A PITCH-AWARE APPROACH TO SINGLE-CHANNEL SPEECH SEPARATION

Ke Wang[†,‡]*, Frank Soong*[‡]*, Lei Xie*[†]

[†] Shaanxi Provincial Key Laboratory of Speech and Image Information Processing,
School of Computer Science, Northwestern Polytechnical University, Xi'an, China
[‡] Microsoft Research Asia (MSRA), Beijing, China

ABSTRACT

Despite significant advancements of deep learning on separating speech sources mixed in a single channel, same gender speaker mix, i.e., male-male or female-female, is still more difficult to separate than the case of opposite gender mix. In this study, we propose a pitch-aware speech separation approach to improve the speech separation performance. The proposed approach performs speech separation in the following steps: 1) training a pre-separation model to separate the mixed sources; 2) training a pitch-tracking network to perform polyphonic pitch tracking; 3) incorporating the estimated pitch for the final pitch-aware speech separation. Experimental results of the new approach, tested on the WSJ0-2mix public dataset, show that the new approach improves speech separation performance for both same and opposite gender mixture. The improved performance in signal-to-distortion (SDR) of 12.0 dB is the best reported result without using any phase enhancement.

Index Terms— speech separation, deep clustering, permutation invariant training, pitch tracking

1. INTRODUCTION

Speech separation [1] has been the focus for over several decades. Thanks to the developments of deep learning, we have witnessed exciting progress towards solving this mysterious *cocktail party* effect. Specially, the invention of deep clustering [2, 3, 4] and permutation invariant training [5] have dramatically improved the performance of single-channel, speaker independent, multi-speaker speech separation. Recently, their combinations, i.e., chimera network [6, 7] and computational auditory scene analysis (CASA) approach [8], have become the state-of-the-art short-time Fourier transform (STFT) based systems. Moreover, another interesting work named time-domain audio separation network (TasNet) [9], which performs speech separation in the time domain instead, established a new state-of-the-art.

In deep clustering (DPCL) [2], a deep recurrent neural network (RNN) is trained to transform each time-frequency (T-F) bin to a high dimensional discriminative embedding space where T-F bins belong to the same speaker to be closer to each other and further away otherwise. In this embedding space, a clustering algorithm (e.g., K-means) is applied to identify the clusters. Finally, a binary mask according to the clustering results is constructed to separate the mixed speech. On the other hand, permutation invariant training (PIT) [5] directly pools over all possible permutation for the mixing sources and uses the permutation with the lowest error to update the network. Then separated speech can be directly available by PIT without an extra clustering step. In addition, PIT shows the comparable performance with DPCL [5].

Although all the aforementioned works have achieved impressive performance over the traditional signal processing methods in single-channel speech separation. We find that the same-gender mixed speech separation is harder than opposite-gender case and the separation result of the same-gender mixture is often worse than the opposite-gender mixture by about 3 dB in signal-to-distortion ratio (SDR) [10]. Moreover, we also find that separating female-female (FF) mixture is more challenging than male-male (MM) mixture.

As we know, voicing is produced by regular opening and closure of vocal folds and the detailed geometry of vocal folds is somewhat speaker-specific [11]. In most cases, the pitch of speech corresponds very nearly to the vibration frequency of vocal folds. Considering the nature of multi-speaker speech separation, which aims at separating overlapped speech from multiple speakers, it's natural to contemplate pitch information from specific speakers to further improve the performance.

We have noticed that there are some studies [12, 13] successfully integrating pitch information to music separation. But to the best of our knowledge, the integration of pitch information in speech separation has not been explored and polyphonic pitch tracking still should be an arguably unsolved open problem. Liu *et al.* [14] have explored uPIT-based multi-speaker pitch tracking and they adopted the pitch tracking as a classification problem. In this paper, we investigate a regression approach to polyphonic pitch tracking and propose a pitch-aware approach to single-channel, speaker independent, multi-speaker speech separation. The proposed approach performs speech separation in the following steps: 1)

This work was done while K. Wang was an intern at MSRA. The corresponding author is L. Xie (e-mail: lxie@nwpu.edu.cn).

training a pre-separation model to separate the mixed sources; 2) training a pitch-tracking network to perform polyphonic pitch tracking; 3) incorporating the estimated pitch for the final pitch-aware speech separation. Our experimental results show that pitch information is a key element to lead tangible improvements when combined with conventional STFT-domain frameworks. We also strongly believe that pitch is also helpful for the time-domain speech separation.

2. MODEL DESCRIPTION

2.1. Pitch-Aware Speech Separation

A block diagram of the proposed two-stage pitch-aware speech separation framework is depicted in Fig. 1. In the first pitch tracking stage, a deep clustering model [2] is trained to do deep embedding. Then a trainable component is trained to do clustering (i.e., learning mask for each source). After masking, another component performs pitch tracking for each source. In the second speech separation stage, the estimated pitch from a well-trained pitch estimation model will be augmented with the corresponding mixture as the input to final separation model.

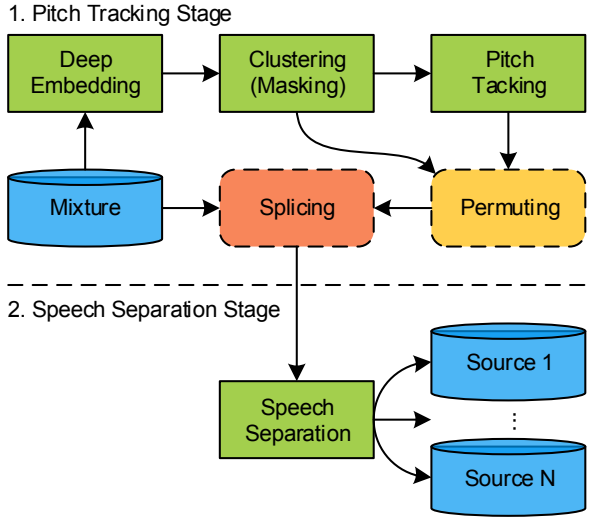

Fig. 1. A block diagram of the proposed framework. The solid green rectangular parts are trainable.

For deep embedding component, the network computes a unit-length embedding vector $v_i \in \mathbb{R}^{1\times D}$ corresponding to the i-th T-F element, and $y_i \in \mathbb{R}^{1\times C}$ is a one-hot label vector indicating which source in the mixture dominates the T-F bin i, where D is the embedding dimension and C is the number of source. Vertically stacking these, we form an embedding matrix $V \in \mathbb{R}^{TF\times D}$, and a label matrix $Y \in \mathbb{R}^{TF\times C}$. The embeddings are learned by minimizing the following objective function:

$$\begin{aligned}\mathcal{L}_{\text{DPCL}}(V,Y) &= \|VV^T - YY^T\|_{\text{F}}^2 \\ &= \|V^TV\|_{\text{F}}^2 - 2\|V^TY\|_{\text{F}}^2 + \|Y^TY\|_{\text{F}}^2,\end{aligned} \tag{1}$$

where $\|\cdot\|_{\text{F}}^2$ is the squared Frobenius norm.

Unlike DPCL++ [3], we adopt 2 feed-forward layers with ReLU activation to perform soft K-means. After clustering, masks are applied to the mixture to get the final pre-separated sources. During clustering component training, we optimize the following phase-sensitive spectrum approximation (PSA) loss function:

$$\begin{aligned}\mathcal{L}_{\text{PSA}} = \min_{\pi\in\mathcal{P}} \frac{1}{B}\sum_c \Big\| &\hat{M}_{\pi(c)} \odot |X| \\ &- (|S_c| \odot \cos(\angle S_c - \angle X)) \Big\|_2^2,\end{aligned} \tag{2}$$

where $\mathcal{P}$ is the set of permutations $\{1,\ldots,C\}$, B is the total number of frames over all sources, $\hat{M}_c$ is the c-th estimated mask, $|X|$ is the mixture magnitude, $|S_c|$ is the magnitude of the c-th reference source, $\odot$ denotes element-wise matrix multiplication, $\angle S_c$ is the phase of the c-th source, $\angle X$ is the mixture phase, and $\|\cdot\|_2^2$ is the squared Euclid norm. For pitch tracking component, the uPIT loss is also adopted. Finally, any sensible separation frameworks can be used for final separating model after augmenting the predicted pitch with the mixture.

2.2. Pitch Tracking and Permutation

Fundamental frequency (F0) is an intrinsic property of periodic signals. Here, as shown in Fig. 2, we can train a shared pitch-tracking model for all the separated sources or train C separate models for each source, where C is the number of sources.

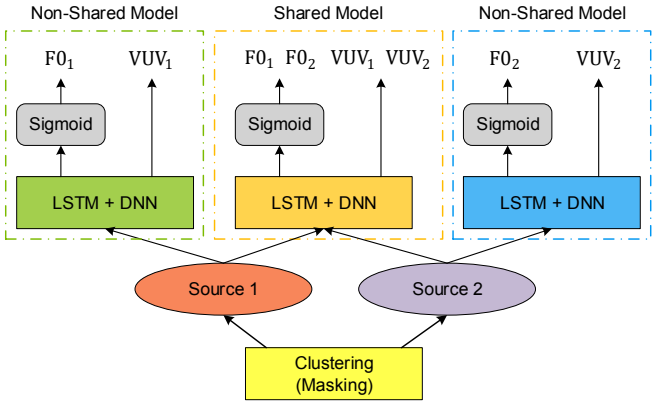

Fig. 2. Pitch-tracking component. *Non-Shared Model*: Separating pitch-tracking model for each source. *Shared-Model*: All sources share the same pitch-tracking model.

Recall that the order of the speakers in the target does not need to be the same as the order of the speakers in the output of network because of the *permutation problem* [2]. After pitch tracking, the order of the predicted pitch is also unknown. If splicing the estimated pitch with the mixture

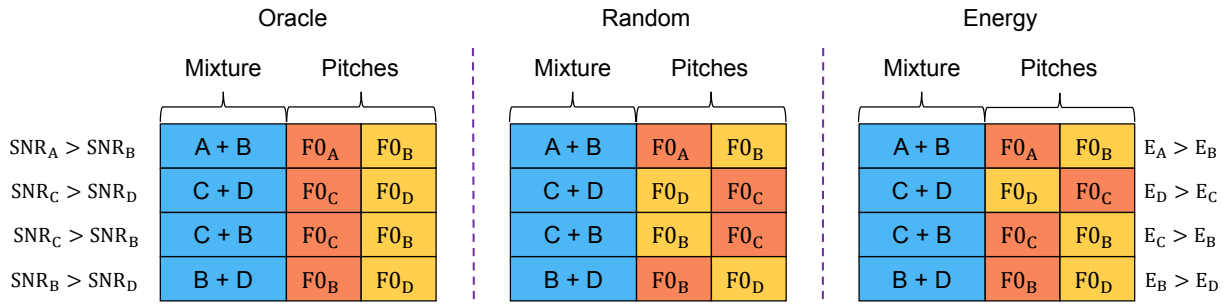

Fig. 3. Three permutation schemes for permuting component.

directly, the neural network may need to learn the relationship between the random-order input pitch and the random-order output separated sources. As shown in Sec. 3.2, this bi-directional uncertainty will make the pitch ineffective in speech separation. Thus for the effective pitch-aware speech separation system, a permutation component is added between two stages.

In Fig 3, three different permutation schemes are shown to set the order of the input in the separating stage.

- *Oracle*: We assume we know the procedure of mixture creation and the pitch of the source with the *ideal* highest signal-to-noise ratio (SNR) will be augmented next to the mixture. Similar to other sources.
- *Random*: The pitches of mixture will be randomly permuted and augmented with the mixture.
- *Energy*: The pitch of the source with the *predicted* highest average energy will be the first augmented feature to the mixture. Similar to other sources.

3. EVALUATION

3.1. Experimental setup

We evaluated our proposed methods on the widely used and publicly available WSJ0-2mix corpus [2]. It contains 20,000, 5,000 and 3,000 two-speaker mixtures in its 30 hours training, 10 hours validation and 5 hours test sets, respectively. The mixtures are generated at random signal-to-noise ratios (SNR) between -2.5 dB and 2.5 dB. Moreover, 49 male and 51 female speakers in training and validation sets are available during training, while 16 speakers in the test set are totally unseen. The sampling rate is 8 kHz. The window length is 32 ms and the hop size is 8 ms. The square rooted Hanning window is employed as the analysis and synthesis window. A 256-point STFT is performed to extract 129-dimensional magnitude input features.

In order to keep the same configurations with previous work [3, 5], our DPCL-based model contains 4 BLSTM layers with 300 units in each direction and PIT-based model contains 3 BLSTM layers with 896 units in each direction. A dropout of 0.3 is applied on each BLSTM layer except the last one and a dropout of 0.1 or 0.4 is applied between the final LSTM layer and feed-forward layer. The networks are trained with 8 full-length utterances parallel processing using Adam [15] algorithm. All systems are implemented using PyTorch [16].

3.2. Separation with Ideal Pitch

In this section, we show the potential of pitch-aware speech separation. When the oracle pitch information is available, the SDR results are reported in Table 1. Note that, for some systems [3, 4, 5], their results are SDR improvement (SDRi), while all of our results and some latest results [6, 7] are reported in SDR. So we manually add 0.2 dB to their final results although the SDR result of the mixture is about 0.15 dB. Moreover, the reproduced result is slightly different from the original DPCL [3] performance which is about 10.5 dB. It may due to the fact that the input feature is linear magnitude, the network is trained starting from random initialization and processing with the full-length utterance instead of 400-frame segments. And for uPIT [5], our configurations of the frame shift, hop size and window type are the same with DPCL instead of the original uPIT. Our reproduced results show that separating the same-gender mixture is more challenging than separating the opposite-gender mixture and separating the female-female mixture is most challenging.

From Table 1, we can find both DPCL-based model and uPIT-based model can give good results. With the ideal pitch-aware augmentation, the gap between the same-gender (SG) mixture, i.e., female-female (FF) and male-male (MM) mixture, and the opposite-gender mixture, i.e., male-female (MF) mixture, is reduced. The average performance can also be further improved. Here, we also find pitch-aware speech separation is sensitive to the order of pitches because of bi-directional uncertainty. Moreover, ideal pitch combined with uPIT performs better than DPCL combined with ideal pitch (13.4 vs. 12.2). This could be because DPCL needs to produce embedding and perform an extra clustering step. The DPCL's final performance may depend on the embedding dimension and the clustering algorithm.

Table 1. SDR (dB) performance with different level of ideal pitch. "Random", "Energy" and "Oracle" represent the order of pitches. "*" means our reproduced results.

Approaches	MF	FF	MM	SG	AVG
DPCL [3]	-	-	-	-	10.5
DPCL*	11.8	8.3	9.2	8.9	10.4
+ Random	11.9	8.5	9.4	9.2	10.5
+ Energy	**12.5**	**12.6**	**11.6**	**11.9**	**12.2**
+ Oracle	12.5	12.5	11.6	11.8	12.2
uPIT [5]	-	-	-	-	9.6
uPIT*	11.3	7.1	7.9	7.7	9.5
+ Random	11.5	7.2	8.1	7.8	9.7
+ Energy	**13.7**	**13.9**	**12.8**	**13.1**	**13.4**
+ Oracle	13.7	13.8	12.8	13.0	13.3

Table 2. VUV error rates (%) and RMSE of F0 for WSJ0-2mix validation and test set in the first pitch estimation stage.

# Parames ($\times 10^6$)		14.48		14.84	
Shared Model		Y	Y	N	N
Joint Training		Y	N	Y	N
Validation	VUV Err (%)	5.9	**5.6**	5.9	5.9
	F0 RMSE (Hz)	**12.6**	12.7	12.9	13.0
Test	VUV Err (%)	6.1	**5.8**	6.2	6.2
	F0 RMSE (Hz)	**14.6**	**14.6**	15.0	14.8

3.3. Pitch-Aware Separation Results

We adopt the baseline DPCL model as our deep embedding model, and two feed-forward layers with 300 units and ReLU activation as our clustering model. As for pitch estimating model, one LSTM layer with 300 hidden cells, one feed-forward layer with sigmoid activation which predicts voiced and unvoiced (VUV) flag, and one linear layer which predicts F0, are adopted. In addition, the parameters of pitch estimating model can be shared or non-shared by each sources. In all of our experiments, we use RAPT [11] algorithm to extract the ground-truth F0 and VUV labels from the clean sources.

For pitch estimation, the deep embedding component, clustering component and pitch tracking component can be trained step by step. Moreover, when training the next component, the previous component(s) can be jointly trained or frozen. If previous part(s) are trainable, we denote this strategy as *joint training*. Table 2 shows the performance of pitch tracking. We can predict F0 and VUV well after pre-separation. When we use shared pitch-tracking model to predict the pitch of each source, the results are slightly better than using 2 separate models to estimate the pitch of each source separately, because the training data for the shared model is twice in size comparing to the non-shared model. Furthermore, joint training does not yield improvement, but with longer training time. Thus we select the shared model without joint training as our final first stage's model.

Table 3. Comparison with other systems on WSJ0-2mix on SDR (dB).

Approaches	MF	FF	MM	SG	AVG
DPCL++ [3]	12.2	-	-	9.6	11.0
ADANet [4]	-	-	-	-	11.0
uPIT-ST [5]	12.4	-	-	7.7	10.2
Chimera++ [6]	-	-	-	-	11.2
CASA-E2E [8]	12.4	-	-	9.8	11.2
one-stage	12.1	9.1	9.6	9.5	10.8
two-stage					
+ DPCL	12.1	9.0	9.7	9.5	10.8
+ uPIT	**13.3**	**10.1**	**10.8**	**10.6**	**12.0**
T/S	11.4	6.5	7.8	7.5	9.4

In Table 3, we list the SDR performance of different systems. If we use the pre-separation stage's masking results as separated sources, the SDR is 10.8 dB, which is slightly worse than Chimera++ [6]. Concatenating predictive pitch and mixed feature as the input to the final separation model, the average SDR is 10.8 dB when we use DPCL as second separation model. When we use uPIT as the final separation model, it improves the performance to 12.0 dB, which is a new state-of-the-art result without using the phase enhancement. Moreover, inspired by parallel wavenet [17, 18], we also tried teacher-student (T/S) learning [19] framework to learn some knowledge from ideal pitch-aware model. The best T/S result, which is shown in Table 3, is worse than that of uPIT.

4. CONCLUSION AND DISCUSSION

In this paper, we investigate a pitch-aware approach for single-channel speech separation. We show that pitch information can be instrumental for further improving the separation performance. Significant improvements are obtained by training the deep separation model with the estimated pitch info. Recently, time-domain separation and phase reconstruction were proposed to further improve the performance by alleviating the phase error in reconstruction and achieved a new state-of-the-art separation performance. Our work here aims at improving the accuracy of separated magnitude, which can be refined in our future work by combining it with the time-domain approach and incorporating the phase reconstruction process.

5. ACKNOWLEDGEMENT

The research work is supported by the National Natural Science Foundation of China (No.61571363) and the authors also would like to thank Zhong-Qiu Wang from The Ohio State University for his helpful comments on this work.

6. REFERENCES

[1] DeLiang Wang and Jitong Chen, "Supervised Speech Separation Based on Deep Learning: An Overview," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 26, pp. 1702–1726, 2018.

[2] John R. Hershey, Zhuo Chen, Jonathan Le Roux, and Shinji Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 31–35.

[3] Yusuf Isik, Jonathan Le Roux, Zhuo Chen, Shinji Watanabe, and John R. Hershey, "Single-channel multi-speaker separation using deep clustering," in *Interspeech 2016*, 2016, pp. 545–549.

[4] Yi Luo, Zhuo Chen, and Nima 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.

[5] Morten Kolbæk, Dong Yu, Zheng Hua Tan, and Jesper 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.

[6] Zhong-Qiu Wang, Jonathan Le Roux, and John R Hershey, "Alternative Objective Functions for Deep Clustering," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 686–690.

[7] Zhong-Qiu Wang, Jonathan Le Roux, DeLiang Wang, and John R. Hershey, "End-to-End Speech Separation with Unfolded Iterative Phase Reconstruction," in *Interspeech 2018*, 2018, pp. 2708–2712.

[8] Yuzhou Liu and DeLiang Wang, "A CASA Approach to Deep Learning Based Speaker-Independent Co-Channel Speech Separation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5399–5403.

[9] Yi Luo and Nima Mesgarani, "TasNet: Surpassing Ideal Time-Frequency Masking for Speech Separation," *arXiv preprint arXiv:1809.07454*, 2018.

[10] Emmanuel Vincent, Rémi Gribonval, and Cédric Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, pp. 1462–1469, 2006.

[11] David Talkin, "A Robust Algorithm for Pitch Tracking (RAPT)," *Speech Coding and Synthesis,*, pp. 495–518, 1995.

[12] Yipeng Li and DeLiang Wang, "Separation of Singing Voice from Music Accompaniment for Monaural Recordings," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 4, pp. 1475–1487, 2007.

[13] Tuomas Virtanen, Annamaria Mesaros, and Matti Ryynänen, "Combining Pitch-Based Inference and Non-Negative Spectrogram Factorization in Separating Vocals from Polyphonic Music," in *SAPA@Interspeech*. 2008, pp. 17–22, ISCA.

[14] Yuzhou Liu and DeLiang Wang, "Permutation Invariant Training for Speaker-Independent Multi-Pitch Tracking," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5594–5598.

[15] Diederik P. Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations (ICLR)*, 2015.

[16] Adam Paszke, Gregory Chanan, Zeming Lin, Sam Gross, Edward Yang, Luca Antiga, and Zachary Devito, "Automatic differentiation in PyTorch," in *NIPS-W*, 2017.

[17] Aaron Van Den Oord, Yazhe Li, Igor Babuschkin, Karen Simonyan, Oriol Vinyals, Koray Kavukcuoglu, and Others, "Parallel WaveNet : Fast High-Fidelity Speech Synthesis," *arXiv preprint arXiv:1711.10433*, 2017.

[18] Wei Ping, Kainan Peng, and Jitong Chen, "ClariNet : Parallel Wave Generation in End-to-End," *arXiv preprint arXiv:1807.07281*, 2018.

[19] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean, "Distilling the Knowledge in a Neural Network," *arXiv preprint arXiv:1503.02531*, 2015.