Speech Enhancement Using Deep Complex Neural Network With Channel Attention

Minghui Hao
Beijing Jiaotong University
Beijing, China
19120008@bjtu.edu.cn

Jingjing Yu
Beijing Jiaotong University
Beijing, China
jjyu@bjtu.edu.cn

***Abstract*—The speech enhancement problem has made great progress with the development of deep learning, however, the current methods rarely consider the weight distribution of the convolution kernel extracted features in the network training process. In this paper, we designed a channel attention (CA) module to assign attention weights to different features of the convolution feature channel, and we embed our CA into the current mainstream deep complex convolution speech enhancement neural network (DCCRN). Particularly, the CA dynamically calculates the weight of the feature channel during the convolution process. We use the DNS-2020 data set to train the model. The experimental results show that the network with the CA module has a significant improvement in PESQ and STOI index, and generally has a certain degree of improvement at different SNR levels.**

Keywords- channel attention, deep complex convolution neural network, assign channel weights, single channel speech enhancement

I. INTRODUCTION

Speech enhancement refers to the task of designing an algorithm to extract the target speech from the speech containing noise (e.g. background noise, stationary/non-stationary noise), suppress the noise and improve the speech quality and intelligibility. Single-channel speech enhancement processes the noisy speech signal collected by a microphone and obtains the processed clean speech. It has important applications in many fields such as automatic speech recognition (ASR) [1], hearing aid [2], and audio surveillance system [3], etc.

Traditional single-channel speech enhancement include spectral subtraction [4], Wiener filtering [5], MMSE estimator [6], etc. In recent years, the data-driven supervised speech enhancement method based on deep neural network (DNN) realizes the mapping of noisy speech to clean speech by establishing a regression model, and has achieved a great improvement in speech enhancement performance, and has gradually become a mainstream research method [7]. [8] used DNN for sub-band decomposition to estimate the ideal binary mask (IBM) for the first time and introduced deep learning to speech enhancement. [9] used a convolutional neural network (CNN) to take the amplitude spectrum of 11 consecutive speech frames after the short-time Fourier Transform (STFT) of the speech signal as the training target, and establish the noisy speech amplitude spectrum to Clean the mapping relationship of the amplitude spectrum of the speech, and use the phase spectrum of the noisy speech and the amplitude spectrum of the enhanced speech to synthesize the enhanced speech. This method uses the powerful modeling capabilities of the deep neural network to achieve a better speech enhancement effect. [10] discussed the shortcomings of excessive neural network parameters in the speech enhancement method based on RNN network, which is difficult to train, and cannot meet the real-time requirements in daily life, proposed convolution recurrent neural network (CRN) and achieved great improvement in real-time performance. [11] uses the same network structure as [10] and uses complex convolution operation and complex LSTM blocks to process the input signal, compared with [10], the [11] can make full use of phase spectrum information for speech enhancement, the experimental results are also show the superiority of this method.

However, none of the above methods considers a situation, in the process of network training, when the convolution kernel extracts the features of the input signal, different features have different effects on the network enhanced speech, that is, different features should have different weights. The attention mechanism selectively pays attention to the input content, and selects the information that needs attention according to certain criteria, and suppresses or removes the information that does not need attention in the system input, this feature can help many tasks achieve better results. There are three main advantages of using attention to model in neural networks [12]: ① The attention neural network model can achieve optimal results in a variety of tasks, such as machine translation, question answering, sentiment analysis and part-of-speech tagging. ② It can improve the interpretability of neural network models instead of treating them as black boxes. ③ It can overcome the problem of system performance degradation caused by the increase of input sequence length and sequence processing order in the convolutional neural network. [13] proposed the use of self-attention mechanism for speech enhancement. The author believes that human beings can focus on a specific area of a piece of audio, such as paying more attention to the target speech, and assigning less attention to noise. [14] integrate the attention gate mechanism into the Unet network structure and apply it to speech enhancement tasks. [15] obtain attention weights and perform speech enhancement by calculating the similarity

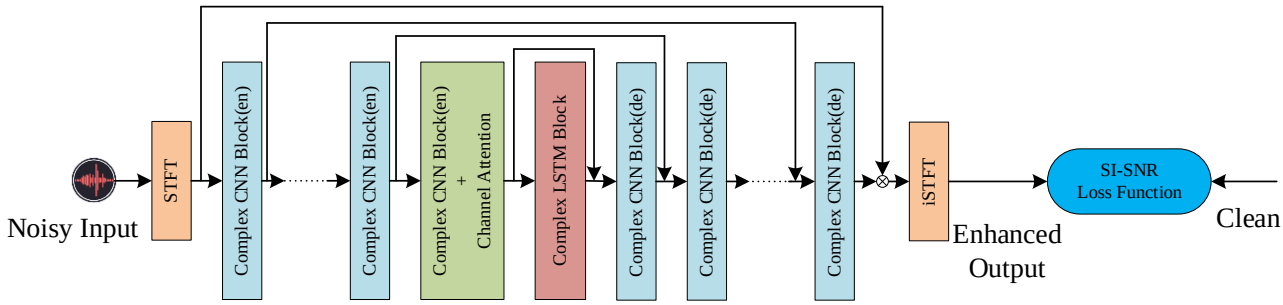

Figure 1. DCCRN-CA model structure

between noisy speech and known features (such as noise). [16] by adding a squeeze-and-excitation block to the output of the convolutional layer to change the weights of different feature channels, the channel attention was studied firstly. [17] improved [16] and perfected the mathematical theory of channel attention. In addition, spatial attention [18], time-restricted attention [19] and other methods have also caused extensive research.

In this paper, we design a channel attention module for single-channel speech enhancement and embed it in the DCCRN network, resulting in DCCRN-CA, aims to dynamically use the CA module to assign different attention weights to different feature channels, so that the network pays more attention to the processing of important features, further enhances the quality and intelligibility of speech, and suppresses noise interference. Our model is almost the same as the original DCCRN model in terms of parameter amount and computational resource consumption, with only a slight increase, but it has achieved a more obvious speech enhancement effect. The structure of this paper is as follows: Section 2 presents proposed framework for DCCRN structure and CA module. Section 3 discusses the experimental results of recovered speech in different SNR levels and network parameters. Section 4 concludes this paper.

II. Proposed Method

A. Framework Overview

The goal of single-channel speech enhancement is to estimate the target clean speech from noisy speech. $\boldsymbol{x} \in \mathbb{R}^{N\times 1}$ represents the time-domain waveform signal of noisy speech, and $\boldsymbol{X} \in \mathbb{C}^{F\times T}$ represents the complex spectrum signal of the received noisy speech after short-time Fourier transform (STFT) processing, where F and T are the number of frequency bins and the number of time frames, respectively. The problem of speech enhancement can be formalized as learning a mask function $h(\cdot)$: $\boldsymbol{X} \in \mathbb{C}^{F\times T} \xrightarrow{h(\cdot)} \hat{\boldsymbol{Y}} \in \mathbb{C}^{F\times T}$, where $\hat{\boldsymbol{Y}}$ is the estimated complex spectrum of clean speech $\boldsymbol{Y}$.

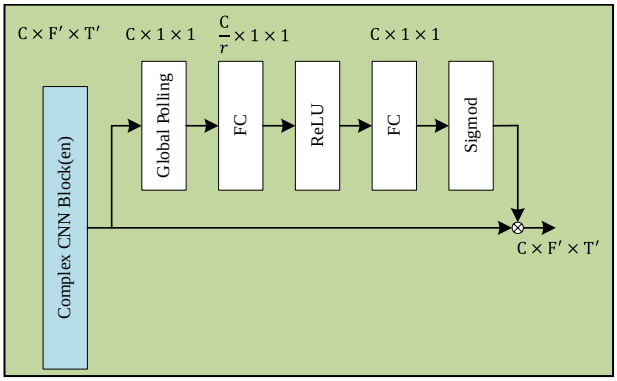

Figure 2. Complex CNN Block (en) and channel attention

The deep complex convolution recurrent network (DCCRN), originally described in [11], it is an extension of CRN. DCCRN consists of a complex encoder block, a complex decoder block and a two-layer complex LSTM block, here, unlike CRN and other networks that use noisy speech amplitude spectrum or power spectrum as network input, lost or fail to effectively use phase spectrum information, DCCRN uses complex spectrum features as network input, the complex convolutional block performs high-dimensional feature extraction on all the information of the input speech signal, and the complex LSTM block models the temporal correlation of the preceding and following speech frames.

As shown in Fig. 1, the model we used contains 6 complex CNN encoder blocks, 2 complex LSTM blocks, and 6 CNN decoder blocks. Here, in the 6th CNN encoder block, we use the channel attention mechanism to extract the correlation of feature channels, in detail, each CNN encoder/decoder block contains complex CNN layer, complex batch normalization layer and activation function layer. The encoder and decoder are symmetrical about the complex LSTM block, and the symmetrical encoder output and decoder input are added with a skip connection to increase the robustness of the model. After the network passes through the last CNN decoder block, it outputs a complex ratio mask (CRM) [20], and the network complex spectrum is multiplied by the CRM to obtain an enhanced complex spectrum. The enhanced wav is obtained after inverse short-time Fourier transform (iSTFT). The difference between the enhanced wav and the clean wav is calculated by the SI-SNR loss function, and then the back propagation of the network is completed.

B. Channel Dependency Extraction

The template is used to format your paper and style the text. All margins, column widths, line spaces, and text fonts are prescribed; please do not alter them. You may note peculiarities. For example, the head margin in this template measures

TABLE II. RESULTS (PESQ AND STOI) ON UNSEEN NOISE TYPES

	PESQ						STOI(%)					
SNR (dB)	-5	0	5	10	15	Ave.	-5	0	5	10	15	Ave.
Noisy	1.4599	1.6413	1.9179	2.2804	2.7233	2.0046	69.49	76.74	83.12	88.05	91.71	81.82
DCCRN	1.7154	1.9712	2.3085	2.6852	3.0472	2.3455	75.71	82.15	86.83	89.94	**92.16**	85.36
DCCRN-CA	**1.7420**	**2.0098**	**2.3528**	**2.7190**	**3.0667**	**2.3781**	**76.03**	**82.36**	**87.06**	**90.03**	92.11	**85.52**

TABLE I. RESULTS (PESQ AND STOI) ON SEEN NOISE TYPES

	Para. (M)	PESQ	STOI(%)
Noisy	-	1.643	78.99
DCCRN	3.76	1.996	84.64
DCCRN-CA	3.77	**2.006**	**84.73**

proportionately more than is customary. This measurement and others are deliberate, using specifications that anticipate your paper as one part of the entire proceedings, and not as an independent document. Please do not revise any of the current designations.

Let $\boldsymbol{H}^{(l)}$ denote the output of the l th complex CNN block(en), the cascaded complex CNN block(en) satisfies

$$\boldsymbol{H}^{(l+1)} = PReLU(CBN(\boldsymbol{H}^{(l)} *_{\mathrm{C}} \boldsymbol{w}^{(l+1)})) \tag{1}$$

where $*_{\mathrm{C}}$ denotes complex convolution operation, $\boldsymbol{H}^{(0)} = \boldsymbol{X}$, $\boldsymbol{w}^{(l+1)}$ denotes $l+1$ th complex convolution kernel of the complex CNN block. CBN and $PReLU$ represents complex batch normalization and parametric rectified linear unit, respectively. In the last block of the encoder (l=5), due to the addition of Channel Attention, the operation of complex CNN block(en) in this module can be calculated as

$$\boldsymbol{U} = PReLU(CBN(\boldsymbol{H}^{(5)} *_{\mathrm{C}} \boldsymbol{w}^{(5+1)})) \tag{2}$$

where $\boldsymbol{U} \in \mathbb{C}^{\mathrm{C}' \times \mathrm{F}' \times \mathrm{T}'}$, C', F', T' represents the number of channels, frequency bins, and time frames output by the module on this block. Here, the complex convolution operation assigns the same weight to each feature channel, however, the high-dimensional features of different channels have different effects on the final speech enhancement effect, so we use the channel attention (CA) mechanism to assign different attention weights to different feature channels. First, we use the global average pooling (GAP) operation to compress all the information of each feature channel into one value

$$\boldsymbol{Z} = \frac{1}{\mathrm{F}' \times \mathrm{T}'} \sum_{f=1}^{\mathrm{F}'} \sum_{t=1}^{\mathrm{T}'} \boldsymbol{U}_{:,f,t} \tag{3}$$

where $\boldsymbol{Z} \in \mathbb{C}^{\mathrm{C} \times 1 \times 1}$ can be viewed as a collection of the local descriptor, the statistical data of these descriptors can represent the entire image. Then, in order to fully extract the channel correlation between the feature channels, we adopted the following operations and finally got the output of the sixth encoder block

$$\boldsymbol{H}^{(6)} = \boldsymbol{U}(Sigmod(\boldsymbol{V}_2 ReLU(\boldsymbol{V}_1 \boldsymbol{Z}))) \tag{4}$$

where $\boldsymbol{V}_1 \in \mathbb{R}^{\frac{\mathrm{C}'}{r} \times \mathrm{C}'}$, $\boldsymbol{V}_2 \in \mathbb{R}^{\mathrm{C}' \times \frac{\mathrm{C}'}{r}}$, $ReLU(\boldsymbol{V}_1 \boldsymbol{Z})$ us used to learn the non-linear interaction between channels, r is the dimensionality-reduction layer with reduction ratio, the channel attention weights are calculated by $Sigmod(\boldsymbol{V}_2 ReLU(\boldsymbol{V}_1 \boldsymbol{Z}))$. Finally, the calculated weight of each feature channel is multiplied by the value of the feature channel to obtain the output of the 6th encoder block.

C. SI-SNR Loss Function

The loss function we use is scale-invariant source-to-noise ratio (SI-SNR), experiments show that the loss function has better performance than the MSE loss function or standard source-to-distortion ratio (SDR), SI-SNR is defined as follows

$$\begin{cases} \boldsymbol{y}_{target} := \dfrac{<\hat{\boldsymbol{y}}, \boldsymbol{y}>\boldsymbol{y}}{\|\boldsymbol{y}\|^2} \\ \boldsymbol{e}_{noise} := \hat{\boldsymbol{y}} - \boldsymbol{y}_{target} \\ \mathrm{SI-SNR} := 10log_{10} \dfrac{\|\boldsymbol{y}_{target}\|^2}{\|\boldsymbol{e}_{noise}\|^2} \end{cases} \tag{5}$$

where $\hat{\boldsymbol{y}} \in \mathbb{R}^{\mathrm{N} \times 1}$, $\boldsymbol{y} \in \mathbb{R}^{\mathrm{N} \times 1}$ are estimated enhanced speech wav and original clean speech time domain wav, respectively. $< \bullet , \bullet >$ denotes the Hadamard product. Scale-invariant is ensured by normalizing $\hat{\boldsymbol{y}}$ and $\boldsymbol{y}$ to zero-mean prior to the calculation.

III. EXPERIMENT

A. Datasets

The Dataset we use to train the model and several baselines is the DNS-2020 dataset [21], which includes a total of 500 hours of clean speech from 2150 speakers. The 180 hours DNS-2020 noise set includes 65000 noise chips from 150 noise classes. In the stage of generating noisy speech, we randomly select and generate noisy speech with a signal-to-noise ratio (SNR) between -5 and 20dB from clean speech and noise. The length of each noisy speech is 3s, and finally we get a total of 24,000 noisy speech chips, a total of 20 hours used as training set. We use two test sets to evaluate the performance of the model, the first test set comes from the noisy-clean speech data pair obtained from the DNS-2020 dataset, a total of 4800 chips, since the test set and the training set come from the same data set, we believe that the noise type of the test set is "seen". The second test set is obtained from MUSAN [22] which contains 6.2 hours free-sound noise and 42.6 hours music, in order to evaluate the speech enhancement performance of the model to different noise intensities, 5 typical SNRs (-5dB, 0dB, 5dB, 10dB, 15dB) were generated in the second test set, the noise types of this set is "unseen". The sampling rate of all audio in the data set is 16kHz.

B. Experimental Setup and Baselines

For all models, when performing STFT on time-domain wavs, the FFT length is 512, the window length and frame shift are 20ms and 10ms, respectively. We use Pytorch to train the models, the optimizer is Adam and the learning rate is set to 0.0005. The speech enhancement performance of each model is assessed by PESQ [23] and STOI [24].

- **CRN**[9]: a real-time model with the best configuration in the original paper. The input and output are amplitude spectrum of the noisy and estimated. The number of output channels of each layer in encoder is {16,32,64,128,256}. The number of LSTM units are 1024. The number of input channels of the decoder is {256+256, 128+128, 64+64,32+32,16+16} since the skip connection. During the network training process, the kernel size and the stride length of the convolution kernel are (3,2) and (2,1), respectively.
- **DCCRN**[8]: the number of output channels of encoder is {16,32,64,128,256,256}. Two complex LSTM layers are adopted and the number of units is 256. The number of input channels of the decoder is {256+256,256+256,128+128, 64+64,32+32,16+16} since the skip connection. During the network training process, the kernel size and the stride length of the convolution kernel are (5,2) and (2,1), respectively.

For our model， We use the same configuration as DCCRN. In our innovative channel attention part, the number of input channels is C=256, the reduction ratio r=16, the units of the two fully connected layers are both 4096.

C. Experimental Results and Discussion

The results in TABLE I shows that compared to the original DCCRN model, DCCRN-CA with channel attention mechanism we proposed can achieve better speech enhancement performance when processing data with unseen noise types. In general, the DCCRN model can achieve a significant increase in PESQ (about 0.3 on average) when processing noisy speech. Compared with the original model, our proposed DCCRN-CA model with channel attention mechanism has an improvement of more than 0.03 in PESQ score and a performance improvement of more than 2% in STOI. In terms of different SNR levels, our method has the most obvious improvement when the SNR is 5dB, and the PESQ scores improves more than 0.04, indicating that our method is suitable for most real-life scenarios. When the SNR is high, the characteristics of the noise are already very inconspicuous, and scenes with a very low SNR are not common in real life.

The results in Table 2 show the comparison results of PESQ and STOI in the seen noise type test set. Since our method recalculates and assigns different weights to different feature channels, it pays more attention to important features and is more effective in information extraction, the experimental results in this part also show the effectiveness of our method on PESQ and STOI index. At the same time, the method we adopted only increased the model parameters by 0.01 million, and there was almost no increase in the computation cost.

IV. Conclusion

This paper designs a channel attention module for speech enhancement, and uses a global pooling operation to obtain the feature descriptor of each feature channel, through a fully connected layer and using the ReLU activation function to learn the nonlinear interaction between feature channels, and a fully connected layer and Sigmod activation function are used to assign learned weights to each feature channel. Finally, the channel feature and the weight are multiplied to get the weighted channel feature. Embed the CA module into the DCCRN model, training with the DNS-2020 data set and compare different test sets, the results show that our method has a significant improvement in PESQ and STOI scores compared to the original DCCRN model in terms of parameter amount and computational consumption. In the future, we will compare the experimental results of adding channel attention to the feature channels of each complex CNN block, we will also try to add convolution path to skip connection to design the new channel attention mechanism.

Acknowledgment

Fundamental Research Funds for Central Universities (Grant No. 2021JBM004) is gratefully acknowledged.

References

[1] P. Cerva, J. Silovsky, J. Zdansky, J. Nouza, and L. Seps, “Speaker-adaptive speech recognition using speaker diarization for improved transcription of large spoken archives,” Speech Communication, vol. 55, no. 10, pp. 1033– 1046, 2013.

[2] Y. Muraki and M. Mori, “Effective tooth excitation position for an implanted dental-bone conduction hearing aid,” 2020 IEEE 9th Global Conference on Consumer Electronics (GCCE), 2020, pp. 24-28.

[3] H. Lim, C. Kim, E. Ekmekcioglu, S. Dogan, A. P. Hill, A. M. Kondoz and X. Shi, “An approach to immersive audio rendering with wave field synthesis for 3D multimedia content,” in 2014 IEEE International Conference on Image Processing (ICIP), 2014, pp. 76-80.

[4] K. Paliwal, K. Wojcicki, and B. Schwerin, “Single-channel speech enhancement using spectral subtraction in the short-time modulation domain,” Speech communication, vol. 52, no. 5, pp. 450–475, 2010.

[5] J. S. Lim and A. V. Oppenheim, “Enhancement and bandwidth compression of noisy speech,” Proceedings of the IEEE, vol. 67, no. 12, pp. 1586–1604, 1979.

[6] Y. Ephraim and D. Malah, “Speech enhancement using a minimum mean-square error log-spectral amplitude estimator,” IEEE transactions on acoustics, speech, and signal processing, vol. 33, no. 2, pp. 443–445, 1985.

[7] D. Wang and J. Chen, “Supervised speech separation based on deep learning: An overview,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 26, no. 10, pp. 1702–1726, 2018.

[8] Y. Wang and D. Wang. “Boosting Classification Based Speech Separation Using Temporal Dynamics,” Proc. Interspeech 2012, 2012, pp. 1528-1531.

[9] T. Kounovsky and J. Malek, “Single channel speech enhancement using convolutional neural network,” 2017 IEEE International Workshop of Electronics, Control, Measurement, Signals and their Application to Mechatronics (ECMSM), 2017, pp. 1-5.

[10] K. Tan, and D. Wang, “A convolutional recurrent neural network for real-time speech enhancement,” Proc. Interspeech 2018, 2018, pp. 3229-3233.

[11] Y. Hu, Y. Liu, S. Lv, M. Xing, S. Zhang, Y. Fu, J. Wu, B. Zhang, and L. Xie, “DCCRN: Deep complex recurrent network for phase-aware speech enhancement,” Proc. Interspeech 2020, 2020, pp. 2472-2476.

[12] S. Chaudhari, V. Mithal, G. Polatkan, and R. Ramanath, “An Attentive Survey of Attention Models,” ACM Transactions on Intelligent Systems and Technology, vol. 12, no. 53, pp. 1-32, 2021.

[13] X. Hao, C. Shan, Y. Xu, S. Sun and L. Xie, “An Attention-based Neural Network Approach for Single Channel Speech Enhancement,” ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 6895-6899.

[14] R. Giri, U. Isik and A. Krishnaswamy, “Attention Wave-U-Net for Speech Enhancement,” Proc. Interspeech 2019, pp. 249-253.

[15] Y. Wang, J. Du, L. Chai, C. Lee, and J. Pan, “A Noise-Aware Memory-Attention Network Architecture for Regression-Based Speech Enhancement,” Proc. Interspeech 2020, pp.4501-4505.

[16] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, “Squeeze-and-excitation networks,” 2019, arXiv: 1709.01507v4.

[17] Z. Qin, P. Zhang, F. Wu, and X. Li, “Fcanet: Frequency channel attention networks,” 2020, arXiv: 2012.11879v2.

[18] S. Xu and E. F. Lussier, “Spatial and Channel Attention Based Convolutional Neural Networks for Modeling Noisy Speech,” ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 6625-6629.

[19] D. Povey, H. Hadian, P. Ghahremani, K. Li and S. Khudanpur, "A Time-Restricted Self-Attention Layer for ASR," 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018, pp. 5874-5878.

[20] D. S. Williamson, Y. Wang, and D. Wang, “Complex ratio masking for monaural speech separation,” IEEE/ACM transactions on audio, speech, and language processing, vol. 24, no. 3, pp. 483-492, 2015.

[21] C. K. Reddy, V. Gopal, R. Cutler, E. Beyrami, R. Cheng, H. Dubey, S. Matusevych, R. Aichner, A. Aazami, S. Braun, P. Rana, S. Srinivasan, and J. Gehrke, “The Interspeech 2020 deep noise suppression challenge: datasets, subjective testing framework, and challenge results,” Proc. Interspeech 2020, 2020, pp. 2492–2496.

[22] D. Snyder, G. Chen, and D. Povey, “MUSAN: A Music, Speech, and Noise Corpus,” 2015, arXiv:1510.08484v1.

[23] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, “Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs,” 2001 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2001, pp. 749–752.

[24] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, “An algorithm for intelligibility prediction of time–frequency weighted noisy speech,” IEEE Transactions on Audio, Speech, and Language Processing, pp. 2125–2136, 2011.