논문 2013-08-27

Improvement of Recognition Performance for Limabeam Algorithm by using MLLR Adaptation

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

Abstract : This paper presents a method using Maximum-Likelihood Linear Regression (MLLR) adaptation to improve recognition performance of Limabeam algorithm for speech recognition using microphone array. From our investigation on Limabeam algorithm, we can see that the performance of filtering optimization depends strongly on the supporting optimal state sequence and this sequence is created by using Viterbi algorithm trained with HMM model. So we propose an approach using MLLR adaptation for the recognition of speech uttered in a new environment to obtain better optimal state sequence that support for the filtering parameters' optimal step. Experimental results show that the system embedded with MLLR adaptation presents the word correct recognition rate 2% higher than that of original calibrate Limabeam and also present 7% higher than that of Delay and Sum algorithm. The best recognition accuracy of 89.4% is obtained when we use 4 microphones with 5 utterances for adaptation.

Keywords : Limabeam, Calibrate Limabeam algorithm, MLLR adaptation

I. INTRODUCTION

In recent years, speech recognition using microphone array has been considered as a useful approach to improve the recognition performance of captured speech. Many methods were proposed for microphone array speech recognition. For example, Delay and Sum (D&S) Beamformer [1], in which delays were inserted in each channel to compensate the difference in travel time between the desired sound source and the various sensors before summing them to give a single enhanced output channel. Post-filtering algorithms proposed by Zelinski [2], McCowan [3] and Leukimmiatis [4] use the input channel auto- and cross-spectral densities to estimate a Wiener post-filter being applied to the beamformer output. All of microphone array processing approaches described above designed for signal enhancement (by the improvement of Signal to Noise Ratio). However, most of speech recognition systems do not interpret waveform-level information directly. It is statistical pattern classifier that operates on a sequence of features derived from the waveform. Therefore, such techniques described above might not improve the Word Error Rate (WER) significantly [5]. Recent work of Michael L. Seltzer shows that in microphone array speech recognition, the WER 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), 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 accurate recognition; no assumption about the

* Corresponding Author (hychung@yu.ac.kr)
Received: 19 Feb. 2013, Revised: 10 Apr. 2013, Accepted: 16 Apr. 2013.
* This work was supported by the 2013 Yeungnam University research grant.

ISSN : 1975-5066

interfering signals; requiring no priori knowledge of microphone array geometry, room configuration, speaker-to-receiver impulse responses, etc. Limabeam is mainly a data-driven approach [5]. But the problem is that the feature parameters used for calculating distance can be distorted by noise in this algorithm [6]. To improve the performance of Limabeam, Cuong etc.[7] proposed FWMD(Feature Weighted Malanobis Distance) method and presented better performance. In FWMD, the performance of optimization strongly depends on the transcription output of the first recognition step. When performing unsupervised Limabeam on an utterance with too few correctly hypothesized labels the system accuracy degrades performance. In order to improve the accuracy of unsupervised Limabeam, the system needs to employ some post-filter techniques to obtain more correct transcription output of the first step.

To solve this problem, we investigate Limabeam algorithm and propose a method to improve recognition performance of this algorithm by using Maximum-Likelihood Linear Regression (MLLR) adaptation for finding the optimal state sequence support for estimation of the FIR filtering parameters.

The organization of the paper is as follows: In section II, we give a brief review of Limabeam algorithm. In section III, we present MLLR technique and a system using MLLR adaptation in Limabeam is described in section IV. A series of experiments are performed in section V. Finally, we will present our conclusion in section VI.

II. SPEECH RECONITION SYSTEM USING LIMABEAM ALGORITHM

The goal of array processing is to produce a distortion-free waveform. On the other hand, the goal of speech recognizer is to hypothesize the correct transcription of the utterance that was spoken. Michael L.Seltzer overcame this problem by proposing Limabeam. In this algorithm, Seltzer do not use the pipeline structure of classical microphone array processing techniques by using the information from the speech recognition system itself to find the array parameters that improve the speech recognition performance. Limabeam uses an adaptive Filter and Sum Beamformer. Assuming the filters have a finite impulse response (FIR), Filter-and-Sum processing is expressed mathematically as

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

Where $h_m[p]$ is the p^{th} tap of the filter associated with microphone m, $x_m[n]$ is the signal received by microphone m, τ_m is the steering delay induced in the signal received by microphone m to align it to the other array channels, and $y[n]$ is the output signal generated by the processing. P is the length FIR filter. All filter coefficients for all microphones are presented by

$$\zeta = [h_o[0], h_o[1], ..., h_{M-1}[P-2], h_{M-1}[P-1]] \quad (2)$$

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

$$\hat{\omega} = \arg\max_{\omega} P(Z(\zeta)|\omega) P(\omega) \quad (3)$$

Limabeam algorithm finds the parameter vector ζ for optimal recognition performance. Let us assume that the correct transcription of the utterance ω_c is known. The optimal parameter vector ζ can be calculated such as

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

Let S_c is the set of all possible HMM state sequences and s is one such state sequence, maximum likelihood estimated of ζ that can be written as

$$\hat{\zeta} = \arg\max_{\zeta, s \in S_c} \left\{ \sum_i \log(P(z_i(\zeta)|s_i)) + \sum_i \log(P(s_i|s_{i-1}, \omega_c)) \right\} \quad (5)$$

According to (5), in order to find $\hat{\zeta}$, the likelihood of the correct transcription must be jointly optimized with respect to both the array parameters and the state sequence.

III. MLLR ADAPTATION

Maximum likelihood linear regression (MLLR)[8] was originally developed for speaker adaptation[9] but can equally be applied to situations of environmental mismatch. A set of transformation matrices are estimated which are applied to the Gaussian mean parameters. We have extended the approach so that Gaussian variances can also be updated.

1. Mean adaptation

Mean adaptation of MLLR is based on affine transformation. Let μ_q and Σ_q be the mean vector and the covariance matrix of the output probability distribution of state q, respectively. For given training samples $O=\{o_1,o_2,...,o_T\}$, the new mean vector $\widehat{\mu_q}$ is estimated as follow:

$$\widehat{\mu_q}=A_q\mu_q+b_q=W_q\zeta_q \tag{6}$$

Where ζ_q is an extended mean vector defined by $\zeta_q=[1\ \mu_q^T]^T$, and W_q is the regression matrix for the mean vector. Define $\gamma_q(t)$ as the probability of generating o_t in state q at the time t, given the observation sequence O and the mode λ:

$$\gamma_q(t)=\frac{1}{P(O|\lambda)}\sum_{\theta\in\Theta}P(O,\theta_t=q|\lambda) \tag{7}$$

Where $\Theta$is the set of all possible state sequence $\theta=\{\theta_1,\theta_2,...,\theta_T\}$ of length T. Then the regression matrix W_q is found by solving the following equation

$$\sum_{t=1}^{T}\sum_{\gamma=1}^{R}\gamma_{qr}(t)\Sigma_{qr}^{-1}o_t\zeta_{qr}^T=\sum_{t=1}^{T}\sum_{\gamma=1}^{R}\gamma_{qr}(t)\Sigma_{qr}^{-1}\zeta_{qr}\zeta_{qr}^T \tag{8}$$

Where R is the number of states which share the regression matrix W_q.

2. Variance adaptation

Covariance matrix adaptation is performed after the mean adaptation. Let $\widehat{\Sigma_q}$ be the adapted covariance matrix

$$\widehat{\Sigma_q}=B_q^TH_qB_q \tag{9}$$

Where $C_q$is the Choleski factor of Σ_q^{-1} and $B_q=C_q^{-1}$. Regression matrix for the covariance matrix H_q is estimated by

$$H_q=\frac{\sum_{\gamma=1}^{R}C_{qr}^T\left[\sum_{t=1}^{T}\gamma_{qr}(t)(o(t)-\widehat{\mu_{qr}})(o(t)-\widehat{\mu_{qr}})^TC_{qr}\right]}{\sum_{\gamma=1}^{R}\sum_{t=1}^{T}\gamma_{qr}(t)} \tag{10}$$

MLLR can be applied in the recognition system to generate the labeling and updates the model parameters adapted with new recognition environment by using the utterance of new recognition condition. In this paper we applied this MLLR to Limbeam algorithm to improve performance of recognition system.

IV. Limabeam with MLLR adaptation

To improve recognition performance a ML LR adaptation was applied to the system with Limabeam algorithm (Limabeam-MLLR) to create better HMM models in optimal filtering process step so that the system can be adapted to a new environment efficiently. By applying this process, a better optimal state sequence will be obtained by Viteribi algorithm and can be used to support for estimating filter parameters. The Limabeam-MLLR algorithms can be stated as follow:

○ *Calibrate Limabeam-MLLR*

1) Time-align the signal from the M microphones.
2) Initialize ζ as $h_i[0]=1/M,\ h_i[k]=0,\ k\neq 0$
3) Process the signal using ξ to generate an output signal
4) Use MLLR adaptation to create new HMM models with output utterance.

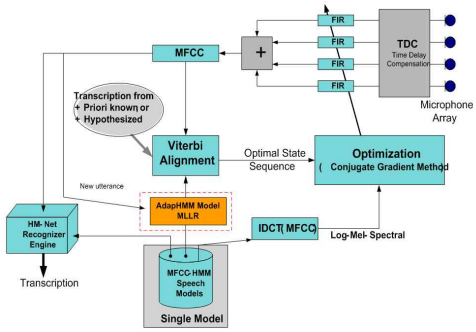

Fig. 1 System block diagram implementing MLLR adaptation to Limabeam algorithm.

5) Determine optimal state sequence through the calibrated transcription, array signal, and a new HMM model as a speech recognizer.
6) If $L(\zeta)$ has not converged, go to step 3.

In the algorithms above, optimization of array parameters is performed in the Log-Mel Spectrum domain. Figure 1 shows the flow chart of our proposed Limabeam-MLLR for improving recognition performance of Limabeam algorithm.

V. Experiments

1. System Implementation

To implement Limabeam algorithm with MLLR adaptation we use the HM-Net speech recognition system including 1000 states (8 Gaussian/state). HM-Net system is trained using TRADE database, a speaker-independent database consisting 8892 utterances spoken by 90 Korean speakers. Each utterance consists of Korean words. The system is trained using 39-dimensional feature vectors consisting of 13 Mel Frequency Cepstral Coefficients (MFCC) parameters, along with their delta and delta-delta parameters. A 25-ms window length and a 10-ms frame shift are used.

In order to investigate the performance of speech recognition with microphone array, we use 4 microphones. The recording conditions

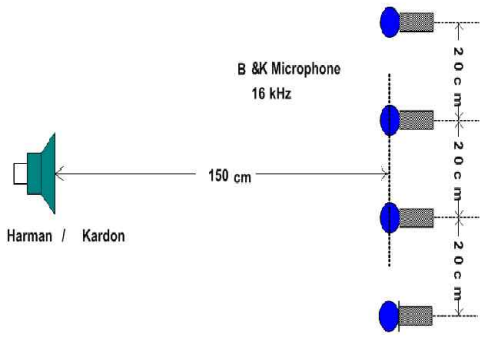

Fig. 2 Recording and recognition condition

are shown in figure 2. The space between each element is 20 cm and the distance between loud-speaker to center microphone is 150 cm. The data is recorded in a room which has many noise sources including computer fan, air-condition, voice, step and slam outside. 596 utterances uttered by 6 speakers that recorded in above environment are used for testing.

For implementation of calibrate Limabeam-MLLR, one utterance from each speaker was used to adapt HMM model by MLLR method, this utterance also used to estimate filter parameters. This constituted first iteration of calibration.. The second iteration of calibration was performed by using estimated filter parameters to initialize in step 2. The calibration process continued in an iterative manner until the overall likelihood converged.

2. The results

In the first experiment, we estimated recognition performance of calibrate Limabeam-MLLR. The experimental results along with the number of taps for filtering for Delay and Sum, calibrate Limabeam, and calibrate Limabeam-MLLR are shown in figure 3.

As we can see in Figure 3, the correct recognition rate for calibrate Limabeam-MLLR shows approximately 2% higher than that of the original calibrate Limabeam and also shows

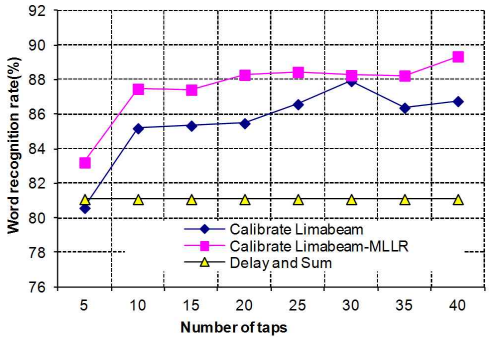

Fig. 3 Compared word recognition rates for Calibrate Limabeam, Calibrate Limabeam -MLLR, and Delay and Sum

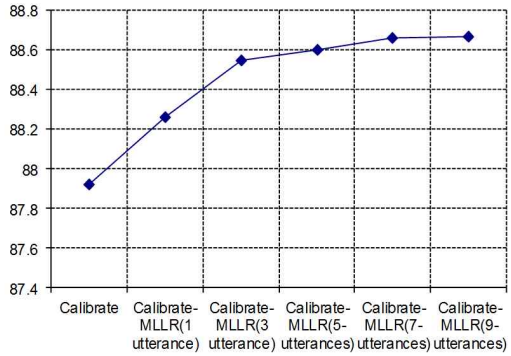

Fig. 4 Word Recognition Rate using Calibrate Limabeam-MLLR with difference of the number of utterances for adaptation.

7% higher than Delay and Sum algorithm. This means that the optimal state sequence of HMM models by Viterbi process is adapted well to a new environment and thus leads to increase the recognition performance of the system when we use MLLR adaptation.

The second experiment was carried out to verify the dependence of performance on the number of utterances for adaptation. Figure 4 shows that number of utterances increases the performance until the number exceeds 5. When the number exceeds 5, no additional performance improvement was achieved due to fully adapted optimal state sequence of HMM model. As we can see in Figure 4, Recognition rate was converged to around 88.7% even if

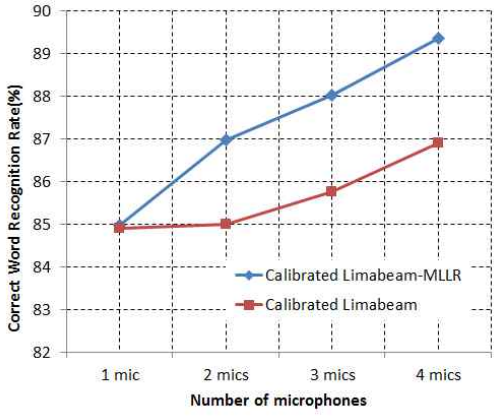

Fig. 5 Word recognition rates for calibrate Limabeam-MLLR and calibrate Limabeam depending on the number of microphones(in case of 40 taps of filters)

we increase the number 7 to 9.

The third experiment was performed to check the dependence of performance on the number of microphones for adaptation of the system. The experimental results are presented in figure 5. A higher recognition performance was obtained when we increased the number of microphones. The best recognition accuracy of 89.4% was obtained when we used 4 microphones with 5 utterances and 40 taps of filers for adaptation.

VI. Conclusion

In this paper, we investigated adding a MLLR adaptation into Limabeam (Limabeam-MLLR) algorithm to improve recognition performance. By using this method, we obtained the optimal state sequence close to new recognition environment. It supports strongly for filtering parameters optimal processing.

To verify the performance improvement of recognition rate of Limabeam-MLLR the compared experiments for Delay and Sum, calibrate Limabeam, and calibrate Limabeam-MLLR were carried out. It turned

out that Limabeam-MLLR gave approximately 2% higher recognition than that of original Limabeam, and is also 7% higher accuracy than that of Delay and Sum algorithm. This means that by using MLLR adaptation the optimal state sequence of HMM models by Viterbi process is adapted well to new environment and thus leads to increase the recognition performance of the system.

In the experiment for verifying the dependency of performance on the number of utterances for adaptation, we found that no additional performance improvement was achieved for the number of utterances exceeds 5. In the experiment to check the dependency of performance on the number of microphones for adaptation, A higher recognition performance was obtained when we increased the number of microphones. The best recognition result of 89.4% was obtained when we used 4 microphones with 5 utterances and 40 taps of filers for adaptation.

References

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

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

[3] I.A. McCowan, H. Bourlard, "Microphone array post-filter for diffuse noise filed," Proceeding of International Conference on Acoustics, Speech, and Signal Processing, Vol. 1, pp.905-908, 2002.

[4] S, Leukimmiatis, "An Optimum Microphone Array Post-Filter for Speech Application," Proceeding of International Conference on Spoken Language Processing, Inter-speech, pp.2142-2145, 2006.

[5] M.L. Seltzer, B. Ra, R.M. Stern, "Likelihood - Maziming Beamforming for Robust Hands-Free Speech Recognition," IEEE Transactions on Speech and Audio Processing, Vol. 12, No. 5, pp.489-498, 2004.

[6] M. Wolfel, H.K. Ekenel, "Feature weighted Mahalanobis distance: improved robustness for gauusian classifiers," Proceedings on European Signal Processing conference, 2005.

[7] 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.

[8] M.J.F. Gales, P.C. Woodland, "Mean and Variance Adaptation within the MLLR Framework," Cambridge Press, 1996.

[9] A.J. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," IEEE Transactions on Information Theory, Vol. 13, No. 2, pp.260-267, 1967.

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