A Waveform Generation Model Based Approach for Segregation of Monaural Mixture Sound

Akira SASOU and Kazuyo TANAKA
National Institute of Advanced Industrial Science and Technology (AIST)
1-1 Central 2, Umezono, Tsukuba, Ibaraki 305-8568, Japan
{a-sasou,kaz.tanaka}@aist.go.jp

ABSTRACT

This paper describes a novel method for segregating concurrent monaural sounds. In a real environment, there are many types of sounds, such as periodic sound, aperiodic sound, impulsive sound and so on, and several sounds usually occur simultaneously. In order to recognize the sounds, it is necessary to be able to model such various type of sounds and segregate the concurrent sounds. The proposed method adopts a waveform generation model consisting of an Auto-Regressive process and a Hidden Markov Model as a template model and achieves segregation of monaural concurrent sounds based on the mixed AR-HMMs. Experiments were conducted to confirm the feasibility of the method using ten types of non-speech sounds. The experimental results indicate that the proposed method is effective for various types of sounds.

1 Introduction

In a real environment, there are many kinds of sounds other than speech and we recognize the environment by both visual and audio cues. However, studies dealing with recognition of non-speech sounds can be rarely seen in contrast with speech recognition. Realizing the ability to recognize the environment by hearing in an engineering sense would be especially helpful in enabling hearing-impaired persons to recognize the environment.

Real environment includes many types of sounds, such as periodic sound, aperiodic sound, impulsive sound and so on, and several sounds usually occur simultaneously. Thus, it is necessary to be able to model such various types of sounds and segregate the concurrent sounds in order to recognize the environment by hearing. In recent years, microphone array systems have been investigated to segregate concurrent sounds[1]. However, microphone array systems need to be large in order to achieve more accurate segregation of sounds in near-field. This sometimes causes difficulties, such as the case for hearing-impaired people to carry the system, for example. This paper thus focuses on segregation of monaural recorded sounds.

A template-matching-based method is one approach to resolving the problem. For instance, a method based on matching between an input waveform and a template waveform of each sound stored in advance has been proposed[2]. This method introduces template waveform adaptation. Since each sound source has several variable elements, such as the gain and/or phase, the observed waveforms may differ, even if the waveforms are obtained from the same sound source. The template waveform therefore needs adaptation to those variable elements. The introduced method achieves the adaptation by applying a linear filter to a raw template waveform. However, for a periodic sound source, this makes it difficult to adapt the pitch of the raw template waveform to that of the observed waveform.

In order to overcome this difficulty, this paper proposes a novel segregation method that adopts a waveform generation model for each sound source as the template model, instead of the raw waveform. The proposed template model is a kind of source-filter model consisting of an Auto-Regressive (AR) process for the articulatory filter and a Hidden Markov Model (HMM) for the excitation source. We hereafter call this model as AR-HMM. The AR-HMM can represent various types of sound sources, such as periodic, aperiodic or impulsive sounds, by means of appropriately designing a network topology of the HMM. In the AR-HMM, the excitation source is separated from the articulatory filter. The gain and phase of the template model can thus easily be adapted to the desired values by adjusting the output gain and the state transition of the HMM. The proposed method segregates monaural concurrent sounds based on the mixed AR-HMMs, and is therefore applicable to various types of sounds. This paper considers the case in which only the gain and phase of the excitation source for each template model can change, i.e., the AR coefficients in each template model are fixed. The sounds corresponding to these conditions are collision sounds, ringing sound of a bell-alarm clock, vowel sounds and so on.

2 Template Model Based on AR-HMM

An Auto-Regressive (AR) process can describe the passive oscillation after an external power excites the object, such as collision sounds caused by knocking on a door, beating a drum and ringing a bell-alarm clock. Thus these sound sources can be adequately represented by a source-filter model consisting of an excitation source and an AR process. A speech signal also corresponds to this kind of sound source since the speech signal is generated by injecting a glottal volume flow into the vocal tract. In the proposed source-filter model, a Hidden Markov Model (HMM) and time-varying output gain are adopted as an excitation source in order to represent such various types of sounds. We call this source-filter model as an AR-HMM.

The output probability distribution of each node of the excitation source HMM is assumed to be a single normal distribution. Figure 1 shows examples of the AR-HMM. The AR-HMM at the top of the figure represents a stationary noise source. If the prediction order is equal to zero, the model represents a white Gaussian noise source. The nodes of the second AR-HMM are concatenated in a ring state, so the state transition occurs in order. That's why this type of AR-HMM can be used to represent periodic and non-

stationary sound sources. An ergodic HMM as shown in the bottom can be used to represent an aperiodic sound source. The numbers of the nodes and the prediction order are determined according to the sound source. Usually, an Akaike Information Criterion (AIC) is employed to determine the model[5]. The AR-HMM parameters are estimated from a sample signal of the sound source so that the likelihood of the parameters is maximized. The details of the estimation algorithm are described in [3, 4].

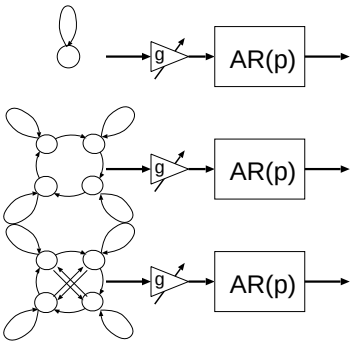

Figure 1: Examples of AR-HMM

3 Gain-adapted AR-HMM Decomposition

3.1 Mixture Model of Concurrent Sounds

Figure 2 shows a model of M mixed sound sources. In this model, the parameters, except the time-varying gain of each AR-HMM, have been estimated by the parameter estimation algorithm[3, 4]. The proposed segregation algorithm consists of the following processes.

1. Adaptively estimate each time-varying gain.
2. Estimate the state transition of each HMM.
3. Segregate a mixture sound based on the adapted AR-HMM.

Although the processes adapting the mixed AR-HMMs are similar to those of the HMM decomposition method[6], the proposed method needs to treat a process adapting time-varying gains. Moreover, since the mixture model output is represented as a summation of all the excitation sources followed by the AR processes, the proposed model needs to also take into account the effects of the AR processes in estimating the state transition. We call this segregation algorithm as a Gain-Adapted AR-HMM (GA-ARHMM) decomposition method.

Like the HMM decomposition method, the GA-ARHMM decomposition method searches for the optimum transition sequence based on a trellis diagram with a dimension of $(M+1)$ and resulting by combining a time-axis and a cartesian product $\Sigma = S_1 \otimes \cdots \otimes S_M$, where S_m denotes a set of each HMM's nodes. Each mixture state is represented by $\mathbf{s} = [s_1, \cdots, s_M] \in \Sigma$, where $s_m \in S_m$.

3.2 Eliminating Effects of the AR Processes

In order to estimate the state transition, the effects of AR processes need to be eliminated from an observed mixture sound $y(t)$. This is achieved as follows. A sound source

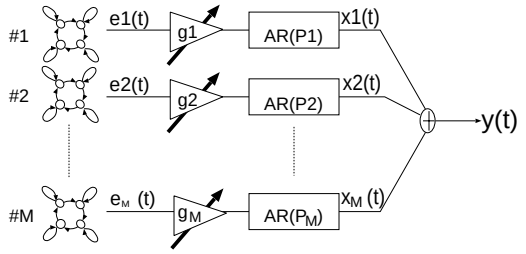

Figure 2: A mixture model of concurrent sounds

signal $x_m(t)$ of the mth AR-HMM is generated according to

$$x_m(t) = \sum_{k=1}^{P_m} a_m(k) x_m(t-k) + g_m(t) e_m(t) \tag{1}$$

where P_m is a prediction order; $a_m(k), k = 1, \cdots, P_m$ are AR coefficients; $g_m(t)$ is a time-varying gain; and $e_m(t)$ is an excitation source. A mixture sound signal is given by

$$y(t) = \sum_{m=1}^{M} \sum_{k=1}^{P_m} a_m(k) x_m(t-k) + \sum_{m=1}^{M} g_m(t) e_m(t) \tag{2}$$

Since the first term on the right side of the equation contains the auto-regressive components of all the AR-HMMs, the state transition cannot be estimated directly from the mixture sound signal like the HMM decomposition method does. Hence, we focus on the residual resulting by subtracting the auto-regressive components from the mixture sound signal. However, the sound signals $x_m(t), m = 1, \cdots, M$ in (2) are the original signals before being mixed and are actually unknown. The segregated signals are therefore used to calculate the residual. A segregated signal of each sound source exists along a path reaching each mixture state. Let $x_m(\mathbf{s}, t)$ denote the segregated signal of the mth sound source in mixture state $\mathbf{s} \in \Sigma$ at time t. We introduce $\Psi(\mathbf{s}, t) \in \Sigma$ as a back tracer representing the mixture state from which the state transition to the mixture state $\mathbf{s}$ at time t occurred. The state transition sequence $\mathbf{s}^*(\mathbf{s}, t, t'),\ t' = 0, \cdots, t$ reaching a mixture state $\mathbf{s}$ at time t is given by

$$\begin{aligned} &\mathbf{s}^*(\mathbf{s}, t, t) = \mathbf{s} \\ &\text{for } \ t' = t-1, t-2, \cdots, 0 \\ &\quad \mathbf{s}^*(\mathbf{s}, t, t') = \Psi(\mathbf{s}^*(\mathbf{s}, t, t'+1), t+1). \end{aligned} \tag{3}$$

The segregated signal $x_m^*(\mathbf{s}, t, t'),\ t' = 0, \cdots, t$ of the mth sound source obtained along the path reaching a mixture state $\mathbf{s}$ at time t is given by

$$x_m^*(\mathbf{s}, t, t') = x_m(\mathbf{s}^*(\mathbf{s}, t, t'), t'). \tag{4}$$

By using (2),(4), the residual $r(\mathbf{s}, t)$ evaluated in a mixture state $\mathbf{s}$ at time t is given by

$$r(\mathbf{s}, t) = y(t) - \sum_{m=1}^{M} \sum_{k=1}^{P_m} a_m(k) x_m^*(\mathbf{s}, t, t-k). \tag{5}$$

If the signals are segregated correctly, that is, $x_m^*(\mathbf{s}, t, t') \approx x_m(t')$, the residual can be represented as

$$r(\mathbf{s}, t) \approx \sum_{m=1}^{M} g_m(t) e_m(t). \tag{6}$$

From this, it is obvious that the residual is free from the effects of the auto-regressive processes. If we assume that all the gains are 1.0, the residual becomes equal to a summation of outputs directly from all the HMMs. By focusing on the residual, we can apply the HMM decomposition method to estimating the state transition. However, the gain assumption is generally not adequate, so we need to estimate the time-varying gains in the residual. The next section describes an adaptive estimation of the gains.

3.3 Adaptive Estimation of Time-varying Gains

First, we summarize the conditions of the time-varying gains as follows.

1. The phase of each excitation source is not affected by the time-varying gain.
2. The gain varies slowly in comparison with the speed of state transition of the excitation source HMM.

The first gain condition can be satisfied by restricting the sign of each gain to a positive sign. That is,

$$g_m(\mathbf{s},t) > 0, \quad for\ \forall m,\ \mathbf{s},\ t. \tag{7}$$

In the following, we describe the adaptive estimation of the time-varying gain. In a state transition from $\mathbf{s}$ at time t to $\mathbf{s}'$ of the next time, the output distribution of the node emitting an excitation source in the mth HMM is given by $N(\mu_m(s_m), \sigma_m^2(s_m))$. As seen in (6), the residual $r(\mathbf{s},t)$ is represented by a linear combination of excitation sources multiplied by the gains. The occurrence probability of the residual r conditioned on the gains taking values of $g_m, m = 1,\cdots,M$ is thus given by

$$O_{\mathbf{s}}(r|g_1,\cdots,g_M) = \frac{1}{\sqrt{2\pi\sum_{m=1}^{M} g_m^2\sigma_m^2(s_m)}} \times \exp\left\{-\frac{(r-\sum_{m=1}^{M} g_m\mu_m(s_m))^2}{2\sum_{m=1}^{M} g_m^2\sigma_m^2(s_m)}\right\}. \tag{8}$$

From another point of view, the equation can be regarded as the likelihood that the evaluated residual was generated by the gains. Due to condition (7), the domain of the likelihood function is restricted to the first quadrant, in which the likelihood function is a convex function with respect to the gain. Hence, the optimal gains maximizing the likelihood can be determined. However, since the optimization of the gains at each time causes discontinuities in the time series of the estimated gains, the second condition for the gains is not satisfied. Instead, we employ an adaptive estimation algorithm based on the gradient of the likelihood function. Since the adaptation process is achieved by slightly modifying each gain at each time, the rapid variations in the time series of the estimated gains should be avoided. The gains estimated along the path corresponding to actual state transitions of the HMMs are expected to converge to the maximum likelihood estimates. The gradient of the logarithmic likelihood with respect to a gain g_l is given by

$$\frac{\partial \ln O_{\mathbf{s}}}{\partial g_l} = \frac{\{r-\sum_{m=1}^{M} g_m\mu_m(s_m)\}\mu_l(s_l) - g_l\sigma_l^2(s_l)}{\sum_{m=1}^{M} g_m^2\sigma_m^2(s_m)} + \frac{\{r-\sum_{m=1}^{M} g_m\mu_m(s_m)\}^2 g_l\sigma_l^2(s_l)}{\{\sum_{m=1}^{M} g_m^2\sigma_m^2(s_m)\}}, \tag{9}$$

and each gain is updated by

$$g_l(\mathbf{s},t) = \alpha\frac{\partial}{\partial g_l}\ln O_{\mathbf{s}}(r|g_1(\mathbf{s}^*,t-1),\cdots,g_M(\mathbf{s}^*,t-1)) + g_l(\mathbf{s}^*,t-1), \quad l=1,\cdots,M \tag{10}$$

where α is the step size and $\mathbf{s}^* = \Psi(\mathbf{s},t)$.

3.4 Selecting the Most Probable Path

Using the estimated gains $g_m(\mathbf{s},t), m = 1,\cdots,M$ for each mixture state $\mathbf{s} \in \Sigma$, the occurrence probability of the residual is given by

$$O_{\mathbf{s}}(r|g_1(\mathbf{s},t),\cdots,g_M(\mathbf{s},t)). \tag{11}$$

The transition probability is given by

$$T(\mathbf{s},\mathbf{s}') = \prod_{m=1}^{M} b_m(s_m,s_m') \tag{12}$$

where $b_m(s,s')$ presents a transition probability from a state s to a state s'. Thus, the most probable path of transition to a mixture state $\mathbf{s}'$ is selected by means of the following equation.

$$\Psi(\mathbf{s}',t+1) = \arg\max_{\mathbf{s}} P(\mathbf{s},t)O_{\mathbf{s}}(r)T(\mathbf{s},\mathbf{s}') \tag{13}$$

where $P(\mathbf{s},t)$ denotes the maximum probability for a state transition sequence ending in $\mathbf{s}$ at time t. The probability of the selected path is given by

$$P(\mathbf{s}',t+1) = \max_{\mathbf{s}} P(\mathbf{s},t)O_{\mathbf{s}}(r)T(\mathbf{s},\mathbf{s}'). \tag{14}$$

3.5 Decomposition of a Residual

The residual is decomposed at each time in order to segregate a mixture sound based on (6), where the estimated gains $g_m(\mathbf{s},t)$ are used instead of the actual gains $g_m(t)$. Hereafter, let q_m denote an excitation source multiplied by the gain as $q_m = g_m(\mathbf{s},t)e_m(t)$, which is a random variable conforming to the normal distribution given by

$$Q'_{\mathbf{s}}(q_m) = \frac{1}{\sqrt{2\pi\sigma_{q_m}^2}}\exp\left\{-\frac{(q_m-\mu_{q_m})^2}{2\sigma_{q_m}^2}\right\} \tag{15}$$

where $\mathbf{s} = \Psi(\mathbf{s}',t+1)$, $\mu_{q_m} = g_m(\mathbf{s},t)\mu_m(s_m)$ and $\sigma_{q_m}^2 = g_m^2(\mathbf{s},t)\sigma_m^2(s_m)$. Since the emissions of excitation sources from all the HMMs are mutually independent events, the joint distribution of all the excitation sources is given by

$$Q_{\mathbf{s}}(q1,\cdots,q_M) = \prod_{m=1}^{M} Q'_{\mathbf{s}}(q_m). \tag{16}$$

We decompose the residual r so that the joint occurrence probability is maximized on condition $r = \sum_{m=1}^{M} q_m$. The decomposed values are obtained by solving the following simultaneous equations.

$$\sigma_{q_M}^2 q_k + \sigma_{q_k}^2\sum_{m=1}^{M} q_m = \sigma_{q_M}^2\mu_{q_k} + \sigma_{q_k}^2(r-\mu_{q_M}) \tag{17}$$

$$q_M = r - \sum_{m=1}^{M} q_m, \qquad k = 1,\cdots,M-1$$

The segregated signal of each sound source is then given by

$$x_m(\mathbf{s}',t+1) = \sum_{k=1}^{P_m} a_m(k)x_m^*(\mathbf{s}',t+1,t+1-k) + q_m. \tag{18}$$

3.6 Flow of the Algorithm

1. Initialization:
   $g_m(\mathbf{s}, -1) = 1 \quad \text{for } \forall m, \mathbf{s} \in \Sigma$
   $x^*_m(\mathbf{s}, 0, t') = 0 \quad \text{for } \forall m, \mathbf{s} \in \Sigma, t' < 0$
2. Recursion:
   for $t = 0, \cdots, T-1$
   for $\mathbf{s}' \in \Sigma$
   for $\mathbf{s} \in \Sigma$
   Evaluation of a residual by (5)
   Adaptation of gains by (9),(11)
   Selection of the path by (13),(14)
   Segregation of mixture sound by (17),(18)
3. Termination:
   $\mathbf{s}^{\#} = \arg\max_{\mathbf{s} \in \Sigma} P(\mathbf{s}, T-1)$

4 Experimental Results

This section presents the experimental results of segregating mixtures of two sound sources in a real environment. The experiments were conducted using the non-speech sound data of the RWCP Sound Scene Database in a Real Acoustic Environment[7]. Ten such sounds (the beating of a wooden board, a metal can, a plastic case, a drum, a handclap, a glass cup, a castanet, and a piece of china and the ringing of a bell-alarm clock and a bicycle bell) were used. We generated a signal for each sound source by concatenating the five successive samples in the data base. The generated signal becomes periodic even if the sound source originally exhibits aperiodicity. Nodes in each AR-HMM are thus concatenated in a ring state. We set the prediction order to 18 and the number of nodes to 10 and estimated the parameters of each AR-HMM using the generated signal.

Mixture sounds were generated by adding one sound source to another one, in which we used ten samples other than the samples used for making the template model. Figure 3 shows an example of segregated waveforms. In this example, we mixed the beating sound of a wooden board and the ringing sound of a bell-alarm clock. The SNRs of the segregated waveforms were 25.1 and 24.1[dB]. The SNRs of all the segregated waveforms are represented by a frequency distribution as shown in Figure 4. The average SNR was 6.07[dB]. The standard deviation was 12.2[dB].

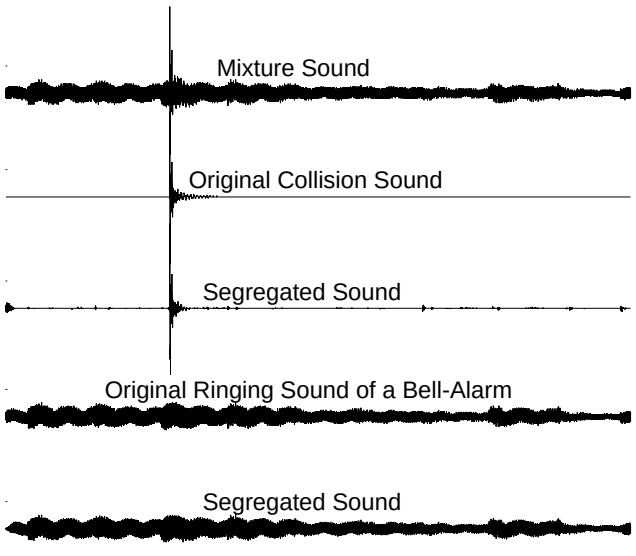

Figure 3: An example of segregated waveforms

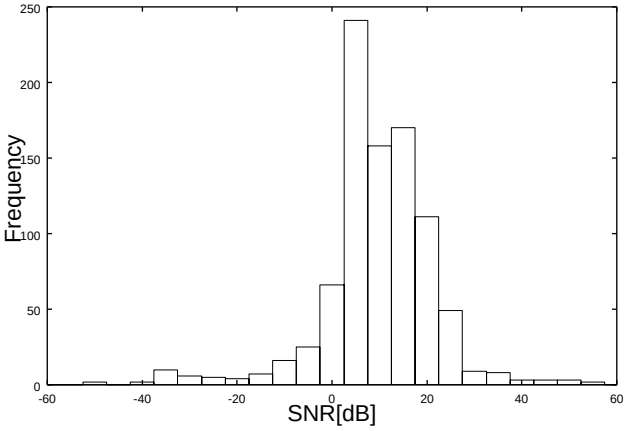

Figure 4: Frequency distribution of SNRs of segregated waveforms

5 Conclusion

This paper proposed a waveform generation model based method of segregating monaural concurrent sounds. The waveform generation model consists of an Auto-Regressive process and a Hidden Markov Model. The AR-HMM can represent various types of sound sources, such as periodic, aperiodic or impulsive sounds, by means of appropriately designing a network topology of the HMM. The proposed segregation method is therefore applicable to various types of sounds. From the experimental results, we were able to confirm the feasibility of the proposed method.

References

[1] D.E.Dudgeon, Array Signal Processing. Englewood Cliffs, NJ: Prentice-Hall, 1993.

[2] K.Kashino,H.Murase, "A sound source identification system for ensemble music based on template adaptation and music stream extraction," Speech Communication, vol.27, no.3-4, pp.337-349, 1999

[3] A.Sasou,K.Tanaka, "Glottal excitation modeling using HMM with application to robust analysis of speech signal," Proc. of ICSLP2000, vol.IV, pp.704-707, Oct.2000

[4] A.Sasou,K.Tanaka, "Robust LP Analysis Using Glottal Source HMM with Application to High-Pitched and Noise-Corrupted Speech," Proc. of EuroSpeech2001, vol.4, pp2443-2446, Sep.2001

[5] H.Akaike, "A new-look at the statistical model identification," IEEE Trans. Autom. Control, vol.AC-19, no.6, pp.716-723, 1974

[6] A.P.Varaga, R.K.Moore, "Hidden Markov model decomposition of speech and noise," Proc. of ICASSP-90, pp.845-848, 1990.

[7] S.Nakamura, K.Hiyane, F.Asano, T.Nishimura, T.Yamada, "Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition," Proc. International Conference on Language Resources and Evaluation, pp.965-968, 2000