CROSS-LINGUAL VOICE CONVERSION WITH BILINGUAL PHONETIC POSTERIORGRAM AND AVERAGE MODELING

Yi Zhou[1], Xiaohai Tian[1], Haihua Xu[2], Rohan Kumar Das[1] and Haizhou Li[1]

[1]Department of Electrical and Computer Engineering, National University of Singapore, Singapore

[2]Temasek Laboratories, Nanyang Technological University, Singapore

OBJECTIVE

This work presents the cross-lingual voice conversion approaches with:

- bilingual Phonetic PosteriorGram (PPG) to represent speaker-independent features of speech signals from different languages in the same feature space.
- the average model to leverage both linguistic and acoustic information from other speakers in different languages, i-vector is used for network adaptation.

1. INTRODUCTION

In cross-lingual voice conversion, the source and target speakers speak in different languages.

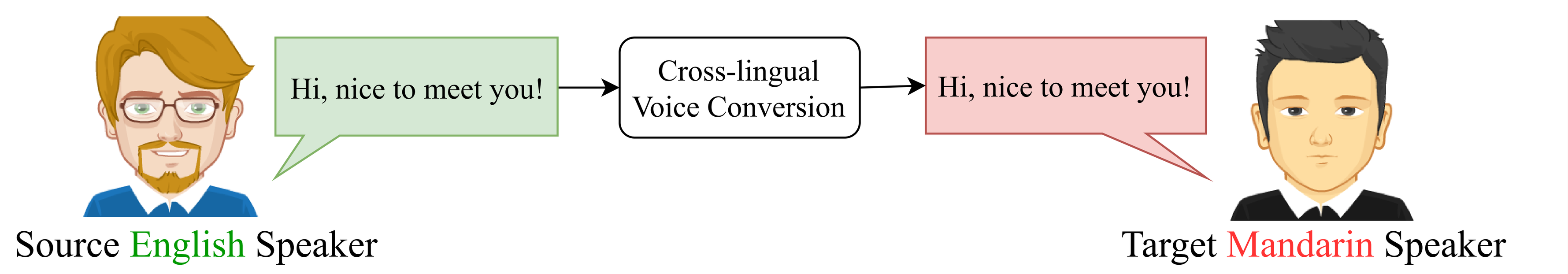

Figure 1: An example to convert from an English source speaker to a Mandarin target speaker.

2. LIMITATIONS OF MONOLINGUAL PPG

- Mismatched phonetic representation, inaccurate linguistic information

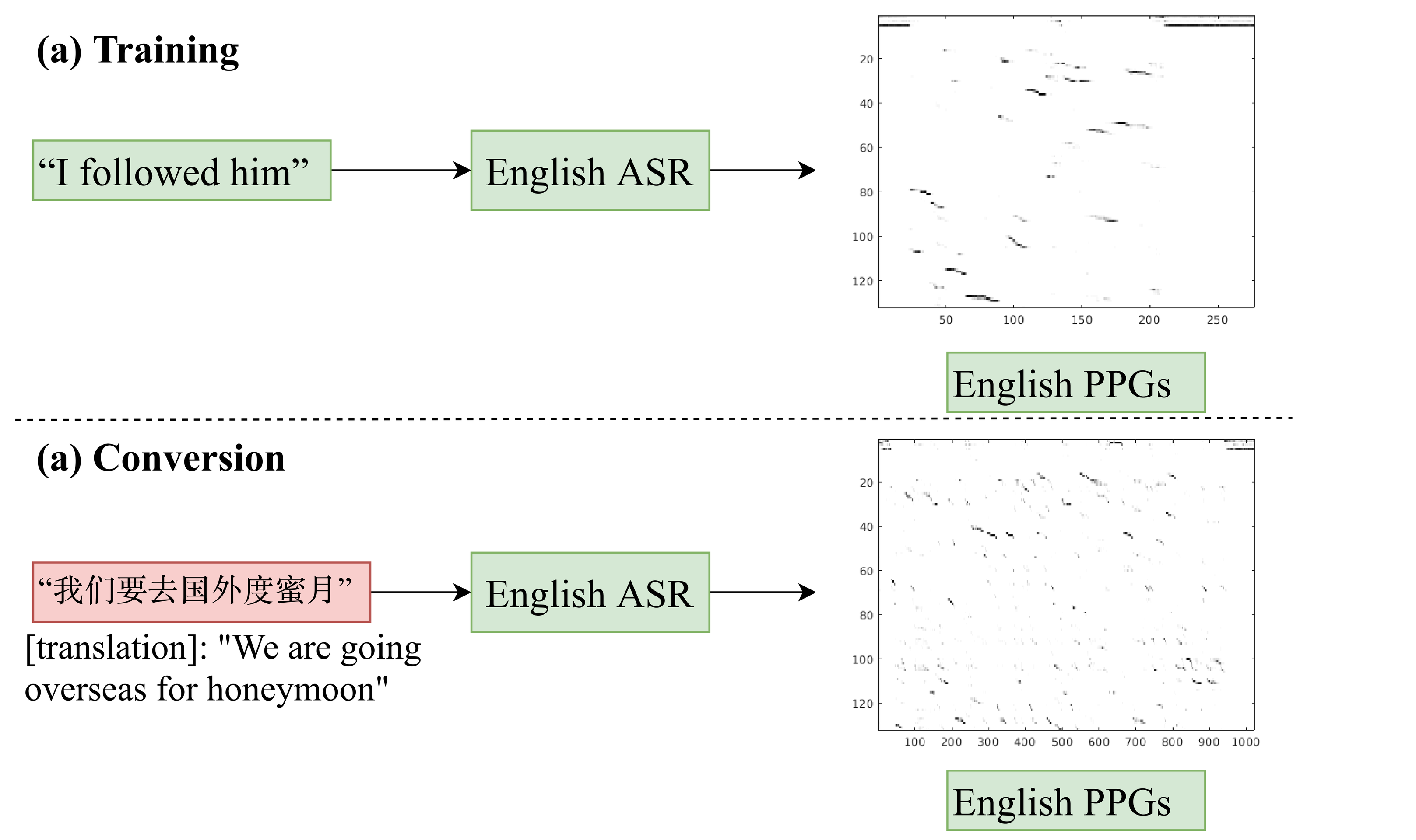

Figure 2: The monolingual PPG example converting Mandarin speech to an English speaker

3. PROPOSED AVERAGE MODELING APPROACH WITH BILINGUAL PPG

- Average Modeling Approach with Bilingual PPG

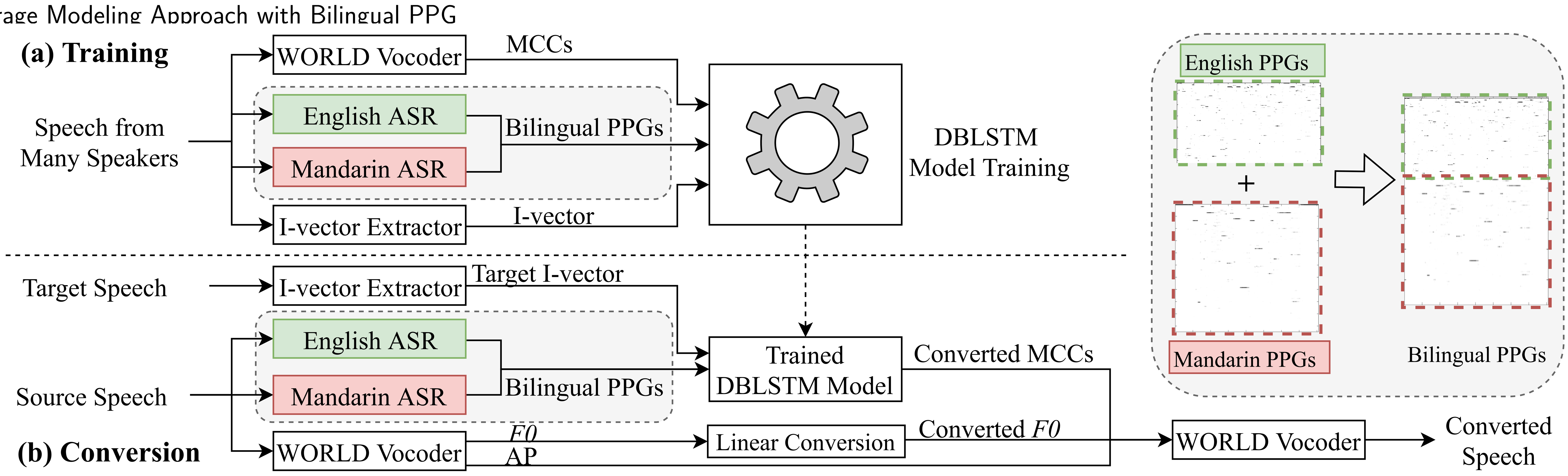

Figure 3: (a) training and (b) conversion workflow of the proposed average modeling approach with bilingual PPGs.

4. EXPERIMENTS AND RESULTS

Database

Table 1: Database used for experiments

	English	Mandarin
ASR	Wall Street Journal (WSJ)	Aishell
Voice Conversion	VCC2016, VCC2018	Library of Average Model [1]

System Descriptions

Table 2: Experimental systems and training data

	System	Training Data
M-PPG	monolingual PPG baseline [2]	150 utterances
B-PPG	proposed bilingual PPG	150 utterances
B-PPG-AMA	proposed average modeling approach with bilingual PPG	1500 utterances 5 English 5 Mandarin speakers

Objective Evaluation for Intralingual Voice Conversion

Table 3: MCD results for intralingual voice conversion

	M-PPG (EN)	M-PPG (CN)	B-PPG
EN2EN	6.486	7.99	**6.339**
CN2CN	8.12	6.759	**6.422**

Converted Samples

Subjective Evaluation for Cross-lingual Voice Conversion

ABX Speaker Similarity Test Result

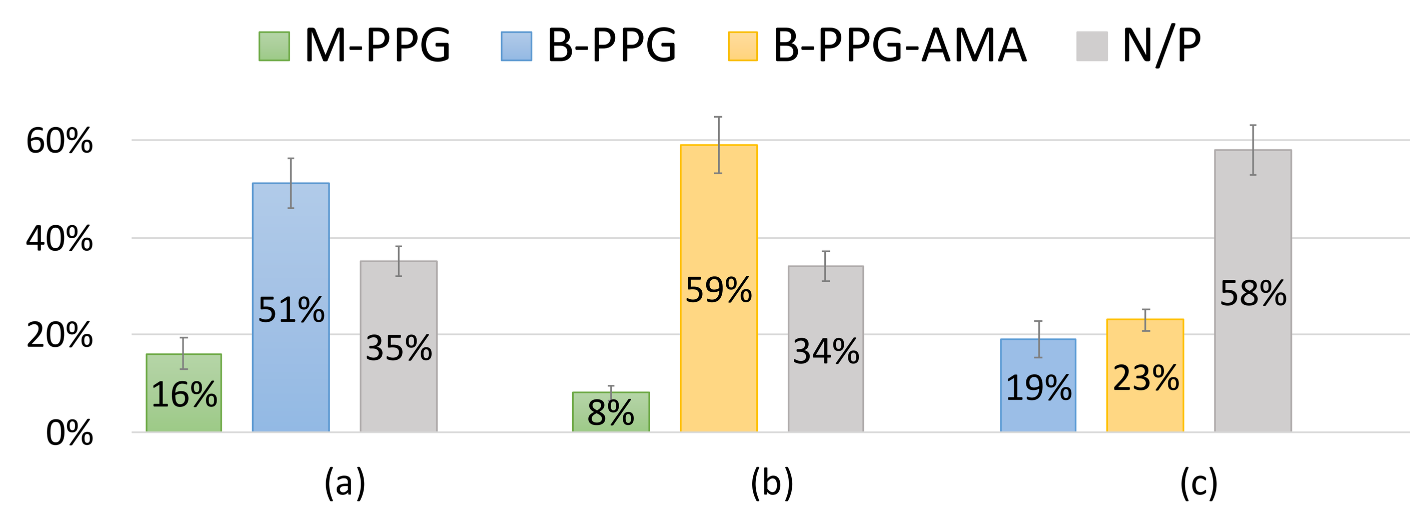

Figure 4: ABX preference test results for speaker similarity

MOS Quality Test Result

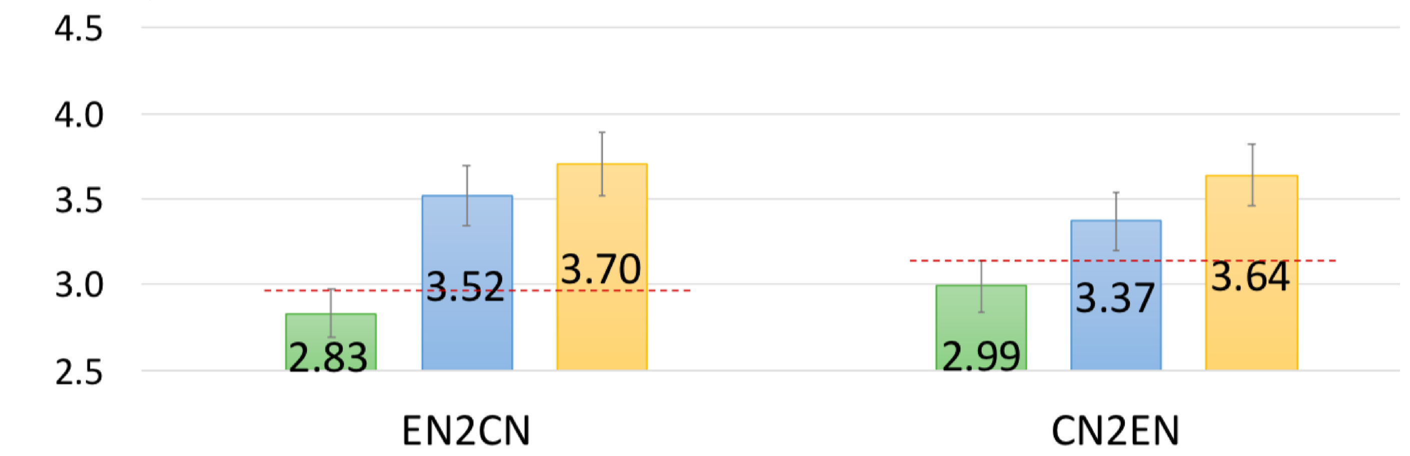

Figure 5: MOS test results for quality and naturalness

[1] http://www.data-baker.com/hc_pm_en.html

[2] L. Sun, H. Wang, S. Kang, K. Li, and H. Meng, "Personalized, cross-lingual TTS using phonetic posteriorgrams," in INTERSPEECH, pp. 322-326, 2016.