Articulatory Synthesis and Perception of Plosive-Vowel Syllables with Virtual Consonant Targets

Peter Birkholz, Bernd J. Kröger, Christiane Neuschaefer-Rube

Department of Phoniatrics, Pedaudiology, and Communication Disorders,
University Hospital Aachen and RWTH Aachen University, Aachen, Germany
pbirkholz@ukaachen.de, bkroeger@ukaachen.de, cneuschaefer@ukaachen.de

Abstract

Virtual articulatory targets are a concept to explain the different trajectories of primary and secondary articulators during consonant production, as well as the different places of the tongue-palate contact depending on the context vowel, for example in [igi] vs. [ugu]. The virtual targets for the tongue tip and the tongue body in apical and dorsal plosives are assumed to lie above the palate, and for bilabial consonants, the target is a negative degree of lip opening. In the present study, we discuss the concept of virtual targets and its application for articulatory speech synthesis. In particular, we examined how the location of virtual targets affects the acoustics and intelligibility of synthetic plosive-vowel syllables. It turned out that virtual targets that lie about 10 mm beyond the consonantal closure location allow a more precise reproduction of natural speech signals than virtual targets at a distance of about 1 mm. However, we found no effect on the intellegibility of the consonants.

Index Terms: articulatory speech synthesis, transition modeling, virtual targets

1. Introduction

In the present study, we discuss the application of virtual consonant targets for articulatory speech synthesis and examine the effect of different virtual target positions on the acoustics and intelligibility of synthetic plosive-vowel syllables. The rest of this section introduces the concept of virtual targets. The articulatory synthesis of the syllables using virtual consonant targets is discussed in Sec. 2 and the perception experiment in Sec. 3. We conclude with a discussion in Sec. 4.

When we observe fleshpoints on the lips and the tongue during transitions between vowels and consonants with oral closures, we notice that the trajectory shape of the primary articulator differs from those of secondary articulators. Figure 1 illustrates the difference by means of EMA-recordings for the utterance [ˌnɔnɔˈnɔnɔ] [1]. The upper and lower solid curves show the displacement of fleshpoints on the tongue tip (primary articulator) and the jaw (secondary articulator) along the main movement direction of these points during the repeated opening and closing gestures. The tongue tip curve has distinct plateaus during the closure intervals. During these intervals, the tongue tip is braced against the palate and hardly moves. The trajectories from the preceding vowel to the closure onset and from the closure offset to the following vowel have an exponential-like shape. Hence, the tongue tip velocity is quite high right before the closure onset and right after the release. In contrast, the opening and closing movements of the jaw are more sigmoidal (s-shaped). These movements are characterized by a smooth acceleration phase followed by a smooth deceleration

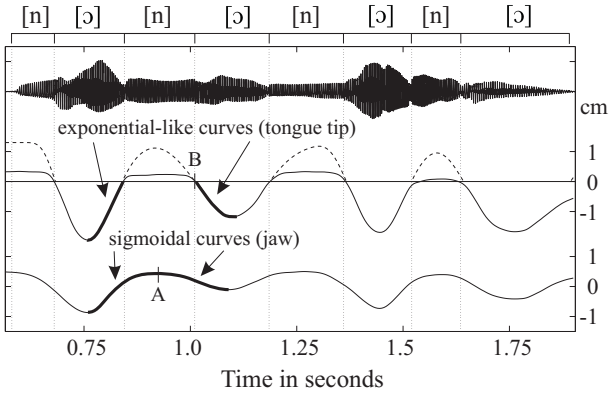

Figure 1: *Audio signal (top), tongue tip trajectory (middle), and jaw trajectory (bottom) for the utterance [ˌnɔnɔˈnɔnɔ]. The trajectories were measured by electromagnetic articulography (EMA) for coils on the tongue tip and the lower incisors [1]. Each trajectory shows the displacement along the first principal component of the original two-dimensional trajectory in the midsagittal plane. The dashed curves show hypothetical continuations of the tongue tip trajectory towards and away from virtual targets during the closure intervals.*

phase. Furthermore, the jaw appears to begin the transition from a plosive to a vowel before the closure offset (mark A vs. mark B in Fig. 1). The patterns described here are typical for all plosives [2].

As explanation for the high velocities of the primary articulators at the closure onset, Löfqvist and Gracco [5] suggest that the articulators are controlled towards *virtual* targets for the consonants. For apical and dorsal plosives, these virtual targets lie above the palate. To reach these targets, the tongue would have to move into the nasal cavity. For bilabial closures, the virtual target is a region of negative lip aperture. That is, to reach their virtual targets, the lips would have to move beyond each other. When we assume a sigmoidal trajectory for the primary articulator from the vowel target to the virtual consonant target analogous to the secondary articulators, the primary articulator would be temporally in the middle part of the trajectory when it hits the opposing wall. At this time, it has the observed high velocity. The advantage of a high velocity right before the closure onset is that the resulting tissue compression ensures the required airtight closure independent from contextual effects [5].

According to the concept of virtual targets, the dashed curves in Fig. 1 show possible virtual continuations of the

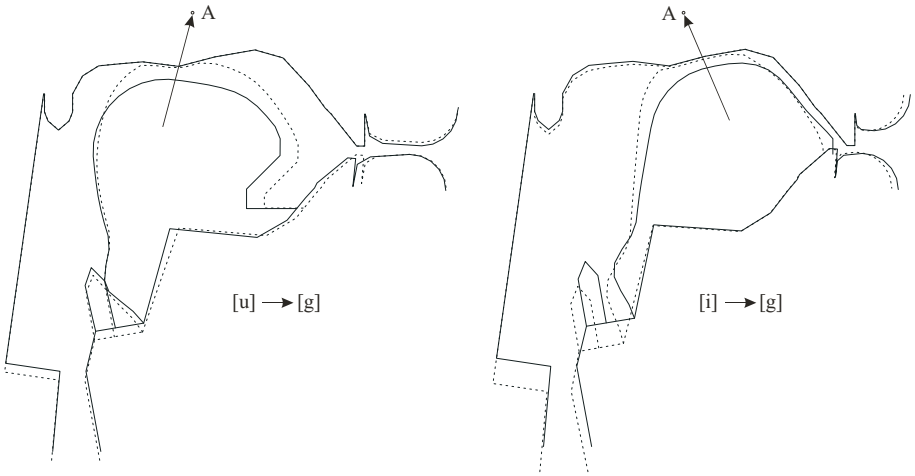

Figure 2: *Left: Vocal tract contour of [u] (solid) and [g] (dashed) in [ugu]. Right: Vocal tract contour of [i] (solid) and [g] (dashed) in [igi]. The contours show model-based reconstructions of vocal tract shapes measured by dynamic MRI [3]. Heike [4] proposes that the different tonge-palate contact positions in [g] result when the tongue body is stopped by the palate on its way from the vowel target to the (context-independent) virtual consonant target A.*

tongue tip movement, if the palate did not stop it. These curves are the result of previous experiments to reproduce natural articulatory trajectories and to estimate the position of virtual targets [6]. The estimated virtual targets were typically located 5–15 mm above the contact location.

Heike [4] proposed the concept of virtual targets in a somewhat different context. He hypothesized a virtual tongue body target above the palate for the closing gesture during [g] to explain the different tongue-palate contact locations observed in [igi] and [ugu]. This idea is illustrated by the vocal tract contours in Fig. 2 that were obtained from (dynamic) MRI measurements [3]. The solid contours on the left and the right show the vowels [u] and [i], respectively. The dashed contours on the left and the right show the vocal tract shapes during the closure phases in [ugu] and [igi], respectively. Obviously, the tongue body during the [g] in [igi] is more anterior than during the [g] in [ugu]. If the tongue body is assumed to move along a straight line from its respective vocalic position to the same virtual target point A until it is stopped by the palate or the velum, the different contact locations can be explained in a purely geometrical way. In reality, the tongue body often follows an elliptical movement path instead of a straight line [7, 8], but the principle remains the same. The higher the position of the virtual target above the contact area, the more the contact locations would differ for [g] in the context of front and back vowels. A virtual target height of about 10 mm as in Fig. 2 appears to be a good approximation for the measured contours.

In the context of articulatory modeling, virtual targets were already previously used by Kröger [9] and Perrier et al. [8]. In these studies, the virtual targets for the tongue body and the tongue tip were assumed to lie only slightly (1–2 mm) above the palate.

2. Articulatory synthesis of plosive-vowel syllables

In contrast to the previous studies by Kröger [9] and Perrier et al. [8], who assumed virtual consonant targets about 1–2 mm beyond the contact location, our own experiments and the work by Heike [4] suggest virtual targets in the region around 10 mm beyond the contact location. In the following, the virtual targets around 1 mm beyond the contact locations will be called *close virtual targets*, and the targets around 10 mm beyond the contact location *distant virtual targets*. In this study, we examined the effect of close and distant virtual consonant targets on the acoustics (this section) and the intelligibility (Sec. 3) of synthetic plosive-vowel syllables.

2.1. Method

We synthesized a corpus of 18 CV syllables with all combinations of the consonants [b,d,g,p,t,k] and the vowels [a,e,o]. Each syllable was synthesized once with a close virtual consonant target and once with a distant consonant target, resulting in 36 items. One mm and 10 mm were chosen as representative virtual target distances for the two groups. The articulatory synthesizer used to generate the stimuli was progressively developed during the last years [10, 11, 3, 12, 13]. A major part of the synthesizer is a three-dimensional geometrical model of the vocal tract [11]. For the present study, the model was extended as follows. Firstly, virtual positions were allowed to be assigned to the tongue tip and the tongue body, i.e. positions above the palatal wall. When a virtual position is assigned to one of these articulators, it is geometrically positioned in direct contact with the palate at the location with the smallest distance to the corresponding virtual point. Secondly, the control parameter that defines the vertical distance between the upper and lower lip was allowed to be set to negative values that represent different degrees of tissue compression for the closed lips.

The basic idea of our modeling approach was to use the same sigmoidal trajectory shape to control the movement of *both* the primary and the secondary articulators during the transition from the plosive to the vowel. As we discussed in Sec. 1, a sigmoidal curve is a good choice for secondary articulators. The exponential-like transitions towards and away from a consonantal closure that are observed for a primary articulator result from the *clipped* sigmoidal trajectories that arise automatically when the articulator is stopped by the walls of the vocal tract model. Mathematically, the trajectories were modeled as step responses of a linear tenth-order system according to [14].

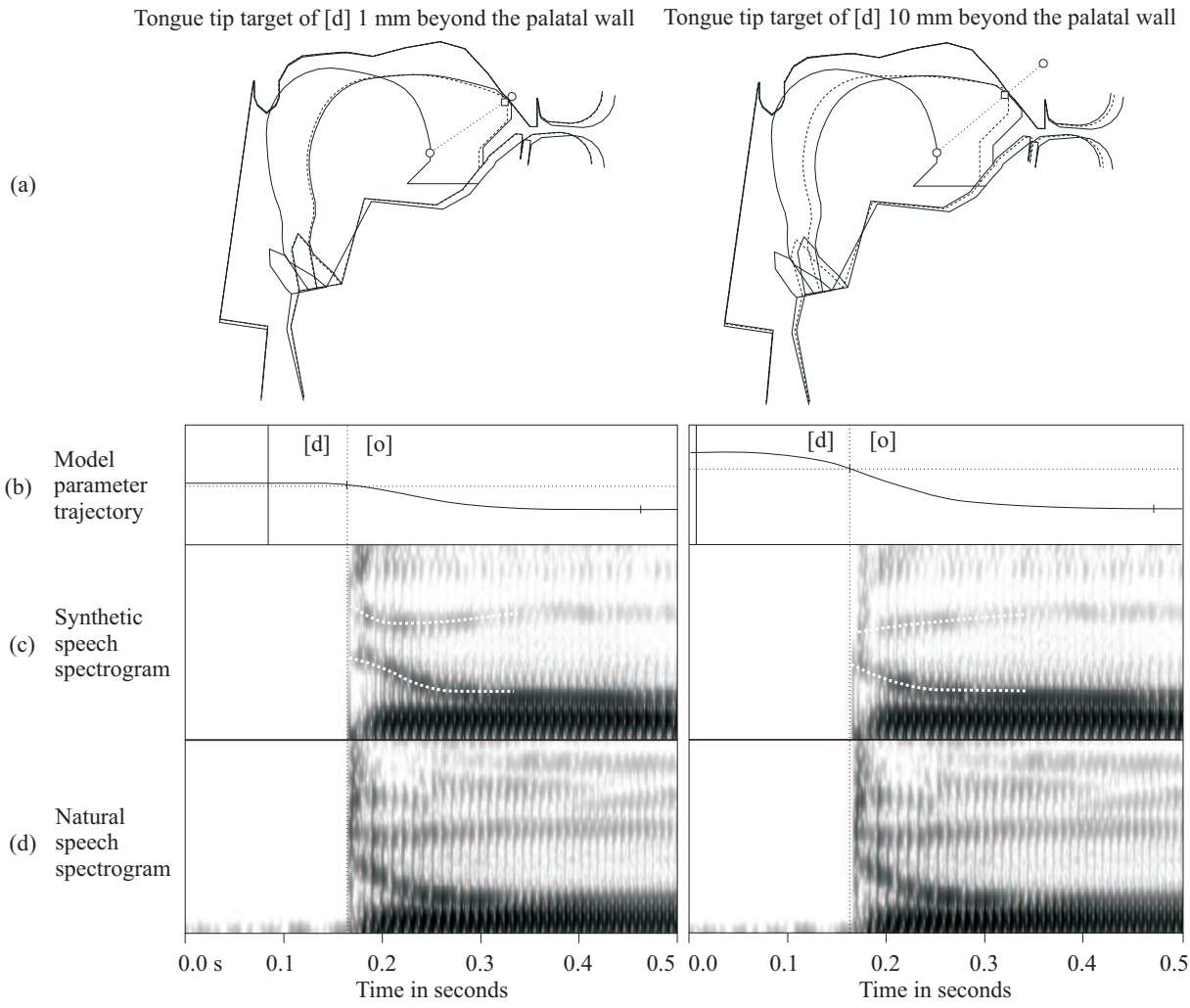

Figure 3: *Vocal tract contours (a), Control parameter trajectories (b), synthetic speech spectrograms (c), and natural speech spectrograms (d) for the syllable [do], synthesized with a close virtual consonant target on the left, and a distant virtual target on the right. The dashed vocal tract contours show the shape at the time of the closure offset.*

Figure 3 b shows the typical sigmoidal shape of these curves. The duration of the transition could be controlled by the time constant of the linear system [14]. In our simulations, all articulators for a given syllable were controlled synchronously, i.e. they started and finished transitioning at the same time.

For each syllable, both the consonant and the vowel were represented by a fixed vocal tract target shape. These shapes were determined previously by means of volumentric and dynamic MR images [3]. Because the vocal tract shape for consonants varies depending on the following vowel (coarticulation), a vowel-dependent vocal tract shape was used for each consonant and context vowel. Furthermore, two variants were created for each consonant shape – one with a close virtual target for the primary articulator, and one with a distant virtual target.

For the synthesized syllables, the duration of the formant transition phase and the voice onset time (for voiceless plosives) was reproduced as closely as possible from natural recordings. The duration of the formant transition phase was adjusted by varying the time constant of the linear system used to generate the articulatory trajectories. The voice onset time was adjusted by varying the relative position of the glottal closing gesture (for voiceless plosives). Both parameters were adjusted manually for the best visual fit between the spectrograms of the natural and synthetic syllables.

2.2. Results

Figure 3 illustrates the effect of the different virtual consonant targets for the synthesis of the syllable [do]. The left column shows the synthesis with the close virtual consonant target, and the right side with the distant target. Figure 3 a shows three vocal tract contours for each variant: the contours at the beginning and the end of the syllable (solid lines), and the contours at the time of the closure offset (dashed lines). The circles mark the (virtual) positions of the tongue tip at the beginning and the end of the syllable. During the transition, the tongue tip moves along the dotted line that connects both circles. For virtual positions, the tongue tip is geometrically set to the closest real position in contact with the palate (white box).

Because the tongue tip starts the CV transition from a higher virtual position on the right side of the figure, the transition must start earlier (solid vertical line in Fig. 3 b) with respect to the closure release (dotted vertical line in Fig. 3 b) than on the

left side. This has the following consequences: (1) The tongue tip has a higher velocity right after the closure offset, making the (clipped) tongue tip trajectory more exponential-like as observed for natural trajectories. This directly affects the shape of the formant trajectories after the closure, as indicated by the white dotted lines in Fig. 3 c. It appears that the formant trajectories resulting from the distant virtual target are more similar to those of the natural utterance in Fig. 3 d. Furthermore, a rapid increase of the area after the closure release results in a different burst than for a slower area increase rate [15]. (2) At the time instant of the closure release, the secondary articulators are displaced further towards their vowel targets, if a distant virtual target is used. This is illustrated by the dotted vocal tract contours in Fig. 3 a and affects the formant frequencies right after the closure release.

All syllables in our corpus were successfully synthesized in a way corresponding to the example presented. For the majority of syllables, the synthesis with distant virtual consonant targets allowed a better spectrographic fit than with the close virtual consonant targets.

3. Identification experiment

The basic question for the perception experiment was, whether or not the better spectrographic reproduction of natural syllables with the distant virtual consonant targets improves the intelligibility of the synthetic syllables. Therefore, all 36 stimuli of the corpus described in Sec. 2 were presented to 13 subjects in randomized order over earphones. The subjects were asked to identify the perceived syllable after each stimulus. When a syllable was perceived as ambiguous, they were asked to state their most probable guess. If requested by the subject, each stimulus could be repeated once.

The vowels were correctly identified in all stimuli by all subjects. However, the consonants were often misunderstood, for example, [ta] was sometimes heard as [ka] or [kɐ] as [pɐ]. Table 1 summarizes the identification errors sorted by the intended consonants.

Table 1: *Absolute number and percentage of misunderstood stimuli.*

Syllable	Close virtual target	Distant virtual target
/pV/	22 (56.4%)	16 (41.0%)
/tV/	9 (23.1%)	8 (20.5%)
/kV/	19 (48.7%)	23 (59.0%)
/bV/	4 (10.3%)	4 (10.3%)
/dV/	1 (2.6%)	4 (10.3%)
/gV/	4 (10.3%)	5 (12.8%)
Sum	59 (25.2%)	60 (25.6%)

The overall error rate did not differ significantly for both virtual target locations. In general, the voiced plosives were better recognized than the voiceless plosives. The synthetic syllables [kV] and [pV] were most frequently confused. Mostly, the [k] was heard as [t] and the [p] as [g] or [k].

4. Discussion and conclusions

Our synthesis experiments with close and distant virtual consonant targets showed that the latter allow a better spectrographic fit of natural plosive-vowel syllables. However, for the intelligibility of the synthetic syllables, the virtual target location appeared not to play a role. For a more detailed evaluation of the perceptual effect of the different virtual target locations, it would be conceivable to not only compare the perceived syllable with the intended syllable, but also to let the listeners rate the degree of ambiguity. The high error rates for voiceless plosives indicate that other aspects of the synthesis that influence their perception need improvement, like the synthesis of aspiration and frication noise, and the onset characteristics of vocal fold vibration.

5. Acknowledgements

This work was supported in part by the German Research Council DFG Grant Nr. KR 1439/15-1.

6. References

[1] S. Fagel, *Audiovisuelle Sprachsynthese*. Logos Verlag Berlin, 2004.

[2] P. Mermelstein, "Articulatory model for the study of speech production," *Journal of the Acoustical Society of America*, vol. 53, no. 4, pp. 1070–1082, 1973.

[3] P. Birkholz and B. J. Kröger, "Vocal tract model adaptation using magnetic resonance imaging," in *7th International Seminar on Speech Production (ISSP'06)*, Ubatuba, Brazil, 2006, pp. 493–500.

[4] G. Heike, "Sprachsynthese auf der Basis eines Artikulationsmodells," in *Fortschritte der Akustik, DAGA '78*, Bochum, Germany, 1978, pp. 467–470.

[5] A. Löfqvist and V. L. Gracco, "Control of oral closure in lingual stop consonant production," *Journal of the Acoustical Society of America*, vol. 111, no. 6, pp. 2811–2827, 2002.

[6] P. Birkholz, B. J. Kröger, and C. Neuschaefer-Rube, "Model-based reproduction of articulatory trajectories for consonant-vowel sequences," *submitted*.

[7] C. Mooshammer, P. Hoole, and B. Kühnert, "On loops," *Journal of Phonetics*, vol. 23, pp. 3–21, 1995.

[8] P. Perrier, Y. Payan, M. Zandipour, and J. Perkell, "Influence of tongue biomechanics on speech movements during the production of velar stop consonants: A modeling study," *Journal of the Acoustical Society of America*, vol. 114, no. 3, pp. 1582–1599, 2003.

[9] B. J. Kröger, *Ein phonetisches Modell der Sprachproduktion*. Max Niemeyer Verlag, Tübingen, 1998.

[10] P. Birkholz, *3D-Artikulatorische Sprachsynthese*. Logos Verlag Berlin, 2005.

[11] P. Birkholz, D. Jackèl, and B. J. Kröger, "Construction and control of a three-dimensional vocal tract model," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP'06)*, Toulouse, France, 2006, pp. 873–876.

[12] ——, "Simulation of losses due to turbulence in the time-varying vocal system," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 4, pp. 1218–1226, 2007.

[13] P. Birkholz, "Control of an articulatory speech synthesizer based on dynamic approximation of spatial articulatory targets," in *Interspeech 2007 - Eurospeech*, Antwerp, Belgium, 2007, pp. 2865–2868.

[14] K. Ogata and Y. Sonoda, "Reproduction of articulatory behavior based on the parameterization of articulatory movements," *Acoustical Science and Technology*, vol. 24, no. 6, pp. 403–405, 2003.

[15] K. N. Stevens, *Acoustic Phonetics*. The MIT-Press, Cambridge, Massachusetts, 1998.