JOINT UNSUPERVISED LEARNING OF HIDDEN MARKOV SOURCE MODELS AND SOURCE LOCATION MODELS FOR MULTICHANNEL SOURCE SEPARATION

Tomohiro Nakatani  Shoko Araki  Takuya Yoshioka  Masakiyo Fujimoto

NTT Communication Science Laboratories, NTT Corporation
2-4, Hikaridai, Seikacho, Sorakugun, Kyoto, 619-0237 Japan

ABSTRACT

This paper discusses a multichannel source separation approach that exploits the statistical characteristics of source location cues characterized by steering vector models (SM) and those of source log spectra characterized by hidden Markov models (spectral HMM). Recently, it was shown that the use of speaker independent spectral HMMs trained in advance substantially improves the quality of speech signals separated based on source location cues in a computationally efficient manner. However, with this approach, mismatches between the spectral HMMs and the observation may substantially degrade the separation quality, which limits the applicability of this approach. To overcome this problem, this paper proposes a method for learning the parameters of the spectral HMMs jointly with those of the SMs from the observed sound mixtures. Experimental results show that the proposed method works effectively for separation of convolutive sound mixtures.

Index Terms: Source separation, unsupervised learning, hidden Markov model, steering vector, log power spectrum

1. INTRODUCTION

Source separation is a technique for estimating individual source signals included in sound mixtures. If we can realize accurate and computationally efficient algorithms for this problem, it will greatly enlarge the application areas of speech signal processing, such as speech enhancement and automatic speech recognition (ASR).

A number of attempts have been made to solve this problem using microphone arrays. One promising approach is based on binary/soft mask estimation (e.g., [1, 2]), where signals are assumed to be sparse so that only one source is dominant at each time-frequency (TF) bin. Source separation is realized by clustering TF bins into individual sources based on their source location cues, such as interchannel phase difference, and many algorithms that can work in computationally efficient ways have been presented. However, because no spectral characteristics are taken into account with this approach, the quality of the separated signals is not necessarily good enough for certain speech applications, such as ASR.

In contrast, it has also been reported that high quality source separation can be accomplished by utilizing statistical source models defined in the log spectral domain and represented by hidden Markov source models (spectral HMM) even from monaural sound mixtures [3, 4]. Simulation experiments have shown that this approach substantially improves the ASR performance for mixed speech signals when used as a front-end [5]. However, this approach requires a huge computational cost because we need to find a good combination of HMM state sequences over all sources. A relatively efficient search algorithm has been proposed [5], however, it is still computationally very expensive. Furthermore, this approach assumes that speaker-dependent source models are available, but this assumption may be impractical for many applications.

Recently, a new approach has been proposed designed to take advantages of the two approaches described above [6]. This approach is referred to as dominance based locational and powerspectral characteristics integration (DOLPHIN) in this paper. With DOLPHIN, speaker-independent spectral HMMs are incorporated and utilized jointly with the source location cues to define a combined optimization criterion for multichannel source separation. The spectral shapes of the source signals can be more accurately estimated with the use of their statistical characteristics than without it. For the computationally efficient decoding of the HMMs, an iterative optimization scheme was introduced based on the expectation maximization (EM) algorithm, where the HMM state sequence of each source can be updated in M-steps based on separate Viterbi decoding. In addition, the use of location cues enables the iterative optimization to avoid effectively the local stationary points that are far from the globally optimal point.

With DOLPHIN, however, certain spectral HMMs with prior training must be provided in advance. In addition, even with spectral HMMs, mismatches between the spectral HMMs and the observation may degrade the separation quality. Mismatches may occur in various aspects of captured utterances, such as in signal levels, channel responses, and speaker characteristics, which cannot be determined in advance in many cases.

To overcome this problem, this paper proposes a method for learning the parameters of the spectral HMMs for respective speakers in an unsupervised manner from the observed sound mixtures during source separation processing. The statistical characteristics of captured utterances for each speaker can be estimated based mainly on the TF bins where the speaker’s utterances dominate in the sound mixtures. The learning is accomplished based on a maximum a posteriori (MAP) estimation given the observation, where the prior of the HMMs can also be obtained from the observed sound mixtures. Experiments show that the proposed method achieves much better quality separated signals even without prior training than the state-of-the-art source separation method based on source location cues when only a few seconds of observed sound mixture is available.

2. OVERVIEW OF DOLPHIN

Suppose \( n = 1, \ldots, N_s \) and \( k = 1, \ldots, N_h \) are time and frequency indices of a TF bin, and that \( N_s \geq 2 \) speech signals \( s_{n,k}^{(1)}, \ldots, s_{n,k}^{(N_s)} \) are mixed and the observed signal \( x_{m,k}^{(m)} \) at a microphone indexed by \( m = 1, \ldots, N_m \) is \( x_{m,k}^{(m)} = \sum_{l=1}^{N_s} h_{l,k}^{(m)} s_{l,k}^{(l)} \), where \( h_{l,m}^{(l,m)} \) is a channel response from the \( l \)-th source to the \( m \)-th microphone at a frequency \( k \). In this paper, for the sake of simplicity, we assume the number of microphone \( N_m \) to be two.

As the observed features, this paper uses the observed signals...
2.2. Model for source spectra – hidden Markov model (HMM)

A spectral HMM is characterized by a state sequence denoted by \( q^{(l)} = \{ q_{n,1}^{(l)}, q_{n,2}^{(l)}, \ldots \} \) including the initial state \( q_{0,1}^{(l)} \), the probability of each initial state \( i \) denoted by \( \pi_i^{(l)} = p(q_{1}^{(l)} = i) \), the state transition probability from \( i \) to \( j \) denoted by \( \alpha_{ij}^{(l)} = p(q_{n+1}^{(l)} = j | q_{n}^{(l)} = i) \), and the output pdf at each state \( i \) defined as

\[
\beta_{i,k}(S_{n,k}^{(l)}) = p \left( S_{n,k}^{(l)} | q_{n+1}^{(l)} = i \right); \quad \mathcal{N}(\phi_{n,k}^{(l)}, \mu_{n,k}^{(l)}, \sigma_{n,k}^{(l)})
\]

where \( \mathcal{N}(\cdot; \mu, \sigma) \) is a Gaussian pdf with mean \( \mu \) and a variance \( \sigma \). Hereafter, \( \psi \) is referred to as a set of model parameters including \( \pi_i^{(l)}, \alpha_{ij}^{(l)}, \mu_{n,k}^{(l)}, \text{and} \sigma_{n,k}^{(l)} \) for all states \( i, j, \) and \( l \).

We adopt the log-max model [3] as an effective way to represent the relationship between the source signal \( S_{n,k}^{(l)} \) and the observed signal, \( X_{n,k}^{(l)} \). This model assumes \( S_{n,k}^{(l)} = S_{n,k}^{(l)}(q_n) \), where \( l_{n,k} \) is the dominant source index and \( S_{n,k}^{(l)} = \max \{ (\gamma_{n,l}^{(1)})^{(l)}(n,k), \ldots, (\gamma_{n,L}^{(1)})^{(l)}(n,k) \} \). Then, the joint pdf of \( X_{n,k}^{(l)} \) and \( l_{n,k} \) given the HMM states, \( q_n = \{ q_{n,1}^{(l)}, q_{n,2}^{(l)}, \ldots, q_{n,N} \} \), for all sources at \( n \) is derived as

\[
p(X_{n,k}^{(l)}, l_{n,k} | q_n; \psi) = \beta_{q_{n,k}}(X_{n,k}^{(l)}) \prod_{i \neq q_{n,k}} \beta_{i,k}(S_{n,k}^{(l)})dS,
\]

2.3. Model parameter estimation using EM algorithm

As discussed in [6], when the model parameters of the HMM, \( \psi \), are given in advance based on prior training, the other model parameters of DOLPHIN, \( \phi \) and \( \{ q_n \} \), can be estimated in a computationally efficient manner based on the EM algorithm, assuming the dominant source indices \( l_{n,k} \) to be hidden variables. Letting \( E_i(\cdot; \psi) \{ \cdot \} \) be a posterior expectation function given the parameter estimates, \( \phi' \) and \( \{ q_n \} \), obtained in the previous EM step, the Q-function for the next step can be derived as

\[
Q = E_i(\phi'; \psi_{n-1}) \{ \log p \left( \{ A_{n,k} \}, \{ X_{n,k} \}, \{ l_{n,k} \}, \{ q_n \}; \phi, \psi \} \}
= \sum_l Q_l^{(1)} + \sum_{l,n} \sum_k M_{n,k}^{(1)} \log \gamma_{l,k}^{(1)}(A_{n,k}), \quad (3)
\]

\[
Q_l^{(1)} = \log(\pi_{i_0}^{(1)}) + \sum_n \log(\alpha_{q_{n-1}^{(1)}, q_{n}^{(1)})} + \sum_n \sum_k Q_{l}^{(1)}
\]

\[
Q_{l}^{(1)} = M_{n,k}^{(1)} \log(\beta_{l,k}^{(1)}(X_{n,k}))
\]

\[+(1 - M_{n,k}^{(1)}) \log \int_{-\infty}^{\infty} \beta_{l,k}^{(1)}(S)dS
\]

\[
M_{n,k}^{(1)} = p(l_{n,k} = l | \{ X_{n,k} \}, \{ A_{n,k} \}, \{ q_n \}; \phi', \psi)
\]

Note that (3) is the sum of \( Q_l^{(1)} \) and \( \gamma_{l,k}^{(1)}(A_{n,k}) \) for all \( l \), each of which only contains either \( \{ q_n^{(1)} \} \) or \( \phi_k^{(1)} \) of a source \( l \). This means that in the M-step, \( \{ q_n^{(1)} \} \) and \( \phi_k^{(1)} \) can be updated separately for each source without considering their combination over different sources. In particular, \( \phi_k^{(1)} \) for each \( l \) can be updated based on Viterbi decoding only using \( Q_l^{(1)} \). In the E-step, on the other hand, \( M_{n,k}^{(1)} \) is updated based on (6), which does not require much computational cost. As a whole, based on the EM algorithm, \( \{ q_n^{(1)} \} \) and \( \{ \phi_k^{(1)} \} \) can be estimated in a computationally efficient manner.

Concrete update equations for \( \{ q_n^{(1)} \} \) and \( \{ M_{n,k}^{(1)} \} \) can be found in [6], and those for \( \phi_k^{(1)} \) can be found in [2].

Fig. 1. Graphical model of DOLPHIN.
3. EXTENSION OF DOLPHIN

In the following, we propose a method for learning the spectral HMM parameters, $\psi$, from the observed mixtures in the course of the EM iteration. By this extension, DOLPHIN becomes applicable to cases where there are substantial mismatches between the spectral HMMs and the observed signals, and/or even to cases where no spectral HMMs with prior training are available.

3.1. Updates of HMM model parameters $\psi$

In the proposed method, $\psi$ is updated separately for each source $l$ in the M-step by maximizing the Q-function (3), or more concretely (4) and (5), after the optimal Viterbi path $\{q_0^{(l)}\}$ is fixed. The way of updating $\pi_i^{(l)}$ and $\alpha_i^{(l)}$ is straightforward, so this paper skips any discussion of this. On the other hand, for the updating of $\mu_{i,k}^{(l)}$ and $\sigma_{i,k}^{(l)}$, the treatment of the second term of (5) is rather complicated. So, the problem here is how to deal with the second term in an efficient manner for the optimization. For this purpose, we introduce two approximation functions for the second term.

Letting $x = (X_{n,k} - \mu_{i,k}^{(l)})/\sigma_{i,k}^{(l)}$, we define a normalized form of the second term as:

$$f(x) = \log \int_{-\infty}^{\infty} \mathcal{N}(s; 0, 1) ds$$

Our preliminary experiments showed that the two functions below well approximate the derivative of $f(x)$, or $df(x)/dx$, as in Fig. 2.

$$g_1(x) = \log(1 + \exp(-x))$$
$$g_2(x) = \begin{cases} -x & \text{for } x < 0 \\ 0 & \text{otherwise} \end{cases}$$

We adopt $g_1(x)$ and $g_2(x)$, respectively, for the updating of $\mu_{i,k}^{(l)}$ and $\sigma_{i,k}^{(l)}$, by which the following advantages can also be obtained.

1. With $g_1(x)$, the convergence of the Newton method is globally guaranteed for the updating of $\mu_{i,k}^{(l)}$. (Proof omitted)

2. $g_2(x)$ is in the same form as the derivative of the first term of (5) for $x < 0$ while it can be disregarded for $x \geq 0$. So, updating of $\sigma_{i,k}^{(l)}$ can be straightforward even with the second term of (5).

As a result, the updating of $\mu_{i,k}^{(l)}$ and $\sigma_{i,k}^{(l)}$ can be computationally very efficient.

Omitting indices, $i$, $n$, $k$ and $l$, from variables, $M_{i,n,k}^{(l)}$, $X_{n,k}$, $\mu_{i,k}^{(l)}$, $\sigma_{i,k}^{(l)}$, $\tilde{\mu}_{i,k}^{(l)}$, and $\tilde{\sigma}_{i,k}^{(l)}$, for brevity, where $\mu_{i,k}^{(l)}$ and $\sigma_{i,k}^{(l)}$ are estimates of $\mu_{i,k}^{(l)}$ and $\sigma_{i,k}^{(l)}$ obtained in the previous EM step, and $\tilde{\mu}_{i,k}^{(l)}$ and $\tilde{\sigma}_{i,k}^{(l)}$ are parameters for a prior term discussed in the next subsection, then the update equations in the M-step become as follows:

$$\mu_{i,k}^{(l)} \leftarrow \mu' = \mu + \frac{\sum_{i,j} \frac{M_{i,n,k}^{(l)}}{\sigma_i^{(l)}} \left( \frac{X_{n,k} - \mu'}{\sigma_i^{(l)}} \right) - \rho \frac{\mu' - \tilde{\mu}}{\sigma_i^{(l)}}}{\sum_{i,j} \frac{M_{i,n,k}^{(l)}}{\sigma_i^{(l)}}}$$
$$\sigma_{i,k}^{(l)} \leftarrow \sigma' = \sigma + \frac{\sum_{i,j} \frac{M_{i,n,k}^{(l)}}{\sigma_i^{(l)}} \left( \frac{X_{n,k} - \mu'}{\sigma_i^{(l)}} \right)^2}{\sum_{i,j} \frac{M_{i,n,k}^{(l)}}{\sigma_i^{(l)}}}$$

where $v_i^{(l)}$ is a set of time indices $n$ that satisfy $q_{0,n}^{(l)} = i$, $g_1^{(l)}(x)$ is the derivative of $g_1(x)$, $\rho$ is the weight of the prior term, and $\kappa_{i,n,k}^{(l)} = M_{i,n,k}^{(l)}$ if $X_{n,k} > \mu_{i,k}^{(l)}$ and $\kappa_{i,n,k}^{(l)} = 1$ otherwise.

3.2. Initialization of spectral HMMs and prior of $\mu_{i,k}^{(l)}$

When any spectral HMMs with prior training are available, they can be used as the initial values of the spectral HMMs for unsupervised learning. Otherwise, we need to initialize the spectral HMMs based only on the observation. In our experiments, a Gaussian mixture model (GMM) was trained on the observed spectra, $X_{n,k}$, and the obtained GMM was commonly used for the initial values of all the spectral HMMs, assuming $\pi_i^{(l)}$ and $\alpha_i^{(l)}$ to be uniformly distributed.

However, even with good initialization, unsupervised learning may degrade the accuracy of the model parameters on certain spectral features when such features are hardly dominant in short observations. In order to avoid this problem, we define prior pdfs of $\mu_{i,k}^{(l)}$ as follows, and add them to $Q_i^{(l)}$ in (4) with a certain weight $\rho$.

$$\log p(\{\mu_{i,k}^{(l)}\}) = \sum_{i} \sum_{k} \log \mathcal{N}(\mu_{i,k}^{(l)}; \tilde{\mu}_{i,k}^{(l)}, \tilde{\sigma}_{i,k}^{(l)})$$

Our preliminary experiments showed that it was effective to determine $\mu_{i,k}^{(l)}$ and $\sigma_{i,k}^{(l)}$ as the mean and variance of each Gaussian in the initialized spectral HMM.

3.3. Processing flow

The following summarizes the overall source separation procedure.

1. Initialize the spectral HMMs. (See Section 3.2)

2. Initialize $M_{i,n,k}^{(l)}$ (See Section 2.4 in [6])

3. Iterate the following until convergence is achieved.

   (a) Update $\{q_{0,n}^{(l)}\}$ for each $l$. (See [6])
   (b) Update $\psi$. (See Section 3.1)
   (c) Update $\phi_k^{(l)}$ for each $l$ and each $k$. (See [2])
   (d) Update $M_{i,n,k}^{(l)}$ for all $n$, $k$, and $l$. (See [6])

4. Estimate $S_{i,n,k}^{(l)}$ for all $n$, $k$, and $l$. (See Section 2.5 in [6])

4. EXPERIMENTS

To evaluate the separation performance, we prepared three test data sets composed with three different room impulse responses (RIR), and referred to as RIR-1, RIR-2, and RIR-3. Each test data set contained 30 mixed utterances, where each mixed utterance contained $N_s = 2$ simultaneous utterances by two of four speakers (4, 5, 6, and 7) in [7]. The average length of a mixed utterance was 2.0 s. With RIR-1, each mixed utterance was generated by simply mixing two utterances with no reverberation. The time differences of arrival of the two utterances were set at $\pm 1.5$ ms. RIR-3 was measured in a room with $RT_{60} = 0.1$ s using a dummy head microphone, and assuming the two sources to be at 30 and 120 degrees, respectively, from the right hand side of the dummy head. RIR-2 was the same as RIR-3 except that it was truncated at a length of 6.3 ms. The
sampling rate was set at 16 kHz. The state number of the spectral HMMs was set at 4, where all the states were fully connected. This HMM structure is shown in [6] to be effective for source separation by DOLPHIN. For the model parameter estimation, the EM iteration number was fixed at 20. The frame size and shift were set at 40 ms and 10 ms, respectively.

4.1. Separated signal quality depending on observation length

Figure 3 shows the quality of the separated signals in terms of the average cepstral distortion (CD) calculated over the 0th to 12th order cepstral coefficients, depending on the length of the observation. The numbers, “1” to “10”, on the horizontal axis were the numbers of mixed utterances used for the model parameter estimation and source separation. The average CDs of the mixed utterances are also shown as “Baseline”. The CDs obtained by DOLPHIN with and without prior training, and those obtained by the method proposed in [2], referred to as SM, are shown for comparison. For DOLPHIN without prior training, the spectral HMMs were initialized using GMMs trained on the observation (see Section 3.2). For DOLPHIN with prior training, we prepared three speaker independent HMMs, respectively, for three test data sets, so that the HMMs were trained under matched RIR conditions. For this purpose, a set of training data were extracted from [7] excluding the speakers in the test data set, convolved with RIR-1, RIR-2, and RIR-3, and used for the prior training. During the model parameter estimation for the source separation, $\psi$ was fixed as in [6] for DOLPHIN with prior training.

According to the figure, DOLPHIN greatly outperformed SM in all cases even without any prior training. In addition, the CDs of signals separated by DOLPHIN without prior training decreased more as the increase of the number of mixed utterances used for the model parameter estimation. In particular, in all cases, the CDs obtained with DOLPHIN without prior training were very close to those obtained with DOLPHIN with prior training, which contains no channel mismatches. This means that the proposed method can appropriately extract the statistical characteristics of speech spectra from only a short observation. By contrast, CDs obtained with SM and DOLPHIN both increased as the reverberation got longer. This suggests that the performance of DOLPHIN in the presence of reverberation depends largely on that of SM. Therefore, we consider that it is important to incorporate a model of reverberation into DOLPHIN to handle the reverberation more appropriately. This should be further investigated in future work.

4.2. Examples of estimated HMM means

Figure 4 shows the means of the spectral HMM obtained by the prior training with RIR-1 and those of the spectral HMM estimated, without prior training, from two mixed utterances with RIR-1. From these results, we can roughly confirm the appropriateness of the estimation when the means are estimated from such a short sound mixture.

5. SUMMARY

This paper proposed a new blind multichannel source separation method that exploits the statistical characteristics of source log-spectra and locations. The two sets of statistical characteristics are estimated from observed sound mixture using spectral HMMs and steering vector models, respectively, in a computationally efficient manner without any prior knowledge about them. In the experiments, the proposed method greatly outperformed the state-of-the-art source separation method based solely on location information, in terms of cepstral distortion, even when only a few seconds of observation was available for the blind estimation.

6. REFERENCES