MICROPHONE ARRAY DESIGN MEASURES FOR HANDS-FREE SPEECH RECOGNITION

Masaaki INOUE, Satoshi NAKAMURA, Takeshi YAMADA, 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

One of the key technologies for natural man-machine interface is hands-free speech recognition. The performance of hands-free distant-talking speech recognition will be seriously degraded by noise and reverberation in real environments. A microphone array is applied to solve the problem. When applying a microphone array to speech recognition, parameters such as number of microphone elements and their spacing interval affect the performance. In order to optimize these parameters, a measure which reflects recognition performance is needed. In this paper, we investigate a measure of a microphone array design for speech recognition through experiments using various kinds of a microphone array design.

1. INTRODUCTION

One of the key technologies for natural man-machine interface is hands-free speech recognition. This realizes so natural and friendly man-machine interface that users are not encumbered by microphone equipments. The accuracy of speaker independent speech recognition is 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. If the speaker inputs his speech from distance or through a telephone channel, the accuracy will be seriously degraded by the influences of the room acoustic distortion or telephone channel distortion and environment noises. Therefore we still have problems of interferences by noise and reverberation.

Many works are presented to solve these problems from viewpoints of speech enhancement and model modification. These approaches, however, only use a monoral speech signal. 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.

In this paper, we tried to solve the problem using the multi-channel signal retrieval by a microphone array. The array signal processing enables high SNR signal retrieval utilizing an information of the speech and noise signal directions. The many researches have been studied [1, 2, 3, 4, 5, 6, 7]. However, when applying a microphone array to speech recognition, parameters such as number of microphone elements and their spacing interval affect the performance. In order to optimize these parameters, a measure which reflects recognition performance is needed. In this paper, we investigate measures of a microphone array design for speech recognition through experiments using various kinds of a microphone array design.

2. DELAY-AND-SUM BEAM-FORMER

The delay-and-sum beam-former is used in order to build a baseline system. The delay-and-sum beam-former realizes super directivity for the sound source direction. Although the subtraction type beam-former which makes a dead angle had also been proposed, the subtraction type beam-former has a problem that it isn't always possible to identify a number of noise sources and their directions. The reverberation and nondirectional ambient noise are the typical situation which the subtraction type beam-former is not able to deal with. On the contrary a speech source is relatively stable and easy to estimate its number and direction. We used a linear array whose microphone elements are arranged in equal distance. Fig.1 indicates a process of the delay-and-sum beam-former with M microphone elements in a equal distance. Where d is a distance of microphones and $x_1(k), \cdots, x_M(k)$ are signals detected by each microphones. The output signal from a direction of θ is $y(k)$. Since the delay-and-sum beam-former assumes a speech as plane wave, the signal $x_{i-1}(k)$ is delayed by $\tau = \frac{d \sin\theta}{c}$ (c = speed of sound) from $x_i(k)$. Then $x_i(k)$ is represented as,

$$\begin{aligned} x_i(t) &= x_{i-1}(t+\tau), \\ &= x_1(t+(i-1)\tau), \; i = 1, 2, \cdots, M \quad (1) \end{aligned}$$

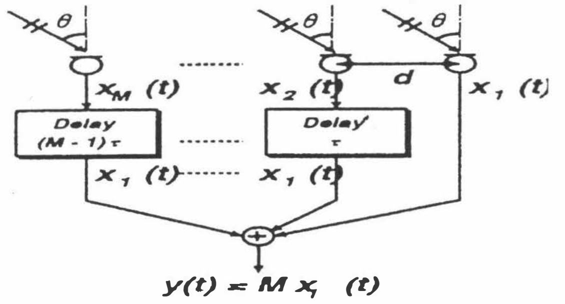

Figure 1. Delay-and-Sum Beam-former

Therefore the signals from the direction θ can be in the same phase by the delay τ as $x_1(k)$. On the other hand the signals from the other directions will not be survived. In general the delay-and-sum beam-former with M-microphone elements is represented as,

$$\begin{aligned} y(t) &= \sum_{i=1}^{M} x_i(t-((i-1)\tau) \\ &= \sum_{i=1}^{M} x_i(t)\exp\left\{-j2\pi f(i-1)\frac{d\sin\theta}{c}\right\} \quad (2) \end{aligned}$$

The received signals $x_1(t)$, $x_2(t)$, $\cdots$, $x_M(t)$ can be regarded as a sampled sequence of complex sin signal by period d. The frequency of complex sin signal is $f_{sp} = -\dfrac{f\sin\theta}{c}$. Then,

$$y(t) = \sum_{i=1}^{M} x_i(t)\exp\{j2\pi f_{sp}(i-1)d\} \quad (3)$$

This equation is formally the same as Fourier transform. For wide-band signal like speech, the derivation above can be easily extended,

$$Y(t,f) = \sum_{i=1}^{M} X_i(t,f)\exp\left\{-j2\pi f(i-1)\frac{d\sin\theta}{c}\right\} \quad (4)$$

$$0 \le f \le f_{\max}$$

Here, $X_1(t,f), X_2(t,f), \cdots, X_M(t,f)$ are Fourier transforms of each microphone signals, and $f_{\max}$ is Nyquist frequency.

Characteristics of delay-and-sum beam-former are summarized as follows;

- According to a number of microphone element and their spacing, beam width of the microphone array becomes sharper.
- The distance of a microphone element, d, must be spaced so that d satisfies $d < \dfrac{c}{2f}$, otherwise spacial aliasing will be occurred. (Spatial Sampling Theorem)
- The main beam width becomes shaper according to frequency.

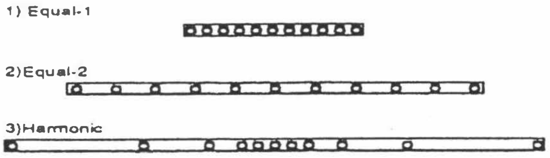

Figure 2. Microphone arrays

Table 1. Specification of Microphone Arrays (Frequency Bands : [kHz])

Array	freq bands	mic spacing	#mic elements	
Equal-1	0~6	2.83cm	11	
Equal-2	0~6	6.00cm	11	
Harmonic	0~0.75	22.67cm	5	11
	0.75~1.5	11.33cm	5	
	1.5~3	5.67cm	5	
	3~6	2.83cm	5	

Table 2. Characteristics of the Arrays

Frequency band	Low	Middle	High
Equal-1	×	○	◎
Equal-2	○	◎	×
Harmonic	◎	○	○

3. COMPARATIVE EXPERIMENTS OF VARIOUS MICROPHONE ARRAYS

3.1. Microphone Arrays

Table 1 shows microphone arrays compared in this paper. These are two equally spaced linear microphone arrays (Equal-1,2) and one harmonic spaced microphone array (Harmonic). A microphone array beam-form is frequency dependent and determined by its spacing interval of microphone elements. The first equally spaced linear microphone array, Equal-1 is designed its spacing interval to Nyquist frequency. Then Equal-1 has a sharp beam-form at Nyquist frequency. The second equally spaced linear microphone array, Equal-2 allows spacial aliasing by spacing beyond 1/2 wavelength. Therefore Equal-2 realizes sharp beam-form in middle frequency while spacial aliasing in high frequency band. The last harmonic spaced microphone array, Harmonic, divides target frequency band logarithmically and assigns the same number of microphone elements. Harmonic realizes the same beam-form for each band. Table 2 summarizes their characteristics and fig.3 shows their directivity pattern to band limited white noise.

Table 3. Speaker Dependent 500word Recognition Accuracy

	Accuracy(%)		Gain(dB)		SDR(dB)		NDSR(dB)	
White Noise(SNR=10dB)	45°	80°	45°	80°	45°	80°	45°	80°
Equal-1	86.8	89.4	−8.90	−10.3	19.9	21.5	−2.127	−1.878
Equal-2	91.8	87.8	−11.9	−9.18	23.1	19.8	−1.774	−2.078
Harmonic	81.8	84.6	−11.8	−12.7	22.3	23.3	−4.052	−2.520
Computer Noise(SNR=10dB)	45°	80°	45°	80°	45°	80°	45°	80°
Equal-1	69.0	70.2	−8.90	−10.3	11.3	11.9	−2.266	−2.159
Equal-2	74.4	72.6	−11.9	−9.18	13.0	14.8	−2.077	−2.076
Harmonic	70.2	89.4	−11.8	−12.7	16.0	19.2	−2.195	−1.829
Reverberation	0.4sec		0.4sec		0.4sec		0.4sec	
Equal-1	66.6		-		−8.55		−2.842	
Equal-2	65.2		-		−8.45		−2.755	
Harmonic	68.2		-		−8.28		−2.745	

3.2. Evaluation Measures

Two kinds of evaluation measure are used as baseline measure such as,

- array gain for noise direction (Gain)
- signal to deviation ratio (SDR).

Gain is a measure commonly used to specify acoustical characteristics of a microphone. SDR is a measure used to evaluate signal quality in signal separation or speech enhancement. The following equation defines SDR.

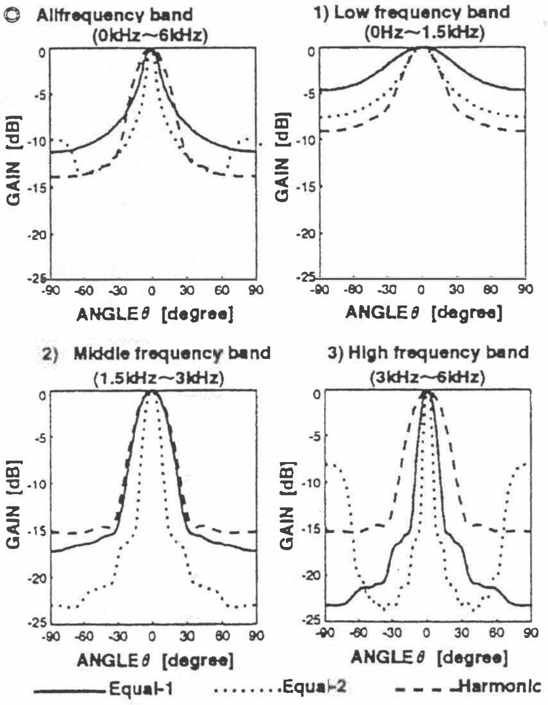

Figure 3. Directivity Patterns

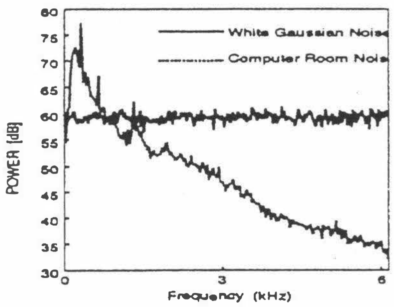

Figure 4. Noise Characteristics

$$SDR = 10\log\frac{\sum_{n=0}^{L-1} s^2(n)}{\sum_{n=0}^{L-1}\{s(n)-\beta\widehat{s}(n)\}^2}[dB]$$

Here, $s(n)$ and $\widehat{s}(n)$ are an original and an enhanced signal, respectively.

3.3. Recognition Experiments

The speaker dependent isolated word recognition experiments are conducted using Japanese 500 words. The experiments aim to clarify relationships between speech recognition accuracy and conventional measures. The speech recognition is carried out using context independent 54 phoneme HMMs with 256 tied-mixture distributions. HMMs are trained by 2620 words out of 5240 words of ATR Set-A MHT database. Other 500 words are used for testing. Parameters for recognition are 16MFCC, 16ΔMFCC and Δenergy. The evaluation is conducted under the following two conditions.

Noisy Environment Speech source is placed in front of microphone array, 90 degree. Noise sources are place in 45 degree or 80 degree. White Gaussian noise and computer room noise in fig.4 are used as noise source. The delays of wave arrival are calculated by computer simulation on the assumption that the wavefront is a plane wave.

Reverberant Environment Reverberant signals are simulated by convolution of an original signals and impulse responses measured in a real 5.83 x 4.33m room using a microphone array. The reverberation time of the room used is about 0.4 sec. The distance between the microphone array and the speaker is 3m.

Table 3 shows the results. Although Harmonic achieves the maximum array gain of -12.7dB for white noise from 80 degree, this microphone array gives the worst performance of 84.6% among evaluated microphone arrays. As far as concerned with noise characteristics, speech recognition accuracy can be varied in spite that the microphone characteristics are the same. It is also observed SDR is insufficient to predict recognition accuracy in the Table. These results suggest that the measure must take into account of following properties,

- characteristics of speech signals
- characteristics of noise signals
- SNR for each frequency band

4. NDSR

In order to provide a good microphone array design measure to predict recognition performance, this paper proposes a new design measure, NDSR (Normalized Distortion to Signal Ratio; NDSR) of a microphone array for speech recognition. This measure takes into account of three properties described in the previous section. NDSR is defined as,

$$NDSR = -10\log\left(\sum_{\omega}\frac{\left|\widehat{S}(\omega) - S(\omega)\right|}{S(\omega)} + 1\right)\ [dB]$$

Here, $S(\omega)$, $\widehat{S}(\omega)$ are power spectra of the target speech and the noise signal, respectively.

Since NDSR is an average SNR for each frequency band, NDSR becomes larger if distortion is small. "NDSR = 0dB" means there is no distortion. While the target speech signals are necessary to calculate NDSR, the target speech signals are not always available. In this paper, a phoneme balanced 216 Japanese words are used in stead of the target speech signals.

Table 3 also shows the results by NDSR. It is confirmed that the proposed measure NDSR provides good prediction of speech recognition accuracy for almost every environments. Only the case NDSR doesn't follow the recognition rate is a reverberation environment using Equal-2. This result suggests that there is a remaining problem for NDSR caused by spatial aliasing.

5. SUMMARY

This paper shows the conventional measures for microphone array are insufficient to predict speech recognition accuracy and to design a good microphone array for speech recognition. To provide the sufficient measure, a new measure, NDSR is proposed. The measure is evaluated through speech recognition experiments. The experiments clarified the effectiveness of the proposed measure. However, we still have a problem that the proposed measure doesn't always predict speech recognition performance in reverberant environments using a microphone array which allows spacial aliasing.

REFERENCES

[1] D.Van Compernolle, W.Ma, F.Xie, M.Van Diest, "Speech recognition in noisy environments with the aid of microphone arrays",Speech Communication, 9(5/6) pp.433-442, 1990

[2] J.L.Flanagan, R.Mammone, G.W.Elko, "Autodirective microphone system for natural communication with speech recognizers", 4th DARPA Workshop, pp.4.8-4.13, 1991.2

[3] H.F.Silverman, S.E.Kirtman, J.E.Adcock, P.C.Meuse, "Experimental results for baseline speech recognition performance using input acquired from a linear microphone array", 5th DARPA Workshop, pp.285-290, 1992

[4] Q.Lin, E.E.Jan, C.Che, B.de Vries, "System of microphone arrays and neural networks for robust speech recognition in multimedia environment", ICSLP94, S22-2, pp. 1247–1250, Sep. 1994.

[5] D.Giuliani, M.Omologo, P.Svaizer, "Talker Localization and Speech Recognition Using a Microphone Array and a Cross-Powerspectrum Phase Analysis", Proc. ICSLP94, S22-1, pp. 1243–1246, Sep. 1994.

[6] S.Nakamura, T.Yamada, T.Takiguchi, K.Shikano, "Hands Free Speech Recognition by a Microphone Array and HMM Composition", Proc.IWHIT95 Aizu,pp.33-38,1995,10

[7] T.Yamada, S.Nakamura, K.Shikano, "Robust speech recognition with speaker localization by a microphone array",ICSLP96 1996,10