International Journal For Recent Developments in Science & Technology
2581-4575
Crossref
A Peer Reviewed Research Journal

ASA BASED UNITARY INPUT MODEL FOR SEQUENTIAL PROCESSING OF SPEECH SEPARATION

[1]VELPULA VIJAY KUMAR, [2]DUBASI KIRTANA, [3]POTAPARTHINI KIRANMAYEE
[1,2,3]ASSISTANT PROFESSOR, ST.MARTIN'S ENGINEERING COLLEGE, KOMPALLY, SECUNDERABAD

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 beenproposed. 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 theregrouped 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 timefrequency 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

International Journal For Recent Developments in Science & Technology
2581-4575
Crossref
A Peer Reviewed Research Journal

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 thesequential 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.

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

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

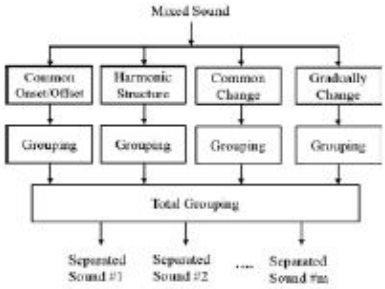

Figure 1. Concept model of ASA.

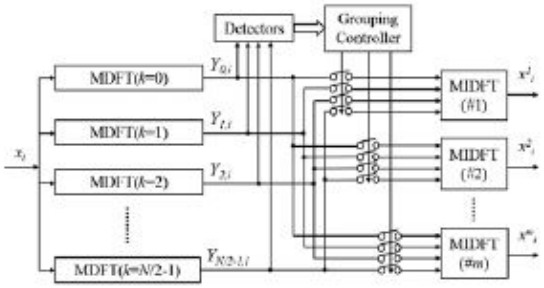

Figure 2. Sequential processing of ASA using a MDFT pair.

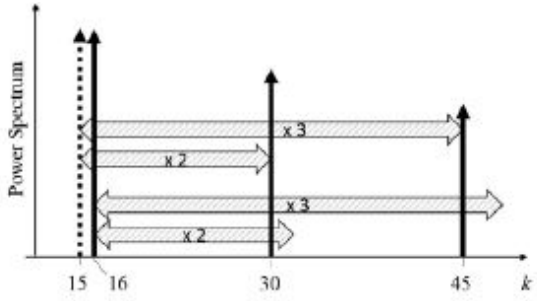

Figure 3. An example of variation of harmonics.

International Journal For Recent Developments in Science & Technology
2581-4575
Crossref
A Peer Reviewed Research Journal

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).

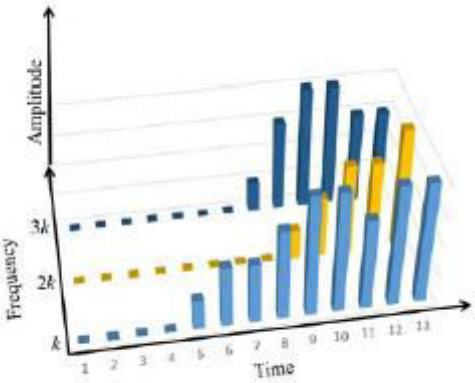

Figure 4. Time variation of the detected spectral peaks.

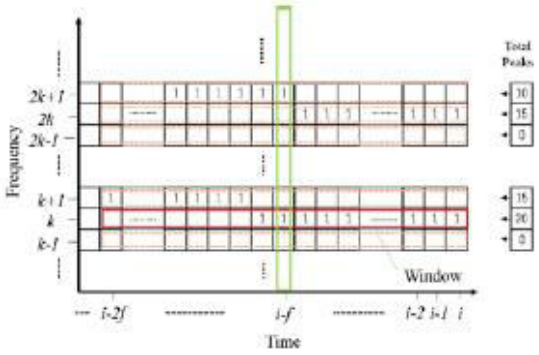

Figure 5. Detection of spectral peaks using a moving window.

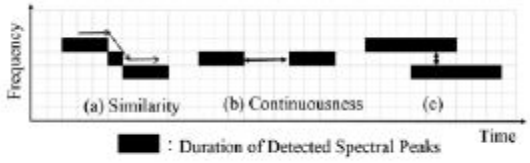

Figure 6. Detection of gradual changes.

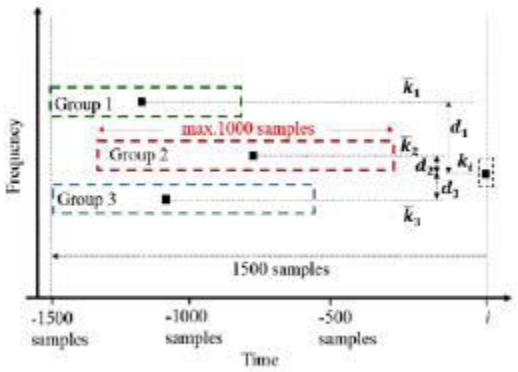

Figure 7. Detection of Continuity.

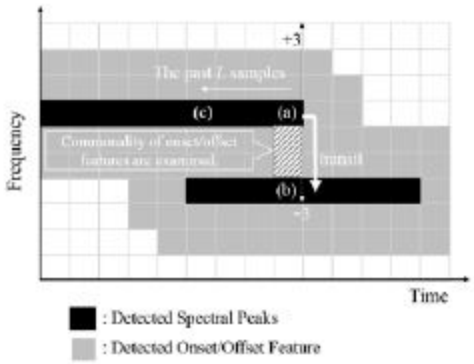

Figure 8. Gradual change detection using the common onset/offset feature.

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

International Journal For Recent Developments in Science & Technology
2581-4575
Crossref
A Peer Reviewed Research Journal

frequencies are concluded to correspond to an identical fundamental frequency.

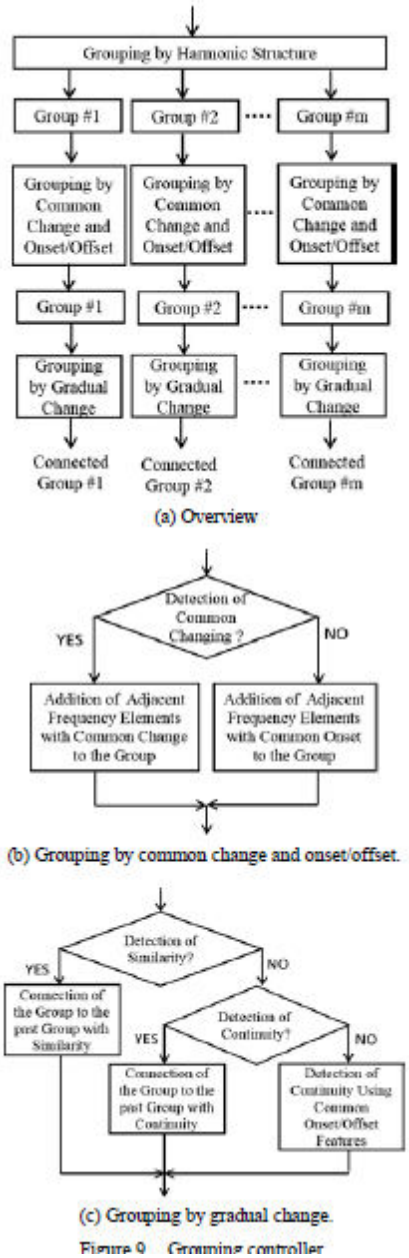

Figure 9 Grouping controller

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

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

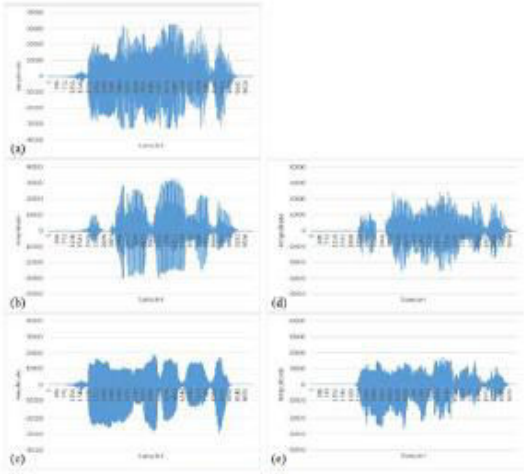

Figure 10. Waveforms (1).

International Journal For Recent Developments in Science & Technology
2581-4575
Crossref
A Peer Reviewed Research Journal

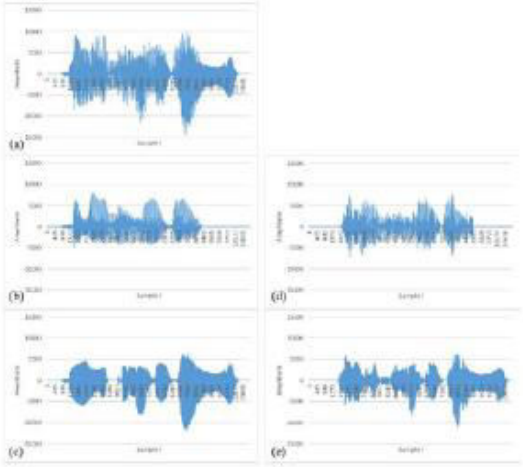

Figure 11. Waveforms (2).

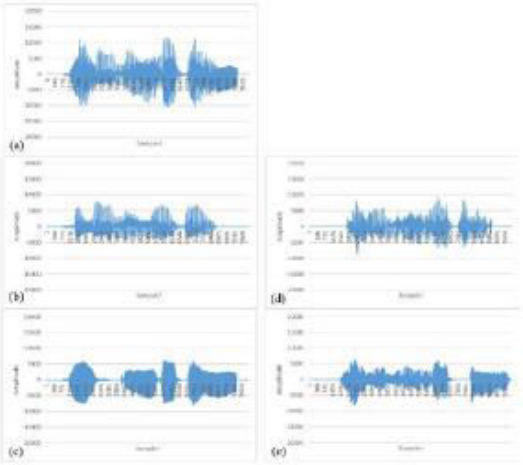

Figure 12. Waveforms (3).

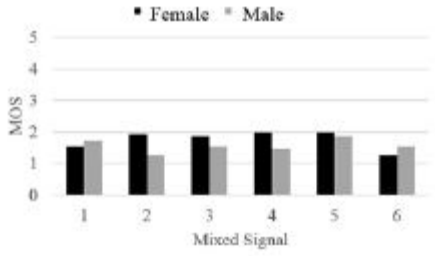

Figure 13. Subjective evaluation using MOS.

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 usingthe proposed model which is based on simple signal processing and rules.

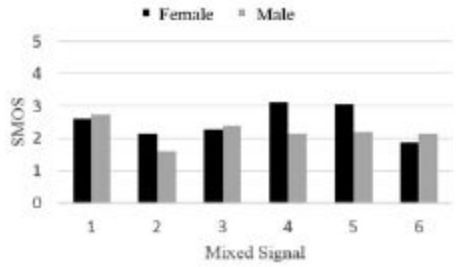

Figure 14. Subjective evaluation using SMOS.

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

International Journal For Recent Developments in Science & Technology
2581-4575
Crossref
A Peer Reviewed Research Journal

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, "Switchedcapacitor 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.