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

\[
X_k(t) = S_k(t) \exp(j\omega_k t + j\phi_k(t)).
\]

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

\[
A_k(t) = S_k(t) \frac{Y_k(t) \cos \phi_k(t) - \sin \phi_k(t)}{(S_k(t) \sqrt{Y_k(t)^2 + 1})},
\]

(4)

\[
B_k(t) = S_k(t) \frac{Y_k(t) \cos \phi_k(t) + \sin \phi_k(t)}{(S_k(t) \sqrt{Y_k(t)^2 + 1})},
\]

(5)

where \( Y_k(t) = \sqrt{2A_k(t)B_k(t)^2 - Z_k(t)^2}/Z_k(t) \)

\[
Z_k(t) = S_k(t)^2 - A_k(t)^2 - B_k(t)^2.
\]

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-posed 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 latter, it is difficult to completely determine the phases, although it can be segregated 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 waveforms, by using constraints on the continuity of instantaneous phases as well as constraints on the continuity of instantaneous amplitudes and fundamental frequencies.
Table 1: Constraints corresponding to Bregman’s psychoacoustical heuristic regularities.

<table>
<thead>
<tr>
<th>Regularity (Bregman, 1993)</th>
<th>Constraint (Ueno and Akagi, 1999)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(i) common onset/offset</td>
<td>synchronous of onset/offset</td>
</tr>
<tr>
<td>(ii) gradualness of change</td>
<td>piecewise-differentiable polynomial approximation (smoothness)</td>
</tr>
<tr>
<td>(iii) harmonicity</td>
<td>multiples of the fundamental frequency</td>
</tr>
<tr>
<td>(iv) changes occurring in</td>
<td>correlation between the instantaneous amplitudes</td>
</tr>
<tr>
<td>the acoustic event</td>
<td></td>
</tr>
</tbody>
</table>

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 $T_b - T_{b-1}$, where $T_b$ 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 (iii) 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 $f_2(t)$.

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.

Figure 1: Auditory sound segregation model.

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.
Figure 2: Signal processing of a separation block.

Next, to show the advantages of the constraints of Table 1, we compared the performance of our method in the following three conditions:

1. extract the harmonics using the Comb filter and predict $A_{k,h}(t)$ and $\theta_{k}(t)$ 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 $\theta_k(t)$ are shown in Fig. 3 H and I, respectively. The segregated signal $f_1(t)$ 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 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