TRACKING HUMAN SPEECH EVENTS USING A PARTICLE FILTER

*H. Asoh, I. Hara, F. Asano**

AIST
Tsukuba, Ibaraki 305-8568, Japan

K. Yamamoto

University of Tsukuba
Tsukbua, Ibaraki 305-8577, Japan

ABSTRACT

A method of detecting and tracking moving human speech events by integrating audio and video signals using a particle filter is proposed and evaluated. Using the particle filter, location, on/off status, and human/non-human determination can be estimated simultaneously for multiple sound sources. Experiments demonstrate that the proposed method performs well for the data obtained in an ordinary meeting room using a microphone array and a monocular camera.

1. INTRODUCTION

Tracking user speech events is a very important function for realizing robust multimedia human-machine interaction in noisy everyday environments. Various interactive systems including personal data assistants, video conferencing systems, and interactive robots, require such a function. The result of tracking can be used as an input for separating the speech signal from noise and for enhancing and recognizing the speech.

In order to distinguish between user speech and audio signals from interference sound sources such as television, we have proposed to combine audio and video information in a Bayesian framework[1]. A microphone array is used to localize sound sources, and a monocular camera to localize humans. These informations are then combined to compute the posterior probability of the time and location of human speech events.

The problem is composed of several sub-problems, such as estimation of the number of sound sources, localization of multiple sound sources, and estimation of the number of humans. In a previous study[2], these sub-problems were solved separately. We first localized multiple sound sources and potential human speakers and then performed tracking of human speech events using the results of the localization.

In this paper, we propose a method by which to solve these problems simultaneously using statistical signal models and a particle filter in the Bayesian framework. Particle filters[3] are very powerful technique for estimating a time series of hidden variables from noisy observations. In contrast to Kalman filters, which are applicable only to linear Gaussian models, particle filters can be applied to very general statistical signal models. In addition, particle filters are very simple and efficient algorithms.

Checka *et al.*[4] applied a particle filter to the problem of multiple human speakers and speaker activity tracking. In the present study, we generalize their framework to include human/non-human discrimination and enhance their statistical signal model in order to estimate human speech intervals more accurately. Accurate estimation of speech intervals is crucial in order to use the tracking results as inputs for speech enhancement and recognition.

The present paper is organized as follows. In Section 2, we introduce a Bayesian framework for tracking human speech events. In Section 3, statistical signal models for audio and video signals are described. Then, in Section 4, the proposed framework and models are applied to experimental data obtained in an ordinary meeting room using a microphone array and a monocular camera mounted on the head of a humanoid robot. In the experiment, the performances of several signal models are compared. Finally, Section 5 presents a discussion and conclusions.

2. BAYESIAN FRAMEWORK

The problem of tracking human speech events can be formulated in a framework of Bayesian estimation of hidden state sequences. As in [4], let the hidden variable vector be

$$\mathbf{X}(t) = (n(t), \chi_1(t), \cdots, \chi_{n(t)}(t)),$$

where $n(t)$ is the number of tracking targets and $\chi_i(t) = (\theta_i, s_i, h_i)$ is the configuration of the ith target. The discrete variable θ_i denotes the 2D direction of the target, the Boolean variable s_i denotes audio activity (on/off). Another Boolean variable, h_i, denoting human/non-human discrimination of the target, is added. Observation variables $\mathbf{Y}(t)$ are composed of audio signals $\mathbf{Y}_a(t)$ from microphones and video signals $\mathbf{Y}_v(t)$ from cameras.

In the Bayesian framework, the relationship between $\mathbf{X}_{1|T} = \mathbf{X}(1), \cdots, \mathbf{X}(T)$ and $\mathbf{Y}_{1|T} = \mathbf{Y}(1), \cdots, \mathbf{Y}(T)$ is modeled by a joint probability distribution $P(\mathbf{X}_{1|T}, \mathbf{Y}_{1|T})$. As is often done, we assumed that the joint probability dis-

*This work is partly supported by JSPS KAKENHI 14208033.

tribution can be decomposed as

$$P(\mathbf{X}_{1|T}, \mathbf{Y}_{1|T}) = \prod_{t=1}^{T} P(\mathbf{Y}(t)|\mathbf{X}(t))P(\mathbf{X}(t)|\mathbf{X}(t-1)).$$

This means that only the state transition probabilities $P(\mathbf{X}(t)|\mathbf{X}(t-1))$ and the observation probabilities (the likelihood of the observation) $P(\mathbf{Y}(t)|\mathbf{X}(t))$ are needed in order to specify the joint probability distribution. When observations $\mathbf{y}_{1|t}$ are given, the posterior probability distribution $P(\mathbf{X}_{1|t}|\mathbf{y}_{1|t})$ is computed.

The particle filter is a kind of Monte Carlo technique for computing and representing the posterior distribution efficiently with random particles distributed in the hidden variable space. Although several sophisticated particle filter algorithms have been proposed, for the present study, we chose the simplest algorithm and use state transition probability as the proposal distribution and the likelihood of the observation as the importance weight[3].

3. STATISTICAL MODELS

3.1. State Transition Probability

As the state transition model, a very simple random model is used in which each target is assumed to move independently from its current location according to a Gaussian distribution with zero mean and common variance σ_l. Speech activity s and human/non-human state h changes randomly according to transition probability. Although, in reality, h does not change, for the sake of implementation we model h as being able to change randomly.

3.2. Audio Signal Model

The audio signal $\mathbf{Y}_a(t)$ is treated in the frequency domain. Let the short-time Fourier transform (STFT) of the M microphone inputs be $\mathbf{z}(\omega, t) = (Z_1(\omega, t), \cdots, Z_M(\omega, t))^T$, where $Z_m(\omega, t)$ is the STFT of the mth microphone input at time t and frequency ω.

Then, for each narrow band ω, $\mathbf{z}$ can be modeled as

$$\mathbf{z} = A\mathbf{s} + \mathbf{n}$$

with location vector matrix $A = (\mathbf{a}(\theta_1), \cdots, \mathbf{a}(\theta_L))$, source spectrum $\mathbf{s} = (S_1, \cdots, S_L)^T$, and background noise spectrum $\mathbf{n} = (N_1, \cdots, N_M)$. Here, L is the number of active sound sources. ω and t are omitted for the sake of simplicity.

As in [5], we assume that both the signal and the noise are 0-mean Gaussian, that is, $E[\mathbf{s}, \cdots, \mathbf{s}^H] = \mathrm{diag}(\gamma_1, \cdots, \gamma_L)$, and $E[\mathbf{n}, \mathbf{n}^H] = \sigma_n \mathbf{I}$. Then the log-likelihood function of $\mathbf{z}$ becomes

$$L(\mathbf{z}|\mathbf{X}) = -\log|det(K_y)| - \frac{1}{2}\mathbf{z}^H K_y^{-1}\mathbf{z}, \quad \text{where}$$

$$K_y = \sum_{l=1}^{L} \gamma_l \, \mathbf{a}(\theta_l)\mathbf{a}(\theta_l)^H + \sigma_n \mathbf{I}.$$

In [4] the likelihood is computed for each single sound frame. In order to stabilize the likelihood values, we take the average over N frames. This means that we use log-likelihood

$$L(\mathbf{z}_{t|t+N}|\mathbf{X}(t)) = -\log|det(K_y(t))| - \frac{1}{2}tr(C_z(t)K_y(t)^{-1}),$$

where

$$C_z(t) = \frac{1}{N}\sum_{k=t}^{t+N} \mathbf{z}(k)\mathbf{z}(k)^H.$$

The observation probability for the broadband signal is computed as

$$P(\mathbf{y}_{a,t|t+N}|\mathbf{X}(t)) = \prod_{\omega} \exp L(\mathbf{z}_{t|t+N}(\omega)|\mathbf{X}(t)).$$

However, in order to evaluate the likelihood, the values of $\gamma_l(\omega, t)$ must be known. Although we can include γ_l among the hidden variables of the particle filter, doing so causes the hidden variable space to become enormous and computation of the posterior distribution becomes impractical. In [4], it is assumed that $\gamma_l = 1$ for active audio sources and $\gamma_l = 0$ for all other audio sources. This is referred to herein as the 0/1-γ model. Because the relative power of each active sound source changes in time, this is a rough approximation. In particular, the power of human speech changes greatly due to the presence of both vowel and consonant sounds.

In order to cope with this problem, we introduced two more elaborate models. The first model is a mixture of multiple models with different γ_l values. In this paper, a mixture of three γ_l values, $\gamma_l = 1, 0.3, 0.1$ was employed. This model is referred to herein as the multiple-γ model. In the second model, we tried to estimate γ_l from the signal[5] as

$$\hat{\gamma}_l = \frac{\mathbf{a}(\theta_l)^H C_z \mathbf{a}(\theta_l)}{|\mathbf{a}(\theta_l)|^4}.$$

This model is referred to as the estimated-γ model. These models were compared experimentally.

3.3. Video Signal Model

For each single camera image, the video observation probability $P(\mathbf{y}_v(t)|\mathbf{X}(t))$ is computed as follows. First, for each given direction $\theta_l(t)$ of a human target, the color distribution around the direction is evaluated using the skin color distribution modeled via a Gaussian distribution. Next, template matching using face templates is executed in the local region, and the minimum distance is evaluated by the distance distribution, which is also modeled by another 0-mean

Gaussian distribution. For the direction $\theta_l(t)$ for a non-human target, the likelihood is assumed to be uniform for all image data. Finally, the likelihood for all of the targets is computed by simply multiplying the likelihood values of the targets.

3.4. Combining Audio and Video Information

For every N audio FFT frame, the total observation probabilities $P(\mathbf{y}_{t|t+N}|\mathbf{X}(t))$ are computed simply by multiplying the audio and video observation probabilities. We introduce a parameter w for balancing information from audio and video. When the number of active sound sources are changes with time, the number of hidden variables in the signal models also changes. Hence the problem of estimating the number of active sound sources should be treated as a kind of model selection problem. So far, there has been proposed several model selection criteria such as AIC and MDL. In this paper we simply discounted the observation probability by the number of hidden variables. That is, we introduced another parameter α, and the observation probabilities finally become

$$P(\mathbf{y}_{t|t+N}|\mathbf{X}(t)) = P(\mathbf{y}_{a,t|t+N}|\mathbf{X}(t))P(\mathbf{y}_v(t)|\mathbf{X}(t))^w \alpha^{N_s(t)},$$

where $N_s(t)$ is the number of active sound sources.

4. EXPERIMENTS

4.1. Conditions

The performance of the proposed framework and models was evaluated using data obtained in a medium-sized meeting room with a reverberation time of approximately 0.5 s. As shown in Figure 1, the microphone array and camera were mounted on the head of a humanoid robot HRP-2 developed in AIST[6]. The robot head was placed on a computer controlled turntable that was rotated at a constant speed. A standing human speaker uttered a number of sentences in Japanese separated by pauses. As the interference sound, a loudspeaker played music continuously. The S/N ratio is about 0 dB. The configuration of the robot head, the human, and the loudspeaker is shown in Figure 1.

4.2. Results

Audio signals from eight microphones were sampled at 16 kHz. The length of the Fourier transform window was 512 and the frame shift was 128. The range of frequency ω was [800, 3000] Hz. The camera provided 320 x 240 images at 30 frames per second. The maximum number of targets in this experiment was two. The direction of targets θ was quantized into 3-degree segments, i.e., 120 direction bins were created. We chose the averaging interval $N = 9$ (about 0.1 sec.) for balancing stability and trackability. For every 0.1 sec. the likelihood was evaluated and state transition was executed. The number of particle was 500, and other parameters were $\sigma_n = \max_l(\gamma_l), w = 2$. Here, α was tuned for each data in order to obtain the best detection rate.

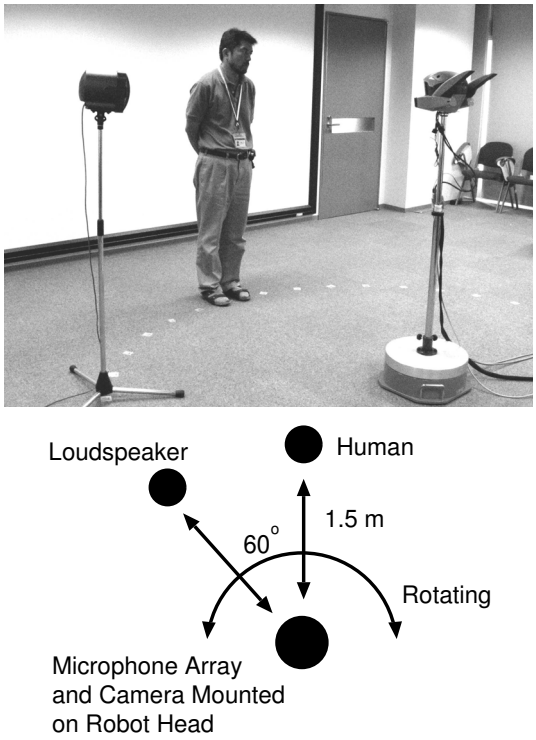

Fig. 1. Experimental setup

Figure2 shows a typical result for posterior probability computation. For this data, the turntable was first rotated counter-clockwise 30 degrees and then clockwise back to the initial position. The rate of rotation was approximately 15 degrees per second. In the figure, (a) shows the probability of target existence, (b) shows the probability of sound existence, and (c) shows the probability of human speech event existence, respectively. The horizontal axis indicates the time, and the vertical axis indicates the direction. Darker colors indicate higher probabilities. You can see that human speech events are clearly detected and tracked in (c) during the interval when the camera can observe the targets. However, for example, in the interval from $t = 13to22$ the sound from loudspeaker is also detected as human speech because the loudspeaker is out of the camera's view angle in the interval and the system could not discriminate whether the sound is from human or not.

Figure3 compared the marginal posterior probability of human speech event existence from different models. In the figure, (a) is ground truth, and (b), (c), and (d) show the posterior probabilities computed using the 0/1-γ model, the multiple-γ model, and the estimated-γ model, respectively. The figure indicates that human speech intervals are detected more stably in (c) and (d) than in (b).

Table1 summarizes the detection error rates for these

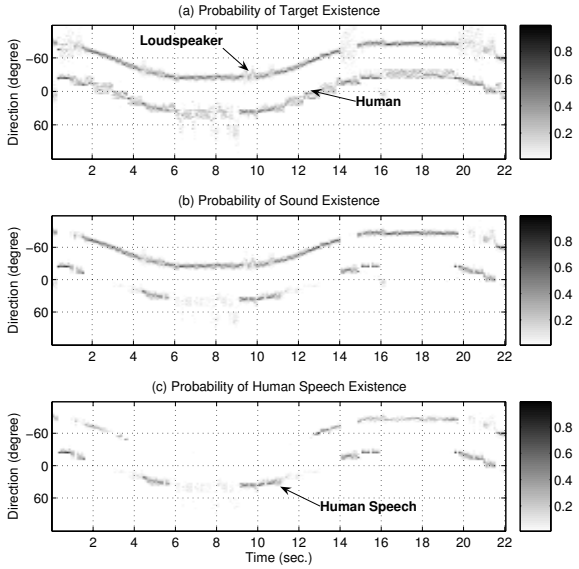

Fig. 2. Posterior probabilities.

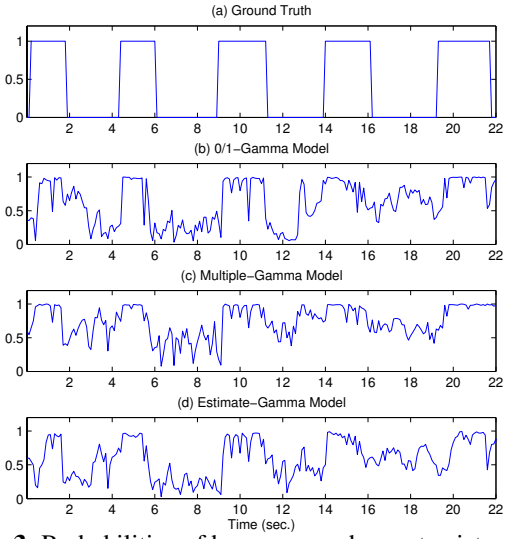

Fig. 3. Probabilities of human speech event existence.

models.

$$E_1 = \frac{\text{Number of undetected speech frames}}{\text{Number of speech frames}},$$

$$E_2 = \frac{\text{Number of misdetected non-speech frames}}{\text{Number of non-speech frames}},$$

and the total error rate are calculated for the time interval where both the human and the loudspeaker are in the camera's view area. E_1 is the error for detecting human speech, while E_2 is the false alarm.

Although the total error rate was almost same for all models, the E_1 decreased to 15% and 13% for the proposed models, while nearly 23% of speech frames were not detected with the conventional model. The reason is that the detection of speech segments with small power such as consonant portion was improved by the multiple-γ model or the estimated-γ model. On the other hand, the false alarm was increased for the proposed models. However, failure in detecting speech segments is considered to be more serious in our applications. Because once speech segments are not detected and discarded at this stage, there would be no chance of recovery at the later stage. In case of false alarm, to the contrary, there would still be a chance of recovery.

Table 1. Error rates for human speech event detection

Error Rate	E_1	E_2	$Total$
0/1-γ model	0.225	0.08	0.144
multiple-γ model	0.15	0.14	0.144
estimated-γ model	0.125	0.18	0.156

5. DISCUSSION AND CONCLUSIONS

A method of tracking human speech events by the fusion of audio and video information using a particle filter was proposed and evaluated. Using the particle filter, several sub-problems can be solved simultaneously and simply. We extended the work of Checka *et al.*[4] for detecting only human speech events. The two elaborated statistical models proposed herein for computing audio signal likelihood performed better than the simpler model.

The proposed framework is easily extended to incorporate other information, such as lip movement detection and human voice detection. In future studies, we intended to enable more robust and accurate results and built the proposed algorithm into various interactive systems, including the humanoid robot.

6. REFERENCES

[1] F. Asano et al., "Detection and separation of speech segment using audio and video information fusion," in *Proceedings of Eurospeech 2003*, 2003, pp. 2257–2260.

[2] H. Asoh et al., "An application of a particle filter to bayesian multiple sound source tracking with audio and video information fusion," in *Proceedings of Fusion 2004*, 2004.

[3] A. Doucet, N. Freitas, and N. Gordon, Eds., *Sequential Monte Carlo Methods in Practice*, Springer-Verlag, 2001.

[4] N. Checka, K. W. Wilson, M. R. Siracusa, and T. Darerll, "Multiple person and spaker activity tracking with a particle filter," in *Proceedings of ICASSP2004*, 2004.

[5] M. I. Miller and D. R. Fuhrmann, "Maximum-likelihood narrow-band direction finding and the em algorithm," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 38, pp. 1560–1577, 1990.

[6] I. Hara et al., "Robust speech interface based on audio and video information fusion for humanoid HRP-2," in *Proceedings of IROS 2004*, to be published.