The 10th ISCSLP
17-20/10/2016
Tianjin, China
ID: 133, P1-8

Perceptual Evaluation of Natural and Synthesized Speech with Prosodic Focus in Mandarin Production of American Learners

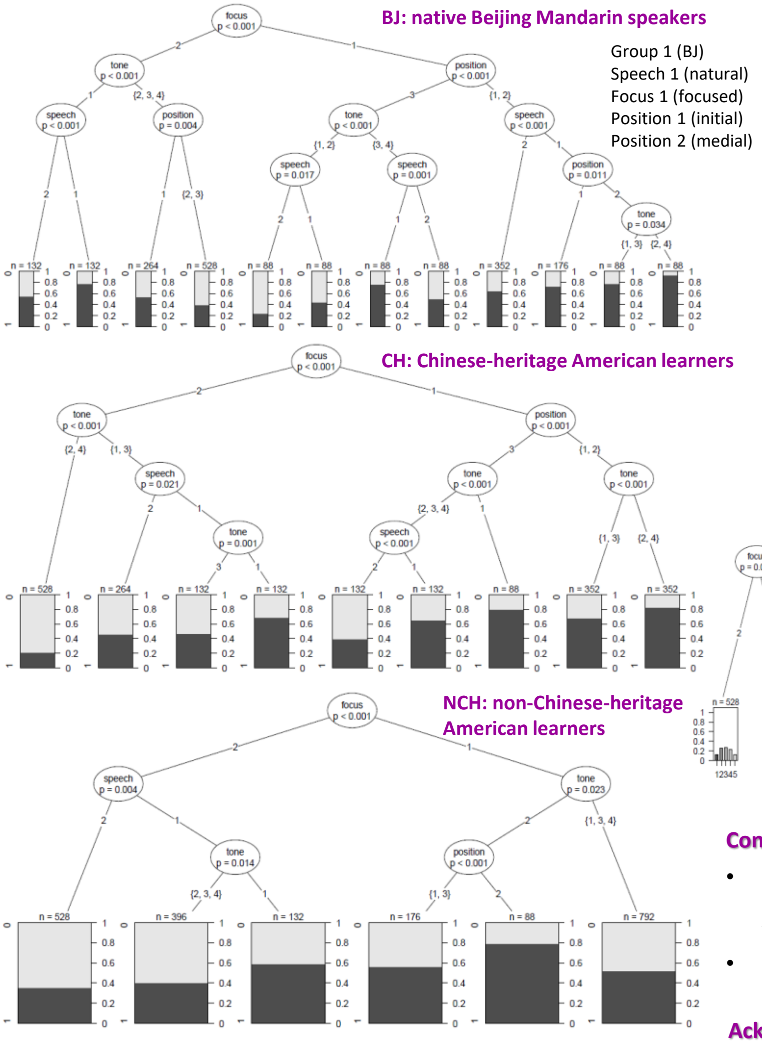
Ying Chen ychen@njust.edu.cn, Li Liu, Xueqin Zhao
School of Foreign Studies, Nanjing University of Science and Technology

Introduction

- Bilingual learners use duration and intensity more than F0 and in-focus expansion more than post-focus compression to code focus in L2 (Wu & Chung, 2011; Chen et al., 2014).
- The Parallel Encoding and Target Approximation (PENTA) model interprets speech prosody as a process of encoding communicative functions in parallel and sequentially realizing pitch target by the surface F0 (Xu, 2005).
- As a data-driven system, PENTAtrainer2 models, learns and synthesizes natural speech based on hierarchically layered functional annotations (Xu & Prom-on, 2014). This work uses PENTAtrainer2 to examine the production of prosodic focus in L2 Mandarin of American learners.

Research questions

- Can American learners' acoustic realization of focus in L2 Mandarin be recognized by native listeners?
- Can native Mandarin listeners recognize focus in speech synthesized based on speaker group modeling?
- How natural do native listeners find the learners' speech and its synthesized version?


BJ: native Beijing Mandarin speakers

Group 1 (BJ)
Speech 1 (natural)
Focus 1 (focused)
Position 1 (initial)
Position 2 (medial)

CH: Chinese-heritage American learners

NCH: non-Chinese-heritage American learners

Focus status identification for groups: Conditional inference trees

Stimuli

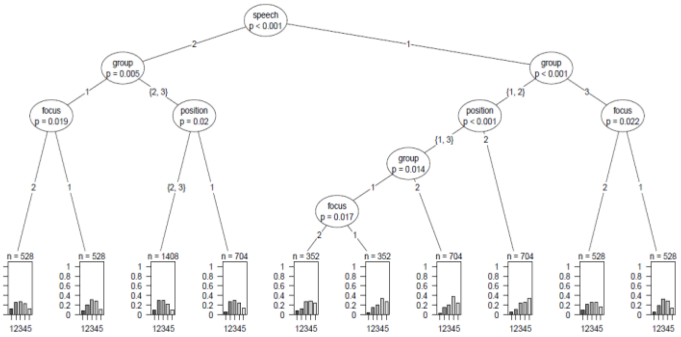
No Focus	Q	ni3 shuo1 shen2me0? 'What did you say?'
	A	*See initial, medial, and final focus sentences below.*
Initial Focus	Q	shui2 mo1 ni1la1? 'Who patted Nila?'
	A	*wu1ma1/liu2ma1/li3ma1/wei4ma1/* mo1 ni1la1 'Wuma/ Liuma/ Lima/ Weima patted Nila.'
Medial Focus	Q	wu1ma1 dui4 ni1la1 zuo4 shen2me0? 'What did Wuma do to Nila?'
	A	wu1ma1 *mo1/ nao2/ lou3/ ma4* ni1la1 'Wuma patted/scratched/hugged/cursed Nila.'
Final Focus	Q	wu1ma1 mo1 shui2? 'Who did Wuma pat?'
	A	wu1ma1 mo1 *ni1la1/ni1lan2/ni1mei3/ni1na4* 'Wuma patted Nila/ Nilan/ Nimei/ Nina'.

Procedures

- Three functional layers were annotated: tone (1, 2, 3 and 4), syllable position (word-final, sentence-final, non-final and monosyllabic) and focus condition (pre-focus, in-focus and post-focus) for the synthesis.
- The natural and synthesized speech of one male speaker and one female speaker, who displayed the median standard deviations of their groups, was selected as the stimuli for the perception experiment.

Focus status identification: Logistic regression results

Variable	*b*	*se(b)*	*z*	*p*
(Intercept)	1.229	0.142	8.649	0.000
Group 2 (CH)	-0.160	0.077	-2.093	0.036
Group 3 (NCH)	-0.411	0.075	-5.458	0.000
Speech 2 (synthesized)	-0.398	0.069	-5.751	0.000
Position 3 (final)	-0.314	0.065	-4.794	0.000
Focus 2 (no focus)	-0.916	0.054	-16.993	0.000
Tone 2	-0.229	0.076	-3.026	0.002
Tone 3	-0.178	0.076	-2.348	0.019
Tone 4	-0.338	0.076	-4.453	0.000



Speech naturalness rating: Conditional inference tree

Conclusions

- Native Mandarin listeners recognized and rated early American learners' (CH) acoustic realization of focus in L2 Mandarin better than that of late learners (NCH).
- They also recognized and rated natural speech better than the synthesized speech modeled by speaker group.

Acknowledgments This work was supported by the National Science Foundation of China 61573187.