Coarticulation as synchronized dimension-specific sequential target approximation: An articulatory synthesis simulation

Anqi Xu[1], Peter Birkholz[2], Yi Xu[1]

Department of Speech, Hearing and Phonetic Sciences, University College London, UK[1],
Institute of Acoustics and Speech Communication, Technische Universität Dresden, Germany[2]
a.xu.17@ucl.ac.uk, peter.birkholz@tu-dresden.de, yi.xu@ucl.ac.uk

ABSTRACT
The present study tested the idea that coarticulation, despite involving overlap of articulatory gestures, is achieved by sequential target approximation at the level of individual articulator dimensions. For example, CV co-onset in a velar stop can be achieved by having the tongue body vertically move upward for a closure contact, while at the same time also moving horizontally to achieve the tongue shape for the vowel, resulting in velar contact locations that vary gradiently with adjacent vowels. We examined this hypothesis in an analysis-by-articulatory-synthesis paradigm, whereby vocal tract parameters were optimized to minimize acoustic differences between synthetic and natural speech. Results of a perceptual identification experiment demonstrated that syllables synthesized with articulatory parameters learned this way had fairly high intelligibility. Potential impacts of the findings on understanding coarticulation, coarticulation resistance and vocal learning as well as on articulatory synthesis are discussed.

Keywords: coarticulation, articulatory synthesis, target approximation, vocal learning

1. INTRODUCTION

The term 'coarticulation' ('Koartikulation') was first proposed by Menzerath and de Lacerda [17] to describe the phenomenon that the movement of the vowel in a consonant-to-vowel (CV) sequence starts at the same time as the consonant [15]. By now, however, it is mostly used to refer to any variation of a segment with adjacent or nearby segments. The contextual variability of segments has intrigued theoretical discussions on how linguistically invariant segments take various articulatory-acoustic manifestations in connected speech. In the task dynamic model [10-11, 26] as well as Articulatory Phonology [7], it is assumed that there are temporal overlaps between linguistically relevant movements of the vocal tract, referred to as gestures. In the window model of coarticulation, a segmental feature has a 'window' consisting of a maximum and a minimum physical value that reflects contextual sensitivity [14]. Bladon & Al-Bamerni [3] has hypothesized that there is a specific "coarticulation resistance" value associated with each segment. Later on, quantitative measurements of coarticulation resistance have been developed based on advances in articulatory imaging techniques, such as the degree of articulatory constraints (DAC) model [24-25], the statistical model [13] and the mutual information (MI) model [12]. These measurements represent to what extent a specific articulator is involved with the presence of distinct surrounding segments. The early philosophical models and the latest measurements tend to focus on the articulatory movement. Difficulties arise, however, when it is implemented in articulatory synthesis, which requires high perceptual accuracy.

Lindbolm [16] applied linear regressions to capture the changes in vowel formant frequencies soon after the consonant release, known as locus equations. As an extension, Öhman [19] proposed a mathematical function that treated the consonant gesture as a diphthongal force that superposes the movements of vowels. His work inspired a growing body of literature that adopted different approaches to calculating vocal tract area functions for the modelling of coarticulation in speech synthesis [1, 8, 27]. Birkholz [1] modelled the vocal tract shapes of context-sensitive consonants based on weighted means of reference shapes of consonants following point vowels (i.e., /a/, /i/ and /u/) via acoustic optimization. This synthesis system relies on articulatory data to pre-define the vocal tract shapes, which is unsatisfactory if articulatory synthesis were to be envisaged as a simulation of vocal learning, as learners would not normally have access to knowledge of articulation.

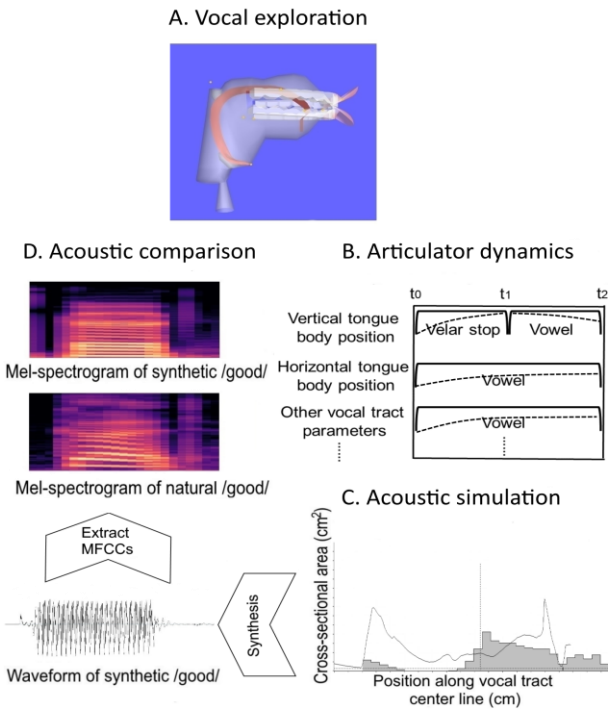
The aim of the current study is to explore how coarticulation can be learned through acoustic imitation by simulating it in articulatory synthesis using VocalTractLab [2]. We tested the hypothesis that a) C and V articulation are fully synchronized at syllable onset, and b) despite the CV overlap, at the level of individual articulator dimensions, target approximation is sequential [28]. The hypothesis differs from the task dynamic model [10-11, 26] in that a) CV synchrony is presumed rather than learned,

and b) there is no blending at the level of articulatory dimension. For example, in /gV/, the tongue body vertically moves upward for a velar contact, while also moving horizontally for the vowel tongue shape, resulting in a velar contact location depending on the vowel context. The articulatory synthesis paradigm has been successfully tested in generating Thai vowels but not CV sequences where the C is an obstruent consonant [20-21]. The presented study is to test the effectiveness of adding dimension-specific sequential target approximation to the paradigm. The performance of the model will be evaluated in terms of perceptual quality of the synthetic sounds via an identification task and in terms of plausibility of the learned articulatory parameters.

2. METHOD

Fig. 1 is an illustration of the learning model. The articulatory targets being learned are parameters of VocalTractLab (Fig. 1A). The dynamics of the articulators are controlled by dimension-specific sequential target approximation (Fig. 1B). The simulated vocal tract parameter curves are then used to calculate area functions for acoustic simulation (Fig. 1C). Mel frequency cepstral coefficients (MFCCs) are extracted from both the target and synthetic sounds and compared (Fig. 1D). The articulatory parameters are optimized iteratively from A to D to minimize the sum of squared errors of MFCCs (i.e., the cepstral distance) between the target and synthetic sounds.

Figure 1: Overview of the learning model.



2.1. Vocal tract model

VocalTractLab 2.2 (www.vocaltractlab.de) [2] calculates area functions for acoustic simulation on the basis of a geometrical 3D vocal tract model, adapted from MRI data of a German male speaker. The simulation involved twenty vocal tract parameters, as shown in Table 1. During speech production, the physiological structure of the vocal tract restricts the articulatory movement. Here, the inter-articulator constraints were simulated by regulating adjacent articulators together. For example, whenever the tongue blade parameters are adjusted, the tongue body parameters move in the same direction by 20%.

Table 1: Vocal tract parameters of the model

Parameter	Description
HX, HY	Horiz. and vert. hyoid positions
JX, JA	Jaw position and Jaw angle
LP, LD	Lip protrusion and vert. lip distance
VS, VO	Velum shape and velum opening
TCX, TCY	Horiz. and vert. tongue body center positions
TTX, TTY	Horiz. and vert. tongue tip positions
TBX, TBY	Horiz. and vert. tongue blade positions
TRX, TRY	Horiz. and vert. tongue root positions
TS1 – TS4	Tongue side elevation from the anterior to the posterior part of the tongue

2.2. Simulation

We ran a series of simulations for learning English monosyllabic CVC words containing bilabial, alveolar and velar stops (Table 2), with the goal that the synthetic words would be correctly identified by naïve listeners without phonetic knowledge. For the target words to be learned, recordings were made by a female American English speaker in a quiet room.

In the simulation, what was being learned were articulatory targets of consonants and vowels. In each learning cycle, a full set of hypothetical targets with pre-specified fixed time intervals were tested. Each articulatory dimension was controlled either by the consonant or by the vowel. As shown in Fig. 1B for /gV/, at t_0 the vertical tongue body parameter was allowed to move towards the consonant target and the other articulator dimensions moved towards the vowel target. The vertical tongue body parameter was allowed to move towards vowel target from t_1, which ended at t_2. Similarly, for /b/, the lip distance was controlled by the consonant before t_1, and for /d/, the

vertical tongue tip and tongue blade positions were controlled by the consonant until t_1.

Table 2: Target words

Vowel	/bv/	/dV/	/gV/
/i/	bead	deed	
/ɪ/	bid	did	
/ɛ/	bed	dead	get
/æ/	bad		
/ɒ/	bod		god
/u/	booed		good
/ʌ/	bud		

2.3. Optimization

The optimization was based on an analysis-by-synthesis paradigm to simulate learners' vocal exploration. We adopted a pure random search algorithm [6] in which the vocal tract configurations were randomly adjusted until a best acoustic match with the target utterance was obtained. MFCC was used as auditory feedback, which is a robust parametric representation widely used in speech recognition and speech synthesis [9]. For each target CV sequence, the model was trained for 3-30k iterations, depending on the learning difficulties. After CV sequences were trained, codas were optimized in the same way. The intonation contours were synthesized based on f_0 targets learned from the target words using PENTAtrainer2 [22-23].

2.4. Identification experiment

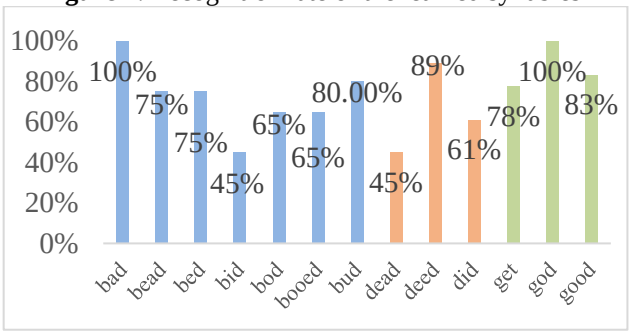
Ten native English speakers (female: 6) participated in the perception experiment. The stimuli were thirteen recorded target sounds and the corresponding synthetic sounds. The stimuli were randomised and presented by the ExperimentMFC function of Praat [4]. The perceptual task was a free identification task in which listeners wrote down what they heard and judged the naturalness of the stimuli on a 1-5 Likert scale with 5 being the most natural and 1 being the most unnatural.

3. RESULTS

3.1. Identification results

The recognition rate was calculated in terms of how many segments were correctly identified. The mean recognition rate of the CV sequences was 97% for the natural syllables and 74% for the synthetic ones. Fig. 2 shows the recognition rates of the synthetic syllables. The average naturalness rating of the natural and synthetic sounds was 4.74 and 1.91 respectively. An ordered logistic regression showed that the naturalness rating of the synthetic speech significantly predicted whether the whole word was correctly perceived ($\beta = 0.808$, SE = 0.355, $t = 2.275$). Sample natural and synthetic sounds are embedded at the end of the PDF.

Figure 2: Recognition rate of the learned syllables



3.2. Learned vocal tract parameters

Fig. 3 displays the optimized vocal tract shapes of the velar stops at the moment of maximal constriction. With regard to the horizontal tongue body position (TCX), the more positive the number, the more forward the tongue position. The learned tongue body targets of velar stops are similar in the vertical position (TCY) but different in the horizontal position (TCX). The tongue body therefore moved upward to contact the soft palate and also horizontally towards the vowel. As the vowel in /get/ is more anterior than in /god/, the place of the closure of /g/ is more anterior in /get/, too.

Figure 3: Learned vocal tract shapes right before the release of the initial consonants in /gV/ sequences. TCX and TCY represent the horizontal and vertical tongue body positions respectively.

A. /god/ B. /get/

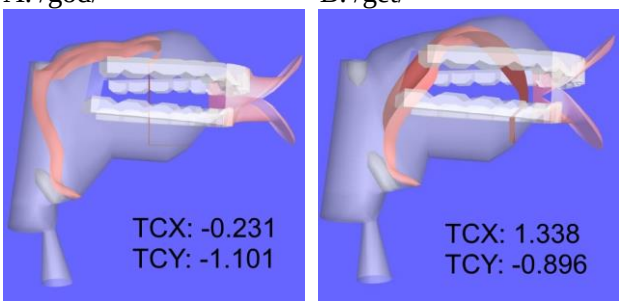


Fig. 4 and Fig. 5 illustrate the learned vocal tract shapes right before the release of the bilabial stops and alveolar stops. The negative lip distance indicates a closed lip target and the negative TTY values indicate that the consonant target forms a constriction at the alveolar ridge. In Fig. 4, although the lips are both closed before the release, the tongue shapes are ready for the vowel. Likewise, in CV sequences

containing alveolar stops (Fig. 5), the anterior part of the tongue is in a similar shape at the moment of the oral contact, the posterior part of the tongue is in a shape similar to the adjacent vowel.

Figure 4: Learned vocal tract shapes right before the release of the initial consonants in /bV/ sequences. LD represents the lip distance.

A. /bod/ B. /booed/

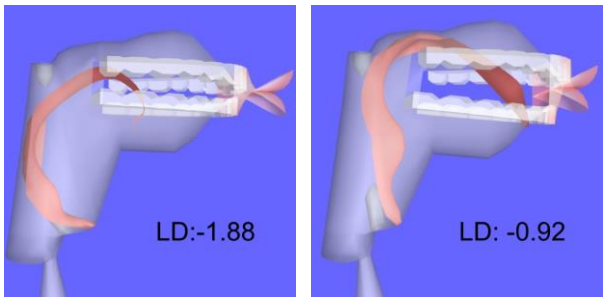
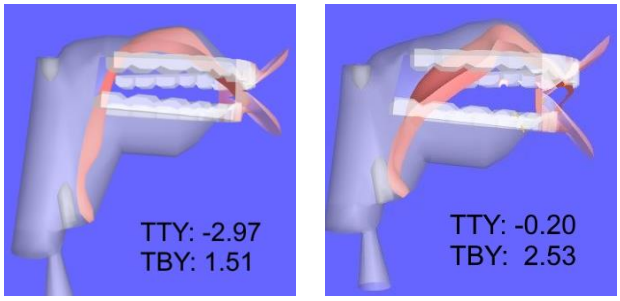


Figure 5: Learned vocal tract shapes right before the release of the initial consonants in /dV/ sequences. TTY and TBY represent the vertical tongue tip positions and the tongue blade positions respectively.

A. /deed/ B. /did/



4. DISCUSSION

It has been well-established that different speech segments affect the articulatory movement of the surrounding segments to a varying degree, known as 'coarticulation resistance' [3, 12-13, 24-25]. The contextual variations have been interpreted as originating from variable temporally overlapping gestures between the consonant and vowel in the task dynamic model and Articulatory Phonology, which may sometimes involve gestural blending [7, 10-11, 19, 26]. In the present study we tested the alternative hypothesis that even when the C and V gestures are overlapped in time, the articulatory execution in the form of target approximation can be serially ordered for specific articulatory dimensions [28]. In /dV/ and /gV/ sequences, the part of the tongue crucial for a consonant moves upwards to make a contact, while the rest of the tongue moves backwards or forwards for the co-produced vowel. In /bV/ sequences, the lip distance is controlled by the consonant, while the entire tongue moves towards the position for the co-produced vowel. Such simulated CV co-onset therefore sheds light on the inner workings behind the observed co-onset of vowel and consonant movements at the beginning of the syllable upon which the term 'koartikulation' was coined [17].

The study also shows that although they have heavily burdened concatenative speech synthesis, contextual variations of phonetic segments can be simulated without excessive amounts of training data. Furthermore, unlike previous articulatory synthesis that relies on articulatory data [1], the present study has tackled the acoustic-to-articulation mapping by implementing analysis-by-articulatory-synthesis with the hypothetical mechanisms of synchronized CV co-onset and dimension-specific target approximation. The proposed methods may eventually lead to high-quality articulatory speech synthesis. Importantly, the methods also showed the ability to address the speaker normalization problem in language acquisition [5, 18], because the target sounds were spoken by a female while the vocal tract model used in the learning was that of a male. Guided by auditory feedback, the model was able to automatically learn to produce intelligible CV sequences despite the anatomical differences. This demonstrates that the learner can discover a motor representation equivalent to the sensory input by repeatedly adjusting vocal tract configurations to match the perceived sounds. The simulation provides a possible solution for the problem of how speech production and perception can be linked by acoustic imitation. Further research, however, will be undertaken to simulate children's vocal learning and how the critical articulator dimensions for consonant targets are discovered by learners rather than being pre-set as was done in the present study.

Another limitation of the current study is that the pure random search used was time-consuming. The intention was to test the power of the synchronization mechanism and the intrinsic articulatory constraints rather than to find the best machine learning algorithms. Work is currently in progress to apply genetic algorithms and neural networks to speed up the optimization process.

Overall, the findings of the study provide support for the hypothesis that CV coarticulation is realized by co-onset of sequential target approximation at the level of individual articulator dimensions. The model succeeded in simulating the learning of contextual articulatory variances and achieved fairly high intelligibility. The findings therefore offer new insight on the basic mechanisms of coarticulation and vocal learning, and may have implications for high-quality articulatory synthesis.

5. REFERENCES

[1] Birkholz, P. 2013. Modeling Consonant-Vowel Coarticulation for Articulatory Speech Synthesis. *PLoS ONE* 8(4), e60603.

[2] Birkholz, P. 2017. VocalTractLab 2.2.

[3] Bladon, R. A., Al-Bamerni, A. 1976. Coarticulation resistance in English /l/. *Journal of Phonetics* 4, 135-150.

[4] Boersma, P., Weenink, D. 2016. Praat: doing phonetics by computer.

[5] Breazeal, C., Scassellati, B. 2002. Robots that imitate humans. *Trends in cognitive sciences* 6, 481-487.

[6] Brooks, S. H. 1958. A Discussion of Random Methods for Seeking Maxima. *Operations Research* 6, 244-251.

[7] Browman, C. Goldstein, L. 1986. Towards an articulatory phonology. *Phonology Yearbook* 3, 219-252.

[8] Carré, R., Chennoukh, S. 1995. Vowel-consonant-vowel modeling by superposition of consonant closure on vowel-to-vowel gestures. *Journal of Phonetics* 23(1–2), 231–241.

[9] Davis, S. B., Mermelstein, P. 1980. Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences. *IEEE Transactions on Audio, Speech and Language Processing* 28(4), 357–366.

[10] Fowler, C. A. 1980. Coarticulation and theories of extrinsic timing. *Journal of Phonetics* 8, 113–133.

[11] Fowler, C. A., Saltzman, E. 1993. Coordination and Coarticulation in Speech Production. *Language and Speech* 36(2–3), 171–195.

[12] Iskarous, K., Mooshammer, C., Hoole, P., Recasens, D., Shadle, C. H., Saltzman, E., Whalen, D. H. 2013. The coarticulation/invariance scale: Mutual information as a measure of coarticulation resistance, motor synergy, and articulatory invariance. *The Journal of the Acoustical Society of America* 134(2), 1271–1282.

[13] Jackson, P. J. B., Singampalli, V. D. 2009. Statistical identification of articulation constraints in the production of speech. *Speech Communication* 51(8), 695–710.

[14] Keating, P. 1990. The Window Model of Coarticulation: Articulatory Evidence. *Papers in Laboratory Phonology I*, 451–470.

[15] Kühnert, B., Nolan, F. 1999. The origin of coarticulation. In: Hardcastle,W. J. & Hewlett, N. (eds), *Coarticulation: Theoretical and Empirical Perspectives*. Cambridge: Cambridge University Press, 7-30.

[16] Lindblom, B. 1963. Spectrographic Study of Vowel Reduction. *The Journal of the Acoustical Society of America* 35(11), 1773–1781.

[17] Menzerath, P., de Lacerda, A. 1933. *Koartikulation, Seuerung und Lautabgrenzung.* Berlin and Bonn: Fred. Dummlers.

[18] Messum, P., Howard, I. S. 2015. Creating the cognitive form of phonological units: The speech sound correspondence problem in infancy could be solved by mirrored vocal interactions rather than by imitation. *Journal of Phonetics* 53, 125-140.

[19] Öhman, S. E. G. 1967. Numerical Model of Coarticulation. *The Journal of the Acoustical Society of America* 41(310), 310–320.

[20] Prom-On, S., Birkholz, P., Xu, Y. 2013. Training an articulatory synthesizer with continuous acoustic data. *Proceedings of INTERSPEECH* Lyon, France, 349–353.

[21] Prom-On, S., Birkholz, P., Xu, Y. 2014. 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*, 23.

[22] Prom-on, S., Xu, Y. 2012a. PENTATrainer2 : A hypothesis-driven prosody modeling tool. *Proceedings of Exling.* Athens, Greece, 93-100.

[23] Prom-on, S., Xu, Y. 2012b. Pitch Target Representation of Thai Tones. *Proceedings of Tal.* Nanjing, China.

[24] Recasens, D. 1985. Coarticulatory patterns and degrees of coarticulatory resistance in Catalan CV sequences. *Language and Speech* 28, 97-114.

[25] Recasens, D., Espinosa, A. 2009. An articulatory investigation of lingual coarticulatory resistance and aggressiveness for consonants and vowels in Catalan. *The Journal of the Acoustical Society of America* 125(4), 2288–2298.

[26] Saltzman, E. L., Munhallt, K. G. 1989. A Dynamical Approach to Gestural Patterning in Speech Production. *Haskins Laboratories Status Report on Speech Research* SR-99, 38–68.

[27] Story, B. H. 2017. An acoustically-driven vocal tract model for stop consonant production. *Speech Communication* 87, 1-17.

[28] Xu, Y. manuscript. Syllable as a synchronization mechanism that makes human speech possible. http://www.homepages.ucl.ac.uk/~uclyyix/Syllable_manuscript.pdf

6. AUDIO EXAMPLES

Target /bad/

Learned /bad/

Target /did/

Learned /did/

Target /good/

Learned /good/