NON-SPEECH SOUND RECOGNITION WITH MICROPHONE ARRAY

Kazuo Hiyane, Jun Iio

RWCP Autonomous Function MRI Laboratory
(Mitsubishi Research Institute, Inc.)
2-3-6 Otemachi, Chiyoda, Tokyo, JAPAN 100-8141
hiya@mri.co.jp, iiojun@mri.co.jp

ABSTRACT
We developed a non-speech sound recognition system with sound source direction estimator. Ten types of single impulsive sounds are recognized at the rate of 80% in a second. A delay-sum beamformer with 16 channels microphone array reduces environmental noise by -10dB and estimates source direction at 10 degree resolution. The system is made on dual pentium PC and the processing is almost realtime. Thousands of non-speech sounds recorded in an unechoic room are gathered into a RWC Database that is open for academic use.

1. INTRODUCTION
Studies of non-speech sound recognition are very few except musical sounds although practical speech recognition systems have appeared recently. We get various information from non-speech sounds in daily life. But, it is impossible to recognize all types of sounds for mainly two reasons. One is almost infinite types of non-speech sounds around us. Namely, we don't have adequate categories for non-speech sounds yet. The other is that different sound sources are often heard as the same sound for human auditory sense. In many cases, vision sense helps us to determine its sound source. So, we must begin with investigation of what non-speech sounds are.

First, we investigated Japanese onomatopoeia that are words expressing sounds. There are over 200 words for non-speech sounds [1]. That means Japanese shares common concept about non-speech sounds. We classified onomatopoeia of short-time sounds by linguistic analysis [2]. Then, we found three top categories; impulsive sounds, movement sounds, and characteristic sounds. Next, we recorded many various typical non-speech sounds in unechoic room. These thousands of sounds are gathered into a database, that was opened for academic [3] [4].

We are developing a non-speech sound recognition algorithm using spectral and temporal features and beamformer using microphone array that estimate the position of sound source and that emphasize sound from specific direction[5]. We combine these two techniques into non-speech sound recognition system with directivity.

2. NON-SPEECH SOUND RECOGNITION
2.1. Spectral Structure of Non-speech Sound
Fig.1 shows spectrograms of typical non-speech sounds. These spectra seem to have their own profiles both in spectral axis and in temporal axis. But, their variations are heavily depended on the difference of objects and the way of making sounds. So, it is necessary that feature values of the recognition algorithm are to be selected for each category.

2.2. Algorithm Overview
We categorized non-speech sounds into several types such as single and double impulsive sounds, narrow-band sounds, and broad-band noise. The first problem is to classify these types. The single impulsive sound are caused by an impact of objects, that has characteristic triangular shape in temporal power profile. Its onset duration is very short. Its attenuation is almost linear at logarithmic scale.

(i) Single Impulsive sounds
A single impulsive sound has characteristic triangular shape in temporal power profile. Its onset duration is very short. Its attenuation is almost linear at logarithmic scale. We define the onset time and reverberation time when their power is lower than that of peak time by 20dB (Fig.2). The criteria of single impulsive sound are that the onset duration is less than 20msec

This work was supported by Real World Computing Partnership (RWCP)
and linear decrease between peak time and reverberation time in temporal power profile.

The classification of single impulsive sounds use two spectra at peak time and reverberation time. Each spectrum is divided into 16 bands of 1/4 octave and normalized by peak power. These 32 feature values are modeled by 32-dimensional Gaussian distribution.

(ii) Narrow-banded sounds

The typical narrow-banded sounds are whistle, electronic sounds, musical instruments that these duration time are rather long and spectra are stable in hundreds of milli-seconds. But, simple spectrum template does not work because the pitch is varied. The classification of narrow-band sounds uses spectrum peak detection. Combination of several features are effective; number of peaks, frequency of peaks, sharpness of peaks, and existence of harmonics.

(iii) Broadbanded sounds

The sounds of gas spouting and friction of objects are typical broadbanded sounds. These spectra are spread whole of frequency. This type of sounds can be classified spectrum profile. Each spectrum is divided into 16 bands of 1/4 octave and normalized by total power. These 16 feature values are modeled by 16-dimensional Gaussian Distribution.

Figure 1: Spectral structure of non-speech sounds (Y-axis: frequency[Hz], X-axis: time[ms])

Figure 2: Temporal power profile and spectrum of single impulsive sounds

3. BEAMFORMING WITH MICROPHONE ARRAY

In a real acoustic environment, presence of background noise and multiple sound sources degrades the quality
of sound capture. For this problem, we adopt beamformer with microphone array that forms superdirectivity to the multiple target directions.

We use the delay-sum based beamformer [7]. Its principle is based on the summation of signals at each microphone with delay of arrival time from a target direction. As for a circular microphone array in Fig.3, the delay of time $\tau_i$ at microphone No.$i$ against the center of array is written in the next equation.

$$\tau_i = \frac{(r/c) \cos(i\theta - \phi)}{s}.$$  \hspace{1cm} (1)

Here, $r$ is diameter of array, $c$ is sound velocity. The output signal $y(t)$ of beamforming is written by input signal $x_i(t)$ at $i$-th microphone as,

$$y(t) = \sum_{i=1}^{18} x_i(t - \tau_i).$$ \hspace{1cm} (2)

![Figure 3: Configuration of microphone array for delay-sum filter](image)

We adopt a 16 channels circular array with a diameter of 30cm (Fig.4). To evaluate noise reduction performance, 1kHz pure tone sound is emitted from two meters distance. The resolution of source direction estimation is less than 10 degrees. But, its SNRs are improved 10dB when other noise sources are located at the direction by 30 degrees apart.

4. DEMONSTRATION SYSTEM

We developed a demonstration system on two PCs with 16 channel microphone array. The beamformed signals of 13 directions at every 10 degrees are simultaneously calculated almost in realtime. Ten types of impulsive and narrow banded sounds can be recognized at the rate of about 80%.

The system consists of a microphone array, microphone amplifier, A/D converter and two PCs for recognition and display (Fig.7). The microphone array consists of 16 omnidirectional electret microphones that are installed on a circle with a diameter of 30 cm.

Figure 6 shows the processing flow of the system. The acoustic signal measured by using the microphone array is fed to the first PC through the microphone amplifier and A/D converter. The virtual beamformer of 10 degrees and 13 directions (-60 to +60 degrees) are simultaneously calculated to yield an acoustical signal from each direction. When the signal from the direction of the largest power is yielded and power thereof continues in a certain value, it is assumed that an acoustical event occurred. After the acoustical signal is judged to be the single impulse sound, spectrum matching is carried out for each sound source previously registered.

The direction of the sound source and the type of the sound source are sent to the second PC for display through a LAN and shown on a window for a spectrogram display and a window for displaying the result of sound source type recognition developed by Java language.

In a calm office environment, the direction presumption shows 10 degrees of resolution performance and noise reduction of about -10 dB. In addition, the average sound source type recognition ratio is about 80%. Recognition processing showed almost real time action and recognition result display in 1 sec or less.

5. RWCP SOUND DATABASE

Besides study of non-speech sound recognition, we developed "RWCP Sound Scene Database in Real Acoustical Environments" [3] [4]. It consists of tens of thousands of non-speech sounds recorded in unechoic room, impulse response of microphone array in various rooms, and speech sounds recorded with microphone array. It will be available in 2001, and you can get it at this workshop!
Acknowledgements

The authors would like to thankfully acknowledge the members of RWC Speech and Acoustic Database Working Group, especially Dr. Satoshi Nakamura of ATR Spoken Language Translation Research Labs. and Dr. Futoshi Asano of Electrotechnical Lab. for their helpful support and advices.

REFERENCES


Figure 6: Processing flow of non-speech sound recognition system

Figure 7: Overview of hardware