L2-ARCTIC:

a non-native English speech corpus

Guanlong Zhao[1], Sinem Sonsaat[2], Alif Silpachai[2], Ivana Lucic[2], Evgeny Chukharev-Hudilainen[2], John Levis[2], and Ricardo Gutierrez-Osuna[1]

Presented by Christopher Liberatore[1]

[1]Department of Computer Science, Texas A&M University, U.S.
[2]Department of English, Iowa State University, U.S.

Disclaimer: this project is not affiliated with the CMU-ARCTIC project

Outline

Introduction

- Target tasks
  - Voice conversion (VC)
  - Accent conversion (AC)
  - Mispronunciation detection (MPD)
- Motivation

Corpus curation

- Design philosophy
- Data collection
- Data annotation

Corpus statistics

Usage examples

- Accent conversion
- MPD based on GOP

Conclusion

Target tasks

Voice conversion for non-native speakers

– Change speaker identity

Accent conversion

– Change speaker accent

Mispronunciation detection

– Detect segmental errors in non-native speech

Motivation

- Past voice conversion corpora focus on native speakers
  - CMU ARCTIC (Kominek & Black, 2004)
  - Voice Conversion Challenge dataset (Toda et al., 2016)
  - VCTK (Veaux et al., 2017)
- Limited non-native English resources
  - Noisy and limited data per speaker: Speech Accent Archive (Weinberger) and International Dialects of English Archive (Meier)
  - Restricted access: The Wildcat (Engen et al, 2010), LDC2007S08 (Lander), and NUFAESD (Bent & Bradlow, 2003)
- Limited open source resources for MPD
  - Restricted access: CU-CHLOE (Li et al., 2017) and College Learners' Spoken English Corpus (Yang & Wei, 2015)
  - Limited accents: ISLE Speech Corpus (Menzel et al., 2000) and SingaKids-Mandarin (Chen et al., 2016)

Design philosophy

- Multi-Dialects
  - Indian
  - Korean
  - Mandarin
  - Spanish
  - Arabic
  - And more…
- Enough data from each speaker
  - Phoneme inventory coverage
  - Pronunciation error elicitation
- Annotations
  - Clear and easy to process

Data collection

Participants

- First release (Apr. 2018): ten speakers from five dialects, one male and one female per dialect
- Second release (Sept. 2018): ten more speakers, two for each dialect
- Medium to high English proficiency

Recording

- Each speaker read 1132 sentences from CMU ARCTIC
- Recorded at ISU in a quite room with a linguist's guidance

Post-processing

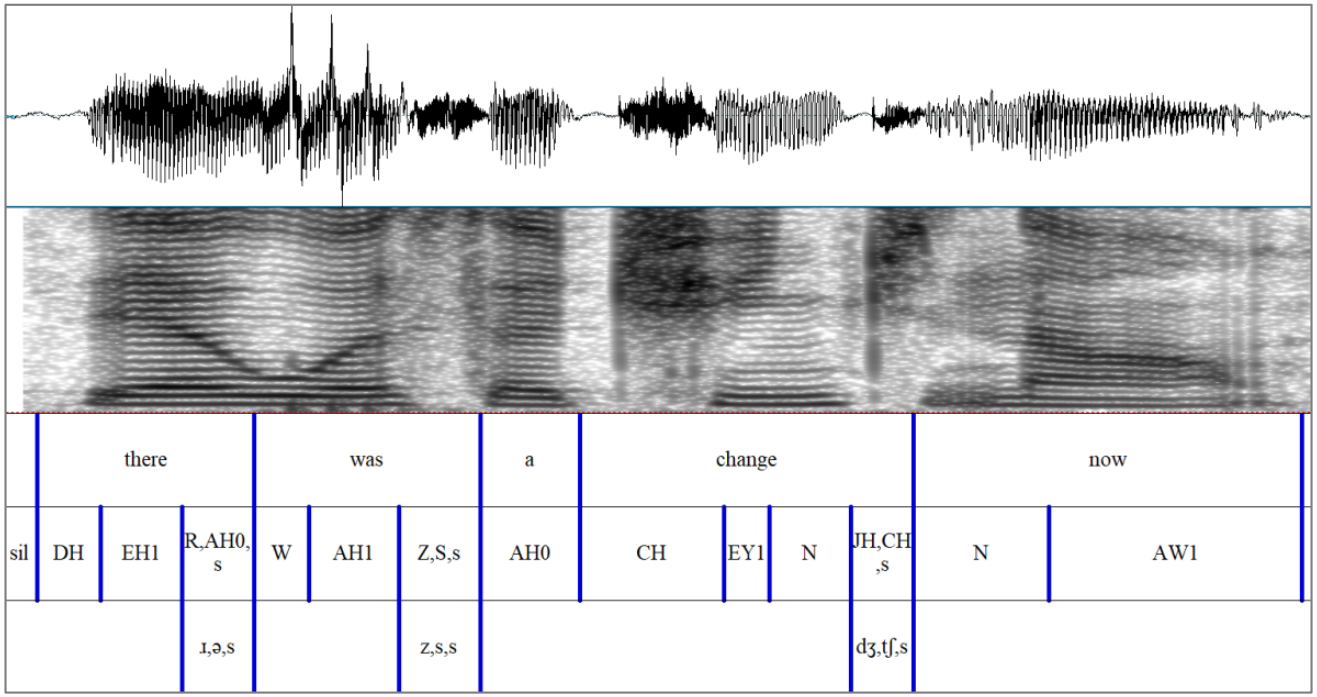
- Removed repetitions and false starts
- Carefully remove the leading and trailing silence and non-speech sounds

Annotations

Orthographic

Forced-alignment

Manual annotations



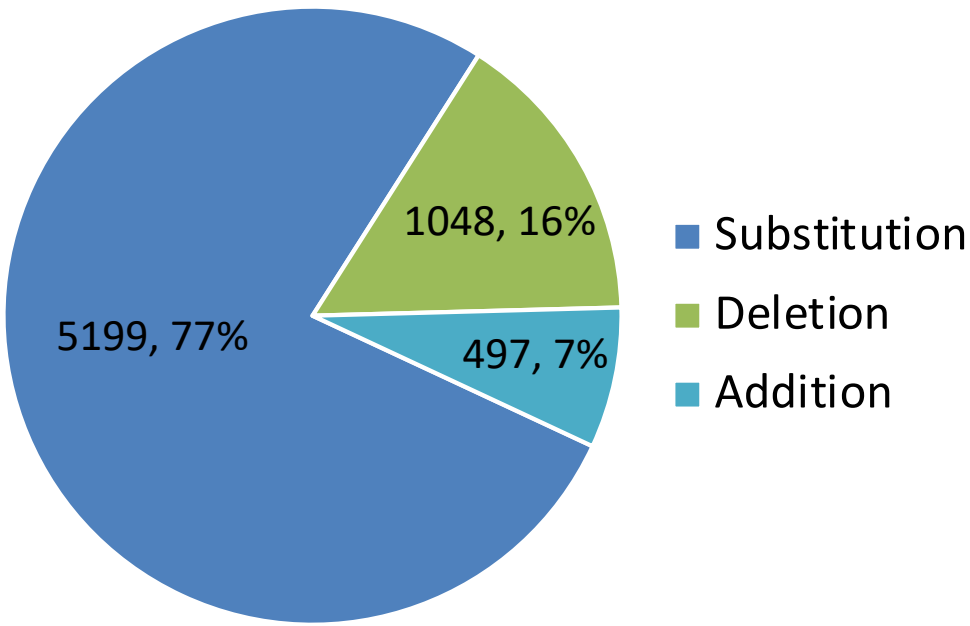
Corpus statistics: overview

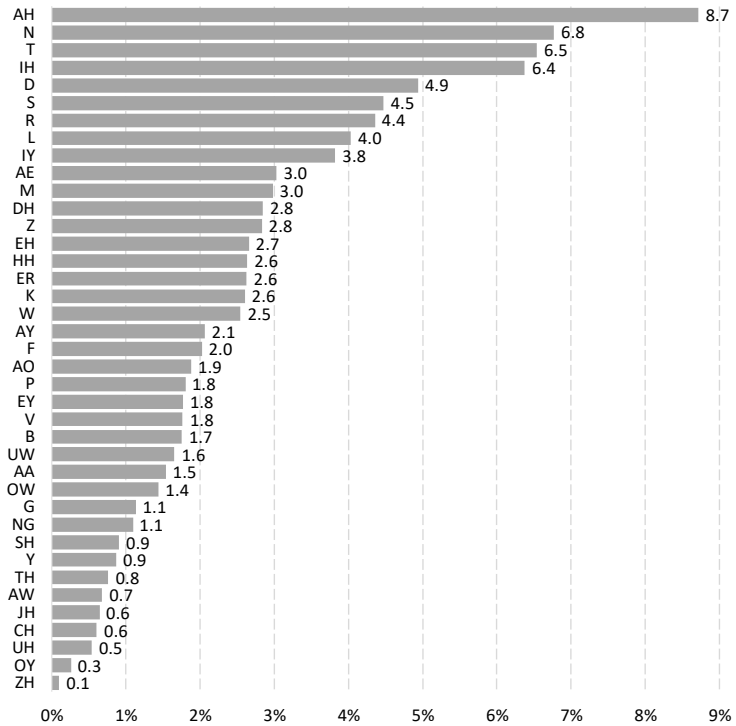
Speech data (first release)

- 11,026 utterances or 11.2 hours of speech
- Around nine words per sentence (~3.7s)

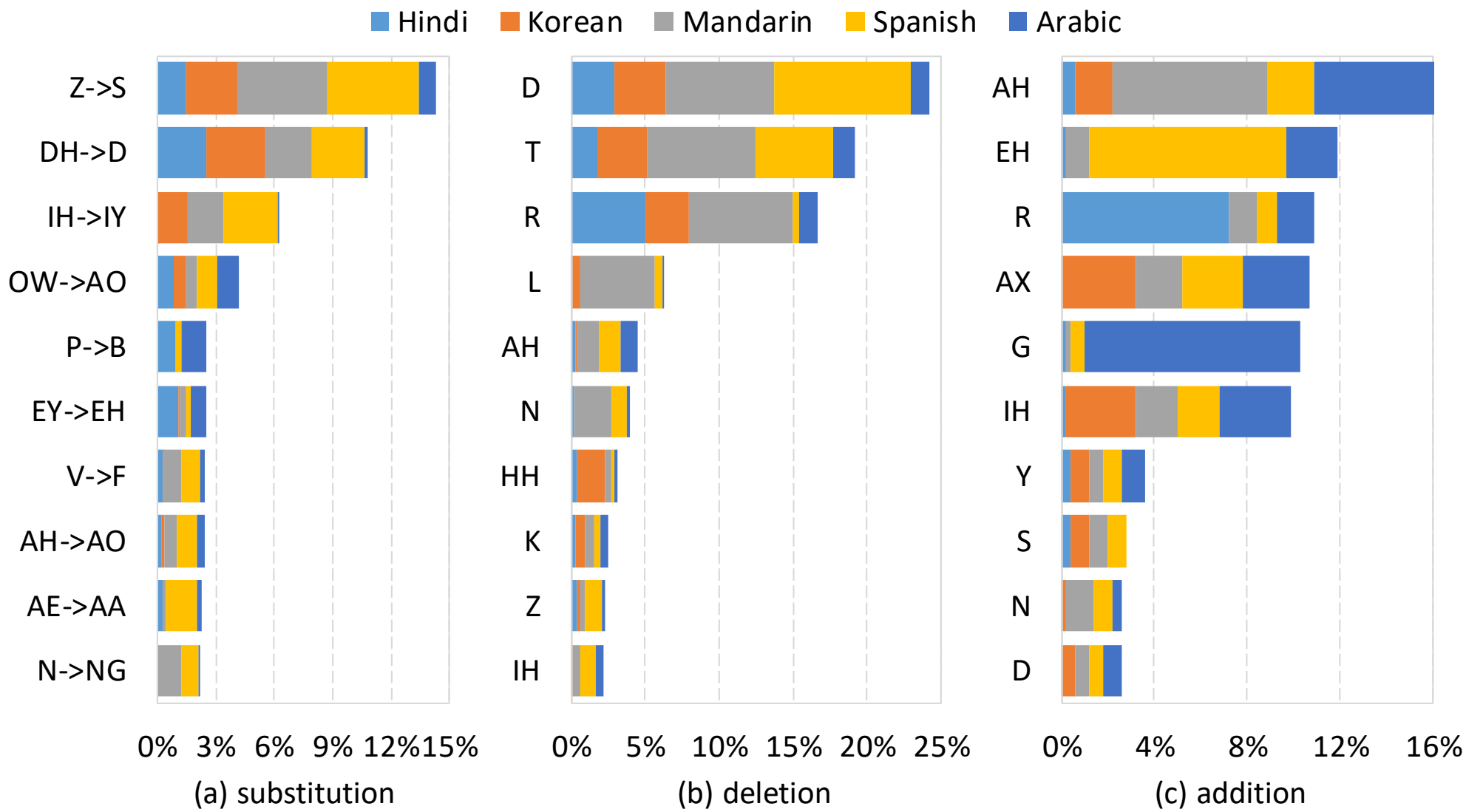
Manual annotation (first release)

- Annotated 1500 sentences (150 per speaker)





Corpus statistics: annotations (1)



Corpus statistics: annotations (2)

High frequency errors

L1	Substitutions	Deletions	Additions
Hindi	DH→D, Z→S, W→V EY→EH, TH→T	R, D, T ER, HH	R, AH, S, Y AA
Korean	DH→D, Z→S, IH→IY OW→AO, EH→AE	D, T, R HH, K	AX, IH, AH, S Y
Mandarin	Z→S, DH→D, IH→IY N→NG, V→F	D, T, R L, N	AH, AX, IH N, R
Spanish	Z→S, IH→IY, DH→D AE→AA, AH→AO	D, T, AH Z, IH	EH, AX, AH IH, IY
Arabic	P→B, OW→AO R→ERR, DH→Z, Z→S	T, R, D AH, IH	G, AH, IH AX, EH

Usage example: accent conversion

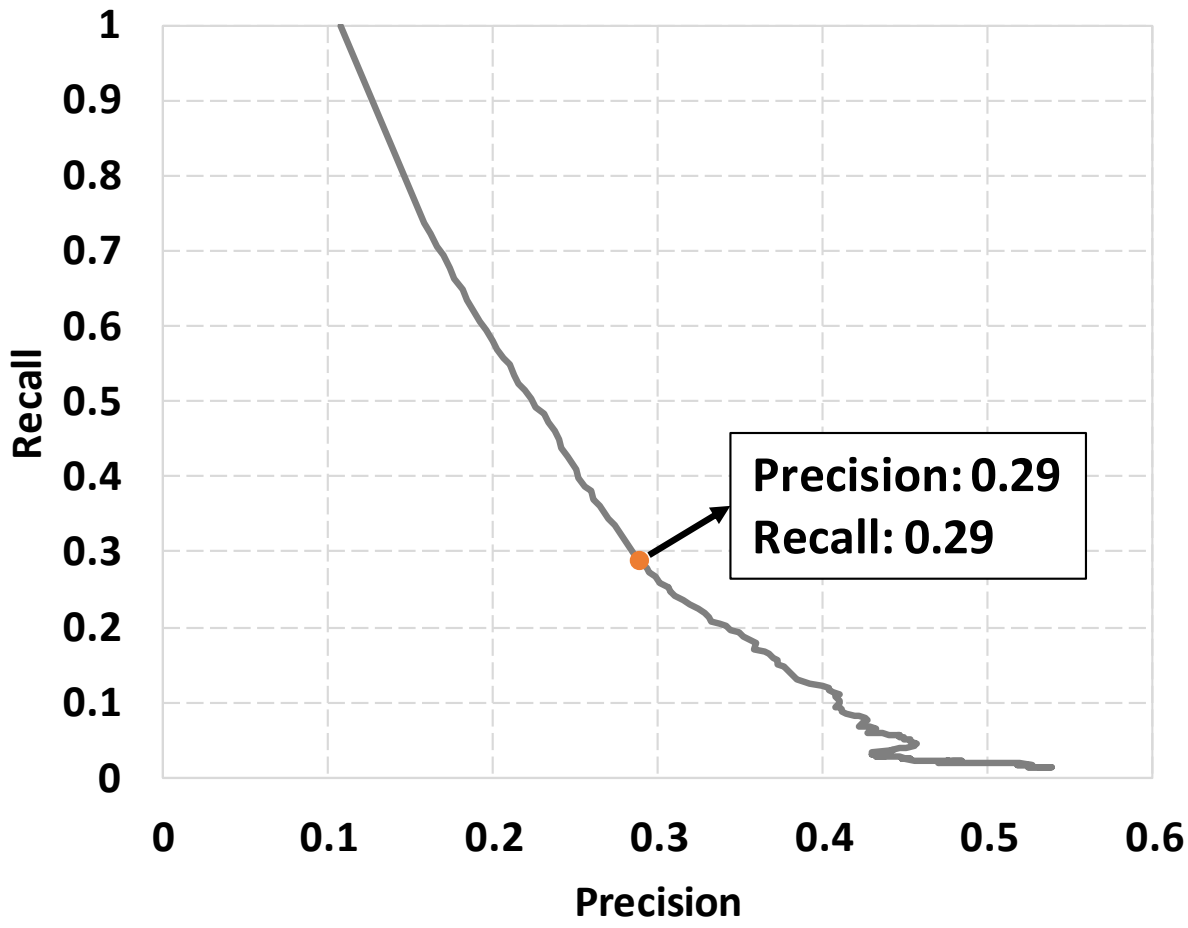
- Published in Accent Conversion Using Phonetic Posteriorgrams (Zhao et al., ICASSP'18)
- Used data from speakers in this corpus
- Samples:

L2 speaker	L1 reference	L2 speech	Accent conversion
ABA			
HKK			
TNI			

Usage example: MPD (1)

- Provide a baseline for MPD on L2-ARCTIC
- Based on a classic Goodness of Pronunciation (GOP) measurement
- Used phone-independent thresholding
- Used 206 utts to determine the search range of the threshold
- Tested on 1293 sentences
  - 41,353 phone samples
  - 4,415 (10%) were tagged as substitution errors

Usage example: MPD (2)



Conclusion

Current status

- Released XX speakers

Future work

- Release more speakers, Vietnamese in progress
- Cross-annotator analysis

License

- CC BY-NC 4.0

Links

- AC samples: http://people.tamu.edu/~guanlong.zhao/icassp18_demo.html
- Corpus: https://psi.engr.tamu.edu/l2-arctic-corpus/

Thanks
Q & A