Interspeech 2018
2-6 September 2018, Hyderabad

A Shifted Delta Coefficient Objective for Monaural Speech Separation using Multi-task Learning

Chenglin Xu[1,2], *Wei Rao*[2], *Eng Siong Chng*[1,2], *Haizhou Li*[2,3]

[1] School of Computer Science and Engineering, Nanyang Technological University, Singapore
[2] Temasek Laboratories@NTU, Nanyang Technological University, Singapore
[3] Department of Electrical and Computer Engineering, National University of Singapore, Singapore
{xuchenglin,raowei,aseschng}@ntu.edu.sg, haizhou.li@nus.edu.sg

Abstract

This paper addresses the problem of monaural speech separation for simultaneous speakers. Recent studies such as uPIT, cuPIT-Grid LSTM and their variants have advanced the state-of-the-art separation models. Delta and acceleration coefficients are typically used in the objective function to capture short time dynamics. We consider that such coefficients don't benefit from the temporal information over a long range such as phoneme and syllable. In this paper, we propose a shifted delta coefficient (SDC) objective to explore the temporal information over a long range of the spectral dynamics. The SDC ensures the temporal continuity of output frames within the same speaker. In addition, we propose a novel multi-task learning framework, that we call SDC-MTL, by extending the SDC objective with a subtask of predicting the time-frequency labels ({silence, single, overlapped}) of the mixture. The experimental results show 11.7% and 3.9% relative improvements on WSJ0-2mix dataset under open conditions over the uPIT and cuPIT-Grid LSTM baselines. A further analysis shows 17.8% and 6.2% relative improvements with speakers of same gender.

Index Terms: Shifted Delta Coefficient, Time-Frequency labelling, Single Channel Speech Separation, Multi-tasking Learning.

1. Introduction

In human speech communication, a listener can easily capture the desired speech channel, even from complex background noise and competing speech. Such a cocktail party problem [1] is not trivial to be solved by a machine. To naturally interact with machines, an automatic solution is required in many real-world applications, such as remote meeting devices, smart speakers and robots.

The cocktail party problem has been studied for decades, from initial works using computational auditory scene analysis (CASA) [2, 3] based on heuristics, e.g., pitch continuity, non-negative matrix factorization (NMF) [4, 5, 6], to probabilistic models such as a factorial GMM-HMM [7]. However, these methods not only rely on accurate trackers (i.e., pitch tracker) but also intensive computation. Furthermore, these methods often either produce poor performance or only work under the conditions with prior knowledge of the speakers.

Recently, the performance of the monaural speech separation has been significantly improved by several deep learning techniques, such as deep clustering (DC) [8, 9], deep attractor network (DANet) [10], permutation invariant training (PIT) [11], utterance level PIT (uPIT) [12] and constrained uPIT (cuPIT) [13]. The DC and DANet techniques project the spectrogram of the mixture to a high dimensional embedding space, where the time-frequency bins belonging to each source are grouped together.

One shortcoming of the DC method is that its objective function is defined in the embedding space, which may not be optimal for the speech separation task. The DANet method addresses this problem by creating attractor points in high dimensional embedding space and estimating mask within the network to separate signals directly. Unfortunately, this approach increases the complexity as it involves the attractor in the runtime process. The PIT, uPIT and cuPIT techniques are a series of approaches to speech separation with end-to-end training by minimizing the cost over all permutations in order to solve the label ambiguity problem. To overcome the frame discontinuity problem during inference in PIT, the uPIT is proposed by forcing the separated frames belonging to the same speaker to be aligned to the same output stream using BLSTM [14] with an utterance level training criterion. The cuPIT [13] method further improves the performance, especially in same gender condition, by adding weighted delta and acceleration of output frames, that we call delta-acceleration coefficients, in the cost function. Such temporal information has improved the end-to-end training. However, the delta-acceleration coefficients are limited to a short-time window of a few speech frames.

Previous studies show that human speech comprehension depends on the integrity of both the spectral content and temporal envelop of the speech signal [15, 16]. Human auditory neurons are tuned to detect local spectro-temporal patterns of speech [17, 18]. Inspired by these findings, we would like to address the discontinuity problem by proposing a shifted delta coefficient (SDC) objective, which uses long contextual temporal dynamics to supervise the mask estimation process during the end-to-end training. The shifted delta coefficients capture the time dynamic of the speech behavior in a long range of 10 frames ($16ms$ shift for one frame). The spectral transitions between phonemes and even syllables are now included in the SDC analysis window. With the dynamics, the spectral frames of same speaker will be aligned to the same side without having to explicitly find and model the spectral structure of the speech signal. In addition, we propose a novel multi-task learning framework that extends the SDC objective in the main task of speech separation with a subtask of predicting the time-frequency labels ({silence, single, overlapped}) of the mixed speech signal. We name the multi-task learning system as SDC-MTL method. Since the masks of the overlapped parts are hard to predict, the proposed multi-task learning architecture improves the mask estimation by explicitly telling which time-frequency bins are overlapped during the training.

Section 2 describes the monaural speech separation problem by using masks to filter the mixture. The details of the

10.21437/Interspeech.2018-1150

proposed novel multi-task learning framework are discussed in Section 3. Section 4 reports the experiments and the results. Finally, we conclude in Section 5.

2. Monaural Speech Separation with Masks

The monaural speech separation aims to separate a linearly mixed single channel microphone signal $y(n)$ into individual source signals $x_s(n), s \in [1, S]$.

$$y[n] = \sum_{s=1}^{S} x_s[n] \tag{1}$$

Since only the mixed signal $y(n)$ is observed, the goal is to estimate $\hat{x}_s[n]$ that is close to $x_s[n]$. This problem is recently formulated as a supervised learning task, which estimates a filter (i.e., mask) for each speaker with the supervised information of ideal binary mask or ideal ratio mask [19, 20, 21, 22]. With the introduction of deep learning techniques, the separation performance has been dramatically improved with the magnitude spectrum approximation loss [11, 13, 23, 24]. To estimate the mask, the mixed speech signal is transformed into spectrogram representation as $Y(t, f)$ for each time-frequency bin (t, f). Then the mask estimation is conducted on the spectrogram of the mixture directly.

By using the predicted masks, the magnitude $|\hat{X}_s(t, f)|$ of individual source is obtained by

$$|\hat{X}_s(t, f)| = M_s(t, f) \odot |Y(t, f)| \tag{2}$$

where $\odot$ indicates element-wise multiply. The estimated magnitude $|\hat{X}_s(t, f)|$ of each speaker and the phase of mixed speech $\angle Y(t, f)$ are used to reconstruct the time domain waveform $\hat{x}_s[n]$ by an inverse discrete Fourier transform and an overlap and add operation. We note that phase estimation remains a challenge in speech separation or speech enhancement. In this paper, we use the phase of mixed speech directly when reconstructing the output signals.

3. Multi-Task Learning with SDC

In this paper, we propose to estimate magnitude spectrum approximation masks for each individual source using a shifted delta coefficient objective along with permutation invariant training to explore the temporal information over a long context. Then we incorporate the proposed objective into a multi-task learning framework with the subtask of predicting the labels for each time-frequency bin, as shown in Figure 1. In this framework, each time-frequency bin is tagged with one of the labels in {silence, single, overlapped}, which contributes to the process of mask learning.

3.1. Shifted Delta Coefficient Objective

Dynamic features [25, 26] and the shifted delta cepstral features [27, 28], which spans the delta cepstra computation across multiple frames of speech, have been proven effective in speech and language recognition. We propose the use of shifted delta coefficient objective in monaural speech separation to capture the long range contextual information. Although this objective is proposed for speech separation, it also can be used in neural network based speech enhancement.

The shifted delta coefficients explore the dynamic information of speech spectrogram. Since it spans across multiple

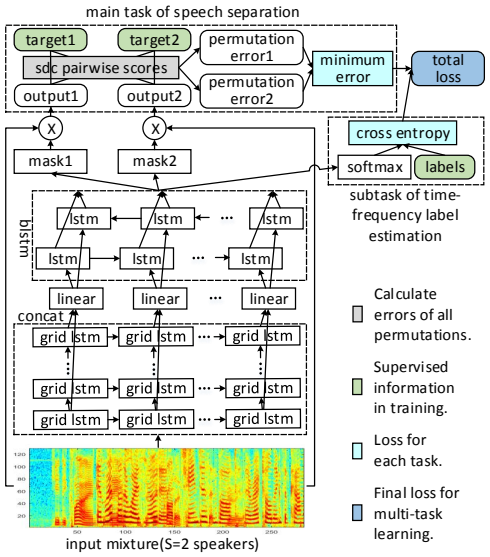

Figure 1: *The proposed multi-task learning framework for monaural speech separation during training. In run-time testing stage, the upper dotted box and the subtask in the upper right dotted box are not necessary. And the system takes input mixture in and outputs the separations of output1 and output2.*

frames, the spectral transitions between phonemes and even syllables are included into its analysis window. A cost function with the shifted delta coefficients ensures that the spectral continuity of speaker is explicitly taken into consideration in the speech separation task.

The computation of shifted delta coefficient is based on the delta coefficient ($f_d(t)$) that can be calculated as follows,

$$f_d(t) = \frac{\sum_{l=1}^{L} l \times (v(t+l) - v(t-l))}{\sum_{l=1}^{L} 2l^2} \tag{3}$$

where L is the order and is set as 2 in this study. $v(t)$ is the spectral feature vector from the frame t of speech. The shifted delta coefficient $f_{sdc}(t)$ extends the delta coefficient by concatenating K (i.e., 4) blocks of delta coefficient with a shift of P (i.e., 2) in this work. For kth coefficient,

$$\begin{aligned} f^k_{sdc}(t) &= f_d(t + (k-1)P) \\ &= \frac{\sum_{l=1}^{L} l \times (v(t + (k-1)P + l) - v(t + (k-1)P - l))}{\sum_{l=1}^{L} 2l^2} \end{aligned} \tag{4}$$

where $k \in \{1, K\}$. In this way, the shifted delta coefficient vector expands multiple frames (i.e., 10 frames) and contains

$$f_{sdc}(t) = [f_d(t), f_d(t+P), \ldots, f_d(t + (K-1)P)] \tag{5}$$

Following previous works [12, 13], the permutation invariant training is implemented to estimate the mask for each speaker with the supervision of the ideal phase sensitive mask (IPSM) [29], which considers the phase difference between the mixture and individual source. When the magnitude spectrum approximation loss is applied, the training criterion will be the cross mean square error of the shifted delta coefficients between the estimated magnitude and true magnitude with phase difference.

$$\begin{aligned} J_{sdc,\phi_p(s)} = \frac{1}{T} \sum_{s=1}^{S} (||f_{sdc}(\hat{M}_s \odot |Y|) - \\ f_{sdc}(|X_{\phi_p(s)}| \odot cos(\theta_y - \theta_{\phi_p(s)}))||^2_F) \end{aligned} \tag{6}$$

where $\phi_p(s), p \in [1, P]$ is an assignment of target source (s) to an output, and $P = S!$ is the number of all permutations. $||\cdot||_F$ is the Frobenius norm.

With the permutation invariant training, the optimal assignment is done by choosing the minimal cost among all permutations (P). For instance, the costs of $2! = 2$ permutations (error1, error2 when p=1 and 2) in the case of 2 speakers are considered in Figure 1.

$$\hat{p} = \arg\min_{p \in P} J_{sdc,\phi_p(s)} \tag{7}$$

And the cost of the shifted delta coefficient objective used to optimize the network is obtained with the optimal assignment.

3.2. Multi-task Learning Framework

Due to the time-frequency sparseness characteristic of speech signal, the time-frequency bins of the mixed speech can be classified into one of the three categories {silence, single, overlapped}. A speech separation task is to separate the overlapped time-frequency bins into individual speakers. The ideal masks for the overlapped parts are always less than 1. For the time-frequency bins tagged with 'single', the ideal masks are 1 for the speaker that they belong to, otherwise, they are 0. The time-frequency labels are information directly related to the masks.

In this paper, we form the time-frequency label estimation as a subtask in monaural speech separation, as shown in Figure 1. To estimate the time-frequency label, the output layer with 3 hidden nodes uses a softmax function to predict the label for each time-frequency bin. The cross entropy loss is calculated over frames.

$$J_{ce} = -\frac{1}{T}\sum_{t=1}^{T}(g_t \times \log \hat{g}_t + (1 - g_t) \times \log(1 - \hat{g}_t)) \tag{8}$$

where g_t is the true probability vector over all frequency bins at frame t. And $\hat{g}_t$ is the predicted probability vector over all frequency bins at frame t.

Then the multi-task learning loss is obtained by weight sum of the shifted delta coefficient objective loss and the cross entropy loss.

$$J_{mtl} = (1 - \lambda) \times J_{sdc,\phi_p(s)} + \lambda \times J_{ce} \tag{9}$$

where the λ is the weight to tune the importance of the two loss.

4. Experiments and Discussion

4.1. Experimental Setup

Same as [13], the proposed methods are evaluated on the WSJ0-2mix dataset [1] [8], which was mixed by randomly choosing utterances of two speakers from the WSJ0 corpus [30]. In this paper, the WSJ0-2mix (two-speaker mixed) dataset was divided into three sets: training set ($20,000$ utterances $\approx 30h$), development set ($5,000$ utterances $\approx 8h$), and test set ($3,000$ utterances $\approx 5h$). Specifically, the utterances from 50 male and 51 female speakers in the WSJ0 training set (si_tr_s) were randomly selected to generate the training and development set in WSJ0-2mix at various signal-to-noise (SNR) ratios uniformly chosen between 0dB and 5dB. Similarly, the test set was created by randomly mixing the utterances from 10 male and 8 female speakers in the WSJ0 development set (si_dt_05) and evaluation set (si_et_05). Since the speakers in the development set of WSJ0-2mix dataset were the same as those in the training set, the development set was used as closed condition (CC) to tune parameters. Moreover, the test set was considered as open condition (OC) evaluation, because the speakers in the test set were different from those in the training and development sets.

[1] Available at: http://www.merl.com/demos/deep-clustering

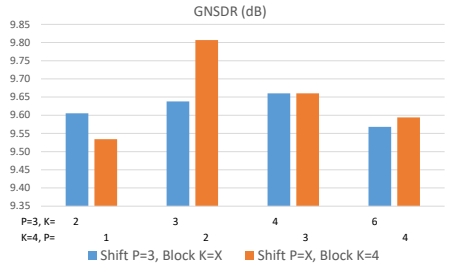

Figure 2: *The GNSDR (dB) with different shift P and block K of the SDC objective in the multi-task learning system on the test set.*

The data was generated at a sampling rate of 8kHz. And a 128-point STFT with the normalized square root of the 32ms length hamming window and 16ms window shift was used to transform the speech signal into frequency domain. The input 129-dim spectral magnitude features were obtained. Since the masks were estimated with magnitude spectrum approximation approach, the magnitudes of two targets were obtained in the same way as the supervision. To obtain the true time-frequency labels, we firstly labeled each time-frequency bin of every target speaker with voiced and unvoiced tags. If the time-frequency bin of both speakers were labeled as voiced, the tag for this time-frequency bin of the input mixture was 'overlapped'. If only one was labeled as voiced, the tag was 'single'. Otherwise, the tag of 'silence' was given.

In order to fairly compare with previous work [12, 13], 3 BLSTM layers with 896 units in each layer were kept same and deployed in our proposed architecture. The units, frequency input dimension and shift in the Grid LSTM cell were set to 64, 29 and 10. And the outputs of the Grid LSTM [31, 32, 33] were reduced by a linear layer from 1408 to 896. The BLSTM and Grid LSTM layers used a random dropout with a dropout rate of 0.5 [2]. The ReLU activation function was used in the mask estimation layer. Since there were 129 frequency bins and each frequency bin had 3 possible tags, the output layer of the subtask had $387 (= 3 * 129)$ nodes. The learning rate was initialized as 0.0005 and scaled down by 0.7 when the training loss increased on the development set. 16 randomly selected utterances were used in each minibatch. The number of minimum epoch was set to 30 and the early stopping criterion was that the relative loss improvement was lower than 0.01. The model was optimized with Adam algorithm [34] and implemented using Tensorflow [3]. We evaluate the performance using global normalized signal-to-distortion ratio (GNSDR, same as "SDR improvement" in [8, 9, 11, 12, 13]) using the toolbox in [35].

4.2. Experimental Results

4.2.1. Effect of Shifted Delta Coefficient Objective

Figure 2 shows the results with different shift P and block K in the SDC objective. To tune K, we firstly fix P to 3 and observe that the best performance is obtained when K grows to 4. Then we fix K to 4 to tune the shift P. With this parameter tuning

[2] The dropout was not applied across time steps, although it was known to be effective and used in [9].

[3] https://www.tensorflow.org/

Table 1: *A comparison of GNSDR (dB) over different objective functions calculated on Magnitude, Delta, Acceleration and SDC on WSJ0-2mix test sets using 3 BLSTM layers in uPIT-BLSTM systems.*

Objective	GNSDR
Magnitude	9.5
Delta	9.7
Acceleration	9.5
SDC	**9.8**

scheme, the best GNSDR (dB) is achieved when P is 2 and K is 4. Therefore, the SDC vector at frame t is $[f_d(t), f_d(t+2), f_d(t+4), f_d(t+6)]$. Since the best order L is tuned as 2 in the delta ($f_d(\cdot)$) computation, the SDC vector expands the long contextual temporal information to maximum of 10 frames.

To evaluate the effectiveness of long contextual temporal information in SDC objective, we compare SDC with Magnitude, Delta, and Acceleration based mean square error (MSE) objectives with permutation invariant training. The Delta and SDC include temporal information over the Magnitude directly and SDC captures the information with a even longer context window than Delta. Table 1 shows that the performance of the proposed SDC objective outperforms others. By comparing with magnitude based MSE objective [12], we confirm that dynamic temporal information, represented by Delta and SDC, is helpful. The long range contextual information (SDC) ensures the speech continuity within each speaker.

4.2.2. *Effect of Multi-task Learning*

By using the time-frequency labels to improve the performance of speech separation, we propose a novel multi-task learning framework to combine the proposed SDC objective and cross entropy of the subtask together. Figure 3 shows the results with different weight λ, which balances the importance between the proposed SDC objective and the cross entropy. We observe that the best performance is achieved when the weight is set to 0.2.

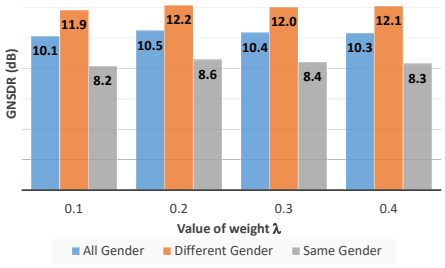

Figure 3: *The GNSDR (dB) of tuning the weight λ in the multi-task learning system using 1 Grid LSTM layer and 3 BLSTM layers evaluated under three conditions on the test set. All gender: all data are used to calculate GNSDR. Different gender: GNSDR is computed on the data that the gender of the input mixture is different. Similarly, same gender: GNSDR is computed on the data that the gender of the input mixture is same.*

From Table 2, we observe that the proposed multi-task learning has improved the performance by 3.1% for all test data, 2.6% for different gender and 5.3% for same gender, over the case where we only use SDC objective as a single task. By adding a Grid LSTM layer at the front of 3 BLSTM layers, the proposed multi-task learning system further improves the separation performance, especially in the same gender condition. This suggests that the proposed time-frequency labels in a multi-task learning framework are effective by telling where the speech is overlapped. We observe a GNSDR relative improvement over [12] by 10.5%, 6.1% and 17.8% for all test data, data with different and same gender, respectively.

Table 2: *GNSDR (dB) in a comparative study of with or without multi-task learning and Grid LSTM on WSJ0-2mix test set. The results of different and same gender are also included for further analysis. The weight λ is set to 0.2 for multi-task learning systems. cMSE means adding weighted delta and acceleration as constraints into MSE objective [13].*

Objective	Method	GNSDR		
		All	Diff.	Same
MSE [12]	uPIT-BLSTM	9.5	11.5	7.3
cMSE [13]	uPIT-BLSTM	9.8	11.7	7.7
cMSE [13]	uPIT-Grid LSTM	10.1	11.8	8.1
SDC	uPIT-BLSTM	9.8	11.7	7.6
SDC-MTL	uPIT-BLSTM	10.1	12.0	8.0
SDC-MTL	uPIT-Grid LSTM	**10.5**	**12.2**	**8.6**

4.2.3. *Comparisons with Other Methods*

Table 3 compares the proposed approach with other state-of-the-art methods on the WSJ0-2mix database. Our proposed SDC-MTL-Grid LSTM method outperforms other methods, such as DC, DANet, uPIT-BLSTM. Compared with our previous proposed cuPIT-Grid LSTM, the proposed SDC-MTL-Grid LSTM uses same network configuration and the difference comes from the SDC objective and multi-task learning. Although the cuPIT-Grid LSTM method uses additional weighted delta and acceleration as constraints in the objective, the proposed SDC-MTL-Grid LSTM outperforms the cuPIT-Grid LSTM by using a long contextual temporal objective to model the spectral dynamics. The time-frequency labels in the multi-task learning also contribute to the improvement by the supervised information of where the overlapped speech is. Examples are available at *https://sites.google.com/site/xuchenglin28/demos/is2018*.

Table 3: *GNSDR (dB) in a comparative study of different separation methods on the WSJ0-2mix dataset with optimal frame level assignment or default assignment on closed (CC) and open (OC) conditions.*

Method	Opt Assign		Def Assign	
	CC	OC	CC	OC
DC [8]	-	-	5.9	5.8
DC+ [9]	-	-	-	9.4
DANet [10]	-	-	-	9.6
PIT-DNN [11]	7.3	7.2	5.7	5.2
PIT-CNN [11]	8.4	8.6	7.7	7.8
uPIT-BLSTM [12]	10.9	10.8	9.4	9.4
cuPIT-Grid LSTM [13]	11.2	11.2	10.2	10.1
SDC-MTL-Grid LSTM	**11.4**	**11.4**	**10.6**	**10.5**
IRM	12.4	12.7	12.4	12.7
IPSM	14.9	15.1	14.9	15.1

5. Conclusion

In this paper, we propose a shifted delta coefficient objective for speech separation and incorporate it into a multi-tasking learning framework with a subtask of predicting time-frequency labels (silence, single and overlapped). Experimental results show that our proposed shifted delta coefficient objective is comparable to our previous proposed constrained cost function with weighted delta and acceleration. By adding the time-frequency labels into the multi-task leaning framework, the proposed SDC-MTL-Grid LSTM method achieves better performance than previous state-of-the-art methods. In addition, the proposed method is also effective on the same gender mixed speech separation task.

6. References

[1] A. W. Bronkhorst, "The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions," *Acta Acustica united with Acustica*, vol. 86, no. 1, pp. 117–128, 2000.

[2] D. P. W. Ellis, "Prediction-driven computational auditory scene analysis," Ph.D. dissertation, Massachusetts Institute of Technology, 1996.

[3] D. Wang and G. J. Brown, *Computational auditory scene analysis: Principles, algorithms, and applications*. Wiley-IEEE press, 2006.

[4] P. O. Hoyer, "Non-negative matrix factorization with sparseness constraints," *Journal of Machine Learning Research*, vol. 5, no. Nov, pp. 1457–1469, 2004.

[5] M. N. Schmidt and R. K. Olsson, "Single-channel speech separation using sparse non-negative matrix factorization," in *Proceedings of INTERSPEECH*, 2006.

[6] P. Smaragdis, "Convolutive speech bases and their application to supervised speech separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 1–12, 2007.

[7] T. Virtanen, "Speech recognition using factorial hidden markov models for separation in the feature space," in *Ninth International Conference on Spoken Language Processing*, 2006.

[8] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proceedings of ICASSP*. IEEE, 2016, pp. 31–35.

[9] Y. Isik, J. L. Roux, Z. Chen, S. Watanabe, and J. R. Hershey, "Single-channel multi-speaker separation using deep clustering," *arXiv preprint arXiv:1607.02173*, 2016.

[10] Z. Chen, Y. Luo, and N. Mesgarani, "Deep attractor network for single-microphone speaker separation," in *Proceedings of ICASSP*. IEEE, 2017, pp. 246–250.

[11] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *Proceedings of ICASSP*. IEEE, 2017, pp. 241–245.

[12] M. Kolbæk, 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.

[13] C. Xu, W. Rao, X. Xiao, E. S. Chng, and H. Li, "Single channel speech separation with constrained utterance level permutation invariant training using grid lstm," in *Proceedings of ICASSP*. IEEE, 2018.

[14] C. Xu, L. Xie, and X. Xiao, "A bidirectional lstm approach with word embeddings for sentence boundary detection," *Journal of Signal Processing Systems*, pp. 1–13, 2017.

[15] E. Ahissar, S. Nagarajan, M. Ahissar, A. Protopapas, H. Mahncke, and M. M. Merzenich, "Speech comprehension is correlated with temporal response patterns recorded from auditory cortex," *Proceedings of the National Academy of Sciences*, vol. 98, no. 23, pp. 13 367–13 372, 2001.

[16] D. H. H. Nguyen, X. Xiao, E. S. Chng, and H. Li, "Feature adaptation using linear spectro-temporal transform for robust speech recognition," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 24, no. 6, pp. 1006–1019, 2016.

[17] F. E. Theunissen, K. Sen, and A. J. Doupe, "Spectral-temporal receptive fields of nonlinear auditory neurons obtained using natural sounds," *Journal of Neuroscience*, vol. 20, no. 6, pp. 2315–2331, 2000.

[18] T. Chi, P. Ru, and S. A. Shamma, "Multiresolution spectrotemporal analysis of complex sounds," *The Journal of the Acoustical Society of America*, vol. 118, no. 2, pp. 887–906, 2005.

[19] A. M. Reddy and B. Raj, "Soft mask methods for single-channel speaker separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 6, pp. 1766–1776, 2007.

[20] M. H. Radfar and R. M. Dansereau, "Single-channel speech separation using soft mask filtering," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2299–2310, 2007.

[21] D. Wang and J. Chen, "Supervised speech separation based on deep learning: an overview," *arXiv preprint arXiv:1708.07524*, 2017.

[22] M. Delfarah and D. Wang, "Features for masking-based monaural speech separation in reverberant conditions," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 5, pp. 1085–1094, 2017.

[23] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 22, no. 12, pp. 1849–1858, 2014.

[24] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Joint optimization of masks and deep recurrent neural networks for monaural source separation," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 23, no. 12, pp. 2136–2147, 2015.

[25] S. Furui, "Speaker-independent isolated word recognition using dynamic features of speech spectrum," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, no. 1, pp. 52–59, 1986.

[26] X. Xiao, S. Zhao, D. H. H. Nguyen, X. Zhong, D. L. Jones, E. S. Chng, and H. Li, "Speech dereverberation for enhancement and recognition using dynamic features constrained deep neural networks and feature adaptation," *EURASIP Journal on Advances in Signal Processing*, vol. 2016, no. 1, p. 4, 2016.

[27] B. Bielefeld, "Language identification using shifted delta cepstrum," *Proceedings of Fourteenth Annual Speech Research Symposium*, 1994.

[28] P. A. Torres-Carrasquillo, E. Singer, M. A. Kohler, R. J. Greene, D. A. Reynolds, and J. R. Deller Jr, "Approaches to language identification using gaussian mixture models and shifted delta cepstral features," in *Seventh International Conference on Spoken Language Processing*, 2002.

[29] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *Proceedings of ICASSP*. IEEE, 2015, pp. 708–712.

[30] J. Garofolo, D. Graff, D. Paul, and D. Pallett, "Csr-i (wsj0) complete ldc93s6a," *Web Download. Philadelphia: Linguistic Data Consortium*, 1993.

[31] N. Kalchbrenner, I. Danihelka, and A. Graves, "Grid long short-term memory," *arXiv preprint arXiv:1507.01526*, 2015.

[32] T. N. Sainath and B. Li, "Modeling time-frequency patterns with lstm vs. convolutional architectures for lvcsr tasks." in *Proceedings of INTERSPEECH*, 2016, pp. 813–817.

[33] S.-Y. Chang, B. Li, T. N. Sainath, G. Simko, and C. Parada, "Endpoint detection using grid long short-term memory networks for streaming speech recognition," *Proceedings of Interspeech 2017*, pp. 3812–3816, 2017.

[34] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

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