MULTI-BAND PIT AND MODEL INTEGRATION FOR IMPROVED MULTI-CHANNEL SPEECH SEPARATION

Lianwu Chen[1*], *Meng Yu*[1*], *Dan Su*[1], *Dong Yu*[2]

[1]Tencent AI Lab, Shenzhen, China [2]Tencent AI Lab, Bellevue, WA, USA

ABSTRACT

The recent exploration of deep learning for supervised speech separation has significantly accelerated the progress on the multi-talker speech separation problem. Multi-channel extension has attracted much research attention due to the benefit of spatial information in far-field acoustic environments. In this paper, We review the most recent models of multi-channel permutation invariant training (PIT), investigate spatial features formed by microphone pairs and their underlying impact and issue, present a multi-band architecture for effective feature encoding, and conduct a model integration between single-channel and multi-channel PIT for resolving the spatial overlapping problem in the conventional multi-channel PIT framework. The evaluation confirms the significant improvement achieved with the proposed model and training approach for the multi-channel speech separation.

Index Terms— speech separation, multi-channel, permutation invariant training, multi-band, model integration.

1. INTRODUCTION

Speech separation is an important task under the cocktail-party condition [1] where a set of source signals are mixed with an unspecified process and recorded at a single or array of microphones. Given the observed mixed signal, the objective is to invert the unknown mixing process and estimate the individual source signals. Great advances were observed in monaural speech separation when the problem is converted into a supervised regression problem in which the optimization objective is closely related to the separation task. The deep learning based techniques, such as deep clustering (DPCL) [2, 3], deep attractor network (DANet) [4, 5] and permutation invariant training (PIT) [6, 7], aim for solving the label permutation issue and work very well when separating multi-talker speech. It is summarized in [8] that these three approaches have strong connections, particularly, PIT is much simpler to implement, easier to integrate with other techniques, and more efficient during testing.

However, these deep learning based monaural speech separation methods are not good enough in real world applications due to the inherent limitation of the separation power. The far field speech processing suffers from the reverberation which blurs speech spectral cues and degrades the single-channel speech separation. Since the microphone array is more widely deployed than before, multi-channel techniques become more and more important.

An array of microphones provides multiple recordings, which contain information indicative of the spatial origin of a sound source. When sound sources are spatially separated, with microphone array inputs one may localize sound sources and then extract the source from the target direction. Binaural features, such as inter-channel time difference (ITD), phase difference (IPD) and level difference (ILD), all extracted from individual Time-Frequency (T-F) unit pairs, were first exploited in supervised speech segregation [9] based on the sparsity assumption of speech signal in T-F representation. The use of spatial information afforded by an array as features in deep learning is a straightforward extension to the deep learning models originally designed for monaural speech separation. Much attention has been paid to the multi-channel integration of PIT for speech separation due to its end-to-end architecture and simple training procedure in the monaural implementation [10, 11, 12].

We review and analyze the most recent multi-channel PIT approaches in Section 2 and make three contributions in this paper. First, we reveal a "spatial overlapping" issue existed in the conventional multi-channel end-to-end PIT framework that multi-channel approaches fail when source speakers are closely located (Section 3). Second, by considering phase wrapping issue in spatial features, we propose a multi-band PIT in which multi-band embeddings are generated with a multi-tower neural network in which each tower is trained for encoding features in an individual sub-band (Section 4). Third, we present a model that integrates the single-channel and multi-channel utterance-based PIT for solving the spatial overlapping problem (Section 5). The experiments are conducted in Section 6. We conclude this paper in Section 7.

2. OVERVIEW OF MULTI-CHANNEL PITS

In order to associate references of S mixing sources to the output layers in a typical monaural speech separation neural network, we compute the total mean square error (MSE) for each of $S!$ possible assignments. The assignment with the

*Both authors contributed equally to this work.

least total MSE is chosen and the model is optimized to reduce this particular MSE. The utterance-level PIT (uPIT) [7], a more effective approach to solve the tracing and label permutation problem than original frame-level PIT [6], extends the frame-level PIT technique with the following utterance-level cost function:

$$\mathcal{J}_{\phi^*} = \frac{1}{T \times F \times S} \sum_{s=1}^{S} \parallel \hat{M}_s \otimes |Y| - |X_{\phi^*(s)}| \parallel_F^2, \quad (1)$$

where $|Y|$ and $|X_s|$ are the spectrograms of the mixture and clean speech reference, respectively, $\hat{M}_s$ is the mask estimation for source s, T and F correspond to the total number of time frames and frequency bands for an utterance, respectively, ϕ^* is the permutation that minimizes the utterance-level separation error defined as

$$\phi^* = \arg\min_{\phi \in \mathcal{P}} \sum_{s=1}^{S} \parallel \hat{M}_s \otimes |Y| - |X_{\phi(s)}| \parallel_F^2, \quad (2)$$

and $\mathcal{P}$ is the set of all $S!$ permutations. With uPIT, the permutation corresponding to the minimum utterance-level separation error is used for all frames in the utterance.

A straightforward end-to-end multi-channel approach was proposed in [10]. They used the magnitude spectra from all the microphones and the IPDs between a reference microphone and each of the other microphones as the input features for training. Chen et al. [11] efficiently integrate fixed beamformers and the monaural PIT model for multi-channel speech separation. The signals from the selected pre-enhanced beams are processed through a single-channel separation model for further enhancement. The system in [12] consists of two neural networks, a dual-channel PIT network for initial mask and source direction of arrival (DOA) estimation, and a multi-channel enhancement network trained to separate the speaker of interest with specific spectral characteristics and arriving from a particular direction.

In general, the two most prominent ways for extending to multi-channel PIT are either converting the multi-channel inputs to single-channel features by means of beamformers so that they could fit the single-channel PIT model, or incorporating spatial features together with spectral features for the separation model training. As an end-to-end approach, the later one is more fundamental for the multi-channel study of speech separation under PIT framework.

3. LEARNING WITH SPATIAL FEATURES

Since room reverberation can substantially deteriorate the ILDs [13, 14], IPDs, ITDs and a few variations have been widely used as inputs for the multi-channel neural network. The recent studies in [10, 12] compute phase differences between microphone pairs and incorporate spatial features in the training scheme to discriminate one source from another through their location differentials. However, no investigation has been conducted on performance variability created by a change in relative location of multiple simultaneous speakers, particularly when two speakers or more are closely located.

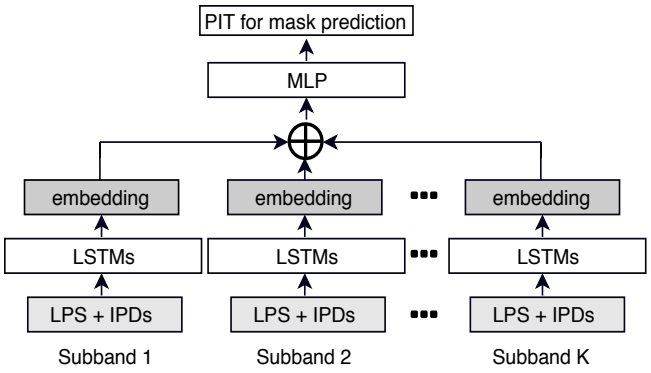

Fig. 1. Multi-band feature encoding for multi-channel PIT.

The binaural spatial feature based classification has been widely used in blind speech separation [14, 15]. Given a microphone pair $\langle p, q \rangle$, to avoid frequency dependence in the IPD we extract frequency normalized IPD $\frac{1}{2\pi f} arg\left[\frac{Y_p(f,t)}{Y_q(f,t)}\right]$ for each T-F bin, which has been utilized in [13] for avoiding frequency permutation issue after feature clustering.

If multiple microphone pairs are enrolled, each speaker has multiple location information by referring to different microphone pairs. In practice, it increases the feature discrimination and avoids failure in the case of location ambiguity that two speakers' directions are symmetric with respect to a certain microphone pair. We thus believe that neural networks could learn to select the most discriminative spatial features from a number of enrolled microphone pairs. Furthermore, spectral features, e.g. log power spectra (LPS) of a reference microphone used in this work, can be tightly integrated with spatial features to improve the system's robustness to the "spatial overlapping" problem when speakers are closely located, in which case the spatial features fail for source discrimination. Unfortunately, even equipped with spectral features, we found out that multi-channel PIT does not perform well in spatial overlapping scenarios as illustrated in Table 1 (schemes 1 vs. 2-5) when the two speakers' directions are less than 15° apart from each other. This situation is even worse when the number of enrolled microphone pairs increases. The monaural model's performance may serve as the upper-bound for the multi-channel models under this circumstance. It indicates from the observation that spatial features may play an overwhelming role in the model training as the source separation task is easier while relying on spatial difference of speaker sources than their spectral characteristics in most cases. Therefore, the model is over adapted to fit spatial features rather than pursuing a balance between the two, and thus fails at the spatial overlapping case. Increasing the size of training set in the category $0^\circ \sim 15^\circ$, i.e. with more spatial overlapped speakers in the training set, does not help to achieve better performance. We propose solutions in the following sections.

4. MULTI-BAND EMBEDDINGS

A variety of inter-channel spatial features including IPD [10], cosIPD & sinIPD [16] and generalized cross-correlation (GCC) [16, 17] have been utilized for the multi-channel model training. However, the main disadvantage of the estimated phase difference is the potential phase wrapping in high frequencies, particularly when the microphone spacing is not sufficiently small. The occurrence of phase wrapping is common when the microphone spacing exceeds $\lambda_{min}/2$, half of the minimum wavelength of the speech signal. In practice, wide spacing of microphones is required to enhance DOA resolution, reduce mutual coupling between microphones, or make the microphone placement physically realizable [18]. For two microphones of 7cm spacing as an example, phase wrapping occurs at around of 2.5k Hz. This implies that IPDs in high frequency bands, no matter in which form they are operated, may have ambiguities and thus are not effective for discriminating sources in terms of their spatial information.

PIT based approaches conduct the end-to-end learning which directly produces the separation masks for the entire frequency bands in a time frame. It could also be considered as an "embedding" based method in the following way. The conventional PIT learns a full-band embedding up to the last projection layer, and the last fully connected layer projects the full-band embedding to each speaker source in every individual frequency band. This motivates us to split the full-band input features (log power spectra + IPDs) into multiple frequency subbands. As illustrated in Figure 1, multiple recurrent neural network towers are jointly trained to generate individual subband embeddings from the corresponding subband input features. Therefore, those subbands with reliable spatial features could leverage them to boost the embedding learning, while high frequency subbands learn to attend more on their spectral features. The dimension of each subband embedding remains the same as the conventional full-band embedding's. Without any change on the projection layer, subband embeddings are summed up to form a new embedding for the mask prediction.

5. MODEL INTEGRATION

Revealed in Section 3, the conventional multi-channel training does not generalize well to the spatial overlapping case in which spatial cues are ineffective while spectral cues are not selectively attended. We propose a simple yet effective algorithm to address the spatial overlapping issue existing in the multi-channel feature encoding and model training. In a multi-task learning framework, the prediction of speakers' relative location is jointly trained on the shared embedding with the multi-channel speech separation model to infer if the included angle of two speakers on the horizontal plane is less than 15°, i.e. spatial overlapping. As shown in Figure 2, with the multi-band feature encoding, the multi-channel PIT has its

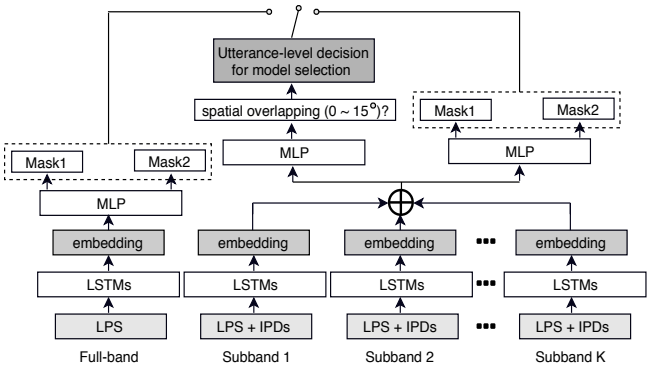

Fig. 2. The architecture of model integration.

own objective while its embedding with spatial features encoded benefits the task of spatial overlapping prediction and vice versa. We convert the frame-level spatial information inference to an utterance-level decision for model selection. Straightforwardly, the monaural PIT is employed if an utterance is identified with spatial overlapped speakers in the testing phase, otherwise the integrated system switches to operate on the multi-channel PIT.

6. EXPERIMENT AND EVALUATION

To create reverberant multi-channel speaker mixtures, we convolve the room impulse responses (RIRs) with the utterances in the WSJ0-2mix dataset [2], which contains single-channel anechoic two-speaker mixtures in its 30-hour training, 10-hour validation and 5-hour test set, respectively. The test speakers are unseen in the training phase. We consider a 6-microphone circular array of 7cm diameter with speakers and the microphone array randomly located in the room. The two speakers and microphone array are on the same plane and all of them are at least 0.3m away from the wall. We employ image method [19] to simulate RIRs randomly from 3000 different room configurations with the size (length$\times$width$\times$height) ranging from $3m \times 3m \times 2.5m$ to $8m \times 10m \times 6m$. The reverberation time RT_{60} is sampled in a range of 0.05s to 0.5s. We generate 30-hour, 10-hour and 5-hour 6-channel utterances for training, validation and testing, respectively. The RIRs used in validation and testing are unseen in the training phase.

The log power spectra and frequency normalized IPDs are computed based on 512-point short-time Fourier transform (STFT) of waveform signal with a 32ms window and 16ms shift. The microphone pairs utilized for computing IPDs for each evaluated configuration are listed in Table 1, where the microphone spacing is 3.5cm for mic pair 1-2, 3-4, 5-6, while it is 7cm for the rest. The baseline PIT networks contain three LSTM layers, each with 512 units, followed by a fully connected layer of 512 hidden units using rectified linear unit (ReLU) nonlinearity and a sigmoid output layer. Phase sensitive approximation [20] infers 257×2 dimensional real mask

Table 1. *Evaluation of different approaches in terms of SDR (dB) on closed condition (CC) and open condition (OC) sets.*

Method	$0^\circ \sim 15^\circ$		$15^\circ \sim 45^\circ$		$45^\circ \sim 90^\circ$		$90^\circ \sim 180^\circ$		Avg.	
	CC	OC	CC	OC	CC	OC	CC	OC	CC	OC
raw	2.1	2.2	2.1	2.1	2.1	2.1	2.1	2.1	2.1	2.1
LPS[1]	9.1	8.4	9.1	8.8	9.1	8.7	9.2	8.9	9.1	8.7
LPS + 1 IPD (mic pair 1-2)[2]	8.9	8.2	9.0	8.8	9.5	9.1	10.1	9.9	9.4	9.1
LPS + 2 IPDs (mic pair 1-2, 1-4)[3]	7.6	7.1	10.0	9.8	11.3	10.9	11.9	11.5	10.5	10.2
LPS + 3 IPDs (mic pair 1-4, 2-5, 3-6)[4]	7.2	6.7	10.2	10.0	11.7	11.3	12.0	11.5	10.7	10.3
LPS + 6 IPDs (mic pair 1-4, 2-5, 3-6, 1-2, 3-4, 5-6)[5]	6.1	5.6	9.6	9.4	11.3	11.0	11.9	11.6	10.2	9.9
LPS + 6 IPDs, two-band (6k Hz)[6]	6.8	6.3	10.0	9.9	11.8	11.4	12.6	12.1	10.8	10.4
LPS + 6 IPDs, two-band (4k Hz)[7]	6.8	6.5	10.5	10.3	12.3	12.0	13.0	12.7	11.1	10.9
LPS + 6 IPDs, two-band (2k Hz)[8]	6.8	6.4	10.8	10.7	12.7	12.3	13.5	13.1	**11.5**	**11.2**
LPS + 6 IPDs, comparable model size[9]	6.4	6.0	9.8	9.7	11.4	11.1	12.1	11.7	10.4	10.1
LPS + 6 IPDs, four-band (2k/4k/6k Hz)[10]	6.5	6.2	10.5	10.5	12.3	12.0	13.1	12.8	11.1	11.0
LPS, two-band (2k Hz)[11]	8.6	7.9	8.4	8.3	8.4	8.2	8.5	8.3	8.4	8.2
LPS + 1 IPD, two-band (2k Hz)[12]	7.5	7.0	9.6	9.4	11.4	11.0	12.4	12.1	10.6	10.3
LPS + 2 IPDs, two-band (2k Hz)[13]	6.7	6.1	10.2	10.0	12.0	11.6	12.9	12.6	11.0	10.6
LPS + 3 IPDs, two-band (2k Hz)[14]	6.9	6.3	10.5	10.4	12.4	12.1	13.0	12.6	11.2	10.9
LPS + 6 IPDs, two-band (2k Hz), multi-task[15]	7.0	6.6	11.0	10.9	12.7	12.4	13.5	13.1	**11.6**	**11.3**
LPS + 6 IPDs, two-band (2k Hz), model integ.[16]	**8.9**	**8.3**	11.0	10.7	12.6	11.9	13.3	12.6	**11.8**	**11.2**

in the output layer for two-speaker mixtures. Multi-band PIT networks are formed by K-tower baseline LSTMs, where K is the number of divided bands. Other parts of the networks remain the same as the baseline networks. All networks are trained with single-frame segment on 30-hour training utterances for 40 epochs using Adam algorithm.

In Table 1 we summarized the signal-to-distortion ratio (SDR) [21] from different schemes for two-talker mixture separation, respectively. The evaluation set is subdivided into four categories based on the two speakers' included angle in the horizontal plane. Since the two speakers' locations in a mixture utterance are uniformly sampled in a certain area of a room, the ratio of utterance samples falling into the four categories is approximately $1:2:2:2$. The same ratio applies to the training set as well. The limitation of the conventional multi-channel PIT employed in [10, 12] in spatial overlapping cases can be observed from the results in category $0^\circ \sim 15^\circ$, although more enrolled spatial features usually lead to better performance for spatially separated sources in other three categories (schemes 1 to 5 in Table 1). One exception is that the baseline method with 6 IPDs performs slightly worse than the one with 3 IPDs, likely indicating the demand for a larger model size as the feature size increases. This is proved in the evaluation of multi-band framework where the model with 6 IPDs achieves better results than others (Schemes 8 vs. 12-14). More importantly, splitting the full-band features at 2k Hz, with two bands 0 to 2k Hz and 2k to 8k Hz, leads to the best result, which is coincident with the phase wrapping frequency 2.5k Hz for this microphone array configuration (schemes 8 vs. 6-7). We also prove that the performance improvement on the multi-band architecture is not due to its larger model size by including a comparison with a baseline model of comparable model size (schemes 8 vs. 9). Meanwhile, splitting the full-band features into four bands at 2k Hz, 4k Hz and 6k Hz does not achieve greater results than the two-band architecture (schemes 8 vs. 10). Furthermore, since the multi-band feature encoding aims for resolving phase wrapping issue in the spatial feature training, monaural PIT with only LPS does not benefit from such approach as shown in scheme 11. We report the exploration of multi-task learning of two-band PIT for speech separation and spatial overlapping prediction. Results in scheme 15 show the merit of multi-task learning for both tasks. The equal-error-rate (EER) of frame-level spatial overlapping prediction learned on the higher band embedding (2k to 8k Hz) in this experiment is about 8%. Finally, with model integration (scheme 16), the spatial overlapping issue is resolved with results in category $0^\circ \sim 15^\circ$ significantly improved.

7. CONCLUSION

By revisiting multi-channel approaches under the PIT framework for speech separation, we reveal two underlying issues in the end-to-end multi-channel PIT: phase wrapping and spatial overlapping. A multi-band PIT for effective feature encoding is proposed to minimize the impact of phase wrapping in spatial features. Furthermore, an integrated PIT system leverages both single-channel and multi-channel models, leading to the significantly improved performance, particularly for the multi-talker mixtures of the spatial overlapped sources. We believe the way of feature encoding is encouraging to merit further exploration.

8. REFERENCES

[1] E. Colin Cherry, "Some experiments on the recognition of speech, with one and with two ears," *The Journal of the Acoustical Society of America*, vol. 25, no. 5, pp. 975–979, 1953.

[2] J. R. Hershey, Z. Chen, J. L. Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *the Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE*, 2016, pp. 31–35.

[3] Y. Isik, J.L. Roux, Z. Z. Chen, and et al., "Single-channel multi-speaker separation using deep clustering.," in *Interspeech*, 2016, pp. 545–549.

[4] Z. Chen, Y. Luo, and N. Mesgarani, "Deep attractor network for single-microphone speaker separation.," in *the Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE*, 2017, pp. 246–250.

[5] Y. Luo, Z. Chen, and N. Mesgarani, "Speaker-independent speech separation with deep attractor network," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 2018.

[6] D. Yu, M. Kolbak, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *the Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE*, 2017.

[7] M. Kolbk, D. Yu, Z.-H. Tan, and J. Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 25, no. 10, pp. 1901–1913, 2017.

[8] Y. Qian, C. Weng, X. Chang, S. Wang, and D. Yu, "Past review, current progress, and challenges ahead on the cocktail party problem," *Frontiers of Information Technology & Electronic Engineering*, vol. 19, no. 1, pp. 40–63, 2018.

[9] N. Roman, D.L. Wang, and G.J. Brown, "Speech segregation based on sound localization," *J. Acoust. Soc. Am.*, vol. 114, pp. 2236–2252, 2003.

[10] T. Yoshioka, H. Erdogan, Z. Chen, and F. Alleva, "Multi-microphone neural speech separation for far-field multi-talker speech recognition," in *the Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE*, 2018.

[11] Z. Chen, T. Yoshioka, X. Xiao, J. Li, M. L. Seltzer, and Y. Gong, "Efficient integration of fixed beamformers and speech separation networks for multi-channel far-field speech separation," in *the Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE*, 2018.

[12] Z. Wang and D-L Wang, "Integrating spectral and spatial features for multi-channel speaker separation," in *Interspeech*, 2018, pp. 2718–2722.

[13] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. Sig. Proc.*, vol. 52, pp. 1830–1847, 2004.

[14] H. Sawada, S. Araki, and S. Makino, "A two-stage frequency-domain blind source separation method for underdetermined convolutive mixtures," in *Proc. WASPAA*, 2007.

[15] A. Jourjine, S. Rickard, and O. Yilmaz, "Blind separation of disjoint orthogonal signals: Demixing n sources from 2 mixtures," in *the Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE*, 2000.

[16] Z.Q. Wang, J. Le Roux, and J. Hershey, "Multi-channel deep clustering: Discriminative spectral and spatial embeddings for speaker-independent speech separation," in *the Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE*, 2018.

[17] X. Xiao, S. Watanabe, H. Erdogan, L. Lu, J. Hershey, M. L. Seltzer, G. Chen, Y. Zhang, M. Mandel, and Dong Yu, "Deep beamforming networks for multi-channel speech recognition," in *the Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE*, 2016, pp. 5745–5749.

[18] T. Ballal and C. J. Bleakley, "Doa estimation of multiple sparse sources using three widely-spaced sensors," in *Proc. the 17th Europ. Signal Process. Conf*, 2009.

[19] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *JASA*, vol. 65, 1979.

[20] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *the Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE*, 2015, pp. 708–712.

[21] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 4, pp. 1462–1469, 2006.