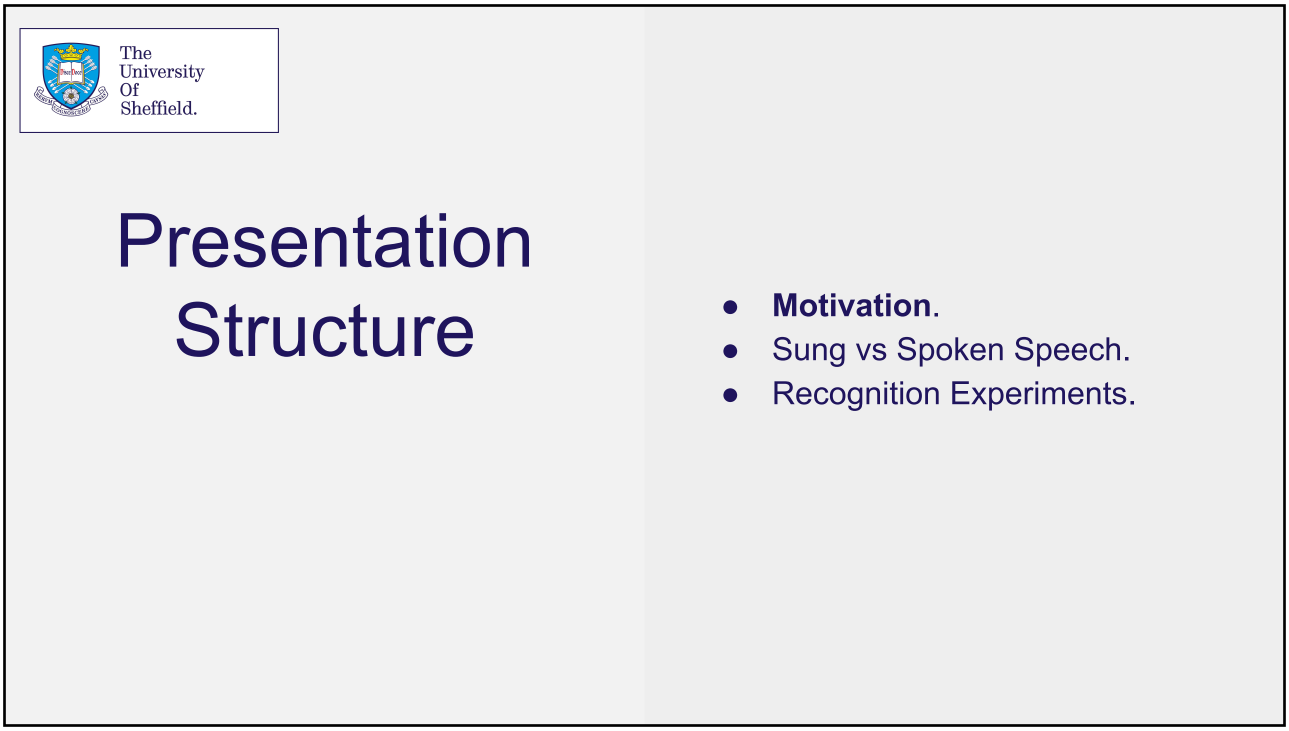
The Use of Voice Source Features for Sung Speech Recognition

Authors: Gerardo Roa Dabike
Jon Barker

Why Sung Speech Recognition?

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

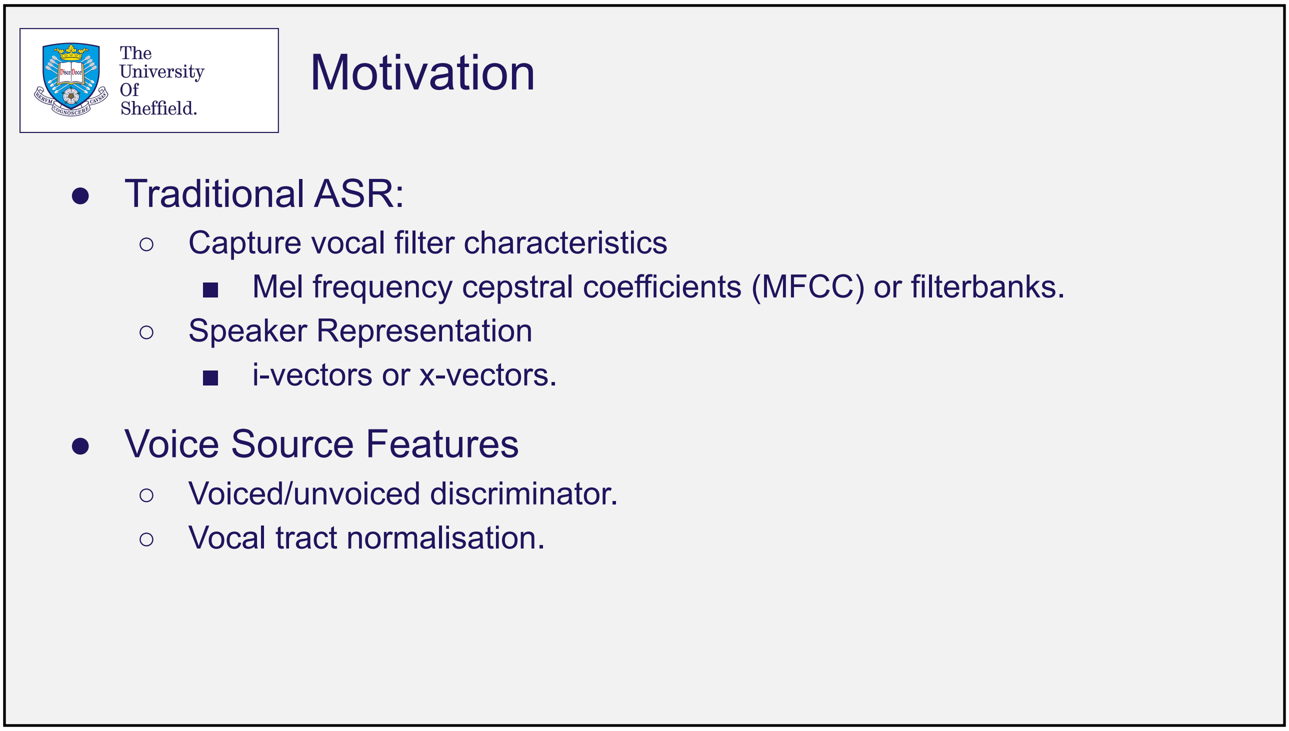
- Less intelligible type of speech.
  - Get insights of how to adapt speech technologies.
  - Other atypical speech (i.e., dysarthric)



The
University
Of
Sheffield.

Presentation Structure

- **Motivation**.
- Sung vs Spoken Speech.
- Recognition Experiments.



The
University
Of
Sheffield.

Motivation

- Traditional ASR:
  - Capture vocal filter characteristics
    - Mel frequency cepstral coefficients (MFCC) or filterbanks.
  - Speaker Representation
    - i-vectors or x-vectors.
- Voice Source Features
  - Voiced/unvoiced discriminator.
  - Vocal tract normalisation.

The University Of Sheffield.

Presentation Structure

- Motivation.
- **Sung vs Spoken Speech.**
- Recognition Experiments.

The University Of Sheffield.

NUS-48E Sung and Spoken Lyrics Corpus [1]

- Sung and Spoken Lyrics.
- 12 speakers.
  - 6 females (3 soprano, 3 alto).
  - 6 males (2 tenor, 3 baritone, 1 bass).
- 48 English songs (20 unique).
- 25,474 phone instances.
- American and Singapore accents.
  - Accent variation is neutralised when singing [2].
  - Move towards American pronunciation [3].

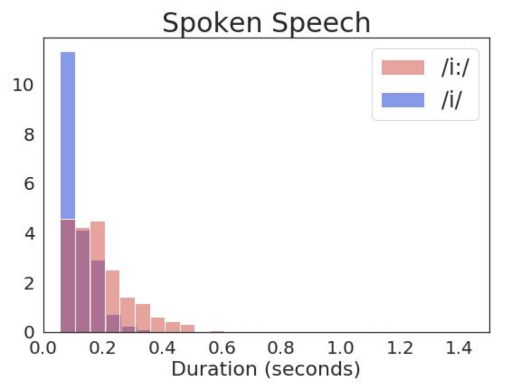
[1] Z. Duan, H. Fang, B. Li, K. C. Sim, and Y. Wang, “The NUS sung and spoken lyrics corpus: A quantitative comparison of singing and speech,” in Proc. APSIPA ASC, 2013

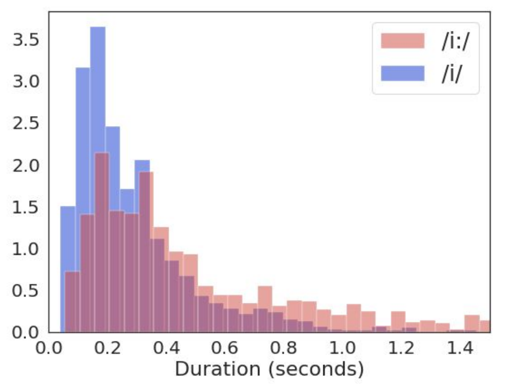
[2] A. Gibson, “Production and perception of vowels in New Zealand popular music,” MPhil Thesis, Auckland University, NewZealand, 2010.

[3] M. Konert-Panek, “Overshooting americanisation. accent stylisation in pop singing acoustic properties of the bath and trap vowels in focus,”Research in Language, vol. 15, pp. 371–384, 12 2017

Sung vs Spoken Speech

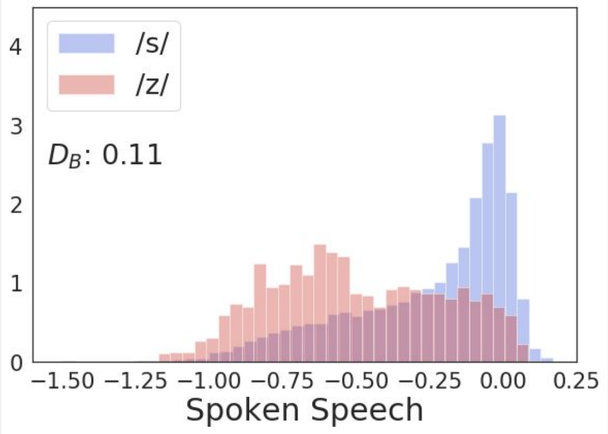
Vowel Duration Bit (short vowel) vs Beat (long vowel)

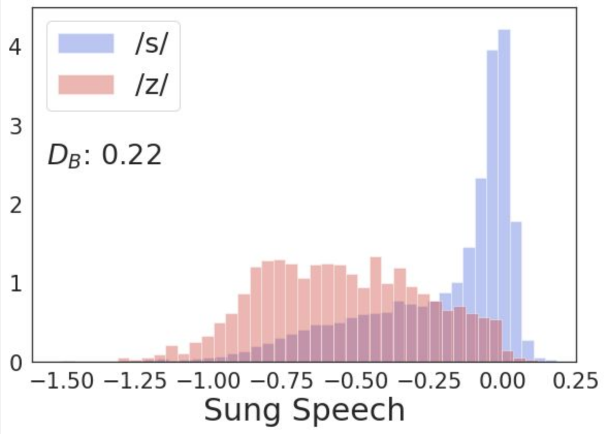




Sung vs Spoken Speech

Degree of Voicing Price (voiceless **/s/**) vs Prize (voiced **/z/**)

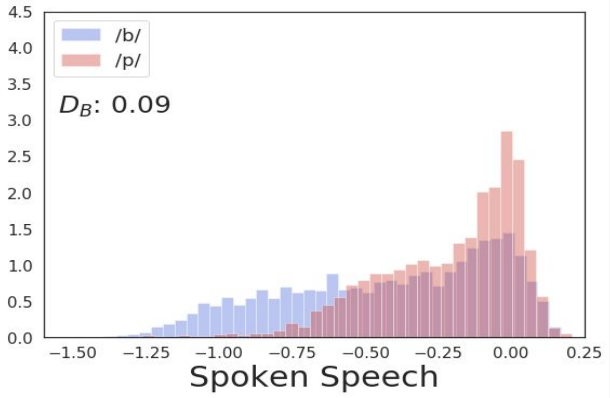


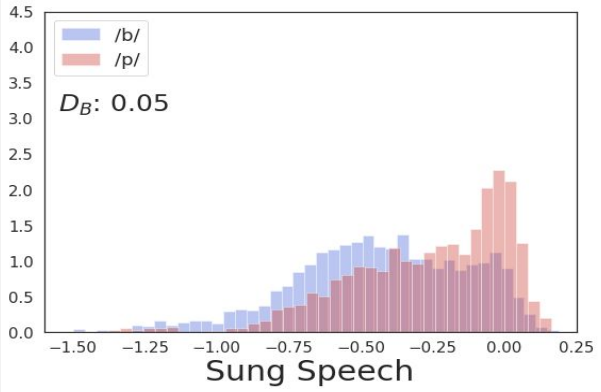


The University Of Sheffield.

Sung vs Spoken Speech

Degree of Voicing Peak (voiceless **/p/**) vs Beak (voiced **/b/**)

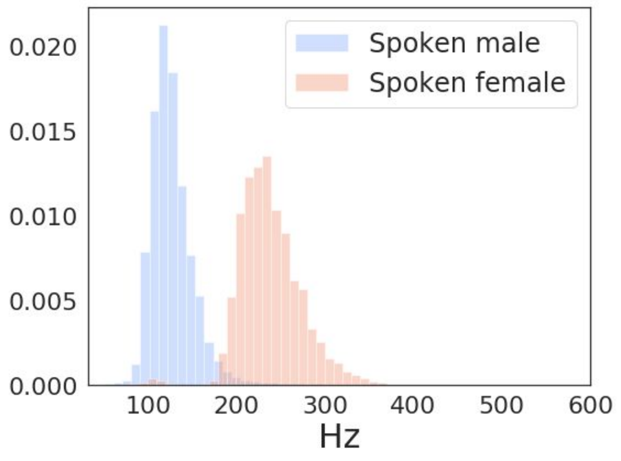


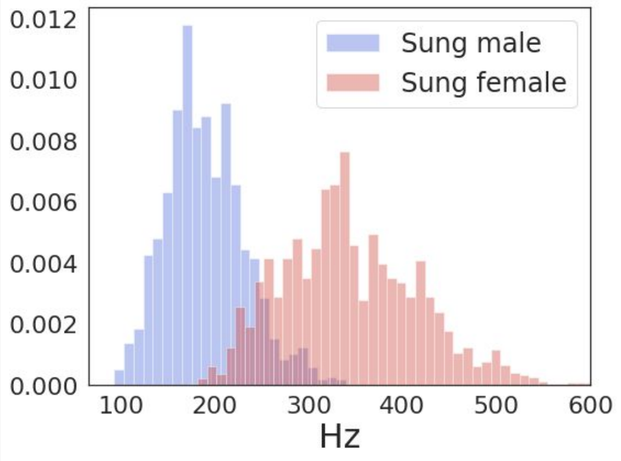


The University Of Sheffield.

Sung vs Spoken Speech

Pitch Range

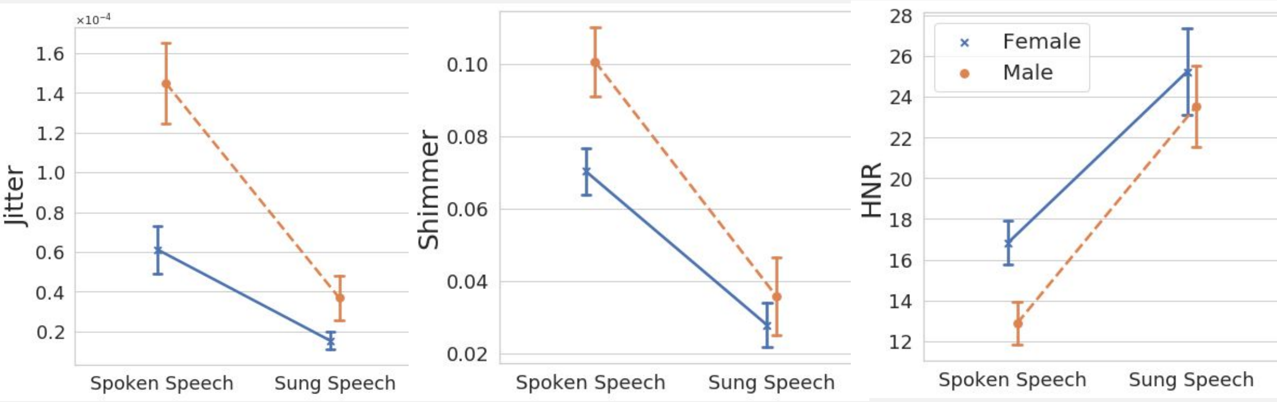




The University Of Sheffield.

Sung vs Spoken Speech

Voice Quality Jitter - Shimmer - HNR



The University Of Sheffield.

Presentation Structure

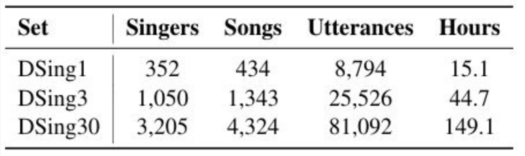
- Motivation.
- Sung vs Spoken Speech.
- **Recognition Experiments.**

The University Of Sheffield.

Speech Sung Corpus

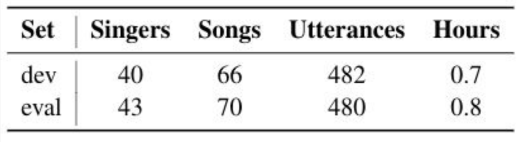
DSing sung speech dataset [1] based on the Smule Sing! 300x30x2 corpus [2].

- Train sets:
  - DSing1 (GB).
  - DSing3 (GB + AU + US).
  - DSing30 (All 30 countries).



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

- Test sets:
  - Development (GB).
  - Evaluation (GB).



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

[1] Roa Dabike, G and Barker, J. “Automatic lyric transcription from Karaoke vocal tracks: Resources and a Baseline System”. Interspeech. 2019

[2] Smule Sing!300x30x2 Dataset, “https://ccrma.standford.edu/damp/”, accessed September 2018.

The University Of Sheffield.

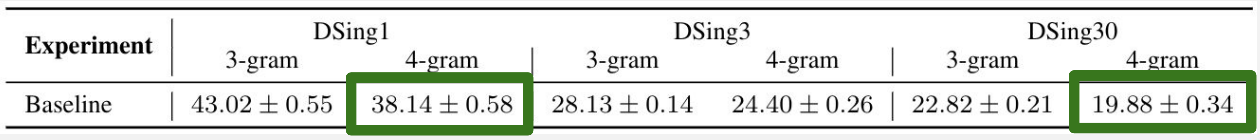
Baseline

Language Model

- Lyrics
  - Artist in DSing3.
  - Billboard’s ‘The Hot 100’
  - 44,287 lyrics.
  - 456 artist.
  - 28K vocabulary.
- 3-gram - 4-gram MaxEnt

Acoustic Model

- Features
  - 40 MFCC + Delta + Delta-Delta
  - 100 i-vectors.
- Factorised TDNN model.
  - Lattice-free MMI.
  - 3 frames of context.



Experiment	DSing1		DSing3		DSing30	
	3-gram	4-gram	3-gram	4-gram	3-gram	4-gram
Baseline	43.02 ± 0.55	38.14 ± 0.58	28.13 ± 0.14	24.40 ± 0.26	22.82 ± 0.21	19.88 ± 0.34

Values express percentage of Word Error Rate (WER)

The University Of Sheffield.

Methodology

1. Experiment 1
   1. MFCC + i-vectors
   2. Pitch
   3. Degree of voicing

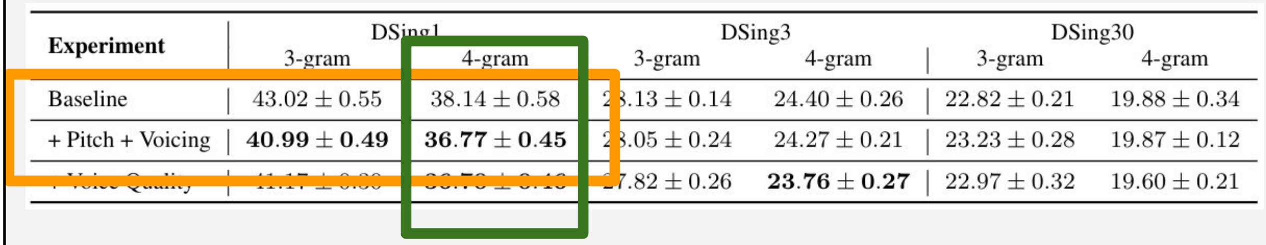
2. Experiment 2
   1. MFCC + i-vectors
   2. Pitch
   3. Degree of voicing
   4. Voice quality
      1. Jitter
      2. Shimmer
      3. HNR

The University Of Sheffield.

ASR Results

Evaluation measured in Word Error Rate

- Pitch + Voicing + VQ obtained best performance in DSing1 and DSing3 (*p*-value < 0.05).
- For the DSing30, no significant improvement was obtained.



Experiment	DSing1		DSing3		DSing30	
	3-gram	4-gram	3-gram	4-gram	3-gram	4-gram
Baseline	43.02 ± 0.55	38.14 ± 0.58	2[illegible].13 ± 0.14	24.40 ± 0.26	22.82 ± 0.21	19.88 ± 0.34
+ Pitch + Voicing	**40.99 ± 0.49**	**36.77 ± 0.45**	2[illegible].05 ± 0.24	24.27 ± 0.21	23.23 ± 0.28	19.87 ± 0.12
+ Voice Quality	41.17 ± 0.30	36.73 ± 0.46	[illegible]7.82 ± 0.26	**23.76 ± 0.27**	22.97 ± 0.32	19.60 ± 0.21

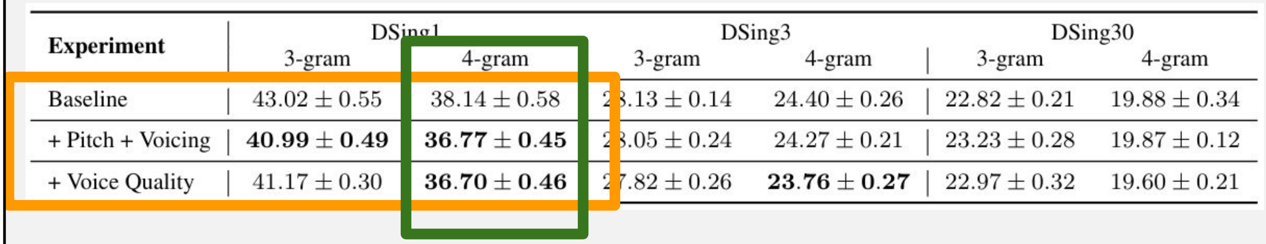
Voice Quality = Jitter, Shimmer and HNR

The University Of Sheffield.

ASR Results

Evaluation measured in Word Error Rate

- Pitch + Voicing + VQ obtained best performance in DSing1 and DSing3 (*p*-value < 0.05).
- For the DSing30, no significant improvement was obtained.



Experiment	DSing1		DSing3		DSing30	
	3-gram	4-gram	3-gram	4-gram	3-gram	4-gram
Baseline	43.02 ± 0.55	38.14 ± 0.58	28.13 ± 0.14	24.40 ± 0.26	22.82 ± 0.21	19.88 ± 0.34
+ Pitch + Voicing	**40.99 ± 0.49**	**36.77 ± 0.45**	28.05 ± 0.24	24.27 ± 0.21	23.23 ± 0.28	19.87 ± 0.12
+ Voice Quality	41.17 ± 0.30	**36.70 ± 0.46**	27.82 ± 0.26	**23.76 ± 0.27**	22.97 ± 0.32	19.60 ± 0.21

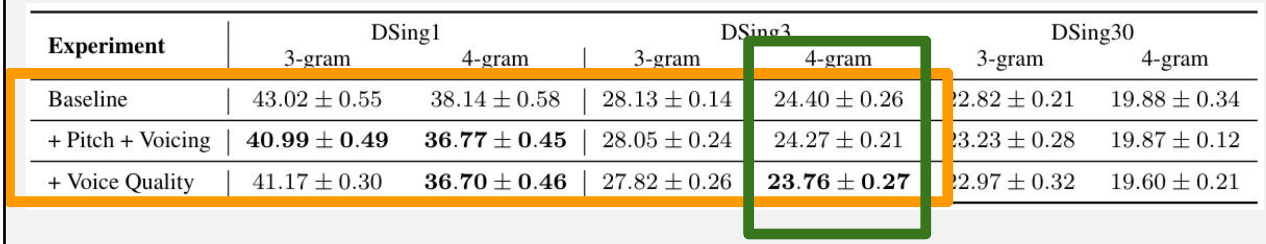
Voice Quality = Jitter, Shimmer and HNR

The University Of Sheffield.

ASR Results

Evaluation measured in Word Error Rate

- Pitch + Voicing + VQ obtained best performance in DSing1 and DSing3 (*p*-value < 0.05).
- For the DSing30, no significant improvement was obtained.



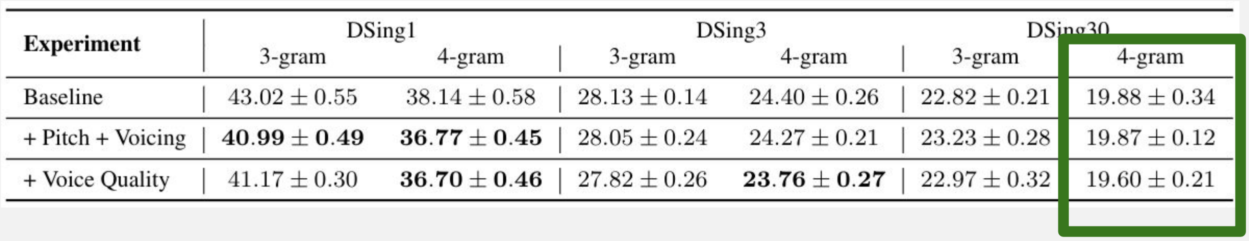
Experiment	DSing1		DSing3		DSing30	
	3-gram	4-gram	3-gram	4-gram	3-gram	4-gram
Baseline	43.02 ± 0.55	38.14 ± 0.58	28.13 ± 0.14	24.40 ± 0.26	22.82 ± 0.21	19.88 ± 0.34
+ Pitch + Voicing	**40.99 ± 0.49**	**36.77 ± 0.45**	28.05 ± 0.24	24.27 ± 0.21	23.23 ± 0.28	19.87 ± 0.12
+ Voice Quality	41.17 ± 0.30	**36.70 ± 0.46**	27.82 ± 0.26	**23.76 ± 0.27**	22.97 ± 0.32	19.60 ± 0.21

Voice Quality = Jitter, Shimmer and HNR

ASR Results

Evaluation measured in Word Error Rate

- Pitch + Voicing + VQ obtained best performance in DSing1 and DSing3 (*p*-value < 0.05).
- For the DSing30, no significant improvement was obtained.



Experiment	DSing1		DSing3		DSing30	
	3-gram	4-gram	3-gram	4-gram	3-gram	4-gram
Baseline	43.02 ± 0.55	38.14 ± 0.58	28.13 ± 0.14	24.40 ± 0.26	22.82 ± 0.21	19.88 ± 0.34
+ Pitch + Voicing	**40.99 ± 0.49**	**36.77 ± 0.45**	28.05 ± 0.24	24.27 ± 0.21	23.23 ± 0.28	19.87 ± 0.12
+ Voice Quality	41.17 ± 0.30	**36.70 ± 0.46**	27.82 ± 0.26	**23.76 ± 0.27**	22.97 ± 0.32	19.60 ± 0.21

Voice Quality = Jitter, Shimmer and HNR

Conclusions

- Sung Speech less intelligible:
  - Larger vowel duration.
  - Different voicing degree.
  - Larger pitch range.

- Voice Source Features:
  - Pitch and Voicing degree more helpful in small resources dataset.
  - May be learned by models when using enough data.

The
University
Of
Sheffield.

Thank you for Watching