Transcribing Lyrics From Commercial Song Audio: The First Step Towards Singing Content Processing

Che-Ping Tsai*, Yi-Lin Tuan* and Lin-shan Lee
National Taiwan University

1. Introduction

- **Singing content processing not yet considered**
  - spoken content can now be successfully retrieved, browsed, summarized and comprehended.
  - ex: songs may be similarly retrieved based on lyrics in addition to melody
- **Songs are human voice carrying plenty of semantics just as speech**
  - core information in lyrics
  - with much more flexible prosody (pitch, duration, pauses, energy)
  - transcribing lyrics is a much more difficult version of automatic speech recognition(ASR)
- **A data set of English commercial songs by original professional singers**
  - closer to the goal towards singing content

2. Data

Acoustic Corpus

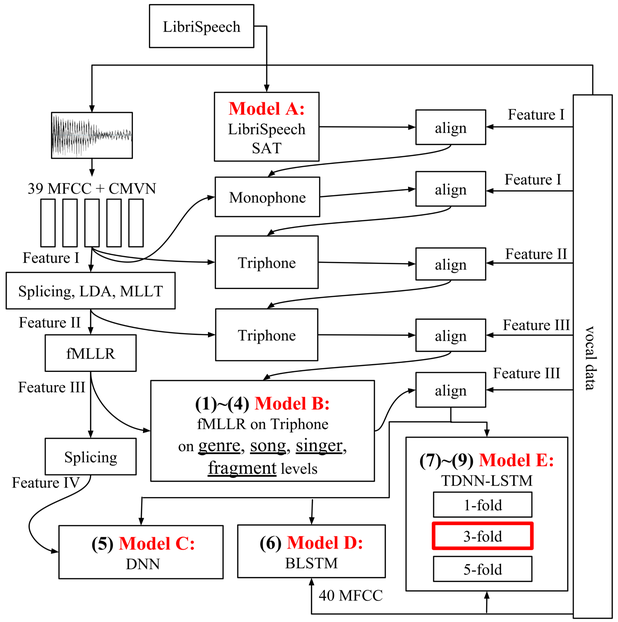
- **Vocal-only English commercial songs** from YouTube.
- **Segmented to fragments** from 10s to 35s, primarily based on silence.

	# Songs	# Singers	# Frag.	Dur. (min)
Training Set	95	49	640	271
Testing Set	15	13	97	42.8

Linguistic Corpus

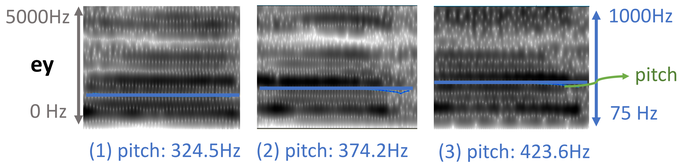
- LibriSpeech : 803M words / 40M sentences
- Lyrics : 129.8M words / 574k pieces

4. System Structure

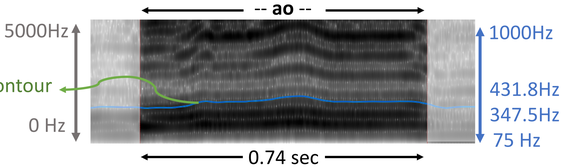


3.Difficulties of Recognition

- Word repetition (e.g. oh oh oh) and meaning less word (e.g. oh)
- Highly flexible pitch contours with much wider range
- Different acoustic characteristics of the same phoneme at **different pitch levels**



- **Prolonged** phoneme duration with **varying pitch**



5. Proposed Approaches and Experiments

	Acoustic Model		WER(%)
HMM-GMM	(baseline) **Model A :** LibriSpeech		96.21
	Model B: fragment-level	(1)Training data: Vocal -> Speech	88.26
		(2) + Lyric LM	80.40
		(3) + Ext Lex	77.08
		(4) + Self-loop prob.	76.62
Deep Learning	(5) **Model C :** DNN		75.56
	(6) **Model D :** BLSTM (3-fold)		74.32
	Model E : TDNN-LSTM	(7) 1-fold	79.01
		(8) 3-fold	**73.90**
		(9) 5-fold	74.53
	(13) fMLLR adaptation on fragment-level		
HMM-GMM	**Model B**	(10) *genre-level	84.24
		(11) singer-level	78.53
		(12) song-level	78.80
		(13) fragment-level	77.08

*genre : pop, electronic, rock, hiphop, R&B

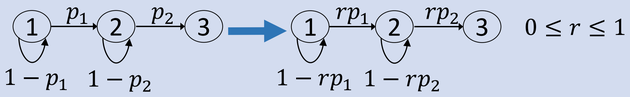
(1) Training data : speech -> vocal

(2) Lyric language model

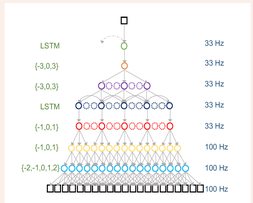
(3) Extended lexicon (vowel can be repeated or not)

Apple : ae-p-ah-l			
ae repeated		ae not repeated	
ah repeated	ah not repeated	ah repeated	ah not repeated
ae-ae-p-ah-ah-l	ae-ae-p-ah-l	ae-p-ah-ah-l	ae-p-ah-l

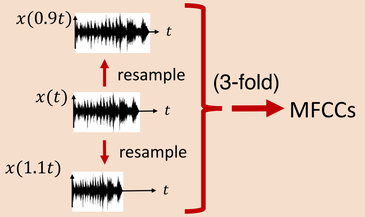
(4) Increased self-loop probability (for vowels HMM)



(7)(8)(9) TDNN-LSTM



(8)(9) 3-fold : resampled expand / compressed in time for speed perturbation at ratio 0.9,1,1.1



6.Conclusion

- 3-fold TDNN-LSTM is the best model.
- The achieved WER was relatively high compared to experiences in speech recognition.
- The results may be better with more training data.