# Influence Relation Estimation based on Lexical Entrainment in Conversation

Tomoharu Iwata[1], Shinji Watanabe[1]

**Abstract**

In conversations, people tend to mimic their companions' behavior depending on their level of trust. This phenomenon is known as entrainment. We propose a probabilistic model for estimating influences among speakers from conversation data with multiple people by modeling lexical entrainment. The proposed model estimates the word use as a function of weighted sum of earlier word use of other speakers. The weights represent influences between speakers. The influences can be efficiently estimated by using the expectation maximization (EM) algorithm. We also develop its online inference procedures for sequentially modeling the dynamics of influence relations. Experiments on two meeting data sets in Japanese and in English demonstrate the effectiveness of the proposed method.

*Keywords:* conversation analysis, influence, latent variable model, entrainment

## 1. Introduction

In conversations, people tend to mimic such aspects of their companions' behavior as postures [1], facial expressions [2], lexicon [3, 4], syntax [5], acoustic, prosodic [6] and amplitude [7]. This phenomenon is known in the literature as entrainment, accommodation, adaptation, or alignment [8]. Entrainment is said to indicate that people are trusting, accommodating and empathic [1, 9].

*Email address:* iwata.tomoharu@lab.ntt.co.jp (Tomoharu Iwata)
*URL:* http://www.kecl.ntt.co.jp/as/members/iwata/index.htm (Tomoharu Iwata)
[1]NTT Communication Science Laboratories

This paper focuses on the entrainment of lexicon in polylogue, or how people are influenced by their companions in terms of word use in conversation with multiple speakers. The degree to which a person exerts an influence and is influenced by others varies from speaker to speaker. A powerful person is likely to be mimicked by others, and a passive person might often be accommodating to others. The influences also differ between pairs depending on their level of trust. For example, Alice might use words spoken by Bob, but not words spoken by Charlie. The influences therefore have an asymmetric nature.

We propose a simple and effective probabilistic model for estimating influence relations among speakers from conversation data with multiple people [10]. With the proposed model, we assume that a speaker's word use (word distribution) depends on the preceding word use of other speakers as well as his/her own preceding word use and the general word distribution. We estimate the strength of influence for each pair of speakers using the expectation maximization (EM) algorithm [11]. We also develop online inference procedures for sequentially modeling the dynamics of influence relations. Note that the proposed model estimates influences on the word use of a speaker from the word use of other speakers.

The remainder of this paper is organized as follows. In Section 2, we briefly review related work. In Section 3, we formulate our proposed probabilistic model for influence estimation. In Section 4, we describe the inference procedures for the proposed model. In Section 5, we extend the inference with online version for efficiency. In Section 6, we demonstrate the effectiveness of the proposed method by analyzing two meeting data sets in Japanese and in English. Finally, we present concluding remarks and a discussion of future work in Section 7.

## 2. Related Work

In recent years, a huge amount of conversation data have been accumulated due to the improvement of recoding devices and automatic speech recognition systems, and there has been great interest in the analysis of conversation [12, 13]. For example, [4] investigated the correlation between task success and similarity of word use, and [9] analyzed the relationship between social game results and word repetition. However, they focused on dyadic conversations, and did not consider the asymmetric nature of influences. On the other hand, we deal with the conversation of multiple people, in which

influence and sensitiveness are assumed to depend on the pair of speakers. In addition, since the proposed method is a probabilistic generative model for conversations, we can efficiently estimate inferences in a principled statistical framework, and use it for a language model of the conversation.

The proposed method is related to speaker role recognition [14, 15], in which each speaker is automatically classified into a role category, because influences depend on the speaker's role. In the role recognition, roles are predefined, and classifiers are trained using labeled data. On the other hand, the proposed method directly estimate influences, and it does not require labeled data.

A number of language models for conversation have been proposed [16, 17]. However, they do not aim to estimate influences between speakers. By using online social network data, [18] proposed a probabilistic model for estimating influences in online behavior, but this is not applied to conversation data.

## 3. Proposed Model

With the proposed model, we assume that the word use of a speaker changes depending on the preceding word use of other speakers as well as the own preceding word use. Figure 1 shows the dynamics of the word use of each speaker in the proposed model, where $\lambda$ represents the weight of the influence.

Let $\boldsymbol{w} = \{w_1, \cdots, w_t, \cdots\}$ be a word sequence of a polylogue, where $w_t$ represents the $t$th word, and let $\boldsymbol{s} = \{s_1, \cdots, s_t, \cdots\}$ be its speaker sequence, where $s_t$ indicates the speaker of the $t$th word. Here, $w_t \in \{1, \cdots, W\}$ and $s_t \in \{1, \cdots, M\}$, where $W$ is the vocabulary size, and $M$ is the number of participants.

In the proposed model, we assume that a speaker's word use depends on the preceding word use of other speakers. The preceding word use of speaker $m$ at position $t$ can be modeled as follows:

$$P_{\mathrm{C}}(w|\boldsymbol{w}_{t-\tau}^{t-1}, m) = \frac{N(t-1, \tau, w, m) + \beta}{\sum_{w'} N(t-1, \tau, w', m) + \beta W}, \tag{1}$$

where $\tau$ represents the period of the influence, $\beta$ is a smoothing parameter, and $N(t-1, \tau, w, m)$ represents the count of word $w$ that is spoken by speaker $m$ from $t-\tau$ to $t-1$. This probability is proportional to the number of times

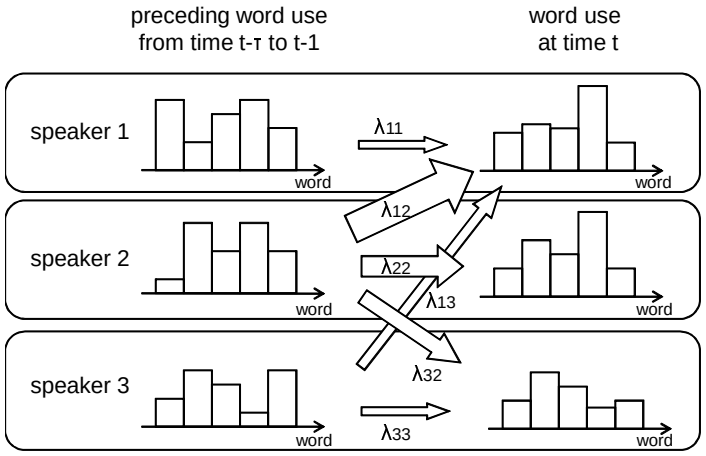



Figure 1: The word use of a speaker changes depending on the preceding word use of other speakers as well as the own preceding word use.

word $w$ is used by speaker $m$ in the preceding period $\tau$. The smoothing parameter $\beta$ is introduced to avoid the zero probability problem.

The word use of speaker $n$ at position $t$ is then modeled by a mixture of the preceding word use of the participants as follows:

$$P(w|\boldsymbol{w}_{t-\tau}^{t-1}, n) = \sum_{m=1}^{M} \lambda_{nm} P_{\mathrm{C}}(w|\boldsymbol{w}_{t-\tau}^{t-1}, m) + (1 - \sum_{m=1}^{M} \lambda_{nm}) P_{\mathrm{G}}(w), \qquad (2)$$

where $\lambda_{nm}$ represents the influence of speaker $m$ on speaker $n$, $0 \leq \lambda_{nm} \leq 1$, and $P_{\mathrm{G}}(w)$ is the general word distribution, which does not depend on the preceding conversation. The general word distribution can be obtained by using other corpora. Speaker $m$ who influences the word distribution of speaker $n$ is not observed, and $m$ is a latent variable.

This proposed model is an extension of cache models [19] for multi-speaker conversations. The cache-based language model integrates short-term patterns of word use into the word distribution by means of a cache component. With the proposed model, we build speaker-specific cache components, and set different influences among pairs of speakers. The proposed model can be easily extended for $n$-gram language models by taking a $n$-gram sequence as input instead of unigram sequence $\boldsymbol{w}$. The other approaches to incorporate sequential information include unigram rescaling [20] and linear interpolation of the proposed model with $n$-gram language models.

## 4. Inference

We estimate the influences $\lambda_{nm}$ based on maximum posterior (MAP) estimation. For simplicity, we rewrite the proposed word distribution in (2) as follows:

$$P(w|\boldsymbol{w}_{t-\tau}^{t-1}, n) = \sum_{m=0}^{M} \lambda_{nm} P_{\mathrm{C}}(w|\boldsymbol{w}_{t-\tau}^{t-1}, m), \tag{3}$$

where we set $P_{\mathrm{C}}(w|\boldsymbol{w}_{t-\tau}^{t-1}, m=0) \equiv P_{\mathrm{G}}(w)$, $\lambda_{n0} \equiv 1 - \sum_{m=1}^{M} \lambda_{nm}$, in which $\lambda_{nm} \geq 0$ and $\sum_{m=0}^{M} \lambda_{nm} = 1$. With this notation, the logarithm of the posterior probability of parameters given the conversation data $\{\boldsymbol{w}_{t=1}^{T}, \boldsymbol{s}_{t=1}^{T}\}$, which is to be maximized, is calculated as follows,

$$L = \sum_{t=1}^{T} \log \sum_{m=0}^{M} \lambda_{s_t m} P_{\mathrm{C}}(w_t|\boldsymbol{w}_{t-\tau}^{t-1}, m) + \sum_{n=1}^{M} \log P(\boldsymbol{\lambda}_n|\alpha), \tag{4}$$

where $T$ is the current position, and the second term represents the prior probability for parameters $\boldsymbol{\lambda}_n = \{\lambda_{nm}\}_{m=0}^{M}$. We use the following Dirichlet prior with hyperparameter $\alpha$:

$$\log P(\boldsymbol{\lambda}_n|\alpha) = \sum_{m=0}^{M} \alpha \log \lambda_{nm}, \tag{5}$$

because it is conjugate to multinomial parameters $\boldsymbol{\lambda}_n$. The inference is made more robust by introducing the priors.

We can efficiently maximize the posterior (4) by using the EM algorithm [11]. The EM algorithm is commonly used for the inference of mixture models. Because the proposed model is a mixture model, the standard EM algorithm can be used for the proposed model. The conditional expectation of the complete-data log likelihood with priors is represented as follows:

$$Q = \sum_{t=1}^{T} \sum_{m=0}^{M} P(m|t) \log \lambda_{s_t m} P_{\mathrm{C}}(w_t|\boldsymbol{w}_{t-\tau}^{t-1}, m) + \sum_{n=1}^{M} \sum_{m=0}^{M} \alpha \log \lambda_{nm}, \tag{6}$$

where $P(m|t)$ is the posterior probability of selecting speaker $m$ given the $t$th word. The derivation of $Q$ is described in Appendix *Appendix A*. It indicates the probability that the $t$th word is influenced by speaker $m$. In the E-step, we compute the probability according to the Bayes rule:

$$P(m|t) = \frac{\lambda_{s_t m} P_{\mathrm{C}}(w_t|\boldsymbol{w}_{t-\tau}^{t-1}, m)}{\sum_{m'=0}^{M} \lambda_{s_t m'} P_{\mathrm{C}}(w_t|\boldsymbol{w}_{t-\tau}^{t-1}, m')}, \tag{7}$$

where $\lambda_{nm}$ is the prior term, and $P_{\mathrm{C}}(w_t|\boldsymbol{w}_{t-\tau}^{t-1}, m)$ is the likelihood term. In the M-step, we obtain the next estimate of influences $\boldsymbol{\lambda}_n$ by maximizing $Q$ w.r.t. $\boldsymbol{\lambda}_n$ subject to $\sum_{m=0}^{M} \lambda_{nm} = 1$:

$$\lambda_{nm} = \frac{\sum_{t=1}^{T} I(s_t = n) P(m|t) + \alpha}{\sum_{m'=0}^{M} \sum_{t=1}^{T} I(s_t = n) P(m'|t) + \alpha(M+1)}, \tag{8}$$

where $I(A)$ represents an indicator function, i.e. $I(A) = 1$ if $A$ is true, $I(A) = 0$ otherwise. Note that the speaker dependent preceding word distribution $P_{\mathrm{C}}(w_t|\boldsymbol{w}_{t-\tau}^{t-1}, m)$ can be calculated by (1) independent of estimating parameters $\boldsymbol{\Lambda} = \{\boldsymbol{\lambda}_n\}_{n=1}^{M}$. The general word distribution $P_{\mathrm{C}}(w_t|\boldsymbol{w}_{t-\tau}^{t-1}, m = 0)$ is assumed to be given in advance. By iterating the E-step and the M-step until convergence, we obtain a local optimum solution for influences $\boldsymbol{\Lambda}$.

## 5. Online Inference

In the previous section, we described the inference procedures using all the data in a conversation session. However, in some real applications, we might need to estimate influences in the middle of the conversation session. Then, we develop online inference procedures for the proposed model for sequentially estimating influences based on the online EM algorithm [21, 22]. With the online inference, the proposed model is sequentially updated using newly obtained data. This means that past conversation data are not required to make the inference, and we can reduce the memory requirement as well as computational time.

With the online inference, sufficient statistics are updated from previous values using newly obtained data and the current estimated model as follows,

$$\bar{\lambda}_{nm}^{(t+1)} = \gamma \bar{\lambda}_{nm}^{(t)} + I(s_t = n) P(m|t), \tag{9}$$

where $\bar{\lambda}_{nm}^{(t)}$ is a sufficient statistic at $t$ and $\gamma$ is the forgetting factor. $P(m|t)$ can be calculated by (7). Using the sufficient statistics $\bar{\lambda}_{nm}^{(t)}$, we can obtain the estimate of influence as follows,

$$\lambda_{nm}^{(t)} = \frac{\bar{\lambda}_{nm}^{(t)}}{\sum_{m'=0}^{M} \bar{\lambda}_{nm'}^{(t)}}. \tag{10}$$

The forgetting factor $\gamma$ represents how likely influences change over time. By controlling $\gamma$, we can model the dynamics of influence relations. As an initial value for the sufficient statistic, we can use $\bar{\lambda}_{nm}^{(0)} = \alpha$.

We can also update the preceding word use in an online fashion. The preceding word use of speaker $m$ at position $t$ of (1) can be rewritten as follows,

$$P_{\mathrm{C}}(w|\boldsymbol{w}_{t-\tau}^{t-1}, m) = \frac{N(t-1,\tau,w,m)+\beta}{\sum_{w'} N(t-1,\tau,w',m)+\beta W}. \tag{11}$$

This number $N(t,\tau,w,m)$ can be calculated using the preceding number $N(t-1,\tau,w,m)$ and data at $t$ as follows,

$$N(t,\tau,w,m) = N(t-1,\tau,w,m)+I(w_t = w \wedge s_t = m)-I(w_{t-\tau} = w \wedge s_{t-\tau} = m). \tag{12}$$

By updating the model using (10) and (12), we can estimate influence relations sequentially.

## 6. Experimental Results

We demonstrate the importance of modeling influence in conversation data with experiments. First, we visualize the influence relations inferred by the proposed method. Second, we analyze influential and sensitive speakers, and show it can be used to identify a chairperson in the conversation. Third, we compare the proposed method with other methods quantitatively using perplexity, which represents the performance of predicting word use. We also analyze perplexities with different numbers of training utterances, with different lengths of effective period, and with different forgetting factors in online inference.

We evaluated the proposed method using the following two real meeting transcription data sets: NTT [23] and RT07 [24]. The NTT data set consists of six sessions in Japanese. In each meeting, one participant mainly talked about a technical topic using slides, and the other participants asked questions spontaneously. RT07 data set is an English corpus of eight conference room meetings, which consists of primarily goal-oriented, decision-making exercises and can vary from moderated meetings to group consensus-building meetings. Four sites contributed two meeting recordings for eight total meetings. For both of the data sets, speakers in a session are different from other sessions. Table 1 shows a summary of the NTT and RT07 data sets, and includes the number of sessions, vocabulary size, and the minimum and maximum number of speakers and utterances for a session. With the proposed model, we used $\alpha = 1$ and $\beta = 10^{-8}$ for the hyperparameters, and modeled the preceding word use by using all preceding utterances in the session,

Table 1: Summary of NTT and RT07 meeting data sets.

| | #session | #speakers | | #utterance | | #vocabulary |
|---|---|---|---|---|---|---|
| | | min | max | min | max | |
| NTT | 6 | 4 | 4 | 560 | 918 | 2,098 |
| RT07 | 8 | 4 | 6 | 337 | 749 | 3,113 |

or $\tau = \infty$. The general word distribution $P_{\mathrm{G}}(w)$ is learned by using other sessions in each data set.

We estimated the influences between speakers using the proposed model. Figure 2 shows the result. Each node represents a speaker, and the width of the arrow represents the strength of the influence, where only influences with $\lambda_{nm} \geq 0.1$ are shown. The self influence is generally strong, which indicates that the word use depends strongly on the speaker's own preceding word use. This is an intuitive result. There are also many influences between speakers. Some speakers are influential, e.g. speaker 1 in Session 7 in RT07, and some speakers are sensitive to other speakers, e.g. speaker 2 in Session 6 in RT07. Most of the influences are asymmetric. This result indicates that it is important to model the direction of the influences.

In all the NTT sessions, speaker 4 was appointed chairperson, and therefore, speaker 4 was influential and not sensitive. The result obtained with the proposed model reveals the influential and non-sensitive characteristics of speaker 4 as shown in Figure 2 (a), where there are more than three arrows from speaker 4 in five out of six sessions, and there is no arrow pointing to speaker 4 from others in all the sessions. Figure 3 shows its quantitative analysis. The influence to other speakers of speaker $n$ in Figures 3 (a) is calculated by summing up the influences to others $\sum_{m \neq n} \lambda_{mn}$. In the same way, the influence from other speakers of speaker $n$ in Figures 3 (b) is calculated by $\sum_{m \neq n} \lambda_{nm}$. The influences to others of speaker 4 are the highest in five sessions, and the influences from others of speaker 4 are the lowest in all six sessions. This result represents influential and non-sensitive characteristics of speaker 4. In this way, the proposed model can use conversation data to analyze the influences between speakers.

For a quantitative evaluation, we compared the following six models:

- **CC** has a common cache that is shared by all speakers, and a common parameter that control the influence of the common cache. The word

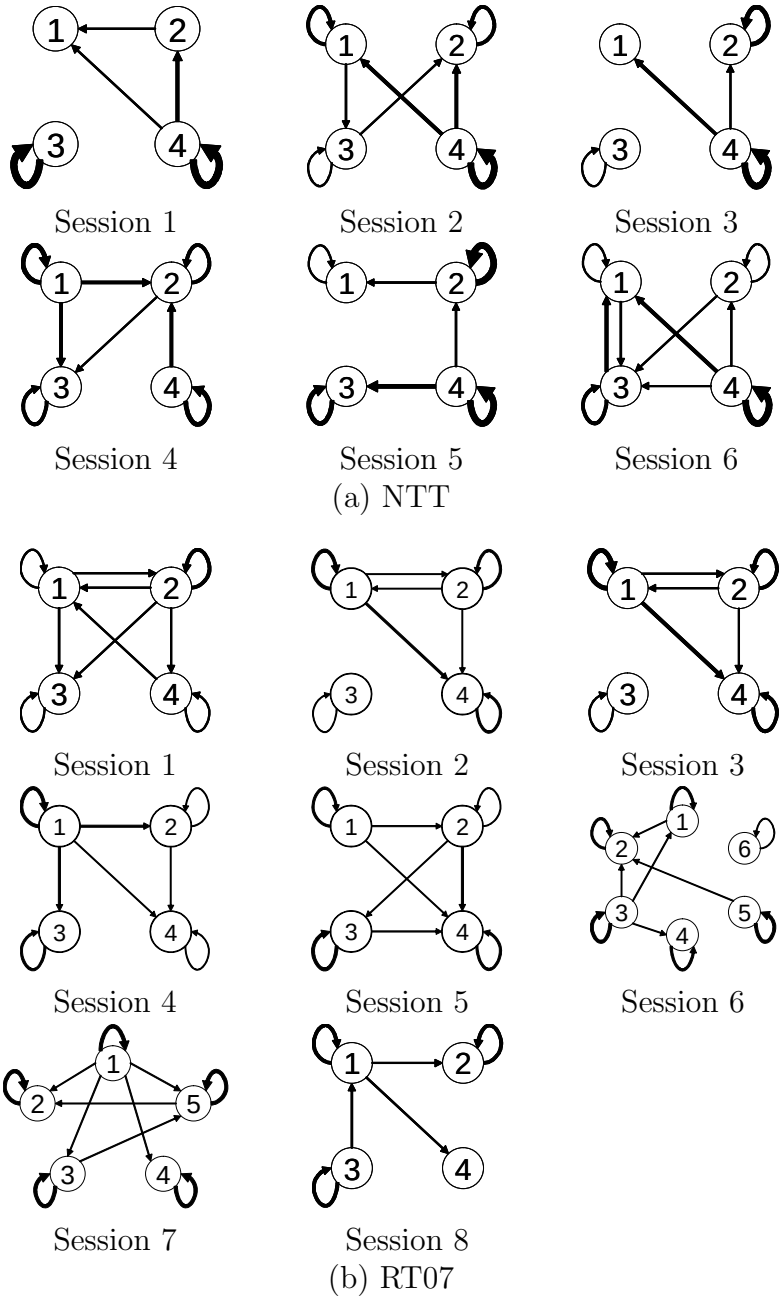
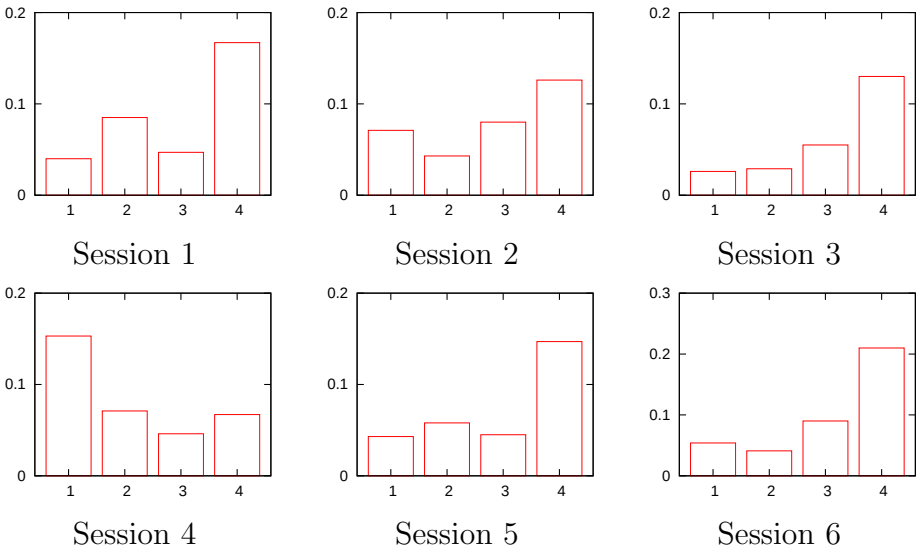


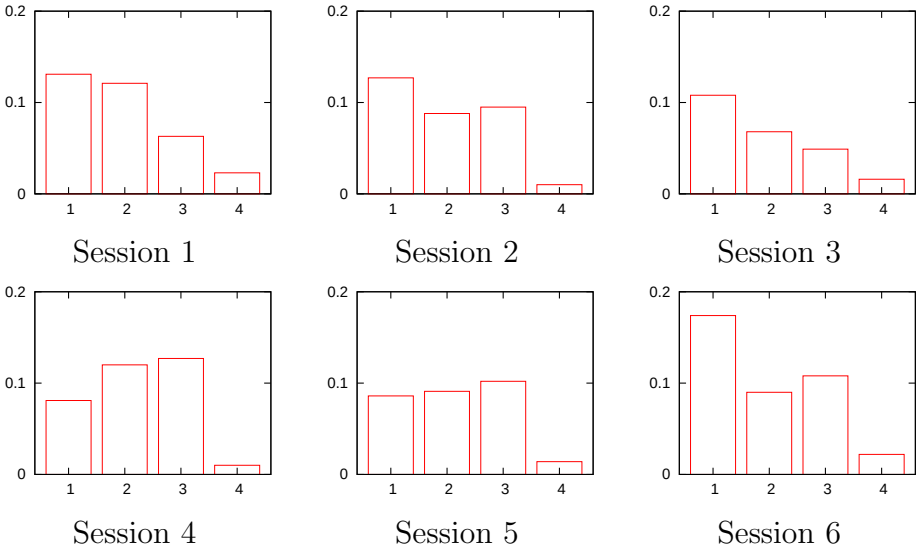

Figure 2: Estimated influences by the proposed model.

distribution is described as follows:

$$P(w|\boldsymbol{w}_{t-\tau}^{t-1}, n) = \lambda P_{\mathrm{C}}(w|\boldsymbol{w}_{t-\tau}^{t-1}) + (1-\lambda)P_{\mathrm{G}}(w), \tag{13}$$




(a) Influence to other speakers.




(b) Influence from other speakers.

Figure 3: Quantitative analysis with NTT data. The x-axis represents the speaker.

where

$$P_{\mathrm{C}}(w|\boldsymbol{w}_{t-\tau}^{t-1}) = \frac{N(t-1,\tau,w)+\beta}{\sum_{w'} N(t-1,\tau,w)+\beta W}, \tag{14}$$

is the common cache. Here, $N(t-1,\tau,w)$ represents the number of times word $w$ is spoken from $t-\tau$ to $t-1$. The CC is the same as the standard cache language model.

- **OC** has the speaker's own caches, and a common parameter that controls the influence of the speaker's own cache. The word distribution is as follows:

$$P(w|\boldsymbol{w}_{t-\tau}^{t-1},n)=\lambda P_{\mathrm{C}}(w|\boldsymbol{w}_{t-\tau}^{t-1},n)+(1-\lambda)P_{\mathrm{G}}(w). \tag{15}$$

  This model assumes that the word use depends only on the speaker's own preceding word use.

- **IC** has individual caches for each speaker, and a common parameter that controls the influence of the speaker dependent caches. The word distribution is as follows:

$$P(w|\boldsymbol{w}_{t-\tau}^{t-1},n)=\sum_{m=1}^{M}\lambda_m P_{\mathrm{C}}(w|\boldsymbol{w}_{t-\tau}^{t-1},m)+(1-\sum_{m=1}^{M}\lambda_m)P_{\mathrm{G}}(w), \tag{16}$$

  where $\lambda_m$ represents the influence of speaker $m$ on all speakers including speaker $m$ himself/herself. This model assumes that the strength of the influence depends on the speaker, but the sensitivity does not differ among speakers.

- **CI** has a common cache, and individual parameters that control the influence of the common cache for each speaker as follows:

$$P(w|\boldsymbol{w}_{t-\tau}^{t-1},n)=\lambda_n P_{\mathrm{C}}(w|\boldsymbol{w}_{t-\tau}^{t-1})+(1-\lambda_n)P_{\mathrm{G}}(w). \tag{17}$$

  This model assumes that the word use depends on all speakers' word use and the degree of dependence differs among speakers.

- **OI** has the speaker's own caches, and individual parameters that control the influence depending on the speakers as follows:

$$P(w|\boldsymbol{w}_{t-\tau}^{t-1},n)=\lambda_n P_{\mathrm{C}}(w|\boldsymbol{w}_{t-\tau}^{t-1},n)+(1-\lambda_n)P_{\mathrm{G}}(w). \tag{18}$$

- **II** has individual caches for each speaker, and individual influence parameters. This is our proposed model in (2).

The first letter of a method's name C/O/I represents common/own/individual caches, respectively, and the second letter of the method's name C/I represents common/individual parameters, respectively. Only the proposed model (II) takes the asymmetricity of influences into account. With all the models, we used $\alpha = 1$, $\beta = 10^{-8}$ and $\tau = \infty$.

In each session, we used data until the $j$th word as training data to learn the parameters, and used words after the $(j+1)$th word as test data. We evaluated the performance of each model using the perplexity of held-out words:

$$\exp\left(-\frac{\sum_{t=j+1}^{T}\log P(w_t|\boldsymbol{w}_1^j, s_i)}{T-j}\right). \tag{19}$$

A lower perplexity represents higher predictive performance.

Table 2 shows the average perplexities for the NTT and RT07 data sets, in which the number of training utterances ranges from $j = 10$ to $j = 300$. Here, an utterance consists of a set of words that are consecutively spoken by a speaker. The proposed model achieved the lowest perplexities in all sessions. This result indicates that it is important to estimate the asymmetric influences between speakers, which only the proposed model considers. Figure 4 show the perplexities with different numbers of training utterances for NTT and RT07 data sets. Generally speaking, the perplexity decreased as the number of training utterances increased because the estimation accuracy of the influences and preceding word use improves. In some sessions, for example Session 3 in the NTT data set, the perplexity increased because of the change of topics. Except when the number of training utterances was small, the perplexity of the proposed model (II) steadily achieved the lowest perplexities. When the number of training utterances is very small, the perplexity of the proposed model (II) was higher than a few other models in some sessions because the number of parameters in the proposed model is more than that of other methods. However, the proposed model achieved the lowest perplexity with small addition of training utterances.

The average computational time for learning parameters in the proposed model with 300 training utterances was 0.01 and 0.02 seconds for the NTT and RT07 data sets, respectively. The proposed model is very efficient, and it can be used in real time applications [25]. Figure 7 shows the log likelihood over iterations with the proposed model. The log likelihood, which is to be maximized, quickly converged.

Figure 8 shows the average perplexities with different lengths of effective

Table 2: Average perplexities for each session.

(a) NTT

| session# | CC | OC | IC | CI | OI | II |
|---|---|---|---|---|---|---|
| 1 | 259.2 | 251.4 | 256.5 | 259.4 | 249.9 | **247.7** |
| 2 | 279.1 | 263.2 | 280.0 | 278.9 | 264.7 | **261.4** |
| 3 | 297.1 | 287.7 | 298.9 | 298.3 | 288.1 | **284.4** |
| 4 | 321.3 | 307.9 | 314.9 | 320.8 | 309.7 | **294.4** |
| 5 | 332.8 | 322.6 | 328.1 | 332.5 | 320.3 | **313.8** |
| 6 | 274.4 | 260.9 | 277.7 | 275.9 | 267.5 | **254.9** |
| average | 294.0 | 282.3 | 292.7 | 294.3 | 283.4 | **276.1** |

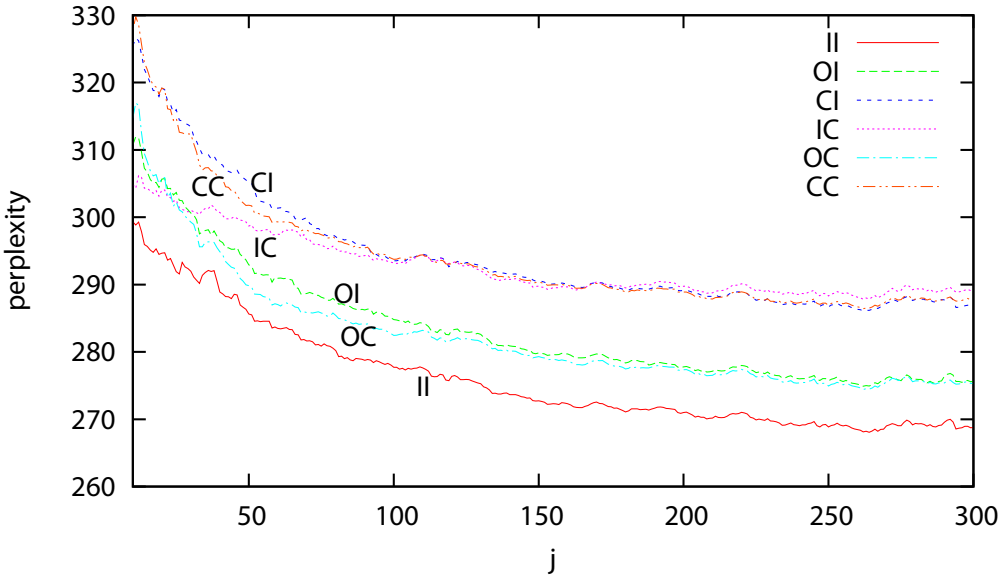
(b) RT07

| session# | CC | OC | IC | CI | OI | II |
|---|---|---|---|---|---|---|
| 1 | 395.7 | 411.8 | 396.7 | 397.1 | 412.7 | **395.6** |
| 2 | 304.3 | 308.5 | 308.6 | 301.6 | 309.4 | **296.1** |
| 3 | 322.6 | 330.6 | 324.4 | 322.9 | 333.6 | **313.2** |
| 4 | 373.0 | 386.1 | 369.0 | 377.2 | 390.7 | **368.7** |
| 5 | 300.5 | 301.9 | 299.9 | 303.2 | 304.4 | **293.4** |
| 6 | 342.5 | 343.7 | 368.2 | 350.4 | 352.3 | **340.4** |
| 7 | 340.6 | 350.7 | 345.4 | 355.8 | 357.4 | **332.3** |
| 8 | 340.7 | 346.7 | 345.7 | 344.9 | 357.9 | **340.3** |
| average | 340.0 | 347.5 | 344.8 | 344.1 | 352.3 | **335.0** |

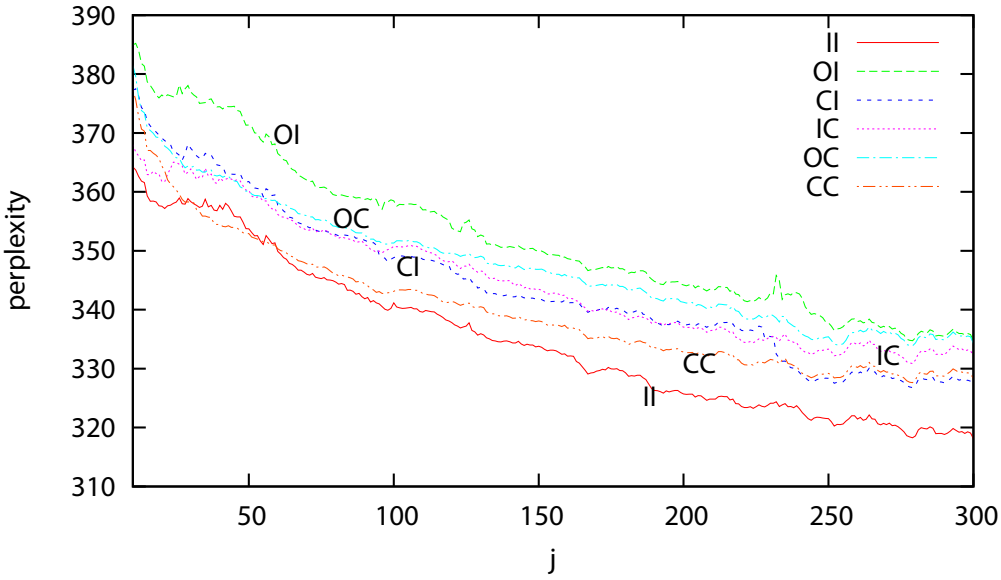
period $\tau$. As the effective period gets longer, the perplexities decreases. This result implies that the speakers in theses data sets were influenced for a long time.

We evaluated the proposed online inference procedures. Figure 9 shows the average perplexities with different forgetting factors $\gamma$. The lowest perplexity was achieved at around $\gamma = 0.1$, which indicates that the tuning of $\gamma$ is important in the online inference. The perplexities achieved by the online inference were higher than those by the batch inference in the both data sets even though the online inference is more efficient than the batch inference.

Figure 10 shows the average perplexities with different numbers of training utterances in online inference with $\gamma = 0.1$. The perplexities decreased as the training utterances increased because the model can use more training data for the inference.




(a) NTT




(a) RT07

Figure 4: Average perplexities with different numbers of training utterances for NTT and RT07 data. The horizontal axis represents the number of training utterances.

## 7. Conclusion

We have proposed a probabilistic model for learning influences from conversation data with multiple speakers. We have confirmed experimentally

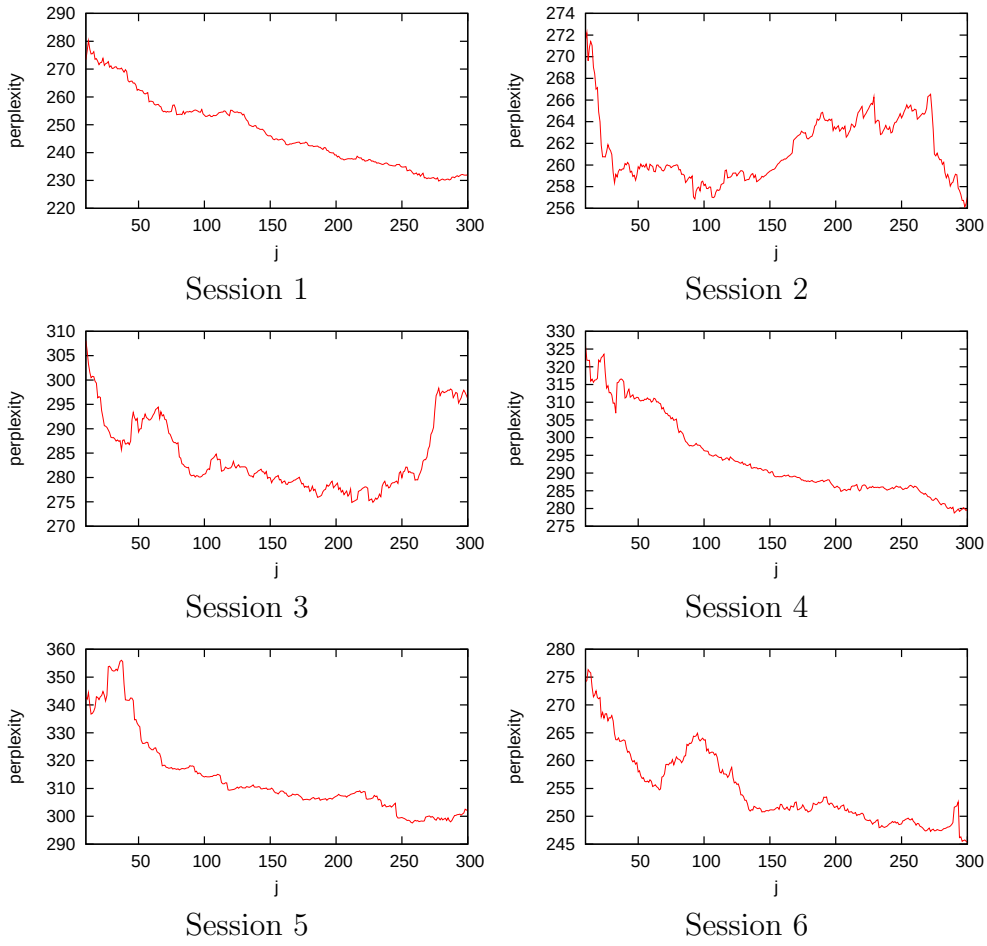



Figure 5: Perplexities with different numbers of training utterances for NTT data. The horizontal axis represents the number of training utterances.

that the proposed model can extract influences between speakers and learn conversation's word use.

Although our results have been encouraging to date, our model can be further improved in a number of ways. First, we would like to estimate influences using other behaviors, such as nonverbal speech acts, posture and eye movement, as well as word use. Second, we would like to extend the proposed model. The proposed model can be extended so that it can incorporate the dynamics of topics by combining it with dynamic models, such as dynamic topic models [26] and topic tracking language models [27]. The

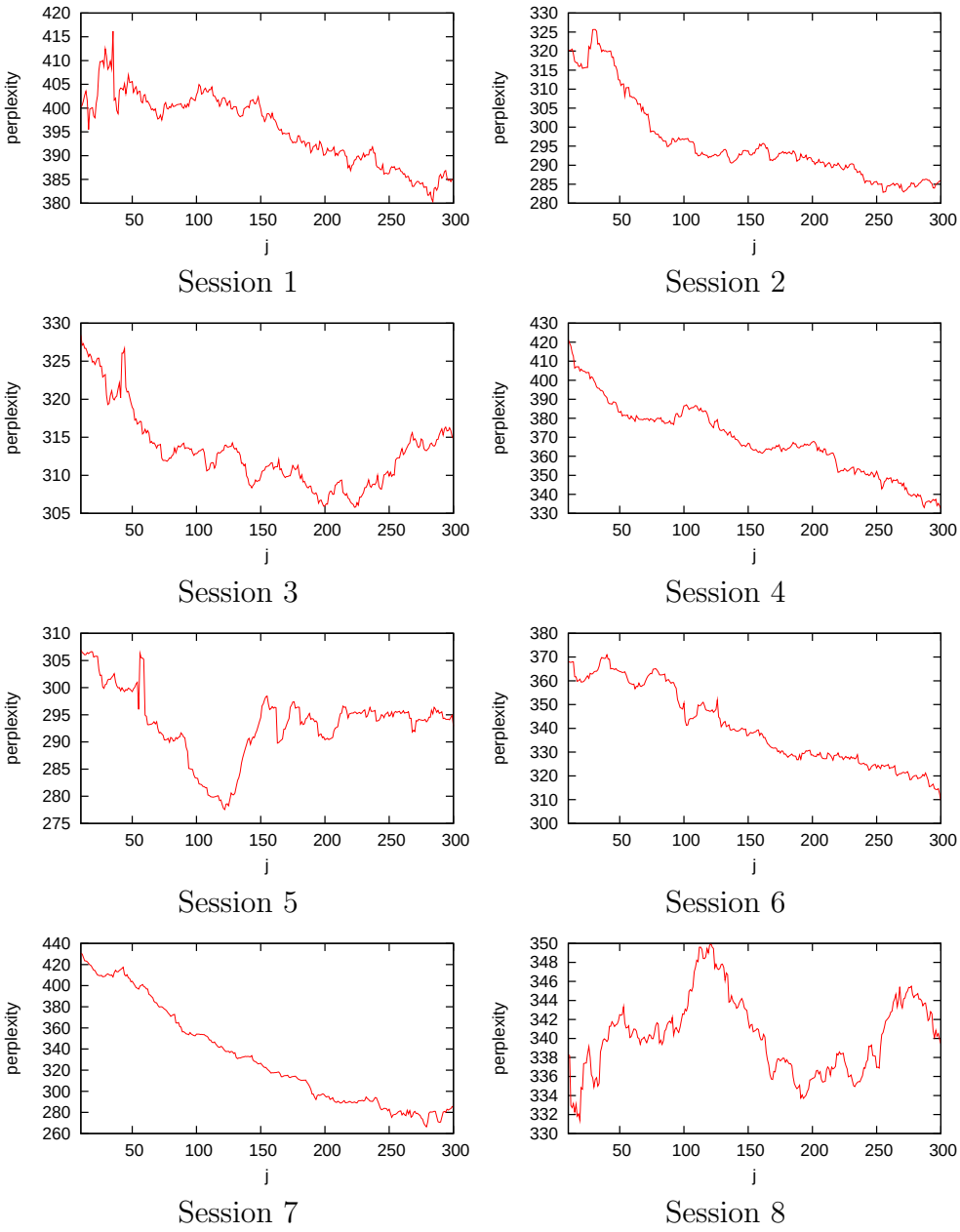
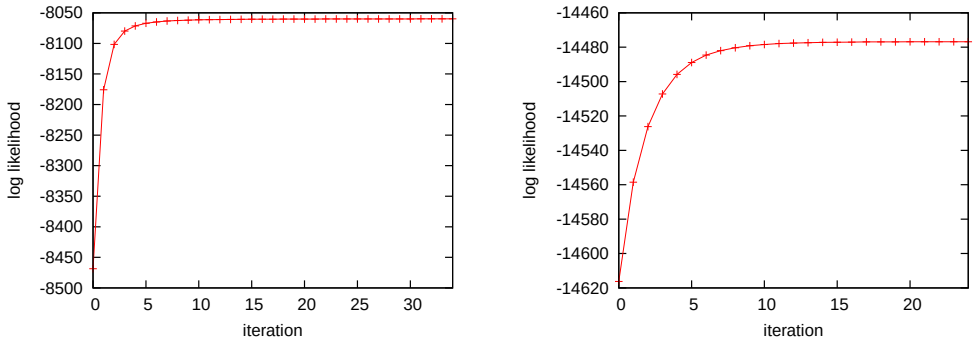
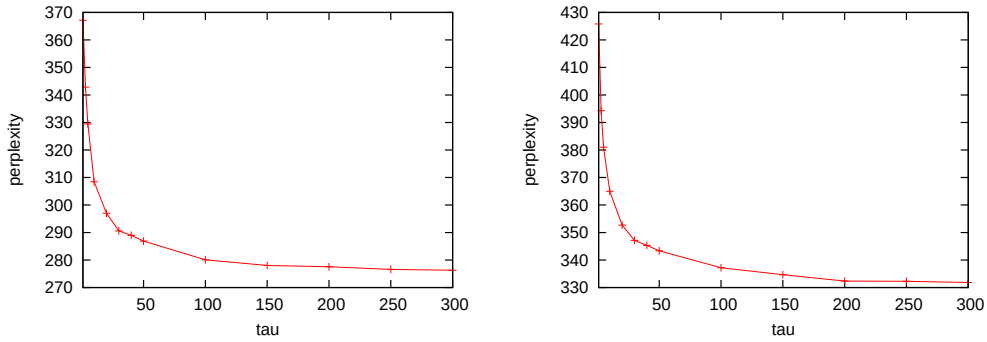



Figure 6: Perplexities with different numbers of training utterances for RT07 data. The horizontal axis represents the number of training utterances.




(a) NTT (b) RT07

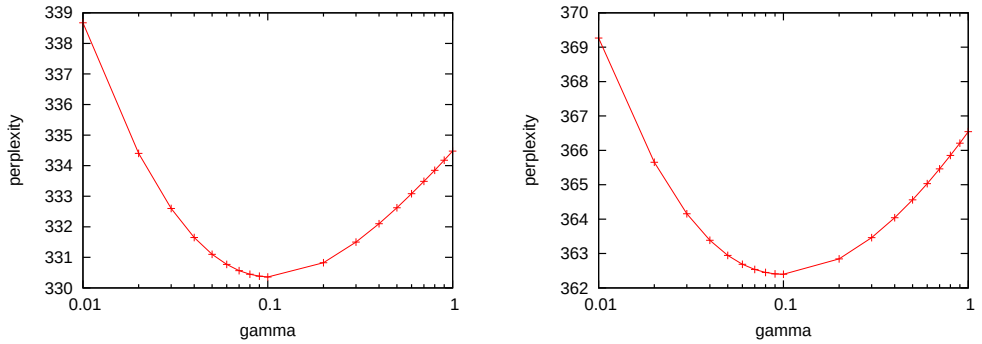
Figure 7: Log likelihoods over iterations.




(a) NTT (b) RT07

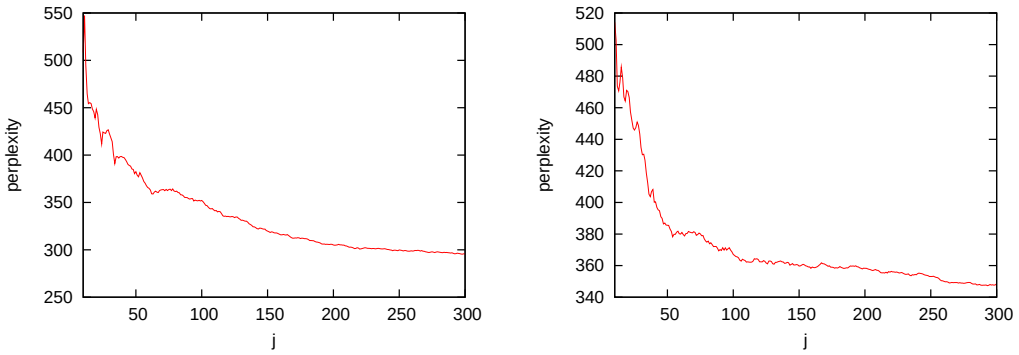
Figure 8: Average perplexities with different lengths of effective period $\tau$.

proposed model assumes that topics do not change over time in a session. The word use can be changed by topics as well as influences. Therefore, we can estimate influences more clearly by modeling topic dynamics. The proposed model can also be extended by using the speaker-specific general word distribution, which can help to estimate influences. Third, we would like to extend the estimation procedure. We must determine the period of the influence automatically. In the proposed model, we used a mixture models with a fixed number of components, where we assumed that all speakers can influence on a speaker. We can extend the model by selecting a model with an arbitrary number of components using model selection techniques. Fourth, we would like to quantitatively evaluate the accuracy of estimated influences by using a mesure that correlates to the actual influences such as




(a) NTT (b) RT07

Figure 9: Average perplexities with different forgetting factors $\gamma$ in online inference.




(a) NTT (b) RT07

Figure 10: Average perplexities with different numbers of training utterances in online inference.

the result of questionnaire. We showed that the proposed model achieved better perplexities than other models. However, the perplexity is a measure for evaluating language models, and not a measure for evaluating influence estimation. Finally, we would like to evaluate the proposed model in an automatic speech recognition system.

## Appendix A. Derivation of (6)

The lower bound of the posterior probability of parameters given the data to be maximized can be obtained as follows,

$$\begin{aligned} L &= \sum_{t=1}^{T} \log \sum_{m=0}^{M} \lambda_{s_t m} P_{\mathrm{C}}(w_t | \boldsymbol{w}_{t-\tau}^{t-1}, m) + \sum_{n=1}^{M} \log P(\boldsymbol{\lambda}_n | \alpha) \\ &= \sum_{t=1}^{T} \log \sum_{m=0}^{M} P(m|t) \frac{\lambda_{s_t m} P_{\mathrm{C}}(w_t | \boldsymbol{w}_{t-\tau}^{t-1}, m)}{P(m|t)} + \sum_{n=1}^{M} \log P(\boldsymbol{\lambda}_n | \alpha) \\ &\geq \sum_{t=1}^{T} \sum_{m=0}^{M} P(m|t) \log \frac{\lambda_{s_t m} P_{\mathrm{C}}(w_t | \boldsymbol{w}_{t-\tau}^{t-1}, m)}{P(m|t)} + \sum_{n=1}^{M} \log P(\boldsymbol{\lambda}_n | \alpha) \\ &= Q - \sum_{t=1}^{T} \sum_{m=0}^{M} P(m|t) \log P(m|t), \end{aligned} \tag{A.1}$$

where we used Jensen's inequality. Therefore, $Q$ in (6) is the lower bound of the objective function with respect to parameters $\lambda_{nm}$.

## References


[1] T. L. Chartrand, J. A. Bargh, The chameleon effect: the perception-behavior link and social interaction, Journal of Personality and Social Psychology 76 (1999) 893–910.

[2] U. Dimberg, Facial reactions to facial expressions, Psychophysiology 19 (1982) 643–647.

[3] S. Brennan, Lexical entrainment in spontaneous dialog, in: ISSD '96, pp. 41–44.

[4] A. Nenkova, A. Gravano, J. Hirschberg, High frequency word entrainment in spoken dialogue, in: ACL '08: HLT, pp. 169–172.

[5] D. Reitter, F. Keller, J. D. Moore, Computational modelling of structural priming in dialogue, in: NAACL '06, pp. 121–124.

[6] R. Levitan, J. Hirschberg, Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions, in: Proceedings of 12th Annual Conference of the International Speech Communication Association, Interspeech '11, pp. 3081–3084.

[7] R. Coulston, S. Oviatt, C. Darves, Amplitude convergence in children's conversational speech with animated personas, in: ICSLP '02, volume 4, pp. 2689–2692.

[8] H. Giles, J. Coupland, N. Coupland, Accommodation theory: Communication, context, and consequence, Cambridge University Press, 1991.

[9] L. E. Scissors, A. J. Gill, D. Gergle, Linguistic mimicry and trust in text-based CMC, in: CSCW '08, pp. 277–280.

[10] T. Iwata, S. Watanabe, Learning influences from word use in polylogue, in: Proceedings of 12th Annual Conference of the International Speech Communication Association, Interspeech '11, pp. 3089–3092.

[11] A. Dempster, N. Laird, D. Rubin, Maximum likelihood from incomplete data via the EM algorithm, Journal of the Royal Statistical Society, Series B 39 (1977) 1–38.

[12] S. Renals, T. Hain, , H. Bourlard, Recognition and interpretation of meetings: The ami and amida projects, in: Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU '07, pp. 238–247.

[13] K. Otsuka, S. Araki, K. Ishizuka, M. Fujimoto, M. Heinrich, J. Yamato, A realtime multimodal system for analyzing group meetings by combining face pose tracking and speaker diarization, in: Proceedings of the 10th international conference on Multimodal interfaces, ICMI '08, ACM, New York, NY, USA, 2008, pp. 257–264.

[14] A. Vinciarelli, Speakers role recognition in multiparty audio recordings using social network analysis and duration distribution modeling, IEEE Transactions on Multimedia 8 (2007) 1215–1226.

[15] F. Valente, A. Vinciarelli, Language-independent socio-emotional role recognition in the ami meetings corpus, in: Proceedings of 12th Annual Conference of the International Speech Communication Association, Interspeech '11, pp. 3077–3080.

[16] G. Ji, J. Bilmes, Multi-speaker language modeling, in: HLT-NAACL '04, pp. 133–136.

[17] M. Purver, T. L. Griffiths, K. P. Körding, J. B. Tenenbaum, Unsupervised topic modelling for multi-party spoken discourse, in: ACL '06, pp. 17–24.

[18] A. Auyeung, T. Iwata, Capturing implicit user influence in online social sharing, in: ACM Conference on Hypertext and Hypermedia (HT '10), pp. 245–254.

[19] R. Kuhn, R. D. Mori, A cache-based natural language model for speech recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 12 (1990) 570–583.

[20] D. Gildea, T. Hofmann, Topic-based language models using em, in: Proceedings of Eurospeech '99, pp. 2167–2170.

[21] R. M. Neal, G. E. Hinton, A view of the em algorithm that justifies incremental, sparse, and other variants, in: Learning in Graphical Models, pp. 355–368.

[22] M. Sato, S. Ishii, On-line em algorithm for the normalized gaussian network, Neural Computation 12 (2000) 407–432.

[23] T. Hori, S. Araki, T. Yoshioka, M. Fujimoto, S. Watanabe, T. Oba, A. Ogawa, K. Otsuka, D. Mikami, K. Kinoshita, T. Nakatani, A. Nakamura, J. Yamato, Low-latency real-time meeting recognition and understanding using distant microphones and omni-directional camera, IEEE Transactions on Audio, Speech, and Language Processing (2011).

[24] J. G. Fiscus, J. Ajot, J. S. Garofolo, The rich transcription 2007 meeting recognition evaluation, in: Multimodal Technologies for Perception of Humans, pp. 373–389.

[25] T. Hori, S. Araki, T. Yoshioka, M. Fujimoto, S. Watanabe, T. Oba, A. Ogawa, K. Otsuka, D. Mikami, K. Kinoshita, T. Nakatani, A. Nakamura, J. Yamato, Real-time meeting recognition and understanding using distant microphones and omni-directional camera, in: SLT '10, pp. 412–417.

[26] D. M. Blei, J. D. Lafferty, Dynamic topic models, in: ICML '06, pp. 113–120.

[27] S. Watanabe, T. Iwata, T. Hori, A. Sako, Y. Ariki, Topic tracking language model for speech recognition, Computer Speech & Language 25 (2011) 440–461.