A Study on Speech Recognition using a Microphone Array and HMM composition

Kiyohiro SHIKANO*, Satoshi NAKAMURA*, Shiro ISE*, Ken'ji KURAKATA*, Masahide SUGIYAMA+, Hikaru DATE+,
*Graduate school of Information Science
Nara Institute of Science and Technology
8916-5 Takayama-machi, Ikoma-shi, Nara 630-01
e-mail: shikano@is.aist-nara.ac.jp
+The University of Aizu,
Aizu-Wakamatsu, Fukushima 965-80, JAPAN

1.Introduction

The speech interface is the most natural and friendly interface for human, because the interface is used for man-man communication in common. The speech interface will also play a important role in multi-modal interface when various modalities such as a keyboard, mouse and image are integrated.

In spite of recent advances, the speech recognition technology did not reach to the practical use in the real world. The reason is that the advances are almost achieved in the error reduction of clean speech where the functions above had not been taken into consideration.

A key issue way to the practical use is a development of a recognition technology of distant talking speech in the noisy and reflective room. This realizes a hands free and eyes free man-machine interface. This paper presents two novel approaches to realize hand free speech recognition, which are a multi-channel signal retrieval by a microphone array and HMM model composition of an additive noise and convolutional distortion.

2. Speaker Localization by a Microphone Array

The most serious problem of previous approaches is to use a monoral speech signal only. According to the fact that a human listeners process speech signals in stereo, the multi-channel signal retrieval would bring more significant information for speech recognition.

This paper presents a robust speech recognition with speaker localization by using arrayed microphones, SLAM. The array signal processing enables high SNR signal retrieval utilizing an information of the difference of the speech and noise signal directions. The several researches have been conducted[Qlin94,Omologo95,Nakamura95]. However previous works didn't take into account of a speaker tracking module which extracts the specific speaker's pitch harmonics.

3 System Architecture

We applied the proposed SLAM method to speech recognition system as described in Fig.1.

The system hypothesizes a sound source direction and retrieves speech signal by beamforming with arrayed microphones. Then, the cepstral analysis are conducted to extract a speaker's pitch harmonics. The system verifies the direction whether the speech signal comes from or not. Finally the speech recognition is conducted by using the extracted speech signal from the estimated speech source direction.

3.1 Delay-and-Sum Beam-former

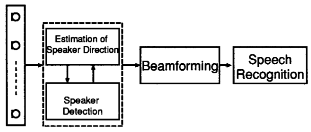

Fig.1: Block Diagram of SLAM Speech Recognizer

We used the delay-and-sum beam-former in order to build a baseline system. Although the subtraction type beam-former which makes a dead angle such as AMNOR[Kaneda92] had also been proposed, it is supposed to be impractical since it is impossible to identify a number of noise sources and their directions. On the contrary a speech source is relatively stable and easy to estimate its number and direction.

We used a linear array for SLAM whose microphone elements are arranged in equal distance.

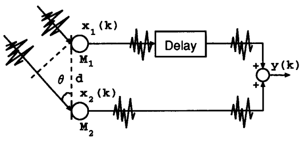

Fig.2: Delay-and-Sum Beam-former

Fig.2 indicates a process of the delay-and-sum beam-former with two microphone elements in distance. Therefore the signals from the target can be in the same phase by the delay, while the signals from the other directions will not be survive. The distance of microphone elements are designed by the frequency band of the target signal.

3.2 Sound Source Localization

For the speech recognition the robust speech sound source localization is required. Almost all methods had been proposed are based on extraction of a direction with a maximum power. However, those methods will be in trouble when SNR becomes worse.

This paper describes an algorithm which makes maximum use of a prior knowledge of the speech. As the prior knowledge, pitch harmonics during voiced speech are utilized. Two problems arise. The first problem is the accurate and stable extraction of pitch harmonics. The second problem is the localization during an unvoiced speech period. For the first problem we adopted the cepstrum analysis. The cepstal analysis captures pitch harmonics stably. For the second problem we assume the temporal continuity of a speech source direction. In our algorithm the speech source direction during an unvoiced region will be set to the direction of the least recent extracted direction.

Fig.3 is an overview of our algorithm. In the algorithm candidates of speech source direction are extracted when the pitch harmonics is lager than preset threshold. Then a direction whose power is maximum is chosen as the speech source direction. If there is no candidate, the algorithm will chose the least recent extracted direction.

3.3 Experiments

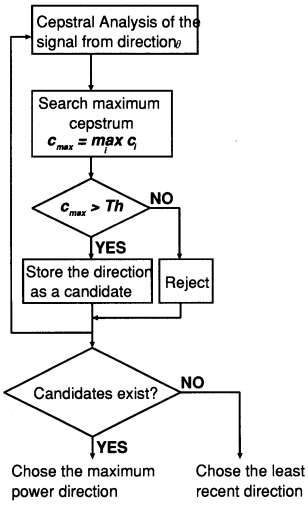

Fig.3: Speech sound localization algorithm

The arrayed microphones used in this paper is 14-element linear array. The distance of the elements is set to 2.83cm for 6kHz Nyquist frequency. The sampling rate is 12 kHz. A/D converter synchronously operates for 14 channels.

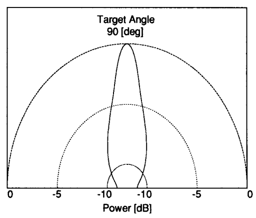

Fig.4: Directional Pattern

Fig.4 shows directional pattern of the arrayed microphones for 6kHz band-limited Gaussian white noise. The arrayed microphone improves SNR by +10dB in 40 degree.

The evaluation are conducted through 500 word recognition experiments. The recognition accuracy, RA, and the accuracy of speaker localization, ASL are used as an evaluation measure. Where the accuracy of speaker localization is defined as,

ASL = (number of frames extracted correctly)/(number of total frames)

Speech corpus for evaluation is ATR Japanese speech database Set-A. This database contains word and sentence utterances by announcers. The recognition algorithm is based on 256 Tied Mixture HMM. Context independent 54 phone models are trained by 2620 words. The other 500 words are used for testing. Speech signal from the speech source direction is windowed by 32msec Hamming window every 8msec. Then FFT is used to calculate 16-order MFCC, 16-order Delta-MFCC and Delta-power.

In this paper we evaluate the system by speaker dependent word recognition experiments, where the recognition rate for clean speech is 96.4%. The recognition rates for noisy speech by clean HMMs are 0.2, 0.8, 23.4 for SNR -10, 0, +10dB, respectively. In the environment the speech source and the Gaussian white noise source are positioned at 90th degree and 40th degree, respectively.

Three methods are compared,

A Speech source direction known.
B Speech source direction not known

B-1 Localization by maximum power
B-2 Localization by maximum amplitude of pitch harmonics and total power

Table.1 shows the results. A is an upper bound of the system. The SLAM can improve the recognition accuracy by 66case that the speaker localization algorithm works perfectly. There remains 7speech. This is caused by the acoustic channel distortion. The degradation will be reduced by improving the speech source localization algorithm. These results clarified the proposed SLAM system improves the recognition accuracy in noisy environments.

4. Speech Recognition of Noisy and Room Acoustics Distorted Speech

Another approach to recognize distant-talking speech is a model-based approach. This section describes that based on HMM composition.

As model-based approaches, the conventional multi-template approach, the HMM model (de-)composition approach[Varga-Moore,Gales92,Martin93] and the HMM model adaptation approach had been proposed. Among these approaches the HMM composition approach is the most promising, because the HMM for noisy speech can be easily generated by composing the speech HMMs and the noise HMM which trained during noise portion. The paper[Gales92,Martin93] showed the composed noisy HMM outperforms very high accuracy.

In this paper, we apply the HMM composition to the recognition of the speech which is contaminated by not only an additive noise but also room acoustics distortion. If the components are independent each other and additive, HMM composition can be adopted. The noise and speech are independent and additive in linear spectral domain. While the transfer function and speech are convolutional in linear spectral domain, they are independent and additive in cepstral domain. Thus the HMM composition is applicable for noise and room acoustics distorted speech.

4.1 HMM Composition

On the assumption that speech and noise signal are independent, the observed signal is represented by

$O(t) = S(t) + N(t)$

The conventional approach estimates noise statistics during noise portion and recognizes an input noisy speech by using the noise added reference patterns. The HMM composition executes an addition in HMM parameter domain instead of the addition in signal domain. Since the signal level is generally different between training and testing, an adjustment factor k is introduced. Thus the observed signal is represented by

$O(t) = S(t) + k*N(t)$

where O(t),S(t) and N(t) are the observed noisy signal, speech signal and noise signal, respectively. Since this relation is preserved in linear spectral domain, we regard O(t),S(t),N(t) as short time linear spectra whose analysis window starts at time t from now on.

Generally, parameters for speech recognition are represented by cepstrum. The parameters have to be transformed to linear domain in order to hold the additive property of speech and noise[Gales92,Martin93].

As for a convolutional distortion, the observed distorted spectrum is represented by

O(t) = H(t) * S(t)

where H(t) is a transfer function from the sound source to the microphone. H(t) is a function of time t since the sound source may move. The multiplication can be converted to sum in cepstral domain as,

Ocep(t) = Hcep(t) + Scep(t)

where, Ocep(t), Hcep(t) and Scep(t) are cepstra for the observed signal, acoustic transfer function and speech signal, respectively. Therefore the observed signal is represented by

O(t) = exp(IFT(Scep(t) + Hcep(t))) + k*N(t)

This procedure is summarized in Fig.5. The cosine transform, inverse cosine transform, exponential transform and log transform are conducted on HMM parameters.

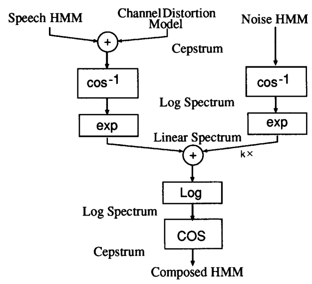

Fig.5: Block Diagram of HMM Composition

The HMM recognizer decodes the observed signal on a trellis diagram according to maximize the log likelihood. Decoded path will bring a optimal combination of a speech, transfer function and noise.

4.2 Experiments

We conducted recognition experiments of noisy speech uttered in nosy room. Fig.6 shows the room used in the experiment. We measured 11 transfer functions from 11 positions to the microphone. The former five positions, h1,...,,h5 are used for the model composition and the latter six positions, p1,...,p6 are used for the recognition tests. The transfer functions are measured by sweep method. The test data are simulated by convolution of speech corpus and measured transfer function. Here we only use 32msec of transfer function adjusting to speech analysis window.

The speaker dependent experiment are carried out. The noise data is collected in computer room and added to the clean speech data or acoustics distorted data as the SNR is 15dB.

We assigned one state for noise HMM and 5 states for acoustics distortion HMM.

For the experiment four kinds of HMMs are prepared listed in Table.2. This table also summarizes the experiment results of 500 word speech recognition. The results show that the proposed HMM composition improves the recognition accuracy from 20.6noise and acoustics distorted speech.

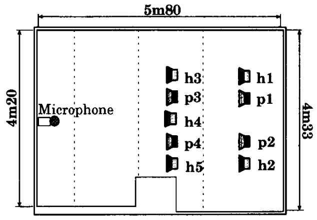

Fig.6: Room Environments

5. Conclusion

A key issue way to the practical use is a development of a recognition technology of distant talking speech in the noisy and reflective room. The SNR will be drastically degraded for the distant talking speech and the speech will be distorted by room acoustics transfer function. This paper proposes two novel approaches and confirmed their effectiveness. Furthermore complementary combination of these approaches will bring more effective hands free speech recognition system.

References

Qlin94: Q.Lin, E-E.Jan, J.Flanagan, "Microphone Arrays and Speaker Identification", IEEE Trans.on Speech and Audio, Vol.2, No.4, pp.622-629, 1994

Omologo95: D.Giuliani, M.Matassoni, M.Omologo, P.Svaizer, "Hands Free Continuous Speech Recognition in Noisy Environment Using a Four Microphone Array", ICASSP95, pp.860-863, 1995

Nakamura95: S.Nakamura, T.Yamada, K.Shikano,"Speech Recognition with Sound Source Localization by Microphone Array", Proc.ASJ meeting, 1-5-8, 1995(in Japanese).

Kaneda92: Y.Kaneda,"Adaptive Microphone Array", IEICE, J75-BII, pp.742-748, 1992(in Japanese)

boll79: S.F.Boll,"Suppression of Acoustic Noise in Speech Using Spectral Subtraction", IEEE, ASSP-27, No.2, 1979

Varga-Moore: A.P.Varga, R.K.Moore, "Hidden Markov Model Decomposition of Speech and Noise", ICASSP90, pp.845-848, 1990

Gales92: M.J.F.Gales, S.J.Young, "An improved Approach to the Hidden Markov Model Decomposition of Speech and Noise",ICASSP92, pp.233-236, 1992

Martin93: F.Martin, K.Shikano, Y.Minami, "Recognition of Noisy Speech by Composition of Hidden Markov Models", EUROSPEECH93, pp.1031-1034, 1993

Keywords: robust speech recognition, hands-free, microphone array, delay-and-sum beamformer, source localization, HMM composition