SIGNAL EXTRACTION FROM NOISY SIGNAL 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
E-mail: {unoki, akagi}@jaist.ac.jp

ABSTRACT

This paper proposes a method of extracting the desired signal from a noisy signal. This method solves the problem of segregating two acoustic sources by using constraints related to the four regularities proposed by Bregman and by making two improvements to our previously proposed method. One is to incorporate a method of estimating the fundamental frequency using the Comb filtering on the filterbank. The other is to reconsider the constraints on the separation block, which constrain the instantaneous amplitude, input phase, and fundamental frequency of the desired signal. Simulations performed to segregate a vowel from a noisy vowel and to compare the results of using all or only some constraints showed that our improved method can segregate real speech precisely using all the constraints related to the four regularities and that the absence some constraints reduces the accuracy.

1. INTRODUCTION

Bregman reported that the human auditory system uses four psychoacoustically heuristic regularities related to acoustic events, to solve the problem of Auditory Scene Analysis (ASA) [1]. If a segregation model was constructed using constraints related to these heuristic 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, 6]. All these models use some of the four regularities, and the amplitude (or power) spectrum as the acoustic feature. Thus they cannot completely extract the desired signal from a 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, addressing the problem of segregating two acoustic sources [8]. This problem is defined as follows [8]. 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 (the number of channels is K). The output of the k-th channel $X_k(t)$ is represented by

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

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

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

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. Here, $\theta_{1k}(t)$ and $\theta_{2k}(t)$ are the instantaneous input phases of $f_1(t)$ and $f_2(t)$, respectively. Finally, $f_1(t)$ and $f_2(t)$ can be reconstructed by using the determined $[A_k(t), \theta_{1k}(t)]$, and $[B_k(t), \theta_{2k}(t)]$ for all channels.

This problem is an ill-inverse problem because there are currently no equations for determining the two instantaneous phases. Therefore, we have proposed a method of solving this problem using constraints related to the four regularities [8]. It was assumed that the fundamental frequency was constant and known, and that $\theta_{1k}(t) = 0$, although this method could extract the synthesized vowel from a noisy synthesized vowel with high accuracy. Here, $\theta_{1k}(t) = 0$ means that each frequency of the signal component that passed through the channel coincides with the center frequency of each channel. Therefore, it is difficult to extract real speech from noisy speech using this method because the fundamental frequency of speech fluctuates, and multiples of the fundamental frequency cannot coincide with the center frequencies of the channels.

This paper proposes a new method for extracting real speech from noisy speech by (1) incorporating of a method of estimating the fundamental frequency and (2) reconsidering the constraint of $\theta_{1k}(t)$.

2. AUDITORY SEGREGATION MODEL

The proposed method is implemented by an auditory segregation model. This model is composed of four blocks: an auditory-motivated filterbank, an $F_0(t)$ estimation block, a separation block, and a grouping block, as shown in Fig. 1. In this paper, the $F_0(t)$ estimation block is incorporated into the previous model [8] and the separation block

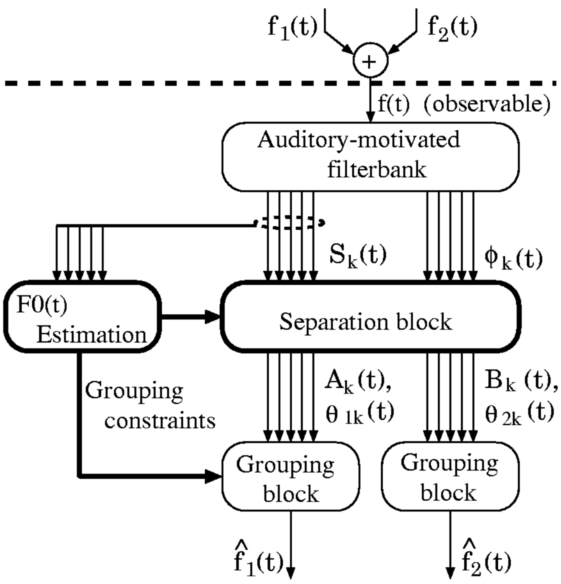

Figure 1: Auditory segregation model.

is improved as shown by the bold boxes in Fig. 1.

2.1. Overview of the proposed model

First, the observed signal $f(t)$ is decomposed into $S_k(t)$ and $\phi_k(t)$ using an auditory-motivated filterbank. This filterbank is implemented as a constant Q gammatone filterbank, constructed with $K = 128$, bandwidth of 60–6000 Hz, and sampling frequency of 20 kHz [8]. Next, the fundamental frequency $F_0(t)$ of the desired signal is determined using an amplitude spectrogram $S_k(t)$s (see Sec. 3.1). Then, the concurrent time-frequency region of the desired signal is determined using constraints (i) and (iii) [8]. In the determined concurrent time-frequency region, $A_k(t)$ and $B_k(t)$ are determined from $S_k(t)$, $\phi_k(t)$, $\theta_{1k}(t)$, and $\theta_{2k}(t)$. $S_k(t)$ and $\phi_k(t)$ are determined by using the amplitude and phase spectra defined by the wavelet transform [8]. $\theta_{1k}(t)$ and $\theta_{2k}(t)$ are determined using constraints (ii) and (iv) (see Sec. 2.2 and 3.2). Finally, $f_1(t)$ and $f_2(t)$ are determined from Eqs. (2) and (3), respectively.

2.2. Assumptions and constraints

In this paper, it is assumed that the desired signal $f_1(t)$ is a harmonic tone, consisting of the fundamental frequency $F_0(t)$ and the harmonic components, which are multiplies of $F_0(t)$. The proposed model segregates the desired signal from the mixed signal by constraining the temporal differentiation of the instantaneous amplitude $A_k(t)$, the instantaneous phase $\theta_{1k}(t)$, and the fundamental frequency $F_0(t)$. Constraints used in this model are shown in Table 1. Constraint (ii) for the above parameters gives $dA_k(t)/dt = C_{k,R}(t)$, $d\theta_{1k(t)} = D_{k,R}(t)$, and $dF_0(t)/dt = E_{0,R}(t)$, where $C_{k,R}(t)$, $D_{k,R}(t)$, and $E_{0,R}(t)$ are R-th-order differentiable piecewise polynomials (using Table 1 (ii)). 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

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

Regularity	Constraint
(i) common onset/offset	synchronous of onset/offset
(ii) gradualness of change	piecewise-differentiable polynomial approximation
(slowness)	(Kalman filtering)
(smoothness)	(spline interpolation)
(iii) harmonicity	multiples of the fundamental frequency
(iv) changes occurring in the acoustic event	correlation between the instantaneous amplitudes

$$\theta_k(t) = \arctan\left(\frac{S_k(t)\sin(\phi_k(t)-\theta_{1k}(t))}{S_k(t)\cos(\phi_k(t)-\theta_{1k}(t))+C_k(t)}\right), \quad (4)$$

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

3. IMPROVEMENTS TO PREVIOUS MODEL

To handle real speech extraction, we improved our previous model [8] in the following two respects.

3.1. $F_0(t)$ estimation block

In the proposed model, the fundamental frequency $F_0(t)$ is estimated using Comb filtering on the auditory-motivated filterbank. This Comb filter is defined by

$$\mathrm{Comb}(k,\ell) = \begin{cases} 2, & \omega_k = n\cdot\omega_\ell, 1 < n < 3 \\ 1, & \omega_k = n\cdot\omega_\ell, 4 \le n < N \\ 0, & \text{otherwise} \end{cases} \quad (5)$$

where k and ℓ are indices, ω_k and ω_ℓ are the center frequencies in channels, and N is the number of harmonics of the highest order. Then, $\hat{\ell}$, which corresponds to the channel containing the fundamental wave, is determined by

$$\hat{\ell} = \arg\max_{\ell\le L}\sum_{k=1}^{K}\mathrm{Comb}(k,\ell)S_k(t), \quad (6)$$

where L is the upper-limited search region of ℓ. The estimated $F_0(t)$ is determined by $\omega_{\hat{\ell}}/2\pi$.

Since the number of channels in the auditory-motivated filterbank is finite, the estimated fundamental frequency $F_0(t)$ takes a discrete value. In addition, the fluctuation of the estimated $F_0(t)$ 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 constraint (ii) for a segment. Here, the above segment is determined using the following equation, as the duration for which the temporal variation of $F_0(t)$ has variance of zero as $F_0(t)$.

$$\frac{1}{T_h - T_{h-1}} \int_{T_{h-1}}^{T_h} \left| F_0(t) - \overline{F_0(t)} \right|^2 dt \leq 0, \tag{7}$$

where the length of the segment is $T_h - T_{h-1}$. In this paper, let the parameters be $N = 10$ and $L = K/4$.

3.2. Separation block

In this paper, in order to reduce the computational cost for estimating $C_{k,R}(t)$ and $D_{k,R}(t)$, we assumes that $C_{k,R}(t)$ is a linear ($R = 1$) polynomial ($dA_k(t)/dt = C_{k,1}(t)$) and $D_{k,R}(t)$ is zero ($d\theta_{1k}(t)/dt = D_{k,0} = 0$) in constraint (i). In this assumption, $A_k(t)$ which can be allowed to undergo a temporal change in region, constrains the second-order polynomial ($A_k(t) = \int C_{k,1}(t)dt + C_{k,0}$). Moreover, $\theta_{1k}(t)$, which is constrained (i.e. $\theta_{1k}(t) = D_{k,0}$), cannot be allowed to temporarily change. Here, if the number of channels K is very large, each frequency of the signal component that passed through the channel approximately coincides with the center frequency of each channel. Even if the above condition is false, its frequency difference can be represented by $D_{k,0}$.

In the segment $T_h - T_{h-1}$, $C_{k,1}(t)$ and $D_{k,0}$ are determined by the following steps. First, let $D_{k,0}$ be any value within $-\pi/2 \leq D_{k,0} \leq \pi/2$. Next, using the Kalman filter, determine the estimated region, $\hat{C}_{k,0}(t) - P_k(t) \leq C_{k,1}(t) \leq \hat{C}_{k,0}(t) + P_k(t)$, where $\hat{C}_{k,0}(t)$ is the estimated value and $P_k(t)$ is the estimated error. Then select candidates of $C_{k,1}(t)$ using the spline interpolation in the estimated error region. Next, determine $C_{k,1}(t)$ using

$$\hat{C}_{k,1} = \mathop{\arg\max}_{\hat{C}_{k,0} - P_k \leq C_{k,1} \leq \hat{C}_{k,0} + P_k} \frac{< \hat{A}_k, \hat{\hat{A}}_k >}{||\hat{A}_k|| \, ||\hat{\hat{A}}_k||}, \tag{8}$$

where $\hat{A}_k(t)$ is obtained by the spline interpolation and $\hat{\hat{A}}_k(t)$ is determined in across-channel which is satisfied constraint (iii). Finally, determine $D_{k,0}$ using

$$\hat{D}_{k,0} = \mathop{\arg\max}_{-\pi/2 \leq D_{k,0} \leq \pi/2} \frac{< \hat{A}_k, \hat{\hat{A}}_k >}{||\hat{A}_k|| \, ||\hat{\hat{A}}_k||}. \tag{9}$$

Then, determining $\theta_{1k}(t) = \hat{D}_{k,0}$ and $\theta_{2k}(t) = \theta_k(t) + \theta_{1k}(t)$, we can determine $A_k(t)$ and $B_k(t)$ from Eqs. (2) and (3).

4. SIMULATIONS

To show that the proposed method can extract the desired signal $f_1(t)$ from the mixed signal $f(t)$, we carried out three simulations using the following signals: (1) noisy AM harmonicity tone; (2) noisy synthesized vowel; and (3) noisy real speech. In simulation 1, $f_1(t)$ was an AM complex tone, where $F_0(t) = 200$ Hz, the tone's instantaneous amplitude was sinusoidal (10 Hz). In simulation 2, $f_1(t)$ was a vowel /a/ synthesized by the log magnitude approximation (LMA) [4], where averaged $\overline{F_0(t)} = 125$ Hz, and jitter was 5 Hz (from 123 to 128 Hz). In simulation 3, $f_1(t)$ was a male vowel /a/ in the ATR database [7]. In all three simulations, $f_2(t)$ was bandpassed pink noise. Five types of $f(t)$ were used as simulation stimuli, where the SNRs of $f(t)$ ranged from 0 to 20 dB in 5-dB steps.

4.1. Evaluation of the estimated $F_0(t)$

We used the root-mean-squared (RMS) error between the reference $F_0(t)$ pattern and the estimated $F_0(t)$ to evaluate the estimation performance of the fundamental frequency $F_0(t)$. The reference pattern was extracted from clean speech using TEMPO [5]. The RMS errors of $F_0(t)$ for simulations 1, 2, and 3 were 3.8, 1.8, and 1.1 Hz, respectively. The results show that the proposed method could estimate $F_0(t)$ with high accuracy.

4.2. Evaluation of the extracted signal

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

$$\sqrt{\frac{1}{W} \sum_{\omega}^{W} \left(20 \log_{10} \frac{\tilde{F}_1(\omega)}{\tilde{\hat{F}}_1(\omega)} \right)^2}, \tag{10}$$

where $\tilde{F}_1(\omega)$ and $\tilde{\hat{F}}_1(\omega)$ are the amplitude spectra of $f_1(t)$ and $\hat{f}_1(t)$, respectively. In the above equation, the frame length is 51.2 ms, the frame shift is 25.6 ms, W is the analyzable bandwidth of the filterbank (about 6 kHz), and the window function is Hamming.

Next, in order to show the advantages of the constraints in Table 1, we compare the performance of our method under the following three conditions: (1) extract the harmonics using the Comb filter and predict $A_k(t)$ 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.

Segregation accuracies of the three simulations are shown in Fig. 2. For example, when the SNR of $f(t)$ was 10 dB as shown in Fig. 3(b) for simulation 3, the proposed method could segregate $A_k(t)$ with high accuracy and could extract $\hat{f}_1(t)$, shown in Fig. 3(d), from $f(t)$. In addition, we compared our proposed method with the other method (under three conditions) for the above simulations. The results show that the segregation accuracy using the pro-

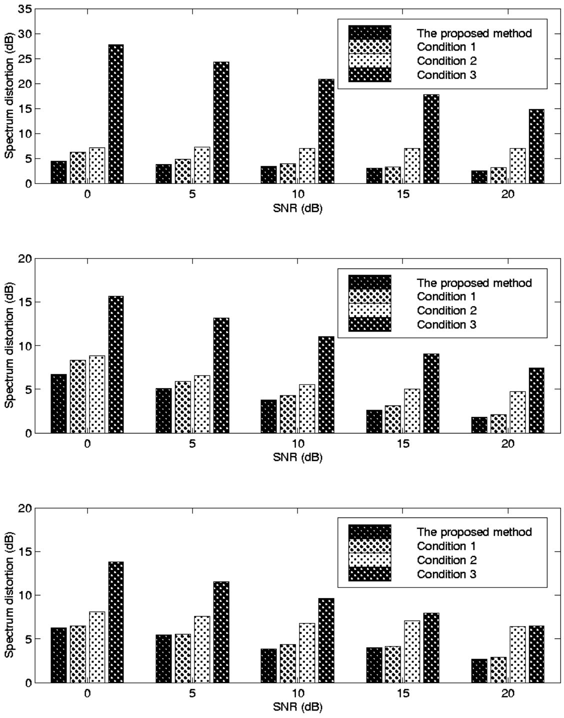

Figure 2: Segregation accuracies for simulations: (top) simulation 1, (middle) simulation 2, (bottom) simulation 3.

posed method was better than that using the other three conditions. The results for these three simulations and three conditions show that the proposed method can segregate the desired harmonic tone from a noisy harmonic tone with high accuracy. Improvements in spectrum distortion for simulations 1, 2, and 3 are about 17.6, 7.0, and 5.4 dB, respectively.

5. CONCLUSIONS

This paper proposed a new method of extracting the desired speech from noisy speech, by improving two aspect of our previously proposed method. One was to incorporate a method of estimating the fundamental frequency $F_0(t)$, by using the Comb filtering on the filterbank. The other was to reconsider the constraints on the separation block, which are the instantaneous amplitude, input phase, and fundamental frequency of the desired signal.

As an example of segregation using the proposed method, we demonstrated three simulations of segregating two acoustic sources. The results for extracting $F_0(t)$ showed that the proposed method can estimate the fundamental frequency with high accuracy. In particular, the results of simulations 1 and 2 examined how to solve the problem of segregating two acoustic sources. The results of simulation 3 showed that the proposed method could also extract the speech signal from noisy speech. Moreover, comparisons under three conditions showed that using the proposed method with three conditions related to the four regularities was better than using the other method under all three conditions.

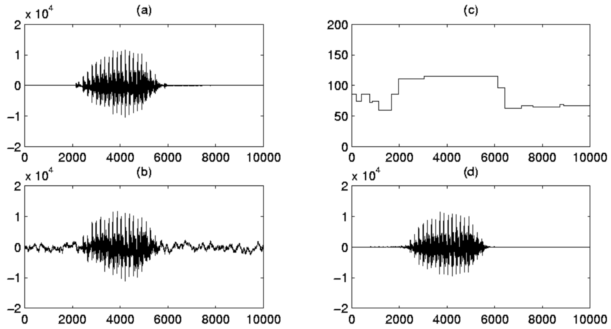

Figure 3: Examples of segregation: (a) original $f_1(t)$, (b) mixed signal $f(t)$, (c) $F_0(t)$, (d) segregated signal $\hat{f}_1(t)$

6. 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.

7. REFERENCES

1. Bregman, A.S. "Auditory Scene Analysis: hearing in complex environments," in Thinking in Sounds, (Eds. S. McAdams and E. Bigand), pp. 10–36, Oxford University Press, New York, 1993.
2. Cooke, M. P. "Modeling Auditory Processing and Organization," Ph.D. Thesis, University of Sheffield, 1991 (Cambridge University Press, Cambridge, 1993).
3. Ellis, D. P. W. "Prediction-driven computational auditory scene analysis," Ph.D. thesis, MIT Media Lab., 1996.
4. Imai, S. "Low bit rate cepstral vocoder using the log magnitude approximation filter," In Proc. ICASSP78, pp. 441-444, 1978.
5. Kawahara, H. "STRAIGHT - TEMPO: A Universal Tool to Manipulate Linguistic and Para-Linguistic Speech Information," In Proc. SMC-97, Oct. 12-15, Orlando, Florida, USA.
6. 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.
7. Takeda, K., Sagisaka, Y., Katagiri, S., Abe, M., and Kuwabara, H. Speech Database User's Manual, ATR Technical Report TR-I-0028, 1988.
8. Unoki, M. and Akagi, M. "A Method of Signal Extraction from Noisy Signal," In Proc. EuroSpeech'97, vol. 5, pp. 2583-2586, RHODOS-GREECE, Sept. 1997.