The Speaker and Language Recognition Workshop (Odyssey 2022)
28 June - 1 July 2022, Beijing, China

SINGLE-CHANNEL TARGET SPEAKER SEPARATION USING JOINT TRAINING WITH TARGET SPEAKER'S PITCH INFORMATION

Jincheng He[1], *Yuanyuan Bao*[1], *Na Xu*[2], *Hongfeng Li*[2],
Shicong Li[2], *Linzhang Wang*[2], *Fei Xiang*[2], *Ming Li*[1,3]

[1]Data Science Research Center, Duke Kunshan University, Kunshan, China
[2]Xiaomi, Beijing, China
[3]School of Computer Science, Wuhan University, Wuhan, China
ming.li369@duke.edu

Abstract

Despite the great progress achieved in the target speaker separation (TSS) task, we are still trying to find other robust ways for performance improvement which are independent of the model architecture and the training loss. Pitch extraction plays an important role in many applications such as speech enhancement and speech separation. It is also a challenging task when there are multiple speakers in the same utterance. In this paper, we explore if the target speaker pitch extraction is possible and how the extracted target pitch could help to improve the TSS performance. A target pitch extraction model is built and incorporated into different TSS models using two different strategies, namely concatenation and joint training. The experimental results on the LibriSpeech dataset show that both training strategies could bring significant improvements to the TSS task, even the precision of the target pitch extraction module is not high enough.

1. Introduction

Target speaker separation (TSS) has attracted much attention in recent years [1, 2, 3, 4, 5, 6, 7, 8, 9]. It is the task which only extracts the speech of the target speaker in the environment with multiple people speaking simultaneously. The general deep neural network based TSS framework could be summarized as an Encoder (including the speech and speaker encoder)-Separator-Decorder architecture, shown as Figure 1.

The related works, such as VoiceFilter [3], Atss-Net [4], spex++ [5, 6, 7], made efforts in different parts of the aforementioned architecture. The Atss-Net introduced attention mechanisms in the separator. The spex++ adopted the time-domain method and made lots of changes in the speech and speaker encoder. All of them contribute a lot to the development of TSS task.

Despite the great progress made, we are motivated to explore useful and robust training strategies that could be applied to different model architectures. For instance, use new feature as one of the inputs of separator.

Pitch, or fundamental frequency, is an important characteristic of speech and music signals. The task of pitch extraction, or pitch tracking has a long history. There are multiple signal processing based methods to extract pitch. A time domain signal processing method is proposed in [10] to estimate the foundamental frequency. A frequency-domain signal processing method is proposed in [11].

Before the usage of DNN methods for extracting pitch, there are some traditional signal processing methods, and although they have the advantage that the algorithms are easy to understand and do not require training data, they have limitations in terms of accuracy especially in complex environments. Hence, many machine learning based algorithms were developed. A supervised machine learning based algorithm in the time domain is proposed in [12]. A self-supervised machine learning based algorithm in the frequency domain is proposed in [13].

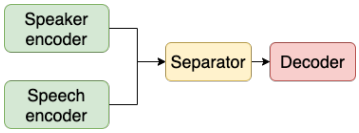

Figure 1: The Encoder-Separator-Decoder architecture

Using pitch information to help speech separation task also attracts a lot of attention in recent years. A pitch extraction module is concatenated with the separation module together to perform the separation task in [14]. A serial model is built and the final loss is designed as a weighted sum of the speech separation loss and pitch loss in [15]. However the serial model in [15] needs to go through the target speaker extraction first and then perform the pitch tracking after the extraction.

In our paper, we propose a target speaker pitch extraction module which can directly estimate the target speaker's pitch from a mixture of utterances from multiple speakers. Then we explore the strategies on how to contribute this target speaker's pitch information to the target speaker separation task. We propose a small scale Multi-Block RNNoise (MBRNN) model (details in 2.3) as our baseline speech separation system. Then we propose two training strategies, namely concatenation training and joint training. We further implement these two strategies on multiple models with different scales and the experiment results show that the joint training of the target pitch extraction model and the target speaker separation model is useful to improve the separation performance. The proposed strategies could make positive impact on the TSS task even though the precision of the target pitch extraction is not high enough. The performance of concatenation with ground-truth pitch information show great potential in utilizing the target speaker's pitch information for the TSS task.

2. Model Architecture

In this section, we give a detailed description of our target pitch extraction module and the training strategies of incorporating

10.21437/Odyssey.2022-42

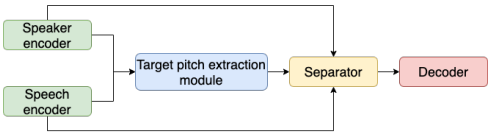

Figure 2: TSS model with pitch concatenation.

pitch information in the speech separation model. We have three modules in our framework: 1) the speaker embedding extraction module, which accepts the enrollment utterance and outputs the 128-dimensional speaker embedding the same way as in [4]. 2) the target pitch extraction module, which accepts the magnitude spectrogram of the mixed utterance and the target speaker embedding, then outputs the 1-dimensional target pitch value; 3) the TSS module, which accepts the mixed utterance, the target speaker embedding, the target pitch information, then ouputs the estimated target speaker's voice.

2.1. Training strategies with pitch

2.1.1. Concatenation

The architecture of the TSS model with simply pitch information concatenated is illustrated in Figure 2.

The extracted target pitch, together with the speaker embedding, are concatenated with the audio feature along the feature axis and then fed into the speech separation model. In this concatenation strategy, the speaker embedding extraction module and the target pitch extraction module are all pre-trained well-performed models. Thus the parameters of both models are frozen in the speech separation stage.

2.1.2. Joint training

The structure of joint training is as same as in section 2.1.1. The difference is that the target pitch extraction model and the TSS model are optimized jointly. The experimental results of this strategy suggest better performance than the concatenation strategy.

2.2. Target pitch extraction module

Our target pitch extraction model is LSTM-based. The structure is shown in Figure 3. The model takes in the spectrogram of the mixed utterance and the target speaker embedding, and outputs the 1-dimensional pitch information $f_0 \in \mathbb{R}^{T\times 1}$ of the target speaker, where T is the number of frames of the mixed utterance. The ground-truth pitch information is extracted using the RAPT algorithm [16] on the pre-mix clean signal, the pitch range is set to $60 \sim 404$ Hz. We see this target pitch extraction work as a regression problem and adopt L1 loss as the loss function:

$$L = min|\hat{f}_0 - f_0| \tag{1}$$

where $\hat{f}_0$, f_0 denote the estimated and the ground-truth pitch respectively.

We employ the precision rate (PR) to evaluate the pitch extraction result [17], which is defined as follow:

$$PR = \frac{N^{0.05}}{N} \tag{2}$$

where $N^{0.05}$ denotes the number of frames in which the estimated pitch deviates less than 5% from the ground-truth pitch, and N denotes the total frames of mixed utterance.

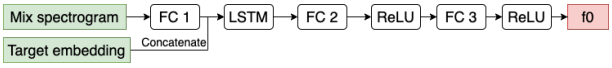

Figure 3: Model structure of pitch extraction model.

2.3. MBRNN

Our proposed small scale separation model, named as Multi-Block RNNoise (MBRNN), is modified from the well-known RNNoise [18] model in speech enhancement. The architecture of MBRNN model is shown in Figure 4. Each RNNblock in MBRNN includes a fully-connected (FC) layer and a RNNoise-like module (shown in Figure 5). After the concatenation between speaker embedding and the magnitude spectrogram, the FC layer is used to compress the feature dimension into a fixed smaller scope. The RNNblock is repeated for 4 times to improve the representational capacity. In order to ease the training problem of deep neural network, a cumulative layer normalizaton (cLN) [19] is adopted between RNNoise blocks.

We use a Conv1D layer to accelerate the computation of STFT. As the regular time-frequency domain method, only the magnitude spectrogram of the mixed uttenrace $X \in \mathbb{R}^{T\times F}$ is fed into the following network, the phase spectrogram $P \in \mathbb{R}^{T\times F}$ is used to reconstruct the estimated target signal at the end, where T denotes the number of frames and F denotes the spectrogram bin axis. The estimated magnitude spectrogram is the element-wise product between the mixed spectrogram and the estimated mask $M \in \mathbb{R}^{T\times F}$. The estimated magnitude spectrogram and the mixed phase spectrogram P are fed into a Conv-Trans1D to perform inverse short-time Fourier transform (iSTFT) to get the estimated target speech. It can be expressed as equation 3:

$$\hat{s} = \text{Conv\_iSTFT}\left(\text{ReLU}\left(M \odot X\right), P\right) \tag{3}$$

And we choose scale-invariant source-to-noise ratio (SI-SNR) as our training target [20].

3. Experiments

3.1. Dataset description

Our experiments are conducted on the LibriSpeech dataset, and we use the same training and testing tuple as same as Google used in VoiceFilter [3]. The mixed utterances are all truncated to 5 seconds in the training stage. We mix the utterances to 0dB in SNR.

3.2. Target pitch training

The hidden units of LSTM in the target pitch extraction is set to 300. The window length and hop size are 25ms and 10ms as same as the speech separation model used. And we perform a 512-point STFT on the mixed utterance. To evaluate the pitch extraction ability of our model, we also trained a clean pitch extractor on single speaker clean data. The PR result is shown in Table 1.

Type	PR(%)
Single speaker pitch extraction on clean data	93.06
Target pitch extractor on mixture data	70.27

Table 1: PR results of different type of pitch extraction models.

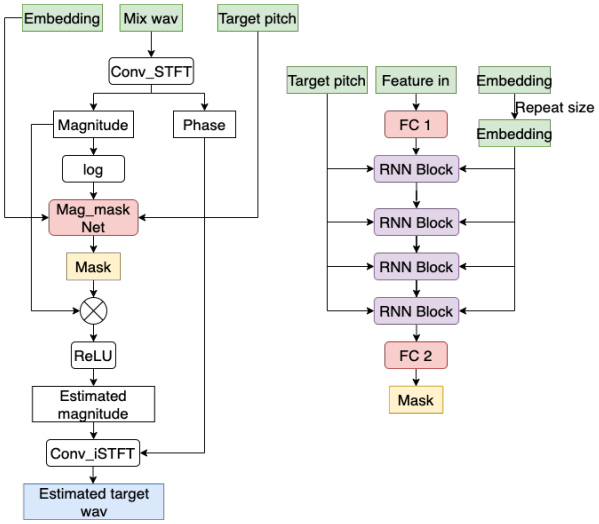

Figure 4: Model architecture of MBRNN model combining with target pitch information, the left side of the figure is the whole architecture of MBRNN with target pitch, the right side of the figure is the detail structure of Mag_mask Net in the left side, for our experiments, we choose the number of RNN Blocks is 4.

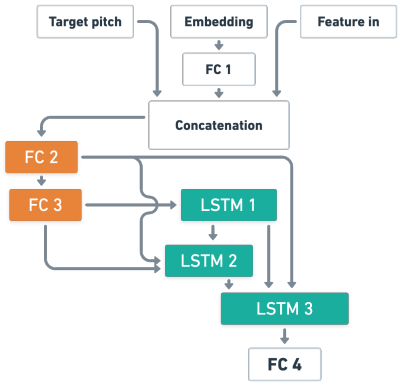

Figure 5: Detailed structure of RNN Block in MBRNN model.

A target pitch extraction examples selected from the test set is shown in Figure 6.

3.3. Implementation details

Our experiments are done using PyTorch [21]. In Table 2, TSS and target pitch extraction (TPE) mean the number of model parameters of speech separation model and target pitch extraction model respectively. MBRNN (Baseline) means the baseline MBRNN model which does not use target pitch information. MBRNN Pitch (Concatenation) and MBRNN Pitch (Joint train) mean the MBRNN model uses target pitch information with the training strategy in Section 2.1.1 and Section 2.1.2 respectively.

We also validate our proposed methods on the well known VoiceFilter baseline model which has larger scale compared to our MBRNN model. For the VoiceFilter model, we use MSE loss instead of SI-SNR loss for all experiments for keeping the same setup as the original one. However for the faster training

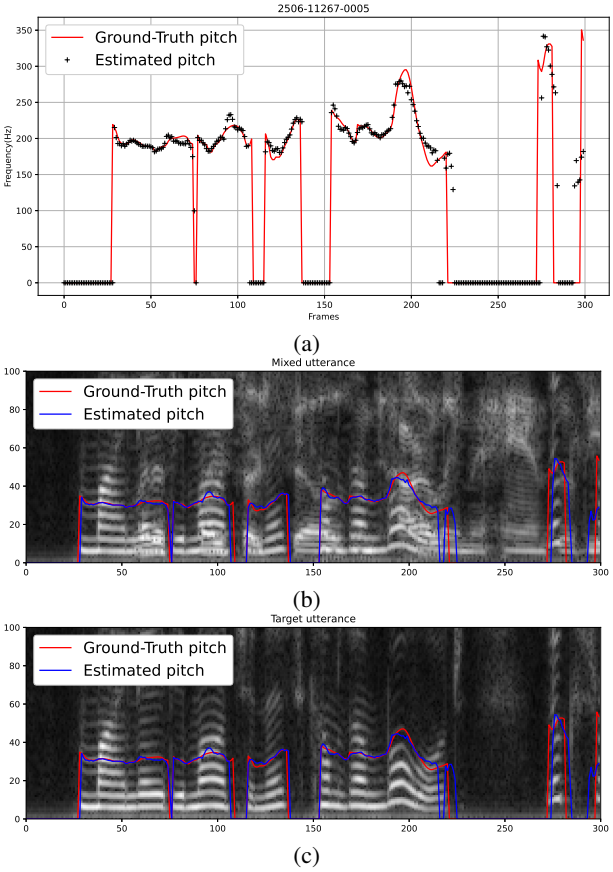

Figure 6: Target pitch extraction result of 2506-11267-0005 item in the test set. (a) Estimated pitch compared with the ground-truth pitch; (b) and (c) Estimated and ground-truth pitch respectively represented on the magnitude spectrogram, only the low frequency part is showed.

time, we change the channels to 16 in CNNs for those which are originally 64 and 8 in VoiceFilter. The re-implementation in [4] of VoiceFilter is 9.04, the re-implementation of VoiceFilter here is 9.02. Both separation models and target pitch extraction model, we did 512-point STFT on the utterance, the window length is 25ms and the hop length is 10ms. The frequency dimension of the spectrogram is 257 and the embedding dimension is 128.

In the target pitch extraction module, the first fully connected layer (FC 1) maps the vector from frequency dimension (257) into embedding dimension (128). Then concatenate with the embedding, so the input dimension into LSTM is $2 \times$ embedding_dim. In LSTM, the hidden dimension is 300, number of recurrent layers is 2, set dropout as 0.3. For FC 2, it maps the dimension from 300 to 128, a ReLU activation is added after FC 2. FC 3 maps the dimension from 128 to 1 and set ReLU activation after FC 3.

In RNN Block in MBRNN model, FC 1 maps from embedding_dim to embedding_dim $-$ 1. Before FC 2, concatenate the three inputs. FC 2 maps from hidden_dim $\times$ amp $+$ embedding_dim to rnn_units $\times$ amp, amp and rnn_units are set to 1 and 38 respectively in our experiments. FC 3 maps to 24. For LSTM 1 - 3, the hidden sizes are 24, 48, 96 respectively. FC 4 maps dimension to 64. The output of each RNN Block is the sum of the output from FC 4 and the output feature from the

Model (strategy)	#PARAM TSS(TPE)	Mean SDR Before	After	Improvement
MBRNN (Baseline)	0.60M	1.26	6.09	4.83
MBRNN Pitch (Concatenation)	0.60M(1.46M)	1.26	6.28	5.02
MBRNN Pitch (Joint train)	0.60M(1.46M)	1.26	**7.11**	**5.85**
MBRNN Pitch (Ground truth pitch)	0.60M	1.26	9.80	8.54
Our implementation of VoiceFilter (Baseline)	15.52M	1.26	9.02	7.76
Our implementation of VoiceFilter Pitch (Joint train)	15.53M(1.46M)	1.26	**9.20**	**7.94**
VoiceFilter [3] (Baseline)	9.45M	1.26	9.04	7.78
VoiceFilter [3] Pitch (Joint train)	9.46M(1.46M)	1.26	**9.60**	**8.34**

Table 2: Results comparison between different training strategies and the baseline with MBRNN and VoiceFilter model. "Before" means the SDR value of the mixed utterance, "After" means the SDR value of the estimated speech.

last layer before RNN Block.

All the experiments use Adam optimizer [22] and 10^{-4} as initial learning rate. For the experiments which choose MBRNN as the baseline, their batch size are 128, and for VoiceFilter, their batch size are 64.

3.4. Results discussion

From Table 2, due to the small model size of MBRNN, the baseline of MBRNN is a little bit weak. But from the results of both MBRNN and VoiceFilter we can see that target pitch information can help TSS model no matter in small model and large model. For the concatenation training strategy in MBRNN, it can improve 5.02 dB compared to 4.83 dB in the baseline, and the joint training strategy can improve 5.85 dB. So joint training strategy is 0.83 dB higher than the concatenation strategy. To validate the idea that the target pitch information is indeed useful for TSS model, we use the ground truth target pitch and found out it can help MBRNN reach 8.54dB improvement which is very significant, 2.69dB higher compared to the joint training strategy. We can see that a high PR target pitch extraction can really help the TSS model, even without high precision target pitch extraction, by using the joint training strategy, it can still help the TSS model improve the performance.

Moreover, from the experiments done on the VoiceFilter baseline, we can show that the joint training strategy on TSS model and pre-trained target pitch extraction model is a robust strategy to help TSS model improve the performance. And our experiments showed that even though the precision of target pitch extraction is not very high, the joint training strategy is better than simply concatenating pitch.

For the future works, we will continue to work on using target pitch on the time-domain model.

4. Conclusions

In this paper, we propose the idea that using target speaker's pitch as an auxiliary feature to improve the performance of target speaker separation. A target pitch extraction model is built and the target pitch information is incorporated with the TSS models in both simply concatenation and joint training strategies. We found that the target pitch information could improve the separation performance even though the pitch precision is not high enough yet. While the performance of concatenation with ground-truth pitch information show the great potency of this approach. The joint training approach yields better performance than simple concatenation. We also explore if joint training would bring improvement to the target pitch extraction, the result shows no obvious help. In the future work, we will continue to improve the precision of target pitch extraction and do more experiments on large scale models to validate the proposed methods.

5. Acknowledgment

This research is funded in part by the National Natural Science Foundation of China (62171207), Science and Technology Program of Guangzhou City (202007030011) and Xiaomi. Many thanks for the computational resource provided by the Advanced Computing East China Sub-Center.

6. References

[1] Kateřina Žmolíková, Marc Delcroix, Keisuke Kinoshita, Tsubasa Ochiai, Tomohiro Nakatani, Lukáš Burget, and Jan Černocký, "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.

[2] Marc Delcroix, Katerina Zmolikova, Tsubasa Ochiai, Keisuke Kinoshita, Shoko Araki, and Tomohiro Nakatani, "Compact network for speakerbeam target speaker extraction," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6965–6969.

[3] Quan Wang, Hannah Muckenhirn, Kevin Wilson, Prashant Sridhar, Zelin Wu, John R. Hershey, Rif A. Saurous, Ron J. Weiss, Ye Jia, and Ignacio Lopez Moreno, "VoiceFilter: Targeted Voice Separation by Speaker-Conditioned Spectrogram Masking," in *Proc. Interspeech 2019*, 2019, pp. 2728–2732.

[4] Tingle Li, Qingjian Lin, Yuanyuan Bao, and Ming Li, "Atss-Net: Target Speaker Separation via Attention-Based Neural Network," in *Proc. Interspeech 2020*, 2020, pp. 1411–1415.

[5] Chenglin Xu, Wei Rao, Eng Siong Chng, and Haizhou Li, "Time-domain speaker extraction network," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019, pp. 327–334.

[6] Chenglin Xu, Wei Rao, Eng Siong Chng, and Haizhou Li, "Spex: Multi-scale time domain speaker extraction network," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1370–1384, 2020.

[7] Meng Ge, Chenglin Xu, Longbiao Wang, Eng Siong Chng, Jianwu Dang, and Haizhou Li, "SpEx+: A Complete Time Domain Speaker Extraction Network," in *Proc. Interspeech 2020*, 2020, pp. 1406–1410.

[8] Shulin He, Hao Li, and Xueliang Zhang, "Speakerfilter: Deep learning-based target speaker extraction using anchor speech," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 376–380.

[9] Shulin He, Hao Li, and Xueliang Zhang, "Speakerfilter-pro: an improved target speaker extractor combines the time domain and frequency domain," 2020.

[10] Alain de Cheveigné and Hideki Kawahara, "Yin, a fundamental frequency estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1917–1930, 2002.

[11] Arturo Camacho and John G. Harris, "A sawtooth waveform inspired pitch estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 124, no. 3, pp. 1638–1652, 2008.

[12] Jong Wook Kim, Justin Salamon, Peter Li, and Juan Pablo Bello, "Crepe: A convolutional representation for pitch estimation," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 161–165.

[13] Beat Gfeller, Christian Frank, Dominik Roblek, Matt Sharifi, Marco Tagliasacchi, and Mihajlo Velimirović, "Spice: Self-supervised pitch estimation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1118–1128, 2020.

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

[15] Yu Jiang, Meng Ge, Longbiao Wang, Jianwu Dang, Kiyoshi Honda, Sulin Zhang, and Bo Yu, "A pitch-aware speaker extraction serial network," in *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2020, pp. 616–620.

[16] David Talkin, "A robust algorithm for pitch tracking (rapt)," 2005.

[17] Jianshu Zhang, Jian Tang, and Li-Rong Dai, "RNN-BLSTM Based Multi-Pitch Estimation," in *Proc. Interspeech 2016*, 2016, pp. 1785–1789.

[18] Jean-Marc Valin, "A hybrid dsp/deep learning approach to real-time full-band speech enhancement," in *2018 IEEE 20th International Workshop on Multimedia Signal Processing (MMSP)*, 2018, pp. 1–5.

[19] Yi Luo and Nima Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[20] Yi Luo and Nima Mesgarani, "Tasnet: Time-domain audio separation network for real-time, single-channel speech separation," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 696–700.

[21] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds. 2019, vol. 32, Curran Associates, Inc.

[22] Diederik P. Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," 2017.