# CONVERSATION CLUSTERING BASED ON PLCA USING WITHIN-CLUSTER SPARSITY CONSTRAINTS

*Yohei Kawaguchi and Masahito Togami*

Central Research Laboratory, Hitachi, Ltd.
Higashi-Koigakubo Kokubunji-shi, Tokyo 185-8601, Japan
{yohei.kawaguchi.xk, masahito.togami.fe}@hitachi.com

## ABSTRACT

We propose a new method for detecting separate conversations between people. In this paper, to model the rules of turn-taking in conversation, we introduce sparsity constraints of temporal activities within each cluster into the probabilistic latent component analysis (PLCA). The proposed method can detect conversation groups by using PLCA on the within-cluster sparsity constraints although the conventional PLCA has no effectiveness in clustering. Our method has two features: First, it can be applied to the cases that more than two speakers participate in the same group, for which the within-cluster sparsity constraints can be defined. Second, it has the practical advantage that it requires no training phase. Despite the lack of any training phase, experimental results indicate that the proposed method remains effective in scenarios where three speakers participate in the same group.

***Index Terms—*** conversation clustering, direction of arrival estimation, probabilistic latent component analysis, turn-taking, sparsity

## 1. INTRODUCTION

Automatic conversation analysis is an important technology to realize speech summarization and robots with communication capabilities. One of the main tasks in automatic conversation analysis is to detect groups of people that participate in the same conversation group under the condition that unplanned multi-groups exist simultaneously. Hereafter, this task is called "conversation clustering."

Several studies exist in the field of conversation clustering [1, 2, 3, 4, 5, 6, 7]. Assuming the condition that all the participants use wearable microphones, Nakakura *et al.* [1] focused on the well-known fact that participants of the same conversation group tend to be near one another, and they assumed that each participant's voice recorded by his/her microphone is louder than the recorded voices of other groups. Based on this assumption, clustering is performed by correlating amplitudes input at microphones. However, the assumption does not hold in the case that participants of different groups are near each other in typical office environments. In most of the studies, conversation groups were detected by focusing on the timing characteristic of utterances in conversation [2, 3, 4, 5]. The typical timing characteristic of utterances is the turn-taking rules that "minimize gap and overlap" between speakers. These studies on conversation clustering employ the mutual information (MI) of voice activity between speakers to model the turn-taking rules. However, these existing approaches have two problems: First, it is necessary to attach wearable microphones to each person in a conversation; second, these approaches cannot be applied in the case that more than two speakers participate in the same group, because MI can only be defined for two speakers. To solve the first problem, direction of arrival (DOA) estimation was combined with MI-based conversation clustering to create so-called "DOA-MI [6]." To solve the second problem, an extension of DOA-MI that matches voice activities with turn-taking of more than two speakers modeled by the Hidden Markov Model (DOA-HMM) was proposed [7]. DOA-HMM is effective in the case that more than two speakers belong to the same group. However, it has the disadvantage that the HMM requires training phases.

In the present study, a new conversation clustering method is proposed. This method has two key features: First, it is applicable in the case that more than two speakers participate in the same group; second, it requires no HMM training phase. To model the turn-taking rules in the cases of two speakers and more than two speakers, sparsity constraints of temporal activities within each cluster are introduced into the probabilistic latent component analysis (PLCA) [8, 9, 10]. The proposed method can detect conversation groups by using PLCA on the within-cluster sparsity constraints, although conventional PLCA has no effectiveness in clustering. It is thus called "DOA-PLCA" and is applicable in the case that a group has more than two speakers, because the within-cluster sparsity constraints can be defined for more than two speakers. Moreover, DOA-PLCA has no training phase because the parameter of the within-cluster sparsity is invariant to changes of speakers. Experimental results indicate that the method is effective in the case of a three-speaker group in spite of the fact that it has no training phase.

## 2. PROBLEM STATEMENTS AND NOTATION

It is assumed that $K$ speakers exist, and the set of the speakers is defined as $\mathcal{S} = \{1, \cdots, K\}$. The voices of the speakers are recorded at a microphone array that consists of $M$ microphones. The recorded signals are analog-to-digital converted and analyzed by the short-time Fourier transform. These multi-channel signals represented as $\boldsymbol{x}(f,\tau) = [x_1(f,\tau)\cdots x_M(f,\tau)]^T$, where $x_m(f,\tau)$ is the input signal of the $m$-th microphone, $f$ is the index of a frequency bin, and $\tau$ is the frame index. $\boldsymbol{x}(f,\tau)$ is modeled as follows:

$$\boldsymbol{x}(f,\tau) = \sum_{k=1}^{K} \boldsymbol{a}_k(f) s_k(f,\tau) + \boldsymbol{b}(f,\tau), \tag{1}$$

where $\boldsymbol{a}_k(f)$ is a complex vector that represents the impulse responses of the frequency domain for the $k$-th speaker, $s_k(f,\tau)$ is the source signal of the $k$-th speaker, and $\boldsymbol{b}(f,\tau)$ is background noise. Here, $\boldsymbol{a}_k(f)$ is the normalized vector such that $|\boldsymbol{a}_k(f)| = 1$.

Next, we estimate the direction of arrival (DOA) $\theta(f,\tau)$ in each $(f,\tau)$ by modified delay-and-sum beamformer (MDSBF) [11] as follows:

$$\hat{\theta}(f,\tau) = \arg\max_{\theta} \left| \boldsymbol{a}_\theta(f)^H \boldsymbol{x}(f,\tau) \right|^2, \tag{2}$$

where $\boldsymbol{a}_\theta(f)$ is the vector of the theoretical impulse responses for discrete direction $\theta$, superscript $H$ represents Hermitian transposition, and this vector can be calculated from the configuration of the microphones. A DOA histogram $H(\theta,\tau)$ is created by voting for direction $\hat{\theta}(f,\tau)$ as follows:

$$H(\theta,\tau) = \sum_f |\boldsymbol{a}_{\hat{\theta}(f,\tau)}(f)^H \boldsymbol{x}(f,\tau)|^2 \tag{3}$$

The goal of conversation clustering is to estimate the set of the directions of each speaker $\mathcal{D} = \{\theta_1 \cdots \theta_K\}$ and the set of the clusters $\mathcal{X} = \{\mathcal{C}_1, \cdots, \mathcal{C}_N\}$ from the given $H(\theta,\tau)$, where $N$ is the number of clusters, each cluster $\mathcal{C}_n$ is a disjoint subset of the set of speakers $\mathcal{S}$, where $n$ is the index of a cluster, and all the speakers in $\mathcal{C}_n$ participate in the conversation group that corresponds to $\mathcal{C}_n$. The clusters are called "conversation clusters" hereafter.

## 3. CONVENTIONAL METHODS

### 3.1. MI-based clustering

MI-based clustering (DOA-MI) [6] assumes the turn-taking rules that "minimize gap and overlap" between speakers, the concept of which was pioneered by Sacks [12]. The turn-taking rules can be interpreted that, in the frames when the voice of one speaker is active, those of the other speakers are inactive at a high probability. DOA-MI utilizes the mutual information to represent this inverse correlation of the voice activities.

First, this approach estimates the directions of each speaker $\hat{\theta}_k$. $\hat{\theta}_k$ can be calculated as the centroid of k-means clustering for $\theta$ weighted by $w(\theta) = \int H(\theta,\tau)d\tau$. Next, the voice activities of each speaker $v_k(\tau) = 0, 1$ in each frame $\tau$ are estimated by voice activity detection (VAD) for $H(\hat{\theta}_k,\tau)$. Then, conversation clusters are detected by agglomerative clustering for MI between speakers. MI between the $k$-th speaker and the $l$-th speaker, $\mu(k,l)$, is defined by Basu [3] as follows:

$$\begin{aligned}\mu(k,l) \quad = \quad & \sum_{b_k, b_l \in \{0,1\}} P(v_k = b_k, v_l = b_l) \\ & \times \log \frac{P(v_k = b_k, v_l = b_l)}{P(v_k = b_k) P(v_l = b_l)},\end{aligned} \tag{4}$$

where

$$P(v_k = b_k, v_l = b_l) = \begin{cases} \frac{1}{T}\sum_{\tau=1}^{T} v_k v_l & \text{if} \quad (b_k, b_l) = (1,1), \\ \frac{1}{T}\sum_{\tau=1}^{T} v_k (1 - v_l) & \text{if} \quad (b_k, b_l) = (1,0), \\ \frac{1}{T}\sum_{\tau=1}^{T} (1 - v_k) v_l & \text{if} \quad (b_k, b_l) = (0,1), \\ \frac{1}{T}\sum_{\tau=1}^{T} (1 - v_k)(1 - v_l) & \text{if} \quad (b_k, b_l) = (0,0), \end{cases}$$

$$P(v_k = b_k) = \begin{cases} \frac{1}{T}\sum_{\tau=1}^{T} v_k & \text{if} \quad b_k = 1, \\ \frac{1}{T}\sum_{\tau=1}^{T} (1 - v_k) & \text{if} \quad b_k = 0, \end{cases}$$

and $T$ is the number of the frames.

DOA-MI is effective for many cases, but cannot be applied in the case that more than two speakers belong to the same cluster because MI can be defined only for two speakers.

### 3.2. HMM-based approach

The HMM-based approach (DOA-HMM) [7] is an extension of DOA-MI to more than two speakers. DOA-HMM models turn-taking within groups as the HMM for each number of speakers that belongs to the same cluster. It detects conversation clusters by matching the HMM to voice activities of combinations of speakers in place of clustering for MI.

Here, we assume $L$ speakers $(p_1, \cdots, p_L)$ participate in a conversation cluster $\mathcal{C}$. We introduce the HMM that models turn-taking within $\mathcal{C} = \{p_1, \cdots, p_L\}$ as follows: The number of states of the HMM is $L$, each state of the HMM represents that the corresponding speaker has a turn, and the HMM outputs the observation symbol $\boldsymbol{V}_\mathcal{C}(\tau) = [v_{p_1}(\tau) \cdots v_{p_L}(\tau)]$ in frame $\tau$, where $v_{p_i}(\tau)$ represents the voice activity of speaker $p_i$. Now, we can observe the voice activities of all the speakers $\boldsymbol{V}(\tau) = [v_1(\tau) \cdots v_K(\tau)]$. Therefore, we can formulate the problem as the maximization of the probability distribution that the combination of the HMM generates the sequence

of observation symbols $\boldsymbol{V}(1), \cdots, \boldsymbol{V}(T)$ as follows:

$$\begin{aligned}\hat{\mathcal{X}} &= \arg\max_{\mathcal{X}} P(\boldsymbol{V}(1), \cdots, \boldsymbol{V}(T)|\mathcal{X}) \\ &= \arg\max_{\mathcal{X}} \prod_{n=1}^{N} P(\boldsymbol{V}_{\mathcal{C}_n}(1), \cdots, \boldsymbol{V}_{\mathcal{C}_n}(T)|\mathcal{C}_n). \quad (5)\end{aligned}$$

$P(\boldsymbol{V}_{\mathcal{C}_n}(1), \cdots, \boldsymbol{V}_{\mathcal{C}_n}(T)|\mathcal{C}_n)$ can be calculated from the HMM defined above. DOA-HMM is effective in the case that more than two speakers take part in the same group. However, it has the disadvantage that the HMM needs training phases.

## 4. CONVERSATION CLUSTERING BASED ON PLCA

### 4.1. PLCA model

A new conversation-clustering method, called DOA-PLCA, which solves the disadvantage of DOA-HMM, is proposed in the following. The turn-taking rules in conversation are modeled as sparsity constraints of temporal activities within each cluster in PLCA [8][9]. DOA-PLCA detects conversation clusters by using PLCA on the within-cluster sparsity constraints. It can be applied in the case that more than two speakers participate in the same group because the within-cluster sparsity constraints can be defined for more than two speakers. Furthermore, it has no training phase because the parameter of the within-cluster sparsity is invariant to changes of speakers.

The observed DOA histogram $H(\theta, \tau)$ can be modeled as a linear combination of non-negative basis components that correspond to speakers, where the voice activity of each speaker is generated probabilistically in each frame, and the mixing weights increase in the frame when the voice of corresponding speaker is active. The generation model of PLCA [8] is such a probabilistic non-negative mixing model. The generation process of $H(\theta, \tau)$ was thus modeled by using the PLCA model as follows:

$$P(H(\theta,\tau)\forall\theta,\tau) = \prod_{\tau}\prod_{\theta}\left\{\sum_{k=1}^{K} P_\tau(k)P(\theta|k)\right\}^{H(\theta,\tau)}, \quad (6)$$

where $P_\tau(k)$ is the probability that the voice of the $k$-th speaker is active in frame $\tau$, and $P(\theta|k)$ represents the probability distribution that the voice activity of the $k$-th speaker votes at $\theta$ in the DOA histogram. $P_\tau(k)$ is called the "probabilistic activity," and $P(\theta|k)$ is called the "probabilistic basis component." Equation (6) leads to the log-likelihood

$$\log P(H(\theta,\tau)\forall\theta,\tau) = \sum_{\tau}\sum_{\theta} H(\theta,\tau)\log\sum_{k=1}^{K} P_\tau(k)P(\theta|k). \quad (7)$$

$P_\tau(k)$ and $P(\theta|k)$ that maximize Eq. (7) can be calculated by the expectation-maximization (EM) algorithm similarly to the conventional PLCA proposed by Raj [8]. However, these estimates of $P_\tau(k)$ and $P(\theta|k)$ are not the solutions of the conversation clustering problem. To detect conversation clusters, "within-cluster sparsity constraints" that model the turn-taking rules in Section 4.2 are introduced in the following.

### 4.2. Solution of PLCA by using within-cluster sparsity constraints

The within-cluster sparsity constraints model the turn-taking rules, which "minimize gap and overlap" between speakers within the same conversation group. The constraints represent that the voice of only one speaker is active in every frame at a high probability in the same conversation cluster. The aim of using the constraints is to correspond the estimates of $P(\theta|k)$ to the indices of speakers of each conversation cluster.

To use sparsity constraints of the whole probabilistic basis component, Shashanka [9] introduced "entropic priors" into PLCA. PLCA has the advantage that it makes it possible to use a priori knowledge of domains like these methods. We also introduce the entropic priors to represent the within-cluster sparsity constraints and solve the clustering problem. The objective function is defined by adding the term of entropic priors to Eq. (7) as follows:

$$\begin{aligned}&J\left(\{P_\tau(k)\}, \{P(\theta|k)\}\right) \\ &= \sum_{\tau}\sum_{\theta} H(\theta,\tau)\log\sum_{k=1}^{K} P_\tau(k)P(\theta|k) \\ &\quad -\beta\sum_{n}\sum_{\tau} E(\{P_\tau(k)\}_{k\in\mathcal{C}_n}), \quad (8)\end{aligned}$$

where $\beta$ is the parameter of the within-cluster sparsity of $P_\tau(k)$, and $E(\{P_\tau(k)\}_{k\in\mathcal{C}_n})$ is the $\alpha$-order Renyi's entropy defined as $E(\{P_\tau(k)\}_{k\in\mathcal{C}_n}) = \frac{1}{1-\alpha}\log\sum_{k\in\mathcal{C}_n} P_\tau(k)^\alpha$. The second term of Eq. (8) corresponds to the within-cluster sparsity of $P_t(k)$. Equation (8) has two notable features. One is that the within-cluster sparsity can be defined in the case that more than two speakers belong to each cluster $\mathcal{C}_n$. DOA-PLCA is thus applicable to the case that more than two speakers participate in the same group. The other feature is that the within-cluster sparsity has only one a priori parameter, $\beta$. Unlike the state transition and emission probabilities of the HMM, $\beta$ is invariant to changes of speakers, and we can use $\beta$ that is tuned once for different scenes. Therefore, DOA-PLCA need no training phase.

By maximizing the objective function $J\left(\{P_\tau(k)\}, \{P(\theta|k)\}\right)$ in Eq. (8), the following EM algorithm is obtained to estimate $P_\tau(k)$ and $P(\theta|k)$:

**E step:**

$$P_\tau(k|\theta) = \frac{P_\tau(k)P(\theta|k)}{\sum_{k'=1}^{K} P_\tau(k')P(\theta|k')}, \quad (9)$$

**M step:**

$$P_\tau(k) = g(\beta, \sum_{\theta} H(\tau,\theta)P_\tau(k|\theta)), \quad (10)$$

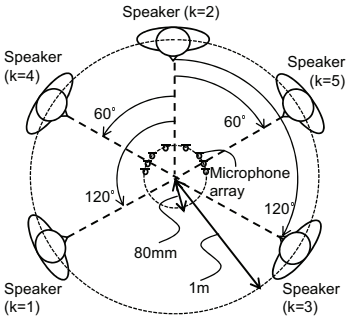


**Fig. 1**. Experimental setup.

$$P(\theta|k) = \frac{\sum_{\tau=1}^{T} H(\tau,\theta) P_\tau(k|\theta)}{\sum_{k'=1}^{K}\sum_{\tau=1}^{T} H(\tau,\theta) P_\tau(k'|\theta)}, \qquad (11)$$

where $g(\beta,\gamma_k)$ is the $\alpha$-order Renyi's entropic prior, which can be calculated by an iteration process as follows:

1. $h(k) = \beta\gamma_k + \dfrac{\alpha}{\alpha-1}\dfrac{g(\beta,\gamma_k)^\alpha}{\sum_{k'\in\mathcal{C}_n s.t. k\in\mathcal{C}_n} g(\beta,\gamma_{k'})^\alpha}$

2. $g(\beta,\gamma_k) = \dfrac{h(k)}{\sum_{k'\in\mathcal{C}_n s.t. k\in\mathcal{C}_n} h(k')}$

3. Return to 1 until convergence.

The difference between the above estimation process and the conventional PLCA [9] is that the areas of the sparsity constraints are limited to within each cluster. The conventional PLCA has no effectiveness in clustering. However, the within-cluster sparsity constraints enable PLCA to perform clustering. This clustering process can detect correspondences between the indices of speakers $k$ and the basis components $P(\theta|k')$ such that the temporal activities of speakers in the same cluster follow the turn-taking rules.

## 5. EXPERIMENTAL RESULTS

The performance of the proposed method was evaluated as follows. Five speakers and a micropone array were configured as shown in Fig. 1. The microphone array consists of eight microphones configured in a semicircle with radius of 80 mm. The reverberation time $RT_{60}$ is 310 milliseconds. Casual conversations between the speakers were recorded at 8 kHz sampling rate and 16 bit-per-sample. The conversations were recorded under the following two conditions:

**Condition 1:** Three-speaker conversation $\mathcal{C}_1 = \{k=1,2,3\}$ and two-speaker conversation $\mathcal{C}_2 = \{k=4,5\}$.

**Condition 2:** Two-speaker conversation $\mathcal{C}_1 = \{k=1,2\}$ and another two-speaker conversation $\mathcal{C}_2 = \{k=4,5\}$.

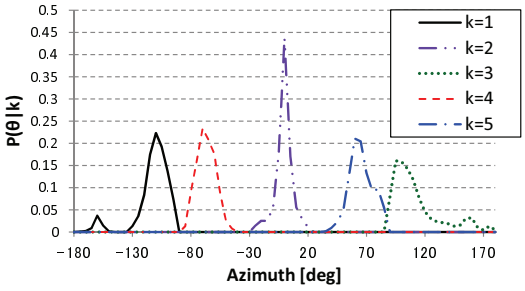


**Fig. 2**. Example of estimates of the basis component $P(\theta|k)$ for Condition 1 ($\mathcal{C}_1 = \{k=1,2,3\}$ and $\mathcal{C}_2 = \{k=4,5\}$). X and Y axis show the azimuth $\theta$ and $P(\theta|k)$. Each line represents the corresponding speaker.

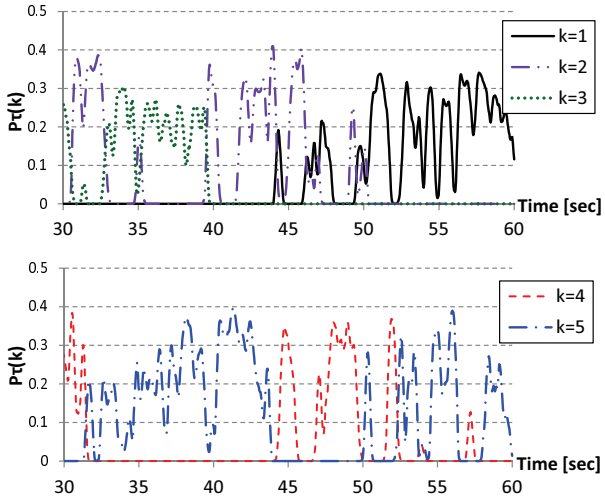


**Fig. 3**. Example of estimates of activity $P_\tau(k)$ for condition 1 ($\mathcal{C}_1 = \{k=1,2,3\}$ and $\mathcal{C}_2 = \{k=4,5\}$). Top: $\mathcal{C}_1$. Bottom: $\mathcal{C}_2$. X and Y axis show frame $\tau$ [sec] and $P_\tau(k)$. Each line represents the corresponding speaker.

32 sessions and 7 sessions were recorded under condition 1 and condition 2 respectively. The length of each session was 60 seconds.

Figure 2 and 3 illustrate an example of the estimation results. As shown in Fig. 2, all the correct directions of the speakers were estimated, and the speakers were corresponded to the correct clusters. According to Fig. 3, the estimated activities followed the turn-taking rules within clusters. These results indicate that the proposed method works well. Next, the accuracy of the proposed method was compared with that of existing methods. The accuracy is calculated by Accuracy $= \frac{N_B}{N_A}$, where $N_A$ is the total number of sessions, and $N_B$ is the number of sessions in which all the speakers are clustered into the correct clusters. "DOA-MI" and "DOA-HMM" represents the conventional methods based on mutual

**Table 1**. Accuracy by using DOA-MI, DOA-HMM, DOA-PLCA without WCSC, and the proposed method (DOA-PLCA with WCSC) for each condition.

| Method | Condition 1 | Condition 2 |
|---|---|---|
| DOA-MI | 0.53 (17/32) | 0.86 (6/7) |
| DOA-HMM | 0.88 (28/32) | 0.86 (6/7) |
| DOA-PLCA without WCSC | 0 (0/32) | 0.29 (2/7) |
| DOA-PLCA with WCSC | 0.81 (26/32) | 0.86 (6/7) |

information [6] and the HMM [7], respectively. "DOA-PLCA with within-cluster sparsity constraints (WCSC)" represents the proposed method. "DOA-PLCA without WCSC" is a version of the proposed method in which the number of clusters is set to one. This version corresponds to conventional PLCA. The HMM parameter set of DOA-HMM was trained by using the Baum-Welch algorithm [13]. The training data were all the data except the test session. In DOA-PLCA with WCSC, the number of clusters was set as $N = 2$. The accuracy of each method is listed in Table 1. Under condition 1 (three-speaker group), the accuracy of DOA-MI is much lower than that of the other methods. This is due to the fact that mutual information is defined only for two speakers. On the other hand, the accuracy of the proposed method is over 80%, and this performance is comparable to that of DOA-HMM, which needs a training phase. As Table 1 shows, DOA-PLCA without WCSC corresponding to the conventional PLCA has no effectiveness in clustering. This result shows that the within-cluster sparsity constraints enable PLCA to perform clustering.

These results indicate that the within-cluster sparsity of the proposed method can model the turn-taking rules in three-speaker groups and that the method is effective in the case that there is a three-speaker group. In addition, the proposed method needs no training phase (unlike DOA-HMM).

## 6. CONCLUSION

A new method for conversation clustering in the case that more than two speakers participate in the same group and in the case that it is difficult to perform training was proposed. To model the turn-taking rules in conversation in the cases of two speakers and more than two speakers, sparsity constraints of temporal activities within each cluster were introduced into PLCA. The proposed method can detect conversation groups by using PLCA on the within-cluster sparsity constraints although conventional PLCA has no effectiveness in clustering. Experimental results indicate that the method is effective in the case of a three-speaker group in spite of the fact that it needs no training phase.

## 7. REFERENCES


[1] T. Nakakura, Y. Sumi, and T. Nishida, "Neary: Conversation field detection based on similarity of auditory situation," in *HotMobile 2009*, 2009.

[2] D. Wyatt, T. Choudhury, and J. Bilmes, "Conversation detection and speaker segmentation in privacy-sensitive situated speech data," in *Interspeech 2007*, 2007.

[3] S. Basu, *Conversational scene analysis*, Ph.D. thesis, MIT, 2002.

[4] T. Choudhury, *Sensing and modeling human networks*, Ph.D. thesis, MIT, 2004.

[5] A. Härmä and K. Pham, "Conversation detection in ambient telephony," in *ICASSP 2009*, 2009.

[6] Y. Kawaguchi, M. Togami, and Y. Obuchi, "Turn taking-based conversation detection by using DOA estimation," in *Interspeech 2010*, 2010.

[7] Y. Kawaguchi and M. Togami, "Conversation clustering by matching models of turn-taking," *IEICE Transactions on fundamentals of electronics communications and computer sciences (Japanese Edition)*, vol. J95-A, no. 2, pp. 217–221, February 2012.

[8] B. Raj, R. Singh, M. Shashanka, and P. Smaragdis, "Bandwidth expansion with a Pólya urn model," in *ICASSP 2007*, 2007.

[9] M. Shashanka, B. Raj, and P. Smaragdis, "Sparse overcomplete latent variable decomposition of counts data," in *NIPS 2007*, 2007.

[10] P. Smaragdis, M. Shashanka, B. Raj, and G.J. Mysore, "Probabilistic factorization of non-negative data with entropic co-occurrence constraints," in *ICA 2009*, 2009.

[11] M. Togami, Y. Obuchi, and A. Amano, "Automatic speech recognition of human-symbiotic robot EMIEW," in *Human Robot Interaction*, N. Sarkar, Ed., pp. 395–404. I-tech Education and Publishing, 2007.

[12] H. Sacks, E. Schegloff, and G. Jefferson, "A simplest systematics for the organization of turn-taking for conversation," *Language*, vol. 50, no. 4, pp. 696–735, 1974.

[13] L.E. Baum, T. Petrie, G. Soules, and N. Weiss, "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains," *The Annals of Mathematical Statistics*, vol. 41, no. 1, pp. 164–171, 1970.