A STUDY ON THE DEEP LEARNING BASED TARGET SPEAKER SPEECH SEPARATION ALGORITHMS

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

[1]Data Science Research Center, Duke Kunshan University, Kunshan, China [2]Xiaomi, Beijing, 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.

Index Terms— target speaker separation, target pitch extraction, joint learning

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 pitches. 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 pitches, 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 based in the time domain is proposed in [12]. A self-supervised machine learning based algorithm in the frequency domain is proposed in [13]. 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 design the final loss as a weighted loss with 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 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.

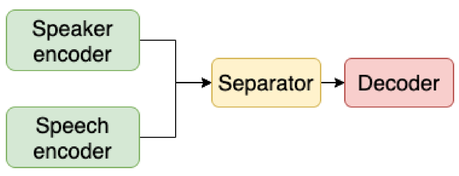

Figure 1. The Encoder-Separator-Decoder architecture

Experiments

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.

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.

A target pitch extraction examples selected from the test set is shown in Figure 2.

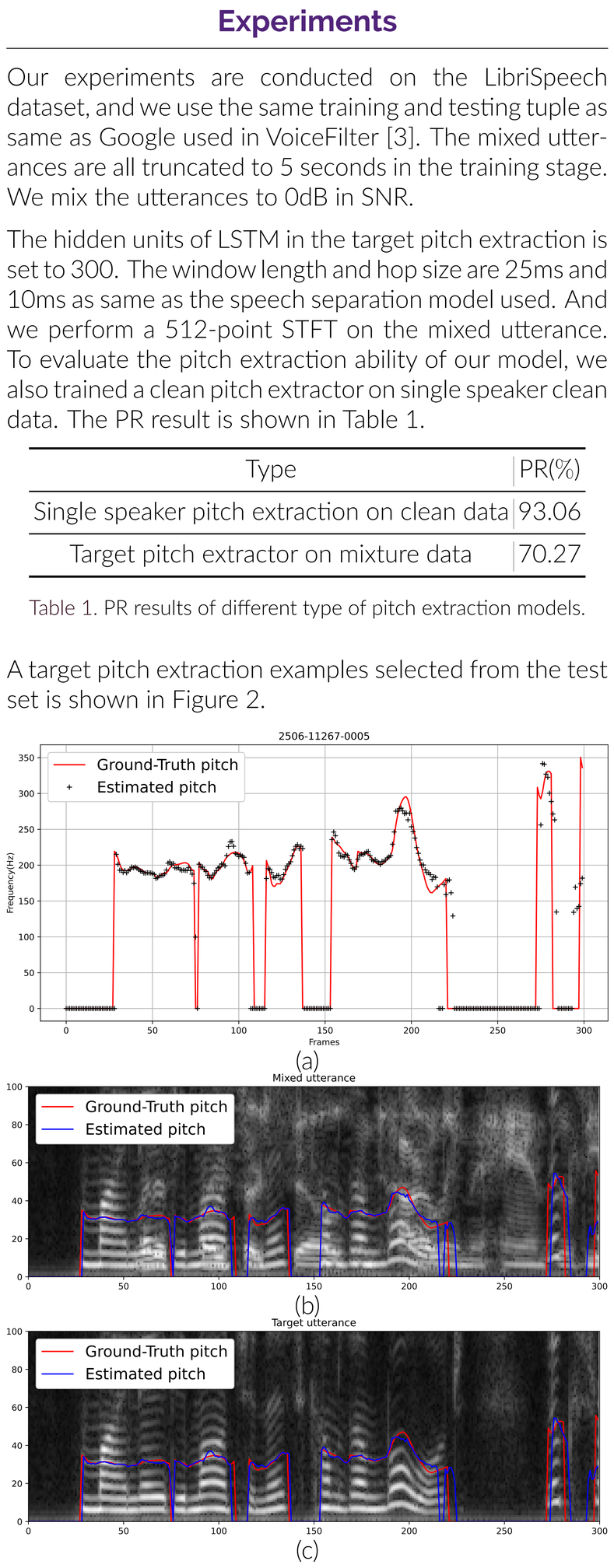

Figure 2. Target pitch extraction result of 2506-11267-0005 item in the test set. (a) Estimated pitch compared with the ground-truth pitch; (b) and (c) Estimated and ground-truth pitch respectively represented on the magnitude spectrogram, only the low frequency part is showed.

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.

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.