EURASIP Journal on
Audio, Speech, and Music Processing
a SpringerOpen Journal

RESEARCH Open Access

Identifying underlying articulatory targets of Thai vowels from acoustic data based on an analysis-by-synthesis approach

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

Abstract

This paper investigates the estimation of underlying articulatory targets of Thai vowels as invariant representation of vocal tract shapes by means of analysis-by-synthesis based on acoustic data. The basic idea is to simulate the process of learning speech production as a distal learning task, with acoustic signals of natural utterances in the form of Mel-frequency cepstral coefficients (MFCCs) as input, VocalTractLab - a 3D articulatory synthesizer controlled by target approximation models as the learner, and stochastic gradient descent as the target training method. To test the effectiveness of this approach, a speech corpus was designed to contain contextual variations of Thai vowels by juxtaposing nine Thai long vowels in two-syllable sequences. A speech corpus consisting of 81 disyllabic utterances was recorded from a native Thai speaker. Nine vocal tract shapes, each corresponding to a vowel, were estimated by optimizing the vocal tract shape parameters of each vowel to minimize the sum of square error of MFCCs between original and synthesized speech. The stochastic gradient descent algorithm was used to iteratively optimize the shape parameters. The optimized vocal tract shapes were then used to synthesize Thai vowels both in monosyllables and in disyllabic sequences. The results, both numerically and perceptually, indicate that this model-based analysis strategy allows us to effectively and economically estimate the vocal tract shapes to synthesize accurate Thai vowels as well as smooth formant transitions between adjacent vowels.

Keywords: Articulatory target; Articulatory synthesis; Target approximation; Acoustic-to-articulatory inversion; Thai vowels

1. Introduction

Speaking requires an accurate control of highly variable successive articulatory movements, each involving simultaneous actions of multiple articulators [1,2], and all of them coordinated in such a way that many layers of meanings are simultaneously encoded [3]. Even more intriguingly, a human child seems to be able to acquire such highly intricate motor skills without specific articulatory instructions and without direct observation of the articulators of the mature speakers other than the visible ones such as the lips. The only definitive input the child receives that is articulatorily (as opposed to meaning-wise) informative is the acoustics of the speech utterances. Understanding how proper articulatory skills can be learned from acoustic data, a task known as acoustic-to-articulatory inversion, is therefore the key to our understanding of the nature of human speech acquisition and production. Such knowledge is also beneficial to both speech recognition [4] and speech synthesis [5].

Different approaches have been proposed to achieve acoustic-to-articulatory inversion [6-16]. These methods rely on either explicit mapping between acoustic and articulatory data [6-14] or optimization of articulatory synthesis model parameters [15,16]. Different methods of mapping between articulatory and acoustic data have been tested using probabilistic models such as hidden Markov models (HMM) [6,7], neural networks [8,9], codebooks [10-13], or filters [14]. Except the methods that are based on the task dynamic (TD) model, most of them, however, share the common drawback of the

* Correspondence: santitham@cpe.kmutt.ac.th
[1]Department of Computer Engineering, Faculty of Engineering, King Mongkut's University of Technology Thonburi, Bangkok 10140, Thailand
[3]Department of Speech, Hearing and Phonetic Sciences, University College London, London WC1N 1PF, UK
Full list of author information is available at the end of the article

mapping paradigm, i.e., the lack of the inclusion of speech production mechanism in the modeling process, in particular, the dynamic movement of speech gestures [1,2] that results in smooth spectral transitions observed in the natural acoustic data. An alternative approach is to use an analysis-by-synthesis strategy [15,16] in which parameters of a synthesis model are iteratively adjusted to minimize a cost function. The cost function can be the error from acoustic comparison between the original speech and speech synthesized with 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.

This paper reports the results of a study based on this alternative approach. The study attempts to identify underlying articulatory targets of Thai vowels by means of model-based optimization. Using the analysis-by-synthesis strategy, the underlying articulatory targets representing the vocal tract shapes are estimated from coarticulated disyllabic vowels. This modeling process iteratively adjusts the articulatory parameters to the optimal condition by minimizing the acoustical error between original and synthesized sounds generated with the tentative vocal tract parameters. The accuracy of these estimated vocal tract shapes are then evaluated by comparing the formants of the synthetic vowels to the formant trajectories of the natural utterances and to those of previous studies and by a listening experiment that compared the perceptual accuracy and naturalness of synthetic to those of natural speech.

2. Methods

2.1 Corpus

The corpus was designed to have full contextual vowel variations in Thai to facilitate the modeling process. Thai has nine vowels, in short and long minimal pairs, which are evenly spread across the vowel space [17]. To estimate the articulatory targets of all the Thai vowels, each utterance was designed to have two syllables consisting of only vowels, in the form of /V1 V2/, where both V1 and V2 are one of the nine long vowels (/a:/, /i:/, /u:/, /e:/, /ɛ:/, /ɯ:/, /ɤ:/, /o:/, /ɔ:/). Thus, there are 81 combinations in total. These disyllabic vowel sequences do not have any meanings. This design allows us to fully study the spectral changes resulting from transitions between vowels and to simulate their dynamics through computational modeling.

Speech data were recorded by a native Thai male 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 tones for both syllables and without pauses between vowels as if they were in a simple noun-verb sentence. This makes the stress placed on the second syllable according to the general rule of Thai pronunciation. No particular normalization was done to remove the effect of stress. Nevertheless, the full factorial design of the corpus balances the occurrence of tones in both positions. The utterances were recorded at a sampling rate of 22.05 kHz and 16-b resolution.

The corpus was annotated using Praat [18]. Syllable boundaries were manually marked according to the concept of target approximation to be detailed later in Section 2.4. Briefly, the articulation of a segment is defined as a unidirectional movement toward its underlying target [19]. As a result, the moment a movement starts to turn away from the segmental target is viewed as the offset of one segment and onset of the next. Therefore, the boundary between two syllables was marked at the point where the spectrogram starts to change. This strategy, which was also used in our previous studies [20-23], differs from the conventional marking of syllable boundaries (*cf* [24] for evidence from production). Since the syllables contained only vowels, no consonantal boundaries were marked.

2.2 Overview of analysis-by-synthesis strategy for target estimation

Figure 1 shows a workflow diagram for the method used in this study. The basic idea, as mentioned in the 'Introduction,' is to estimate the underlying articulatory targets based on a distal learning strategy [25], in such a way that the learner (optimization system) is able to utilize the speech production system (articulatory synthesizer) to generate acoustic data that can be compared to the original speech. Vocal tract shapes as articulatory vowel targets are estimated for each utterance, starting from neutral positions and then iteratively adjusted until the overall acoustic error converges or the maximum number of steps is reached.

2.3 VocalTractLab, the articulatory synthesizer

The core of the analysis-by-synthesis strategy in this paper is VocalTractLab [26] - an articulatory synthesizer that can generate acoustic output based on articulatory parameters. VocalTractLab is capable of generating a full range of speech sounds by controlling vocal tract shapes, aerodynamics, and voice quality [27-29]. It consists of a detailed 3D model of the vocal tract that can be configured to fit the anatomy of any specific speaker, an advanced self-oscillating model of the vocal folds, and an efficient method for aeroacoustic simulation of the speech signal.

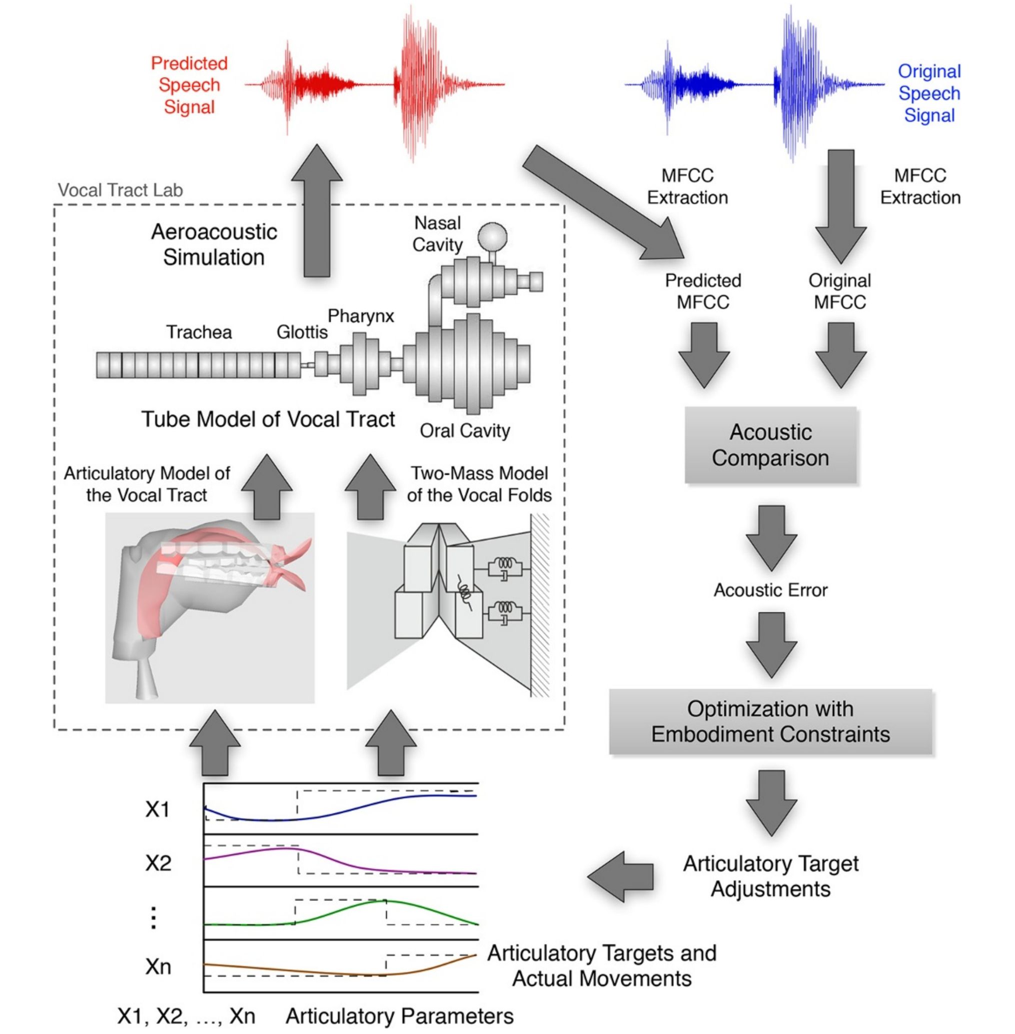

Figure 1 A schematic diagram of the articulatory target estimation.

The acoustic simulation of VocalTractLab approximates the trachea, the glottis, and the vocal tract as a series of cylindrical tube sections with variable lengths as shown in Figure 1. The aeroacoustic simulation is based on a transmission-line circuit representation of the tube system (see [30] for a detailed derivation). The simulation considers fluid dynamic losses at constrictions, as well as losses due to radiation, soft walls, and viscous friction.

2.4 Target approximation model

The control of the dynamics of VocalTractLab is based on the concept of sequential target approximation (TA), which has previously been implemented in various forms, as reviewed in [28]. The TA model implemented in VocalTractLab is illustrated in Figure 2. Compared to other articulatory models, TA is most similar to the task dynamic model [2], but with critical differences in several respects. Like the task dynamic model, TA simulates continuous articulatory trajectories as asymptotic movements toward underlying targets. Unlike the task dynamic model, however, TA does not assume that the target is always reached at the end of each target approximation interval. The unfinished target approximation movement thus often results in a highly dynamic articulatory trajectory that needs to be transferred to the next target approximation interval as its initial state. Such a transfer, as can be seen around the vertical dashed line in Figure 2, guarantees smooth and continuous articulatory movements across the TA boundaries.

Mathematically, VocalTractLab models articulatory trajectories as responses of target-driven high-order critically damped linear dynamic system [28]. The input to the system is a sequence of articulatory target positions. The dynamic system has another parameter, the time constant τ that controls the rate of target approximation. The basic

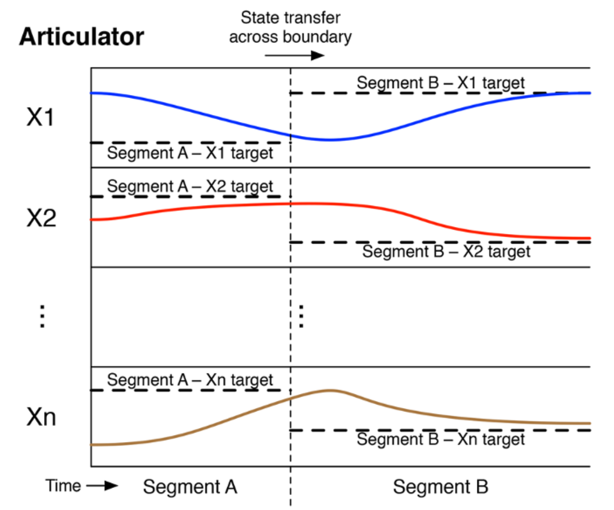

Figure 2 Target approximation model for controlling articulatory movements.

idea is to describe the dynamics of an articulator by a cascade of several identical first-order linear systems. The transfer function of such a system is

$$H(s) = \frac{Y(s)}{X(s)} = \frac{1}{(1 + s\tau)^N} \tag{1}$$

where N is the order of the system and s is the complex frequency. The higher N, the more bell-shaped is the velocity profile associated with a step response of the system, a shape which is typically found in human target-directed movements [28]. However, with increasing order, the delay between input (target) and output (action) also increases. In VocalTractLab, sixth-order systems are used as a compromise.

The time-domain representation of Equation 1 can be derived using the inverse Laplace transform [28], which results in

$$y(t) = \left(c_0 + c_1 t + \cdots + c_{N-1} t^{N-1}\right) e^{-t/\tau} + x(t) \tag{2}$$

where for a static target, $x(t) = b$ is the position of the articulatory target. The time t is relative to the onset of the target interval. The coefficients c_i are calculated based on the initial conditions at the onset of the target [28], as can be shown by the following equation:

$$c_i = \begin{cases} y(0) - b & n = 0 \\ \left(y^{(n)}(0) - \sum_{i=0}^{n-1} c_i a^{n-i} \binom{n}{i} i!\right) / n! & 0 < n < N \end{cases} \tag{3}$$

Also, as a result of this purely 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. Any seeming overlap between adjacent movements is simulated by the combined effect of cross-boundary state transfer and the articulation strength of the later movement that determines how quickly it can eliminate the carryover effect exerted by the transferred state.

A key advantage of TA is that it allows the mapping of variant surface trajectories due to phonetic context, stress, speech rate, etc. to a single invariant target as demonstrated in our previous studies in prosody modeling [3,20-23,31]. TA simplifies the problem of inverse mapping from acoustics to underlying articulatory targets. This can be achieved because of the clear separation of the transient and the steady-state responses in the TA representation of the articulatory movements. The estimation of articulatory targets underlying the movements due to these factors allows us to capture the trend of the variability, which can be then summarized into a single contextually invariant target by the analysis of parameter distribution [21,22]. This approach of using an invariant target in inverse mapping is different from the Directions Into Velocities of Articulator (DIVA) framework that defines articulatory targets as regions [32,33]. The DIVA framework relies on a neural network to map the associations between acoustic and articulatory data. In this sense, DIVA is largely a mapping method. In the TA framework, there is only one invariant articulatory target corresponding to a specific functional condition. The contextual variability due to the transition from one target to another is modeled as a transient response which is a by-product of the transition. This means that for each phonetic unit a single target or a single compound target can be learned from its many context-sensitive realizations. The feasibility of this approach has been seen in our recent work on F_0 modeling [21-23].

Another critical strategy in the present implementation of the TA model is a novel segmentation method, that is, a segmental interval is defined as the time period during which its canonical pattern is unidirectionally approached [19]. As a result, the point where a segment best approximates its canonical pattern is marked as its offset rather than onset or center, as shown in Figure 3.

Table 1 shows the list of articulatory target parameters that define vocal tract configurations. Beside these positional parameters, VocalTractLab also requires the specification of articulatory strength for each TA movement. In the target approximation framework [21,22,28,31], this strength determines how fast the articulatory target is approached. The adjustment of articulatory strength would directly affect the rate of articulatory movement and indirectly affect the rate of formant changes in the spectrograms. 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 a strength

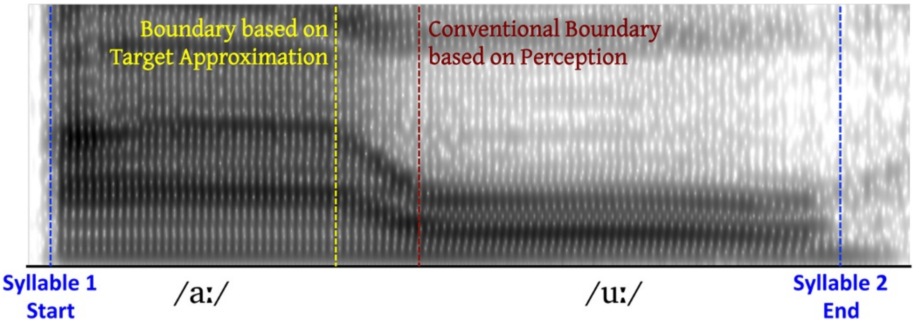

Figure 3 **Syllable boundary marking based on target approximation framework comparing to that from perception.**

parameter. Because each utterance consists of two vowels, 38 parameters need to be learned in total in each simulation run.

2.5 Optimization via analysis-by-synthesis

One of the critical issues in modeling the learning process is to determine the level of representation of the observable data. In this paper, we explicitly assume that the calculation of the comparison is done only at the acoustic level. This may not entirely cover the range of inputs that a child receives in the actual learning process, which may also involve orofacial features [34-36]. But it allows us to systematically and separately test the effectiveness of information that may be present in acoustic data independent of the visual features. Therefore, parameters of the visible articulators such as the lips and the jaw are acoustically optimized. Note that this strategy does not prevent future studies from including visual information as additional training input.

The representation of the acoustic kinematics should be sufficiently detailed to allow accurate analysis-by-synthesis but not so detailed as to make the computation infeasible. For segmental learning, we need to identify a spectral representation that best captures the articulatory changes and reflects human speech perception. A good candidate is Mel-frequency cepstral coefficients (MFCCs), which have been successfully used in speech recognition and HMM-based synthesis [37]. In this paper, MFCCs of the surface acoustics of both natural and synthetic speech are calculated using a standard setting in Praat [18], and the difference between the two are used as errors in the optimization of the articulatory targets.

A set of articulatory targets is searched for each segmental interval, whose boundaries are manually defined before optimization. The optimization process is assumed to follow a simple gradient descent search algorithm with random adjustments. The purpose of this simple design is so that the result of the study would only reflect the effectiveness of the overall strategy rather than that of the optimization algorithm *per se*. Further improvement of the optimization process can be done in the future study by testing various optimization modules that implement the same input and output interfaces. For each segment, articulatory targets in the form of vocal tract shapes are optimized iteratively to minimize the total sum of square errors, E, of MFCC between original and synthesized sounds, which can be described as follows:

$$E = \sum_{i=1}^{n} \sum_{j=1}^{m} \left(c_{ij} - \hat{c}_{ij}\right)^2$$

where n is the number of acoustic feature timeframes, m is the number of MFCC coefficients, c_{ij} is the jth cepstral coefficient of the ith frame in the natural utterance, and $\hat{c}_{ij}$ is the jth cepstral coefficient of the ith frame in the synthetic utterance.

Some articulatory parameters may not be entirely independent of others, however. For example, the tongue parameters have been found to be positively correlated with each other in articulatory movements for certain places of articulation, such as alveolar, palatal, and velar [38]. This correlation suggests that there is a constraint weakly tying these parameters together so that the changes in one parameter also affect other parameters, depending on the physiological locations. Such an embodiment relationship can be used to help the optimization process so

Table 1 List of articulatory target parameters of VocalTractLab

Parameter	Description	Parameter	Description
HX, HY	Hyoid positions	VO	Velic opening
JX	Horizontal jaw position	TTX, TTY	Tongue tip positions
JA	Jaw angle	TBX, TBY	Tongue blade positions
LP	Lip protrusion	TCX, TCY	Tongue body positions
LD	Lip distance	TS1-TS4	Tongue side elevation 1-4
VS	Velum shape		

that the parameter adjustment is more realistic. In this paper, we modeled this embodiment constraint by co-adjusting nearby articulators. For example, whenever the tongue blade parameters (TBX/TBY) were adjusted, those of tongue tip and tongue body (TTX/TTY, TCX/TCY) were also modified by a small amount (20%) in the direction of the main adjustment.

2.6 Numerical assessment and perceptual evaluation

After obtaining the optimal target values, the accuracy of the estimated articulatory targets was assessed by comparing the formant tracks of the synthesized utterances with the original formant tracks. Time-normalized formant tracks (F1-F3) of both synthesized and original utterances were extracted using FormantPro [39], a Praat script for large-scale systematic analysis of continuous formant movements. The comparison was done by measuring for each syllable the root mean square error (RMSE) between the synthesized and original utterances.

To assess the synthesis quality, a listening experiment was conducted with native Thai participants to identify the synthetic vowels and evaluate their naturalness. Target parameters of the same vowel optimized through the analysis-by-synthesis strategy are averaged together across multiple contexts as the underlying representation of that vowel. A monosyllabic word of each individual vowel was predictively (since no monosyllables were used in the training) synthesized using the estimated articulatory parameters. The standard vocal tract configuration of a male German speaker provided by VocalTractLab version 2.1 [26] was used to generate the stimuli. It should be noted that while the original vocal tract configuration is derived from a German speaker, the articulatory parameters and the synthesis process are language dependent. As controls, the natural stimuli of the same words were recorded by the same speaker used in the training corpus at a sampling rate of 22.05 kHz and 16-b resolution. Recording was done in a sound-treated room at the King Mongkut's University of Technology Thonburi. Both natural and synthetic stimuli had their intensities normalized to 70 dB using Praat. In total, there were 18 stimuli.

Twenty native Thai listeners participated in the experiment, which was conducted with the ExperimentMFC function of Praat. The stimulus words were randomly presented to the listeners, who were asked to first identify the Thai word they heard and then select a naturalness score on a five-level scale from terrible (1) to excellent (5). Listeners were allowed to listen to the stimuli as many times as they preferred.

3. Results

3.1 Synthesis accuracy of estimated vowel targets

Table 2 shows the mean formant RMSE of each vowel compared between the synthetic and original utterances. Low RMSEs can be observed for all vowels compared to the average RMSE levels reported in the previous studies [40]. While there are no significant difference in RMSEs of F1 and F2 between the first and second syllables (F1, $t(8) = -0.528$, $p = 0.306$; F2, $t(8) = -0.403$, $p = 0.349$), RMSE of F3 is slightly higher in the second syllable than in the first (F3, $t(8) = -2.328$, $p = 0.024$). It should also be noted that a certain level of RMSE values might be attributed to the difference in the vocal tract anatomy of the speaker and the template used in learning. By further comparing these synthesized vowels to the empirical findings reported in [41], we find that both F1 and F2 of synthesized vowels are comparable to those of the natural ones as shown in Figure 4. This indicates that the optimization can effectively estimate the underlying articulatory targets of both syllables.

Table 2 Mean formant RMSEs in percentage of each vowel in each syllable

Vowel	First syllable			Second syllable		
	F1	F2	F3	F1	F2	F3
/aː/	9.8	4.0	6.0	8.7	4.2	6.2
/iː/	8.1	3.7	8.3	6.9	5.4	7.1
/uː/	4.9	8.4	6.9	6.9	7.9	8.0
/eː/	2.8	5.1	4.1	4.3	5.1	4.6
/ɛː/	9.9	7.8	3.7	7.6	6.6	5.2
/ɯː/	7.3	6.2	4.3	7.1	5.3	5.1
/ɤː/	4.1	3.2	2.3	5.0	3.9	3.9
/oː/	6.4	4.2	7.8	6.0	4.4	8.6
/ɔː/	8.3	2.9	5.9	6.8	3.8	6.4

The percentage value is calculated relative to the original formant averages.

Figure 5 shows examples of spectrograms of a vowel sequence in the corpus and that of a synthetic one generated with optimized articulatory targets. Further comparisons of formant frequency contours of exemplar utterances (obtained with FormantPro [39]) are shown in Figure 6. 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 (lower panels in Figure 5, solid lines in Figure 6), just as in the natural utterances (upper panels in Figure 5, dotted lines in Figure 6). Visual inspection of spectrograms of all other cases also confirmed the accuracy in representing the formant patterns of 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 6, between the original and synthesized formant frequencies, these mismatches are evened out once they

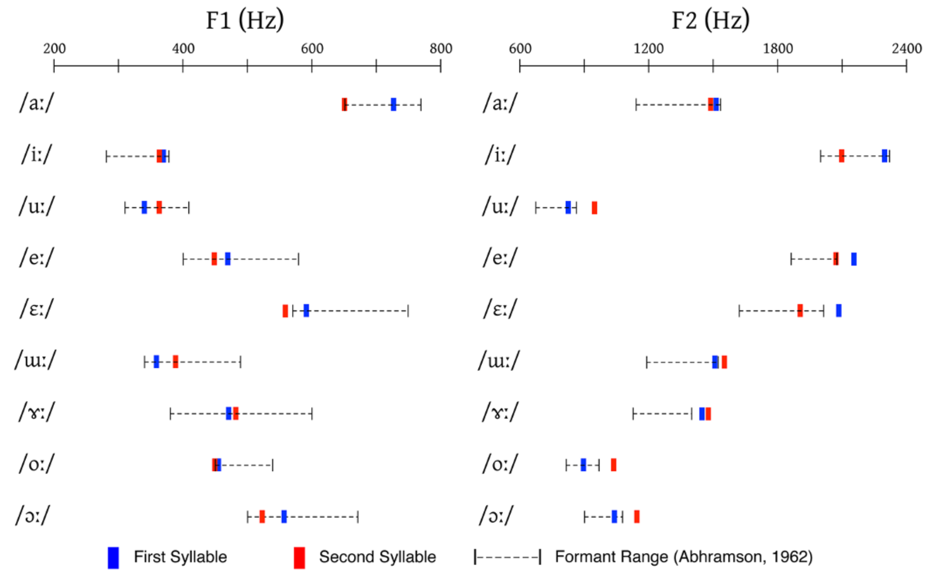

Figure 4 **Comparison of F1 and F2 formant frequencies of synthetic vowels and the empirical formant ranges [41].**

are averaged together with other cases (e.g., for /uː/, /aː uː/ and /uːɤː/ as shown in Figure 6). Further evaluation of synthesis quality will be later shown by a listening test.

3.1 Vocal tract shapes of Thai vowels

Figure 7 shows the articulatory targets as vocal tract shapes of Thai vowels reconstructed from the optimized articulatory targets by averaging the parameters across multiple instances of each vowel in both syllable locations. The shapes are largely consistent with previously reported Thai phonetic descriptions. Both vowel backness and vowel height of estimated vocal tract shapes are consistent with their theoretical locations [17]. Using these vocal tract shapes, we can synthesize vowel sequences either in isolation or in context. It should be noted that these shape targets are estimated without direct knowledge of the actual articulatory position, but through the utilization of the articulatory synthesizer and the acoustic optimization, starting from the vocal tract shape of a schwa. Distinctive vocal tract estimates that contain the gradient in terms of tongue backness and vocal tract openness as shown in Figure 7 suggest that the strategy is effective in obtaining the articulatory parameters capable of accurately generating the acoustic responses. These underlying targets of Thai vowels will be further used in the perceptual evaluation.

3.3 Synthesis quality

Figure 8 shows the results of the listening experiment on synthetic Thai vowels. Listeners could identify vowels equally well for both natural and synthetic vowels, with no significant difference between the two ($t(8) = 1.25$,

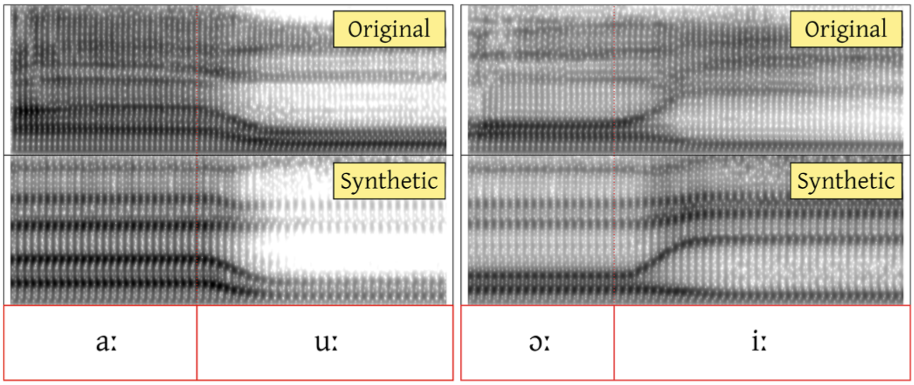

Figure 5 **Spectrogram comparisons of example original and synthetic utterances.**

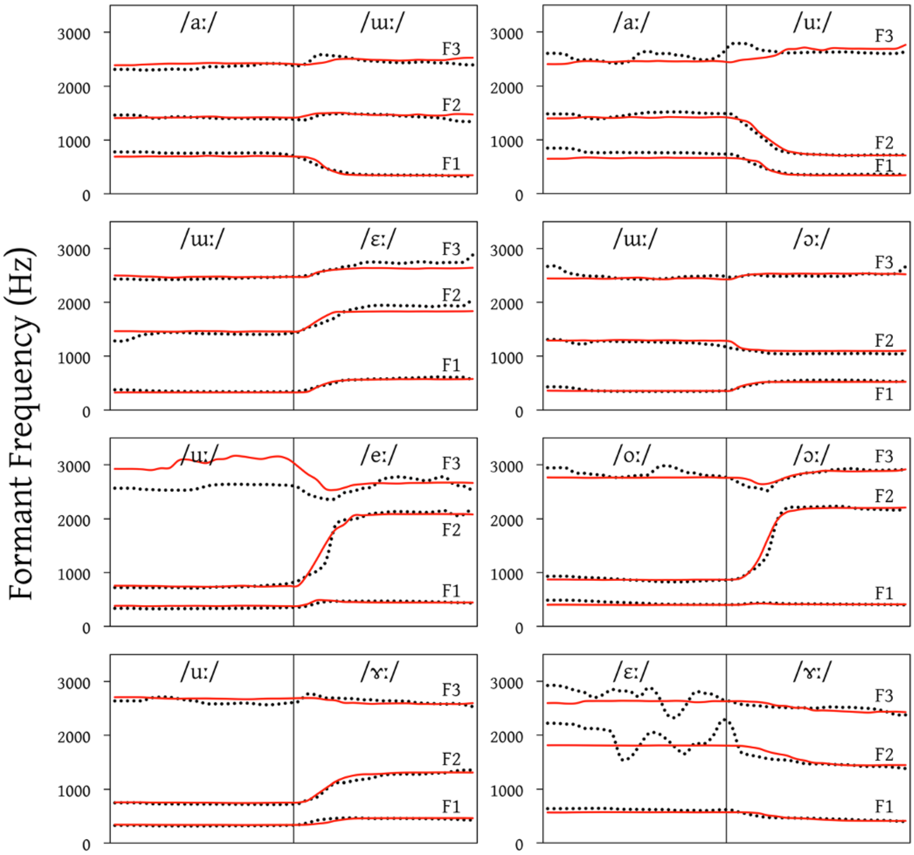

Figure 6 Time-normalized formant frequency contour comparisons of example original (dotted lines) and synthetic (solid lines) utterances. Each utterance consists of two syllables, each marked with the vowel symbol and separated by the boundary as the vertical line.

$p = 0.247$). All misidentifications were confusions with the neighboring vowels. It should be noted that in this study, one misidentified case is equivalent to 5% in the identification rate. This indicates the effectiveness of the proposed method for estimation of articulatory targets in representing vowels. 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. This is possibly because the estimated velum parameters are continuous rather than discrete values, which makes it difficult for an optimization algorithm to get a full closure of the velum opening. Since the natural /u:/ does not require the usage of the nasal track, the nasality in the synthetic /u:/ vowel may cause it to be confusable with /o:/, which contains a higher degree of nasality due to the openness of the vowel. Naturalness scores of natural vowels were generally better than those of synthetic vowels ($t(8) = 3.24$, $p = 0.012$). However, once we compare the range of the naturalness score, listeners identified both of them in roughly the same score ranges, 4.2 to 5 for natural stimuli and 3.9 to 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. Also, the slightly lower naturalness scores may be partially related to the specific voice quality used in generating the synthetic vowels, which we did not adjust to match the voice quality of the real speaker.

4. Discussion

The results of the present study have shown that it is possible to estimate the underlying articulatory targets of vowels from surface acoustics of continuous speech and predefined annotated segmental boundaries as the input, with an articulatory synthesizer controlled by target approximation models as the learner, and analysis-by-synthesis optimization as the training regimen. The numerical assessment as shown in Table 2 and Figure 4 indicates that the learned targets can be used to consistently synthesize the acoustic data that closely approximate the natural utterances. The visual impression of the synthesis examples as shown in Figures 5 and 6 suggests a good match in the dynamics of the acoustic data. The perceptual evaluation shows that the underlying articulatory targets learned this way can be used

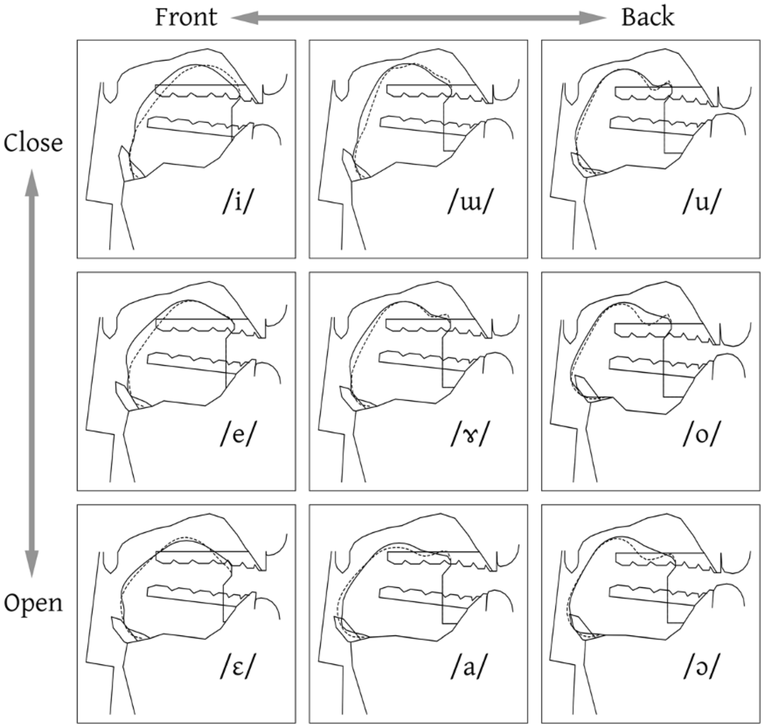

Figure 7 Vocal tract shapes reconstructed from estimated articulatory target of each Thai vowel. Dotted curves represent the tongue side elevations displayed relative to tongue body.

to generate isolated vowels that approximate to the original both perceptually as shown in Figure 8 and in terms of articulation as shown in Figure 7. All these results indicate that the estimated articulatory parameters closely represent the underlying targets of the Thai vowels.

One advantage of the present approach over the previous attempts is the decoupling of the observed data and the speech production mechanism. Compared to the mapping approaches [6-14], the present study does not use any actual articulatory data in the optimization process but

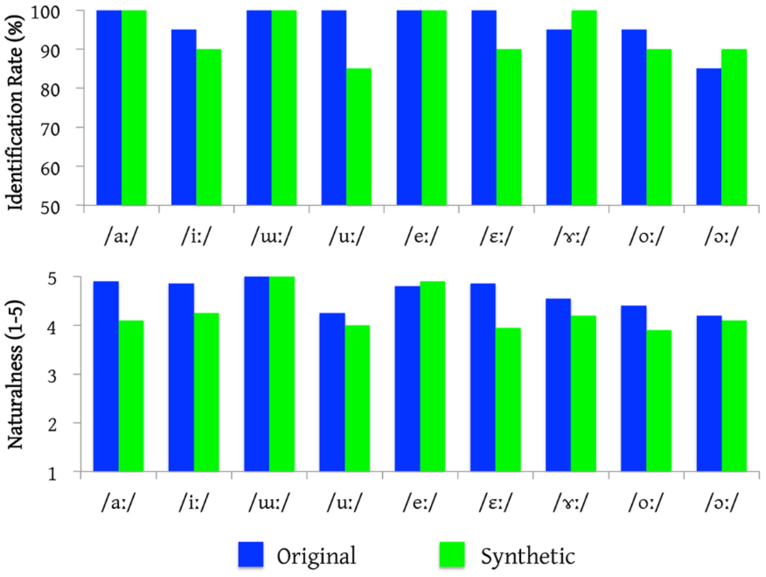

Figure 8 Mean vowel identification rate (top) and naturalness score (bottom) for both original and synthetic stimuli.

utilizes the knowledge of speech production mechanisms incorporated in the articulatory synthesis. This also provides an option that further studies may integrate the visible articulator data into the scheme to reduce the total degree of freedom. Another advantage of the present approach over the mapping approach is on the emergence of new data that are unknown to the system. The mapping approach would have difficulties in immediately applying the learned model to the new data since it has to be trained on a specific speaker. In contrast, the present approach has shown that even with the German vocal tract configuration, the system can learn Thai vowels that are numerically and perceptually accurate. Such decoupling of the articulatory mechanisms from the linguistic information enables us to apply this approach to different languages.

Another main feature of the present approach is the use of the TA model. With the TA model, the optimization only needs to estimate the targets and let the transient responses of the articulatory trajectories be calculated by the TA model. This significantly reduces the degrees of freedom of the estimation to only a set of articulatory parameters, along with a time constant, which need to be optimized, instead of having to map from every frame of acoustic data to the articulatory estimates. The usage of the TA model also allows us to simulate smooth transitions in the acoustic data as observed in natural utterances. While this feature may be present in different forms in the previous attempts such as generalized smoothness constraint [14], state transition in HMM [6,7], or smoothing algorithm [11-14], the direct incorporation of TA model into the articulatory synthesizer provides a simple yet effective strategy in the simulation of articulatory dynamics.

Some mapping studies that are based on the TD model [9,10] do take dynamic gestural control into consideration. 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 always completed and adjacent gesture movements are assumed to be overlapped. This is in contrast to TA, which assumes that targets are not always reached, and allows remaining momentum at the end of a target approximation movement to be transferred to the next interval as its initial conditions.

Further development of the framework for organizing trained targets is still needed. First, no consonants have been simulated, so the effect of gestural overlap between consonant and vowel has not yet been modeled. Strategies have yet to be developed to simulate the learning of the fully overlapped CV gestures [19]. Second, 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. Third, the visible articulatory data can be directly incorporated into the learning strategy. This will further reduce the degree of freedom of the optimization process and may further improve the effectiveness of the system. Finally, the acoustic-to-articulatory inversion in the current study is not fully complete, as the segmentation of continuous utterances into discrete unidirectional movements is done manually. The underlying assumption is that the learning of perceptual segmentation is achieved prior to the learning of the articulatory targets. But the validity of this assumption is not fully established and has to be addressed in future studies.

5. Conclusions

In this study, we explored the estimation of articulatory targets of Thai vowels as a distal learning task using a model-based analysis-by-synthesis strategy. Articulatory targets as vocal tract parameters of each vowel were iteratively optimized by minimizing the acoustic error between original and synthetic utterances. The estimated vocal tract shape targets were used to synthesize the acoustical vowels, and the perceptual evaluation confirmed the synthesis quality. These results demonstrate that distal learning with an articulatory synthesizer that incorporates knowledge of speech production mechanisms is an effective strategy for the simulation of speech production acquisition.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

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 New Researcher (TRG5680096 to SP).

Author details

[1]Department of Computer Engineering, Faculty of Engineering, King Mongkut's University of Technology Thonburi, Bangkok 10140, Thailand. [2]Department for Phoniatrics, Pedaudiology and Communication Disorders, University Hospital Aachen and RWTH Aachen University, Aachen 52074, Germany. [3]Department of Speech, Hearing and Phonetic Sciences, University College London, London WC1N 1PF, UK.

Received: 26 December 2013 Accepted: 17 April 2014
Published: 8 May 2014

References

1. P Mermelstein, Articulatory model for the study of speech production. J. Acoust. Soc. Am. **53**(4), 1070–1082 (1973). doi:10.1121/1.1913427
2. EL Saltzman, KG Munhall, A dynamical approach to gestural patterning in speech production. Ecol. Psychol. **1**, 333–382 (1989). doi:10.1207/s15326969eco0104_2
3. Y Xu, Speech melody as articulatorily implemented communicative functions. Speech Commun. **46**, 220–251 (2005). doi:10.1016/j.specom.2005.02.014
4. J Sun, L Deng, An overlapping-feature-based phonological model incorporating linguistic constraints: applications to speech recognition. J. Acoust. Soc. Am. **111**, 1086 (2002). doi:10.1121/1.1420380
5. Z Ling, K Richmond, J Yamagishi, R Wang, Integrating articulatory features into HMM-based parametric speech synthesis. IEEE Audio Speech Lang. Process. **17**(6), 1171–1185 (2009). doi:10.1109/TASL.2009.2014796

6. G Hofer, J Yamagishi, H Shimodaira, Speech-driven lip motion generation with a trajectory HMM, in *Proceedings of the 9th Annual Conference of the International Speech Communication Association* (Interspeech, Brisbane, 2008). 22–26 September 2008, pp. 2314–2317
7. M Tamura, S Kondo, T Masuko, T Kobayashi, Text-to-visual speech synthesis based on parameter generation from HMM, in *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP 98), Seattle, WA. 12–15 May 1998, pp. 3745–3748
8. B Uria, S Renal, K Richmond, A deep neural network for acoustic-articulatory speech inversion, in *Proceedings of the NIPS 2011 Workshop on Deep Learning and Unsupervised Feature Learning*, Sierra Nevada, Spain. 16 December 2011. http://www.cstr.ed.ac.uk/downloads/publications/2011/articulatory_inversion.pdf
9. V Mitra, H Nam, CY Espy-Wilson, E Saltzman, L Goldstein, Retrieve tract variables from acoustics: a comparison of different machine learning strategies. IEEE J. Sel. Topics Signal Process. **4**(6), 1027–1045 (2010). doi:10.1109/JSTSP.2010.2076013
10. H Nam, V Mitra, M Tiede, M Hasegawa-Johnson, C Espy-Wilson, E Saltzman, L Goldstein, A procedure for estimating gestural scores from speech acoustics. J. Acoust. Soc. Am. **132**(6), 3980–3989 (2012). doi:10.1121/1.4763545
11. J Schroeter, MM Sondhi, Dynamic programming search of articulatory codebooks, in *Proceedings of the 1989 IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP, 1989). vol. 1, Glasgow, UK, 23–26 May 1989, pp. 588–591
12. S Ouni, Y Laprie, Modeling the articulatory space using a hypercube codebook for acoustic-to-articulatory inversion. J. Acoust. Soc. Am. **118**(1), 444–460 (2005). doi:10.1121/1.1921448
13. B Potard, Y Laprie, S Ouni, Incorporation of phonetic constraints in acoustic-to-articulatory inversion. J. Acoust. Soc. Am. **123**(4), 2310–2323 (2008). doi:10.1121/1.2885747
14. PK Ghosh, S Narayanan, A generalized smoothness criterion for acoustic-to-articulatory inversion. J. Acoust. Soc. Am. **128**(4), 2162–2172 (2010). doi:10.1121/1.3455847
15. S Panchapagesan, A Alwan, 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). doi:10.1121/1.3514544
16. R McGowan, Recovering articulatory movement from formant frequency trajectories using task dynamics and a genetic algorithm: preliminary model test. Speech. Commun. **14**, 19–48 (1994). doi:10.1016/0167-6393(94)90055-8
17. K Tingsabadh, AS Abhramson, Thai. J Int. Phon. Assoc. **22**(1), 24–48 (1993). doi:10.1017/S0025100300004746
18. P Boersma, Praat, a system for doing phonetics by computer. Glot. Int. **5**(9/10), 314–345 (2001)
19. Y Xu, F Liu, Tonal alignment, syllable structure and coarticulation: toward an integrated model. Italian. J. Linguist. **18**, 125–159 (2006)
20. S Prom-on, P Birkholz, Y Xu, Training an articulatory synthesizer with continuous acoustic data, in *Proceedings of the 14th Annual Conference of the International Speech Communication Association* (Interspeech, Lyon, 2013). 25–29 August 2013, pp. 349–353
21. S Prom-on, B Thipakorn, Y Xu, Modeling tone and intonation in Mandarin and English as a process of target approximation. J. Acoust. Soc. Am. **125**, 405–424 (2009). doi:10.1121/1.3037222
22. Y Xu, S Prom-on, Toward invariant functional representations of variable surface fundamental frequency contours: synthesizing speech melody via model-based stochastic learning. Speech Commun. **57**, 181–208 (2014). doi:10.1016/j.specom.2013.09.013
23. S Prom-on, F Liu, Y Xu, Post-low bouncing in Mandarin Chinese: acoustic analysis and computational modeling. J. Acoust. Soc. Am. **132**, 421–432 (2012). doi:10.1121/1.4725762
24. Y Xu, F Liu, Determining the temporal interval of segments with the help of F_0 contours. J. Phon. **35**, 398–420 (2007). doi:10.1016/j.wocn.2006.06.002
25. MI Jordan, DE Rumelhart, Forward models: supervised learning with a distal teacher. Cogn. Sci. **16**, 307–354 (1992). doi:10.1207/s15516709cog1603_1
26. P Birkholz, VocalTractLab 2.1 for Windows (2013). http://www.vocaltractlab.de. Accessed 17 December 2013
27. P Birkholz, Modeling consonant-vowel coarticulation for articulatory speech synthesis. PLoS One **8**(4), e60603 (2013). doi:10.1371/journal.pone.0060603
28. P Birkholz, BJ Kröger, C Neuschaefer-Rube, Model-based reproduction of articulatory trajectories for consonantal-vowel sequences. IEEE Audio, Speech and Lang. Process **19**(5), 1422–1433 (2011). doi:10.1109/TASL.2010.2091632
29. P Birkholz, BJ Kröger, C Neuschaefer-Rube, Proceedings of the 12th Annual Conference of the International Speech Communication Association (Interspeech 2011). Florence **28–31**, 2681–2684 (August 2011)
30. P Birkholz, D Jackèl, BJ Kröger, Simulation of losses due to turbulence in the time-varying vocal system. IEEE Audio, Speech and Lang. Process **15**(4), 1218–1226 (2007). doi:10.1109/TASL.2006.889731
31. Y Xu, QE Wang, Pitch targets and their realization: evidence from Mandarin Chinese. Speech. Commun. **33**, 319–337 (2001). doi:10.1016/S0167-6393(00)00063-7
32. FH Guenther, T Vladusich, A neural theory of speech acquisition and production. J. Neurolinguist **25**, 402–422 (2012). doi:10.1016/j.jneuroling.2009.08.006
33. FH Guenther, Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production. Psychol. Rev. **102**, 594–621 (1995). doi:10.1016/j.jneuroling.2009.08.006
34. JR Green, CA Moore, M Higashikawa, RW Steeve, The physiologic development of speech motor control: lip and jaw coordination. J. Speech Lang. Hear. Res. **43**, 239–255 (2000). PMCID: PMC2890218
35. JR Green, CA Moore, KJ Reilly, The sequential development of jaw and lip control for speech. J. Speech Lang. Res **45**, 66–79 (2002). PMCID: PMC2890215
36. MP Harold, SM Barlow, Effects of environmental stimulation on infant vocalizations and orofacial dynamics at the onset of canonical babbling. Infant Behav. Dev. **36**, 84–93 (2013). doi:10.1016/j.infbeh.2012.10.001
37. P Taylor, *Text-to-Speech Synthesis* (Cambridge University Press, Cambridge, 2009)
38. JR Green, YT Wang, Tongue-surface movement patterns during speech and swallowing. J. Acoust. Soc. Am. **113**(5), 2820–2833 (2009). doi:10.1121/1.1562646
39. Y Xu, FormantPro Version 1.1, http://www.phon.ucl.ac.uk/home/yi/FormantPro. Accessed 24 December 2013
40. RS McGowan, MA Berger, Acoustic-articulatory mapping in vowels by locally weighted regression. J. Acoust. Soc. Am. **126**(4), 2011–2032 (2009). doi:10.1121/1.3184581
41. AS Abramson, *The vowels and tones of standard Thai: acoustical measurements and experiments* (Indiana University Research Center in Anthropology, Folklore, and Linguistics, Pub. 20, Bloomington, 1962). http://www.haskins.yale.edu/Reprints/HL0035.pdf. Accessed 26 February 2014

doi:10.1186/1687-4722-2014-23
Cite this article as: Prom-on *et al.*: Identifying underlying articulatory targets of Thai vowels from acoustic data based on an analysis-by-synthesis approach. *EURASIP Journal on Audio, Speech, and Music Processing* 2014 **2014**:23.