AN ARTICULATORY-FUNCTIONAL APPROACH TO MODELING PERSIAN FOCUS PROSODY

Mortaza Taheri-Ardali, Yi Xu

Department of Linguistics, Institute for Humanities and Cultural Studies, Tehran, Iran
Department of Speech, Hearing and Phonetic Sciences, University College London, UK
m.taheri@ihcs.ac.ir, yi.xu@ucl.ac.uk

ABSTRACT

This paper is an attempt to test PENTA, an articulatory-functional model, on Persian focus prosody. The test was done on a corpus consisting of utterances with different focus conditions using PENTAtrainer2, a trainable prosody synthesizer that optimizes categorical pitch targets each corresponding to multiple communicative functions. The evaluation was done by comparing the F_0 contours generated by the extracted pitch targets to those of natural utterances through numerical and perceptual evaluations. The numerical results showed that the synthesized F_0 was close to the natural contour in terms of RMSE (= 1.94) and Pearson's r (= 0.84). Perceptual evaluation showed that the rate of focus identification and naturalness judgement by native Persian listeners were highly similar between synthetic and natural F_0 contours.

Keywords: PENTA model, focus, quantitative target approximation, Persian.

1. INTRODUCTION

Prosody as a key component of speech has always been a hard challenge for speech technology. For instance, in text-to-speech synthesis, it is still an unresolved issue how to generate rich human-sounding prosody. Finding a solution to this problem will not only facilitate the progress of speech technology but also improve the theoretical understanding of speech.

The main acoustic correlate of speech prosody is F_0, and most research effort has been spent on trying to achieve acceptable computational modeling of pitch contours. Although variability and uncertainty of F_0 makes its modeling really difficult, there have been many attempts to achieve this goal in recent years as reviewed in [16].

Previous approaches can be largely divided into two general categories, namely, those that model F_0 contours directly and those that attempt to simulate the underlying mechanisms of F_0 production [10]. One of the latter approaches is the parallel encoding and target approximation (PENTA) model [14], which is currently realized by the quantitative target approximation (qTA) model [10]. It has been shown that high-accuracy predictive synthesis of F_0 in languages like English, Mandarin and Thai can be achieved with this approach [16]. The current contribution is to assess PENTA's ability in modeling Persian F_0 in a small corpus containing focal and non-focal utterances.

Persian prosody has generally been categorized as a stress language at word level [4, 5, 6, 7] and consisting of accentual phrases with a single pitch accent and high or low boundary tones at sentence level [1, 3, 9, 11]. Under focus, Persian prosody changes dramatically with higher F_0 in on-focus elements and a significant decrease in post-focal words, hence a PFC (post-focus compression) language [11, 12, 13, 15].

2. PENTA MODEL

Drawing on an articulatory-functional view of speech, PENTA is a framework for linking communicative meanings to fine-grained prosodic details [14, 16]. From its inception, the model has been focused on two aspects of speech prosody, namely, communicative functions and articulatory mechanisms [14], each of which will be discussed briefly below.

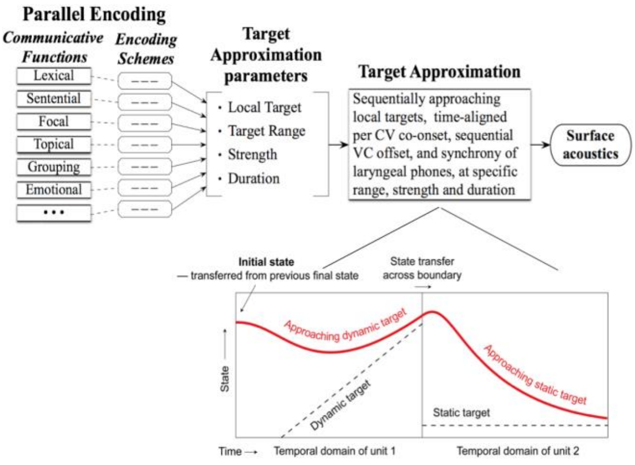
Parallel Encoding
Communicative Functions
Encoding Schemes
Lexical
Sentential
Focal
Topical
Grouping
Emotional
• • •
Target Approximation parameters
• Local Target
• Target Range
• Strength
• Duration
Target Approximation
Sequentially approaching local targets, time-aligned per CV co-onset, sequential VC offset, and synchrony of laryngeal phones, at specific range, strength and duration
Surface acoustics
Initial state — transferred from previous final state
State transfer across boundary
Approaching dynamic target
Approaching static target
Dynamic target
Static target
State
Time
Temporal domain of unit 1
Temporal domain of unit 2

Figure 1: A sketch of Parallel Encoding and Target Approximation (PENTA) model [14].

2.1. Communicative functions

In PENTA, a communicative function is a specific communicative meaning that the speaker intends to convey to the listener through speech prosody. As shown in Figure 1 the stacked boxes on the far left conceptualizes the individual functions as the driving force of the model. Each of these functions has a unique encoding scheme (the second stack of boxes from the left). These encoding schemes are composed of specifications of the pitch targets shown in the open box in the middle. These targets are then articulatorily implemented through target approximation, which ultimately generates the observed continuous F_0 contours, as shown in the two boxes on the right. The PENTA framework thus describes the generation of speech prosody as a process of encoding communicative functions through target approximation.

2.2. Target approximation

The lower part of Figure 1 illustrates the target approximation (TA) process proposed by [17]. The red solid curve is the F_0 contour that asymptotically approaches two successive pitch targets, one dynamic and the other static, represented by the dashed lines. The three grey vertical lines represent syllable boundaries. This conceptual model was later implemented as the quantitative Target Approximation (qTA) model in [10]. In qTA, the F_0 of each syllable is represented by the following third-order critically damped linear equation:

(1) $$f_0(t) = (mt + b)(c_1 + c_2t + c_3t^2)e^{-\lambda t}$$

where m and b denote the slope and height of the pitch targets, respectively, and λ represents the rate or strength of target approximation. In addition, the three transient coefficients in (1) are computed from the following formulae:

(2) $c_1 = f_0(0) - b$

(3) $c_2 = f_0'(0) + c_1\lambda - m$

(4) $c_3 = (f_0''(0) + 2c_2\lambda - c_1\lambda^2)/2$

qTA therefore uses three model parameters, m, b and λ, to control the F_0 trajectory of each syllable. Syllable is assumed as the basic prosody carrier in the PENTA model.

2.3. PENTAtrainers

The target parameters of qTA can be obtained in various ways. They can be specified arbitrarily (e.g., for purpose of a perception experiment), or obtained automatically through training on real speech data. To achieve automatic parameter estimation, two Praat-script-controlled [2] programs have been developed, PENTAtrainer1 [10] and PENTAtrainer2 [16]. Both trainers extract target parameters via analysis-by-synthesis, but they differ in the manner of optimization. PENTAtrainer1 performs an exhaustive search, i.e., testing all the possible targets within a range and then selecting the one that generates the best fit to the natural F_0 of each individual syllable. In this way, however, categorical targets corresponding to specific communicative functions could be obtained only by post-hoc averaging of targets belonging to the same categories [10]. In contrast, PENTAtrainer2 obtains optimal categorical targets directly by performing a global stochastic search over an entire corpus [16]. The present study used PENTAtrainer2 to extract categorical targets from a Persian corpus originally collected for a study of focus prosody.

3. METHOD

3.1. Corpus

The corpus was made of utterances originally collected for a study of the production and perception of focus in Persian [12, 13]. The target sentence is shown in Table 1. Five male speakers produced six different versions of this sentence in various focus conditions, which in total make up 6 foci x 5 repetitions x 5 speakers = 150 utterances.

Table 1: The target sentence of the experiment

W1	W2	W3	W4	W5
maha	baba-ye	nili-ro	lændæn	didim
we-PL	father-EZ	Nili-DO	London	see.PST-1PL

To compare the current perception results with those on natural utterances, we selected utterances from three of the speakers that were used in the perception experiment in [12]. The utterances of these speakers had the minimum, maximum and median standard deviations of F_0 of all the five speakers. In total, therefore, there were 6 foci x 5 repetitions x 3 speakers = 90 tokens.

3.2. Functional annotation and modeling

Following PENTA's assumption of parallel encoding of communicative functions, the corpus was annotated with three functional layers: stress (Stressed / Unstressed), syllable position (Initial / penultimate / Final) and focus condition (Pre-focus / On-focus / Post-focus), as shown in Figure 2. All syllable boundaries were marked manually. In addition, F_0 rectification was done manually with the help of the annotation tool to check the vocal cycles in the wave form.

After the annotation, the learn tool in PENTAtrainer2 was used to obtain all the multi-functional pitch targets (19 in total) which were then used by the synthesis tool to obtain the final synthesis results. The training and synthesis were performed in a speaker-independent manner. That is, the pitch targets were first extracted from each speaker, and then averaged across the speakers. The averaged targets were then used to perform synthesis on the utterances of all speakers.

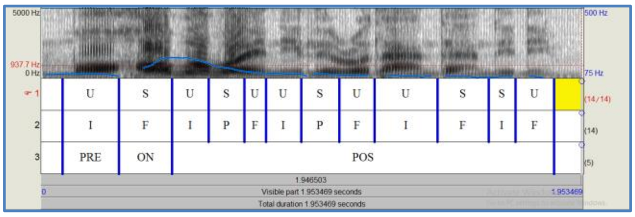

Figure 2: The layered annotation of communicative functions. The layers from top to down are stress, syllable position and focus condition. U and S denote unstressed and stressed, I, F and P denote initial, final and penultimate, PRE, ON and POS denote pre-focus, on-focus and post-focus, respectively.

3.3. Perceptual evaluation

Five males and five females with the same language background as those in [12] were recruited from an educational centre to perform the perception evaluation. They had no self-reported speech and hearing disorders, and they were paid for their participation.

ExperimentMFC in Praat was used to carry out two separate experiments on focus identification and naturalness judgement, respectively. Listeners were instructed to identify the focused word in one task and to judge whether the utterance was natural or synthetic in another task. They had an optional practice run before doing the main tests.

3.4. Results

Figure 3 shows the demo window of the synthesis tool in PENTAtrainer2 for an utterance by one of the speakers.

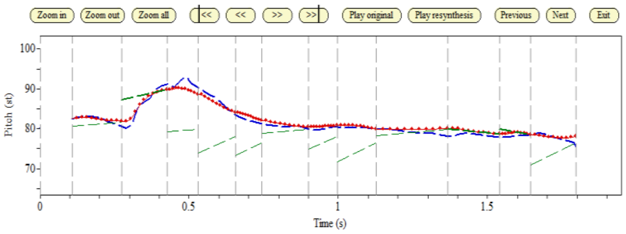

Figure 3: The demo window of the synthesis tool of PENTAtrainer2, showing the original contours of an utterances (dotted blue curve), the learned pitch targets (green dashed lines), and the synthetic contours (red dotted curve).

The blue curve, green line and the red trajectory are the original pitch contour, the learned pitch targets and the synthesized contour, respectively. Figure 4 shows examples of mean time-normalized synthetic and original F_0 contours.

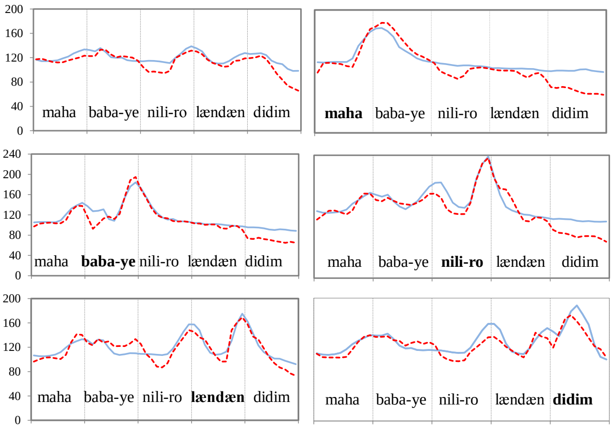

Figure 4: Mean time-normalized original (blue solid line) and synthetic (red dotted line) F_0 contours across five repetitions and three speakers. Bold-face indicates a focus location.

Table 2 shows the results of RMSE and Pearson's r for neutral-focus and focused utterances, which indicate the goodness of fit between the synthetic and original F_0. These values show very good synthesis performance, comparable to those reported in previous studies [8, 16], indicating that the reconstructed F_0 fits the original trajectory quite well.

Table 2: Average RMSE and Pearson's r correlation coefficients between two types of neutral focus and focused utterances.

Sentence Type	RMSE	Correlation
Neutral focus	1.62	0.76
Focus	2.01	0.86

Figure 5 shows the rate of focus identification in the listening experiment. The highest and lowest rate of focus recognition belongs to utterances with focus on the second and fifth (last) word, respectively. It is comparable to the results of [12] which were obtained with the same methodology from natural utterances.

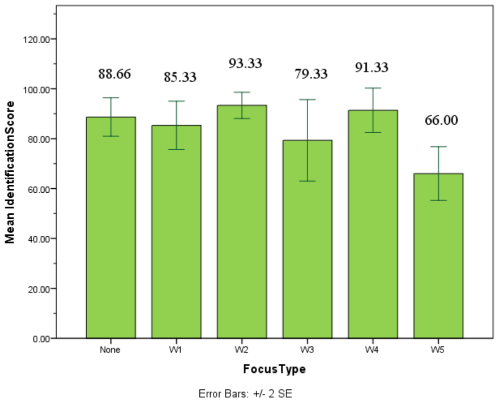

Figure 5: Percentage (numbers above the bars) of correct identification of neutral focus and focus on word 1-5. The error bars represent standard errors.

Table 3 shows the confusion matrix of focus perception. The main difference is between focus on word 5 and the other focus locations.

Table 3: Confusion matrix of focus perception of synthesized utterances (in percentage of identification). Bold face indicates correct focus identification.

heard as / original	none	W1	W2	W3	W4	W5
None	**88.6**	3.6	0.0	1.6	3.6	1.6
W1	12	**85.3**	2	0.0	0.0	0.0
W2	4.6	2	**93.3**	0.0	0.0	0.0
W3	13.3	1.3	3.3	**79.3**	2.6	0.0
W4	8.0	0.0	0.0	0.0	**91.3**	0.6
W5	28.6	0.6	0.0	0.0	5.0	**66.0**

Table 4 shows the results of post-hoc pairwise comparisons with Bonferroni adjustments. The only significant difference is between the focus on word 5 and focus on words 2 and 4. In other words, the rate of correct focus recognition for word 5 is significantly lower than the focus on words 2 and 4. This is also comparable to the results of the same experiment for natural utterances reported in [12].

Table 4: Results of post-hoc pairwise comparisons. The mean difference is significant at the .05 level.

Focus Type (I)	Focus Type (J)	Mean Difference (I-J)	Std. Error	Sig.
None	W1 W2 W3 W4 W5	5.333 -2.669 11.333 -.668 25.334	5.692 4.355 8.836 5.918 6.426	1.000 1.000 1.000 1.000 .051
W1	W2 W3 W4 W5	-8.002 6.000 -6.001 20.001	4.074 7.533 5.484 8.835	1.000 1.000 1.000 .748
W2	W3 W4 W5	14.002 2.001 28.003^*	6.399 3.151 6.111	.846 1.000 **.020**
W3	W4 W5	-12.001 14.001	4.423 8.343	.358 1.000
W4	W5	26.002^*	5.208	**.011**

The only difference between the two studies is that in [12] there was also a significant difference between focus on word 5 and neutral focus condition. However, it is interesting to see that the rate of focus identification for synthesized utterances was higher than the natural utterances in [12] for all focus conditions (84.0% vs. 75.5%).

Figure 6 shows the results of naturalness judgement. There is no significant difference in either neutral-focus [$F(1,9) = 0.87$, $p = 0.775$] or focused utterances [$F(1,9) = 2.969$, $p = 0.119$].

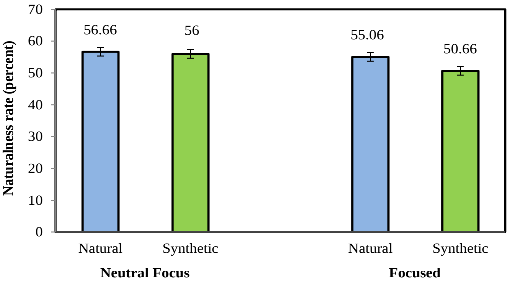

Figure 6: Means (bars and numbers above them) and standard errors (vertical lines) of naturalness evaluation of synthesized utterances in focus and neutral focus condition.

4. DISCUSSION AND CONCLUSION

The results reported above demonstrate that it is possible to achieve high quality synthetic F_0 in non-tonal languages like Persian with PENTAtrainer2, a semi-automatic software package for studying speech prosody which combines simulation of articulatory mechanisms of pitch production, functional annotation and stochastic optimization. Subjective and objective evaluation tests showed good results, which were comparable to previous ones on modeling Mandarin, Thai, English and Japanese [8, 16]. It is especially worth noting that perception of focus was better for synthetic prosody in this study than for natural prosody in [12].

PENTAtrainer2 was therefore found to be an effective tool for simulating focus prosody in Persian. For future research we would like to test this model with a large scale Persian database designed for a text-to-speech application to check the predictive power of this framework in speech synthesis systems.

5. REFERENCES

[1] Abolhasanizadeh, V., Bijankhan, M. & Gussenhoven, C. 2012. The Persian pitch accent and its retention after the focus. *Lingua* 122, 1380-1394.

[2] Boersma, P. & Weenink, D. 2001. Praat, a system for doing phonetics by computer. *Glot international*. 5, (2001), 341–345.

[3] Eslami, M. 2000. Šenaxt-e næva-ye goftar-e zæban-e farsi væ karbord-e an dær bazsazi væ bazšenasi-ye rayane'i-ye goftar [The prosody of the Persian language and its application in computer-aided speech recognition]. Unpublished PhD thesis, Tehran University.

[4] Ferguson, C. 1957. Word stress in Persian. *Language* 33, 123-135.

[5] Jun, S.-A. 2005. *Prosodic Typology: The Phonology of Intonation and Phrasing.* Oxford: Oxford University Press.

[6] Kahnemuyipour, A. 2003. Syntactic categories and Persian stress. *Natural Language & Linguistic Theory*, *21*(2), 333-379.

[7] Lazard, G. 1957. Grammaire du Persan Contemporain. Klincksieck, Paris, New Edition published by Peeters, Paris, 2006.

[8] Lee, A., Xu, Y., & Prom-on, S. 2014. Modeling Japanese F_0 contours using the PENTAtrainers and AMtrainer. In *Fourth International Symposium on Tonal Aspects of Languages*, Nijmegen, Netherlands. 2014.

[9] Mahjani, B. 2003. *An Instrumental Study of Prosodic Features and Intonation in Modern Farsi (Persian)*, M.Sc. thesis, retrieved from: http://www.ling.ed.ac.uk/teaching/postgrad/mscslp/archive/dissertations/2002-3/behzad_mahjani.pdf.

[10] Prom-on, S., Xu, Y. and Thipakorn, B. 2009. Modeling tone and intonation in Mandarin and English as a process of target approximation. *The Journal of the Acoustical Society of America*. 125, 1 (Jan. 2009), 405–424.

[11] Sadat-Tehrani, N. 2007. *Intonational Grammar of Persian*, doctoral dissertation. Manitoba: University of Manitoba.

[12] Taheri-Ardali, M., Rahmani, H. & Xu, Y. 2014. The perception of prosodic focus in Persian. In *Proceedings of Speech Prosody 2014,* Dublin: 515-519.

[13] Taheri-Ardali, M. & Xu, Y. 2012. Phonetic realization of prosodic focus in Persian. In *Proceedings of Speech Prosody 2012,* Shanghai: 326-329.

[14] Xu, Y. 2005. Speech melody as articulatorily implemented communicative functions. *Speech Communication* 46, 220–251.

[15] Xu, Y. 2011. Post-focus compression: Cross-linguistic distribution and historical origin. In *Proceedings of 17th International Congress of Phonetic Sciences*, Hong Kong.

[16] Xu, Y. & Prom-on, S. 2014. Toward invariant functional representations of variable surface fundamental frequency contours: Synthesizing speech melody via model-based stochastic learning. *Speech Communication* 57, 181–208.

[17] Xu, Y., Wang, Q. E. 2001. Pitch targets and their realization: Evidence from Mandarin Chinese. *Speech Communication* 33: 319-337.