Segregation of vowel in background noise using the model of segregating two acoustic sources based on auditory scene analysis

Masashi Unoki and Masato Akagi

Japan Advanced Institute of Science and Technology
1-1 Asahidai, Tatsunokuchi, Nomi, Ishikawa, 923-1292 Japan
Email: {unoki,akagi}@jaist.ac.jp

ABSTRACT

This paper proposes an auditory sound segregation model based on auditory scene analysis. It solves the problem of segregating two acoustic sources by using constraints related to the heuristic regularities proposed by Bregman and by making an improvement to our previously proposed model. The improvement is to reconsider constraints on the continuity of instantaneous phases as well as constraints on the continuity of instantaneous amplitudes and fundamental frequencies in order to segregate the desired signal from a noisy signal precisely even in waveforms. Simulations performed to segregate a real vowel from a noisy vowel and to compare the results of using all or only some constraints showed that our improved model can segregate real speech precisely even in waveforms using all the constraints related to the four regularities, and that the absence of some constraints reduces the segregation accuracy.

1. INTRODUCTION

Bregman has reported that the human auditory system uses four psychoacoustically heuristic regularities related to acoustic events to solve the problem of auditory scene analysis [1]. If an auditory sound segregation model was constructed using these regularities, it would be applicable not only to a preprocessor for robust speech recognition systems but also to various types of signal processing.

Some ASA-based segregation models already exist. There are two main types of models, based on either bottom-up [2] or top-down processes [3, 4]. All these models use some of the four regularities, and the amplitude (or power) spectrum as the acoustic feature. Thus they cannot completely segregate the desired signal from noisy signal when the signal and noise exist in the same frequency region.

In contrast, we have discussed the need to use not only the amplitude spectrum but also the phase spectrum in order to completely extract the desired signal from a noisy signal, thus addressing the problem of segregating two acoustic sources [5]. This problem is defined as follows [5, 7]. First, only the mixed signal $f(t)$, where $f(t) = f_1(t) + f_2(t)$, can be observed. Next, $f(t)$ is decomposed into its frequency components by a filterbank (K channels). The output of the k-th channel $X_k(t)$ is represented by

$$X_k(t) = S_k(t)\exp(j\omega_k t + j\phi_k(t)). \quad (1)$$

Here, if the outputs of the k-th channel, which correspond to $f_1(t)$ and $f_2(t)$, are assumed to be $A_k(t)\exp(j\omega_k t + j\theta_{1k}(t))$ and $B_k(t)\exp(j\omega_k t + j\theta_{2k}(t))$, then instantaneous amplitudes $A_k(t)$ and $B_k(t)$ can be determined by

$$A_k(t) = S_k(t)\sin(\theta_{2k}(t) - \phi_k(t))/\sin\theta_k(t), \quad (2)$$
$$B_k(t) = S_k(t)\sin(\phi_k(t) - \theta_{1k}(t))/\sin\theta_k(t), \quad (3)$$

where $\theta_k(t) = \theta_{2k}(t) - \theta_{1k}(t)$, $\theta_k(t) \neq n\pi, n \in \mathbf{Z}$, and ω_k is the center frequency of the k-th channel. Instantaneous phases $\theta_{1k}(t)$ and $\theta_{2k}(t)$ can be determined by

$$\theta_{1k}(t) = -\arctan\left(\frac{Y_k(t)\cos\phi_k(t) - \sin\phi_k(t)}{Y_k(t)\sin\phi_k(t) + \cos\phi_k(t)}\right) + \arcsin\left(\frac{A_k(t)Y_k(t)}{S_k(t)\sqrt{Y_k(t)^2+1}}\right), \quad (4)$$

$$\theta_{2k}(t) = -\arctan\left(\frac{Y_k(t)\cos\phi_k(t) + \sin\phi_k(t)}{Y_k(t)\sin\phi_k(t) - \cos\phi_k(t)}\right) + \arcsin\left(-\frac{B_k(t)Y_k(t)}{S_k(t)\sqrt{Y_k(t)^2+1}}\right), \quad (5)$$

where $Y_k(t) = \sqrt{(2A_k(t)B_k(t))^2 - Z_k(t)^2}/Z_k(t)$ and $Z_k(t) = S_k(t)^2 - A_k(t)^2 - B_k(t)^2$. Hence, $f_1(t)$ and $f_2(t)$ can be reconstructed by using the determined pair of $[A_k(t)$ and $\theta_{1k}(t)]$ and the determined pair of $[B_k(t)$ and $\theta_{2k}(t)]$ for all channels. However, $A_k(t)$, $B_k(t)$, $\theta_{1k}(t)$, and $\theta_{2k}(t)$ cannot be uniquely determined without some constraints as is easily understood from the above equations. Therefore, this problem is an ill-inverse problem.

To solve this problem, we have proposed a basic method of solving it using constraints related to the four regularities [5] and the improved method [6]. However, the former cannot deal with the variation of the fundamental frequency, although it can segregate the synthesized signal from the noise-added signal. Additionally, for the later, it is difficult to completely determine the phases, although it can be segregated vowel from noisy vowel precisely at certain amplitudes by constraining the continuity of the instantaneous amplitudes and fundamental frequencies.

This paper proposes a new sound segregation method to deal with real speech and noise precisely even in wave-

Table 1: Constraints corresponding to Bregman's psychoacoustical heuristic regularities.

Regularity (Bregman, 1993)	Constraint (Unoki and Akagi, 1999)	
(i) common onset/offset	synchronous of onset/offset	$\lvert T_{\rm S} - T_{k,{\rm on}}\rvert \leq \Delta T_{\rm S}, \lvert T_{\rm E} - T_{k,{\rm off}}\rvert \leq \Delta T_{\rm E}$
(ii) gradualness of change	piecewise-differentiable polynomial approximation	$dA_k(t)/dt = C_{k,R}(t), d\theta_{1k}(t)dt = D_{k,R}(t)$ $dF_0(t)/dt = E_{0,R}(t)$
(smoothness)	(spline interpolation)	$\sigma_A = \int_{t_a}^{t_b}[A_k^{(R+1)}(t)]^2 dt \Rightarrow \min$ $\sigma_\theta = \int_{t_a}^{t_b}[\theta_{1k}^{(R+1)}(t)]^2 dt \Rightarrow \min$ (**new**)
(iii) harmonicity	multiples of the fundamental frequency	$n \times F_0(t), \quad n = 1, 2, \cdots, N_{F_0}$
(iv) changes occurring in the acoustic event	correlation between the instantaneous amplitudes	$\frac{A_k(t)}{\lVert A_k(t)\rVert} \approx \frac{A_\ell(t)}{\lVert A_\ell(t)\rVert}, \quad k \neq \ell$

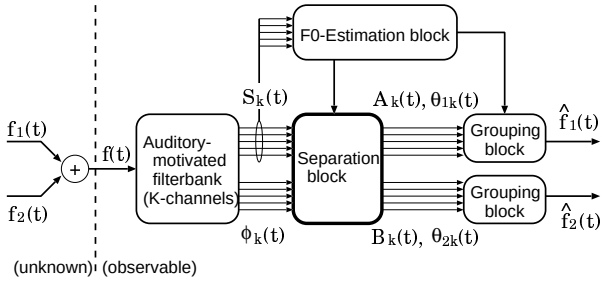

Figure 1: Auditory sound segregation model.

forms, by using constraints on the continuity of instantaneous phases as well as constraints on the continuity of instantaneous amplitudes and fundamental frequencies.

2. AUDITORY SOUND SEGREGATION MODEL

In this paper, it is assumed that the desired signal $f_1(t)$ is a harmonic complex tone, where $F_0(t)$ is the fundamental frequency. The proposed model segregates the desired signal from the mixed signal by constraining the temporal differentiation of $A_k(t)$, $\theta_{1k}(t)$, and $F_0(t)$.

The proposed model is composed of four blocks: an auditory-motivated filterbank, an F_0 estimation block, a separation block, and a grouping block, as shown in Fig. 1. Constraints used in this model are shown in Table 1.

2.1. Auditory-motivated filterbank

The auditory-motivated filterbank decomposes the observed signal $f(t)$ into complex spectra $X_k(t)$. This filterbank is implemented as a constant Q gammatone filterbank, constructed with $K = 128$, a bandwidth of 60–6000 Hz, and a sampling frequency of 20 kHz [5]. $S_k(t)$ and $\phi_k(t)$ are determined by using the amplitude and phase spectra defined by the wavelet transform [5].

2.2. F_0 estimation block

The F_0 estimation block determines the fundamental frequency of $f_1(t)$. This block is implemented as the Comb filtering on an amplitude spectrogram $S_k(t)$s [6]. Since the number of channels in the $X_k(t)$ is finite, the estimated $F_0(t)$ takes a discrete value. In addition, the fluctuation of $F_0(t)$ behaves like a stair shape and the temporal differentiation of $F_0(t)$ is zero at any segment. Therefore, this paper assumes that $E_{0,R}(t) = 0$ in Table 1 (ii) for a segment. Let the length of the above segment be $T_h - T_{h-1}$, where T_h is the continuous point of $F_0(t)$.

2.3. Grouping block

The grouping block determines the concurrent time-frequency region of the desired signal using constraints (i) and (ii) in Table 1, and then reconstructs the segregated instantaneous amplitude and phase using the inverse wavelet transform [7]. $\hat{f}_1(t)$ and $\hat{f}_2(t)$ are the reconstructed $f_1(t)$ and $f_2(t)$. Constraint (i) is implemented by comparing the onset/offset $(T_{k,{\rm on}}, T_{k,{\rm off}})$ of $X_k(t)$ with the onset/offset $(T_{\rm S}, T_{\rm E})$ of $X_{\hat{\ell}}(t)$ corresponding to $F_0(t)$, where $\Delta T_{\rm S} = 25$ msec and $\Delta T_{\rm E} = 50$ msec [7]. Constraint (iii) is implemented by determining the channel number corresponding to the integer multiples of $F_0(t)$.

2.4. Separation block

The separation block determines $A_k(t)$, $B_k(t)$, $\theta_{1k}(t)$, and $\theta_{2k}(t)$ from $S_k(t)$ and $\phi_k(t)$ using constraints (ii) and (iv) in the determined concurrent time-frequency region. In this paper, the improvement of the auditory sound segregation model is to reconsider the constraints on the continuity of $\theta_{1k}(t)$ as well as the constraints on the continuity of $A_k(t)$ and $F_0(t)$. Constraint (ii) is implemented such that $C_{k,R}(t)$ and $D_{k,R}(t)$ are linear ($R = 1$) polynomials, in order to reduce the computational cost of estimating $C_{k,R}(t)$ and $D_{k,R}(t)$. In this assumption, $A_k(t)$ and $\theta_{1k}(t)$, which can be allowed to undergo a temporal change in region, constrain the second-order polynomials ($A_k(t) = \int C_{k,1}(t)dt + C'_{k,0}$ and $\theta_{1k}(t) = \int D_{k,1}(t) + D'_{k,0}$). Then, substituting $dA_k(t)/dt = C_{k,R}(t)$ into Eq. (2), we get the linear differential equation of the input phase difference $\theta_k(t) = \theta_{2k}(t) - \theta_{1k}(t)$. By solving this equation, a general solution is determined by

$$\theta_k(t) = \arctan\left(\frac{S_k(t)\sin(\phi_k(t) - \theta_{1k}(t))}{S_k(t)\cos(\phi_k(t) - \theta_{1k}(t)) + C_k(t)}\right), \tag{6}$$

where $C_k(t) = -\int C_{k,R}(t)dt - C_{k,0} = -A_k(t)$ [7].

The signal flow of the separation block is shown in Fig. 2. In the segment $T_h - T_{h-1}$ which can be determined by $E_{0,R}(t) = 0$, $A_k(t)$, $B_k(t)$, $\theta_{1k}(t)$, and $\theta_{2k}(t)$ are determined by the following steps. First, the estimated

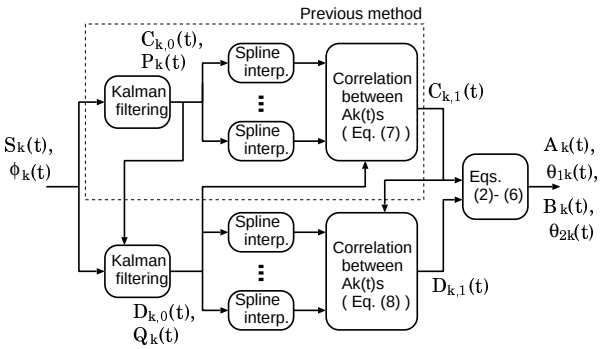

Figure 2: Signal processing of a separation block.

regions, $\hat{C}_{k,0}(t) - P_k(t) \leq C_{k,1}(t) \leq \hat{C}_{k,0}(t) + P_k(t)$ and $\hat{D}_{k,0}(t) - Q_k(t) \leq D_{k,1}(t) \leq \hat{D}_{k,0}(t) + Q_k(t)$, are determined by using the Kalman filter, where $\hat{C}_{k,0}(t)$ and $\hat{D}_{k,0}(t)$ are the estimated values and $P_k(t)$ and $Q_k(t)$ are the estimated errors. Next, the candidates of $C_{k,1}(t)$ at any $D_{k,1}(t)$ are selected by using the spline interpolation in the estimated error region. Then, $\hat{C}_{k,1}(t)$ is determined by using

$$\hat{C}_{k,1} = \underset{\hat{C}_{k,0}-P_k \leq C_{k,1} \leq \hat{C}_{k,0}+P_k}{\arg\max} \frac{< \hat{A}_k, \hat{\hat{A}}_k >}{||\hat{A}_k|| ||\hat{\hat{A}}_k||}, \qquad (7)$$

where $\hat{A}_k(t)$ is obtained by the spline interpolation and $\hat{\hat{A}}_k(t)$ is determined in across-channel that satisfies constraint (iii). Finally, $\hat{D}_{k,1}(t)$ is determined by using

$$\hat{D}_{k,1} = \underset{\hat{D}_{k,0}-Q_k \leq D_{k,1} \leq \hat{D}_{k,0}+Q_k}{\arg\max} \frac{< \hat{A}_k, \hat{\hat{A}}_k >}{||\hat{A}_k|| ||\hat{\hat{A}}_k||}. \qquad (8)$$

Since $\theta_{1k}(t)$ and $\theta_k(t)$ are determined from $\hat{D}_{k,1}(t)$ and $\hat{C}_{k,1}(t)$, $A_k(t)$, $B_k(t)$, and $\theta_{2k}(t)$ can be determined from Eq. (2), Eq. (3), and $\theta_{2k}(t) = \theta_k(t) + \theta_{1k}(t)$, respectively.

3. SIMULATIONS

To show that the proposed method can segregate the desired signal $f_1(t)$ from a noisy signal $f(t)$ precisely even in waveforms, we ran three simulations using the following signals:
(a) noisy synthesized AM–FM harmonic complex tone [6];
(b) noisy real vowel (/a/, /i/, /u/, /e/, /o/); and
(c) noisy real continuous vowel (/aoi/),
where noise was a pink noise and the SNRs of noisy signals were from 5 to 20 dB in 5-dB steps. The speech signals were the Japanese vowels of four speakers (two males and two females) in the ATR-database [8].

We used segregation accuracy to evaluate the segregation performance of the proposed method, as defined by

$$10 \log_{10} \frac{\int_0^T f_1(t)^2 dt}{\int_0^T (f_1(t) - \hat{f}_1(t))^2 dt} \quad \text{(dB)}. \qquad (9)$$

Next, to show the advantages of the constraints in Table 1, we compared the performance of our method in the following three conditions:
(1) extract the harmonics using the Comb filter and predict $A_k(t)$ and $\theta_{1k}(t)$ using the Kalman filtering;
(2) extract the harmonics using the Comb filter; and
(3) do nothing.
Here, condition 1 corresponds to the smoothness of constraint (ii) being omitted; condition 2 corresponds to constraints (ii) and (iii) being omitted; and condition 3 corresponds to no constraints being applied at all.

3.1. Overview of signal processing

An overview of signal processing of the proposed model is shown in Fig. 3. First, noisy vowel /a/ $f(t)$ shown in Fig. 3 A (the SNR of $f(t)$ is 10 dB) for simulation 2 is decomposed into $S_k(t)$ and $\phi_k(t)$ as shown in Fig. 3 B and C, respectively. Next, $F_0(t)$ is estimated as shown in Fig. 3 D. The concurrent time-frequency region of the desired signal $f_1(t)$ is determined using constraints (i) and (iii) as shown in Fig. 3 E and F. Finally, the instantaneous amplitudes and the instantaneous phases of two signals are determined from $S_k(t)$ and $\phi_k(t)$ using constraints (ii) and (iv). The determined $A_k(t)$ and $\theta_{1k}(t)$ are shown in Fig. 3 H and I, respectively. The segregated signal $\hat{f}_1(t)$ is shown in Fig. 3 J.

3.2. Results and considerations

The segregation accuracy of the three simulations and the four comparisons is shown in Fig. 4. In this figure, the bar height shows the mean of segregation accuracy and the error bar shows the standard deviation of segregation accuracy. The results show that the segregation accuracy using the proposed model was better than that using the other three methods. These results show that the proposed model can segregate the desired vowel from a noisy vowel precisely even in waveforms. In addition, the result of the comparison between the proposed model and (2) shows that the simultaneous signals can be precisely segregated using the instantaneous amplitude and phase. As a result of comparison between the proposed method and (3), improvements in segregation accuracies at the SNR of 5 dB for simulations 1, 2, and 3 are about 10 dB, about 9 dB, and about 7 dB, respectively.

4. CONCLUSIONS

We have proposed a new method of extracting the desired speech from noisy speech precisely even in waveforms by using constraints on the continuity of the instantaneous phases as well as constraints on the continuity of the instantaneous amplitudes and the fundamental frequency.

To show that the proposed model can extract real speech from noisy speech precisely even in waveforms, we demonstrated one evaluation and three simulations of segregating two acoustic sources. The result of evaluation showed

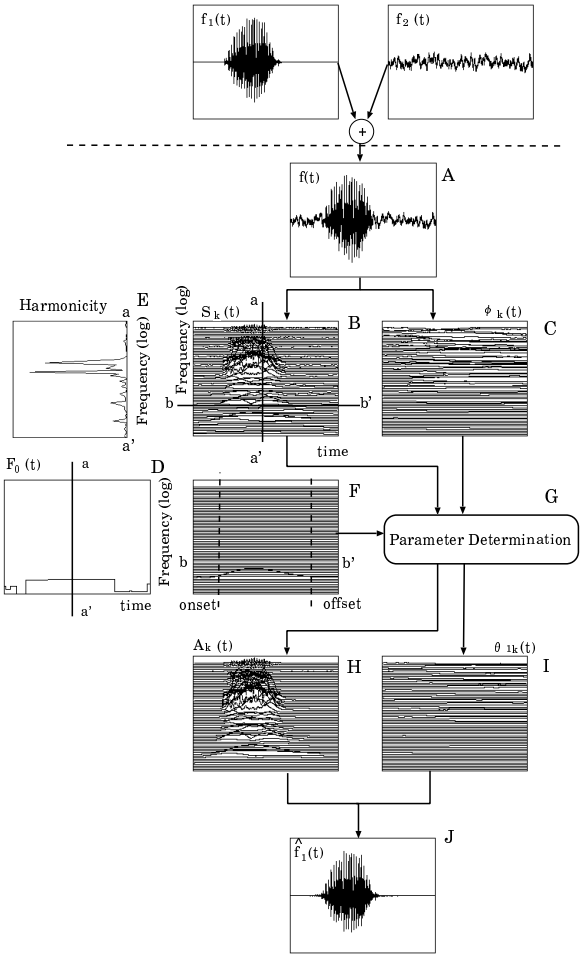

Figure 3: Overview of signal processing of the proposed model.

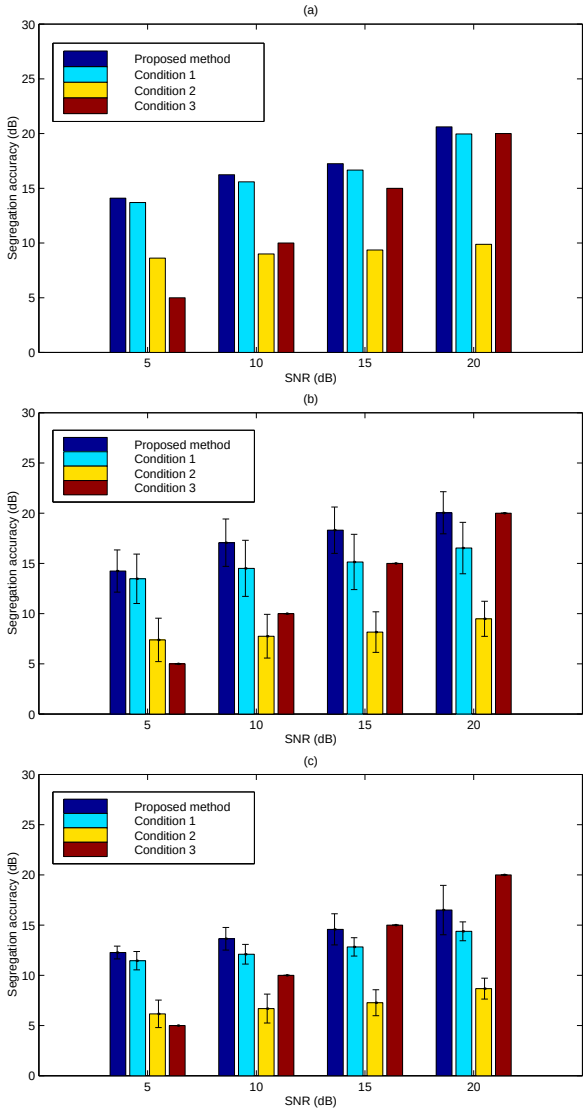

Figure 4: Segregation accuracies for simulations. (a) AM–FM complex tone, (b) vowel, (c) continuous vowel.

that all constraints related to the four regularities are useful in order to segregate the desired vowel from noisy vowel. The results of the three simulations showed that the proposed method can segregate the desired vowel from noisy vowel precisely even in waveforms. It was also shown that the proposed method can precisely segregate the desired signal from the simultaneous signals using the instantaneous amplitude and phase.

5. ACKNOWLEDGMENTS

This work was supported by Grant-in-Aid for Science research from the Ministry of Education (Research Fellowships of the Japan Society for the Promotion of Science for Young Scientists and No. 10680374) and by CREST.

6. REFERENCES

1. Bregman, A.S. “Auditory Scene Analysis: hearing in complex environments,” in Thinking in Sounds, pp. 10–36, Oxford University Press, New York, 1993.
2. Cooke, M. P. and Brown, G.J., “Computational auditory scene analysis : Exploiting principles of perceived continuity,” Speech Communication, vol. 13, pp. 391-399, Dec. 1993.
3. Ellis, D. P. W. “Prediction-driven computational auditory scene analysis,” Ph.D. thesis, MIT Media Lab., 1996.
4. Nakatani, T., Okuno, H. G., and Kawabata, T. “Unified Architecture for Auditory Scene Analysis and Spoken Language Processing,” In Proc. of ICSLP ’94, 24, 3, 1994.
5. Unoki, M. and Akagi, M. “A Method of Signal Extraction from Noisy Signal,” In Proc. EuroSpeech’97, vol. 5, pp. 2583-2586, Sept. 1997.
6. Unoki, M. and Akagi, M. “Signal Extraction from Noisy Signal based on Auditory Scene Analysis,” In Proc. ICSLP’98, vol. 4, pp. 1515–1518, Dec. 1998.
7. Unoki, M. and Akagi, M. “Signal Extraction from Noisy Signal based on Auditory Scene Analysis,” Speech Communication, vol. 27, no. 3 , pp. 261–279, April. 1999.
8. Takeda, K., Sagisaka, Y., Katagiri, S., Abe, M., and Kuwabara, H. Speech Database User’s Manual, ATR Technical Report TR-I-0028, 1988.