Exploring Aligned Lyrics-Informed Singing Voice Separation

Chang-Bin Jeon, Hyeong-Seok Choi and Kyogu Lee
Seoul National University, Music and Audio Research Group

contact : vinyne@snu.ac.kr
paper : preprint

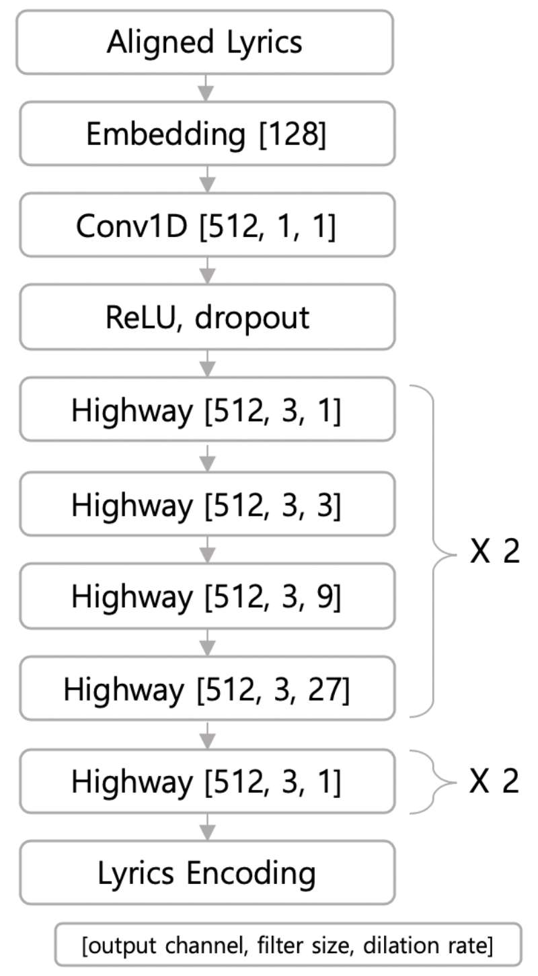
Motivation

- · Singing voice separation with additional information
  - Still, few deep learning-based methods.
- · In case of music signals, scores and lyrics can be used
  - Lyrics are more easy to collect.
  - Lyrics have rich information, such as formant frequencies.
  - Let's use lyrics as additional information.
  - We assumed that lyrics are pre-aligned.

Contribution

- · We proposed the singing voice separation network utilizing the additional aligned lyrics information.
  - The highway network-based lyrics encoder + state-of-the-art *Open-unmix* separation network
- · We analyzed the cause of the performance improvement.
  - Aligned lyrics have both vocal activity and phonetic information.
  - We checked the performance gain is due to the phonetic information of aligned lyrics.

Model – Lyrics Encoder

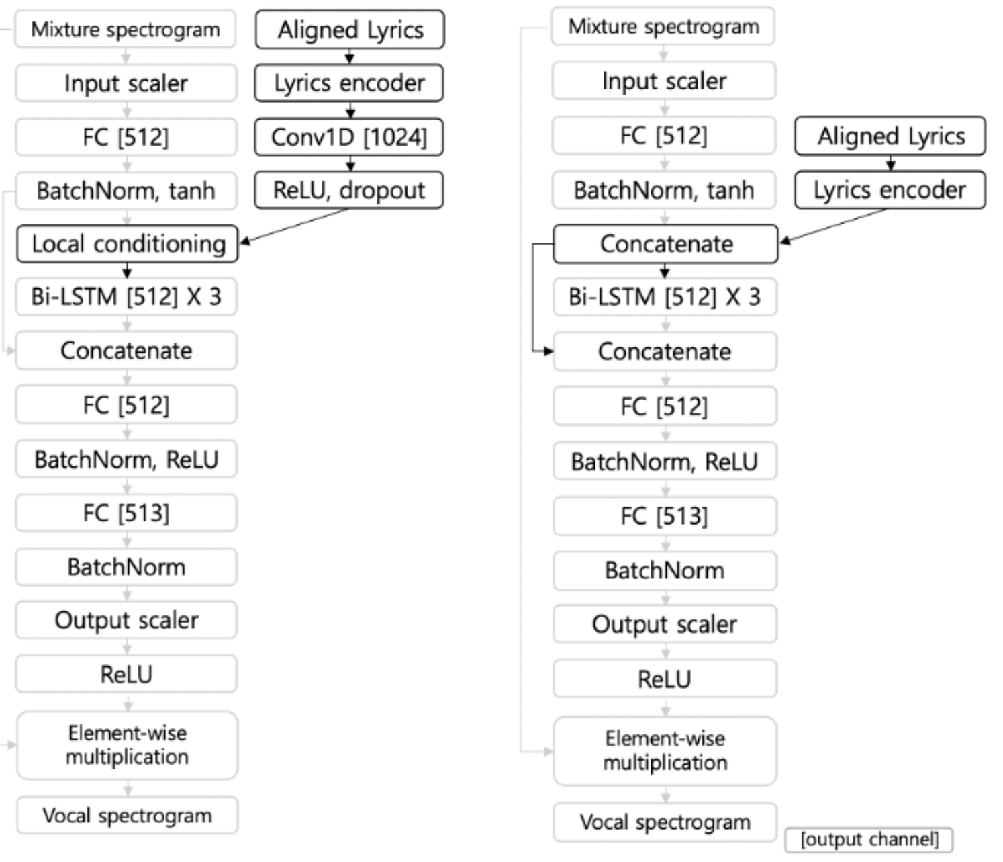


Aligned Lyrics
Embedding [128]
Conv1D [512, 1, 1]
ReLU, dropout
Highway [512, 3, 1]
Highway [512, 3, 3]
Highway [512, 3, 9]
Highway [512, 3, 27]
X 2
Highway [512, 3, 1]
X 2
Lyrics Encoding
[output channel, filter size, dilation rate]

· As a lyrics encoder, we used highway network-based model, which was also used in TTS and singing synthesis network.

· Like voice synthesis model, we assumed that it would be also capable of modelling the phonetic information in singing voice separation task.

Model



Mixture spectrogram
Input scaler
FC [512]
BatchNorm, tanh
Local conditioning
Bi-LSTM [512] X 3
Concatenate
FC [512]
BatchNorm, ReLU
FC [513]
BatchNorm
Output scaler
ReLU
Element-wise multiplication
Vocal spectrogram

Aligned Lyrics
Lyrics encoder
Conv1D [1024]
ReLU, dropout

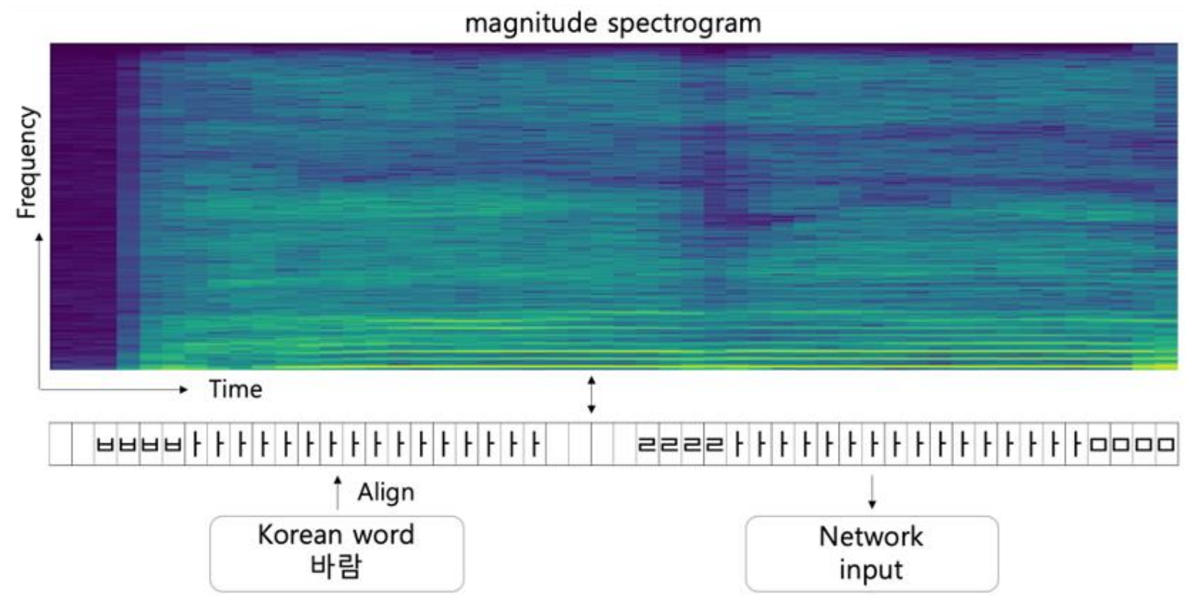
Mixture spectrogram
Input scaler
FC [512]
BatchNorm, tanh
Concatenate
Bi-LSTM [512] X 3
Concatenate
FC [512]
BatchNorm, ReLU
FC [513]
BatchNorm
Output scaler
ReLU
Element-wise multiplication
Vocal spectrogram

Aligned Lyrics
Lyrics encoder

[output channel]

- · Lyrics encoder + *Open-unmix*
- · We used two different conditioning methods for experiments
  - Local conditioning
  - Concatenation

Dataset



- · 201 Korean songs dataset
  - train -> 162 songs (9h)
  - validation -> 19 songs (1h)
  - test -> 20 songs (1h)
- · Manually annotated lyrics
- · Training with 19,113 instrumental tracks

Experiments

Model name	Inputs to the lyrics encoder
model 1	None
model 2	Meaningless inputs (all 0)
model 3	Vocal activity information
model 4	Aligned lyrics

* *LC* – Local conditioning

* *CC* - Concatenation

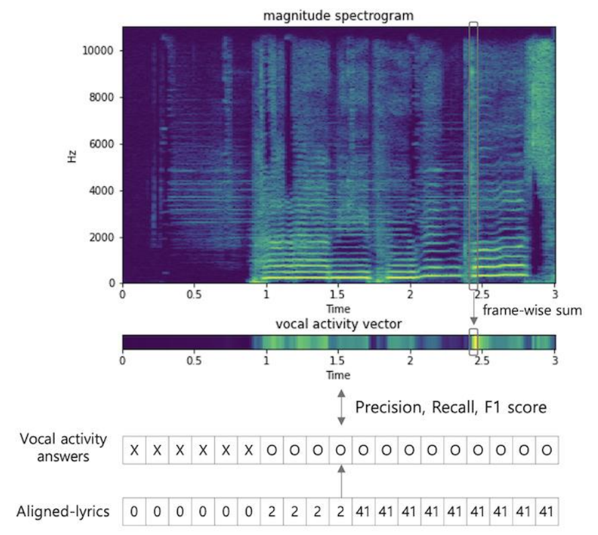
- · We trained total 7 models for the experiments.
- · *model 1*
  - baseline Open-unmix
- · *model 2 -> {LC, CC}*
  - For checking the effect of increased network capacity
- · *model 3 -> {LC, CC}*
  - For checking the effect of vocal activity information
- · *model 4 -> {LC, CC}*
  - With aligned lyrics, our proposed model

Experiments – Performance Evaluation

Models	Median			Mean		
	SDR	SIR	SAR	SDR	SIR	SAR
model 1	9.956	18.674	9.847	8.595	16.062	9.145
LC-model 2	10.140	18.465	9.766	8.589	16.001	9.093
LC-model 3	10.090	18.713	9.763	9.250	16.298	9.153
LC-model 4	**10.767**	19.505	10.223	9.723	17.116	9.699
CC-model 2	10.110	18.434	9.909	8.691	16.164	9.207
CC-model 3	10.444	19.328	10.169	9.718	17.031	9.609
CC-model 4	10.757	**19.623**	**10.371**	**9.752**	**17.250**	**9.803**

- · Both *LC* and *CC* were effective
- · Phonetic information is helpful for performance gain -> SDR 0.8dB ↑

Experiments – Analysis of vocal activity usage



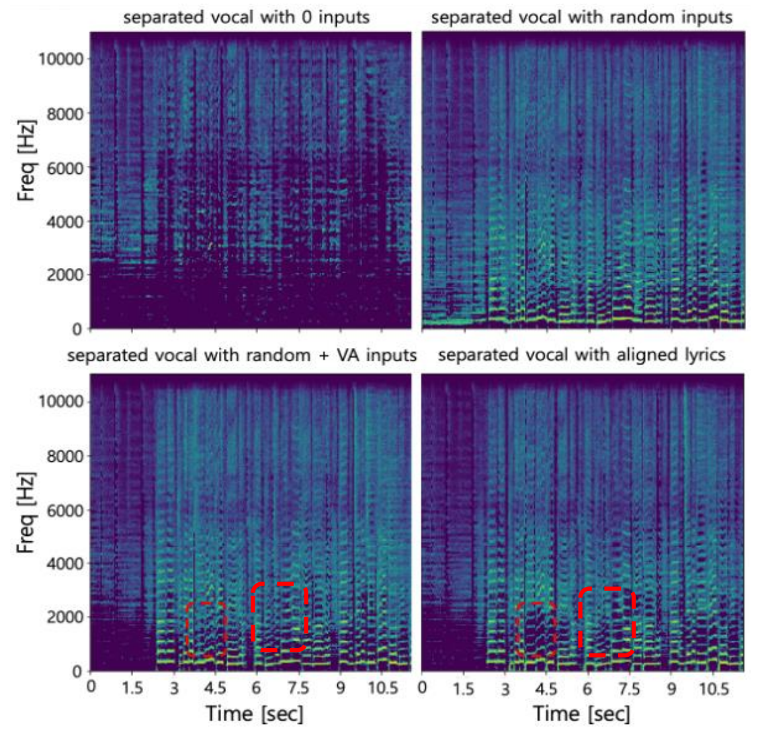
* *VA* – Vocal Activity

Models	Precision	Recall	F1 score
model 1	0.807	0.853	0.828
LC-model 2	0.810	0.852	0.830
LC-model 3	0.887	**0.857**	0.872
LC-model 4	0.876	0.854	0.865
CC-model 2	0.814	0.853	0.833
CC-model 3	**0.896**	0.855	**0.875**
CC-model 4	0.879	0.855	0.867

- · Both *VA* and phonetic information are important
- · Can the model leverage *VA* information well?
  - Yes. *model 4* can reflect *VA* information as well as *model 3.*
- · We extracted the *VA* from the separated vocal and compared it with ground truth *VA.*

Experiments – Analysis with using incorrect lyrics

Models	Inputs	SDR	SIR	SAR
LC-model 4	Zero	0.001	8.040	-4.286
	Random	5.317	15.802	4.845
	VA+Random	7.899	19.270	6.403
	AlignedLyrics	**10.767**	19.505	10.223
CC-model 4	Zero	0.002	7.246	-3.671
	Random	0.946	14.837	0.341
	VA+Random	7.164	19.545	6.290
	AlignedLyrics	10.757	**19.623**	**10.371**



- · To check if *model 4* effectively uses the information in lyrics, we gave the model incorrect lyrics
- · Four different lyrics inputs
  - Zero : 0 value (means silence) in training
  - Random : random lyrics
  - *VA* + Random : random lyrics with *VA* information
  - AlignedLyrics : Correct lyrics (proposed)
- · Critical performance degradation occurred
- · Dashed line shows the enhanced parts when the correct aligned lyrics are used

Conclusions

- · We proposed an integrated framework of combining the lyrics encoder into the *Open-unmix* separation network.
- · By various experiments, we confirmed that the phonetic information of lyrics is helpful for the singing voice separation network.
- · We are planning to use the un-aligned lyrics for the future work.