# A Sequential Processing Model for Speech Separation Based on Auditory Scene Analysis

Isao Nakanishi and Junichi Hanada
Graduate School of Engineering
Tottori University
4-101 Koyama-minami, Tottori 680-8552, Japan

***Abstract*—Speech separation based on auditory scene analysis (ASA) has been widely studied. We propose a sequential processing model of computational ASA (CASA), in which a mixed speech is sequentially decomposed into frequency signals using modified Discrete Fourier Transform (DFT), four features in ASA are extracted from the decomposed frequency signals, the frequency signals are regrouped by examining the extracted features, and each separated speech is obtained by recomposing the frequency signals in a group. In this paper, we attempt to separate speeches only using the harmonic structure, which is one of the features and regarded as the backbone in our sequential implementation model.**

***Keywords—speech separation; auditory scene analysis; sequential processing model***

## I. INTRODUCTION

Speech separation based on auditory scene analysis (ASA) [1] is widely studied. Human beings can hear a specific speech in an environment where many people are simultaneously speaking. This ability is famous as cocktail party effect. ASA gives us a model which psychologically explains the cocktail party effect. Concretely, speech separation is achieved by extracting four features: common onset/offset, harmonic structure, common change, and gradual change from a mixed speech and then grouping frequency signals which have the common features.

Computational ASA (CASA) is to implement ASA in computer systems [2], which is generally based on time-frequency analysis (like a spectrogram) obtained in block processing. In addition, learning functions are recently installed for improving the reproducibility of original speeches [3], [4], [5], [6], [7], [8], [9], [10], [11], in which all features are extracted and learned in advance of separation. However, these approaches are not suitable for real-time processing.

Thus, we propose to sequentially implement CASA for achieving real-time processing of ASA. Here, the "sequential" means a processing at every sampled time. As far as I know, this is the first approach to sequentially implement CASA. The speech separation ability of the proposed method is certainly limited since the information which can be obtained sample by sample is less than the information obtained by block-processing. On the other hand, the proposed method can be achieved in a low computational complexity.

In the proposed approach, a mixture signal is sequentially decomposed into frequency signals by using a modified DFT (MDFT) [12]. Four features in ASA are sequentially extracted from the decomposed frequency signals. By using the extracted features, decomposed frequency signals which were assumed to compose an unmixed signal are grouped. Original signals are obtained by recomposing the grouped frequency signals. In this paper, we attempt to separate a mixed speech only using the harmonic structure, which is one of four features and regarded as a backbone in our approach.

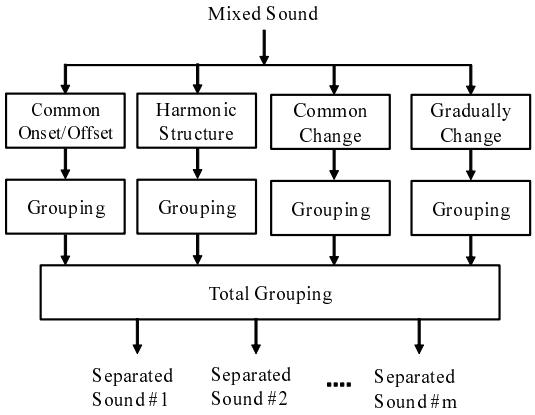



Fig. 1. Single input model of ASA.

## II. AUDITORY SCENE ANALYSIS

ASA psychologically explains the auditory mechanism of human beings and gives us a framework for implementing the auditory function of human beings in computational systems [1], [2]. The single input model of ASA is illustrated in Fig. 1. From a mixed sound, four features: common onset/offset, harmonic structure, common change, and gradual change, are extracted and then grouping is achieved by using individual feature. Total grouping is performed by considering the grouping results and generates separated sounds.

## III. SEQUENTIAL PROCESSING OF ASA USING MDFT PAIR

In order to implement ASA in real-time processing, we adopt a modified DFT (MDFT) pair[12]. The MDFT pair is obtained by simplifying an original DFT pair and defined as

$$Y_{k,i} = \sum_{n=0}^{N-1} x_{i-n} \ \cos(2\pi nk/N) \tag{1}$$

$$x_i^{`} = \frac{Y_{0,i}}{N} + \frac{2}{N} \sum_{k=1}^{N/2-1} Y_{k,i} \tag{2}$$

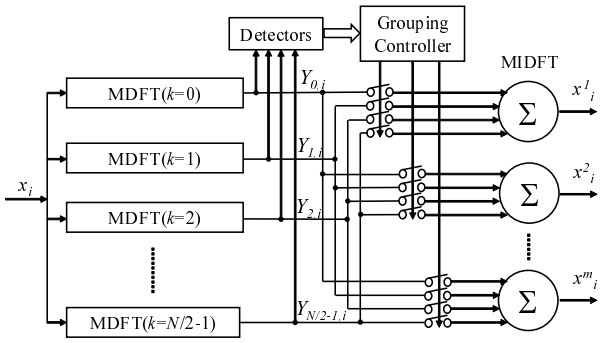



Fig. 2. Sequential implementation of ASA using MDFT pair.

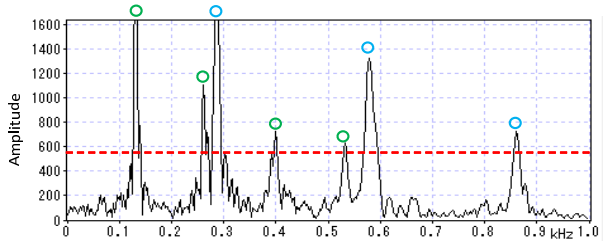



Fig. 3. Detection of spectral peaks.

where $N$ is the number of samples for DFT analysis and assumed to be even hereafter. MDFT requires only real number calculations and MIDFT is achieved by summing MDFT outputs. In other words, MDFT sequentially decomposes an input signal into frequency signals while maintaining their phase differences; therefore, MIDFT is achieved, that is, the input signal is recomposed only by adding the frequency signals. Please refer to Ref. [12] about the details of a MDFT pair.

Moreover, in the case of applying a window function, MDFT is rewritten as [13]

$$Y_{k,i} = \sum_{n'=-N/2}^{N/2-1} x_{i-(n'+N/2)} \cos(2\pi n' k/N) \ , \qquad (3)$$

where $n' = n - N/2$. This modification causes the delay of $N/2$ in an output.

Sequential implementation of ASA using MDFT pair is described in Fig. 2. A mixed sound $x_i$ is sequentially decomposed into frequency signals by using MDFT. From the frequency signals, four features of ASA are extracted by the Detectors. In the Grouping Controller, it is determined which group each frequency signal belongs to by using the extracted features. The grouped frequency signals are composed in MIDFT and a separated signal $x_i^m$ is obtained.

It can be also performed to extract frequency signals from a mixed signal by using a filter bank or a sinusoidal modeling. However, it is important to guarantee linear phase in extracted frequency signals. MDFT perfectly guarantees the linear phase characteristic.

## IV. Detection of Harmonic Structures

It is natural that frequency signals which compose a speech have different amplitudes and phases. This fact greatly influences on sequential detection of commonality features: common onset/offset and common change; therefore, it is expected that the detection accuracy of commonality features is low.

Thus, we regard the harmonic structure of ASA features as the backbone of the processing and supplementarily use other features. This point greatly differs from the conventional CASA based on block processing, and is first found by trying to sequentially implement ASA. In this paper, we attempt to separate a mixed speech only using the harmonic structure feature.

### A. Preprocessing of decomposed frequency signals

First, decomposed frequency signals are processed using a moving average with 100 samples in order to simply suppress noises. Sampling rate is assumed to be 8 kHz. Next, their envelopes are extracted by using the signal level detector, which had been proposed in Ref. [14].

### B. Detection of spectral peaks

From the preprocessed frequency signals, spectral peaks are sequentially detected. The detection method is not novel and its concept is described in Ref. [15] for example. Since the purpose of this research is to sequentially implement CASA, it will be welcome to introduce other harmonics detectors with high accuracy.

The procedure is explained using Fig. 3, where the horizontal axis indicates frequency and the vertical axis is amplitude at a sampling time. If the amplitude of an enveloped frequency signal is larger than those of frequency signals at the both adjacent frequencies, there assumed to be a spectral peak at the frequency. However, we want to detect only discriminative spectral peaks. By setting a threshold that is shown as the dashed line, the spectral peaks with ∘ mark that are larger than the threshold are finally remained. Note that it is necessary for us to adjust the threshold to the amplitude of an input.

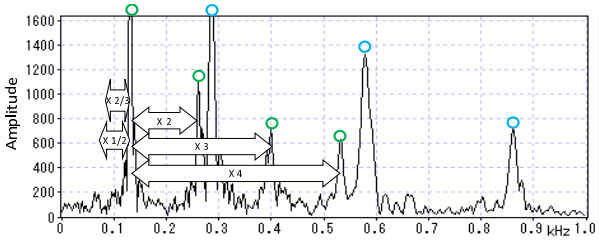
### C. Grouping by a harmonic structure

In general, a sound consists of a fundamental frequency signal and its harmonics. By applying such relationship to detected peaks, we can extract a harmonic structure.

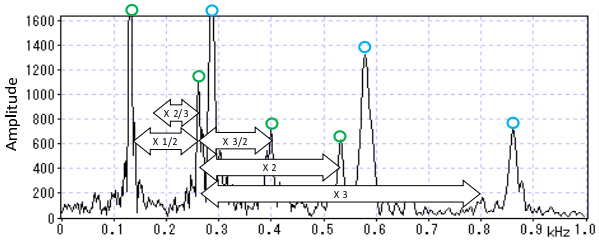
Focusing on a detected spectral peak, it is examined whether other spectral peaks exist on integral or fractional multiple frequencies. The searching procedure is explained using Fig. 4.

Getting the frequency of the first spectral peak as shown in (a), the existence of other spectral peaks on its integral multiple and fractional multiple frequencies is examined. In this case, there are spectral peaks on twice, three times, and four times the frequency. As a result, the frequency signals of the first, second, fourth, and fifth peaks become candidates for a harmonic structure.

The similar searching is performed on the second peak as shown in (b). There are spectral peaks on half, $3/2$ times,

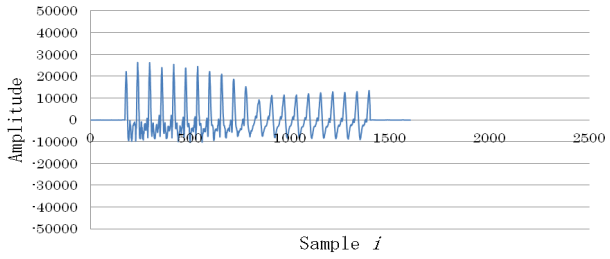



(a)

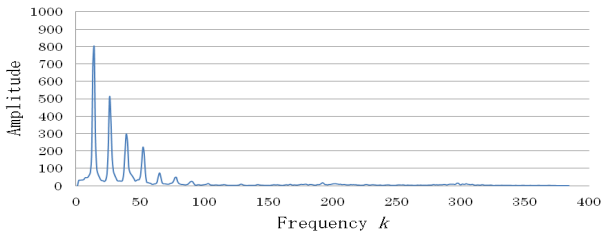



(b)

Fig. 4. Detection of harmonics structure.




(a) Speech signal (/te/)




(b) Spectrum

Fig. 5. A male speech and its spectrum.

and twice the frequency, and then the frequency signals of the first, second, fourth, and fifth peaks become candidates for a harmonic structure.

This searching is performed on all detected spectral peaks. As a result, spectral peaks that are commonly candidates are regarded as of a harmonic structure. In Fig. 4, frequency signals of the first, second, fourth, and fifth peaks are considered to compose a harmonic structure.

However, if the number of spectral peaks that are regarded to compose a harmonic structure is less than half number of all detected peaks, the harmonic structure is not adopted. This prevents the false detection of harmonic structures.

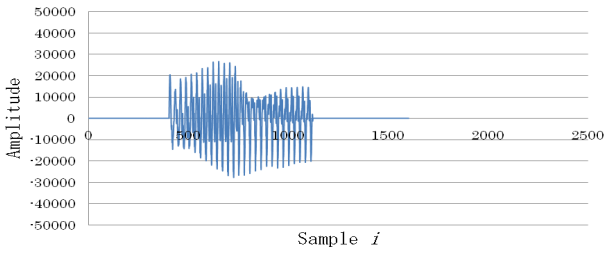
Frequency signals with a common harmonic structure are grouped and then recomposed to generate a separated sound. However, an original sound is never reconstructed by gathering only harmonics. Frequency signals that are adjacent to the harmonics are also necessary to reconstruct an original sound. In this paper, adjacent frequency signals ($k \pm 2$) to harmonics are grouped in total grouping.

## V. Evaluation in a Simulation

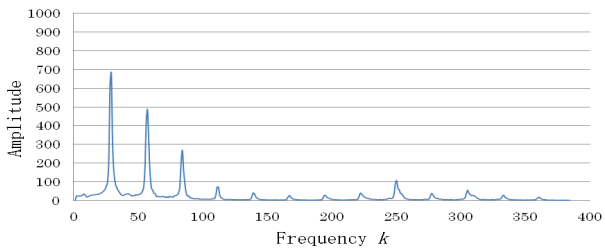
In order to evaluate the ability of the proposed sequential model of ASA, especially only by using the feature of harmonic structure, we carried out a computer simulation.

We prepared two speeches (phonemes) pronounced by a Japanese male and a Japanese female from a speech database. By adding two speech signals, we obtained a mixed speech. The speech signals and the spectra of the male, female and mixed signal are shown in Fig. 5, 6, and 7, respectively.

The number of samples for MDFT was $N = 768$; therefore, the maximum frequency is $N/2-1 = 383$. The threshold



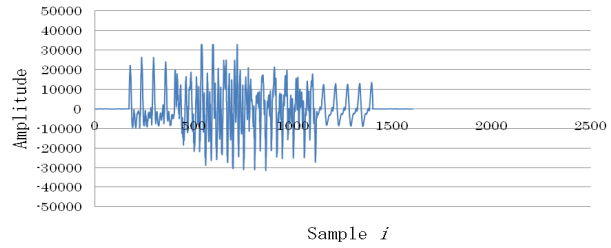

(a) Speech signal (/se/)



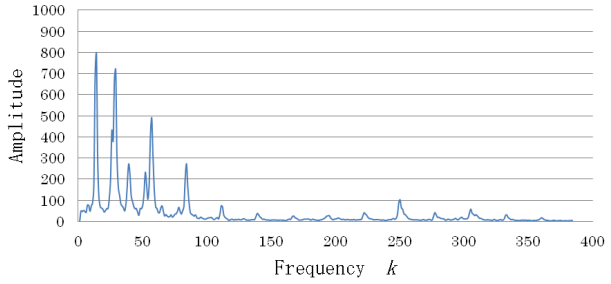

(b) Spectrum

Fig. 6. A female speech and its spectrum.

for extracting spectral peaks was set to 100 plus twice the mean of the input spectrum. Considering the pitch of human voices, the search range of a harmonic structure was $k \geq 7$, which was $\approx 73$ Hz. In general, the number of main harmonics is approximately four in a human voice; therefore, the searching range for the detection of harmonic structure was from $1/4$ to 4. Concretely, integral multiple and fractional multiple values

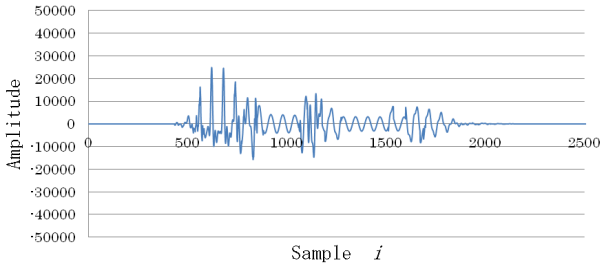



(a) Mixed speech signal

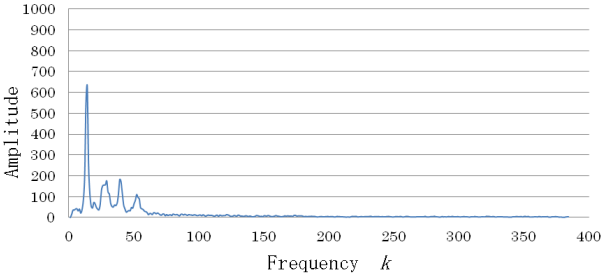



(b) Spectrum

Fig. 7. A mixed speech and its spectrum.

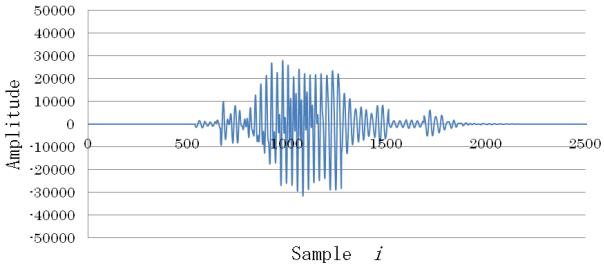



(a) Separated speech

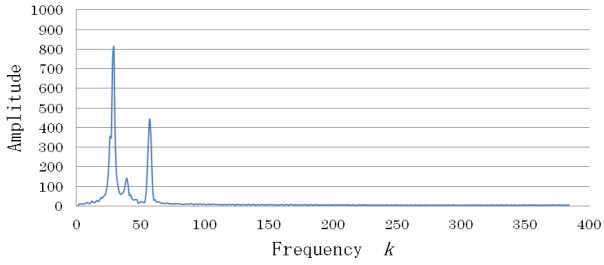



(b) Spectrum

Fig. 8. The separated speech and its spectrum.




(a) Separated speech




(b) Spectrum

Fig. 9. The separated another speech and its spectrum.

for searching the harmonic structure was 1/4, 1/3, 1/2, 2/3, 3/4, 1, 5/4, 4/3, 3/2, 5/3, 7/4, 2, 9/4, 7/3,5/2, 8/3, 11/4, 3, 13/4, 10/3, 7/2, 11/3, 15/4, and 4.

Figures 8 and 9 show separated speeches and their spectra. Comparing an unmixed speech in Fig. 5 or 6 with a separated one in Fig. 8 or 9, respectively, it is confirmed that there was the time-delay of approximately 400 samples. This is due to the processing delay in MDFT and the preprocessing of frequency signals.

Comparing the spectra of unmixed speeches with those of separated speeches, it is confirmed that the spectral elements could be roughly separated. However, there are some breaks in the separated signals. In addition, it is noticeable that higher harmonics than fourth were not reconstructed in the male case. This is due to setting of the search range. In addition, spectral peaks are detected in higher frequency range but their amplitudes are greatly attenuated. In this paper, adjacent frequency signals ($k \pm 2$) to harmonics are automatically grouped in the total grouping. Some adaptive method for controlling adjacent frequency signals may solve the problem. Other features in ASA which were not implemented in this paper are expected to be applied to the method.

## VI. Conclusions

We have studied to sequentially implement ASA. In our model, a mixture signal is sequentially decomposed into frequency signals in parallel by using MDFT, four features of ASA are extracted from the decomposed frequency signals, the decomposed frequency signals are grouped by using the extracted features, and original signals are reconstructed by composing the grouped frequency signals. In particular, the harmonic structure becomes the backbone of processing and other features are used supplementarily.

Thus, in this paper, we evaluated the reconstruction ability of separated speeches only by using the harmonic structure feature. As a result, it was confirmed that the spectral elements

of each speech could be roughly separated but there were some breaks in the separated signals. There are many issues to be re-considered.

Our challenge is quite primitive. To implement other feature extraction and to group using all features are also urgent issues to be studied. After implementing all functions, we will have to evaluate the speech separation performance using various speeches.

## REFERENCES


[1] A. S. Bregman, "Auditory Scene Analysis: Hearing in Complex Environments", Chapter 2 in S. McAdams and E. Bigand (Eds.), Thinking in Sound, Oxford Univ. Press, 1992.

[2] D. Wang and G. J. Brown, "Computational Auditory Scene Analysis", IEEE Inc., 2006.

[3] Y. Shao, S. Srinivasan, Z. Jin, and D. Wang, "A Computational Auditory Scene Analysis System For Speech Segregation and Robust Speech Recognition", Computer Speech and Language, vol. 24, pp. 77–93, 2010.

[4] P. Li, Y. Guan , S. Wang , B. Xu, and W. Liu, "Monaural Speech Separation Based on MAXVQ and CASA for Robust Speech Recognition", Computer Speech and Language, vol. 24, pp. 30–44, 2010.

[5] M. Cooke, J. R. Hershey , and S. J. Rennie, "Monaural Speech Separation and Recognition Challenge", Computer Speech and Language, vol. 24, pp. 1–15, 2010.

[6] C. Hsu, and J. R. Jang, "On the Improvement of Singing Voice Separation for Monaural Recordings Using the MIR-1K Dataset", IEEE Trans on Audio, Speech, and Language Processing, vol. 18, No. 2, pp. 310–319, 2010.

[7] G. Hu and D. Wang, "A Tandem Algorithm for Pitch Estimation and Voiced Speech Segregation", IEEE Trans on Audio, Speech, and Language Processing, vol. 18, No. 8, pp. 2067–2079, 2010.

[8] Z. Jin and D. Wang, "Reverberant Speech Segregation Based on Multipitch Tracking and Classification", IEEE Trans on Audio, Speech, and Language Processing, vol. 19, No. 8, pp. 2328–2337, 2011.

[9] A. Rabiee, S. Setayeshi, and S. Lee, "A Harmonic-Based Biologically Inspired Approach to Monaural Speech Separation", IEEE Signal Processing Letters, vol. 19, No. 9, pp. 559–562, 2012.

[10] A. Mahmoodzadeh, H. R. Abutalebi, H. Soltanian-Zadeh, and H. Sheikhzadeh, "Single Channel Speech Separation in Modulation Frequency Domain based on a Novel Pitch Range Estimation Method", EURASIP Journal on Advances in Signal Processing, 2012.

[11] W. Yu, L. Jiajun, C. Ning and Y. Wenhao, "Improved Monaural Speech Segregation Based on Computational Auditory Scene Analysis"C EURASIP Journal on Audio, Speech, and Music Processing , 2013.

[12] S. Yoneda, I. Nakanishi, I. Sasaki, and A. Ogihara, "Switched-capacitor DFT and IDFT circuit", Int. J. Electronics, vol. 67, no. 6, pp. 839–851, Dec. 1989.

[13] I. Nakanishi, Y. Nagata, T. Asakura, Y. Itoh and Y. Fukui, "Speech Noise Reduction System Based on Frequency Domain ALE Using Windowed Modified DFT Pair", IEICE Trans. Fundamentals, vol. E89-A, no. 4, pp. 950–959, Apr. 2006.

[14] Y. Minato and I. Nakanishi, "Noise Reduction System Using Signal and Noise Level Detectors in Frequency Domain", Proc. of 2008 IEEE International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS2008), pp. 180–183, Mar. 2009.

[15] X. Serra and J. O. Smith, "Spectral Modeling Synthesis", Proc. of International Computer and Music Conference, pp. 281–284, Nov. 1989.