跨语言的语音转换

Zhizheng Wu

https://drwuz.com/

The Chinese University of Hong Kong, Shenzhen

11/13/2022

Why do we need voice conversion? Voice dubbing case

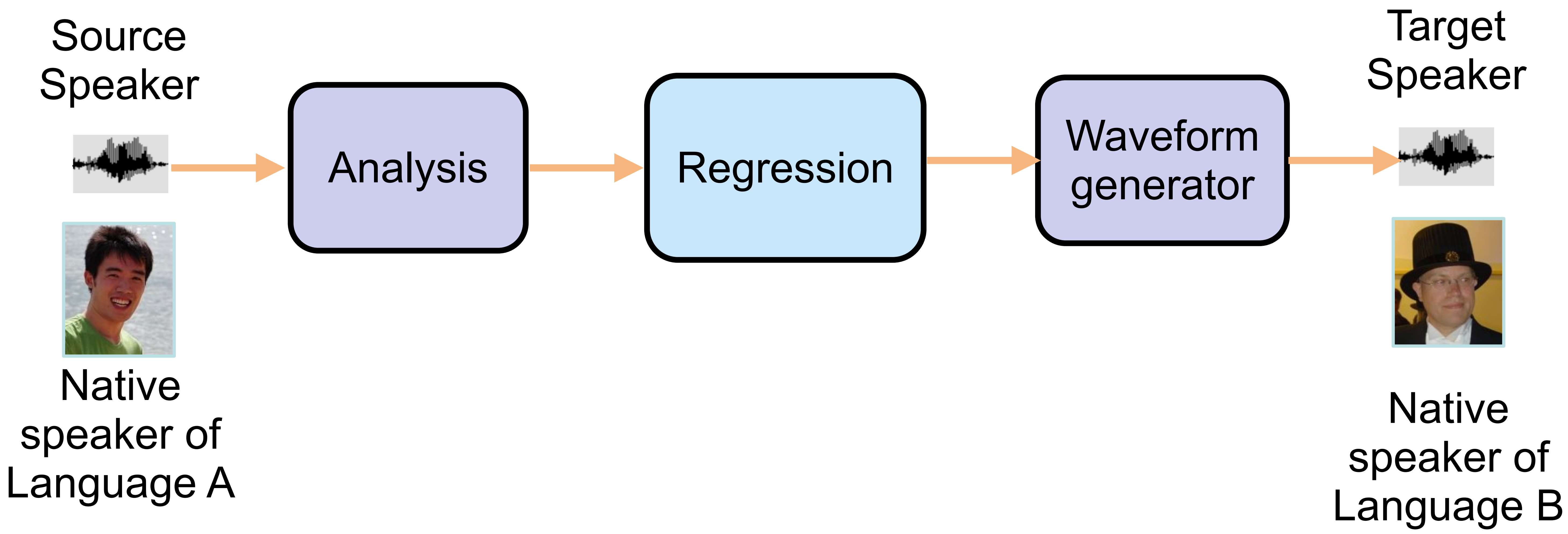




Voice dubbing in a different language

- The original movie actor may not speak different languages
- A native voice actor is needed
- However the voice timber between the native voice actor and the original movie actor is different

XVC: Cross-Lingual Voice Conversion



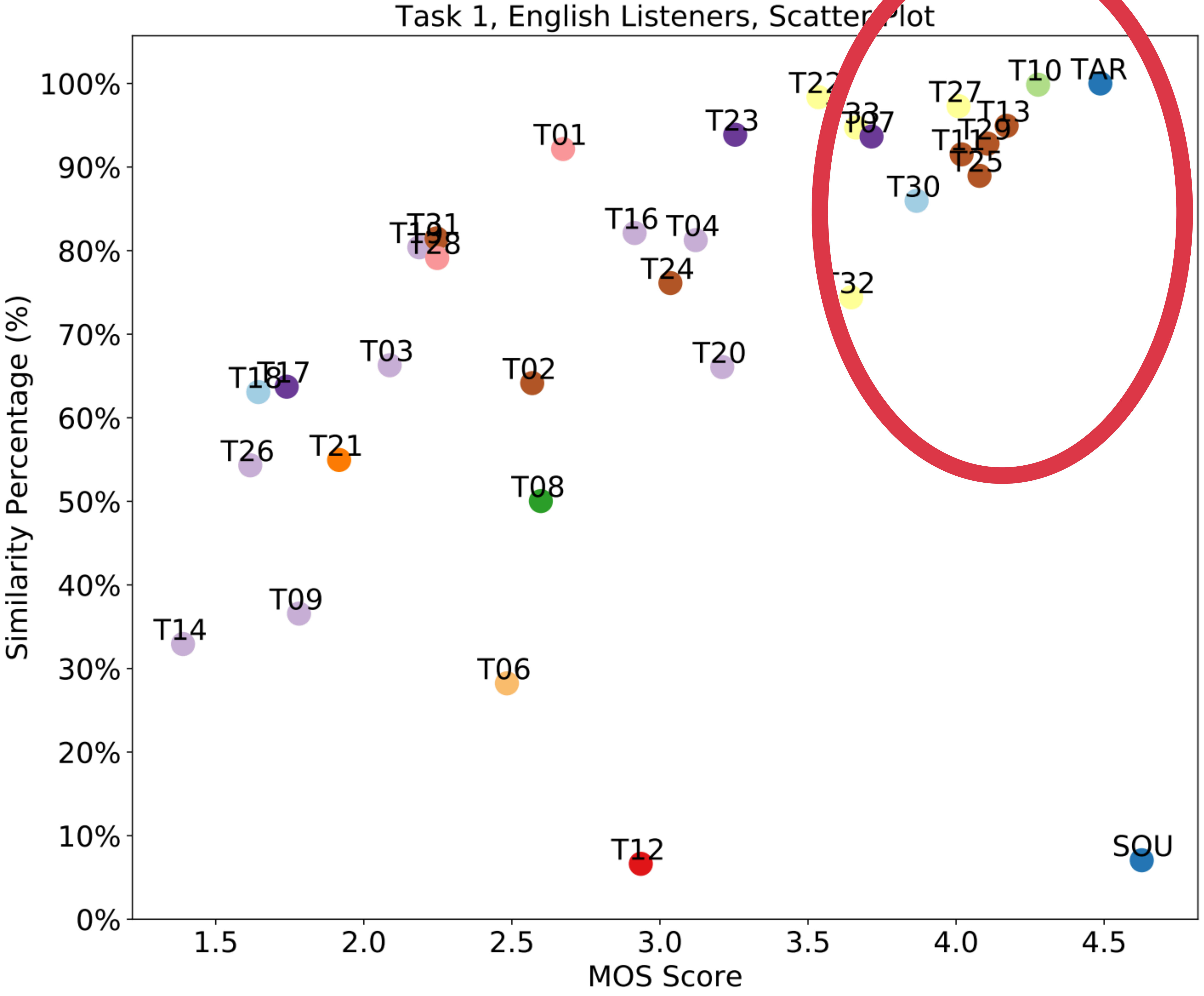
Voice Conversion: State of the Art

Team ID	Task 1		Task 2	
	VC model	Vocoder	VC model	Vocoder
T01	PPG-VC (Tacotron)	Parallel WaveGAN	N/A	N/A
T02	PPG-VC (Tacotron)	WaveGlow	PPG-VC (Tacotron)	WaveGlow
T03	AutoVC	WaveRNN	AutoVC	WaveRNN
T04	VQVAE	WaveNet	N/A	N/A
T05	N/A	N/A	PPG-VC (IAF)	WORLD & WaveGlow
T06	StarGAN	WORLD	StarGAN	WORLD
T07	NAUTILUS (Jointly trained TTS VC)	WaveNet	NAUTILUS (Jointly trained TTS VC)	WaveNet
T08	VTLN + Spectral differential	WORLD	VTLN + Spectral differential	WORLD
T09	AutoVC	Parallel WaveGAN	AutoVC	Parallel WaveGAN
T10	ASR-TTS (Transformer) / PPG-VC (LSTM)	WaveNet	PPG-VC (LSTM)	WaveNet
T11	PPG-VC (LSTM)	WaveNet	PPG-VC (LSTM)	WaveNet
T12	ADAGAN	AHOcoder	ADAGAN	AHOcoder
T13	PPG-VC (Tacotron)	WaveNet	PPG-VC (Tacotron)	WaveNet
T14	One shot VC	NSF	N/A	N/A
T15	N/A	N/A	AutoVC	MelGAN
T16	CycleVAE	Parallel WaveGAN	CycleVAE	Parallel WaveGAN
T17	Cotatron	MelGAN	N/A	N/A
T19	VQVAE	Parallel WaveGAN	VQVAE	Parallel WaveGAN
T20	VQVAE	Parallel WaveGAN	VQVAE	Parallel WaveGAN
T21	CycleGAN	MelGAN	N/A	N/A
T22	ASR-TTS (Transformer)	Parallel WaveGAN	ASR-TTS (Transformer)	Parallel WaveGAN
T23	Transformer VC (Jointly trained TTS VC)	Parallel WaveGAN	CycleVAE	WaveNet
T24	PPG-VC (Tacotron)	LPCNet	PPG-VC (Tacotron)	LPCNet
T25	PPG-VC (CBHG)	WaveRNN	PPG-VC (CBHG)	WaveRNN
T26	One shot VC	Griffin-Lim	One shot VC	Griffin-Lim
T27	ASR-TTS (Transformer)	Parallel WaveGAN	PPG-VC / ASR-TTS (Transformer)	Parallel WaveGAN
T28	Tacotron	WaveRNN	Tacotron	WaveRNN
T29	PPG-VC (CBHG)	LPCNet	PPG-VC (CBHG)	LPCNet
T31	Multi-speaker Parrotron	WaveGlow	Multi-speaker Parrotron	WaveGlow
T32	ASR-TTS (Tacotron)	WaveRNN	ASR-TTS (Tacotron)	WaveRNN
T33	ASR-TTS (Tacotron)	Parallel WaveGAN	PPG-VC (Transformer)	Parallel WaveGAN

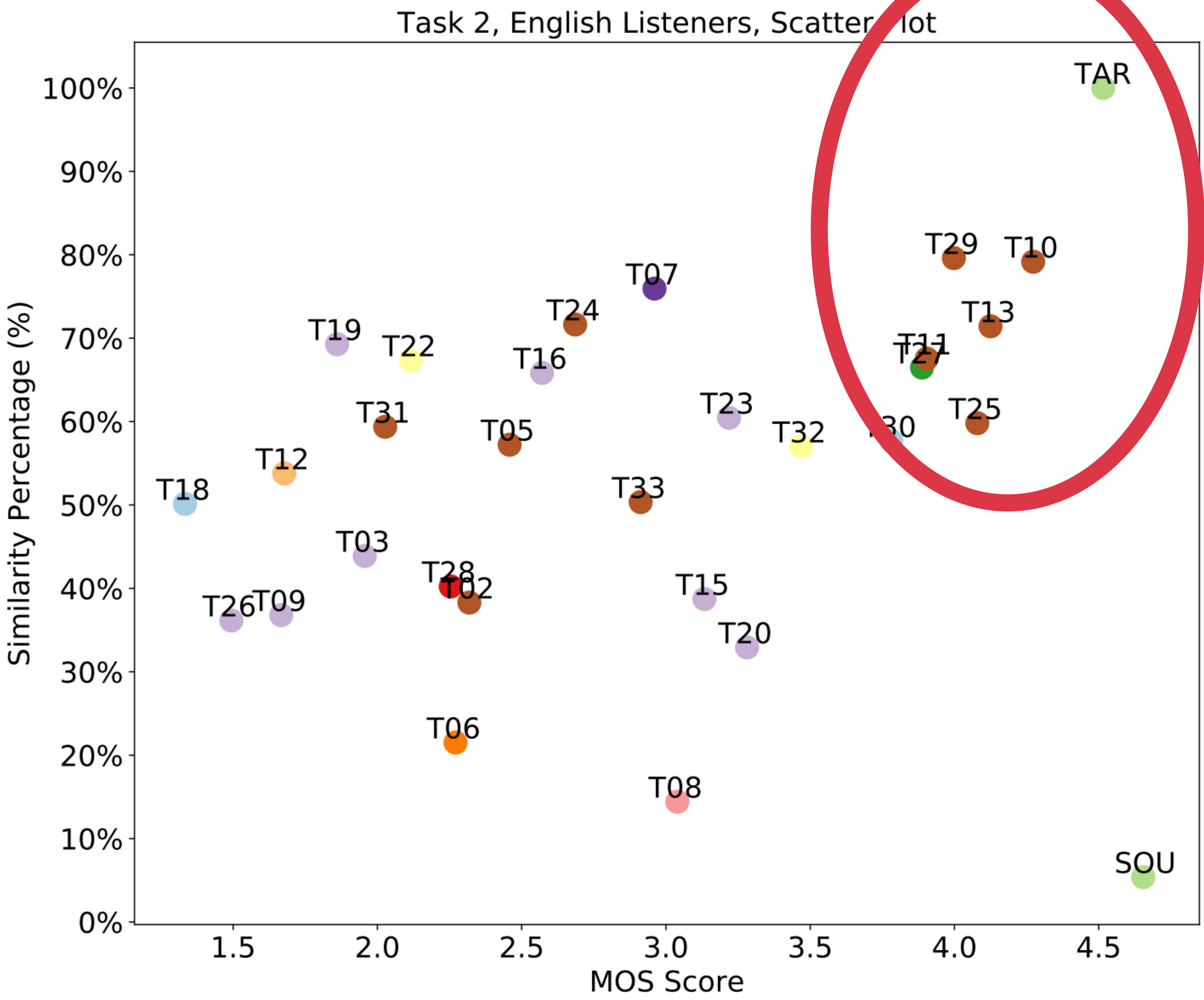
Zhao, Y., Huang, W.C., Tian, X., Yamagishi, J., Das, R.K., Kinnunen, T., Ling, Z. and Toda, T., 2020. Voice conversion challenge 2020: Intra-lingual semi-parallel and cross-lingual voice conversion. ISCA Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020

Voice Conversion: State of the Art

Intra Lingual



Cross Lingual



Zhao, Y., Huang, W.C., Tian, X., Yamagishi, J., Das, R.K., Kinnunen, T., Ling, Z. and Toda, T., 2020. Voice conversion challenge 2020: Intra-lingual semi-parallel and cross-lingual voice conversion. ISCA Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020

Opportunities in Cross-Lingual Voice Conversion

Speech Intelligibility：Objective Measure

Word Error Rate (WER): The lower the better

VCC2020 top systems

	Source	T10	T13	T25	T29	Average
Intra-Lingual	13.79	11.26	18.69	20.19	23.33	18.37
Cross-Lingual		15.11	22.99	24.68	31.48	23.57

In-house system

Voice Conversion	ENG WER (%)	MAN CER (%)	Average
Source	14.61	12.11	13.36
Intra-Lingual	24.01	21.68	22.85
Cross-Lingual	35.66	29.87	32.77

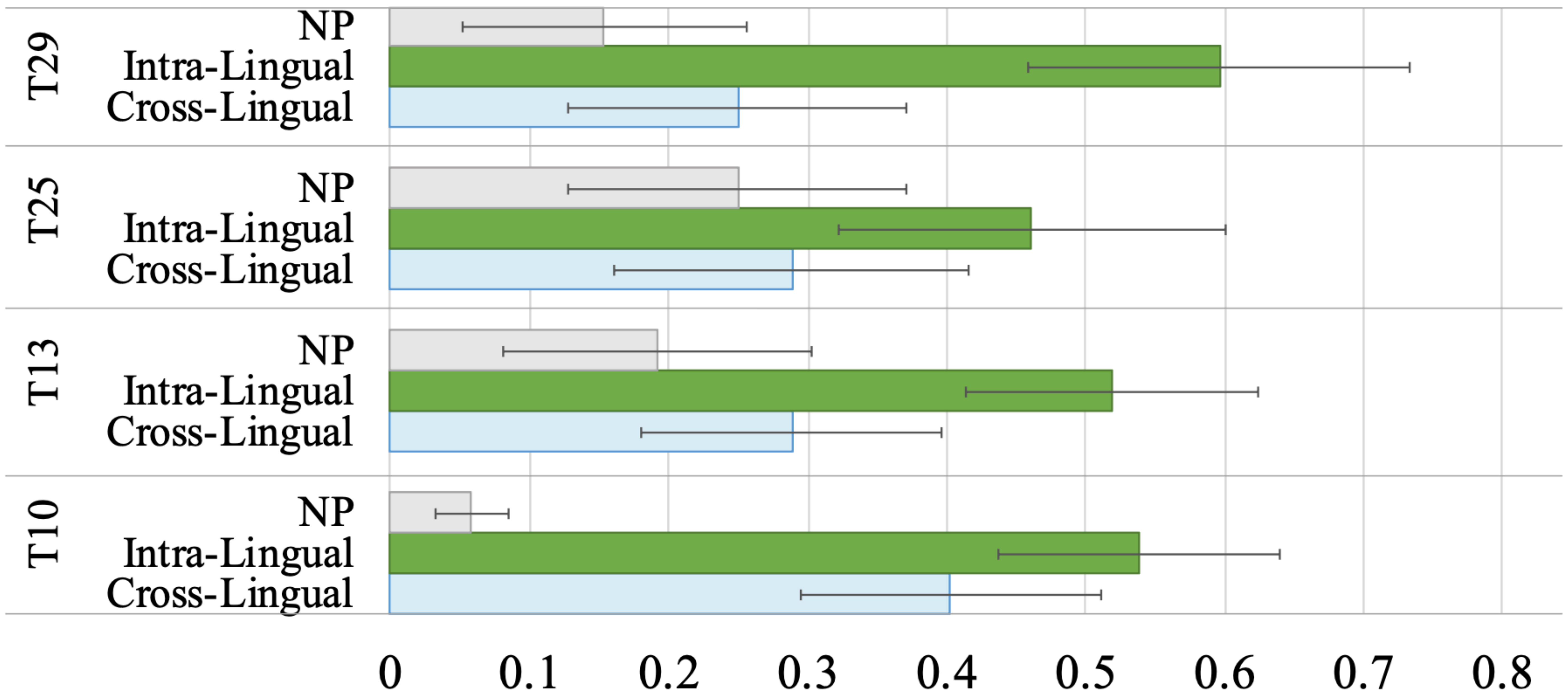
https://cloud.google.com/speech-to-text

Yi Zhou, Zhizheng Wu, Xiaohai Tian, Haizhou Li, “Optimization of Cross-Lingual Voice Conversion with Linguistics Losses to Reduce Foreign Accents”, submitted to IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING

Speech Intelligibility: Subjective Measure

Subjective preference test: VCC 2020 top 4 systems

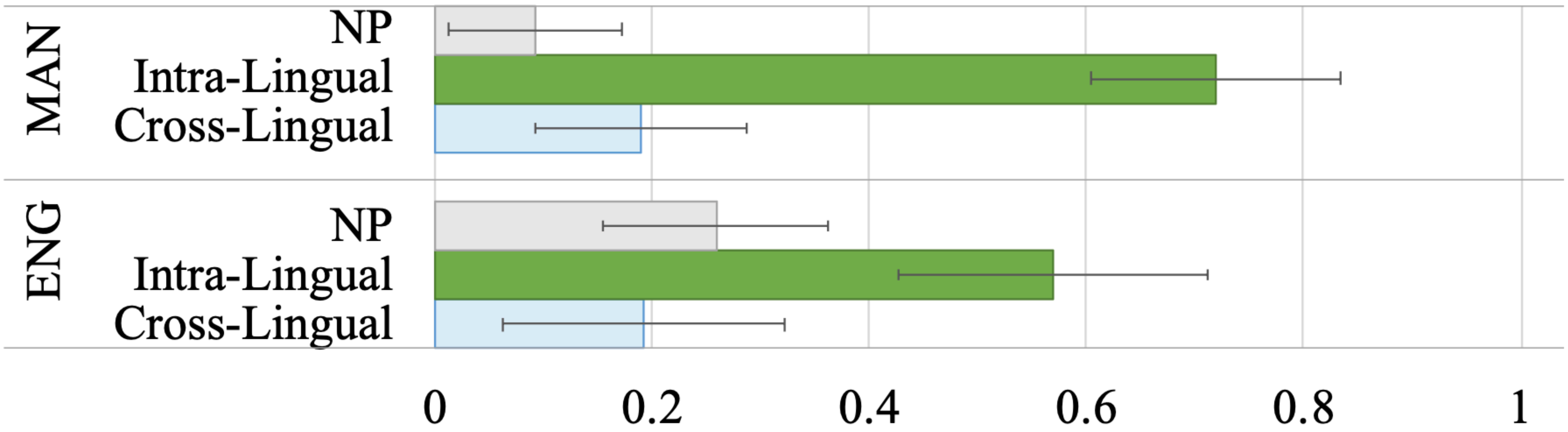
- 20 listeners, each evaluated 20 pairs



Yi Zhou, Zhizheng Wu, Xiaohai Tian, Haizhou Li, "Optimization of Cross-Lingual Voice Conversion with Linguistics Losses to Reduce Foreign Accents", submitted to IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING

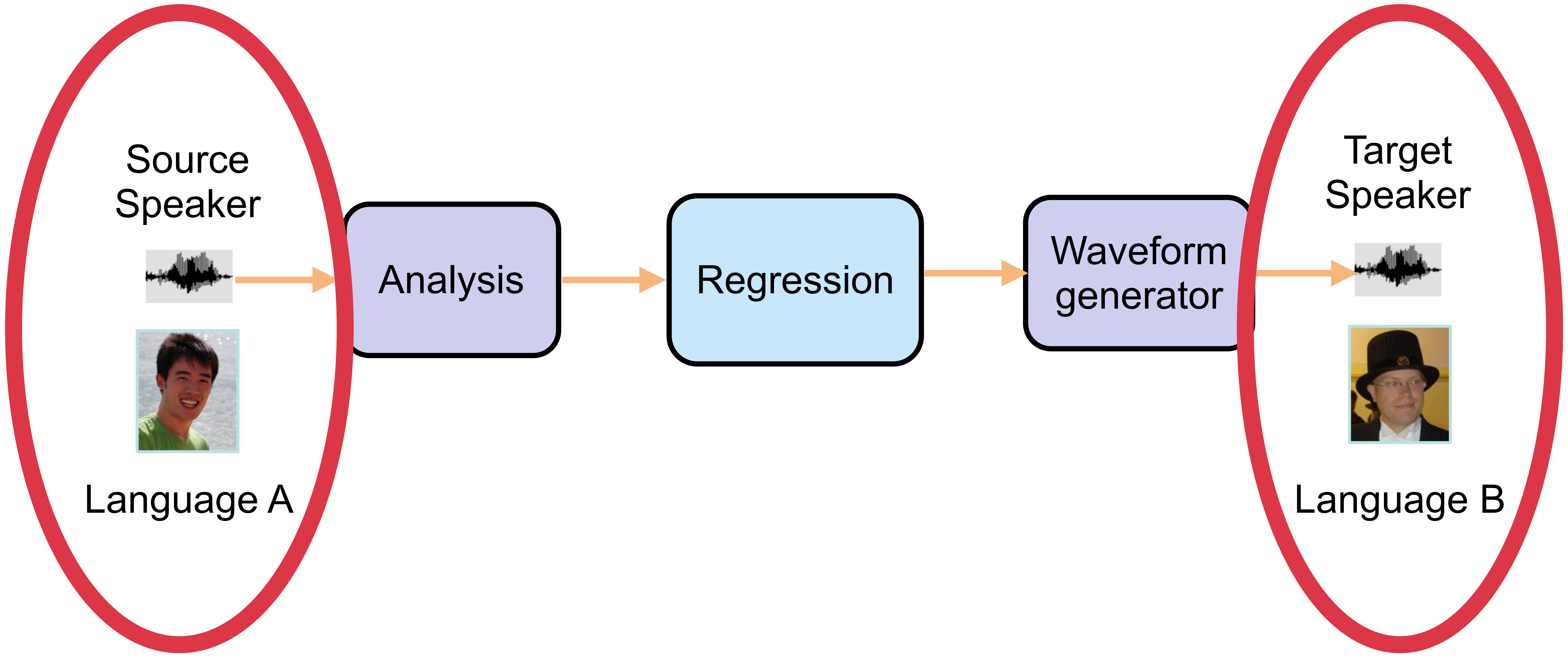
Speech Intelligibility: Subjective Measure

Subjective preference test: In-house XVC system



Yi Zhou, Zhizheng Wu, Xiaohai Tian, Haizhou Li, “Optimization of Cross-Lingual Voice Conversion with Linguistics Losses to Reduce Foreign Accents”, submitted to IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING

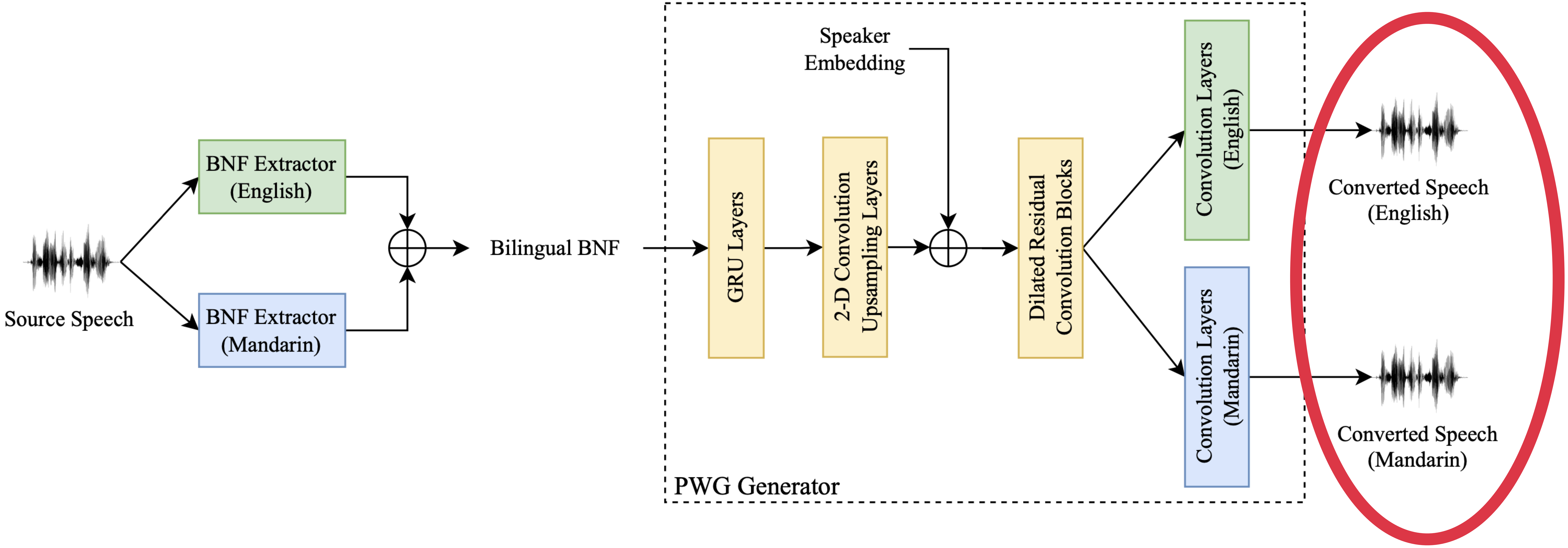
XVC Speech Intelligibility



WER: 13.79%

WER: 23.57%

XVC system



Yi Zhou, Zhizheng Wu, Xiaohai Tian, Haizhou Li, “Optimization of Cross-Lingual Voice Conversion with Linguistics Losses to Reduce Foreign Accents”, submitted to IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING

XVC system with linguistic loss

Yi Zhou, Zhizheng Wu, Xiaohai Tian, Haizhou Li, "Optimization of Cross-Lingual Voice Conversion with Linguistics Losses to Reduce Foreign Accents", submitted to IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING

XVC system with linguistic loss

$$\mathcal{L}_{Total} = \mathcal{L}_{PWG} + \lambda_{BNF}\mathcal{L}_{BNF} + \lambda_{ASR}\mathcal{L}_{ASR}$$

Waveform and GAN loss

Frame-by-frame BNF loss

Segment ASR loss

Yi Zhou, Zhizheng Wu, Xiaohai Tian, Haizhou Li, "Optimization of Cross-Lingual Voice Conversion with Linguistics Losses to Reduce Foreign Accents", submitted to IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING

Experimental setup

- BNF extractors
  - English: 460-hour Librispeech
  - Mandarin: 1,238 hours of speech
    - AIDataTang, AISHELL-1, MagicData, PrimeWords, ST-CMDS, and THCHS-30
- ASR systems
  - English: 460-hour Librispeech
    - WER: 10.12%
  - Mandarin: 151-hour AISHELL-1
    - CER: 5.72%

Experimental setup

- Speaker embedding: 256 dimension
  - Pre-trained on the AISHELL-2 database [59]
  - Fine-tuned with English and Mandarin speech data from 100 speakers

- XVC system
  - 50 English speakers are randomly selected from the VCTK database
  - 50 Mandarin speakers from Data-Baker Mandarin Corpus
  - Each speaker has 150 utterances

Experimental setup

- XVC system
  - 4 bilingual speakers (MF2, MF4, MM1, MM2) from the EMIME database for testing
  - Each speaker 20 English utterances and 20 Mandarin utterances

Source-Target Speaker Pairing
MF2 $\rightarrow$ MM2 (Female $\rightarrow$ Male)
MF4 $\rightarrow$ MF2 (Female $\rightarrow$ Female)
MM1 $\rightarrow$ MF4 (Male $\rightarrow$ Female)
MM2 $\rightarrow$ MM1 (Male $\rightarrow$ Male)

Experimental setup

- XVC system
  - BNF2Wav: Baseline system which take bottleneck feature as input and predict waveform directly
  - BNF2Wav-BNF-ASR(F0): Bottleneck feature and F0 as input. Using both BNF and ASR losses

	Experimental System	Configuration				MCD			RMSE			WER/CER (%)		
		Input	Output	BNF Loss	ASR Loss	ENG	MAN	Avg	ENG	MAN	Avg	ENG	MAN	Avg
	Natural Source Speech	N.A.				8.71	8.94	8.83	18.08	19.93	19.01	8.21	3.75	5.98
1)	BNF2Wav	BNF	Wav	×	×	8.77	9.01	8.89	13.24	13.41	13.33	21.66	17.83	19.75
2)	BNF2Wav($F0$)	BNF$\bigoplus F0$	Wav	×	×	8.69	8.81	8.75	13.19	13.17	13.18	21.68	17.77	19.73
3)	BNF2Mel-BNF	BNF	Mel	✓	×	8.71	8.78	8.75	12.73	12.89	12.81	15.46	10.01	12.74
4)	BNF2Wav-BNF	BNF	Wav	✓	×	7.85	7.96	7.91	12.52	12.40	12.46	12.10	9.98	11.04
5)	BNF2Wav-ASR	BNF	Wav	×	✓	8.66	8.63	8.65	12.55	12.61	12.58	11.06	9.25	10.16
6)	BNF2Wav-BNF-ASR	BNF	Wav	✓	✓	8.01	8.24	8.13	12.46	12.38	12.42	11.33	9.02	10.18
7)	BNF2Wav-BNF-ASR($F0$)	BNF$\bigoplus F0$	Wav	✓	✓	7.96	7.99	7.98	12.49	12.31	12.40	11.28	9.13	10.21

Subjective test

- Speech quality MUSHRA
  - 20 listeners, each listen to 20 samples

Subjective test

- Speaker similarity
  - 20 listeners, each listen to 20 samples

1) BNF2Wav 2) BNF2Wav(F0) 3) BNF2Mel-BNF 4) BNF2Wav-BNF 5) BNF2Wav-ASR 6) BNF2Wav-BNF-ASR 7) BNF2Wav-BNF-ASR(F0)

Samples

	Female-Female	Female-Male	Male-Male	Male-Female
Source				
Target				
Baseline				
Proposed				

https://vcsamples.github.io/taslp2022/

Summary

- There are opportunities in state-of-the-art XVC systems, especially intelligibility
- With additional linguistic loss, the converted samples are more intelligible
- The speech quality is also improved