A Pitch-aware Speaker Extraction Serial Network

Yu Jiang[1], Meng Ge[1], Longbiao Wang[1,*], Jianwu Dang[1,2], Kiyoshi Honda[1], Sulin Zhang[3], Bo Yu[3]
[1] Tianjin Key Laboratory of Cognitive Computing and Application,
College of Intelligence and Computing, Tianjin University, Tianjin, China
[2] Japan Advanced Institute of Science and Technology, Ishikawa, Japan
[3] Automotive Data of China Co., Ltd
{yu_jiang, gemeng, longbiao_wang}@tju.edu.cn, jdang@jaist.ac.jp, khonda@sannet.ne.jp

***Abstract*—Despite deep learning has an excellent performance in monaural speaker extraction, it's still a challenge to extract speakers when facing the same gender, i.e., male-male and female-female. On the other hand, it has been proved that pitch tracking is effective for same-gender speech separation. In this study, we proposed a pitch-aware speaker extraction serial network (PSESNet) to improve extraction performance. We designed a serial system and compared it with multi-task learning, we tried to use the target speaker's pitch information to optimize the loss function rather than as input to the extraction network. The extraction part uses SpeakerBeam-FE (SBF) with magnitude and temporal spectrum approximation loss (MTSAL) and speaker embedding concatenation. After extracting the spectrogram of the target speaker, we connected the spectrogram to predict the pitch information to do further optimization. Experimental results show that serial system performs better than multi-task learning and proposed method improves performance in both same and opposite gender conditions. On average, PSESNet achieves 4.7% and 3.8% relative improvements on WSJ0 dataset over the SBF-MTSAL-Concat baseline on signal-to-distortion ratio (SDR) under both closed and open condition.**

I. Introduction

Speech separation can be traced back to the cocktail party problem, which was put forward in 1953 by Cherry [1]. In a noisy cocktail party, there are many different sources making sounds at the same time: multiple speakers, noise, and the reflected sounds from the wall and objects in the room. However, people can selectively focus on interested speakers and ignore other speakers. For human beings, powerful auditory mechanisms can help us process this acoustic information perfectly, while for machines, speech separation and speaker extraction techniques play a role in interacting naturally with the target speaker in the speech overlapping condition.

Recently, some deep learning based speech separation methods have been proposed, such as Deep Clustering (DPCL) [2], [3], [4], Deep Attractor Network (DANet) [5], [6], Permutation Invariant Training (PIT) [7], [8], [9]. Although the above approaches significantly improve the performance of speech separation, they all face a common problem. In most real conditions, we can't get the number of speakers from mixture. PIT must know the above information during the training stage to solve the permutation problem and the importance of this information is also illustrated from the clustering perspective of DPCL and DANet, because during inference stage we need to form clusters that equal to the number of speakers.

*Corresponding author. Yu Jiang and Meng Ge contributed equally.

To address this limitation, target speaker extraction has attracted much attention [10], [11], [12], [13], [14]. Target speaker extraction needs not to know the number of speakers by only focusing on the interested speakers and ignoring interference speakers. The key is to use utterances of the target speaker that differ from the mixture to form the auxiliary network. In [10], the author extracts the target speaker based on SpeakerBeam (SB) and explores the extraction capabilities of SB with different configurations. In [11], with a mask-based speaker extraction front-end (SpeakerBeam-F), the author proposes a magnitude and temporal spectrum approximation loss (MTSAL) and concatenates the speaker embeddings from the auxiliary network with the mixture representations in the mask estimation network repeatedly, extraction capacity has been further improved, which is called SBF-MTSAL-Concat.

Although the above models have achieved impressive performance in speaker extraction, they lack further research on the same-gender conditions. To our knowledge, the result in SDR of the same-gender extraction is lower than that of the opposite-gender [11] and the gap is about 4dB. On the other hand, it has been proved that pitch information is effective for speech and music separation [15], [16], [17], [18], especially in the same-gender case [19]. In [20], the author solves the polyphonic pitch tracking by a regression approach and investigates a pitch-aware approach to single-channel speech separation. However, these studies only use pitch information as input to the network and achieve separation on this basis, pitch information does not participate in the optimization of the whole separation loss function and pitch tracking has not been applied to target speaker extraction.

To address these problems, we propose a pitch-aware speaker extraction serial network (PSESNet). We try to introduce pitch tracking into the target speaker extraction network and participate in loss optimization rather than as input to the network. Besides, we design a serial system and a parallel system and compare the performance of the two systems. After comparison, the serial system is selected as our framework because it can enhance the accuracy of target pitch tracking. The PSESNet system first estimates the spectrogram of the target speaker by a speaker extraction block and then connects the predicted spectrogram with a pitch-aware block to predict the target pitch information to optimize the total loss.

This paper is organized as follows. Section II describes the general speaker extraction framework with mask. Section III

introduces our model architecture in detail. Dataset, experimental setup and results are shown and discussed in Section IV with conclusions provided in Section V.

II. Speaker Extraction with Mask

The goal of the speaker extraction network is to obtain the spectrogram of the target speaker from the mixture signal. The mask estimation network and the auxiliary network work together to estimate the mask of the target speaker.

Given the mixed speech $y(n)$ and the auxiliary utterance $a(n)$, where the signal $y(n)$ is a mixture from the target speaker $x(n)$ and other interference speakers, and $a(n)$ is the speech of the target speaker different from the above, we want to get the estimated signal $\hat{x}(n)$ which is very close to the clean speech $x(n)$. The output of the network is usually a filter called mask, and we use the mask and the spectrogram of the mixed speech to do element-wise multiply to obtain the target speaker's spectrogram. The above calculation process can be summarized as follows:

$$|\hat{X}(t,f)| = M(t,f) \odot |Y(t,f)| \tag{1}$$

where $M(t,f)$ is the mask such as ideal binary mask (IBM) and $\odot$ represents element-wise multiplication. Usually auxiliary network uses $a(n)$ to obtain the target speaker characteristics and learn adaptation weights for sub-layers in the adaptation layer of the mask estimation network.

III. Model Architecture

For the problems existing in the previous studies, especially the low extraction capability of the same gender mixture and the lack of the loss function optimization with pitch tracking, we proposed a pitch-aware speaker extraction serial network, PSESNet. Since pitch information is useful for same-gender speaker separation, we use the estimated pitch information to optimize the overall loss after getting the spectrogram of the target speaker. At the same time, we proposed a series system rather than multi-task learning, which can improve the accuracy of information estimation of target pitch.

A. Framework Selection

At the beginning of the experimental design, two kinds of system frameworks were considered, as shown in Fig. 1. In the first structure, after obtaining the mask, we directly use it and the mixed signal to get the spectrogram of the target speaker, and then input the spectrogram into the pitch-aware network to predict the pitch information to optimize the loss, which is called the serial system. In the second structure, we use multi-task learning to optimize the loss by dividing mask estimation and pitch estimation into two tasks and we call it parallel system. Considering that the latter estimates the target speaker information from the mixed speech during the pitch-aware network, it is more difficult to get the estimated target pitch compared with the former. Therefore, we consider the serial system as the proposed model, and the parallel system is considered as a comparative experiment in section IV.

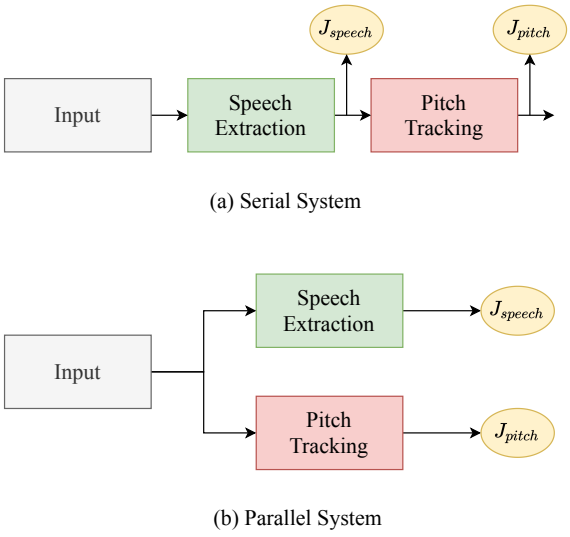

Fig. 1. (a) Serial System: After obtaining the target speaker's spectrogram, the pitch information of the target is estimated. (b) Parallel System: Target pitch information is estimated from the mixture and then combined with the extraction module to optimize loss.

We propose a pitch-aware speaker extraction serial network with two objectives. By introducing a hyperparameter, we have improved the performance of the extraction network. The total loss function considers the MTSAL in the target extraction part and the cross entropy loss in the pitch-aware module. The final proposed loss function is briefly defined as follows:

$$J = \alpha J_{speech} + (1-\alpha) J_{pitch} \tag{2}$$

where α is the hyperparameter, and we adjust the value of α to get the best result. With the addition of pitch loss, we believe there will be improvement in same-gender extraction.

B. Target Speech Extraction

The proposed system PSESNet consists of a target speech extraction block with green color and a pitch-aware network in red color as shown in Fig. 2. Different from reference [10], we use a magnitude and temporal spectrum approximation loss instead of the original mask approximation loss. The newly proposed loss calculates the signal reconstruction error between the extracted magnitude and clean label with phase difference and also computes errors across dynamic information such as delta and acceleration. Besides, the phase sensitive mask (PSM) is used in our experiment rather than the original IBM to enhance the performance [21].

In addition to the loss optimization, we redesign the output of the auxiliary network. Unlike the adaptive weights used in SBF, our auxiliary network uses the target speech, which is different from the utterances in the mixed signal, to obtain the target speaker embeddings. Then the target speaker embeddings are repeatedly concatenated with the mixture representations in the mask estimation network.

The MTSAL function contains a wealth of information. In addition to the original PSM loss, it also contains errors cross

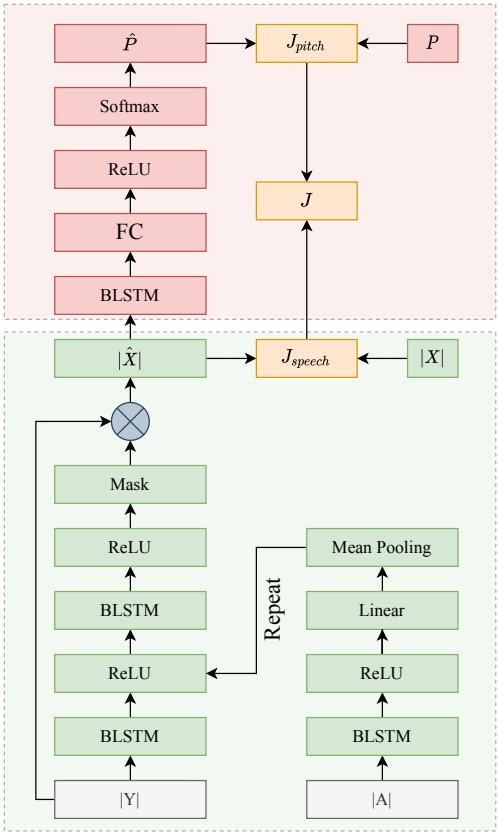

Fig. 2. The diagram of the proposed PSESNet system. $|Y|$: input mixture, $|A|$: auxiliary speech, $|X|$: clean target speech, $|\hat{X}|$: predicted speech, $\hat{P}$: estimated pitch, P: clean target pitch, FC: Fully Connected Layer. The target speaker extraction is green block and the pitch-aware network is red part.

dynamic information [22]:

$$\begin{aligned} J_{speech} = \frac{1}{T}\sum(||M \odot |Y| - |X| \odot cos(\theta_y - \theta_x)||_F^2 \\ + w_d||f_d(M \odot |Y|) - f_d(|X| \odot cos(\theta_y - \theta_x))||_F^2 \\ + w_a||f_a(M \odot |Y|) - f_a(|X| \odot cos(\theta_y - \theta_x))||_F^2) \end{aligned} \tag{3}$$

Where θ_y and θ_x represent the phase angles of the mixed signals and the target speaker's clean utterances, respectively. w_d and w_a are weighting coefficients, which are generally fixed (set as 4.5 and 10.0). f_d and f_a are formulas for calculating delta and acceleration. There is a mathematical relationship between delta and acceleration, i.e. the acceleration value is obtained by computing delta twice. So we just need to provide the delta formula, where $u(t)$ represents a time frame of magnitude, and L is the contextual window (set as 0.2):

$$f_d(u(t)) = \frac{\sum_{l=1}^{L} l \times (u(t+l) - u(t-l))}{\sum_{l=1}^{L} 2l^2} \tag{4}$$

C. Pitch Tracking

The pitch-aware network aims to make the estimated target speaker pitch information approximate to the groundtruth pitch. The network is trained to generate the posterior probabilities that the target pitch states occur at the corresponding frame. In order to simplify the calculation, we need to quantify the pitch information [23]. The pitch frequency (range from 60 to 404Hz) is converted to the corresponding 67 units using the formula $60 \times 2^{(m-1)/24} Hz(m = 1, ..., 67)$. In other words, the above formula is used to divide the frequency range into 67 quantized frequency bins $s_1, .., s_{67}$. In addition, for silent or speech-free segments, using the frequency bin s_0 represents a non-pitch state. A total of 68 quantized states are obtained by this method, and the output vector of the pitch estimation network contains 68 elements.

During the training stage, BLSTM network is used and the pitch information from clean utterances of the target speaker is extracted by Praat [24], which is a cross-platform software that can analyze speech signals. Besides, we use a fully connected layer in this block for further processing of the output dimension, as the dimension of groundtruth pitch is 68. Pitch-aware network uses cross entropy loss, which is a classic solution for classification problems. The function is defined as follows:

$$J_{pitch} = -\sum_{m=0}^{M} p_m log(s_m) \tag{5}$$

where p_m represents the real pitch state distribution, s_m is the actual output result from the network output layer and the range of m is 0 to 67, which is the result of pitch states quantization. As mentioned in part C, we have an excellent quantification method to help model grasp pitch information more specifically.

IV. Experiments and Discussion

A. Data

We used the WSJ0 database [25] to simulate the data we needed[1]. The corpus was used to simulate a two speakers mixture database with sampling rate of 8kHz, and the simulated database was divided into training set of 20,000 utterances, development set of 5,000 utterances and test set of 3000 utterances. During the simulation, the first speaker was selected as the target speaker and the other one was interference. Meanwhile, the utterance of target speaker from WSJ0 was selected as the input of the auxiliary network to obtain the acoustic information of the target speaker, which was different from the one used to generate the mixture.

The utterances from two speakers were randomly selected in WSJ0 "si_tr_s" set to generate the training and the development set with SNR between 0dB and 5dB. Likewise, the test set was generated by randomly selecting utterances from two speakers in the WSJ0 "si_dt_ 05" and "si_et_05" sets and mixing them. Since the development and the training set had same speakers, the development set can be regarded as closed condition (CC) to tune parameters while the speakers of the test set were different from the training and the development set, so it was considered as open condition (OC). Because

[1]https://github.com/xuchenglin28/speaker_extraction

TABLE I
SDR(DB) OF EXTRACTED SPEECH FOR PROPOSED METHOD WITH DIFFERENT VALUE OF α AND BASELINE UNDER CLOSED AND OPEN CONDITION.

Methods	Pitch Aware	Balance (α)	CC				OC			
			FM	FF	MM	AVG	FM	FF	MM	AVG
SBF-MTSAL-Concat [11]	NO	-	12.83	9.84	8.49	11.01	12.45	8.04	9.09	10.66
Serial System (i.e. PSESNet)	YES	0.9	12.95	9.82	8.69	11.11	12.56	7.81	9.26	10.74
		0.8	13.19	10.11	8.83	11.35	12.78	7.90	9.17	10.84
		0.7	12.82	9.62	8.04	10.85	12.46	8.01	8.71	10.51
		0.6	12.96	9.95	8.71	11.16	12.59	8.03	9.18	10.75
		0.5	13.07	10.10	9.10	11.35	12.73	**8.24**	9.35	10.92
		0.4	12.76	9.90	8.40	10.97	12.41	7.96	8.74	10.50
		0.3	12.79	9.64	8.33	10.91	12.43	7.69	8.93	10.54
		0.2	**13.21**	**10.30**	**9.33**	**11.53**	**12.80**	7.98	**9.78**	**11.07**

of the randomness of selection, the system had a significant performance for both the same and opposite gender extraction.

B. Experimental Setup

The auxiliary network used a BLSTM with 256 cells in each forward and backward direction and the following feed-forward relu hidden layer and the linear layer had 256 and 30 nodes respectively. The output of the auxiliary network was a 30 dimensional speaker embedding after a mean pooling over all frames, containing the target speaker characteristics, which was repeatedly concatenated to the activation of the BLSTM in the mask estimation network. Then the concatenated outputs were fed back to a feed-forward relu hidden layer, a BLSTM layer and another relu hidden layer with 512 cells or nodes in the mask network. The final output of the mask network was a PSM with 129 dimensions. The configuration of BLSTM used in the pitch-aware network was the same as that of the mask network, and relu hidden layer had 256 nodes here. Connecting softmax after the relu hidden layer due to the multi-classification problem. In addition, we added a fully connected layer before relu to map the dimension to 68.

In this experiment, the learning rate was initially set to 0.0005. The minibatch size was set to 8 with a minimum training of 30 epochs. We used the Adam algorithm to optimize the network and the signal distortion ratio (SDR) to evaluate performance. It should be emphasized that we reimplement SBF-MTSAL-Concat model as the baseline and train the parallel system as comparative experiment. The configuration of these two experiments is the same as the corresponding part mentioned above.

C. Experimental Results

The experimental results are shown in Table I and Table II. Table I summarizes the average SDR performance of the SBF-MTASL-Concat baseline and our proposed PSESNet with different value of α. Testing the extraction capability in both closed and open conditions, PSESNet system achieves the best performance when α is 0.2, which enhances 0.52dB and 0.41dB compared with the baseline, respectively. We pay more attention to the improvement of the open condition and can still get better performance when the speakers are unseen during training in open condition. It's worth noting that even if α is 0.2, the overall loss is dominated by the extraction part since J_{speech} is 30 times larger than J_{pitch} itself. And the value of α indicates that pitch tracking is significant in optimizing loss.

In order to investigate the performance of the PSESNet system on detailed gender related extraction, we also compare the extraction performance of the same and different gender in the above two conditions. Same-gender case is divided into male-male and female-female mixtures. Overall, the relative improvement of same-gender condition is better than opposite-gender case. We find that when adding pitch tracking and participating in loss optimization, there is improvement in almost all cases. In the closed condition, when α is 0.2, the best results are obtained in all three cases. For open conditions, except the female-female case, we still get the best results when α is set as 0.2. The possible reason is that female-female is more difficult to extract than male-male mixture, especially the speakers are unseen in open condition.

Table II verifies our conclusion in section III. Without loss of generality, we randomly choose two values of α for comparison. In the several experiments we implemented, the performance of the serial system is always optimal under both closed and open condition. The reasonable explanation is that the estimation of target pitch information from the mixed spectrogram is more difficult than estimating directly from the masked speaker spectrogram.

TABLE II
EXPERIMENTAL RESULTS OF THE TWO SYSTEMS

Model		Balance(α)	CC	OC
			AVG	AVG
SBF-MTSAL-Concat [11]		-	11.01	10.66
Pitch Aware	Serial System	0.8	**11.35**	**10.84**
		0.5	**11.35**	**10.92**
	Parallel System	0.8	10.35	9.96
		0.5	11.15	10.79

V. CONCLUSIONS

In this paper, we proposed a pitch-aware speaker extraction serial network. We combine pitch tracking with target speaker extraction and make pitch information participate in the optimization of extraction loss. Besides, compared with parallel system, serial system with better performance is proposed to improve the accuracy of pitch estimation. Experimental results show that PSESNet can achieve better SDR performance in the same and opposite gender extraction.

REFERENCES

[1] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *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 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 31–35.

[3] Y. Isik, J. L. Roux, C. Zhuo, S. Watanabe, and J. R. Hershey, "Single-channel multi-speaker separation using deep clustering," in *Interspeech*, 2016, pp. 545–549.

[4] Z. Wang, J. L. Roux, and J. R. Hershey, "Alternative objective functions for deep clustering," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 686–690.

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

[6] Y. Luo, Z. Chen, and N. Mesgarani, "Speaker–independent speech separation with deep attractor network," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 4, pp. 787–796, 2018.

[7] 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 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 6–10.

[8] M. Kolbæk, D. Yu, Z. 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.

[9] D. Yu, M. Kolbæk, Z. Tan, and J. 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.

[10] M. Delcroix, K. Zmolikova, K. Kinoshita, A. Ogawa, and T. Nakatani, "Single channel target speaker extraction and recognition with speaker beam," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5554–5558.

[11] C. Xu, W. Rao, E. S. Chng, and H. Li, "Optimization of speaker extraction neural network with magnitude and temporal spectrum approximation loss," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6990–6994.

[12] K. Žmolíková, M. Delcroix, K. Kinoshita, T. Ochiai, T. Nakatani, L. Burget, and J. Černockỳ, "SpeakerBeam: Speaker aware neural network for target speaker extraction in speech mixtures," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 4, pp. 800–814, 2019.

[13] M. Delcroix, K. Zmolikova, T. Ochiai, K. Kinoshita, S. Araki, and T. Nakatani, "Compact network for speakerbeam target speaker extraction," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6965–6969.

[14] M. Ge, C. Xu, L. Wang, E. S. Chng, J. Dang, and H. Li, "Spex+: A complete time domain speaker extraction network," in *Proc. Interspeech*, 2020.

[15] G. Hu and D. Wang, "A tandem algorithm for pitch estimation and voiced speech segregation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 8, pp. 2067–2079, 2010.

[16] T. Virtanen, A. Mesaros, and M. Ryynanen, "Combining pitch-based inference and non-negative spectrogram factorization in separating vocals from polyphonic music." in *Interspeech*, 2008, pp. 17–22.

[17] S. Lin, "Jointly tracking and separating speech sources using multiple features and the generalized labeled multi-bernoulli framework," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 3211–3215.

[18] S. W. Lee, F. K. Soong, P. C. Ching, and L. Tan, "Pitch tracking for model-based speech separation," in *International Symposium on Chinese Spoken Language Processing*, 2008.

[19] Y. Liu and D. Wang, "Permutation invariant training for speaker-independent multi-pitch tracking," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5594–5598.

[20] K. Wang, F. K. Soong, and L. Xie, "A pitch-aware approach to single-channel speech separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 296–300.

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

[22] 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, p. 52–59, 1986.

[23] J. Zhang, T. Jian, and L. R. Dai, "Rnn-blstm based multi-pitch estimation," in *Interspeech*, 2016, pp. 1785–1789.

[24] S. Gonzalez and M. Brookes, "Pefac - a pitch estimation algorithm robust to high levels of noise," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 2, pp. 518–530, 2014.

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