A Pitch-Aware Approach to Single-Channel Speech Separation

Ke Wang[†,‡], Frank K. Soong[‡], Lei Xie[†]
[†] Northwestern Polytechnical University, Xi'an, China
[‡] Microsoft Research Asia (MSRA), Beijing, China

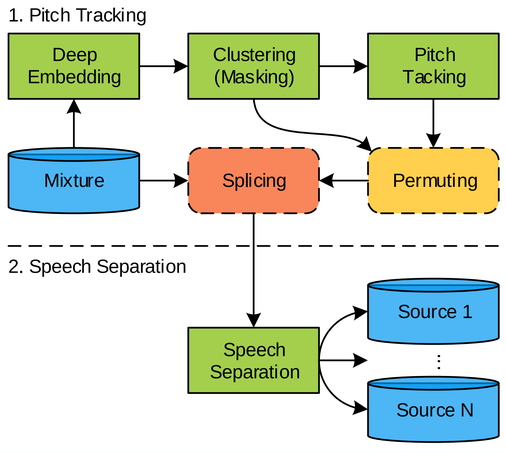
1. Introduction

Propose a pitch-aware speech separation approach:

1. training a pre-separation model to separate mixed sources
2. training a polyphonic pitch-tracking network
3. incorporating estimated pitches for final pitch-aware speech separation

Tested on WSJ0-2mix, the improved performance of 12.0 dB SDR is the best-reported result without using any phase information.

2. Model



3. Objective Function

⇒ **Deep Clustering (DPCL)** [1, 2]

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

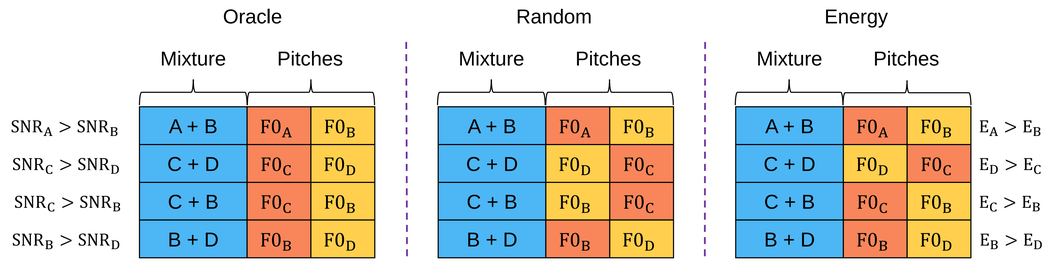
where V: embedding matrix and Y: corresponding label matrix (i.e., IBM).

⇒ **Utterance-level Permutation Invariant Training (uPIT)** [3]

$$\mathcal{L}_{\text{PSA}} = \min_{\pi\in\mathcal{P}} \frac{1}{B}\sum_{c} \left\| \hat{M}_{\pi(c)} \odot |X| - (|S_c| \odot \cos(\angle S_c - \angle X)) \right\|_2^2,$$

where $\hat{M}_c$: c-th estimated mask; $|X|$: mixture magnitude; $|S_c|$: magnitude of c-th reference source; $\angle S_c$: phase of c-th source; $\angle X$: mixture noisy phase.

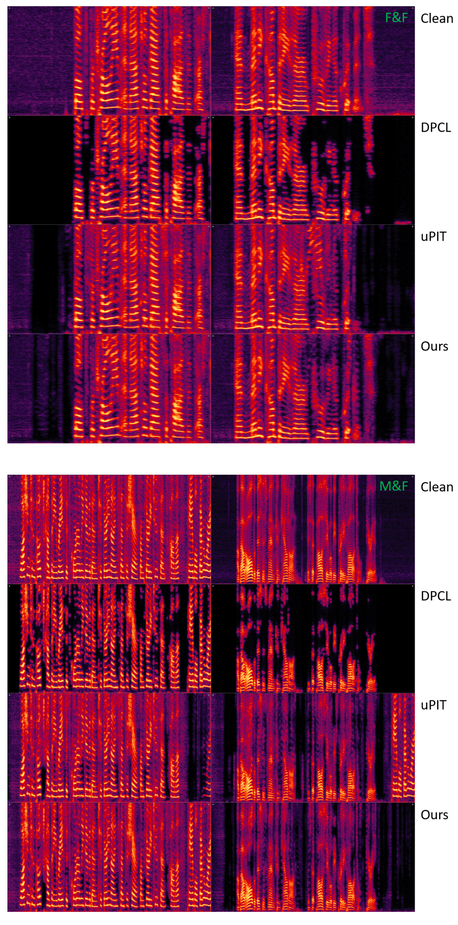
4. Permutation Schemes



5. Experimental Results

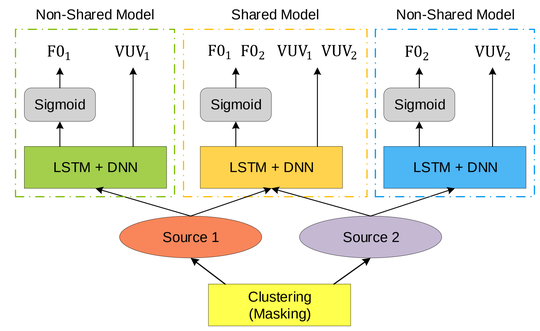
Approaches	MF	FF	MM	SG	AVG
DPCL	-	-	-	-	10.5
DPCL (Our)	11.8	8.3	9.2	8.9	10.4
+ Random	11.9	8.5	9.4	9.2	10.5
+ Energy	**12.5**	**12.6**	**11.6**	**11.9**	**12.2**
+ Oracle	12.5	12.5	11.6	11.8	12.2
uPIT	-	-	-	-	9.6
uPIT (Our)	11.3	7.1	7.9	7.7	9.5
+ Random	11.5	7.2	8.1	7.8	9.7
+ Energy	**13.7**	**13.9**	**12.8**	**13.1**	**13.4**
+ Oracle	13.7	13.8	12.8	13.0	13.3
DPCL++	12.2	-	-	9.6	11.0
ADANet	-	-	-	-	11.0
uPIT-ST	12.4	-	-	7.7	10.2
Chimera++	-	-	-	-	11.2
CASA-E2E	12.4	-	-	9.8	11.2
two-stage					
+ DPCL	12.1	9.0	9.7	9.5	10.8
+ uPIT	**13.3**	**10.1**	**10.8**	**10.6**	**12.0**

6. Separation Examples



https://wangkenpu.github.io/jekyll/update/2019/05/03/ICASSP19_Demo.html

7. Pitch Tracking



# Parames ($\times 10^6$)		14.48		14.84	
Shared Model		Y	Y	N	N
Joint Training		Y	N	Y	N
Valid	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

8. Conclusion & Discussion

- Pitch is instrumental for speech separation.
- Significant improvements obtained with estimated pitch.
- Reasonable permutation strategy is the key to make pitch-aware approach effective.
- To be combined with time-domain approach or phase reconstruction process in the future.

9. Key References

[1] J. R. Hershey *et al.*, "Deep Clustering: Discriminative Embeddings for Segmentation and Separation," in *ICASSP*, 2016.

[2] Y. Isik *et al.*, "Single-channel multi-speaker separation using deep clustering," in *Interspeech*, 2016.

[3] M. Kolbæk *et al.*, "Multi-Talker Speech Separation With Utterance-Level Permutation Invariant Training of Deep Recurrent Neural Networks," *TASLP*, 2017.

[4] D. Talkin, "A Robust Algorithm for Pitch Tracking (RAPT)," *Speech Coding and Synthesis*,, 1995.