ASA Based Unitary Input Model for Sequential Processing of Speech Separation

Isao Nakanishi, Motohiro Ichikawa, and Naoto Sasaoka
Graduate School of Engineering, Tottori University, Tottori, Japan
Email: nakanishi@tottori-u.ac.jp

***Abstract*—Speech separation based on auditory scene analysis (ASA) has been widely studied. In this study a computational ASA model, in which a mixed signal is sequentially decomposed into frequency signals using a modified discrete Fourier transform (MDFT), has been proposed. Four feature types of ASA are extracted from the decomposed frequency signals based on simple rules, and the decomposed frequency signals are regrouped by examining the characteristics of the extracted features. Finally separated speeches are obtained by adding the regrouped frequency signals in a modified inverse DFT. The separation performance of the proposed model is examined via computer simulations and subjective evaluations.**

***Index Terms*—speech separation, auditory scene analysis, unitary input, sequential processing, modified discrete Fourier transform, subjective evaluation**

I. Introduction

Speech separation is actively studied worldwide. It can be applied to the hearing function of a robot, automatic generation of conference minutes, and automatic scoring of music. Speech separation involves two techniques that use multiple and unitary inputs (microphones).

As the multi-input method, blind source separation (BSS), which is a statistical method based on the independent component analysis (ICA), has gained attention. The transform (mixture) matrix from multiple inputs to measured data is estimated; then, speech separation is performed using its inverse matrix. BSS achieves superior separation performance; however, it requires an assumption that multiple sound sources are independent and that the number of microphones is greater than or equal to the number of sources.

Auditory scene analysis (ASA) is proposed as the unitary input method [1]. Human beings can hear specific speeches in an environment where people speak simultaneously. This ability is well known as the cocktail party effect. The ASA psychologically explains the auditory mechanism of human beings. A mixed speech can be separated by extracting four features: common onset/offset, harmonic structure, common changes, and gradual changes. Then, the extracted features are grouped.

Computational ASA (CASA) processes ASA in a computational algorithm [2], which is based on the time-frequency analysis (spectrogram) obtained via block processing. In addition, the separation performance of mixed speeches and the reproducibility of original speeches will be improved by adopting a leaning function [3]-[11], in which all features are extracted in advance for separation. The unitary input method can eliminate the condition that the number of microphones has to be greater than or equal to that of the sources.

This study aims to realize CASA in sequential processing based on simple rules, which is more suitable for real-time processing than block processing. In contrast, the separation performance may be degraded as the available features in a sampling period are restricted compared with block-processing models. This study also aims to investigate how the four features of ASA are implemented in the sequential processing and to clarify what the sequential processing of CASA can and cannot accomplish.

A basic model for the sequential processing of CASA has been proposed previously [12], [13]. However, the separation performance has been evaluated using only a mixed speech; therefore, the effectiveness of the proposed model has not been fully investigated. In this paper, the proposed model is re-explained in detail and the robustness of the settings for the proposed model is visually evaluated in the results using several mixed speeches. In addition, the separation performance is subjectively evaluated using Separation Mean Opinion Score (SMOS) as a new evaluation criterion.

II. Sequential Processing Model of ASA

ASA has been proposed to provide a framework for clarifying the auditory function of human beings [1]. In ASA, four physical features in a mixed signal, namely "common onset/offset," "harmonic structure," "common change," and "gradual change" play prominent roles. The concept model for ASA is described in Fig. 1.

We have proposed to realize sequentially the unitary input model of ASA using a modified discrete Fourier transform (MDFT) pair [12], which is illustrated in Fig. 2. The MDFT pair is defined as the following equations [14]. The MDFT is realized using FIR filter bank and the modified inverse DFT (MIDFT) is realized only by adding the MDFT outputs.

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

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

Manuscript received January 7, 2019; revised June 26, 2019.

doi: 10.18178/ijsps.7.3.78-84

A mixed speech x_i is sequentially decomposed into frequency signals $Y_{k,i}$ using MDFT. From the frequency signals, the four features of ASA are extracted by the detectors. In the grouping controller, the group to which each frequency signal belongs to is determined using the extracted features. The grouped signals are added in the MIDFT and a separated sound x^m_i is generated.

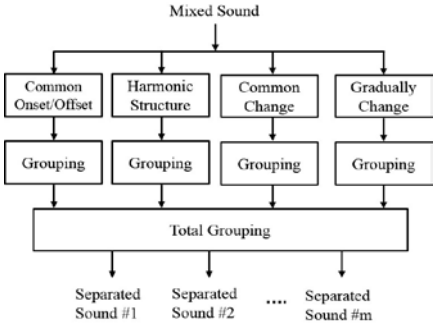


Figure 1. Concept model of ASA.

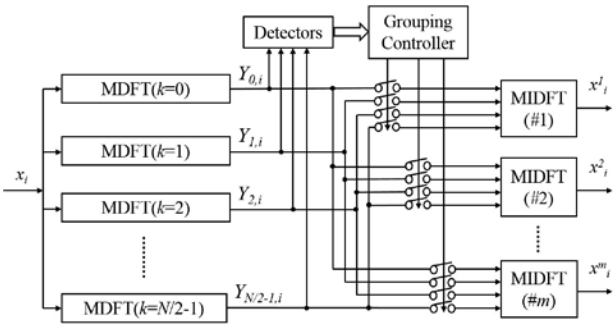


Figure 2. Sequential processing of ASA using a MDFT pair.

A. Detectors

Before the detection of four features, the frequency signals decomposed from an input signal are averaged using a moving window with 100 samples. Then, their envelopes are detected using the signal level detectors proposed in [15] to examine the global behavior of each signal.

1) Detection of common onset/offset

For detecting the common onset/offset features in frequency signals, the onset/offset points in each signal are determined. If the amplitude of a frequency signal at a given time is smaller or greater than a threshold, a label with the value "0" or "1," respectively, is applied. The time when the value "0" changes to "1" is regarded as an onset point. Inversely, the time when the value "1" changes to "0" is regarded as an offset point. The threshold value must be adjusted according to the amplitude of an input. Frequency signals with the same value are regarded as being common.

2) Detection of common changes

The variation of a frequency signal is represented by an increase or decrease of the signal. The amplitude of a frequency signal at a given time is subtracted from that in previous 100 sampling periods. If the result is positively larger than a threshold, "+1" is assigned at the frequency, whereas "-1" is assigned in the negative case. If the result is smaller than the threshold, "0" is assigned. If the frequency signals have the same value at the same time, they are regarded as being commonly changing.

3) Detection of harmonic structure

The harmonic structure is the backbone of processing in the proposed model and the detection accuracy of this structure greatly influences the speech separation performance. In [12], harmonic structures are extracted and the fundamental frequencies are determined using the harmonics; however, this caused misdetection wherein the grouped harmonics included unnecessary frequency signals. In this study, an improved detection method of harmonic structures is introduced to detect the fundamental frequencies.

The extraction method of spectral peaks is identical to that used in the conventional method, which is not novel and its concept is described in [16] for example. However, the detected peaks under the decided frequency are regarded as the fundamental frequencies in the proposed method, whereas all detected spectral peaks are candidates for the fundamental frequencies in the conventional method. Harmonic frequencies are estimated based on the phenomenon that the frequencies of harmonics are integral multiples of the fundamental frequency. However, this phenomenon is not always true. The frequency values of harmonics slightly vary in actual voicing samples.

Let us explain the issue using Fig. 3 where the spectral peaks are detected at frequency k=16, 30, and 45 and then the fundamental frequency is regarded as 16. However, if the harmonics searching using the fundamental frequency of k=16 never detect its harmonics of k=30 and 45. The fundamental frequency may be changed from its true value 15 because of its variation.

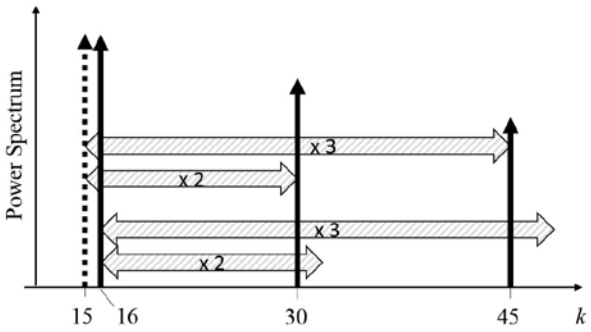


Figure 3. An example of variation of harmonics.

For addressing this problem, ±1 frequencies of the integral-multiple frequencies of the fundamental frequency are also regarded as harmonics. The fundamental frequency also varies slightly; therefore, the fundamental frequency and its ±1 frequencies are used for estimating harmonics. In other words, the above integral multiplication is always achieved at three frequencies (the fundamental frequency and its plus and minus 1 ones). From the three candidates obtained, the frequency with the largest number of harmonics is determined as the true fundamental frequency and its integral-multiple frequencies are detected as harmonics.

Another issue is that the spectral peaks of all harmonics are never simultaneously detected. Harmonics in a real speech does not always vary simultaneously as they have different amplitudes and phases. Figure 4 shows an instance of the time variation of the detected spectral peaks at frequencies k, $2k$, and $3k$, which correspond to the harmonics in a real speech. Even in harmonics, their spectral peaks can never be extracted synchronously. This phenomenon poses a problem in sequential processing as it causes the misdetection of harmonics.

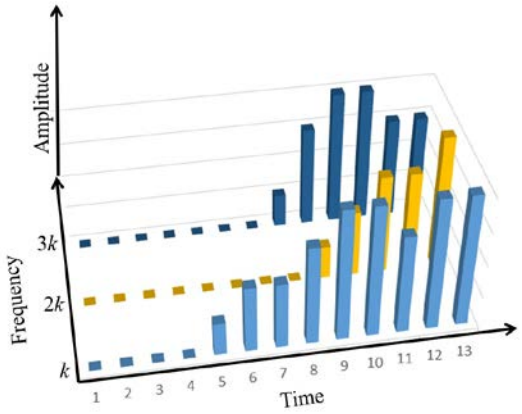


Figure 4. Time variation of the detected spectral peaks.

In this study, a moving-window method is introduced for mitigating the misdetection. Let us explain the proposed method using Fig. 5, where "1" indicates the existence of spectral peaks and *2f* denotes the window size.

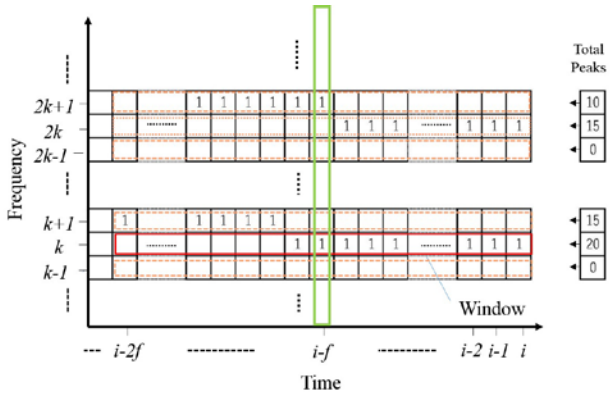


Figure 5. Detection of spectral peaks using a moving window.

Assuming the current time is i, the detection of the harmonic structure is always examined at i-f. In Fig. 5, it is assumed that there are two spectral peaks at k and $2k$+1 at i-f and the lower peak corresponds to the fundamental frequency. In addition, the number of spectral peaks in a window is assumed to be 20 at k and those at k+1 and k-1 are 15 and 0, respectively. In this case, the maximum value is 20 at k. Next, $2k$ is assumed as the second harmonic. Assuming that the number of spectral peaks in a window at $2k$, $2k$+1, and $2k$-1 are 15, 10 and 0, respectively, the maximum value is 15 at $2k$. By adding both the maximum values, 20 and 15, the total value of 35 is obtained for k.

For dealing with the variation in the fundamental frequency, ±1 frequencies of the fundamental frequency k are also investigated using the above procedure. In Fig. 5, in the case of k+1, the number of spectral peaks in a window is assumed to be 15 and those at the second harmonic to be $2(k+1)$ and its ±1 frequencies, $2(k+1)+1$ and $2(k+1)-1$, are 0; 0; and 10, respectively. The total number of spectral peaks is 25 at k+1. In the case of k-1, the total value is assumed to be 0. Comparing the total values, k with the maximum total value of 35 is regarded as the true fundamental frequency. Harmonics are sought using this fundamental frequency.

The window size is considered as a grace period for determining the fundamental frequency. In contrast, such a grace period causes a delay in processing and should be minimized; in the proposed method, f sampling periods are grace periods.

4) Gradual change

In ASA, the gradual change comprises two characteristics, namely, "similarity" and "continuity" as shown in Fig. 6 (a) and (b), respectively. Similarity is defined as the connectedness of sound for a short time (e.g., in a phoneme), and the continuity provides a criterion for the connectedness of sound for a long time (e.g., in successive phonemes).

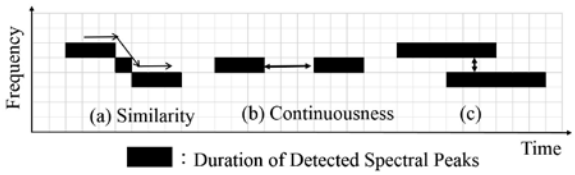


Figure 6. Detection of gradual changes.

a) Detection of similarity

Similarity is based on the phenomenon that the fundamental frequency does not change considerably. The fundamental frequency can be obtained via the detection of harmonic structures. If the detected fundamental frequencies are successive and have the same value, they are regarded identical. However, even if the successive fundamental frequencies are different and have small variation width, they are regarded as identical. When the fundamental frequency is k at time i and that at the previous sampling period i-1 is within k±1, both fundamental frequencies are regarded as connected.

b) Detection of continuity

The continuity is also based on the phenomenon that the characteristic of a sound never changes considerably. The fundamental frequency never changes suddenly in a few phonemes. However, if the phonemes include silent zones, the method for detecting the similarity cannot be used to detect the continuity and another method for detecting continuity is required. The key point is to utilize the information of fundamental frequencies that were included in previously grouped harmonics.

Let us explain the procedure using Fig. 7. When a new harmonic structure with a fundamental frequency k_i is

detected at *i* and there is no fundamental frequency before just one sampling period, it is examined whether harmonic groups are present during the past 1500 sample periods from the current time. The duration corresponds to the number of samples in a phoneme when the sampling frequency is 8 kHz. If there are harmonic groups, the fundamental frequencies during the maximum 1000 sampling periods are averaged in each harmonic group. If the difference d_n between the fundamental frequency k_i of the detected harmonic structure and the averaged frequency k^*_n is within ± 4 and is the smallest, the harmonic structure with the smallest difference is regarded as connected to the detected harmonic structure.

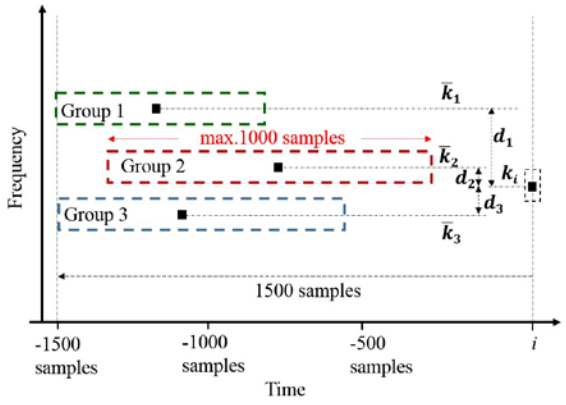


Figure 7. Detection of Continuity.

Moreover, another issue has to be resolved. Two fundamental frequency peaks are simultaneously detected at different frequencies; however, these are originally identical as indicated in Fig. 5 (c). This phenomenon is caused by a variation in the fundamental frequency and the characteristic of the signal level detector, which has a large time constant for a decreasing signal [15]. For instance, it can be assumed that two phonemes are successive; then, the former's fundamental frequency is changed in the latter phoneme. Even after the former phoneme ends, the frequency signal as an output of the signal level detector is continued because it is filtered using a large time constant. Therefore, an originally identical fundamental frequency is detected at two frequencies.

To solve this problem, the onset/offset feature is utilized. Let us explain that using Fig. 8. Actual voiced speeches do not comprise only line spectra; therefore, they comprise main lobes and side lobes. The detected spectral peaks correspond to the main lobes. Preferring to the onset/offset feature of the side lobes, continuity is detected even if an originally identical fundamental frequency is detected at two frequencies as described below.

When a spectral peak of the fundamental frequency ends at (a) in Fig. 8, the similarity and the continuity at this point are investigated first. Even if they are not detected, the presence or absence of another spectral peak is examined within its ±3 frequencies. If there is another spectral peak at (b), the onset/offset features of the fundamental frequency and those of the frequency of another peak are investigated. If the onset features are detected, those spectral peaks are regarded to correspond to an identical fundamental frequency, i.e., the spectral peaks are continued.

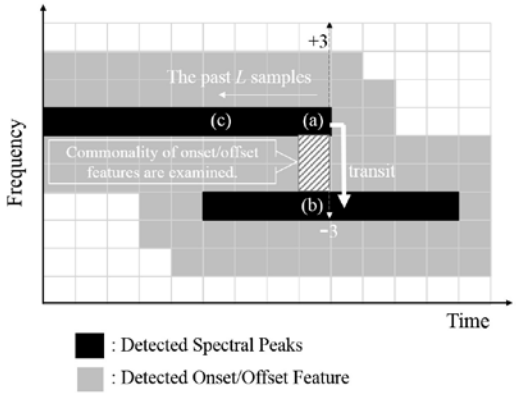


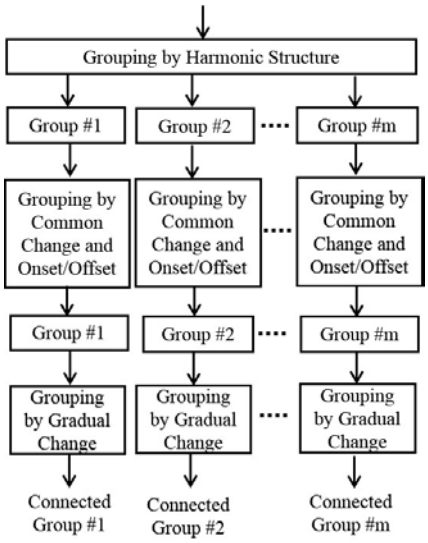
Figure 8. Gradual change detection using the common onset/offset feature.

In contrast, if the detected spectral peaks at different frequencies do not correspond to the identical fundamental frequency but to different fundamental frequencies, the above processing causes the misdetection of continuity. To address this problem, even if the continuity is detected after the above processing, the amplitudes of two frequency signals are compared with each other. However, this comparison is performed not at the end-point of a detected spectral peak (a) but at a point beyond the past *L* sampled point at (c) from the end-point. If the difference between the amplitude levels is less than a quarter of a major one, the two detected spectral peaks at different frequencies are concluded to correspond to an identical fundamental frequency.

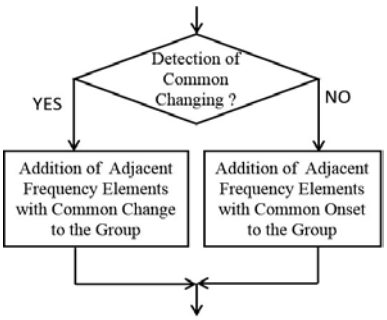
B. Grouping Controller

The procedures in the grouping controller is described in Fig. 9. Harmonic structures are first extracted from spectral peaks based on the method explained in Sect. II-A3 and frequency elements (harmonics) in each harmonic structure are grouped. Next, the common change is examined in each group as described in Sect. II-A2. Concretely, the number of "+1"s, "-1"s, or "0"s of the spectral peaks is investigated in each group. If the number of "+1"s is larger than half of the number of all peaks, the group is regarded as increasing and all spectral peaks are grouped even if some spectral peaks have "-1" or "0." This processing for grouping is performed up to ±3 frequencies of each spectral peak. However, if the number of "+1"s of the frequency element is less than half of the number of all spectral peaks, this processing is aborted. The above processing is also achieved similarly for "-1" and "0." If the number of spectral peaks with the same value is less than half of the number of all peaks, the detected onset/offset is examined described in Sect. II-A1. If the harmonics are labeled as the onset, they are also grouped. This processing is performed at ±3 frequencies of each harmonic. Finally, the similarity and continuity are investigated as described in Sect. II-A4 and

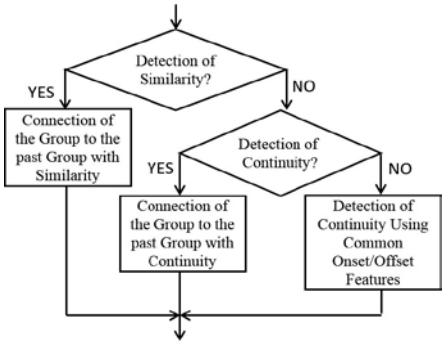
the time-connectedness of grouped harmonics is guaranteed.



(a) Overview



(b) Grouping by common change and onset/offset.



(c) Grouping by gradual change.

Figure 9. Grouping controller.

III. Evaluation in Experiments

A. Visual Evaluation

Fig. 10 (a), (b), and (c) show the waveforms of a mixed signal, male speech, and female speech used for performance evaluation, respectively. The number of samples for MDFT was N=768; therefore, the maximum frequency is N/2-1=383. The upper limit for estimating a fundamental frequency was k=40. The threshold for extracting spectral peaks was the sum of 100 and the twofold mean of the input spectrum. L for comparing the amplitude levels was 100. The moving-window size $2f$ for detecting the fundamental frequency was 101; therefore, a processing delay of 50 sampled periods, i.e., 6.25 ms is necessary. The thresholds for detecting the common change and the common onset/offset are set to

$$2X+100+ (400/k) \text{ for } 1 \leq k< 90,$$
$$2X+50+ (50/ (k-89)) \text{ for } 90 \leq k,$$

where X is the mean value of the amplitude spectrum.

The results are shown in Figs. 10 (d) and (e). The proposed method cannot determine which separated signal corresponds to the original speech. The separated signals (d) and (e) seem to be the original signals (b) and (c), respectively. It is roughly confirmed that speech separation based on ASA can be achieved sequentially using the proposed model. Degradation naturally occurs in the separated signals as the information that can be used in the proposed sequential processing is restricted in contrast with that of the conventional block-processing methods [3]-[11].

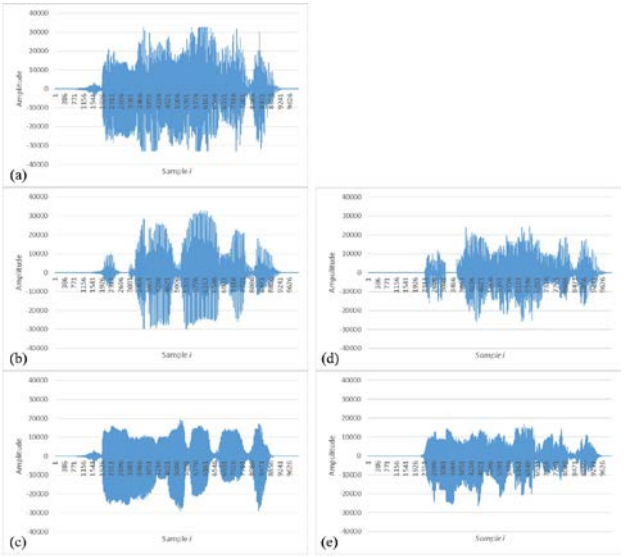


Figure 10. Waveforms (1).

Using the same conditions, another mixed signal was processed. The waveforms are shown in Fig. 11, where a mixed signal (a), male speech (b), female speech (c), and the processed signals (d) and (e). It is confirmed that the settings in the proposed model is robust. The separation performance was conducted using other mixed signals and confirmed that the proposed sequential processing model worked well despite it is simple rule-based processing. As the performance of detectors for four features is improved, the separation performance will be improved. However, there was a fact that some mixed signals cannot be separated well by the proposed model. The example is shown in Fig. 12. The waves which are not included in the original speech (c) are presented in the separated signal in (e). The reason is that the frequencies of a harmonic in one speech were the same as those in the other speech; therefore, the onset/offset features in each speech could not be detected. It is difficult for the

proposed model to separate them since only the phase information is not utilized. This is an apparent limitation of not only the proposed model but also the unitary input models.

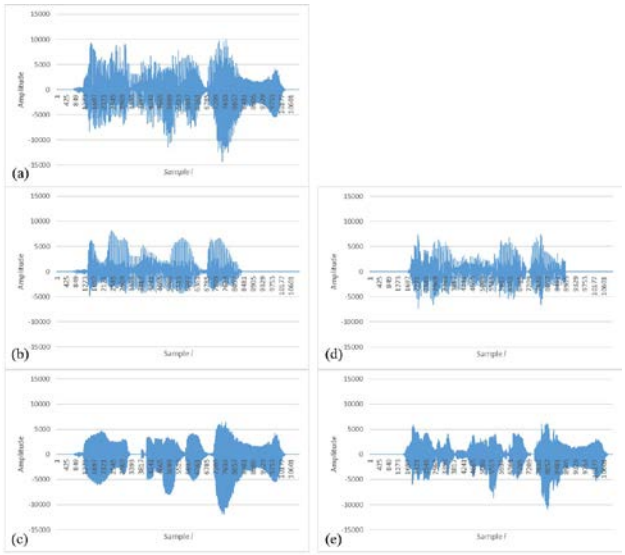


Figure 11. Waveforms (2).

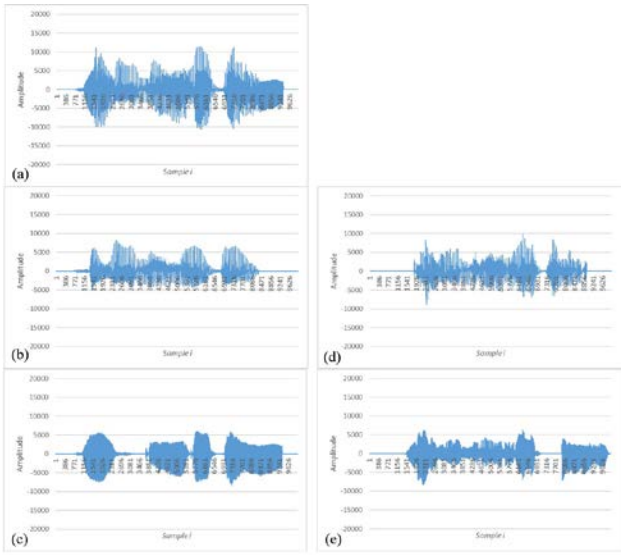


Figure 12. Waveforms (3).

B. *Subjective Evaluation*

In order to evaluate the separation performance, the subjective evaluation using Mean Opinion Score (MOS) was conducted. The number of subjects was fifteen. All were male Japanese students of the authors' university. They were required to rate the separated sounds by five categories, 5: "Excellent," 4: "Good," 3: "Fair," 2: "Poor," and 1: "Bad." With the exception of the mixed signals that could not be separated well by the proposed model, six mixed signals were selected from all processed signals for the subjective evaluation. The results are shown in Fig. 13.

Totally, the MOS was lower than 2 that is ranked as "Poor"; therefore, it is confirmed that the reproducibility in separated signals of the proposed model is low.

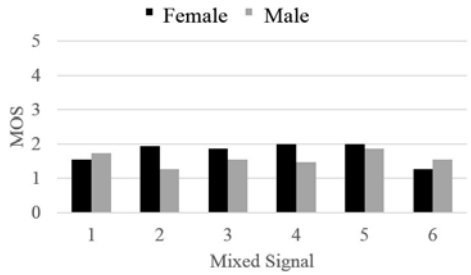


Figure 13. Subjective evaluation using MOS.

On the other hand, the subjective evaluation using the MOS is to evaluate the quality of sounds and not to evaluate the separation performance. Therefore, the authors defined Separation Mean Opinion Score (SMOS) as a new evaluation criterion. In the SMOS, before evaluating separated sounds, the subjects listened to their original sounds. After that, they listened to the separated sounds and evaluated how they were separated by five rating scale as shown in Table 1.

TABLE I. FIVE RATING SCALE OF SMOS

	Score
Completely separated	5
Fairly separated	4
Perceptively separated	3
Not sufficiently separated	2
Completely unseparated	1

The evaluation results are shown in Fig. 14. Total score in SMOS became higher than the MOS; however, it was not sufficient. On the other hand, there was no separated sound which was rated as "completely unseparated". It is confirmed that the speech separation is roughly achieved by even using the proposed model which is based on simple signal processing and rules.

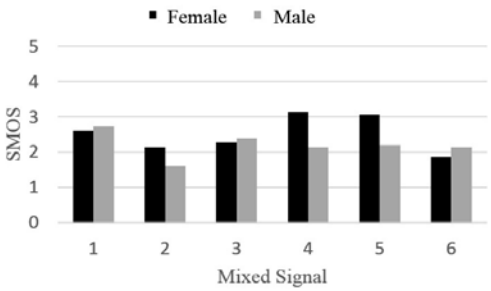


Figure 14. Subjective evaluation using SMOS.

IV. CONCLUSIONS

A unitary input model for the sequential processing of CASA has been proposed. In the conventional studies, the separation performance had been evaluated using only a mixed speech. In this study, the robustness of the settings for the proposed model was visually evaluated in the results using other mixed speeches. In addition, the separation performance was subjectively evaluated using SMOS as a new evaluation criterion.

The results confirmed that the sequential processing of CASA was feasible even if the proposed model was based on simple signal processing and rules; however, speech separation could not be completely achieved using the proposed model.

A future study must involve the verification of the separation performance using various mixed signals and conditions.

REFERENCES

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

[2] D. Wang and G. J. Brown, *Computational Auditory Scene Analysis*, USA: IEEE Press 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," *EURASIP Journal on Audio, Speech, and Music Processing*, 2013.

[12] I. Nakanishi and J. Hanada, "A sequential processing model for speech separation based on auditory scene analysis," in *Proc. 2015 IEEE International Symposium on Intelligent Signal Processing and Communication Systems*, pp. 124-128, 2015.

[13] M. Ichikawa, N. Sasaoka, and I. Nakanishi, "A single input model for sequential processing of speech separation," in *Proc. 2018 IEEE International Conference on Information Communication and Signal Processing*, pp. 108-112, 2018.

[14] 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.

[15] Y. Minato and I. Nakanishi, "Noise reduction system using signal and noise level detectors in frequency domain," in *Proc. 2008 IEEE International Symposium on Intelligent Signal Processing and Communication Systems*, pp. 180-183, 2009.

[16] X. Serra and J. O. Smith, "Spectral modeling synthesis," in *Proc. International Computer and Music Conference*, pp. 281-284, 1989.

Isao Nakanishi was born in Osaka, Japan in 27 De c. 1961. He received his B. E., M. E., and Dr. E. degrees in Electrical Engineering from Osaka Prefecture University, Japan in 1984, 1986, and 1997, respectively. He is now a professor in the Faculty of Engineering, Tottori University, Japan. His research interests are in digital signal processing and biometrics. He is a senior member of the IEEE, the Institute of Electronics, Information and Communication Engineers (IEICE) and a member of the Information Processing Society of Japan (IPSJ).

Motohiro Ichikawa was born in Okayama, Japan in 25 Oct. 1987. He received his B. E. and M. E. degrees in Electrical and Electronics Engineering from Tottori University, Japan in 2016 and 2018, respectively. His research interests are in speech signal processing.

Naoto Sasaoka received his B. E., M. E., and Dr. E. degrees in Electrical and Electronics Engineering from Tottori University, Japan in 2002, 2004, and 2006, respectively. He is now an associate professor in the Faculty of Engineering, Tottori University, Japan. His research interests are in speech signal processing and digital communication. He is a member of the IEEE, and the Institute of Electronics, Information and Communication Engineers (IEICE).