ENHANCING END-TO-END MULTI-CHANNEL SPEECH SEPARATION VIA SPATIAL FEATURE LEARNING

Rongzhi Gu[1]   *Shi-Xiong Zhang*[2]   *Lianwu Chen*[3]   *Yong Xu*[2]   *Meng Yu*[2]
Dan Su[3]   *Yuexian Zou* [1,4*]   *Dong Yu*[2]

[1] ADSPLAB, School of ECE, Peking University, Shenzhen, China
[2]Tencent AI Lab, Bellevue, WA, USA
[3]Tencent AI Lab, Shenzhen, China
[4]Peng Cheng Laboratory, Shenzhen, China

ABSTRACT

Hand-crafted spatial features (e.g., inter-channel phase difference, IPD) play a fundamental role in recent deep learning based multi-channel speech separation (MCSS) methods. However, these manually designed spatial features are hard to incorporate into the end-to-end optimized MCSS framework. In this work, we propose an integrated architecture for learning spatial features directly from the multi-channel speech waveforms within an end-to-end speech separation framework. In this architecture, time-domain filters spanning signal channels are trained to perform adaptive spatial filtering. These filters are implemented by a 2d convolution (conv2d) layer and their parameters are optimized using a speech separation objective function in a purely data-driven fashion. Furthermore, inspired by the IPD formulation, we design a conv2d kernel to compute the inter-channel convolution differences (ICDs), which are expected to provide the spatial cues that help to distinguish the directional sources. Evaluation results on simulated multi-channel reverberant WSJ0 2-mix dataset demonstrate that our proposed ICD based MCSS model improves the overall signal-to-distortion ratio by 10.4% over the IPD based MCSS model.

Index Terms— multi-channel speech separation, spatial features, end-to-end, inter-channel convolution differences
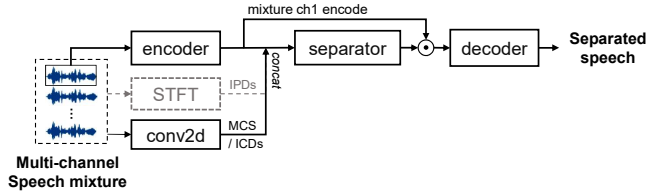
1. INTRODUCTION

Speech separation refers to recovering the voice of each speaker from overlapped speech mixture. It is also known as cocktail party problem [1], which has been studied in signal processing literature for decades. Leveraging the power of deep learning, many methods have been proposed for multi-channel speech separation (MCSS), including time-frequency (T-F) masking [2–5], integration of T-F masking and beamforming [6, 7], and end-to-end approaches [8]. T-F masking based methods formulate speech separation as a supervised learning task in frequency domain. The network learns to estimate a T-F mask for each speaker based on the magnitude spectrogram and interaural differences calculated from the complex spectrograms of observed multi-channel mixture signals, such as the phase difference between two microphone channels, which is known as the interaural phase difference (IPD).

However, one limitation for T-F masking based methods is the phase reconstruction problem. To avoid the complex phase estimation, time-domain speech separation has attracted increasing focus recently. A single-channel time-domain state-of-the-art approach, referred as SC-Conv-TasNet [9], replaces the short time Fourier transform (STFT)-inverse STFT with an encoder-decoder structure. Under the supervision from clean waveforms of speakers, SC-Conv-TasNet's encoder learns to construct an audio representation that optimized for speech separation. However, the performance of SC-Conv-TasNet is still limited under far-field scenario due to the smearing effects brought by reverberation. To tackle with this problem, in [8], we proposed a new MCSS solution, in which hand-crafted IPD features are used to provide spatial characteristic difference information between directional sources. With the aid of additional spatial cues, improved performances have been observed. However, the IPDs are computed in frequency domain with fixed complex filters (i.e., STFT) while the encoder output is learned in the data-driven manner. This causes a data mismatch, which indicates that IPDs may not be the optimal spatial features to incorporate into the end-to-end MCSS framework.

Bearing above discussions in mind, this work aims to design an end-to-end MCSS model, which are endowed with the capability to learn effective spatial cues using a speech separation objective function in a purely data-driven fashion. As illustrated in Figure 1, inspired by the success of SC-Conv-TasNet and [8], the main body of our proposed MCSS model adopts an encoder-decoder structure. In this design, the time-domain filters spanning all signal channels are trained to perform spatial filtering for multi-channel setting. These filters are implemented by a 2d convolution (conv2d) layer to extract the spatial features. Furthermore, inspired by the formulation of IPD, a novel conv2d kernel is designed to

Fig. 1. The diagram of our proposed end-to-end multi-channel speech separation system.

Fig. 2. (a) Visualization of encoder learned from far-field data. The top figure plots encoder's basis functions and the bottom is the corresponding FFT magnitudes. These basis functions are sorted by the frequency bin containing the peak response. (b) FFT magnitudes of STFT kernel functions.

compute the inter-channel convolution differences (ICDs). It is noted that ICDs are learned in data-driven manner and are expected to provide the spatial cues that help to distinguish the directional sources without bringing any data mismatch issue compared with the hand-crafted spatial features. In the end, an end-to-end MCSS model is trained with the scale invariant signal-to-distortion ratio (SI-SDR) loss function. Performance evaluation is conducted on a simulated spatialized WSJ0 2-mix dataset. Experimental results demonstrate that our proposed ICDs based MCSS model outperforms IPD based MCSS model by 10.4% in terms of SI-SDRi.

The rest of the paper is organized as follows. Section 2 introduces our proposed architecture in detail. Experimental procedure and result analysis is presented in Section 3. Section 4 concludes the paper.

2. PROPOSED ARCHITECTURE

2.1. Multi-channel speech separation

The baseline MCSS separation system [8] adopts an encoder-decoder structure, where the data-driven encoder and decoder respectively replaces the STFT and iSTFT operation in existing speech separation pipelines, as shown in Figure 1. Firstly, the encoder transforms each frame of first (reference) channel's mixture waveform $\mathbf{y}_1$ to the mixture encode in a real-valued feature space. Specifically, the learned encoder consists of a set of basis functions, as illustrated in Figure 2 (a). Most learned filters are tuned to lower frequencies, which shares the similar property with mel filter banks [10] and frequency distribution of human auditory system [11]. Secondly, IPDs computed by STFT and the mixture encode are concatenated along the feature dimension and fed into the separation module. The separation module learns to estimate a mask in encoder output domain for each speaker, which shares the similar concept with T-F masking based methods. Finally, the decoder reconstructs the separated speech waveform from the masked mixture encode for each speaker. To optimize the network end-to-end, scale-invariant signal-to-distortion ratio (SI-SDR) [12] is utilized as the training objective:

$$\text{SI-SDR} := 10\log_{10}\frac{\|x_{\text{target}}\|_2^2}{\|e_{\text{noise}}\|_2^2} \tag{1}$$

where $x_{\text{target}} := \langle \hat{x}, x \rangle\, x/\|x\|_2^2$, $e_{\text{noise}} := \hat{x} - x_{\text{target}}$, x and $\hat{x}$ are the reverberant clean and estimated source waveform,
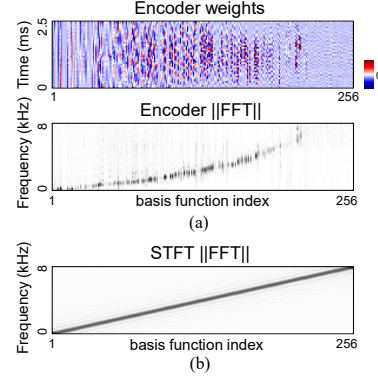
respectively. The zero-mean normalization is applied to x and $\hat{x}$ to guarantee the scale invariance.

However, the combination of IPD and encoder output may cause a data mismatch. Different from the encoder which is learned in a data-driven way, the IPD is calculated with complex fixed filters (i.e., STFT), the center frequencies of which are evenly distributed, as illustrated in Figure 1 (b). Also, as [9] points out, STFT is a generic transformation for signal analysis that may not be necessarily optimal for speech separation.

2.2. Spatial feature learning

To perform the spatial feature learning jointly with the rest of the network, we propose to learn spatial features directly from multi-channel waveforms with an integrated architecture. The main idea is to learn time-domain filters spanning all signal channels to perform adaptive spatial filtering [13–15]. These filters parameters are jointly optimized with the encoder using Eq. 1 in a purely data-driven fashion.

Denote these filters as $\mathbf{K} = \{\mathbf{k}^{(n)}\} \in \mathbb{R}^{C \times L \times N}$, where $\mathbf{k}^{(n)} = [..., k_c^{(n)}, ...] \in \mathbb{R}^{C \times L}$ is a set of filters spanning C signal channels with window size of L. Then, the multi-channel features are computed by summing up the convolution products between the c-th channel mixture signal $\mathbf{y}_c$ and filter $k_c^{(n)}$ along signal channel c, named as multi-channel convolution sum (MCS):

$$\text{MCS}^{(n)} = \sum_{c=1}^{C} \mathbf{y}_c \circledast k_c^{(n)} \tag{2}$$

where $\circledast$ denotes the convolution operation. The design principle lies in Eq. 2 is similar to that of delay-and-sum beam-former, where signal arriving at each microphone are summed up with certain time delays to emphasize sound from a particular direction. Each set of filters $\mathbf{k}^{(n)}$ is expected to steer at a different direction, therefore different spatial views of the

7320
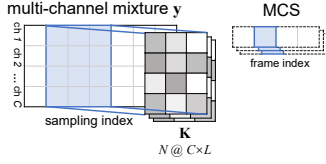
Fig. 3. Conceptual illustration of the conv2d and generation of different groups of MCS.
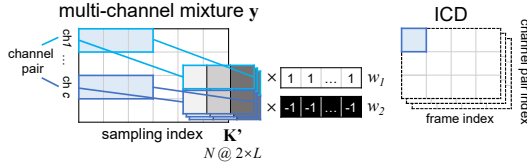
Fig. 4. Conceptual illustration of our designed conv2d kernel and generation of ICDs. To facinate training, w_1 is fixed as 1 while $w_2 \in \mathbb{R}^{1 \times L}$ is a learnable parameter initialized as -1.

multi-channel mixture signals can be obtained by MCS and therefore enhancing the separation accuracy.

To implement these learnable filters within the network, we employ a 2d convolution (conv2d) layer. The generation of MCS with conv2d is illustrated in Figure 3. The kernel size is $C \times L$ (height×width) and there are N convolution channels in total. The conv2d layer's stride along width axis represents the hop size and is fixed as $L/2$ in our experiments.

Furthermore, inspired by the formulation of interaural differences (e.g., IPDs), we design a special conv2d kernel to extract inter-channel convolution differences (ICDs). As we know, IPD is a well-established frequency domain feature widely used for spatial clustering algorithms and recent deep learning based MCSS methods. The rationale lies in that, the IPDs of T-F bins that dominated by the same source will naturally form a cluster within each frequency band, since their time delays are approximately the same. The standard IPD is computed by the phase difference between channels of complex spectrogram as $\text{IPD}_m = \angle \mathbf{Y}_{m_1} - \angle \mathbf{Y}_{m_2}$, where $\mathbf{Y}$ is the mutli-channel complex spectrogram computed by STFT of multi-channel waveform $\mathbf{y}$, m_1 and m_2 represent two microphones' indexes of the m-th microphone pair.

Following this concept, the n-th ICD between the m-th pair of signal channels can be computed by:

$$\text{ICD}_m^{(n)} = \sum_{c=1}^{2} w_c \cdot \left(\mathbf{y}_{m_c} \circledast k'^{(n)} \right) \tag{3}$$

where $k'^{(n)} \in \mathbb{R}^{1 \times L}$ is a filter shared among all signal channels to ensure identical mapping, $w_c \in \mathbb{R}^{1 \times L}$ is a window function designed to smooth the ICD and prevent potential spectrum leakage. When w_1 is fixed as full ones and w_2 as full negative ones, Eq. 3 calculates the exact inter-channel difference between the m-th microphone pair.

Figure 4 illustrates our designed conv2d kernel and generation of different pairs of ICDs. The conv2d kernel height
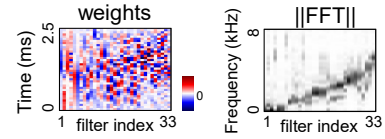
Fig. 5. Visualization of learned filters $\mathbf{K}'$. Each column represents a filter, and the right plot is its corresponding FFT magnitude response. These filters are sorted by the frequency bin containing the peak response.

is set as 2 to span a microphone pair. Note that different configurations of dilation d and stride s on the kernel height axis can extract ICDs from different pairs of signal channels, i.e., $m_1 = 1 + (m-1)s, m_2 = 2 + d + (m-1)s$. For example, for a 6-channel signal, setting dilation as 3 and stride as 1, we can obtain the three pairs of channels: (1, 4), (2, 5) and (3, 6).

To shed light on the property of learned filters $\mathbf{K}' = \{k'^{(n)}\}$, we visualize these filters in Figure 5. It can be observed that these learned filters show similar frequency tuning characteristics with the encoder (Figure 2 (a)). This suggests that the learned ICD may be more coincident with the encoder output and enables more efficient feature incorporation.

3. EXPERIMENTS AND RESULT ANALYSIS

3.1. Dataset

We simulated a spatialized reverberant dataset derived from Wall Street Journal 0 (WSJ0) 2-mix corpus, which are open and well-studied datasets used for speech separation [9, 16–18]. There are 20,000, 5,000 and 3,000 multi-channel, reverberant, two-speaker mixed speech in training, development and test set respectively. All the data is sampling at 16kHz. The performance evaluation is all done on test set, the speakers in which are all unseen during training. In this study, we take a 6-microphone circular array of 7cm diameter with speakers and the microphone array randomly located in the room. The two speakers and the microphone array are on the same plane and all of them are at least 0.3m away from the wall. The image method [19] is employed to simulate RIRs randomly from 3000 different room configurations with the size (length-width-height) ranging from 3m-3m-2.5m to 8m-10m-6m. The reverberation time T60 is sampled in a range of 0.05s to 0.5s. Samples with angle difference between two simultaneous speakers of 0-15°, 15-45°, 45-90° and 90-180° respectively account for 16%, 29%, 26% and 29%.

3.2. Network and Training details

All hyper-parameters are the same with the best setup of Conv-TasNet version 2 in [20], except L is set to 40 and encoder stride is 20. Batch normalization (BN) is used in all the experiments to speed up the separation process.

The microphone pairs for extracting IPDs and ICDs are (1, 4), (2, 5), (3, 6), (1, 2), (3, 4) and (5, 6) in all experiments. These pairs are selected because the distance of microphones

7321

Table 1. SDRi (dB) and SI-SDRi (dB) performances with different configurations of conv2d layer on far-field WSJ0 2-mix.

Setup	window w	# filters N	SI-SDRi (dB)					SDRi (dB)
			$<15°$	$15°$-$45°$	$45°$-$90°$	$>90°$	Ave.	
Single-channel Conv-TasNet	-	-	8.5	9.0	9.1	9.3	9.1	9.4
+MCS (conv2d (6×40))	-	256	5.7	10.3	11.9	12.9	10.8	11.2
+ICD (conv2d (2×40))	fix -1	256	5.5	10.9	12.3	12.9	11.0	11.4
+ICD (conv2d (2×40))	init. -1	256	6.2	11.2	12.6	13.2	11.4	11.8
+ICD (conv2d (2×40))	init. randomly	33	8.2	8.1	9.0	9.1	8.9	9.2
+ICD (conv2d (2×40))	fix -1	33	6.9	11.1	12.3	12.9	11.3	11.7
+ICD (conv2d (2×40))	init. -1	33	6.7	11.7	13.1	13.9	**11.9**	**12.3**

in between each pair is either the furthest or nearest. In this case, there are two setups of dilation d and stride s for the conv2d layer, respectively $d = 1, s = 2$ and $d = 3, s = 1$. The first channel of mixture waveform is set as the reference channel as the encoder input. To match the encoder output's time steps, both IPDs and ICDs are extracted with 2.5ms (40-point) window length L and 1.25ms (20-point) hop size with 64 FFT points. SI-SDR (Eq. 1) is utilized as training objective. The training uses chunks with 4.0 seconds duration. The batch size is set to 32. Permutation invariant training [17] is adopted to tackle with label permutation problem.

3.3. Result Analysis

Following the common speech separation metrics [12,21], we adopt average SI-SDR and SDR improvement over mixture as the evaluation metrics. We also report the performances under different ranges of angle difference between speakers to give a more comprehensive assessment for the model.

Different configurations for conv2d layer. We explore different conv2d configurations for computing the ICD, including different numbers of filters and initialization methods of window function w (w_2 in section 2.2). The number of filters are chosen to be 256 and 33, where 256 matches the basis function number of encoder, 33 is the number of bins for 64-point FFT size, which is the closest exponential of 2 for 40-point frame length. The results are listed in Table 1. SC-Conv-TasNet is served as the baseline system, achieving 9.1dB of SI-SDRi on the far-field dataset. By learning spatial filters, the MCS based model outperforms the baseline by 1.7dB of SI-SDRi. For ICD setups, we found that the performances with 33 filters are relatively superior to those with 256 filters. One possible reason is that, according to sampling theorem, the highest frequency resolution can achieve with sampling rate of 16kHz and frame length of 40 is limited.

Furthermore, the value of w contributes significantly to the separation performance (9.2dB v.s. 12.3dB for model with 33 filters). If w is randomly initialized, or in other words, there is no explicit subtraction operation between signal channels, the model will not be able to automatically learn useful spatial cues. If w is initialized and fixed as -1 (fix -1), this indicates that the exact convolution difference operation between signal channels is computed as the ICD. Furthermore, relaxing w to be learnable (init. -1) produces a much better

Table 2. SDRi (dB) and SI-SDRi (dB) performances of IPD, ICD-based separation systems on far-field WSJ0 2-mix.

Features	SI-SDRi (dB)		
	$<15°$	$>15°$	Ave.
cosIPD, sinIPD	7.7	12.2	11.5
cosIPD, sinIPD (trainable kernel) [8]	7.9	12.3	11.6
ICD	6.7	12.9	11.9
ICD, cosIPD, sinIPD	8.1	13.2	**12.4**

result, which demonstrates the validity of ICD's formulation.
IPD versus ICD. We examine the performance of IPD versus proposed ICD for MCSS and report the results in Table 2. In addition, the performance of IPD with trainable kernel based MCSS model [8] is listed for comparison. Specifically, in [8], the standard STFT operation is reformulated as a function of time domain convolution with a trainable kernel, which is optimized for the speech separation task. Combining the cosIPD and sinIPD we can obtain SI-SDRi of 11.5dB, which has 2.4dB gain over the single-channel baseline. It suggests IPDs can provide beneficial spatial information of sources. With the trainable kernel, the performance improves slightly. The proposed ICD based separation model obtains 0.4dB improvement over cosIPD+sinIPD based, benefiting from the data-driven learning fashion. Note that the performance under 15° for ICD based model is worse than that of IPD based. One possible reason is that the portion of data under 15° is relatively few hence causing difficulty in learning effective ICDs. The incorporation of ICDs and IPDs achieves further 0.5dB improvement. In this case, we found that almost all filters are tuned to relatively low frequency. This indicates that the ICDs learn complementary spatial information to compensate the IPD ambiguity in low frequencies.

4. CONCLUSION

This work proposes an end-to-end multi-channel speech separation model, which is able to learn effective spatial cues directly from the multi-channel speech waveforms in a purely data-driven fashion. Experimental results demonstrated the MCSS model based on learned ICDs outperforms that based on well established IPDs.

5. REFERENCES

[1] Colin Cherry and J. A. Bowles, "Contribution to a study of the cocktail party problem," *Journal of the Acoustical Society of America*, vol. 32, no. 7, pp. 884–884, 1960.

[2] Zhong-Qiu Wang, Jonathan Le Roux, and John R Hershey, "Multi-channel deep clustering: Discriminative spectral and spatial embeddings for speaker-independent speech separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 1–5.

[3] Lianwu Chen, Meng Yu, Dan Su, and Dong Yu, "Multi-band pit and model integration for improved multichannel speech separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 705–709.

[4] Zhuo Chen, Xiong Xiao, Takuya Yoshioka, Hakan Erdogan, Jinyu Li, and Yifan Gong, "Multi-channel overlapped speech recognition with location guided speech extraction network," in *IEEE Spoken Language Technology Workshop (SLT)*, 2018, pp. 558–565.

[5] Zhong-Qiu Wang and DeLiang Wang, "On spatial features for supervised speech separation and its application to beamforming and robust asr," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5709–5713.

[6] Zhuo Chen, Takuya Yoshioka, Xiong Xiao, Linyu Li, Michael L Seltzer, and Yifan Gong, "Efficient integration of fixed beamformers and speech separation networks for multi-channel far-field speech separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5384–5388.

[7] Lukas Drude and Reinhold Haeb-Umbach, "Tight integration of spatial and spectral features for bss with deep clustering embeddings.," in *Interspeech*, 2017, pp. 2650–2654.

[8] Rongzhi Gu, Jian Wu, Shi-Xiong Zhang, Lianwu Chen, Yong Xu, Meng Yu, Dan Su, Yuexian Zou, and Dong Yu, "End-to-end multi-channel speech separation," *arXiv preprint arXiv:1905.06286*, 2019.

[9] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal timefrequency magnitude masking for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 27, no. 8, pp. 1256–1266, Aug 2019.

[10] Satoshi Imai, "Cepstral analysis synthesis on the mel frequency scale," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1983, vol. 8, pp. 93–96.

[11] Colin Humphries, Einat Liebenthal, and Jeffrey R Binder, "Tonotopic organization of human auditory cortex," *Neuroimage*, vol. 50, no. 3, pp. 1202–1211, 2010.

[12] Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R Hershey, "Sdr–half-baked or well done?," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 626–630.

[13] Yedid Hoshen, Ron J Weiss, and Kevin W Wilson, "Speech acoustic modeling from raw multichannel waveforms," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 4624–4628.

[14] Bo Li, Tara N Sainath, Ron J Weiss, Kevin W Wilson, and Michiel Bacchiani, "Neural network adaptive beamforming for robust multichannel speech recognition," *Interspeech 2016*, pp. 1976–1980, 2016.

[15] Tara N Sainath, Ron J Weiss, Kevin W Wilson, Arun Narayanan, Michiel Bacchiani, Bo Li, Ehsan Variani, Izhak Shafran, Andrew Senior, Kean Chin, et al., "Raw multichannel processing using deep neural networks," in *New Era for Robust Speech Recognition*, pp. 105–133. Springer, 2017.

[16] John R Hershey, Zhuo Chen, Jonathan Le Roux, and Shinji Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 31–35.

[17] Dong Yu, Morten Kolbæk, Zheng-Hua Tan, and Jesper Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 241–245.

[18] Zhong-Qiu Wang and DeLiang Wang, "Combining spectral and spatial features for deep learning based blind speaker separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 27, no. 2, pp. 457–468, 2019.

[19] Jont B Allen and David A Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[20] Yi Luo and NimaR Mesgaran1, "Tasnet: Surpassing ideal time-frequency masking for speech separation," *arXiv preprint arXiv:1809.07454*, 2018.

[21] Emmanuel Vincent, Rémi Gribonval, and Cédric Févotte, "Performance measurement in blind audio source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 14, no. 4, pp. 1462–1469, 2006.