TARGET SPEECH EXTRACTION WITH LEARNED SPECTRAL BASES

Sunho Park, Jiho Yoo, and Seungjin Choi

Department of Computer Science, POSTECH, Korea
{titan,zentasis,seungjin}@postech.ac.kr

ABSTRACT

In this paper we present a method for extracting a speech signal of target speaker from noisy convolutive mixtures of target speech and an interference source, when training utterances of the target speaker are available. We incorporate a statistical latent variable model into blind source separation (BSS), where we make use of spectral bases learned from the training utterances of the target speaker to identify which source corresponds to the target speaker. Combined with any existing BSS methods, our post-processing (which is the main contribution) consists of two steps: (1) channel selection where we identify the source corresponding to the target speaker; (2) enhancement where we further suppress the remaining interference. Numerical experiments confirm that our method substantially improves the separation quality of existing BSS methods and successfully restores the target speaker's speech.

Index Terms— Blind source separation, speech extraction, speech segregation

1. INTRODUCTION

The extraction of a target speaker's speech from noisy mixtures in real-world environment is a challenging problem. Such a task involves convolutive blind source separation (BSS) as well as denoising to remove background noise. For instance, an automatic speech recognition or a voice command system requires the enhanced target speaker's speech in the presence of noise and interfering sound, in order to improve the recognition performance. BSS can be directly applied to solve this problem [13, 14, 18, 1, 5]. However, due to inherent ambiguities in BSS (ordering and scale ambiguities), it is not possible to identify which is a signal of interest even after mixtures are separated out. In practice, the effect of interference source often remains up to some extent after BSS is applied, although a perfect separation is expected theoretically. The performance of blind technique (which does not resort to any prior knowledge or side information) is limited in practice. In the case where some side information is available, it is desirable to incorporate such information into BSS to improve the final performance.

There are a few work on the target source extraction [16, 12]. In [12], the problem of target source extraction was considered for the case where interference is babble noise. BSS was applied to separate mixtures out, followed by a kurtosis-based selection was used to choose the target source. In [16], the target source was assumed to have dominant power. The selection of target source was formulated as a permutation problem, where the power dominance was used to select the target source.

Compared to these existing methods, our method directly makes use of the characteristics of target speaker's speech. We assume that a target speaker's training utterances are available, which are not necessarily overlapped with speech signals that are supposed to be separated out. From the training utterances, we learn spectral bases by nonnegative matrix factorization (NMF) of spectrograms or a statistical latent variable model. The main contribution of this paper is in the post-processing of BSS where we use spectral bases learned by a statistical latent variable model in order to identify the source corresponding to the target and to enhance the target speech in the presence of background noise. There exist some post-processing methods for pursuing the same goal but with different approaches [20, 9]. Our post-processing consists of two steps: (1) channel selection step where we identify the source corresponding to the target speaker; (2) enhancement step where we further suppress the remaining interference. These steps are carried out, making use of learned spectral bases. We show that our method substantially improves the separation quality of existing BSS methods and successfully restores the target speaker's speech.

2. PROBLEM FORMULATION

We consider a noisy environment where an interference source (e.g., noise from a printer or music sound from a radio) is convolved with a target speech and a background noise also exists, as shown in Fig. 1. Measurement signals picked up by microphones are assumed to be generated by

$$\boldsymbol{x}_t = \sum_{\tau=0}^{P} \boldsymbol{A}_\tau \boldsymbol{s}_{t-\tau} + \boldsymbol{n}_t, \tag{1}$$

where $\boldsymbol{x}_t = [x_{1,t}, ..., x_{m,t}]^\top \in \mathbb{R}^m$ is an m-dimensional observation vector at time t, $\boldsymbol{s}_t = [s_{1,t}, ..., s_{n,t}]^\top \in \mathbb{R}^n$ is an n-dimensional source vector, $\{\boldsymbol{A}_\tau\}$ is a set of multivariate FIR filter coefficients which models multipath propagation in a room, P is the filter length, and $\boldsymbol{n}_t$ is the background noise (which is typically assumed to be white Gaussian noise).

In this paper we particulary consider the case where a target speaker's speech is convolved with an interference source ($m = n = 2$), in the presence of background noise. The task of target speech extraction is to restore the source corresponding to the target speech from mixtures measured at microphones, in the presence of noise, while methods of BSS restore all independent sources from mixtures, without knowing which source corresponds to the target speech. Therefore, the target speech extraction requires some side information to identify the source corresponding to the target speech. The power dominance of the target source can be used to select the target source from sources restored by BSS, as in [20]. Such prior knowledge is often not satisfied in practice. In this paper, we accomplish the target speech extraction, through learning spectral characteristics by a statistical latent variable model from training utterances of target speaker. The schematic diagram of our proposed method is shown in Fig. 1, where the role of each module is briefly described in figure caption.

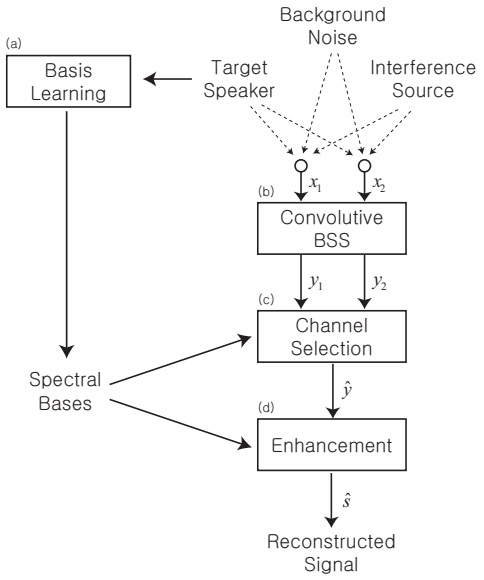


Fig. 1. The schematic diagram of our proposed method for target speech extraction is shown: (a) learning spectral bases given training utterances described in Sec. 3; (b) convolutive BSS to restore independent sources from convolutive mixtures, where we used methods exploiting nonstationarity [13, 14, 6]; (c) channel selection takes two outputs y_1 and y_2 (outputs of BSS) as inputs to identify the source corresponding to the target speaker, yielding an estimate of target speech, $\widehat{y}$, described in Sec. 4.1; (d) enhancement where we suppress the interference and noise remaining in $\widehat{y}$, yielding the final estimate $\widehat{s}$ of target speech, described in Sec. 4.2.

3. LEARNING SPECTRAL BASES

We denote by $\boldsymbol{M} \in \mathbb{R}^{F\times T}$ a spectrogram where each row is associated with a frequency profile which reflects how the power of a spectral component (out of F frequency bins) varies across T time frames, i.e., M_{ij} represents the power of spectral component i in time frame j. Learning spectral bases from the spectrogram $\boldsymbol{M}$ involves the following matrix decomposition,

$$\boldsymbol{M} \approx \boldsymbol{U}\boldsymbol{V}, \tag{2}$$

where $\boldsymbol{U} \in \mathbb{R}^{F\times K}$ is the *basis matrix* containing spectral bases in its columns, $\boldsymbol{V} \in \mathbb{R}^{K\times T}$ is the associated *encoding matrix*, and K corresponds to the intrinsic dimension (the number of latent variables). The decomposition (2) is solved by determining factor matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ which minimizes a discrepancy measure between the data $\boldsymbol{M}$ and the model $\boldsymbol{U}\boldsymbol{V}$.

Nonnegative matrix factorization (NMF) [10] is an appropriate technique to learn the decomposition (2) since the spectrogram is a nonnegative matrix and nonnegativity constraints imposed on factor matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ yield a fruitful representation. NMF was successfully applied to learn spectral bases from audio for sound classification [4] Spectral bases learned from EEG data were shown to be useful in extracting discriminative features for EEG classification [11]. NMF is closely related to probabilistic latent semantic analysis (PLSA) [8] where a statistical latent variable model, known as *aspect model*, is learned by EM optimization [7]. The aspect model was recently extended, relating the generalization to probabilistic matrix tri-factorization [21].

A speaker-specific latent variable was introduced on the top of the aspect model, in order to tackle a problem of single channel speaker separation [15]. The model in [15] treated the spectrogram $\boldsymbol{M}$ as *dyadic data*, where the (i, j)-entry of the spectrogram was interpreted as the count of frequency i in time frame j. It was also further elaborated with an overcomplete representation [17], incorporating the entropic prior [2]. We also adopt the probabilistic model in [15, 17] to learn spectral bases.

4. POST-PROCESSING WITH SPECTRAL BASES

We provide a detailed description of our post-processing method involving channel selection and enhancement, each of which is explained in Sec. 4.1 and Sec. 4.2, respectively. The core ingredient in our post-processing method is spectral bases $\boldsymbol{U}$ learned from training utterances of target speaker. We denote by $\widehat{\boldsymbol{M}}^{(1)} \in \mathbb{R}^{F\times\widehat{T}}$ and $\widehat{\boldsymbol{M}}^{(2)} \in \mathbb{R}^{F\times\widehat{T}}$ the spectrograms of y_1 and y_2 that are sources recovered by a BSS method, respectively.

4.1. Channel Selection

Channel selection aims at identifying the source corresponding to the target speaker, given sources restored by a BSS algorithm. To this end, we first apply the probabilistic decomposition method [15] to spectrograms $\boldsymbol{M}^{(1)}$ and $\boldsymbol{M}^{(2)}$.

The tth column of a spectrogram is interpreted as the histogram of an independent set of draws from an underlying multinomial distribution $P_t(f)$ over F discrete values. With an appropriate normalization, we can relate columm t of the spectrogram to $P_t(f)$. Let $s \in \{s^*, s^\sharp\}$ be a latent variable indicating target speaker (s^*) or interference source ($s^\sharp$). We denote by $z \in \{\boldsymbol{z}_s\}$ be a latent variable (specific to source s) which manipulating the generation of frequency f. Then, $P_t(f)$ is decomposed as [15]

$$P_t(f) = \sum_{s\in\{s^*,s^\sharp\}} P_t(s) \sum_{z\in\boldsymbol{z}_s} P_t(z|s)P_s(f|z), \tag{3}$$

where $P_t(s)$ indicates a priori probability of s-th source, $P_t(z|s)$ is a mixing weight varying to time and $P_s(f|z)$ is a basis function. In eq. (3), we use the learned bases in Sec. 3 as the basis function for s^*, i.e., $P_{s^*}(f|z) = [\boldsymbol{U}]_{f,z}$, where $[\boldsymbol{U}]_{f,z}$ is an (f,z) element of $\boldsymbol{U}$, where z is treated as an index. Thus the parameters to be estimated here are $\theta = \{P_t(s), P_t(z|s), P_{s^\sharp}(f|z)\}$. They are efficiently estimated by an expectation and maximization (EM), where the E-step estimates the posterior $P_t(s, z|f)$ while M-step updates the parameters θ [15]:

- E-step

$$P_t(s, z|f) = \frac{P_t(s)P_t(z|s)P_s(f|z)}{\sum_{s'} P_t(s') \sum_{z'\in\boldsymbol{z}_s} P_t(z'|s')P_{s'}(f|z')},$$

- M-step

$$\begin{aligned}
P_t(s) &= \frac{\sum_{z\in\boldsymbol{z}_s}\sum_f P_t(s, z|f)[\widehat{\boldsymbol{M}}]_{f,t}}{\sum_{s'}\sum_{z\in\boldsymbol{z}_s}\sum_f P_t(s, z|f)[\widehat{\boldsymbol{M}}]_{f,t}},\\
P_t(z|s) &= \frac{\sum_f P_t(s, z|f)[\widehat{\boldsymbol{M}}]_{f,t}}{\sum_{z'\in\boldsymbol{z}_s}\sum_t P_t(s, z'|f)[\widehat{\boldsymbol{M}}]_{f,t}},\\
P_{s^\sharp}(f|z) &= \frac{\sum_t P_t(s^\sharp, z|f)[\widehat{\boldsymbol{M}}]_{f,t}}{\sum_{f'}\sum_t P_t(s^\sharp, z|f')[\widehat{\boldsymbol{M}}]_{f',t}}.
\end{aligned}$$

We apply the above EM method to $\widehat{M}^{(1)}$ and $\widehat{M}^{(2)}$ in order to obtain $\widehat{\theta}^{(1)}$ and $\widehat{\theta}^{(2)}$ respectively, where $\widehat{\theta}^{(i)}$ means the parameters estimated from $\widehat{M}^{(i)}$. After the decompositions of $\widehat{M}^{(1)}$ and $\widehat{M}^{(2)}$, the priori probability of the target speaker s^* is averaged over the time frames as

$$\hat{P}^{(i)}(s^*) = \frac{1}{\hat{T}} \sum_{t=1}^{\hat{T}} P_t^{(i)}(s^*). \tag{4}$$

and used as the measure for the channel selection. Intuitively we can conclude that the larger $\hat{P}(s^*)$ in (4), the higher contribution of the target speech to the spectrogram. When $\widehat{P}^{(1)}(s^*)$ and $\widehat{P}^{(2)}(s^*)$ are calculated, the criterion of channel selection is presented by

$$\left\{\widehat{\boldsymbol{M}}, \widehat{\theta}\right\} = \begin{cases} \left\{\widehat{\boldsymbol{M}}^{(1)}, \widehat{\theta}^{(1)}\right\} & \text{if} \quad \hat{P}^{(1)}(s^*) \geq \hat{P}^{(2)}(s^*), \\ \left\{\widehat{\boldsymbol{M}}^{(2)}, \widehat{\theta}^{(2)}\right\} & \text{otherwise.} \end{cases} \tag{5}$$

4.2. Enhancement

We recover the spectrogram for the target speaker based on $\left\{\widehat{\boldsymbol{M}}, \widehat{\theta}\right\}$ and the learned spectral bases $(P_{s^*}(f|z) = [\boldsymbol{U}]_{f,z})$. The clean magnitude of the spectrogram $[\widehat{\boldsymbol{M}}^*]_{f,t}$ for the target speaker is estimated by a mean value of binomial distribution $\mathcal{B}\left([\widehat{\boldsymbol{M}}]_{f,t}, P_t(s^*|f)\right)$ [15]:

$$[\widehat{\boldsymbol{M}}^*]_{f,t} = \frac{\widehat{P}_t(s^*)\widehat{P}_t(f|s^*)}{\sum_{s'} \widehat{P}_t(s')\widehat{P}_t(f|s')} [\widehat{\boldsymbol{M}}]_{f,t}, \tag{6}$$

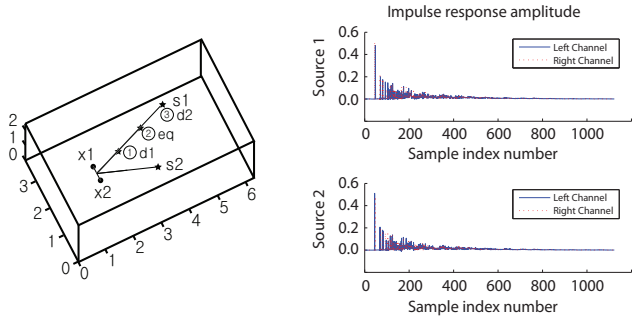
where $\widehat{P}_t(f|s^*) = \sum_{z \in \boldsymbol{z}_{s^*}} \widehat{P}_t(z|s^*) P_{s^*}(f|z)$. Only the learned bases and corresponding encodings are used to reconstruct the target speech. After this step, the remaining interference source and background noise are automatically suppressed.

Speech signal is obtained by using inverse short time Fourier transform with overlapping windows. The phase information of the mixture is used to tune the phase of the reconstructed signal.

5. NUMERICAL EXPERIMENTS

We investigate a performance of our method applied to the target speaker's speech extraction problem in the noisy reverberation environment. Speech signal for the target speaker (male or female) is convolutively mixed with the interference source in noisy environment. The interference source is set to printer noise or trumpet music sound. All signal is resampled at 8000 Hz. In order to learn the spectral bases, we use 30 seconds of the training utterances from each speaker. The spectrogram of the training utterances is generated by the short-term Fourier transformation with widow size 1024, hop size of 256 between frames and a hannig window. The settings for the latent variable model are as follows: the number of the latent variables is set to $K = 1000$; parameters for the entropic prior are set to the values in [17].

In the extraction experiments, 5 seconds of the test speech signal and of the interference source are convolutively mixed in noisy environment. Roomsim [3] is used to generate the convolutive mixtures of the sources. The size of room is set to 6.25 m (width) by 3.75 m (depth) by 2.5 m (height). The 1/4 width and 1/2 depth point of the room is served as a reference point. The interference source is located 2 m away from the reference point. The location of the target source is 1) 1 m, 2) 2 m, and 3) 3 m away from the reference point, each represents target dominant case (d1), approximately equal case (eq), and interference dominant case (d2) respectively. The inner angle of the target source and the interference source is 30 degree. The overall settings are illustrated in Fig. 2(a). The resulting room impulse response are displayed in Fig. 2(b), which shows that the length of the mixing filters are 1123 in 8000 Hz environment. White Gaussian background noise is added to the obtained convolutive mixture.



(a) Room (b) Simulated room impulse response

Fig. 2. Simulation settings in Roomsim.

A source to distortion ratio (SDR) [19] is used to investigate the quality of the separation results in noisy environment. The SDR is a measure representing the ratio between target source part and unwanted signal part in the estimated signal. We used BSS_EVAL toolbox [19] to decompose given signal to the target source part and other signal part. The table 1 shows the SDRs in both cases of the female and the male speaker. We evaluate SDRs at each step of the experiments (see Fig. 1): SDR_x for the value of SDR of the mixture, where $SDR_x = \max(SDR_{x1}, SDR_{x2})$; $SDR_{\hat{y}}$ for the value of SDR of the output from the channel selection; $SDR_{\hat{s}}$ for the value of SDR of the estimated target speaker' speech from the enhancement. Note that, BSS without our postprocessing can not give SDRs superior to $SDR_{\hat{y}}$. From the SDR values of the estimated target speech, $SDR_{\hat{s}}$, of all experiments, we found that our method substantially improves the separation quality of BSS.

Table 1. The SDR values at each step (refer Fig. 1). Pr and tr indicate the interference sources, pr (printer noise) and tr (trumpet music sound) while eq, d1 and d2 mean the dominance of the target speaker respect with to the interference source (see Fig. 2(a)).

	SDR_x		$SDR_{\hat{y}}$		$SDR_{\hat{s}}$	
	female	*male*	*female*	*male*	*female*	*male*
pr-eq	0.53	-1.59	12.55	1.08	16.76	5.68
pr-d1	3.39	1.09	5.08	2.17	17.50	6.85
pr-d2	-1.38	-3.16	9.14	0.85	13.80	4.08
tr-eq	0.36	-1.13	3.89	-3.14	16.66	2.87
tr-d1	2.83	1.18	10.30	1.73	18.39	7.77
tr-d2	-1.93	-3.32	-0.94	-3.81	6.78	2.85
average	0.63	-1.16	6.67	-0.19	14.98	5.02
increase	-	-	6.04	0.97	8.31	5.20

6. CONCLUSIONS

We have presented a method of extracting target speech from convolved mixtures with interference sound or noise, whereas general BSS aimed at restoring every sources from their mixtures. A statistical latent variable model was learned to capture the characteristics of target speech, given training utterances of target speech (side information) that were not necessarily overlapped with speech in consideration when BSS was applied. Spectral bases learned by the overcomplete aspect model with sparse prior led us to identify which restored source corresponds to target speech and to suppress remaining interference as well. The main contribution of this paper was in the post-processing of BSS, consisting of two steps: (1) channel selection where we identified the source corresponding to the target speaker; (2) target speech enhancement where we further suppressed the remaining interference. Numerical experiments in real-world noisy environment confirmed that our post-processing indeed much improved the performance, restoring much more clean target speech, compared to the one determined by the conventional BSS which still suffered from remaining interference and noise. Our post-processing works well only when the spectral profiles of target speech differ from those of an interference. In the case where one male (or female) speaker is associated with target speech and the other male (female) speaker corresponds to interference, the performance of our post-processing is degraded, since spectral profiles of male speakers' speech are similar each other. This difficult case is a challenging problem that we are currently working on.

Acknowledgments: This work was supported by Korea Ministry of Knowledge Economy under the ITRC support program supervised by the IITA (IITA-2008-C1090-0801-0045) and KOSEF WCU Program (Project No. R31-2008-000-10100-0).

7. REFERENCES

[1] S. Amari, S. C. Douglas, A. Cichocki, and H. H. Yang, "Multichannel blind deconvolution and equalization using the natural gradient," in *Proceedings of the IEEE International Conference on Signal Processing Advances in Wireless Communications (SPAWC)*, Paris, France, 1997, pp. 101–104.

[2] M. Brand, "Structure learning in conditional probability models via an entropic prior and parameter extinction," *Neural Computation*, vol. 11, no. 5, pp. 1155–1182, 1999.

[3] D. R. Campbell, K. J. Palomaki, and G. J. Brown, "A MATLAB simulation of "Shoebox" room acoustics for use in research and teaching," *Computing and Information Systems Journal*, vol. 9, no. 3, 2005.

[4] Y. C. Cho and S. Choi, "Nonnegative features of spectro-temporal sounds for classfication," *Pattern Recognition Letters*, vol. 26, no. 9, pp. 1327–1336, 2005.

[5] S. Choi, S. Amari, A. Cichocki, and R. Liu, "Natural gradient learning with a nonholonomic constraint for blind deconvolution of multiple channels," in *Proceedings of the International Conference on Independent Component Analysis and Blind Signal Separation (ICA)*, Aussois, France, 1999, pp. 371–376.

[6] S. Choi, A. Cichocki, and A. Belouchrani, "Second order nonstationary source separation," *Journal of VLSI Signal Processing*, vol. 32, pp. 93–104, 2002.

[7] E. Gaussier and C. Goutte, "Relation between PLSA and NMF and implications," in *Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval (SIGIR)*, Salvador, Brazil, 2005.

[8] T. Hofmann, "Probablistic latent semantic analysis," in *Proceedings of the International Conference on Uncertainty in Artificial Intelligence (UAI)*, 1999.

[9] J. Kocinski, "Speech intelligibility improvement using convolutive blind source separation assisted by denoising algorithms," *Speech Communication*, vol. 50, no. 1, pp. 29–37, 2008.

[10] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in Neural Information Processing Systems (NIPS)*, vol. 13. MIT Press, 2001.

[11] H. Lee, A. Cichocki, and S. Choi, "Nonnegative matrix factorization for motor imagery EEG classification," in *Proceedings of the International Conference on Artificial Neural Networks (ICANN)*. Athens, Greece: Springer, 2006.

[12] S. Y. Low, T. Roberto, and N. Sven, "Spatio-temporal processing for distant speech recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Montreal, Canada, 2004.

[13] L. C. Parra and C. Spence, "Convolutive blind source separation of non-stationary sources," *IEEE Transactions on Speech and Audio Processing*, pp. 320–327, May 2000.

[14] D. T. Pham, C. Serviere, and H. Boumaraf, "Blind separation of convolutive audio mixtures using nonstationarity," in *Proceedings of the International Conference on Independent Component Analysis and Blind Signal Separation (ICA)*, Nara, Japan, 2003, pp. 107–110.

[15] B. Raj and P. Smaragdis, "Latent variable decomposition of dpectrograms for single channel speaker separation," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2005, pp. 17–20.

[16] H. Sawada, S. Araki, R. Mukai, and S. Makino, "Blind extraction of a dominant sourcee signal from mixtures of many sources using ICA and time-frequency masking," in *Proceedings of the IEEE International Symposium on Circuits and Systems (ISCAS)*, Kobe, Japan, 2005, pp. 5882–5885.

[17] M. Shashanka and P. Smaragdis, "Sparse overcomplete decomposition for single channel speaker separation," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Honolulu, Hawaii, 2007.

[18] K. Torkkola, "Blind separation of convolved sources based on information maximization," in *Proceedings of the IEEE Workshop on Neural Networks for Signal Processing*, 1996, pp. 423–432.

[19] E. Vincent, C. Fevotte, and R. Gribonval, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[20] E. Visser, M. Otsuka, and T. W. Lee, "A spatio-temporal speech enhancement scheme for robust speech recognition in noisy environments," *Speech Communication*, vol. 41, pp. 393–407, 2003.

[21] J. Yoo and S. Choi, "Probabilistic matrix tri-factorization," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Taipei, Taiwan, 2009.