에코제거기와 MAP 추정에 기초한 핸즈프리 음성인식

Hands-free Speech Recognition based on Echo Canceller and MAP Estimation

김 성 일*, 신 위 재*
Sung-ill Kim, Wee-jae Shin

요약

핸즈프리 마이크를 이용한 원격회의나 원격 통신 시스템과 같은 몇 가지의 응용분야에서, 음성 신호는 주위 잡음뿐만 아니라 마이크와 스피커사이의 결합에 의해 발생하는 에코에 의해서 왜곡되기 쉽다. 게다가 채널 왜곡이나 부가적인 잡음을 포함한 환경 잡음들은 원래의 입력 음성신호에 영향을 미치리라 고려된다. 본 논문에서는, 이러한 핸즈프리 음성에 있어서의 음성 인식률을 향상시키기 위해 에코 제거기와 최대 사후 추정(MAP)을 이용한 새로운 접근방식을 소개한다. 이 접근방식에서, 제안된 시스템이 에코를 포함한 주위 잡음 환경에서의 핸즈프리 음성인식에 효과적이라는 것을 보여준다. 또한, 실험 결과는 에코 제거기와 MAP 환경적응 기술의 결합 시스템이 에코와 잡음 환경에 잘 적응하는 것을 보여준다.

Abstract

For some applications such as teleconference or telecommunication systems using a distant-talking hands-free microphone, the near-end speech signals to be transmitted is disturbed by an ambient noise and by an echo which is due to the coupling between the microphone and the loudspeaker. Furthermore, the environmental noise including channel distortion or additive noise is assumed to affect the original input speech. In the present paper, a new approach using echo canceller and maximum a posteriori(MAP) estimation is introduced to improve the accuracy of hands-free speech recognition. In this approach, it was shown that the proposed system was effective for hands-free speech recognition in ambient noise environment including echo. The experimental results also showed that the combination system between echo canceller and MAP environmental adaptation technique were well adapted to echo and noise environment.

Key words : Hands-free speech, Speech Recognition, MAP estimation, Echo canceller, Noise Environment

Ⅰ. Introduction

In the past few years, many works have been performed in hidden Markov model(HMM) to improve speech recognition accuracy with a close-talking microphone. Especially recently, however, the dissemination of hands-free communication systems requires to provide users with some comfort. Therefore, the problems of reverberant speech recognition have to be solved to obtain a good speech recognition accuracy when using a distant- talking microphone. In those hands-free speech recognition[1,2,3,4], the sound from the loudspeaker is picked up by the microphone directly or indirectly since the microphone and the loudspeaker are coupled. This is heard by the recognizer as echo, causing unwanted recognition results.

If the hands-free mode is to be used, we inevitably face with the problem of environmental noise including channel distortion or additive noise in addition to the acoustic echo. Consequently, the observed signal is

*경남대학교 공과대학 전기전자공학부
접수 일자 : 2003. 4. 16 수정 완료 : 2003. 7. 21
논문 번호 : 2003-2-10
※This work is supported by the Kyungnam University research fund, 2003

$$y(t) = s(t) * h(t) + n(t) \quad (1)$$

where n(t) is the total additive noise and h(t) is the channel mismatch. Fig. 1 shows how the convolutional channel noise and the additive noise are assumed to affect the clean speech. Fig. 2 shows one of the application example of the teleconference in which hands-free microphone picks up at the remote end both the desired speech and the undesired ambient room noise such as air conditioning, heating or ventilating systems and computer fans, etc.

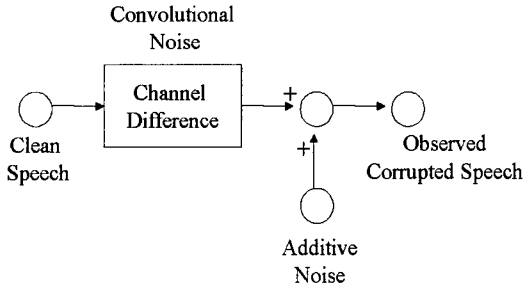

Fig. 1. Channel distortion and additive noise environments

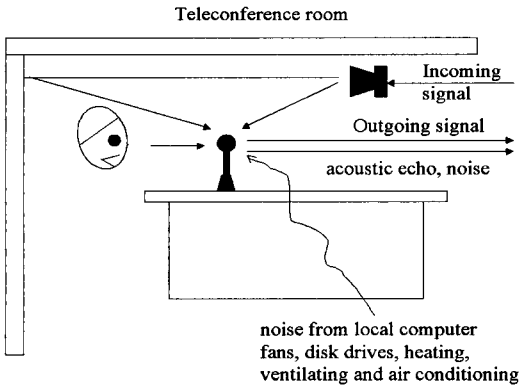

Fig. 2. Echo and ambient noise environments in teleconference application

Furthermore, this figure shows that the speech would be mixed with echo from the speaker as well as noise present within the teleconference room. The echo generally find it difficult to eliminate it if there are high level background noise.

In this paper, we report the implementation of new approach for hands-free speech recognition using echo canceller and maximum a posteriori(MAP) estimation. The acoustic echo canceller[5,6,7,8] is used in the experiment so that hands-free microphone picks up only the desired speech and removes the undesired echo. The speech recognizer using HMM then improves the echo-cancelled speech recognition accuracy by using environmental adaptation technique in the noisy environment with channel distortion and additive noise. For this purpose, the MAP estimation technique [9,10,11] is extended to an environmental adaptation[12] for the high performance of hands-free speech recognition. In addition, we report the speech recognition rates in hands-free mode in comparison with several kinds of modes.

II. The Basic idea of Echo Canceller

Acoustic echo was first encountered with the early video/audio conference studios. In this situation, the sound from the loudspeaker is heard by listener as intended. However, this same sound is also picked up by the microphone both directly and indirectly. The result of this reflection is the creation of multipath echo and multiple harmonics of echo, unless there are eliminated, which are transmitted back to the microphone and are heard by talker as echo. To eliminate it, the echo cancellation device at the near end uses a complex adaptive digital filter on the transmitted signal that models the returning echo and cancels it by subtracting it from the returned signal. Words echoed from an outgoing prompt, for example, may be incorrectly recognised as having been said by the caller. In such an instance, excessive echo can reduce the effectiveness of automatic speech recognition systems.

Fig. 3 shows the basic operation of acoustic echo canceller.

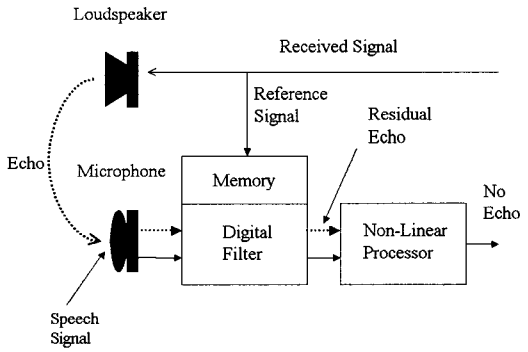

Fig. 3. Hardware design of an acoustic echo canceller

In the echo cancellation, the complex algorithmic procedures are used to compute speech models. This

involves the system generating the sum from the reflected echoes of the original speech, then subtracting this from any microphone signal it picks up. The result is the purified speech of the person talking. The format of this echo prediction is learned by echo canceller in a process known as adaptation. It might be said that the parameters learned from the adaptation process generate the prediction of the echo signal which then forms an audio picture of the room in which the microphone is located.

During the conversation period, this audio picture constantly alters and in turn the canceller has to adapt continually. Therefore, the convergence time is required for the echo canceller to fully learn the acoustic picture of the room. Other important performance criteria involve the acoustic echo canceller's ability to handle acoustic tail circuit delay. This is the time span of the acoustic picture and roughly represents the delay in time for the last significant echo to arrive at the microphone. Another important factor is acoustic echo return loss enhancement(AERLE). This is the amount of attenuation which is applied to the echo signal in the process of echo cancellation. If no attenuation is applied, for example, full echo will be heard. The canceller's performance also relies heavily on the efficiency of a device called the center clipper or non-linear processor. This needs to be adaptive and has a direct bearing on the level of AERLE that can be achieved.

Ⅲ. Environmental Adaptation based on Maximum A Posteriori

The MAP estimation is also called bayesian successive estimation of HMM parameters for the new speaker in a framework. The estimated mean vector value after given N samples is shown as

$$\hat{\mu}_N = \frac{\alpha\mu_o + \sum_{i=1}^{N} X_i}{\alpha + N} \qquad (2)$$

where α is an adaptation parameter. The estimated covariance matrix using N samples is

$$\sum_N = \frac{1}{\beta + N}\{X_N X_N^T - (\alpha + N)\mu_N \mu_N^T + (\beta + N - 1)\sum_{N-1} + (\alpha + N - 1)\mu_{N-1}\mu_{N-1}^T\} \qquad (3)$$

where β is a coefficient. In the present experiment, the values of αβ were set at 15 and 50 respectively, which were determined experimently.

Speaker adaptation is performed with successive training of speaker independent(SI) models using small amounts of adaptation speech samples. Generally, there are two adaptation methods. One is well-known as the supervised speaker adaptation which achieves the adaptation in accordance with correct label sequences. Another method is known as the unsupervised speaker adaptation. In this method, the label sequences are provided automatically by the viterbi segmentation as shown in Fig. 4. Moreover, the specific speaker's utterances are captured under the noisy environment to be given to MAP estimation algorithm so that the parameters of standard acoustic models are also adapted to the acoustic environment that speech recognition is performed as well as the utterances of specific speaker.

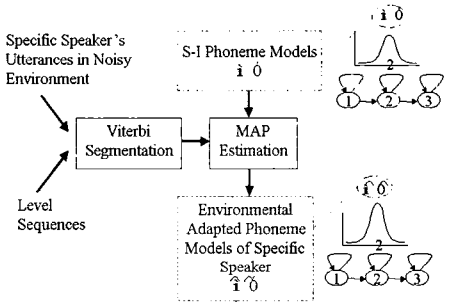

Fig. 4. Block diagram of unsupervised speaker adaptation under the noisy environment

Therefore, a block diagram of the overall hands-free speech recognition system based on acoustic echo canceller and MAP environmental adaptation is shown in Fig. 5.

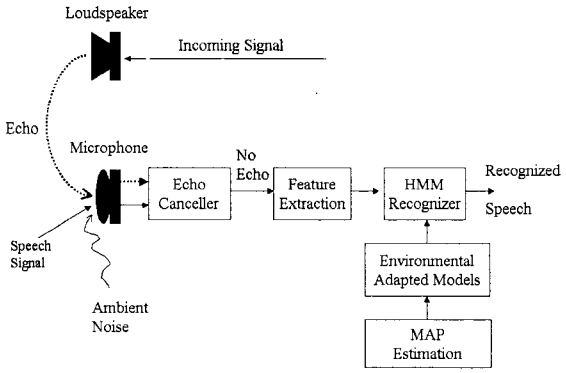

Fig. 5. Overall block diagram of hands-free speech recognition system

Ⅳ. Experimental Conditions

The Japanese 40 phoneme models were trained using 5240 labeled word utterances of 10 male speaker and 503 sentences of 6 male speakers from the ATR Japanese speech database, where 100 ATR phoneme balanced words were used for the test experiment. A set of 20 dimensional observation sequences including discrete duration information are obtained for recognition. Table 1 shows the preprocessing analysis condition of the speech signals and HMM topology.

sampling rate	16 KHz, 16 bit
preemphasis	0.97
window function	16msec Hamming window
frame period	5 ms
feature parameters	10 order MFCC +10 order delta MFCC +log power +delta log power +discrete duration information
model topology	3 state left-light phoneme model

Table 1. Analysis of speech signals

Fig. 6. shows the experimental conditions for three kind of speech-talking modes

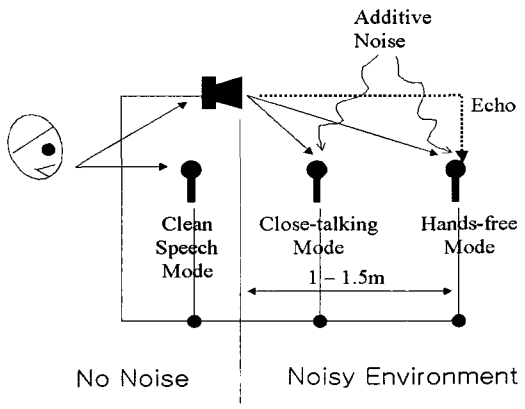

Fig. 6. Experimental conditions for three kinds of speech-talking modes

As shown in this Fig., the clean speech mode is the method using clean speech data and the models trained by clean speech database. In both close-talking and hands-free modes, the test data were also recorded using clean speech in laboratory room environments including ambient noise such as conditioning system, fans, and disk drives, etc as well as acoustic echo controlled by digital audio tape recorder. In hands-free mode for this experiment, the speech signals were captured by using a desk-top microphone that was located at a distance of 1.0-1.5m from the loudspeaker in the echo and noise environment.

Ⅴ. Experiments And Results

Table 2 illustrates the recognition rates of clean and close-talking modes in terms of the number of adaptation words. As can be seen in this table, the performance of the close-talking mode under the noisy environment is degraded in recognition rates in comparison of the clean speech mode. The experimental results also show that the clean speech mode had just 10% improvement in recognition rates in accordance with the increase of the number of adaptation samples, whereas it showed 76% dramatic improvement in the close-talking mode by increasing adaptation samples. Accordingly, it is revealed that the adaptation based on the algorithm of MAP estimation is well adapted to the noisy environment.

Number of Adaptation Samples	Speaking Mode	
	Clean	Close-talking
baseline	85	5
7	87	35
13	90	45
25	94	71
50	95	81

Table 2. Recognition rates in clean and close-talking modes

Table 3 shows the recognition rates of hands-free mode when using echo canceller and not using one. When the number of adaptation samples was increased to 50, for example, the recognition rates using echo canceller shows the significant improvement of 52% in comparison of the rates with no use.

Number of Adaptation Samples	Hands-free mode	
	no echo canceller	+ echo canceller
baseline	25	1
7	31	14
13	43	14
25	64	18
50	71	19

Table 3. Recognition rates in hands-free mode when using echo canceller and not using one

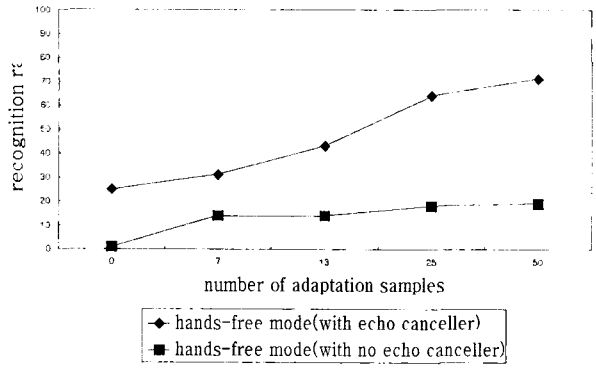

Fig. 7. Comparison of recognition rates between the system with echo canceller and the system with no use

Fig. 7 illustrates the comparison of recognition rates between the system with echo canceller and the system with no use. It shows that the rate with echo canceller is 24% higher that the one with no use in baseline(without any adaptation samples). Moreover, we can see that the echo-cancelled speech in hands-free mode is relatively well adapted to the noisy environment by MAP environmental adaptation technique. However, it might be thought that the rates with echo canceller are not so much satisfactory. It is mainly due to the over- cancel of input speech signals as well as the high level of echo and additive noise.

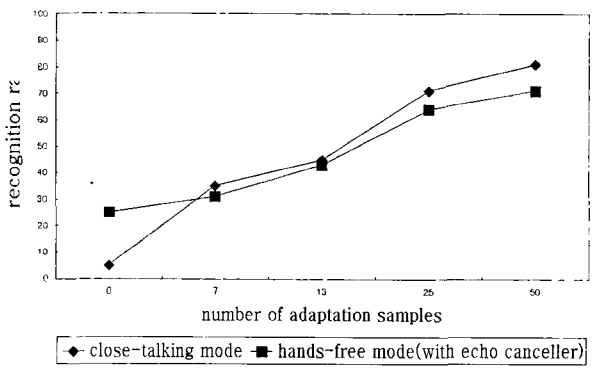

Fig. 8. Comparison of recognition rates between close-talking mode and hands-free mode with echo canceller in noisy environment

Fig. 8 shows the comparison of recognition rates between close-talking and hands-free mode with echo canceller in noisy environment. it represents that the recognition rate of the proposed hands-free mode is similar to the one of close-talking mode.

Ⅵ. Conclusions

This paper describes the new method of hands-free speech recognition based on acoustic echo canceller and MAP environmental adaptation technique. The experimental results indicated that the combination of echo canceller and MAP adaptation technique showed the possibility of the improvement in performance in echo and noisy environment. The results also showed that the acoustic models using the combination method were well adapted to the echo and noisy environment.

As one of the future works, it should be proved that the proposed system can also be optimized in the Korean speech recognition.

References

1. D.Giuliani., M.Matassoni, M.Omologo, and P.Svaizer: "Hands Free Continuous Speech Recognition in Noisy Environment using a Four Microphone Array", Proc. ICASSP, pp.860-863, 1995

2. T.Takiguchi, S.Nakamura, Q.Huo, and K.Shikano: "Model Adaptation based on HMM Decomposition for Reverberant Speech Recognition", Proc. ICASSP, pp.827-830, 1997

3. D.Giuliani, M.Matassoni, M.Omologo and P.Svaizer: "Training of HMM with Filtered Speech Material for Hands-free Recognition", Proc. of ICASSP, Vol. 1, pp.449-452, 1999,

4. J.Bitzer, K.U.Simmer, K.Kammeyer: "Multi-microphone noise reduction techniques for handsfree speech recognition - a comparative study", Robust Methods for Speech Recognition in Adverse Conditions (ROBUST-99), pp.171-174, 1999.

5. R.Martin, J.Altenhoner: "Coupled Adaptive Filters for Acoustic Echo Control and Noise Reduction", Proc. ICASSP, pp.3043-3046, 1995

6. M.Siqueira and A.Alwan: "New Techniques for Adaptive Filtering Applied to Speech Echo Cancellation", Proc. IEEE ICASSP, pp.265-268, 1994

7. S.Makino and Y.Kaneda: "Acoustic echo canceller algorithm based on the variation characteristics of a room impulse response," Proc. ICASSP'90, pp. 1133-1136, 1990
8. W.Kellerman: "Analysis and Design of Multirate Systems for Cancellation of Acoustical Echoes", Proc. ICASSP, pp.2570-2573, 1988
9. Y.Tsurumi, S.Nakagawa: "An Unsupervised Speaker Adaptation Method for Continuous Parameter HMM by Maximum A Posteriori Probability Estimation", Proc. ICSLP,pp.431-434, 1994
10. C.J.Leggetter and P.C.Woodland: "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models", Computer Speech and Language, Vol. 9, pp.171-185, 1995
11. C.Chesta, O.Siohan, C.-H.Lee: "Maximum a posteriori linear regression for hidden Markov model adaptation", Proc. European Conference on Speech Communication and Technology, vol. 1, pp.211-214, 1999.
12. J.H.Lee, B.K.Kim, H.Y.Chung: "Environmen- tal Adaptation using A Posteriori Estimation for Korean Word Recognition", Proc. IEEE Invited Workshop on Pattern Recognition for Multimedia Techniques, pp.49-52, October, 1996

Sung-Ill Kim
Member

He received his B.S. and M.S. degrees in the Department of Electronics Engineering from Yeungnam University, in 1997, and Ph.D. degree in the Department of Computer Science & Systems Engineering from Miyazaki University, Japan, in 2000. During 2000 to 2001, he was a postdoctoral researcher in the National Institute for Longevity Sciences, Japan. He worked in the Center of Speech Technology, Tsinghua University, China during 2001 to 2003. Currently, he is full-time lecturer in the Division of Electrical & Electronic Engineering, Kyungnam University since 2003. His research interests include speech/emotion recognition, neural networks, and multimedia signal processing.

E-mail: kimstar@kyungnam.ac.kr

Wee-Jae Shin
Member

He received B.S. degree in the department of electronics engineering from Dong-A University in 1975.

Also, M.S. degree from Dong-A University in 1979.
He received Ph.D. degree in the department of electronics engineering from Dong-A University in 1989.
He is currently a professor in the division of electronic engineering Kyungnam University. he work a vice-president of a KISPS.
his research interests include the areas of System Intelligence Control, Signal Processing.