Automatic Lyric Transcription from Karaoke Vocal Tracks: Resources and a Baseline System

Gerardo Roa Dabike, Jon Barker

{groadabike1, j.p.barker}@sheffield.ac.uk

Department of Computer Science, The University of Sheffield

Motivation

- Sung speech recognition is a challenging task.
  - The intelligibility is secondary to the musical quality.
  - Large range in pitch and loudness.
  - Large variability in vocal style.
- ... but little previous research [1, 2, 3]
  - Lack of readily available data.
  - No commonly agreed evaluation framework.

The Karaoke Dataset

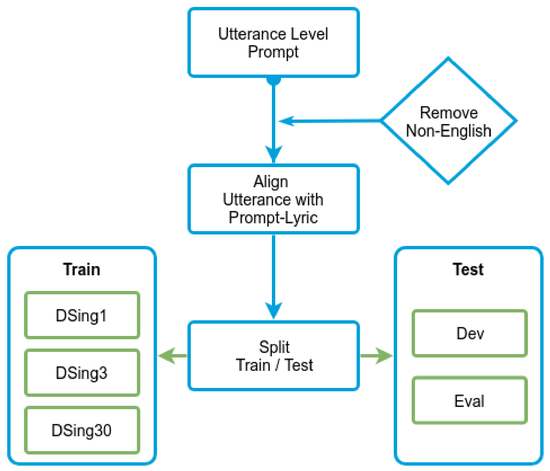
Smule Sing!300x30x2 Dataset (Sing!) [3]

- Two most popular performances (1 male, 1 female), from the 300 most popular arrangements, from 30 countries.
- Total 18,767 multilingual performances, 13,154 singers and 5,690 songs.
- Contains prompt-lyric and prompt-timing per arrangement.

Data Pre-processing

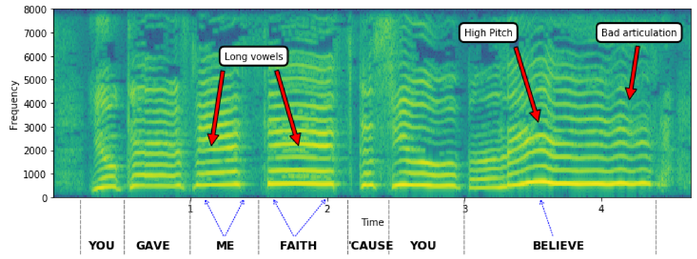
Transform Sing! into an ASR-oriented dataset.

- Generate utterance level prompt.
- Remove non-English songs.
- Segment utterance and align to prompt-lyric.
- Split Train-Test sets.



Sung Speech Challenges

Acoustic Modelling



- Other challenges [1, 2, 3]
  - Vibrato and falsetto.
  - High pitch range.
  - Variation of tempo, dynamics and singing styles.
  - ...

Language Modelling

- Poetic vs prosaic.
- Metaphorical rather literal.
- Word repetition,
  e.g. Karma Chameleon - Culture Club
  ...Karma, karma, karma, karma, karma chameleon...
- Sentence Repetition,
  e.g. All You Need Is Love - The Beatles
  ..All you need is love...(more than 20 times)
- Low semantic predictability.
- Large variation in lyric style.
- Vowel repetition to match with vowel extension in singing,
  e.g. Love Can Move Mountains - Celine Dion
  Faith, Trust, Looooooove...
- Meaningless words and phrases,
  e.g. Pinhead - The Ramones
  ...Gabba gabba hey!...

DSing ASR Task

- Training Datasets
  - **DSing1**, English recordings from users located in GB.
  - **DSing3**, English recordings from users located in GB, Australia and USA.
  - **DSing30**, English recordings from users from all 30 countries.

Table 1: Description of the DSing training sets.

Set	Singers	Songs	Utterances	Hours
DSing1	352	434	8,794	15.1
DSing3	1,050	1,343	25,526	44.7
DSing30	3,205	4,324	81,092	149.1

- Development and Evaluation Dataset.
  - English language recordings from users located in GB.
  - Manually corrected.
    - Endponting, e.g., errors in alignment.
    - Transcriptions, e.g., mis-read lyrics.

Table 2: Description of the hand-corrected development and evaluation sets.

Set	Singers	Songs	Utterances	Hours
dev	40	66	482	0.7
eval	43	70	480	0.8

- The training and test sets are disjoint with respect to singers and songs.

Baseline ASR

System built using Kaldi [5].

- Features: 40 MFCC + iVectors.
- Acoustic Model: TDNN-F.
- Language Model: 3-gram/4-gram MaxEnt based on 44,287 song lyrics (from lyrics.fandom.com) and 28K vocabulary.

Table 3: WER results per training set.

Train Set	AM	LM	dev	eval
DSing1	TDNN-F	3-gram	46.0	42.3
	TDNN-F	4-gram	**41.2**	**37.6**
DSing3	TDNN-F	3-gram	33.0	28.7
	TDNN-F	4-gram	**29.6**	**24.3**
DSing30	TDNN-F	3-gram	26.2	22.3
	TDNN-F	4-gram	**23.3**	**19.6**

Summary

1. Constructed DSing ASR task from the Sing! dataset.
2. TDNN-F Acoustic Model and in-domain Language Model.
3. 19.6% WER, new state of the art for sung speech recognition.
4. Best system is presented as a baseline to the community.
5. **Code and system freely available.**

References and Code

[1] Mesaros, A. et al. (2010). Automatic Recognition of Lyrics in Singing. EURASIP Journal on Audio, Speech, and Music Processing, 2010, 1-11.
[2] Kruspe, A. M. (2016). Retrieval of textual song lyrics from sung inputs. INTERSPEECH 2016.
[3] Tsai, C. et al. (2018). Transcribing Lyrics from Commercial Song Audio: The First Step Towards Singing Content Processing. ICASSP 2018.
[4] Smule Sing! 300x30x2 Dataset, https://ccrma.stanford.edu/damp, accessed September, 2018.
[5] Povey D. et al. (2011) *The Kaldi Speech Recognition Toolkit*. ASRU 2011.

SCAN ME

https://l.ead.me/DSingtask