Study on the Relation of Fundamental and Formant Frequencies for Affective Speech Synthesis

Bogu Li[1], Zhilei Liu[1], Jianwu Dang[1, 2]

[1]Tianjin Key Lab. of Cognitive Computing and Application, Tianjin University, Tianjin, China

[2]Japan Advanced Institute of Science and Technology, Ishikawa, Japan

E-mail: {libogu, zhileiliu, dangjianwu}@tju.edu.cn

Aims

1. To investigate the process of affective speech production based on the combination of fundamental frequency (F0) and formant frequencies;
2. To investigate the relations between F0 and formants of emotional speech;
3. The relations are investigated using the logistic regression (LR). For a given emotion-related F0, the formants can be predicted correctly using the LR models;
4. Experiments on affective speech synthesis were conducted on three different emotional speech datasets,.

Methods

Traditional emotional formant synthesis just modifies F0 contour to generate speech with different emotions, in this paper, the modification process of formants is illustrated in the bottom of Figure 1 with three steps as follows:

- The F1, F2 trajectories of the neutral speech and F0 contour of the positive or negative speech are adopted as the initial input of the logistic regression model;
- The gradient ascent algorithm is adopted to modify F1 and F2 in the corresponding LR models;
- The learned negative or positive F0 contour and the modified formant trajectories are used to get the synthesized emotional speech.

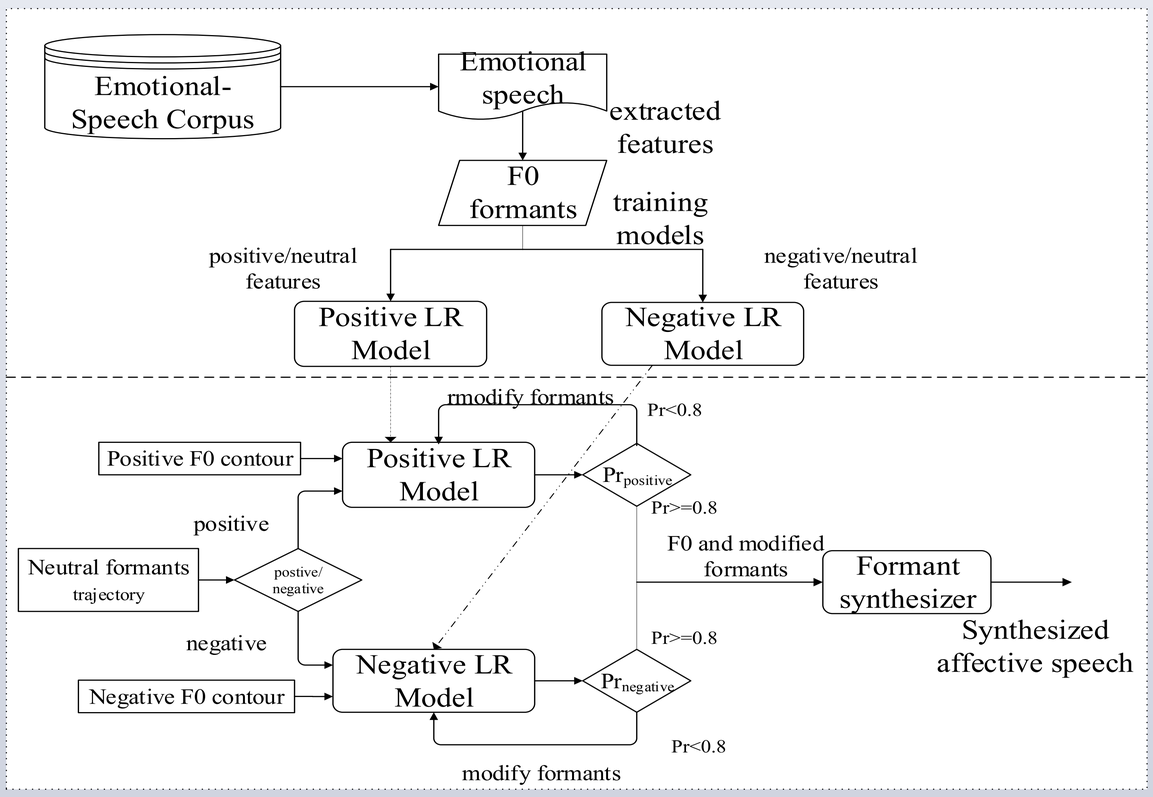

Figure 1: The proposed framework for affective speech synthesis.

Result 1:The coefficients of the learned logistic regression model

Databases	Variables	Coefficients	
		Positive-Neutral LR	Negative-Neutral LR
LANG	B	-10.36 (-)	-8.58 (-)
	F0	25.53 (+)	24.53 (+)
	F1	-2.31 (-)	3.39 (+)
	F2	-1.58 (-)	-1.78 (-)
Berlin	B	-3.81 (-)	-6.59 (-)
	F0	18.78 (+)	16.44 (+)
	F1	-2.27 (-)	6.09 (+)
	F2	-3.47 (-)	-4.56 (-)
CASS	B	-4.21 (-)	-6.13 (-)
	F0	14.61 (+)	12.43 (+)
	F1	-1.21 (-)	3.21 (+)
	F2	-2.24 (-)	-5.38 (-)

Result 2: The recognition result on synthesized speech WITHOUT/WITH modification of F1, F2 (Ne: negative, Po: positive, Nu: neutral, AC: accuracy)

Databases	Emotions	Predicted			SUM	AC
		Ne	Po	Nu		
LANG	Ne	61	/	19	80	76%
	Po	/	63	17	80	79%
Berlin	Ne	45	/	25	70	64%
	Po	/	46	24	70	66%
CASS	Ne	56	/	55	111	50%
	Po	/	58	53	111	52%

WITHOUT modification

Databases	Emotions	Predicted			SUM	AC
		Ne	Po	Nu		
LANG	Ne	75	/	5	80	94%
	Po	/	76	4	80	95%
Berlin	Ne	66	/	4	70	94%
	Po	/	67	3	70	96%
CASS	Ne	110	/	1	111	99%
	Po	/	109	2	111	98%

WITH modification

Result 3: The acoustic features of synthesized speech and target speech(Ne: negative, Po: positive, Nu: neutral, MF0: mean f0 (Hz), MF1: mean f1 (Hz), MF2: mean f2 (Hz)).

Databases	Variables	Synthesized Speech		Target Speech		
		Po	Ne	Po	Ne	Nu
LANG	MF0	165.9	182.5	157.7	191.8	147.0
	MF1	764.5	781.3	776.6	798.2	779.6
	MF2	1671.2	1578.6	1692.1	1589.2	1671.9
Berlin	MF0	285.9	203.3	281.0	216.9	132.1
	MF1	447.9	475.6	438.9	481.8	456.6
	MF2	1381.1	1329.6	1427.7	1386.5	1422.4
CASS	MF0	241.6	220.4	258.8	215.4	165.4
	MF1	360.3	387.9	354.9	410.3	362.7
	MF2	1396.4	1351.9	1407.8	1362.2	1373.2

Conclusions

- Experiment results demonstrate that positive speech has lower F1 and the negative speech has lower F2 and higher F1, which are consistent with [1];
- The recognition results of the synthesis emotional speech with/without formant modification in logistic regression models verify the effectiveness of our proposed affective speech synthesis method.

Reference

[1] Erickson, D., et al. "Some non-f0 cues to emotional speech: an experiment with morphing," *Proc. speech prosody,* 2008, pp. 677-680.

Acknowledgements

This work is supported by the National Basic Research Program of China (No. 2013CB329301), the National Natural Science Foundation of China (No. 61233009 and No. 61503277). The study is supported partially by JSPS KAKENHI Grant (16K00297)