VOICE SEPARATION BASED ON MULTI-CHANNEL CORRELATION AND COMPONENTS TRACING

Jie Huang† and Noboru Ohnishi†‡

†Bio-Mimetic Control Research Center
The Institute of Physical and Chemical Research (RIKEN)

‡Department of Information Engineering
Faculty of Engineering, Nagoya University

ABSTRACT - This paper presents a novel sound separation method for voices pronounced by multiple persons from different directions. The method uses multi-channel signals from microphones located on different positions. Signals from all microphones are divided into narrow sub-bands by a set of band-pass filter. Envelope section curves over frequencies are calculated at each decimated sampling point. We assume that the voice energy meanly exists in the peaks of envelope section curves. The separation task, then, is to group all of those peaks. Arrival temporal disparities are used to group peaks in the method. Grouping operation, in order to cope with echo, is applied only when a peak just appears. It is because reflected sound arrives later than direct sound, and thus onset only contains sound directly from its source. The peaks are traced after onset and the grouping is maintained. Voice separation experiments were conducted in both an anechoic chamber and a normal room enclosed by concrete walls. The availability of the method was demonstrated.

INTRODUCTION

The ability of a human to separate a particular sound from a noisy environment is called the "cocktail party effect" (Cherry 1953). Some proposed speech separation methods utilize monaural cues like the harmonic and synchronous characteristics of speech (Parsons 1976, Cooke 1992). A single monaural cue, however, is usually not perfectly sufficient for sound separation. In speech only vowels have harmonic characteristic, and synchronous will not be maintained in all frequency components. It is well know that the "cocktail party effect" is more effective when listening binaurally than listening monaurally, i.e., multiple channel signal is more effective than single channel signal. However, multi-channel based sound separation methods usually focus on sound parameters estimation by using total sound components. In a very complicated environment, they need a large computation power and are difficult to build a mathematical model. They will be not robust against noise, and are difficult to adapt to multi-path environments.

Our approach is to segment sound into small pieces in both time and frequency domain and then group them into several groups by the difference of arrival temporal disparities (ATD) between different microphones. Each sound piece contains narrow band frequency component and has an onset as starting point and an offset as ending point. A new method named as Group Onset Trace Ongoing Method (GOTO Method) was proposed. It groups sound pieces at onsets by ATDs and then traces the peaks of envelope section curve to keep the grouping. The benefit of sound segmentation is each sound piece can be considered as coming from single sound source; and grouping sound pieces by onset ATDs can eliminate the adverse influence of echoes.

GROUP ONSET TRACE ONGOING METHOD

Envelope Section Curve and Peak Detection

Sound signals are past to a set of narrow band-pass filter. Envelope of each sub-band signal is calculated and decimated by a factor D. The Fourier transform was not used because the averaging nature and time-frequency resolution trade-off characteristic. Envelope section curves over frequencies are also introduced. Each of them has the values of all sub-band envelope curves at each sampling point. Peaks and dominant regions are determined at envelope section curves by a detection algorithm (Figure 1).

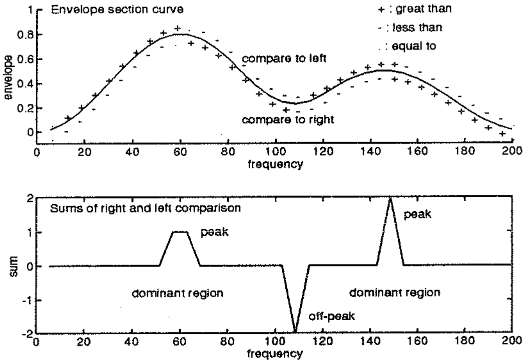

Figure 1: Peak detection algorithm

In this algorithm, each point is compared to its left and right, Each comparison takes value of 1, -1 or 0 when it is great than (+), less than (-) or equal to (.), respectively. It will seldom occur in practical data that neighboring points equal to each other (see left peak in Figure 1), and no consideration was taken for this so far. A peak is when the sum of left and right comparison is great than 0, and an off-peak when less than 0. The region from a peak to off-peaks in both sides is called the dominant region of the peak. The peaks, instead of dominant regions, will be used to approximate total sound signal. The approximation is sufficient because sound energy meanly exists on its envelope peaks.

Onset Detection and Grouping

How does a human localize and separate sounds in an echoic environment? The "precedence effect" (Wallach 1949) shows that the human auditory system can detect the beginning of a sound and mask the subsequent portion of that sound. This phenomenon is concerned to eliminate the influence of echo. It is because reflected sound arrives later than direct sound and thus sound onset only contains sound directly from its source.

We defined sound onset, ongoing and offset in a heuristic manner (Figure 2). A starting portion of a direct sound should be an onset. A starting portion of a long term echo comes in the silent portions of direct sounds can also be an onset, because a long term echo can be considered as another sound source distinguished from the direct sound source. The criterion for onset points detection are (1) a point where sound level becomes higher than a threshold level; or (2) a point where sound level becomes L (a constant) times higher than the previous point. The second criterion is prepared to detect onsets of direct sound in the ongoing of long term echo sounds.

Arrival temporal disparities (ATDs) are calculated in sound onsets. Histograms are produced from the ATDs and weighted with their level (Figure 3). The histograms are contributed by only one value from a same frequency component. When a new onset appears in a same frequency band, the previous one will be removed from the histograms. The histograms will also be renewed by a certain time period. The histograms are smoothed and mapped to a azimuth histogram by the Histogram Mapping Method (Huang *et al.* 1994). Three ATD histograms

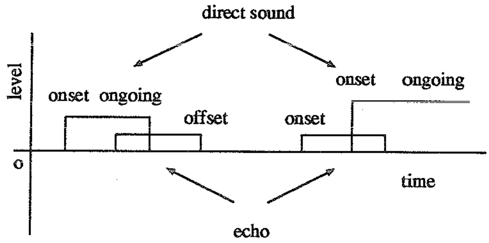

Figure 2: Definition of onset, ongoing and offset protions

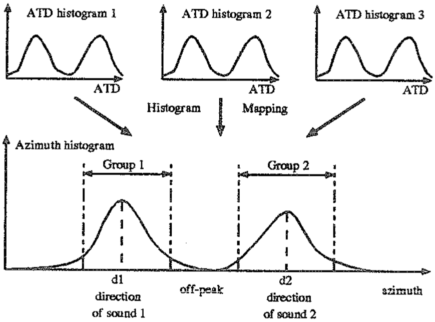

Figure 3: The histogram mapping method

from three microphones are usually used in the method. It is because the mapping is a one to many mapping, a single ATD histogram is not enough for sound localization. Again, the azimuth histogram is smoothed and peaks and off-peaks are detected by the same method mentioned in the previous section. In the azimuth histogram, the number of peaks is the number of sound sources, and each azimuth where a peak appears is the direction of the sound source. Then, sound onsets can be divided into N groups same as the number of sound sources. All onsets contributed to a same dominant region (usually a 80% region is used) of a peak are then grouped to a same group. The sound source number N is not limited because the nature of the method.

Peak Tracing

The ongoing portion is the mixture of direct sound and echo sound. Arrival temporal disparity (ATD) or phase disparity at this portion will be drastically distorted by the echo sound. A onset portion and its subsequence ongoing portion, however, are usually coming from a same sound source. There is no need to group ongoing portions, the only thing to do is to keep the grouping in the periods of sound ongoing.

A peak tracing method is illustrated in Figure 4. In the figure, the solid lines and dotted lines indicate the peak tracing lines and off-peaks lines respectively. The circle marks represent the onsets. The peaks are traced on the conditions that the previous peak is located in the dominant region of the next peak, and the next peak is located in the dominant region of the previous

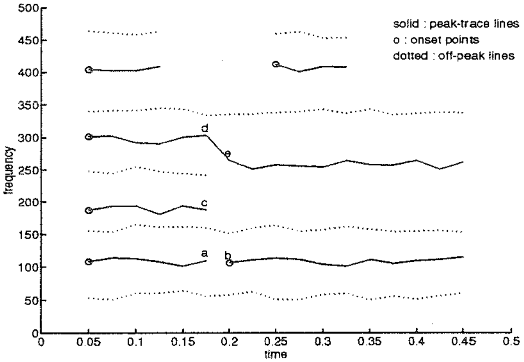

Figure 4: Peak trace method

peak. For example, both peak 'd' and 'c' are located in the dominant region of peak 'e', but peak 'e' is only located in the dominant region of peak 'd'. Thus, peak 'd' traced to peak 'e', and peak 'c' is an offset point. Another case is peak 'a' and peak 'b', where peak 'b' is great than L times peak 'a'. The tracing stops on peak 'a' as an offset, and peak 'b' is an onset point.

So long, the sound signals are segmented into pieces by the Group Onset Trace Ongoing Method in both frequency domain and time domain. Each piece is refer to a continue peak series determined by the peak tracing method. It has an onset as starting point and an offset as ending point. The data processing flow is shown in Figure 5.

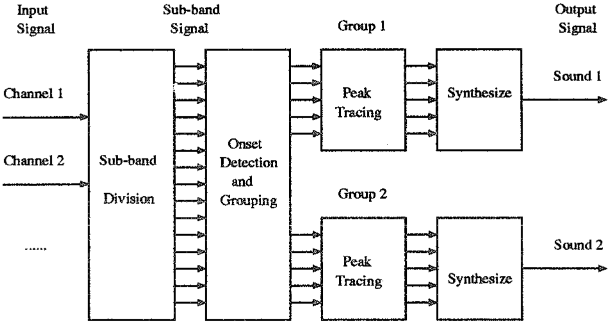

Figure 5: Data processing flow

EXPERIMENTS AND RESULTS

Experiments

Experiments were conducted in an anechoic chamber and a normal room enclosed by concrete walls (Figure 6). The anechoic chamber dimensions are $5.5 \times 5.5 \times 5.5m^3$. The cutoff frequency is 200Hz and the background noise is 20dB. The normal room is an empty room enclosed by concrete walls and no acoustic treatment was applied. The room area is about $30m^2$. The sound sources were, (1) radio weather forecast by a male announcer (speaker 1), and (2) radio talk show by male and female hosts (speaker 2). Each sound had a period of about 20 seconds. The azimuth between sound source 1 and sound source 2 was about 38 degrees. Each microphone (ml, m2 and m3 in Figure 6) was set at a vertex of an equilateral triangle (side length 13.5cm). The triangle could be turned around its center to change the directions of the sound sources. The distance from sound source to microphones is about 2.9m.

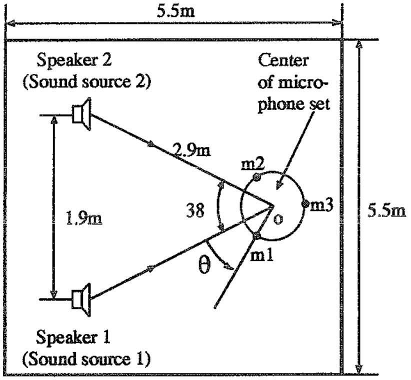

Figure 6: Experimental setup in room

Results and Discussions

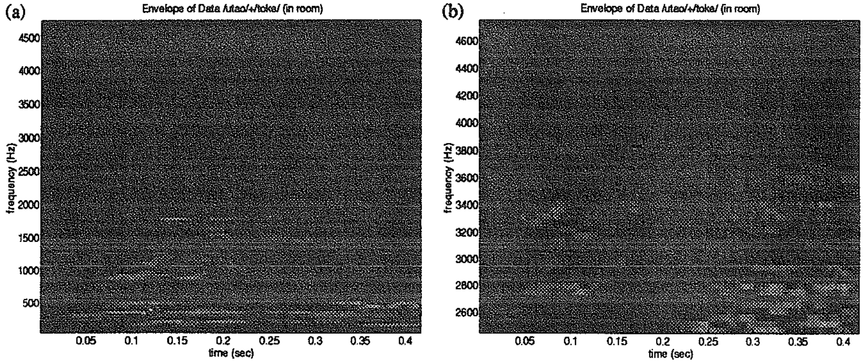

Figure 7: Envelopes of sub-band signals

Sound signals were divided into narrow sub-bands by a set of band-pass filter with 48Hz bandwidth. Envelopes of sub-band signals were calculated. A example of the envelopes (with decimation factor 10) is shown in Figure 7, where Figure (b) enlarges Figure (a) above 2500Hz. It is a similar one to the spectrogram, but is not averaged. The components seem easier to be traced under about 2500Hz, because the outlines of peaks are clearly in line shapes. The components above 2500Hz, however, have few continue peaks.

A example of the peak tracing is shown in Figure 8. It was shown that the major components have been traced over all frequencies. The peaks that can not be traced are also be grouped as

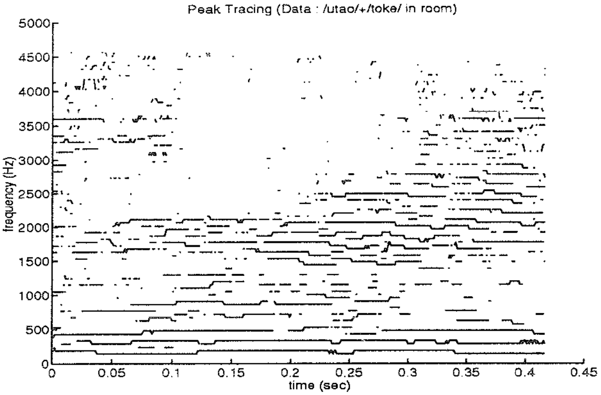

Figure 8: Results in Peak Tracing

isolated points.

An old version program using the Fourier transform and phase disparities is tested for sound separation. It works good on sound signals recorded in the anechoic chamber. It, however, can not separate the sound signals recorded in the normal room. The method mentioned in this paper is to improve the time resolution and the averaging nature of the Fourier transform. It is expected to solve the above mentioned problems.

CONCLUSION

We have proposed the Group Onset Trace Ongoing Method. The target sound of the method is voice or speech. The method is available for sound separation in a multi-path environment or in a normal room. It groups sounds into groups according to how many sound groups exit. Experiments are conducted and being conducted to demonstrate the availability of the method.

REFERENCES

Cherry, E. C. (1953), *Some experiments on the recognition of speech with one and with two ears*, J. Acoust. Soc. Am., vol. 25, pp. 975-979.

Cooke, M. P. (1992), *An explicit time-frequency characterization of synchrony in an auditory model*, Computer Speech and Language, vol. 6, pp. 153-173.

Huang, J. and Ohnishi, N.(1994), *A Biomimetic System for Localization and Separation of Multiple Sound Sources*, Proceedings of IEEE, IMTC/94 pp.967-970 (Hamamatsu, Japan).

Parsons, T. W. (1976), *Separation of speech from interfering speech by means of harmonic selection*, J. Acoust. Soc. Am., vol. 6 no. 4, pp. 911-918.

Wallach, H., Newman, E. B. and Rosenzweig, M. R. (1949), *The precedence effect in sound localization*, J. Psychol. Am., vol. 62-3, pp. 315-331.