CBLDNN-BASED SPEAKER-INDEPENDENT SPEECH SEPARATION VIA GENERATIVE ADVERSARIAL TRAINING

{Chenxing Li, Lei Zhu, Shuang Xu, Peng Gao, Bo Xu} CASIA & UCAS & ROKID

Objectives

In recent years, speaker-independent multi-speaker separation has attracted more and more attention, which has not been solved well by previous methods. An effective monaural speech separation system (CBLDNN-GAT) is designed to improve the performance:

- A more sophisticated structure, CBLDNN ;
- Multi-task learning strategy;
- Generative adversarial training.

Our system aims at capturing perceptually relevant differences and obtaining better speech quality instead of only minimizing a mean square error.

Experimental Setup

CLDNN-GAT is evaluated on WSJ0-2mix. SDR improvement is used to evaluate the performance. WSJ0-2mix contains 30 hours of training data, 10 hours of development data and 5 hours of test data.

The input features of generator and discriminator are 129-dimensional STFT spectral magnitude computed with a frame size of 32ms and 16ms shift. For MTL, 40-dimensional fBank features and 3-dimensional pitch features are extracted. The phase of the source signal is used to build PSM-based loss function, and the phase of the mixed speech is used to restore the speech.

Multi-task Learning & Baselines Results

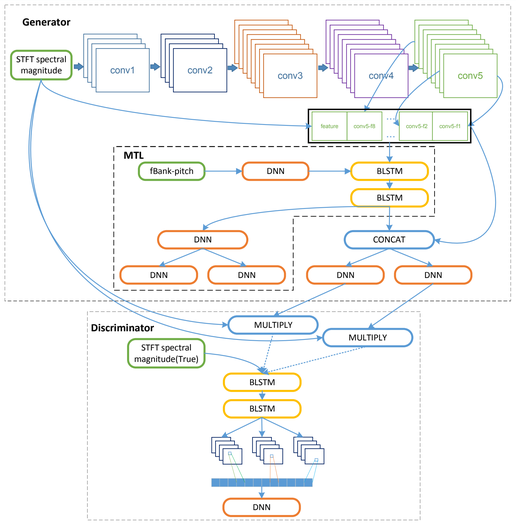
Model	Mask	Activation	SDR Imp. (dB)	
			Dev set	Test set
CBLDNN	IAM	Sigmoid	8.2	8.3
		ReLU	8.7	8.9
		Softmax	8.9	8.9
	PSM	Sigmoid	9.4	9.3
		ReLU	**9.6**	**9.7**
		Softmax	9.6	9.6
CBLDNN-MTL	IAM	Sigmoid	8.8	8.8
		ReLU	9.0	9.1
		Softmax	9.2	9.2
	PSM	Sigmoid	9.6	9.6
		ReLU	9.7	9.7
		Softmax	**9.8**	**9.8**

Table 1: SDR improvement for different separation methods.

MTL is an effective approach to improve the performance of a single task with the help of other related tasks. The human perception of the frequency contents of speech signals is nonlinear. FBank is based on the human peripheral auditory system. Pitch is an important cue in CASA.

Several CBLDNN-based baselines are conducted by using utterance-level PIT. We aim to separate the speech not only with considerable numerical error reduction but with high quality. FBank-pitch-based speech separation is utilized as another task in MTL. Besides, we only use MTL for 10 rounds to warm up the network.

CLDNN-GAT Network



For solving tracing and permutation problem, an utterance-level CLDNN-based generator is proposed. And an utterance-level multi-scale BLCDNN-based discriminator is utilized.

Loss Function

We train the network by GAT, and LSGAN-based method is utilized. L_1-regularization is utilized to guide the training.

$$\min_D \mathcal{L}(D) = \mathbb{E}_{|X|\sim p_{data}(|X|)}[(D(|X|)-1)^2] + \mathbb{E}_{|Y|\sim p_{data}(|Y|)}[(D(G(|Y|)\times|Y|))^2],$$

$$\min_G \mathcal{L}(G) = \mathbb{E}_{|Y|\sim p_{data}(|Y|)}[(D(G(|Y|)\times|Y|)-1)^2] + \lambda\mathcal{L}_1^{PSM}.$$

After employing MTL, final loss has the form:

$$\mathcal{L} = \begin{cases} \mathcal{L}_{T_1} + \mu\mathcal{L}_{T_2}, & if\ epoch \le 10 \\ \mathcal{L}_{T_1}, & if\ epoch > 10 \end{cases},$$

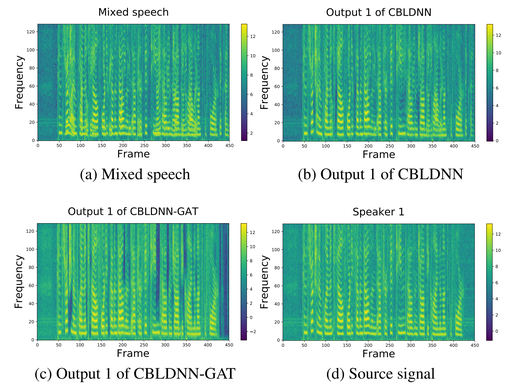
where PSM-based L_1-regularization is:

$$\mathcal{L} = \frac{1}{N}\sum_{s=1}^{S} ||M_s^{PSM}\times|Y| - |X_{\phi^*}|\cos(\theta_y - \theta\phi(s))||.$$

CBLDNN-GAT Separation Systems

Model	Activation	SDR Imp. (dB)	
		Dev set	Test set
CBLDNN-L1 loss	Sigmoid	9.7	9.7
	ReLU	9.6	9.6
	Softmax	9.5	9.6
CBLDNN-GAT	Sigmoid	**11.0**	**11.0**
	ReLU	10.7	10.8
	Softmax	10.6	10.6
DPCL	-	5.9	5.8
DPCL+	-	-	9.1
DPCL++	-	-	10.8
DANet	-	-	9.6
DANet-6 anchor	-	-	10.4
uPIT-BLSTM	ReLU	9.4	9.4
uPIT-BLSTM-ST	ReLU	10.0	10.0
IRM	-	12.4	12.7

Table 2: SDR improvement for different separation methods. Only PSM-based methods are evaluated in this section



(a) Mixed speech (b) Output 1 of CBLDNN

(c) Output 1 of CBLDNN-GAT (d) Source signal

An example is presented, which is randomly selected. The output of CBLDNN achieves 10.02dB SDR improvement, and the output of CBLDNN-GAT achieves 11.5dB SDR improvement. More examples of separated speech are provided at https://github.com/chenxinglili/SpeechSeparationExamples

References

[1] Dong Yu et al. Permutation invariant training of deep models for speaker-independent multi-talker speech separation. In *ICASSP*, 2017.

[2] Ian Goodfellow et al. Generative adversarial nets. In *Advances in neural information processing systems*, 2014.

Future Research

Speech separation remains an important issue in many applications, such as conference transcription system and large vocabulary continuous speech recognition (LVCSR) system. In the future, we will study into: **(1).** increasing the number of training epochs of MTL; **(2).** extending the experiment to three-speaker mix task; **(3).** testing the speech recognition performance to explore the impact of spectrums miss and evaluating the speech quality in objective standards.

Contact Information

Web http://english.ia.cas.cn/
Email {lichenxing2015,shuang.xu,xubo}@ia.ac.cn
Email {zhulei,gaopeng}@rokid.com
Phone +86-18810689316
WeChat lichenxing007 (lower-case)