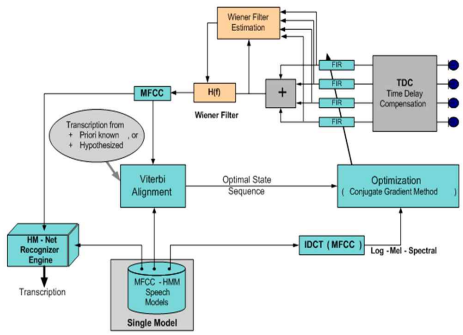

Fig. 4 Flowchart of Unsupervised Limabeam combined with post filtering schema in HMNet speech recognition system

to generate an initial hypothesized transcription. This signal was enhanced by a Wiener filter. Using this hypothesized transcription and the features derived from the enhanced output signal, the state sequence was estimated, based on this state sequence. A second iteration could be performed by generating the hypothesized transcription from these optimal filters. This process could be iterated until the likelihood converged.

V. Experimental Evaluations

HM-Net speech recognition system was used for all experiment in this paper. 1004 states (8 Gaussians/state) HM-Net system were trained using Trade database, a speaker-independent database consisting of 8892 utterances uttered by 90 speakers. The system was trained using 39-dimensional feature vectors consisting of 13 MFCC parameters, along with their delta and delta-delta parameters. A 25-ms window length and a 10-ms frame shift were used.

In order to investigate the performance of speech recognition with microphone array, we employed two microphone array databases recorded at Yeungnam University. In the first database, YUM4-6, we play backed Trade6 databases (596 utterances uttered by 6 speakers) through an Harman/Kardon loudspeaker and used a linear B&K microphone array with 4 elements spaced 20cm apart for recording. The distance between loudspeaker and microphone array is 150cm. This database was recorded in an office room which many noise sources, including computer fans, air-condition, voice, step and slam outside. Recording conditions are shown in Figure 5(a).

In the second database, YUM4-4, we used Trade4 database (396 utterances uttered by 4 speakers) to playback through one of four loudspeakers (2 Harman/Kardon and 2 Creative loudspeakers). The active speaker was selected randomly during recording. The microphones were put randomly on the table. This database was also recorded in the same room as before. Recording conditions are presented in Figure 5(b). The beam-pattern of linear microphone array is shown in Figure 6.

5.1 Experiments Using D&S beamforming

In the experiment, the channels were aligned based on time delays estimated by GCC-PHAT method. The aligned channels were then averaged to generate the delay-and-sum beamforming output signal. MFCC features were extracted from the output signal and passed to the recognizer for decoding. The single channel, Trade6 and Trade4 database were also decoded for comparison. The results are shown in Figure 7. As the figure indicates, D&S beamforming can improve the performance over single channel in both corpora (6.3% absolute for YUM4-6 corpus and 15.2% absolute for YUM4-4 corpus). In the figure, we can also see the results of closed tests representing 98.2% for Trade6 and 98.7% for Trade4.

5.2 Experiments Using Proposed Methods

In this case, filter parameters were optimized afresh for each utterance in the following manner. D&S beamforming was used to process the array signals in order to

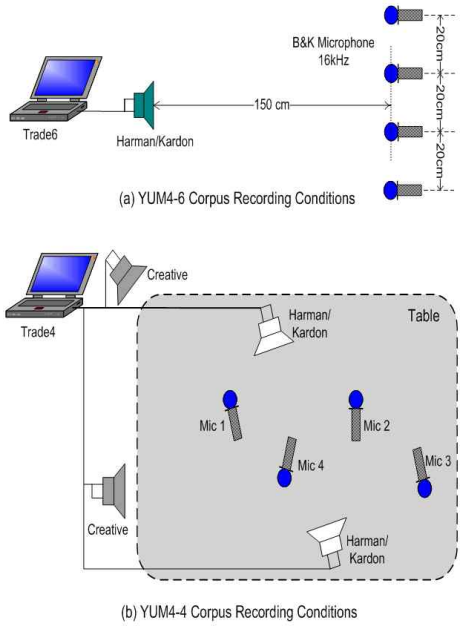

Fig. 5 YUM4-6 and YUM4-4 corpora recording conditions.

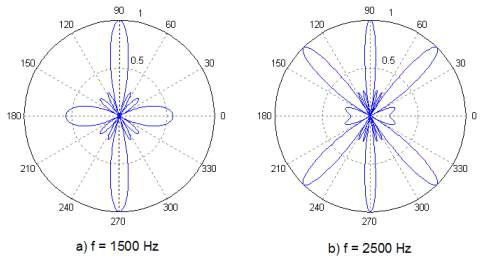

Fig. 6 The beam-pattern of linear microphone array, this array has 4-elements which were distributed equally at 20cm apart.

generate an initial hypothesized transcription. Using this hypothesized transcription and the features derived from the delay-and-sum output, the state sequence was estimated via forced alignment. The filter parameters were optimized based on this sate sequence. A second iteration can be performed by generating the hypothesized transcription from these optimal filters. This process can be interated until the likelihood converges.

Unsupervised Limabeam(USL) algorithm with

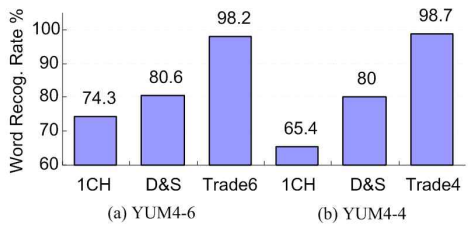

Fig. 7 Word Recognition Rate obtained using Delay-and-Sum Beamforming on the (a)YUM4-6 and (b)YUM4-4 corpora.

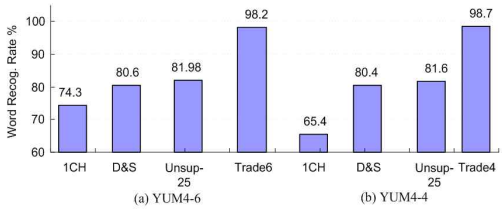

Fig. 8 Word Recognition Rate obtained using Unsupervised Limabeam on the YUM4-6 and YUM4-4 corpora with 25taps FIR filters. The results of single channel, D&S are also shown for references.

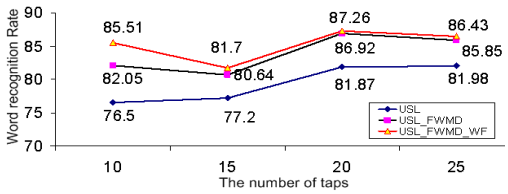

Fig. 9 Word Recognition Rate using USL-FWMD-WF compared with USL-FWMD and USL

25taps of FIR filters was estimated on both YUM4-6 and YUM4-4. The recognition results are presented in Figure 8. In both test sets, USL can improve the performance average 1.3% absolute over that of D&S

In the second experiment, our proposed Unsupervised Limabeam algorithm which combined FWMD with Wiener Filter(USL-FWMD-WF) was estimated. The results are compared to those from

Unsupervised Limabeam algorithm with FWMD (USL-FWMD). The experimental results are shown in Figure 9. We can see that the correct recognition rate of USL-FWMD-WF increases 1.4 % average approximately compared to USL-FWMD and the effectiveness of post filtering is proved. This is considered that when we use Wiener filter for the output signal, the incoherent signal components (noise and reverberant speech) are suppressed and the highly coherent speech signals are passed. So, the fluctuations in output signal that are caused by the higher recognition performance. Experimental results also show that proposed USL-FWMD-WF give approximately 5.8% higher recognition performance than USL.

Ⅵ. Conclusion

In this paper, we proposed an alternative way to implement Limabeam algorithm in Hidden Markov Network speech recognizer for efficient implementation and we proposed to add a post filter technique with Feature Weighted Mahalanobis Distance to Limabeam algorithm in order to improve recognition performance. From our prior investigation for the unsupervised Limabeam, we could see that because the performance of optimization depended strongly on the transcription output of the first recognition step, the output became unstable and that caused to lead lower performance. When we applied a post filter for the output signal, the incoherent signal components (noise and reverberant speech) were suppressed and the highly coherent speech signals were passed.

We estimated recognition performance of our proposed Unsupervised Limabeam algorithm which combined FWMD with Wiener Filter (USL-FWMD-WF). The results were compared to those from Unsupervised Limabeam algorithm with FWMD (USL-FWMD). We could see that the correct recognition rate of USL-FWMD-WF was increased approximately 1.4% average compared to USL-FWMD, 5.8% to USL, and the effectiveness of post filtering was proved. Experimental results also showed that USL-FWMD-WF gave approximately 5.8% higher recognition performance compared to D&S algorithm in case of 25 taps filtering.

References

[1] M.F. Font, “Multi-microphone signal processing for automatic speech recognition in meeting rooms,” Master thesis, Berkeley, California, 2005.

[2] P.J. Moreno, “Speech recognition in noisy environments,” Doctoral dissertation, Carnegie Mellon University, Pittsburgh, PA, 1996.

[3] F.H. Liu, “Environmental adaptation for robust speech recognition,” Doctoral dissertation, Carnegie Mellon University, Pittsburgh, PA, 1994.

[4] A. Acero, “Acoustical and environmental robustness in automatic speech recognition,” Doctoral dissertation, Carnegie Mellon University, Pittsburgh, PA, 1990.

[5] L.J. Griffiths, C.W. Jim, “An alternative approach to linearly constrained adaptive beamforming,” IEEE Transaction on Antennas and Propagation, Vol. AP-30, No. 1, pp.27-34, 1982.

[6] M. Seltzer, “Microphone Array Processing for Robust Speech Recognition,” Doctoral dissertation, Carnegie Mellon University, Pittsburgh, PA, 2003.

[7] P. Placeway, S. Chen, M. Eskenazi, U. Jain, V. Parikh, B. Raj, M. Ravishankar, R. Ronsenfeld, K. Seymore, M. Siegler, R. Stern, E. Thayer, “The 1996 hub-4 sphinx-3 system,” Proceedings on the DARPA Speech Recognition workshop, Vol. 1, pp.243-252, 1997.

[8] S.J. Oh, C.J. Hwang, H.Y. Jung, H.Y. Chung, “A study on statistical language models for large vocabulary continuous speech recognition system,” Proceedings on ICSP, Vol. 1, pp.

113-119, 1999.

[9] W.H. Press, B.P. Flannery, S.A. Teukolsky, W.T. Vetterling, "Numerical Recipes in C: The Art of Scientific Computin," New York: Cambridge University Press, 1998.

[10] L. Rabiner, B.H. Juang, "Fundamentals of Speech Recognition," New Jersey: Prentice Hall, 1993.

[11] A.J. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," IEEE Transactions on Information Theory, Vol. 13, No. 2, pp.260-269, 1967.

[12] P. Morento, B. Raj, R.M. Stern, "A Unified Approach for Robust Speech Recognition," Proceedings of Eurospeech, Vol. 1, pp.481-485, 1995.

[13] R. Zenlinski, "A microphone array with adaptive post-filtering for noise reduction in reverberant room," Proceedings on International Conference of Acoustics, Speech, and Signal Processing, Vol. 5, pp.2578-2581, 1988.

[14] I.A. McCowan, "Robust speech recognition using microphone arrays," Doctoral dissertation, Queen land University of Technology, Australia, 2001.

[15] C.H. Knapp, C. Carter, "The generalized correlation method for estimation of time delay," IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. 24, No. 4, pp.320-327, 1976.

[16] D.C. Nguyen, H.Y. Chung, "Performance Improvement of Microphone Array Speech Recognition using Feature Weighted Mahalanobis Distance," The Journal of the Acoustical Society of Korea, Vol. 29, No. 1E, pp.45-53, 2010.

[17] N.D. Cuong, S. Guanghu, J.H. Youl, C.H. Yeol, "Performace improvement of speech recognition system using microphone array," Proceedings on IEEE International Conference of Research, Innovation and Vision for the Future, pp.91-95, 2008.

[18] L. Brayda, C. Wellekens, M. Omologo, "N-Best Parallel Maximum Likelihood Beamformers for Robust Speech Recognition," Proceedings of European Signal Processing Conference, 2006.

[19] K. Bahram, B. Hamidreza, R. Farbod, "Improvement in speech recognition using phone-based filter and sum parameter optimuization," IEICE Electronics Express, Vol. 6, No. 8, pp.437-442, 2009.

저 자 소 개

Dinh Cuong Nguyen

2003: B.E. degree in computer science from Hanoi university of technology, Vietnam

2008: M.S. degree in the department of information and communication from Yeungnam university, Korea.

Current: Ph.D. candidate in the department of information and communication engineering from Yeungnam university, Korea.

Research Interests: include artificial intelligence, computer vision, and speech recognition.

Email: cuongnd@ntu.edu.vn

Suk Nam Choi

1995: B.E. degree in the department of electronics engineering from Yeungnam university.

2007: M.S. degree in the department of information and communication engineering from Yeungnam university.

Current: Ph.D. candidate in the department of information and communication engineering, Yeungnam university.

Research interests: include speech recognition in adverse environment.

Email: suknam69@nate.com

Hyun Yeol Chung

1975: B.E. in the department electronics engineering from Yeungnam university

1981: M.S. degree in the department of electronics engineering from Yeungnam university.

1989: Ph.D. degree in the department of information engineering from Tohoku university, Japan.

Current: professor in the department of the information and communication, Yeungnam University.

Research interests: include speech analysis, synthesis and recognition.

Email: hychung@yu.ac.kr