The University Of Sheffield.

SPandH

The Use of Voice Source Features for Sung Speech Recognition

Gerardo Roa Dabike, Jon Barker

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

Department of Computer Science, The University of Sheffield

3256

Why Sung Speech Recognition?

Music Information Retrieval applications

- Retrieving lyrics by recognising segments.
- Retrieving songs information by query-by-singing.
- Indexing databases by lyrics keywords.
- Lyrics alignment for Karaoke applications.

Less intelligible type of speech.

- Get insights of how to adapt speech technologies.
- Other less intelligible speech:
  - Dysarthric speech.
  - Casual speech

Research Question

1. What are the key differences between the voice source properties of spoken and sung speech?
2. What implications do these differences have for sung speech recognition?
3. To what extent can useful ASR features be extracted from the voice source?

DSing Corpus

Training Datasets.

Set	Singers	Songs	Utt	Hrs
DSing1 (GB)	352	434	8,794	15.1
DSing3 (GB-AU-US)	1,050	1,343	25,526	44.7
DSing30 (30 ctry)	3,205	4,324	81,092	149.1

Development and Evaluation Datasets.

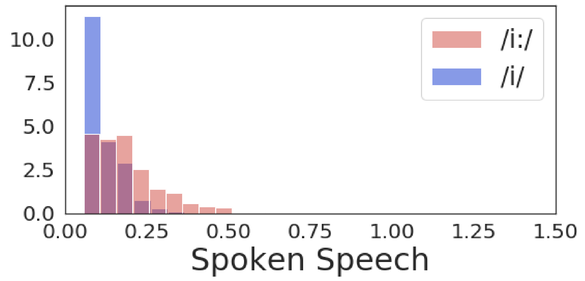
- English language recordings from users located in GB.
- Manually corrected.
  - Endponting, e.g., errors in alignment.
  - Transcriptions, e.g., mis-read lyrics.

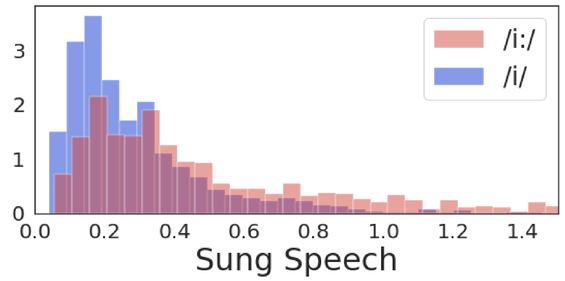
Set	Singers	Songs	Utt	Hrs
dev	40	66	482	0.7
eval	43	70	480	0.8

Sung vs Spoken Speech

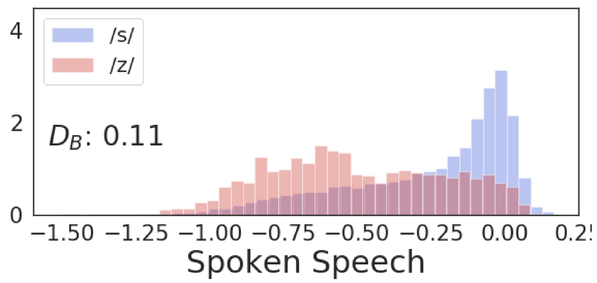
Analysis made using the NUS-48E sung and spoken lyrics.

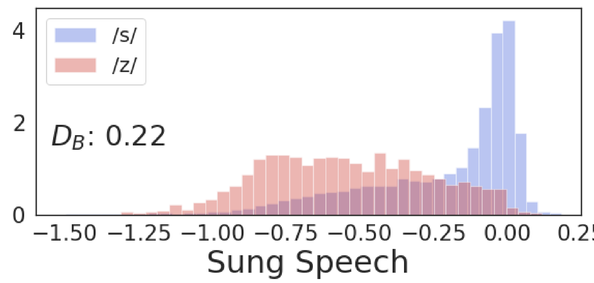
Vowel Duration.



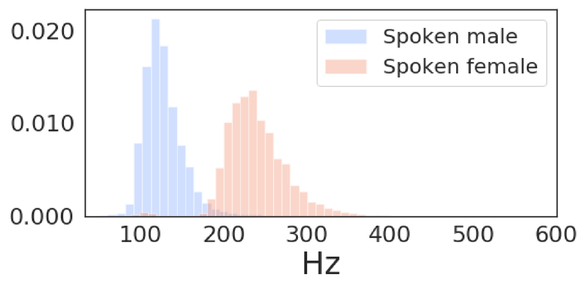


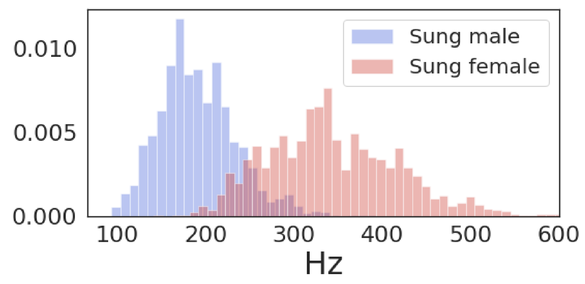
Degree of Voicing.



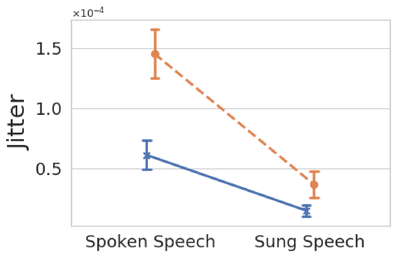


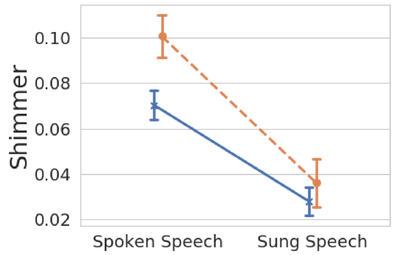
Pitch Range.

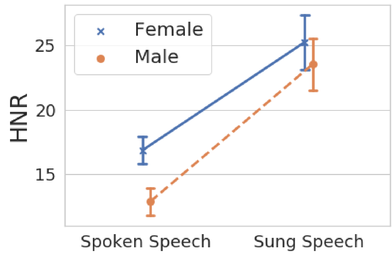




Voice Quality.







Baseline ASR

System built using Kaldi.

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

Methodology

Train new AM by expanding the 40 MFCC feature vector.

- Experiment +Pitch.
  - Expand MFCC feature with pitch and degree of voicing.
- Experiment +VQ.
  - Expand MFCC with pitch, voicing degree and the three VQ features (jitter, shimmer and HNR).

Results

Table 1: Average WER results per training set and experiments.

Train	LM	Baseline	+Pitch	+VQ
DSing1	3-gram	43.02	40.99	41.17
	4-gram	38.14	36.77	36.70
DSing3	3-gram	28.13	28.05	27.82
	4-gram	24.40	24.27	23.76
DSing30	3-gram	22.82	23.23	22.97
	4-gram	19.88	19.87	19.60

Summary

- Sung Speech less intelligible.
- Phonemes less discriminable.
  - Smaller inter-class separation - vowel duration, degree of voicing.
  - Larger intra-class spread - pitch range
- Voice Source Features for ASR.
  - Pitch and voicing degree more helpful in small resources dataset.
  - May be learned by models when using enough data.