Bayesian Integration of Sound Source Separation and Speech Recognition: A New Approach to Simultaneous Speech Recognition

Kousuke Itakura, Izaya Nishimuta, Yoshiaki Bando, Katsutoshi Itoyama, and Kazuyoshi Yoshii

Graduate School of Informatics, Kyoto University, Japan

`{itakura,nisimuta,yoshiaki,itoyama,yoshii}@kuis.kyoto-u.ac.jp`

Abstract

This paper presents a novel Bayesian method that can directly recognize overlapping utterances without explicitly separating mixture signals into their independent components in advance of speech recognition. The conventional approach to contaminated speech recognition in real environments uniquely extracts the clean isolated signals of individual sources (*e.g.*, by noise reduction, dereverberation, and source separation). One of the main limitations of this cascading approach is that the accuracy of speech recognition is upper bounded by the accuracy of preprocessing. To overcome this limitation, our method *marginalizes out* uncertain isolated speech signals by integrating source separation and speech recognition in a Bayesian manner. A sufficient number of samples are drawn from the posterior distribution of isolated speech signals by using a Markov chain Monte Carlo method, and then the posterior distributions of uttered texts for those samples are integrated. Under a certain condition, this Monte Carlo integration is shown to reduce to the well-known method called ROVER that integrates recognized texts obtained from sampled speech signals. Results of simultaneous speech recognition experiments showed that in terms of word accuracy the proposed method significantly outperformed conventional cascading methods.

Index Terms: simultaneous speech recognition, sound source separation, Bayesian modeling, MCMC, ROVER

1. Introduction

It should be noted that although in our daily lives we always hear mixed sounds, it sometimes seems that there is only a single sound source. For example, we are able to focus our auditory attention on a specific person to talk with in a noisy room while ignoring irrelevant sounds such as the utterances of the other people, environmental noise, background music, and reverberant sounds. Such selective attention is well known as the *cocktail party effect* [1] and contributes to our capability of robust speech recognition in noisy environments. Moreover, we can to some extent recognize the overlapping utterances made by two or three people [2, 3]. An interesting observation is that in both cases the recognition results of contaminated speech signals immediately come into our awareness with some confidence even though clean speech signals remain unknown.

The conventional way to improve the accuracy of automatic speech recognition (ASR) in a noisy environment [4–10] is to take a cascading approach as follows:

$$\boldsymbol{S}^* = \underset{\boldsymbol{S}}{\arg\max}\, p(\boldsymbol{S}|\boldsymbol{X}), \qquad (1)$$

$$\boldsymbol{Z}^* = \underset{\boldsymbol{Z}}{\arg\max}\, p(\boldsymbol{Z}|\boldsymbol{S}^*). \qquad (2)$$

where $\boldsymbol{X}$, $\boldsymbol{S}$, and $\boldsymbol{Z}$ are contaminated mixture signals, isolated speech signals, and uttered texts, respectively. Eq. (1) repre-
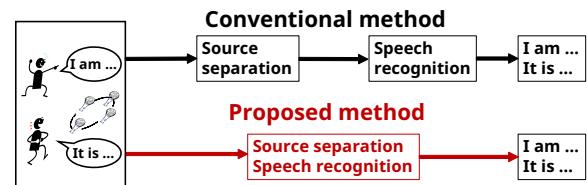

Figure 1: The proposed method directly estimates uttered texts without uniquely estimating isolated speech signals.

sents a preprocessing step (*e.g.*, noise reduction, dereverberation, and/or source separation) that produces maximum a posteriori (MAP) estimates of clean isolated speech signals, $\boldsymbol{S}^*$, from the input signals $\boldsymbol{X}$, and Eq. (2) represents a subsequent ASR step that produces MAP estimates of uttered texts, $\boldsymbol{Z}^*$, from the isolated speech signals $\boldsymbol{S}^*$. Note that isolated speech signals are determined uniquely as intermediate products $\boldsymbol{S}^*$ even though all we are interested in are final recognition results $\boldsymbol{Z}^*$. A critical problem with this approach is that errors of the preprocessing step directly have a negative impact on speech recognition because they cannot be corrected in the ASR step. Furthermore, the estimated isolated speech signals $\boldsymbol{S}^*$ are not guaranteed to be optimal for speech recognition.

To solve these problems, we estimate the final recognition results $\boldsymbol{Z}^*$ directly from the input signals $\boldsymbol{X}$ by integrating the preprocessing and ASR steps in a Bayesian manner as follows:

$$p(\boldsymbol{Z}|\boldsymbol{X}) = \int p(\boldsymbol{Z}|\boldsymbol{S}) p(\boldsymbol{S}|\boldsymbol{X}) d\boldsymbol{S}, \qquad (3)$$

$$\boldsymbol{Z}^* = \underset{\boldsymbol{Z}}{\arg\max}\, p(\boldsymbol{Z}|\boldsymbol{X}), \qquad (4)$$

where Eq. (3) represents marginalization over all possible isolated speech signals $\boldsymbol{S}$ for avoiding uniquely determining $\boldsymbol{S}^*$. Since in general Eq. (3) is analytically intractable, we instead perform Monte Carlo integration [11] as follows:

$$\boldsymbol{S}_l \sim p(\boldsymbol{S}_l|\boldsymbol{X}), \qquad (5)$$

$$p(\boldsymbol{Z}|\boldsymbol{X}) = \frac{1}{L} \sum_{l=1}^{L} p(\boldsymbol{Z}|\boldsymbol{S}_l), \qquad (6)$$

where $\boldsymbol{S}_l$ is a random sample drawn from the posterior distribution of $\boldsymbol{S}$ in the preprocessing step and L is the total number of samples. In the ASR step the posterior distributions of $\boldsymbol{Z}$ are averaged over L samples.

In this paper we focus on source separation as preprocessing and propose a novel method of simultaneous speech recognition for directly recognizing overlapping utterances $\boldsymbol{Z}^*$ contained in mixture speech signals $\boldsymbol{X}$ (Fig. 1). All possibilities of uncertain isolated speech signals $\boldsymbol{S}$ can be taken into account by integrating source separation $p(\boldsymbol{S}|\boldsymbol{X})$ and speech recognition $p(\boldsymbol{Z}|\boldsymbol{S})$ in a Bayesian manner. More specifically, we use a nonparametric Bayesian method of microphone-array-based

source separation [12] for sampling isolated speech signals $\boldsymbol{S}_l$ according to Eq. (5), and for ASR we use an open-source software called Julius [13]. Since Eqs. (4) and (6) are still hard to solve, for mathematical convenience we make two assumptions in the ASR step. The first is that the posterior distributions of individual words are independent of each other, and the second is that the recognition results can be determined with absolute confidence in the ASR step. These assumptions make Eqs. (4) and (6) equivalent to integration of the recognition results for L sampled speech signals based on a multistage recognizer output voting error reduction (ROVER) method [14].

2. Related work

This section introduces related work on source separation and speech recognition. The advance in source separation enabled us to accurately estimate the speech signal of each speaker. The advance in speech recognition, on the other hand, enabled us to accurately recognize distorted and/or noisy speech signals. Although these advances led to improvement of the recognition performance, independent improvement of these steps has a limitation for improving recognition performance. Therefore, a speech recognition method that directly models simultaneous speeches has also been studied.

2.1. Source separation

Various methods of source separation have been proposed [12, 15, 16] as a basic technique of audio analysis. To cluster time-frequency bins of multi-channel mixture signals into individual source signals, Otsuka *et al.* [12] proposed a notable method of microphone array processing based on a hierarchical Dirichlet process extension of the covariance model [17] (HDP-CM). Barker *et al.* [15] proposed a method of monaural sound source separation that uses the modulation spectrogram as a feature for nonnegative tensor factorization. To improve the robustness of speech recognition, Shao *et al.* [16] proposed a source separation method that uses the periodicity information to segregate voiced portions of individual sources in each time frame and the onset/offset information to segregate unvoiced portions.

2.2. Speech recognition

To make speech recognition robust to noise and reverberation, many studies have attempted to improve acoustic models. For example, acoustic models have often been trained from not only clean speech data but also contaminated speech data [5, 6]. A major limitation of such multi-condition training is that acoustic models should be retrained if recording environments are changed. Another popular approach is model adaptation that tries to modify an acoustic model trained from clean speech data according to recording environments [7, 8]. Recently, deep neural networks (DNNs) have widely been used for substantially improving the generalization capability of acoustic models [9]. Michael *et al.* [10] empirically showed the noise robustness of DNN-based acoustic models.

2.3. Simultaneous speech recognition

Varga *et al.* [18] and Deoras *et al.* [19] proposed a method of simultaneous speech recognition that can directly recognize overlapping utterances of short words (digits) by using a factorial hidden Markov model (FHMM) that consists of multiple hidden Markov chains corresponding to individual phoneme sequences. This method, however, is computationally prohibitive for large-vocabulary continuous speech recognition. Several methods [20, 21] perform source separation and then execute speech recognition using only reliable acoustic features of separated speech signals according to the missing feature theory.
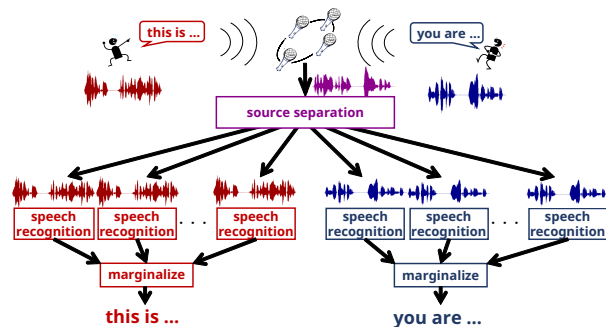

Figure 2: Architecture of the proposed system.

3. Proposed method

This section explains a proposed method of simultaneous speech recognition based on Bayesian integration of source separation and speech recognition according to Eqs. (4), (5), and (6). Since the optimization problem given by Eq. (4) is still hard to solve, in this paper we pose two assumptions for mathematical convenience (explained later). This makes the optimization problem tractable as follows (Fig. 2):

1. Computing Eq. (5)
   A sufficient number of isolated speech signals $\{\boldsymbol{S}_l\}_{l=1}^L$ are randomly sampled from the posterior distribution of those signals. The state-of-the-art probabilistic model of source separation called HDP-CM [12] is used for calculating the posterior distribution.

2. Computing Eqs. (4) and (6).
   Each separated speech signal is independently recognized by using a well-known speech recognizer called Julius [13]. To uniquely determine the most likely recognition results for the simultaneous speech signals, the recognition results for all sampled speech signals are integrated by using a multistage recognizer output voting error reduction (ROVER) method [14].

Note that the proposed method does not uniquely determine hypothetical isolated speech signals, *i.e.*, those signals can be marginalized out in a Bayesian manner. This leads to the performance superiority over conventional methods that just perform source separation and speech recognition in a cascading manner.

3.1. Derivation of optimization algorithm

To derive a tractable algorithm that solves Eqs. (4), (5), and (6), we pose two assumptions just for the sake of mathematical convenience. The first assumption is that the posterior distributions of individual words are independent of each other as follows:

$$p(\boldsymbol{Z}|\boldsymbol{X}) = \prod_{k=1}^K p(Z_k|\boldsymbol{X}), \qquad (7)$$

where Z_k is the kth word in the recognition result $\boldsymbol{Z}$ and K is the number of words contained in $\boldsymbol{Z}$. We aim to estimate the optimal word Z_k^* for each k because maximization of $p(\boldsymbol{Z}|\boldsymbol{X})$ is equivalent to independent maximization of $p(Z_k|\boldsymbol{X})$.

The second assumption is that speech recognition itself can be performed with absolute confidence. In other words, we assume a "function" $f(\boldsymbol{S}) = \boldsymbol{Z}$ that converts isolated speech signals $\boldsymbol{S}$ into texts $\boldsymbol{Z}$ in a deterministic way as follows:

$$p(Z_k|\boldsymbol{S}_l) = 1_{f_k(\boldsymbol{S}_l)}(Z_k), \qquad (8)$$

where $1_{f_k(\boldsymbol{S}_l)}(Z_k)$ is a delta function given by

$$1_{f_k(\boldsymbol{S}_l)}(Z_k) = \begin{cases} 1 & (f_k(\boldsymbol{S}_l) = Z_k), \\ 0 & (\text{otherwise}), \end{cases} \qquad (9)$$

where $f_k(\boldsymbol{S}_l)$ is the kth word in $f(\boldsymbol{S}_l)$. Using these assumptions, the Eq. (6) becomes,

$$Z_k^* \approx \operatorname*{argmax}_{Z_k} \frac{1}{L} \sum_{l=1}^{L} 1_{f_k(\boldsymbol{S}_l)}(Z_k). \qquad (10)$$

This means that each optimal word Z_k^* can be obtained by a majority vote of the recognized words $\{f_k(\boldsymbol{S}_l)\}_{l=1}^{L}$ for L sampled speech signals. This is equivalent to a standard variant of the ROVER method [14] based on word counts in multiple recognition candidates. The speech recognition performance can be improved by dealing with the recognition confidence of each word. We integrate the recognized words $\{f_k(\boldsymbol{S}_l)\}_{l=1}^{L}$ by using the ROVER method with the confidence values.

3.2. Implementation of optimization algorithm

The proposed method consists of the following two steps: 1) taking samples of isolated speech signals (separated sounds) and 2) recognizing those signals independently and integrating the recognition results.

3.2.1. Taking samples of isolated speech signals

A sufficient number of samples of separated sounds $\{\boldsymbol{S}_l\}_{l=1}^{L}$ are taken by using HDP-CM [12]. HDP-CM makes an assumption of the spectral sparsity of each source signal in the time-frequency domain. This enables one to assume that only one sound source is likely to be dominant at each time-frequency bin. HDP-CM separates a mixed sound by clustering time-frequency bins into individual sound sources and localizes each source by assigning a certain direction to each cluster (Fig. 3).

First of all, the input audio signal is converted into the time-frequency domain by taking the short-time Fourier transform (STFT) [22]. Let $\boldsymbol{x}_{tf}$ be the multi-channel observed spectra at time frame t and frequency bin f. When the mixed sound consisting of N sources is observed with M microphones under an anechoic condition, $\boldsymbol{x}_{tf} \in \mathbb{C}^M$ is an M-dimensional complex-valued vector given by

$$\boldsymbol{x}_{tf} = \boldsymbol{B}_f \boldsymbol{s}_{tf}. \qquad (11)$$

Note that $\boldsymbol{s}_{tf} \in \mathbb{C}^N$ is the source spectra and that $\boldsymbol{B}_f \in \mathbb{C}^{M \times N}$ are the instantaneous mixing coefficients. The nth element of $\boldsymbol{s}_{tf}$ is the source spectrum arising from the nth source. Let $b_{f,mn}$ be the element of $\boldsymbol{B}_f$ at the mth row and nth column. $b_{f,mn}$ represents the wave-propagation characteristics from the nth source to the mth microphone. We assume that $\boldsymbol{s}_{tf}$ follows a Gaussian distribution as follows:

$$\boldsymbol{s}_{tf} \sim \mathcal{N}_{\mathbb{C}}(\boldsymbol{s}_{tf} | \boldsymbol{0}, \lambda_{tf}^{-1} \boldsymbol{I}), \qquad (12)$$

where $\mathcal{N}_{\mathbb{C}}(\boldsymbol{x}_{tf} | \boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})$ is the multivariate complex normal distribution with mean $\boldsymbol{\mu}$ and precision $\boldsymbol{\Lambda}$ [23]. λ_{tf} is treated as a fixed value, *i.e.*, $\lambda_{tf} = |\boldsymbol{x}_{tf}|^{-1}$. $\boldsymbol{I}$ represents a identify matrix. Since the spectral components of the sound sources are sparsely distributed in the time-frequency domain, it is assumed that for each time frame t and frequency bin f only one sound source is dominant. This enables us to establish the following equation.

$$\boldsymbol{x}_{tf} = \boldsymbol{b}_{fk_{tf}} s_{tf}^{k_{tf}}, \qquad (13)$$

where k_{tf} denotes a source index that is dominant at time frame t and frequency bin f. $\boldsymbol{b}_{fk_{tf}}$ corresponds to the k_{tf}th column vector of $\boldsymbol{B}_f$, and $s_{tf}^{k_{tf}}$ corresponds to the k_{tf}th element of $\boldsymbol{s}_{tf}$. The assumption of the source sparsity means that $s_{tf}^{k'} = 0$ if $k_{tf} = k$ and $k' \neq k$. Focusing on the linear relationship between $\boldsymbol{x}_{tf}$ and $\boldsymbol{s}_{tf}$, Eqs. (12) and (13) lead to the likelihood function for $\boldsymbol{x}_{tf}$ as follows:

$$\boldsymbol{x}_{tf} | z_{tf}, \boldsymbol{w}, \lambda_{tf}, \boldsymbol{H} \sim \mathcal{N}_{\mathbb{C}}(\boldsymbol{x}_{tf} | \boldsymbol{0}, (\lambda_{tf} \boldsymbol{H}_{fw_{z_{tf}}})^{-1}). \quad (14)$$
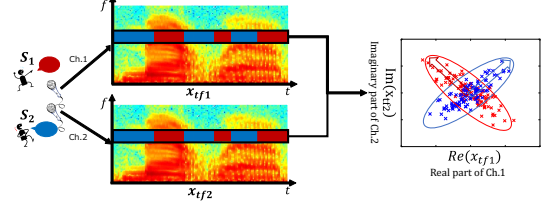

Figure 3: HDP-CM. The right graph is a plot of complex-valued multi-channel spectra of two sources at a particular frequency.

This observation model corresponds to the covariance model [17]. z_{tf} represents the cluster index of $\boldsymbol{x}_{tf}$, w_k represents the direction of the kth cluster, and $\boldsymbol{H}_{fw_{z_{tf}}}^{-1} \approx \boldsymbol{b}_{fw_{z_{tf}}} \boldsymbol{b}_{fw_{z_{tf}}}^H$. In addition, we assume that $\boldsymbol{H}_{fd}$ follows a complex Wishart distribution [24], which is the conjugate prior distribution, that is, $\boldsymbol{H}_{fd}$ follows the following process:

$$\boldsymbol{H}_{fd} \sim \mathcal{W}_{\mathbb{C}}(\boldsymbol{H}_{fd} | \boldsymbol{\nu}_{fd}, \boldsymbol{G}_{fd}). \qquad (15)$$

The hyperparameters are set as $\boldsymbol{G}_{fd} = (\boldsymbol{q}_{fd} \boldsymbol{q}_{fd}^H + \varepsilon \boldsymbol{I})^{-1}$ and $\boldsymbol{\nu}_{fd} = M$. The given steering vectors $\boldsymbol{q}_{fd}$ are normalized such that $\boldsymbol{q}_{fd}^H \boldsymbol{q}_{fd} = 1$, and ε is set to 0.01 in order to enable inverse operation. The posterior distribution of Eq. (14) is $p(\boldsymbol{Z}, \boldsymbol{W}, \boldsymbol{\lambda}, \boldsymbol{H} | \boldsymbol{X})$. $p(\boldsymbol{Z}, \boldsymbol{W} | \boldsymbol{X})$ is calculated from this posterior distribution by assuming $\boldsymbol{\lambda}$ to be a fixed value and marginalizing out $\boldsymbol{H}_{fk_{tf}}$. As illustrated in Fig. 3, clustering of z_{tf} and w_k is equivalent to source separation and localization. This is performed by sampling the cluster using $p(\boldsymbol{Z}, \boldsymbol{W} | \boldsymbol{X})$. When z_{tf}^i and w_k^i denote the ith sample and $\boldsymbol{x}_{tf}^d$ denotes the power arising from the source at the direction d in time frame t and frequency bin f, $\boldsymbol{x}_{tf}^d$ is calculated as follows:

$$\boldsymbol{x}_{tf}^d = \frac{1}{I} \sum_{i=1}^{I} \delta(w_{z_{tf}^i}^i, d) \boldsymbol{x}_{tf}, \qquad (16)$$

where $\delta(w_k, d)$ is defined as follows:

$$\delta(w_k, d) = \begin{cases} 1 & (w_k = d), \\ 0 & (\text{otherwise}). \end{cases} \qquad (17)$$

The separated sound signal arising from the source at direction d is determined by calculating $\boldsymbol{x}_{tf}^d$ for each time frame t and frequency bin f. The source signals $\boldsymbol{S}_l$ are obtained by converting $\boldsymbol{x}_{tf}^d$ into the time domain.

3.2.2. Integrating recognition results

Speech recognition for separated speech signals $\boldsymbol{S}_l$ is independently performed by using a standard speech recognizer called Julius [13]. Julius is capable of estimating the confidence for each word during speech recognition [25]. The recognition results for all L samples are integrated into one recognition result by the ROVER method, which is a popular method to integrate several recognition results. The basic flow of the ROVER method is shown in Fig. 4.

The first step of the ROVER method is to make a set of words located at the same position by aligning candidate recognition results for all L samples. Since joint alignment of more than two sentences is difficult, those sentences are aligned one by one using a two-dimensional dynamic programming.

The second step of the ROVER method is to determine the recognition result by a majority vote, *i.e.*, to select the most highly scored word from each set of words. The score of each word is calculated using both the appearance frequency and the word confidence. One word is selected from each set of words. $\text{Score}(w)$, which is the score of word w, is calculated as follows:

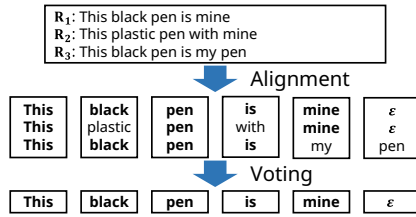$$\text{Score}(w) = \alpha \frac{N_w}{N} + (1 - \alpha) C(w), \qquad (18)$$

Figure 4: The ROVER method. R_i denotes the ith candidate recognition result. Sets of words are constructed by aligning candidate recognition results, and then a recognition result is determined by a majority vote on each set of words.

where N_w is the number of times that w included in the set of words, N is the number of words included in the set of words, α is a parameter used to define the ratio of the appearance frequency and the word confidence, and $C(w)$ is calculated using the appearance frequency as follows:

$$C(w) = \frac{1}{N_w} \sum_{w_i = w} \text{confidence}(w_i). \tag{19}$$

Note that w_i denotes the ith word in the set of words and that confidence(w_i) denotes the word confidence for w_i.

4. Evaluation

This section reports comparative experiments that were conducted for evaluating the performance of the proposed method of simultaneous speech recognition.

4.1. Experimental conditions

4-channel simultaneous speech signals were synthesized by using the transfer function of an anechoic room and two or three isolated speech signals randomly selected from the ATR phonetically balanced Japanese utterances (a01–a50) [26]. Sound sources were positioned 150 cm away from a microphone array. We observed two sound sources with the interval $\theta = 30°$ or $60°$, and three sound sources with the interval $\theta = 30°$. We made 50 sets of simultaneous speech signals. Those audio signals were sampled at 16 kHz, and the STFT was calculated with a Hamming window of 512 samples and a shifting interval of 128 samples. We compared the proposed method with conventional methods that performed source separation and speech recognition in a cascading manner. Two source separation methods (IVA [27] and HDP-CM [12]) were used as conventional source separation methods. Julius was used for speech recognition (julius-dictation-kit-v4.3.1). In MCMC sampling for the proposed method, each sample S_l of isolated speech signals was obtained by taking the average over 20 successive MCMC samples. L was set to 50. In the ROVER method, α was set to 0.5 and word confidence for ε was set to 1.0. Word accuracy [28] was used as a measure of recognition performance:

$$\text{word accuracy} = \frac{C - I}{T} \times 100, \tag{20}$$

where C is the number of correct words, I is the number of insertion errors, and T is the number of words included in ground-truth sentences.

4.2. Experimental results

The results are listed in Table 1. Clean+Julius means the recognition performance of Julius for the clean isolated signals. The proposed method HDP-CM+Julius+ROVER achieved the highest word accuracy for both two and three overlapped utterances. Fig. 5 and Fig. 6 show the word accuracies when the number of samples L was changed from 1 to 50. Although a larger value of

Table 1: Word accuracies for overlapping utterances.

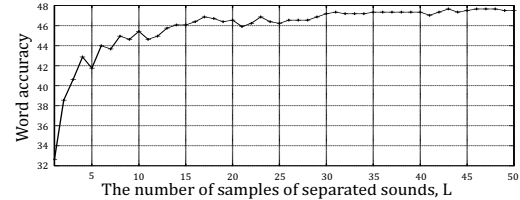| Method	Two	Three
Clean+Julius (upper bound)	66.8	66.8
IVA+Julius	29.4	−17.6
HDP-CM+Julius	32.6	6.17
HDP-CM+Julius+ROVER	**47.5**	**27.1**

Figure 5: The experimental results of recognizing two overlapping utterances according to the number of samples L.
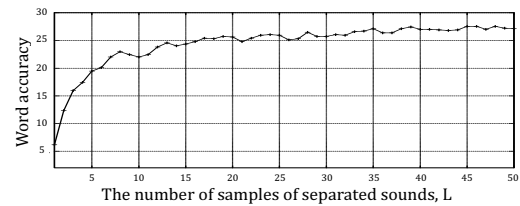

Figure 6: The experimental results of recognizing three overlapping utterances according to the number of samples L.

L tends to yield a better performance, the tradeoff between computational cost and performance should be taken into account. We confirmed that the proposed method was effective for recognizing simultaneous speech including two or three overlapping utterances. The fact that the word accuracy obtained by the proposed method was lower than that obtained by Clean+Julius indicates that there would be much room for improvement.

5. Conclusion

This paper presented a method that integrates source separation and speech recognition in a Bayesian framework. The method can directly recognize overlapping utterances without uniquely determining separated speech signals. Many samples are drawn from the posterior distribution of isolated speech signals by using an MCMC method, and then the posterior distributions of uttered texts for those samples are integrated. Under a certain condition, this Monte Carlo integration was shown to reduce to the well-known method called ROVER that integrates recognized texts obtained from sampled speech signals. Results of simultaneous speech recognition experiments showed that in terms of word accuracy the proposed method significantly outperformed conventional cascading methods.

We plan to relax the two assumptions made just for mathematical convenience in this paper: that the posterior distributions of individual words are independent on each other and that the recognition results have no uncertainty. Moreover, we try to develop a method that integrates noise reduction, dereverberation, source separation, and speech recognition into a unified Bayesian framework for directly recognizing mixture signals containing noise and reverberation in real environments.

6. Acknowledgements

7. References

[1] A. Conway, N. Cowan, and M. Bunting, "The cocktail party phenomenon revisited: The importance of working memory capacity," *Psychonomic Bulletin and Review*, vol. 8, no. 2, pp. 331–335, 2001.

[2] B. G. Shinn-Cunningham and A. Ihlefeld, "Selective and divided attention: Extracting information from simultaneous sound sources." in *Proc. of ICAD*, pp. 967–979, 2004.

[3] A. Ihlefeld and B. Shinn-Cunningham, "Spatial release from energetic and informational masking in a divided speech identification taska," *The Journal of the Acoustical Society of America*, vol. 123, no. 6, pp. 4380–4392, 2008.

[4] K. Nakadai, T. Takahashi, H. G. Okuno, H. Nakajima, Y. Hasegawa, and H. Tsujino, "Design and implementation of robot audition system 'HARK' - open source software for listening to three simultaneous speakers," *Advanced Robotics*, vol. 24, no. 5-6, 2010.

[5] P. Rajan, T. Kinnunen, and V. Hautamäki, "Effect of multicondition training on i-vector plda configurations for speaker recognition." in *Proc. of INTERSPEECH*, pp. 3694–3697, 2013.

[6] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *Proc. of ICASSP*, pp. 7092–7096, 2013.

[7] S. M. Ban and H. S. Kim, "Instantaneous model adaptation method for reverberant speech recognition," *Electronics Letters*, vol. 51, no. 6, pp. 528–530, 2015.

[8] W. Kim and J. H. Hansen, "Gaussian map based acoustic model adaptation using untranscribed data for speech recognition in severely adverse environments." in *Proc. of INTERSPEECH*, pp. 1762–1765, 2012.

[9] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[10] M. Seltzer, D. Yu, and Y. Wang, "An investigation of deep neural networks for noise robust speech recognition," in *Proc. of ICASSP*, pp. 7398–7402, 2013.

[11] W. R. Gilks, *Markov Chain Monte Carlo*. John Wiley and Sons, Ltd, 2005.

[12] T. Otsuka, K. Ishiguro, H. Sawada, and H. Okuno, "Bayesian non-parametrics for microphone array processing," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 2, pp. 493–504, 2014.

[13] A. Lee and T. Kawahara, "Recent development of open-source speech recognition engine Julius," in *Proc. of APSIPA ASC*, pp. 131–137, 2009.

[14] J. G. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER)," in *Proc. of ASRU*, pp. 347–354, 1997.

[15] T. Barker and T. Virtanen, "Non-negative tensor factorisation of modulation spectrograms for monaural sound source separation." in *Proc. of INTERSPEECH*, pp. 827–831, 2013.

[16] Y. Shao, S. Srinivasan, Z. Jin, and D. Wang, "A computational auditory scene analysis system for speech segregation and robust speech recognition," *Computer Speech & Language*, vol. 24, no. 1, pp. 77–93, 2010.

[17] N. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1830–1840, 2010.

[18] A. Varga and R. K. Moore, "Simultaneous recognition of concurrent speech signals using hidden markov model decomposition," in *Proc. ofSecond European Conference on Speech Communication and Technology*, pp. 1175–1178, 1991.

[19] A. Deoras and M. Hasegawa-Johnson, "A factorial HMM approach to simultaneous recognition of isolated digits spoken by multiple talkers on one audio channel," in *Proc. of ICASSP*, vol. 1, pp. 861–864, 2004.

[20] R. Philippe, V. Rolf, and K. Jens, "Robust speech recognition using missing feature theory and vector quantization." in *Proc. of INTERSPEECH*, pp. 1107–1110, 2001.

[21] S. Yamamoto, J. Valin, K. Nakadai, J. Rouat, F. Michaud, T. Ogata, and H. Okuno, "Enhanced robot speech recognition based on microphone array source separation and missing feature theory," in *Proc. of ICRA*, pp. 1477–1482, 2005.

[22] L. Cohen, *Time-frequency analysis*. Englewood Cliffs, NJ: Prentice-Hall, 1995.

[23] A. van den Bos, "The multivariate complex normal distribution-a generalization," *IEEE Transactions on Information Theory*, 1995.

[24] K. Conradsen, A. Nielsen, J. Schou, and H. Skriver, "A test statistic in the complex wishart distribution and its application to change detection in polarimetric sar data," in *Proc. of IEEE Transactions on Geoscience and Remote Sensing*, vol. 41, no. 1, pp. 4–19, 2003.

[25] A. Lee, K. Shikano, and T. Kawahara, "Real-time word confidence scoring using local posterior probabilities on tree trellis search," in *Proc. of ICASSP*, vol. 1, pp. 793–796, 2004.

[26] S. Itahashi, "On recent speech corpora activities in Japan," *Journal of the Acoustical Society of Japan (E)*, vol. 20, no. 3, pp. 163–169, 1999.

[27] N. Ono, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," in *Proc. of WASPAA*, pp. 189–192, 2011.

[28] K.-F. Lee, *Automatic Speech Recognition: The Development of the SPHINX Recognition System*. Springer, 1989, vol. 62.