ESTIMATING VOCAL TRACT SHAPES OF THAI VOWELS FROM CONTEXTUAL TONAL VARIATION

Santitham Prom-on[1,2]*, Peter Birkholz*[3]*, Yi Xu*[2]

[1] Department of Computer Engineering, King Mongkut's University of Technology Thonburi, Thailand
[2] Department of Speech, Hearing and Phonetic Sciences, University College London, United Kingdom
[3] Institute of Acoustics and Speech Communication, Technische Universität Dresden, Germany
santitham@cpe.kmutt.ac.th, peter.birkholz@tu-dresden.de, yi.xu@ucl.ac.uk

ABSTRACT

This paper presents a computational estimation of vocal tract shape parameters as articulatory targets of Thai vowels in an articulatory synthesizer, by means of analysis-by-synthesis with acoustic data as input. A speech corpus designed to capture the contextual variation of nine Thai long vowels, consisting of 81 disyllabic utterances, was recorded by a native Thai speaker. For each utterance, two targets, one for each syllable, were estimated by optimizing the target parameters to minimize the MFCC error between original and synthesized speech. Stochastic gradient descent was used to iteratively optimize the shape parameters. The estimated targets of each vowel type were then averaged, resulting in nine articulatory targets, each corresponding to a vowel. The optimized targets were then used to synthesize Thai vowels both in monosyllables and in disyllabic sequences. The results, both numerically and perceptually, indicate that the estimated targets effectively represent the underlying articulatory goals of Thai vowels.

Index Terms— Articulatory synthesis, vocal tract shape, Thai vowel, optimization, target approximation

1. INTRODUCTION

In speech acquisition, acoustics of speech utterances is the definitive input for children, which seems to enable them to acquire highly intricate speaking skills without detailed articulatory instructions, and without direct observation of the articulators of the mature speakers other than the visible ones such as the lips and the jaw. Understanding how these intricate movements can be learned from acoustic data is thus the important step toward our understanding of the process of speech acquisition, and probably the basic mechanism of speech production.

To computationally learn vocal tract configurations from acoustic data, different methods have been previously proposed, based on either mapping [1-8] or optimization [9-10] strategies. Mapping strategies follow either probabilistic approaches such as hidden Markov models [1,2] or neural network [3], or codebook-based approaches [6-8]. Except those that are based on the task-dynamic (TD) model [4,5], these mapping approaches share the common drawback of not simulating the dynamic movement of speech gestures [11,12] that results in smooth spectral transitions in the acoustic data. In the optimization strategy, parameters of a synthesis model are iteratively adjusted to minimize a cost function [9,10]. The cost function can be the error from acoustic comparison between the original speech and speech synthesized with the optimized parameters. This strategy, when implemented with an articulatory synthesizer with sufficient capacity to generate acoustic data from model parameters, may have the potential to achieve the closest simulation of speech learning behavior.

The mapping studies that are based on the TD model [4,5] do take dynamic gestural control into consideration. They rely on neural network [4], or discrete-time warping [5] to perform estimation. The TD model provides a mechanism for generating movements of tract variables. It uses a critically-damped second-order system to describe the movement. In TD, gestural movements are assumed to be completed and adjacent gesture movements are assumed to be overlapped.

The present study adopts a strategy that combines optimization with consideration of articulatory dynamics. The dynamic model used is the Target Approximation (TA). TA differs from TD in that it does not assume that targets are always reached, and it allows remaining momentum at the end of a target approximation movement to be transferred to the next interval as its initial conditions. This strategy has been implemented in our recent work [16] on training a TA-based articulatory synthesizer with acoustic data to model vowels. The current study is an extension of this work. Here we attempt to identify underlying articulatory targets of Thai vowels by means of model-based optimization using the default vocal tract anatomy provided by VocalTractLab [17]. Thai has nine static vowels, in short and long minimal pairs, which are evenly spread across the vowel space [18]. This makes them ideal cases for testing the idea of target estimation for vowels. The estimated articulatory targets of Thai vowel are evaluated (a) numerically by comparing the formant trajectories of the synthetic vowels to those of natural utterances and (b)

perceptually by a listening experiment that compares the perceptual accuracy and naturalness of synthetic and natural speech.

2. METHOD

2.1. Corpus

The corpus was designed to have full contextual variations in Thai vowels. The sentences are composed of two syllables consisting of only vowels, in the form of /V1 V2/, where both V1 and V2 are one of the nine long vowels (/aː/, /iː/, /uː/, /eː/, /ɛː/, /ɯː/, /ɤː/, /oː/, /ɔː/). Thus there are 81 disyllabic sequences in total.

A longitudinal design is often used in acoustic-to-articulation studies [1-10], thus only few subjects were used. This is because the estimated vocal tract shape represents only an individual. In this study, speech data were recorded from a native male Thai speaker who had been living in the Greater Bangkok region in the past 20 years and had no self-reported speech or hearing disabilities. Recordings were done in a sound-treated room at the King Mongkut's University of Technology Thonburi, Bangkok, Thailand. The speaker was instructed to produce the disyllabic vowel sequences in a continuous manner with the mid tone on both syllables. No further instruction on the stress placement was given, so stress was placed on the second syllable according to the general rule of Thai pronunciation. The utterances were recorded at a sampling rate of 22.05 kHz and 16-bit resolution.

2.2. Annotation

The corpus was annotated using Praat [19]. Syllable boundaries were manually marked according to the concept of target approximation (TA) [14-16,20]. That is, the articulation of a segment is considered as a unidirectional movement toward its underlying target. Therefore the boundary between two vowels should be marked at the point where the spectrogram starts to change toward the target of the next vowel, as shown in Figure 1. This strategy, which was also used in our previous studies [14-16,20-22], differs from the conventional segmentation of demarcating a vowel by its steady-state interval [23]. Since the syllables contained only vowels, no consonantal boundaries were marked.

2.3. Optimization of Articulatory Targets

The optimization procedure is based on our previous study [16]. The basic idea is to estimate the underlying articulatory targets with an analysis-by-synthesis approach, in which the articulatory synthesizer is used to repeatedly generate acoustic data that can be compared to the original

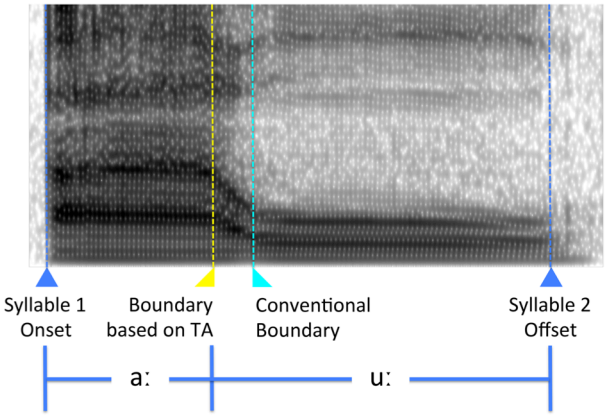

Figure 1: *Annotation scheme based on TA framework.*

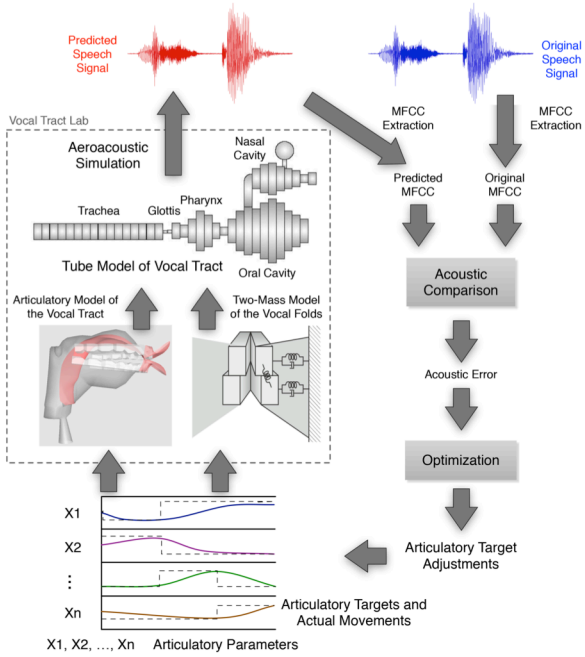

Figure 2: *Scheme for the optimization of articulatory targets* [16].

speech. The optimization process, as shown in Figure 2, encapsulates the synthesizer in an analysis-by-synthesis loop. For each vowel, the vocal tract shape is initialized with a neutral configuration, and then iteratively adjusted until the overall acoustic error is reduced to below a given threshold or until the maximum number of steps is reached.

The articulatory synthesizer used is VocalTractLab 2.1 [17], which is a 3D articulatory synthesizer capable of generating a full range of speech sounds by controlling vocal tract shapes, aerodynamics and voice quality [24-26]. VocalTractLab simulates the acoustic signal by approximating the trachea, the glottis, and the vocal tract as

a series of cylindrical tube sections with variable lengths as shown in Figure 2. The aerodynamic-acoustic simulation is based on a transmission-line circuit model of the tube system [27].

The objective of optimization is to estimate for each vowel the underlying articulatory target parameters, as shown in Table 1, that define the vocal tract configuration. Beside these positional parameters, VocalTractLab also requires the specification of a time constant parameter (τ) for each interval. In the TA framework [14,15,25], this time constant value determines how fast the articulatory target is approached. The adjustment of τ would directly affect the rate of articulatory movement and in turn affect the rate of formant change. The modeling process therefore needs to learn, for each vowel, a vocal tract shape associated with 18 articulatory parameters, as shown in Table 1, and the time constant τ that defines the transition time from the first to the second vowel. Because each utterance consists of two vowels, 37 parameters need to be learned in total in each simulation run.

Table 1. *Articulatory targets of each TA movement* [24].

Parameter	Description
HX, HY	Horz. and vert. hyoid positions
JX, JA	Horz. jaw position and jaw angle
LP, LD	Lip protrusion and vert. lip distance
VS, VO	Velum shape and velum opening
TTX, TTY	Horz. and vert. tongue tip positions
TBX, TBY	Horz. and vert. tongue blade positions
TCX, TCY	Horz. and vert. tongue center positions
TS1 – TS4	Tongue side elevation at 4 positions

Mathematically, VocalTractLab models articulatory trajectories as responses of target-driven sixth-order critically damped linear dynamic system [25]. The input to this model is a sequence of articulatory targets. For each syllable, VocalTractLab generates articulatory movement of all articulatory parameters according to the initial articulatory dynamic condition of each parameter, the given target and the time constant. This process significantly reduces the degrees of freedom of the optimization problem, as for each syllable only one set of articulatory parameters needs to be optimized. Also, based on sequential target approximation, no gestural overlap is assumed as far as any particular articulator is concerned. A target approximation movement does not start until the previous one is over. To ensure the smoothness of the articulatory movement at the interval boundary, up to sixth-order dynamic states are transferred from the end of the preceding syllable to the beginning of the following syllable. Figure 3 illustrates the movements of articulators according to the target approximation model.

The optimization process uses a stochastic gradient descent algorithm. In the process, the articulatory parameters are initially set to neutral positions and the error is computed as the sum of square errors of Mel-Frequency Cepstral Coefficients (MFCCs) between original and synthesized data. Then, in each iteration, parameters are randomly adjusted one-by-one and used to generate the synthetic utterance that is compared with the original data. Any adjustment that results in a worse error is rejected. This process repeats until the error converges or the maximum number of iterations is reached.

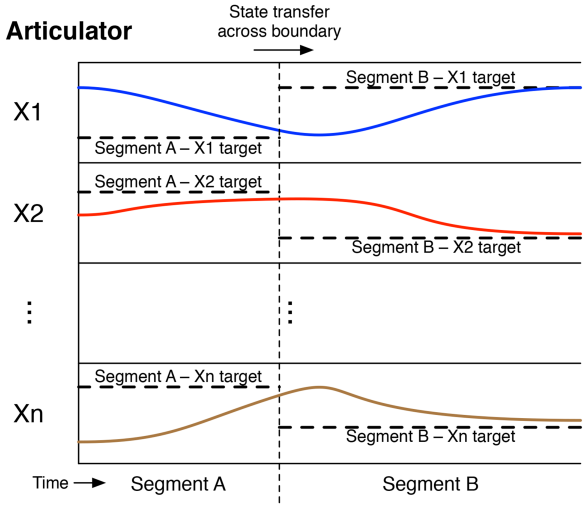

Figure 3: *An illustration of articulatory movements based on the target approximation model.*

Table 2. *Mean formant RMSE in percent of the vowels. Errors were averaged over all vowel instances in both syllables.*

Vowel	Formant RMSE (%)		
	F1	F2	F3
/aː/	9.3	4.1	6.1
/iː/	7.5	4.6	7.7
/uː/	5.9	8.2	7.5
/eː/	3.6	5.1	4.4
/ɛː/	8.8	7.2	4.5
/ɯː/	7.2	5.8	4.7
/ɤː/	4.6	3.6	3.1
/oː/	6.2	4.3	8.2
/ɔː/	7.6	3.4	6.2

3. RESULTS

3.1. Numerical Assessment

After the optimization, the estimated targets were used to resynthesize speech utterances. The accuracy of the estimated targets was assessed by comparing the formant tracks (F1-F3) of synthesized and original utterances, using FormantPro [28], a Praat script for large-scale systematic analysis of continuous formant movements. Root Mean Square Error (RMSE) of both syllables was calculated, as shown in Table 2. Relatively low RMSEs can be observed for all vowels compared to the average RMSE levels reported in the previous study [29]. This indicates that the estimated articulatory parameters effectively represent vocal tract shape of each vowel and are able to accurately generate the acoustic patterns that are close to the natural ones.

3.2. Graphical Comparison

Figure 4 shows examples of formant contour comparisons obtained with FormantPro [28]. Note that each vowel is annotated to terminate at a point where its target is best achieved, so that the formants in each segment move unidirectionally toward an ideal pattern. Smooth formant transitions from one vowel to another can be observed in the synthetic utterances (solid lines in Figure 4), just as in the natural utterances (dotted lines in Figure 4). Visual inspection of spectrograms of all other cases also confirmed the accuracy of the formant patterns generated by the estimated vocal tract shapes. The smooth synthetic formant movements are thanks to the TA dynamics of all the articulators involved. Note that while there are certain mismatches, for example in F3 of /u:/ in /u:e:/ as shown in Figure 4, between the original and synthesized formant frequencies, these mismatches are evened out once they are averaged together with other cases (e.g. for /u:/, /a:u:/ and /u:ɤ:/ as shown in Figure 4).

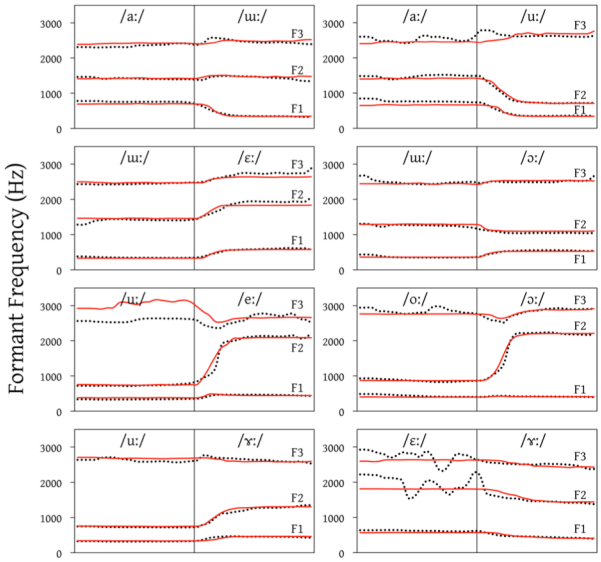

Figure 4: *Time-normalized comparisons of formant movements of example original (dotted black lines) and synthetic (solid red lines) utterances.*

3.3. Perceptual Evaluation

We further assessed the quality of the synthetic Thai vowels by a listening experiment. Target parameters of the same vowel were averaged together across multiple contexts as the underlying representation of that vowel. They were used to synthesize monosyllabic words made of each individual vowel. The natural stimuli of the same words were recorded as references by the same speaker used in the training corpus at a sampling rate of 22.05 kHz and 16-bit resolution.

Twenty native Thai listeners participated in the listening experiment, which was conducted with the ExperimentMFC function of Praat. All natural and synthetic vowels were presented to the listeners in randomized order over earphones, and for each item, the listeners were asked

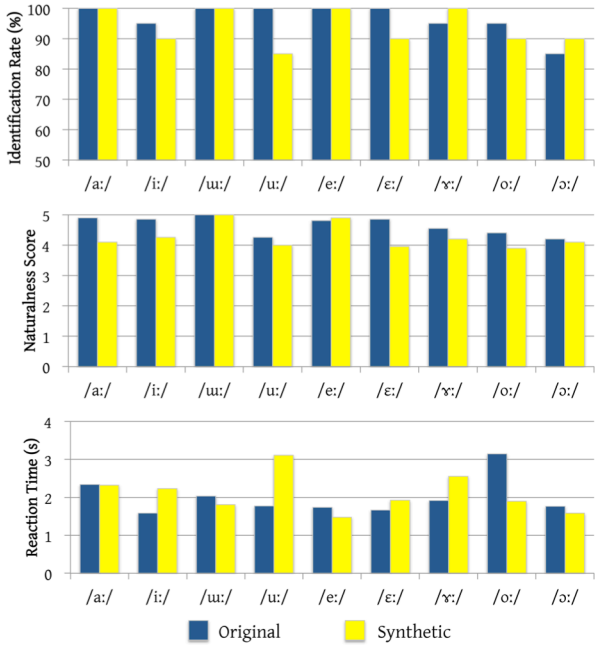

Figure 5: *Mean vowel identification rate, naturalness score and listener reaction time of each vowel for both original and synthetic utterances.*

to identify the presented vowel in a forced choice manner and judge its naturalness on a five-point Likert scale (1=very unnatural; 5=very natural). Furthermore, the reaction time was measured for each rating. Figure 5 shows the results of the experiment. Listeners could identify

vowels equally well for both natural and synthetic vowels ($t(8) = 1.25$, $p = 0.247$). The lowest identification rate of synthetic vowels is that of /u:/ which was perceived by three listeners as /o:/, while the perception of the natural /u:/ was perfect. The reaction time, which indicates the cognitive load, is also higher for the synthetic /u:/. This indicates the artifact in synthesis and possibly due to the continuous nature of the estimated velum parameters. This makes it difficult for an optimization to fully close the velo-pharyngeal port in /u:/. Listeners identified both the natural and synthetic vowels in roughly the same score ranges, 4.2-5 for natural stimuli and 3.9-5 for synthetic stimuli. This result indicates that the present method could generate close-to-natural vowels with underlying articulatory targets learned from natural speech.

4. DISCUSSION AND CONCLUSIONS

This study explored the estimation of articulatory targets of Thai vowels using a model-based analysis-by-synthesis strategy. The results show that it is possible to estimate the underlying articulatory targets of vowels using such a strategy with surface acoustics of continuous speech and TA-based segmentations as the input. The numerical assessment as shown in Table 2 and the visual impression as shown in Figure 4 indicate that the learned targets can be used to consistently synthesize acoustic data that closely approximate those of the natural utterances. The perceptual evaluation shows that underlying articulatory targets learned this way can be used to effectively generate isolated vowels that are perceptually close to their natural counterparts, as shown in Figure 5. All these results indicate that the estimated articulatory parameters closely represent the underlying targets of the Thai vowels.

A further development of the framework for organizing trained targets is still needed. Strategies have yet to be developed to simulate the learning of overlapped CV gestures. The incorporation of a timing model in the articulatory synthesis is also needed, as timing specifications of the segments are required prior to the generation process. The incorporation of the visible articulatory data (e.g. lip and jaw movement) into the process is also required to fully emulate actual speech acquisition.

5. ACKNOWLEDGEMENT

We would like to thank the Royal Academy of Engineering (UK) for financial support through the Newton International Fellowship Alumni follow-on funding, and the Thai Research Fund (Thailand) through the Research Grant for

10. REFERENCES

[1] Hofer, G., Yamagishi, J. and Shimodaira, H., "Speech-driven lip motion generation with a trajectory HMM", Proc. Interspeech 2008, Brisbane, Australia, pp. 2314–2317, 2008.

[2] Tamura, M., Kondo, S., Masuko, T. and Kobayashi, T, "Text-to-visual speech synthesis based on parameter generation from HMM", Proc. ICASSP 98, Seattle, WA, pp. 3745–3748, 1998.

[3] Uria, B., Renal, S. and Richmond, K., "A deep neural network for acoustic-articulatory speech inversion", Proc. NIPS 2011 Workshop on Deep Learning and Unsupervised Feature Learning, Sierra Nevada, Spain, 2011.

[4] Mitra, V., Nam, H., Espy-Wilson, C.Y., Saltzman, E. and Goldstein, L., "Retrieve tract variables from acoustics: a comparison of different machine learning strategies", IEEE J. Sel. Topics Signal Process. 4(6): 1027-1045, 2010.

[5] Nam, H., Mitra, V., Tiede, M., Hasegawa-Johnson, M., Espy-Wilson, C., Saltzman, E. and Goldstein, L., "A procedure for estimating gestural scores from speech acoustics", J. Acoust. Soc. Am. 132(6): 3980-3989, 2012.

[6] Schroeter, J. and Sondhi, M.M., Dynamic programming search of articulatory codebooks. in Proceedings of the 1989 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 1989), Glasgow, UK, May 22-25, 1989, vol. 1, pp. 588–591.

[7] Ouni, S. and Laprie, Y., "Modeling the articulatory space using a hypercube codebook for acoustic-to-articulatory inversion", J. Acoust. Soc. Am. 118(1): 444-460, 2005.

[8] Potard, B., Laprie, Y. and Ouni, S., "Incorporation of phonetic constraints in acoustic-to-articulatory inversion", J. Acoust. Soc. Am. 123(4): 2310-2323, 2008.

[9] Panchapagesan, S. and Alwan, A., "A study of acoustic-to-articulatory inversion of speech by analysis-by-synthesis using chain matrices and the Maeda articulatory model", J. Acoust. Soc. Am. 129(4): 2144-2162, 2011.

[10] McGowan, R., "Recovering articulatory movement from formant frequency trajectories using task dynamics and a genetic algorithm: Preliminary model test", Speech Commun. 14: 19-48, 1994.

[11] Mermelstein, P., "Articulatory model for the study of speech production", J. Acoust. Soc. Am. 53(4): 1070-1082, 1973.

[12] Saltzman, E.L. and Munhall, K.G., "A dynamical approach to gestural patterning in speech production", Ecol. Psychol. 1: 333-382, 1989.

[13] Hanson, H.M. and Steven, K.N., "A quasiarticulatory approach to controlling acoustic source parameters in a Klatt-type formant synthesizer using HLsyn", J. Acoust. Soc. Am. 112(3): 1158-1182, 2002.

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

[15] Prom-on, S., Thipakorn, B. and Xu, Y., "Modeling tone and intonation in Mandarin and English as a process of target approximation", J. Acoust. Soc. Am. 125(1): 405-424, 2009.

[16] Prom-on, S., Birkholz, P. and Xu, Y., "Training an articulatory synthesizer with continuous acoustic data", Proc. Interspeech 2013, Lyon, France, pp. 349-353, 2013.

[17] Birkholz, P., VocalTractLab 2.1 for Windows. Online: http://www.vocaltractlab.de, accessed 17 December 2013

[18] Tingsabadh, K. and Abhramson, A.S., "Thai", J. Int. Phon. Assoc. 22(1): 24-48, 1993.

[19] Boersma, P., "Praat, a system for doing phonetics by computer", Glot International 5(9/10): 314–345, 2001

[20] Xu, Y. and Prom-on, S., "Toward invariant functional representations of variable surface fundamental frequency contours: Synthesizing speech melody via model-based stochastic learning", Speech Commun. 57: 181-208, 2014.

[21] Prom-on, S., Liu, F. and Xu, Y., "Post-low bouncing in Mandarin Chinese: Acoustic analysis and computational modeling", J. Acoust. Soc. Am. 132: 421-432, 2012.

[22] Xu, Y. and Liu, F., “Determining the temporal interval of segments with the help of F0 contours” J. Phon. 35: 398-420, 2007.
[23] Lehiste, I. and Peterson, G.E., “Transitions, glides and diphthongs,” J. Acoust. Soc. Am. 33: 268-277, 1961.
[24] Birkholz, P., “Modeling consonant-vowel coarticulation for articulatory speech synthesis”, PLOS ONE 8(4): e60603, 2013.
[25] Birkholz, P., Kröger, B.J. and Neuschaefer-Rube, C., “Model-based reproduction of articulatory trajectories for consonantal-vowel sequences”, IEEE Audio, Speech and Lang. Process. 19(5): 1422-1`433, 2011.
[26] Birkholz, P., Kröger, B.J. and Neuschaefer-Rube, C., “Synthesis of breathy, normal, and pressed phonation using a two-mass model with a triangular glottis”, Proc. Interspeech 2011, Florence, Italy, pp. 2681-2684, 2011.
[27] Birkholz, P., Jackèl, D. and Kröger, B.J., “Simulation of losses due to turbulence in the time-varying vocal system”, IEEE Audio, Speech and Lang. Process. 15(4): 1218-1226, 2007.
[28] Xu Y., FormantPro Version 1.1, Online: http://www.phon.ucl.ac.uk/home/yi/FormantPro, accessed 24 December 2013
[29] McGowan, R.S., Berger, M.A., “Acoustic-articulatory mapping in vowels by locally weighted regression”, J. Acoust. Soc. Am. 126(4): 2011-2032, 2009.