ROOM ACOUSTICS AND REVERBERATION: IMPACT ON HANDS-FREE RECOGNITION

Satoshi NAKAMURA and Kiyohiro SHIKANO
Graduate School of Information Science, Nara Institute of Science and Technology
8916-5, Takayama-cho, Ikoma-shi, Nara, 630-01, JAPAN
nakamura@is.aist-nara.ac.jp

ABSTRACT

Hands-free speech recognition is a very important issue for a natural human machine interface. The distant talking speech in real environments is distorted by noise and reverberation of the room. This paper introduces characteristics of the room acoustical distortion and their influences on speech recognition accuracy. Then the paper tries to give a prospect of the solution based on previous studies and our research efforts. Especially a microphone array based-method and a model adaptation method are discussed. The microphone array can reduce the influences of the acoustical distortion by beam-forming. On the other hand, the model adaptation method can estimate the acoustical transfer function and adapt the speech models against the distorted observation signals. Furthermore, this paper also addresses hands-free speech recognition by incorporating automatic lip reading.

1. INTRODUCTION

Speech interface is the most important and natural interface for human to communicate intention each other. This speech interface is important not only for human-to-human communication but also human-to-machine communication. Hands-free speech recognition is an essential technology to realize the natural speech interface. The hands-free speech recognition realizes so natural and friendly man-machine interface that users are not encumbered by microphone equipments and that users can utter from distance while moving. This hands-free speech recognition is actually an urgent technology for the hands-free interface of a car navigation system and a cellular telephone in the car.

The accuracy of speaker independent speech recognition has been made a remarkable progress by the arrival of stochastic modeling of speech, HMM, and its training algorithms. Although the HMM brought a high recognition accuracy, a speaker must be equipped a close-talking microphone or forced to use a desk-top microphone. If the speaker inputs his speech from distance, the accuracy will be seriously degraded by the influences of the noise and reverberation of the room.

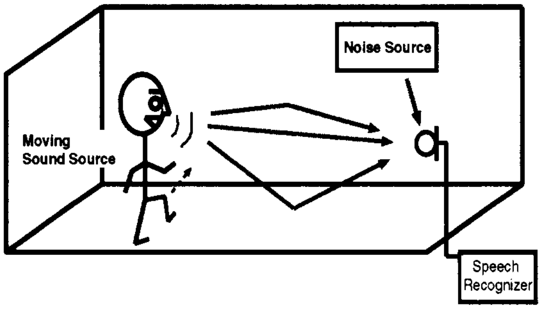

Figure 1. Real environments

The speech recognition performance even using a desktop microphone will be also varied if the distance between a mouth and a microphone is changed, and if the speaker turn his face to another direction. The fundamental problems of hands-free recognition already have lain in the previous speech recognition framework. To these problems, the following technologies are required,

- Robustness to directional noise and omnidirectional noise (diffuse noise) in the room.
- Robustness to acoustical reflection and reverberation in the room.
- Localization, tracing and recognition of the speaker among many sound sources including other speakers and noise.

These problems are quite new ones which previous studies haven't been considered. In fact, performance of current LVCSR will be seriously degraded if used in this hands-free context.

This paper introduces approaches and studies for hands-free speech recognition from a viewpoint of acoustical pre-processing and model adaptation. As acoustical pre-processing, a study using a microphone array, which can utilize a spatial difference of sound source is introduced[7, 8, 9, 11, 12, 14]. As for the microphone array research, detailed survey is reported by Omologo in [1]. The model adaptation approach is an extension of the technology used in speaker adaptation and channel adaptation.

This paper also addresses a different approach to hands-free speech recognition. That is multi-modal integration of audio and visual information. It is known that human integrates audio and visual information, called McGurk effect. It is also known that a human's eye is focused to mouth movements if audio information is insufficient. The paper briefly introduces studies of speech recognition incorporating automatic lip reading.

2. NOISE AND REVERBERATION

Factors which degrades speech recognition performance in the real room are summarized as influences by noise and reverberation. Noise are classified into directional

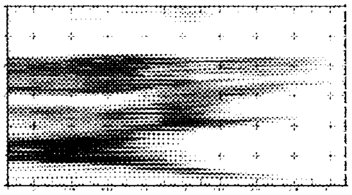

Figure 2. spectrogram of clean speech /ai/

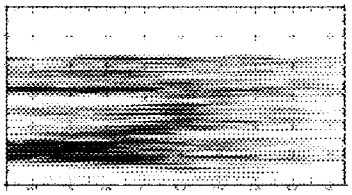

Figure 3. spectrogram of reverberant speech /ai/: T_{60}=600msec

noise and omni directional noise, which is called diffuse noise. The noises such as other speaker's speech and hard disk noise of a workstation is directional noise. On the other hands, the diffuse noise is the noise propagating from every direction to the microphone.

These kinds of noise affect serious influences on the speech recognition performance. Many works are presented. These approaches are summarized as a speech enhancement approach and a model adaptation approach. As for the speech enhancement approach, the spectral subtraction method for additive noise and the cepstral mean normalization method for multiplicative noise had been proposed and confirmed their effectiveness[17, 18]. As for the model modification approach, the conventional multi-template approach, and model adaptation approach[22] and the model (de-)composition approach[19, 21] have been proposed. However, these are based on a single channel microphone. If the noise source is directional, the target speech signal can be enhanced efficiently by a microphone array which has super directivity. We describe about a microphone array in later section.

The next problem is reverberation. The speech uttered from distance in the reverberant room will be corrupted by reflection of walls and delay of sound propagation. Fig.2 and fig.3 shows an example spectrogram of clean speech and reverberant speech. The reverberation is defined by impulse response. The influence of reverberation is also described by a scaler index of reverberation time, T_{60}, and DR-ratio which is a ratio of power of direct path and reflection. In general, impulse response will change according to not only shape of the room but also temperature and humidity. Since measurement of impulse response is difficult and troublesome, the image method are often used to simulate the impulse response.

The differences of a telephone channel, a microphone and a handset are also multiplicative factor and then degrade the speech recognition performance. However, the impulse response is relatively short compared to reverberation. The distortion can be restored by the processes within one analysis window such as cepstral mean normalization, stochastic matching, SDCN, CDCN, RATZ, PMC and HMM composition[18, 22, 3, 20, 21, 24].

The reverberation may affect beyond the analysis window of speech recognition. The basic methods for restoration of the reverberant speech are based on deconvolution by an estimated inverse filter. However the inverse filter can not be calculated stably since the impulse response is not generally minimal phase. The previous works have been studied from viewpoints of acoustical signal processing, which estimate the inverse filter by separating the impulse response into all pass filter and minimal phase filter, by multiple input and output system and by neural networks[26, 27, 28, 29]. Recently speech researchers began to look at this reverberation problem from viewpoints of auditory processing (RASTA[25]) and of model adaptation. In this paper, two methods based on a microphone array and HMM composition are described.

3. ACOUSTICAL PRE-PROCESSING

This section describes methods which deal with noise and reverberation in the real room by acoustical preprocessing. As noted in the previous section, the super directive microphone extracts the target speaker's signal reducing the influences of not only directional noise but also diffuse noise and reverberation. Furthermore, the moving speaker can be traced if accurate source localization is established. The super directivity and source localization can be realized by a digitally steerable microphone array. The microphone array has two functions such as source localization and beam forming. Source localization is carried out by estimating delay between outputs of microphones from target sound source. For this purpose relatively small number of microphones are used with wide aperture for precise spatial resolution. The delay is estimated by cross correlation and cross power spectrum phase[12, 13]. However, it is needed to utilize characteristics of speech like harmonics under the very low SNR condition[14]. On the other hand, large number of microphone is required to realize sharp beam form. The microphone element must be so spaced that the distance is shorter than half wavelength of target frequency(Spatial Sampling Theorem). The studies which use over 500 microphone elements[10] and optimize their spacing for speech recognition[16] have been reported.

4. 3-D SEARCH IN TIME, SPACE AND HMM STATES

In the conventional systems, speech recognition is carried out after localizing a speaker direction. However these two procedures should be performed simultaneously like procedures in human perception. A method to deal with speaker localization and speech recognition simultaneously in a unified framework is reported in [15]. This method finds an optimal combination of a transition of speaker directions and a phoneme sequence of speech. In general, an HMM-based speech recognition algorithm performs Viterbi search on trellis plane composed of input frames and HMM states. As an extension of this algorithm, a speech recognition is carried out based on Viterbi search on 3-D trellis space composed of speaker directions, input frames, and HMM states. Figure 4 shows 3-D trellis space, where x-, y-, and z-axis represent input frames,

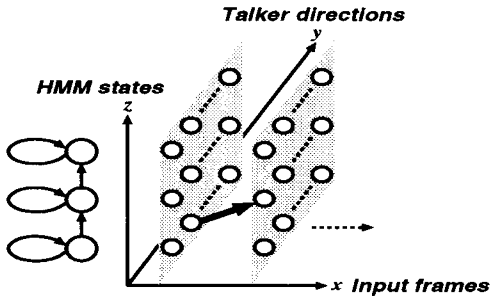

Figure 4. Viterbi search on 3-D trellis space

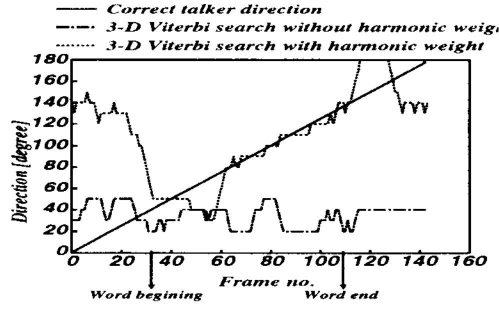

Figure 5. Example of a transition of talker directions obtained by 3-D Viterbi search 1 and 2 in SNR 20 dB

talker directions, and HMM states, respectively. As a result, a transition of speaker directions and a phoneme sequence of speech are obtained by finding an optimal path with the highest likelihood. This method can be extend to the simultaneous recognition of moving speakers and multiple speakers by N-best decoding. Fig.4 shows the 3-D trellis space and fig.5 shows the decoding result with harmonic weighting when the speaker moves from 0 to 180 degree while noise is located in 40 degree.

5. MODEL ADAPTATION

Another approach to deal with noise and multiplicative distortion is the model adaptation which adjusts speech model to the observed speech. The good review can be found in [2, 3, 4, 5, 6]. In this section we introduce our research effort applying model adaptation to noise and reverberation in the room.

The observed signal in the real room is represented by,

$$y(t) = S(t) * H(t) + N(t). \tag{1}$$

Here, $H(t)$ is a spectrum of acoustical transfer function depends on locations of sound source and a microphone in the room. $H(t)$ is function of t since speaker may move around. $N(t)$ is noise in the room. In this approach a model for the observed signal is estimated by composition of HMMs of speech, acoustical transfer function and noise. It is easy to estimate the noise model using signals during the noise period. As for the acoustical transfer function, we measured fixed number of representative impulse response in the room and assign their cepstral representations as mean values of ergodic acoustical transfer function HMM. The composed HMMs of clean speech, noise and acoustical transfer function improves the speech recognition performance, although its approximation is very rough[23, 24]. Then another trial is made by representing variation from preceeding frames by variances of acoustical transfer function HMMs[23]. However, the results are not so good as expected. These efforts so far show the model adaptation approach partly improves the degradation by the room reverberation. However, a further sophisticated method should be developed to deal with influences beyond analysis window length fundamentally.

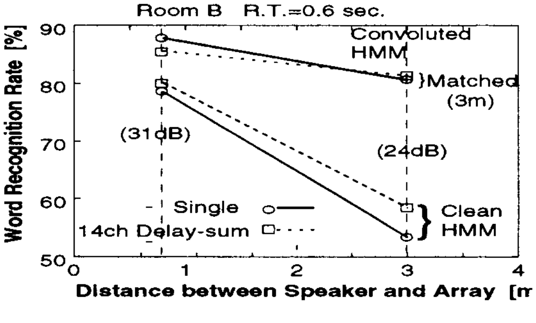

Figure 6. Speaker Dependent 500 Word Recognition

Fig. 6 shows 500 word recognition results in the T_{60}=600msec room. The result by clean speech HMM is seriously degraded according to the distance from the speaker to the microphone. On the other hand the result by HMMs trained by distorted speech in 3m is very good. It is noticed that the recognition result is also improved for the speech uttered in 0.8m. This fact implies the influences by impulse response on speech recognition is not so precise and can be restored by taken into account of a rough estimate of reverberation using several preceding frames. [11] shows a trial to this direction incorporating neural network mapping from reverberant cepstra to clean cepstrum.

6. MULTI-MODAL PROCESSING

Another approach to hands-free speech recognition is multi-modal integration of audio information and visual information such as gesture, face and lip movements. Humans pay attention not only to speaker's speech but also to speaker's mouth in the adverse environments. This suggests a fact that hands-free speech recognition can be improved by incorporating mouth images. Many studies have been presented related to improvements of speech recognition by automatic lip reading [30, 31, 32, 33, 34]. Our experiments through 100 word test show the performance of 85% by lipreading alone[35]. It is also shown that tied-mixture HMMs improve the lip reading accuracy. The speech recognition experiments are also carried out over various SNR integrating audio-visual information. Fig.7 shows the results by early integration based on likelihood using HMMs trained composite vectors of audio and visual vectors, and by late integration based on merging of both results from audio HMMs and visual HMMs. The results show the late integration always realizes better performance than that using either audio or visual information. More sophisticated audio-visual source localization, segmentation of a mouth image and normalization of size, angle and lighting should be needed for real use.

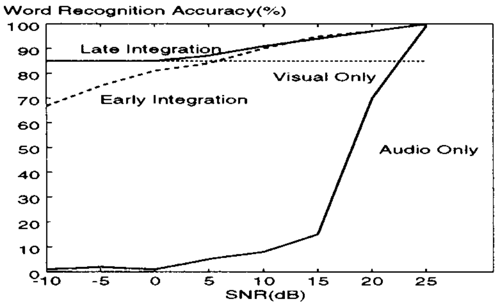

Figure 7. Comparison of Integration Methods

7. CONCLUSION

This paper introduces characteristics of the room acoustical distortion and their influences on speech recognition accuracy. Then the paper also introduces researches to overcome the problems such as acoustical pre-processing, model adaptation and audio-visual integration. To investigate the problems and develop methods, the large database is necessary. We are currently planning to collect impulse responses in various kinds of real rooms and the distorted speech in the room using a microphone array.

REFERENCES

[1] M.Omologo,"On the future trends of hands-free ASR:variabilities in the environmental conditions and in the acoustic transduction", ESCA-NATO Workshop on robust speech recognition for unknown communication channels pp.67-73, 1997,4

[2] S.Furui,"Recent adavances in robust speech recognition", ESCA-NATO Workshop on robust speech recognition for unknown communication channels, pp.11-20, 1997,4

[3] R.M.Stern,et al,"Compensation for environmental degradation in automatic speech recognition", ESCA-NATO Workshop on robust speech recognition for unknown communication channels, pp.33-42, 1997,4

[4] N.Morgan,"Robust feature and model compensation:a few comments", ESCA-NATO Workshopon robust speech recognition for unknown communication channels, pp.43-44, 1997,4

[5] C.H.Lee,"On feature and model compensation approach to robust speech recognition", ESCA-NATO Workshop on robust speech recognition for unknown communication channels, pp.45-54, 1997,4

[6] M.J.F.Gales,"Nice model-based compensation schemes for robust speech recognition", ESCA-NATO Workshop on robust speech recognition for unknown communication channels, pp.55-64, 1997,4

[7] D.Van Compernolle, et al, "Speech recognition in noisy environments with the aid of microphone arrays",Speech Communication, 9(5/6) pp.433-442, 1990

[8] J.L.Flanagan, et al, "Autodirective microphone system for natural communication with speech recognizers", 4th DARPA Workshop, pp.4.8-4.13, 1991.2

[9] H.F.Silverman,et al, "Experimental results for baseline speech recognition performance using input acquired from a linear microphone array", 5th DARPA Workshop, pp.285-290, 1992

[10] H.Silverman,et al,"A digital processing system for source location and sound capture by large microphone arrays", ICASSP97, 1997

[11] Q.Lin, et al, "System of microphone arrays and neural networks for robust speech recognition in multimedia environment", ICSLP94, S22-2, pp. 1247–1250, Sep. 1994.

[12] D.Giuliani,et al, "Talker localization and speech recognition using a microphone array and a crosspowerspectrum phase analysis", Proc. ICSLP94, S22-1, pp. 1243–1246, Sep. 1994.

[13] M.Brandstein,"A framework for speech source localization using sensor arrays", Ph.D thesis, Brown University, 1995.

[14] T.Yamada, et al, "Robust speech recognition with speaker localization by a microphone array",ICSLP96 1996,10

[15] T.Yamada, et al, "Speech recognition of a moving talker based on 3-D viterbi search using a microphone array", IJCAI Workshop of Computational Auditory Analysis, 1997.8.

[16] M.Inoue, et al,"Microphone array design measures for hands-free speech recognition", EUROSPEECH97, 1997.

[17] S.F.Boll,"Suppression of acoustic noise in speech using spectral subtraction", IEEE,ASSP-27,No.2,1979

[18] A.Acero,*Acoustical and environmental robustness in automatic speech recognition.* Ph.D Dissertation, ECE Department, CMU, Sept.1990

[19] A.P.Varga, et al,"Hidden Markov model decomposition of speech and noise", ICASSP90, pp.845-848,1990

[20] M.J.F.Gales,et al,"PMC for speech recognition in additive and convolutional noise",CUED-F-INFENG-TR154, 12, 1993

[21] F.Martin,et al,"Recognition of noisy speech by composition of hidden Markov models",EUROSPEECH93,pp.1031-1034,1993

[22] A.Sankar,et al,"Robust speech recognition based on stochastic matching",ICASSP95,pp.121-124,1995

[23] S.Nakamura,et al, "Noise and room acoustics distorted speech recognition by HMM composition", ICASSP96,1996,6

[24] T.Takiguchi,.et al, "Adaptation of Model Parameters by HMM Decomposition in Noisy Reverberant Environments", ESCA-NATO Workshop in pont-a-mousson, pp.155-158, 1997,4

[25] C.Avendano,et al,"Study on The Dereverberation of Speech based on temporal envelope filtering", ICSLP96, 1996

[26] M.Tohyama,et al,"Source waveform recovery in a reverberant space by cepstrum dereverberation", ICASSP93, 1993

[27] M.Miyoshi,et al,"Inverse filtering of room acoustics", ASSP,36(2):145-152

[28] D.Bees,et al,"Reverberant speech enhancement using cepstral processing", ICASSP91, 1991

[29] Q.G.Liu,et al,"A microphone array processing techniques for speech enhancement in a reverberant space", speech communication, 18:317-334

[30] D.G.Stork, M.E.Hennecke, "Speechreading by Humans and Machines", NATO ASI Series, Springer, 1995

[31] E.Petajan, "Automatic Lipreading to Enhance Speech Recognition", Proc.CVPR'85

[32] C.Bregler, et al, "Improving Connected Letter Recognition by Lipreading", Proc.IEEE ICSLP93

[33] A.Adjoudani,et al, "Audio-Visual Speech Recognition Compared Across Two Architectures", Proc.EUROSPEECH95

[34] M.Alissali, et al, "Asynchronous Integration of Visual Information in an Automatic Speech RecognitionSystem", Proc.ICSLP96

[35] S.Nakamura,et al,"Improved bimodal speech recognition using tied-mixture HMMs and 5000 word audio-visual syncronous database", EUROSPEECH97, 1997