# Space discriminative function for microphone array robust speech recognition[①]

Zhao Xianyu (赵贤宇)[②], Ou Zhijian, Wang Zuoying
(Department of Electronic Engineering, Tsinghua University, Beijing 100084, P.R.China)

**Abstract**

Based on W-disjoint orthogonality of speech mixtures, a space discriminative function was proposed to enumerate and localize competing speakers in the surrounding environments. Then, a Wiener-like post-filterer was developed to adaptively suppress interferences. Experimental results with a hands-free speech recognizer under various SNR and competing speakers settings show that nearly 69% error reduction can be obtained with a two-channel small aperture microphone array against the conventional single microphone baseline system. Comparisons were made against traditional delay-and-sum and Griffiths-Jim adaptive beamforming techniques to further assess the effectiveness of this method.

**Key words**: speech recognition, array signal processing, microphone array, source localization, adaptive filtering

## 0 Introduction

Under the condition of existing competing speakers, the performance of a speech recognition system degrades seriously. With its capability to provide hands-free acquisition of speech and directional discrimination, microphone array has become widely used in many robust ASR front-end[1-3].

Adaptive beamforming realizes notches in the directions of interferences in current working environment by adapting its weights according to some optimum criterion[4]. Adaptive microphone array can realize good interference suppression with a comparable small number of microphone cells when compared with array with fixed beam pattern[5,6].

However, some problems still exist with the conventional adaptive beamforming, such as the negative impact of room reverberation, the potential cancellation and distortion of target signal, etc.[7]. In Ref. [7], a pseudo adaptive microphone array processing strategy is proposed and SNR improvements are obtained in the experiments. In this approach, instead of using some criterion such as MMSE to adapt the response of microphone array directly, array processing is divided into two steps. First, an explicit estimate of current working environment is made. CPSP is used to estimate the directions of multiple sources. Then, a spatial filter, constructed in the frequency domain based on the result of the first step, is used to suppress existing interferences. However CPSP is not suitable for the near-field case and usually demands a large number of microphones and large aperture (8 microphones are used in Ref. [7], with an array aperture of 0.8m). In this paper, a new space discriminative function is proposed to enumerate and localize multiple near-field sources using a small aperture microphone array with only two microphone cells. Based on the location and spectral information of the interferences provided by the discriminative function, a straightforward realization of frequency domain post-filter for interference suppression is implemented. And it is shown that this post-filter is an approximation to Wiener filter for MMSE filtering.

## 1 Near-field propagation model and space discriminative function

The microphone array near-field propagation model used in this paper is shown in Fig.1, where $M_1$ and $M_2$ are two omni-directional microphones, S is the speaker.

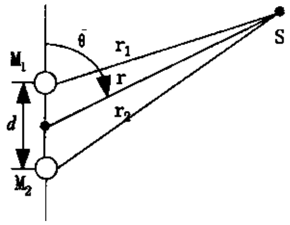

**Fig.1** Microphone array near-field propagation model

Under anechoic environment, the relation between the outputs $\{x_k(t);k=1,2\}$ and source signal $s(t)$ is

$$x_k(t) = s(t - r_k/c)/r_k \qquad (1)$$

[①] Supported by the High Technology Research and Development Programme of China (No. 2001AA114071).

[②] To whom correspondence should be addressed. E-mail: xianyu.zhao@francetelecom.com

Received on Apr. 14, 2004

In frequency domain, the relationship can be written as

$$X_k(\omega) = S(\omega) \cdot \exp(-j\omega r_k / c) / r_k \tag{2}$$

From these, define $a$ and $\tau$ as follows,

$$\begin{cases} a = | X_2(\omega) | / | X_1(\omega) | \\ \tau = [\measuredangle X_2(\omega) - \measuredangle X_1(\omega)] / (-\omega) \end{cases} \tag{3}$$

where $|\cdot|$ and $\measuredangle(\cdot)$ stands for the amplitude and phase angle of a complex variable. It can be easily seen that the relationship between (a,$\tau$) and source location ($\mathrm{r}_1$,$\mathrm{r}_2$) is as follows

$$\begin{cases} a = r_1 / r_2 \\ \tau = (r_2 - r_1) / c \end{cases} \tag{4}$$

Due to the short time stationary property of speech signal, short time Fourier spectrum transform is usually used to calculate $X^n = \{X_k^n(\omega); k = 1,2\}$ for each frame of microphone outputs. In real applications, it is not accurate and robust enough to use only one frame of signals to determine the location of source through Eq. (4) for the reason of windowing, noise and the existence of multiple active sources. In Ref. [8], an off-line method is proposed to determine the locations of multiple sources based on a weighted histogram on the ($a$, $\tau$) plane. In this paper, a new kind of space discriminative function, $f_D(a, \tau | X^n, \cdots, X^0 |)$ based on current and past signal frames is proposed to provide real-time estimation of the number and locations of existing sources. Through updating this discriminative function frame by frame, the movement of various sources can be effectively tracked. We describe the definition and calculation of this function in the following.

First, the prior information about the locations of the potential sources can be incorporated into the space discriminative function $f_D(a, \tau | X_0)$. In the ignorant case, we can set $f_D(a, \tau | X_D)$ to be

$$f(a, \tau | X_0) = 1/M \tag{5}$$

where $M$ is the number of distinctive space cells in the observed area. While in some more informative cases, for example in the car environment, it can be set as follows:

$$f_D(a, \tau | X_0) = \begin{cases} 1/K, & \mathit{if}\ (a, \tau) \in S \\ 0, & \mathit{otherwise} \end{cases} \tag{6}$$

where $S = \{(a_1, \tau_1), \cdots, (a_K, \tau_K)\}$ correspond to the set of locations of various potential sound sources in the car, such as radio, motor, air conditioner, etc.

When acquiring the $n$-th frame of the microphone array outputs, the frequency-dependent instant space discriminative function $f_{D,\omega}(a, \tau | X^n)$ at every frequency bin is calculated as

$$f_{D,\omega}(a, \tau | X^n) = C_\omega^n \cdot [1 / | a \cdot e^{-j\omega\tau} X_1^n(\omega) - X_2^n(\omega) |^2] \tag{7}$$

where $C_\omega^n$ is a normalization coefficient chosen to be

$$C_\omega^n = \sum_a \sum_\tau [1 / | a \cdot e^{-jw\tau} X_1^n(\omega) - X_2^n(\omega) |^2] \tag{8}$$

According to the W-disjoint orthogonality of speech mixtures[8], $f_{D,\omega}(a, \tau | X^n)$ would be significantly large when the source at ($a$, $\tau$) is dominant among others at the frequency $\omega$. So the spike of $f_{D,\omega}(a, \tau | X^n)$ could indicate that there exists a potential source at ($a$, $\tau$). Compared with the 0-1 indicator function in Ref. [8], $f_{D,\omega}(a, \tau | X^n)$ is smoother.

The full-band instant space discriminative function $f_D(a, \tau | X^n)$ that combines different frequency bins is calculated to be

$$f_D(a, \tau | X^n) = \sum_\omega W_\omega^n f_{D,\omega}(a, \tau | X^n) \tag{9}$$

where $W_\omega^n$ is the frequency-dependent weight coefficient chosen according to the energy distribution in the frequency domain, i.e.,

$$W_\omega^n = [| X_1^n(\omega) \cdot X_2^n(\omega) |] / \sum_\omega | X_1^n(\omega) \cdot X_2^n(\omega) | \tag{10}$$

Finally, space discriminative information provided by past frames and the current frame are combined through

$$f_D(a, \tau | X^n, \cdots, X^0) = \beta \cdot f_D(a, \tau | X^n) + (1 - \beta) \cdot f_D(a, \tau | X^{n-1}, \cdots, X^0) \tag{11}$$

where $\beta$ is a forgetting factor.

For two sources located at $r_1 = 0.5\mathrm{m}$, $\theta_1 = 90°$ and $r_2 = 0.5\mathrm{m}$, $\theta_2 = 120°$ in a 3m × 4m × 5m room, the space discriminative function calculated after observing 20 frames is depicted in Fig.2. According to its peaks, two sources are identified to be located at $\hat{r}_1 = 0.5\mathrm{m}$, $\hat{\theta}_1 = 88$ and $\hat{r}_2 = 0.6\mathrm{m}$, $\hat{\theta}_2 = 117$.

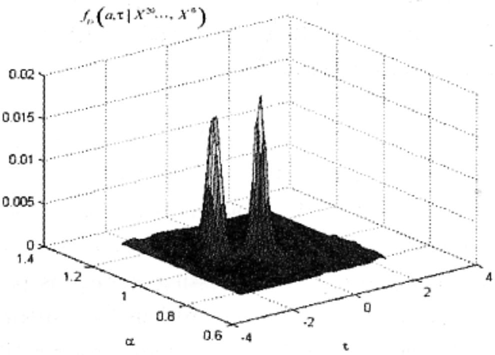

**Fig.2** Example Space discriminative function

## 2 Frequency domain post-filtering for interference suppression

According to the spikes of space discriminative function, several prominent sources are identified at locations corresponding to ($\alpha_1$, $\tau_2$)...($\alpha_N$, $\tau_N$). Based on the prior knowledge, one of these sources can be identified as the target signal, while others are treated as interferences. Without loss of generality, the source located at $(\alpha_1, \tau_1)$ is denoted as the target. Then a frequency domain post-filter for the delay-and-sum beamformer is constructed as

$$H(\omega) = f_{D,\omega}(\alpha_1, \tau_1 \mid X^n) / \sum_{i=1}^{N} f_{D,\omega}(\alpha_j, \tau_j \mid X^n) \tag{12}$$

and the output of the microphone array interference suppression system, $Y(\omega)$, becomes

$$Y(\omega) = H(\omega) \cdot (X_1(\omega) + X_2(\omega) \cdot \exp(j\omega\tau_1))/2 \tag{13}$$

For the case of two sources $S_1$ and $S_2$, the corresponding space discriminative functions are shown to be

$$\begin{aligned} f_{D,\omega}(a_1, \tau_1 | X^n) &= C_\omega^n \cdot [1/| a_1 \cdot e^{-j\omega\tau_1} \cdot X_1^n(\omega) - \\ &\quad X_2^n(\omega) |^2] \\ &= C_\omega^n \cdot [1/[| S_2^n(\omega) |^2 | a_1 \cdot e^{-j\omega\tau_1} - \\ &\quad a_2 \cdot e^{-j\omega\tau_2} |^2]]^{\text{and}} \end{aligned}$$

$$\begin{aligned} f_{D,\omega}(a_2, \tau_2 | X^n) &= C_\omega^n \cdot [1/| a_2 \cdot e^{-j\omega\tau_2} \cdot X_1^n(\omega) - \\ &\quad X_2^k(\omega) |^2] \\ &= C_\omega^n \cdot [1/[| S_1^n(\omega) |^2 | a_2 \cdot e^{-j\omega\tau_2} - \\ &\quad a_1 \cdot e^{-j\omega\tau_1} |^2]] \end{aligned}$$

In this case, $H(e^{j\omega})$ is calculated as

$$\begin{aligned} H(e^{j\omega}) &= \frac{f_{D,\omega}(\alpha_1, \tau_1 \mid X^n)}{f_{D,\omega}(\alpha_1, \tau_1 \mid X^n) + f_{D,\omega}(\alpha_2, \tau_2 \mid X^n)} \\ &= \frac{| S_1^n(\omega) |^2}{| S_1^n(\omega) |^2 + | S_2^n(\omega) |^2} \end{aligned} \tag{14}$$

From this, it can be seen that this post-filter becomes a Wiener filter, and the power spectrum densities used in the Wiener filter are estimated through space discriminative function implicitly. In the cases of more sources and reverberant environment, $H(e^{j\omega})$ is an approximation to the optimum Wiener filter.

## 3 Experiments and results

A baseline hands-free speech recognition system is constructed to recognize continuous digits. This system consists of 10 HMM's to model Chinese digits from zero to nine, each HMM has six emitting states and 14-mixture GMM is used for each state. An additional state is used to model the silence. The feature vector is formed by 14 MFCC's, energy plus their first and second order differentials. The speech frame length is 20 ms and there exists 10ms overlap between adjacent frames. The microphone array front end uses two omni-directional Panasonic WM54B microphones. The space interval between them is 0.05m. The space location settings for the target speaker, interferences and microphone array are depicted in Fig.3.

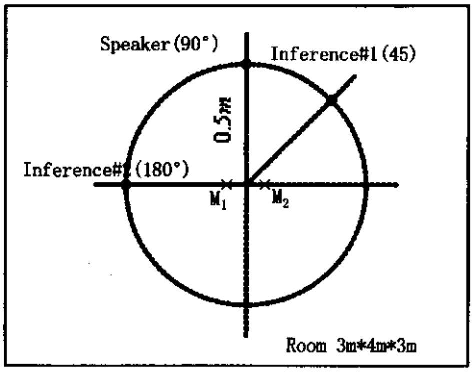

**Fig.3** Experimental location settings for microphone array cells, target speaker and interference speakers

Experiments are carried out in a 3m × 4m × 3m room with reverberation time to be about 0.3s which is typical for the office environment. The results are presented as word error rates (WER), which is given by

$$WER = \frac{D + I + S}{N} \times 100\% \tag{15}$$

where $D$ is the number of deletions, $I$ the number of insertions, $S$ the number of substitutions and $N$ the total number of words. A total of 220 utterances of Chinese digit sequences of different lengths are used to assess the ASR performance. The interference or disturbance signals are Chinese sentences uttered by different speakers with randomly chosen contents.

In the following experiments,

- Single stands for the ASR system using only one microphone without any interference suppression front-end
- DAS for the system using delay-and-sum beamforming
- GJBF for the system using robust Griffiths-Jim adaptive beamforming techniques (filter taps: 64) proposed in Ref. [6]
- MAIS for the system using techniques proposed in this paper

Different microphone array signal processing techniques are compared under different scenario settings. Table 1 summarizes these results.

Table 1 Comparison of WER results (%)

| Scenario | Single | DAS | GJBF | MAIS |
|---|---|---|---|---|
| Speaker + Interference # 1 | 74.39 | 66.75 | 26.41 | 16.8 |
| Speaker + Interference # 2 | 63.7 | 55.22 | 21.47 | 19.73 |
| Speaker + Interference # 1 + Interference # 2 | 94.12 | 86.99 | 73.24 | 37.21 |

When no interference exists this ASR can achieve average WER of 3.41%. Comparing this result with Table 1，it can be seen that the existence of one interference （with approximate SNR to be 0dB） and two interference （with approximate SNR to be -3dB） ASR performance degrades seriously. The use of various microphone array interference suppression techniques improves the performance. With this small microphone array，delay-and-sum beamforming can achieve only limited noise suppression effects. Comparing the two adaptive methods，the method proposed in this paper （MAIS） works much better than the conventional Griffiths-Jim algorithm （GJBF），especially when the interference number increases and SNR becomes lower.

## 4 Conclusion

In this paper，a near-field source localization algorithm applicable to a small microphone array is proposed. Through the use of the space discriminative function defined by short time Fourier transform of microphone array outputs，we can enumerate the number of concurrent speakers，locate and track their locations effectively and robustly. Based on information provided by this space discriminative function，a frequency-domain post-filter is constructed to suppress interferences. In the case to two concurrent speakers，this post-filter becomes the Wiener filter，and the power spectrum densities in the Wiener filter are estimated through the space discriminative function implicitly. Experiments show that this microphone array front end achieves significant WER reductions when compared with the single microphone，traditional delay-and-sum and Griffiths-Jim beamforming techniques. In future work，the extension of this method to microphone array with more sensors and its performance under more adverse and complicated environmental settings are to be investigated.

**References**

[1] Lin Q，Che C W，Yuk D S，et al. Robust distant talking speech recognition. In：Proc. International Conference on Acoustics，Speech，and Singnal Processing ’96，1996. 21

[2] Omologo M，Matassoni M，Svaizer P，et al. Microphone array based speech recognition with different talker-array positions. In：Proc. International Conference on Acoustics，Speech，and Singnal Processing ’97，1997. 227

[3] Moore D C，McCowan I A. Microphone array speech recognition：experiments on overlapping speech in meetings. In：Proc. International Conference on Acoustics，Speech，and Singnal Processing ’2003，2003. 497

[4] Griffiths L J，Jim C W. An alternative approach to linear constrained adaptive Beamforming. *IEEE Trans Antenna Propagation*，1982，30（1）：27

[5] Kennedy R A，Abhayapala P T D，Ward D B. Broadband near field beamforming using a radial beampattern transformation. *IEEE Trans Signal Processing*，1998，46（8）：2147

[6] Hoshuyama O，Sugiyama A，et al. A robust adaptive beamformer for microphone arrays with a blocking matrix using constrained adaptive filters. *IEEE Trans Signal Processing*，1999，47（10）：2677

[7] Chen J，Louis S，Sun H. A pseudo adaptive microphone array. In：Proc. International Conference on Acoustics，Speech，and Singnal Processing ’2003，2003. 469

[8] Jourjine A，Rickard S，Yilmaz O. Blind separation of disjoint orthogonal signals：demixing N sources from 2 mixtures. In：Proc. International Conference on Acoustics，Speech，and Singnal Processing ’2000，2000. 2985

**Zhao Xianyu**，born in 1976，is a Ph.D. student of Tsinghua University now with major in electronics engineering. His main research interests include array signal processing，adaptive signal processing and robust speech recognition.