LABIALIZATION DURING /k/ FOLLOWED BY A ROUNDED VOWEL IS NOT ANTICIPATION BUT THE AUDITORILY REQUIRED ARTICULATION

Shinji MAEDA
Ecole Nationale Supérieure de Télécommunication, Département TSI
and Centre National de la Recherche Scientifique (CNRS), URA 820
Paris, France

ABSTRACT

Vowel-Consonant-Vowel (V_1CV_2) sequences are synthesized using a vocal-tract synthesizer. A deliberately simple phoneme concatenation by interpolation of their target area functions is employed to derive the time-varying area function, which is fed into the synthesizer. The sequences involve three vowels, /i/, /a/, and /u/, and stops. First, a uniform tube having a single constriction section at an appropriate place is used as the target area for the consonant. Listening tests indicate that when V_2 is /u/, the synthesized [k] is not always satisfactory. Second, in order to improve the phonetic value of [k], we modify the consonantal target to take into account for the "anticipatory" effects of [u], such that the lip section during the k-closure is already rounded. After the modification, [k] is judged highly intelligible by all the listeners. It is concluded then that the libialization is not anticipatory coarticulation, but the auditorily required articulation for [k].

1. INTRODUCTION

In speech production, movements of individual articulators such as the jaw, tongue, and lips are not synchronized to each other [1]. Because of this asynchrony, acoustic characteristics of successive phonemes will be fused, *i.e.* coarticulated. Moreover, different articulators can acoustically compensate for each other in the production of certain phonemes [9, 10]. Inter-articulator asynchrony and compensation make articulatory movements immensely variable and complex.

It seems reasonable to question, however, whether all of complex movements are perceptually relevant and are necessary to convey the identity of phonemes. Some of observed movements could be the consequence of constraints imposed by the biomechanical and neurophysiological machinery. In fact, a study by Carré et al. [2], for example, has suggested that in vowel identification tasks, the ears are not so sensitive to certain articulatory variations, including an inter-articulator phasing (or asynchrony). It could well be that certain articulatory movements are critical in encoding phoneme identity into streams of speech sounds and other movements are not. How do we distinguish the critical movements from the non-critical ones?

In order to answer this question, we are carrying out V_1CV_2 synthesis experiments using a vocal-tract syntesizer. The idea is the following: Time-varying area function for a V_1CV_2 sequence is calculated by temporarily interpolating between the two target area functions of successive phonemes. The whole part of the vocal tract varies synchronously from one configuration to another. There is no particular spatiotemporal organization here, but only smooth temporal transitions from one sound to another. Such extremely simple scheme is bounded to fail, at least for some V_1CV_2 sequences, which is exactly the point of this experiment. If a synthesized sequence is not correctly and "easily" identifiable by listeners, some modifications in the specification of time-varying area function are in order. The necessary sophistication that makes it closer to a more realistic articulation signifies a perceptually critical articulatory maneuver.

In this paper, we focus our attention at an [aku] sequence. A particularity of this sequence is that the lip is often rounded already during the k-closure. Many authors interpret this rounding as an anticipatory coarticulation of the following vowel [u] (e.g., page 378 in [7] and [18]). We would like to show that the rounding is a part of the explicit articulation necessary for the consonant to be perceived as [k].

2. VOCAL TRACT SYNTHESIZER

Since, the human vocal tract constitutes a narrow tube, the generation and propagation of sounds inside the tract can be described by a set of one-dimensional aerodynamic and acoustic equations, and equivalently simulated by a lumped transmission line [e.g., 3, 5, 8]. The main reason for using the acoustic simulation is that it allows a straightforward segment concatenation with area functions, which are only the input to the synthesizer. The vocal tract during speech production is nothing but a smoothly time-varying acoustic tube. Such a physical characteristic can be mimicked by smooth temporal interpolations between successive target area functions, for example, by a cosine law. The burst and fricative noise at the supraglottal constriction can be automatically generated whenever the aerodynamic condition is appropriate for. It may be noted that targets for both consonants and vowels are specified by the same area functions having different shapes, the interpolations therefore can be explicitly defined across a vowel and a consonant. In a formant synthesis, for example, this is not always the case.

In the vocal tract synthesizer, the noise is "automatically" generated as a function of tract shape, *i.e.*, of area function, as it occurs in the human vocal tract. In synthesis, the noise is nothing but a band-pass filtered sequence of random numbers, which is injected at the exit of the constriction or at some point

in the downstream from the constriction [6, 12, 13]. The magnitude of noise is modulated by a function of the cross sectional area of the constriction and the airflow level. According to a square law proposed by Flanagan [5], the magnitude is proportional to the square of airflow and inversely proportional to the cross-section area of the constriction.

The airflow which is necessary to determined the noise magnitude can be calculated by using a so-called low frequency model [e.g., 6, 11, 12], where airflow is determined by the function of the sub-glottal air pressure, P_S, which is fixed at 8 cmH_2O in our simulation, and the flow resistance at the two major constrictions, one at the glottis and the other in the supraglottal tract. The flow resistance can be approximated by the sum of the Bernoulli kinetic resistance and viscous resistance, which is a function of the constriction geometry and the airflow itself.

The value of the scaling coefficient for the noise magnitude is empirically determined so that the level of the synthesized burst noise relative to that of surrounding vowels is realistic. Although the coefficient value (and also the spectral characteristics of noise source) varies in a function of detailed constriction shapes and of airflow level [e.g., 13, 14, 16], we use a fixed scaling and a fixed noise-source spectrum shape for different consonants, just for the sake of simplicity.

3. TARGET INTERPOLATION

3.1. Area functions

Each phonemic segment (henceforth, 'phoneme' for short) is specified by a predetermined area function as its target. The area function is specified by a fixed number of sections in which the k-th section is defined by the cross-sectional area, A(k, t), and the length, x(k, t). Obviously, we must have transition period from one phoneme to another so that the area function smoothly varies between phonemes. The interpolation of A and x is done section by section synchronously using cosine law. The stationary part of each phoneme is specified by the same target value at the onset and offset. This is all we need to concatenate V_1CV_2 sequences.

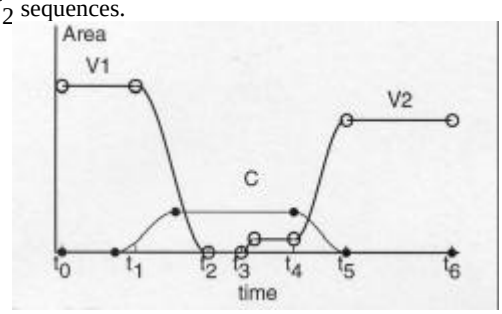

Figure 1. The temporal variation of a vocal-tract section area, corresponding to the constriction, in V_1-stop-V_2 is indicated by the thick line with open circles. The thin line with filled circles indicates the variation of a slow time-varying component of the glottis area. The markers specify points at which a target area is specified. Note that vocal-tract sections and the glottis are not necessarily always synchronized.

Figure 1 illustrates a detail temporal pattern of a section corresponding to the consonantal constriction.. The target area of the initial vowel (V_1) is specified at the onset (t_0) and offset point (t_1), indicated by the corresponding open circles. During this interval, the section area is kept at the target value. The area becomes zero at the closure onset (t_2). The transition from the V_1 offset to the closure onset is specified by the cosine law mentioned before. The stop closure duration is fixed to 80 ms and release duration to 17 ms with the transition interval of 3 ms in between. The total consonant duration equals to 100 ms. Note that a stop consonant is specified by two target area function, one for closure and the other for release. The voice onset of V_2 occurs always at 230 ms, the VOT becomes 30 ms regardless of place of articulation in this sturdy. Since the second syllable (CV_2) is stressed, the V_2 duration is fixed to 150 ms, which is considerably longer than that of V_1.

It may be noted here that the target area function for a consonant is a uniform tube having a constriction shaped by a single section at an appropriate place for that consonant. Such a model has been employed by Stevens [15] in the theoretical analysis of consonantal acoustics. In principle, a variety of consonants can be synthesized by just varying the position of the constriction section along the length of the vocal tract. As described later however, for certain consonants in a certain vowel context, it was necessary to modify the target area as a function of V_2 identity. For vowels, we use realistic area functions as their targets.

3.2. Glottal section

The temporal pattern of the glottal section is specified by the sum of slow and fast time-varying components. The muscular adjustments in the laryngeal system determine the slow component A_{g0}. When certain aerodynamic and biomechanical conditions are met, the vocal folds vibrate, which is described by the fast pulsating oscillation of the glottal section in the simulation.

The adjustment of A_{g0} is important for the generation of the burst and fricative noise. Since the airflow inside the vocal tract is approximately equal regardless of the position and the noise magnitude is inversely proportional to the constriction area, as mentioned earlier, the burst and frication noise dominate over the aspiration noise at the glottis only when the glottis area becomes greater than the supra glottal constriction area, as shown in Figure 1. During the stationary part of vowels, the value of A_{g0} is kept at zero in this synthesis experiment with a male voice.

The vocal fold oscillation is specified by using a descriptive model proposed by Fant [4]. Since we assume a fixed glottal pulse shape, a pulse train is determined by only two parameters, voice fundamental frequency (F0 Hz) and peak value of the glottal pulse, Ap (cm^2). In the synthesis, we specify the target values of these two parameters at appropriate transition points and then their values at any given time is calculated by linear interpolations. An example of calculated temporal variations of glottis section are shown in Figure 2a.

4. SYNTHESIS OF [aku]

The target area function for [k] is specified by modifying a uniform tube so that it has a constriction at 6 cm from the lip opening. Figure 2a shows the temporal variations of the glottal area Ag and the constriction area Ac superimposed for [aku]. Ac is visible only at the vicinity of the closure-release portion. The calculated airflow using the low-frequency model and radiated sound are indicated, respectively in (b) and in (c). An informal listening indicated that the quality of this [aku] token was inadequate. This failure is puzzling, because when the V_2 =[a], this token sounds perfectly [aka]. What went wrong?

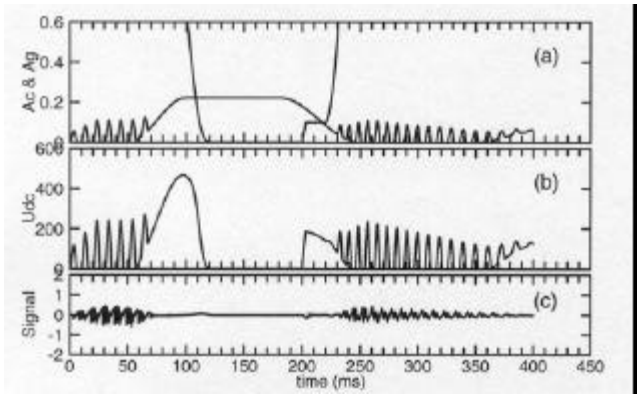

Figure 2. Simulation of [aku]: Temporal variations of specified glottal area, Ag cm^2, and constriction area, Ac cm^2, in (a), airflow, Udc cm^3/s, calculated by a low-frequency model in (b), and radiated sound signal in arbitrary units in (c).

It appears that the occurrence of a prominent peak (or a concentration of energy) in the k-burst spectrum at the vicinity of F2 onset of V_2 is the critical attribute for the consonant to be perceived as a [k] [17]. This was the case for the successful [aka] token. Moreover, the prominent peak of [k] followed by [u] in natural utterances often occurs around 1 kHz, which is close to the F2 vowel onset, as shown in Figure 3a as an example. The synthetic [aku] sequence lacks this critical attribute in its spectrogram, as seen in Figure 3b. The prominent peak, corresponding to the first quarter-wave length resonance of the front cavity, occurs about at 1.5 kHz, which seems too high.

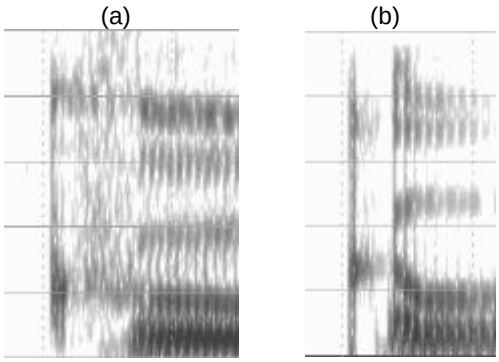

Figure 3. Spectrograms of consonant-vowel segments extracted from natural [aku] utterance in (a) and from synthetic one with a 6 cm long uniform front cavity for the consonant in (b). (The horizontal grids are spaced by 1 kHz, and vertical grids, dashed lines, by 100 ms.)

The most acoustically effective means to lower the first resonance frequency of the front cavity is to form a constriction at the lip opening, forming a Helmholtz resonator. As mentioned before, the rounded lip during the k-closure is often observed in natural speech when it is followed by [u]. As shown in Figure 4, with the constricted lip section, the prominent burst peak is now positioned close to the F2 onset frequency of the vowel [u]. Having the front cavity length of 6 cm, such low first resonance frequency is possible only by the concomitant constrictions at the lips and in the velar region to form the Helmholtz resonator. In an informal listening, this synthetic token is perceived as [aku].

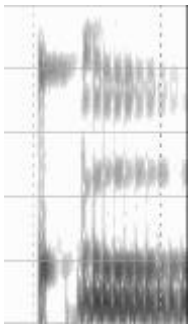

Figure 4. Spectrogram of consonant-vowel segment extracted from a synthesized [aku] sequence with a 6 cm long uniform front-cavity having a constricted lip section, which is indeed perceived as [aku].

5. PERCEPTION TEST

In order to confirm the informal impression of these two synthesized [aku] tokens, listening tests are conducted. To distinguish two tokens, which are different only in the front cavity geometry of target area function of the intervocalic consonant, let us use the following notations, 'aC_6u' for one with the 6 cm long uniform tube and '$aC_{6r}u$' for the other with the uniform tube having the same length but with the constricted (i.e., rounded) lip section.

Since it is not so appropriate to perform tests with only a single pair of stimuli, other two pairs of tokens related to [aki] and [aka] are added. In one case for [aki] tokens, a 5 cm long uniform tube is used as the front cavity of the consonant target and denoted as 'aC_5i'. Due to its uniform front cavity, the first prominent spectral peak of the burst, corresponding to the first quarter-wave length resonance, should occur about at 1.75 kHz. This peak position is too low in comparison with the F2 onset frequency of the second vowel [i]. We expect therefore a poor rating of this token. In other case, the front cavity having the same length is tapered out to form a conical horn, denoted as '$aC_{5c}i$'. The conical expansion grossly approximates the tapered opening of the front cavity after constriction, which occurs during speech production. The acoustic consequence of this modification is that all the resonance of the front cavity would shift up to higher frequencies [3]. In our simulation, the first prominent spectral peak appears at 2.4 kHz, which is somewhere between F2 and F3 of the second vowel [i]. We expect, therefore, a high rating for this token as an [aki] utterance.

In [aka] tokens, the constriction section is place at 6 cm from the lips in one case, denoted as 'aC_6a', and at 2 cm in the other case, noted as 'aC_2a', resulting [aka] and [ata], respectively. When the second vowel is [a], the simple uniform tube as the front cavity is adequate as the target area functions of these two consonants. In total, we have six different stimuli.

Eight listeners participated in the test. The stimuli are repeated 10 times in random order. The listeners are asked to judge, after each presentation of a stimulus token, whether the intervocalic consonant is a good [k], [t], or something else. The percentage scores across listeners are listed in Table 1. The full score (100%) corresponds to 80 votes (=10 repetitions x 8 listeners).

	'a $C_{6r}u$'	'a C_6 u'	'a C_{5c} i'	'a C_5 i'	'a C_6 a'	'a C_2 a'
[k]	98	29	94	41	94	4
[t]	1	51	0	4	4	96
else	1	20	6	55	2	0

Table 1. Percentage scores across eight listeners. The subscript of 'C' of stimulus labels indicates the length of front cavity (in cm) of the consonant target area function. (See text for detail).

Since the consistency of responses of the individual listeners is relatively high, we interpret the scores such that all the listeners judges the stimulus synthesized with the front cavity having rounded lip, 'a C_{6r}u', as [aku]. The stimulus with uniform front cavity, 'a C_6 u', is judged either [t], [k], or something else depending on listeners. The result confirm, therefore, that a "clear" [k] requires the lip rounding when it precedes [u].

Similar results for [aCi] tokens: with the conical front cavity, 'a C_{5c} i', all the listeners judged the stimulus as [k], but with the uniform front cavity, 'a C_5 i', as [k] or something else. In the case of [aCa] tokens, all the listeners judged [k] when the front-cavity length is 6 cm, and [t] when the length is 2 cm.

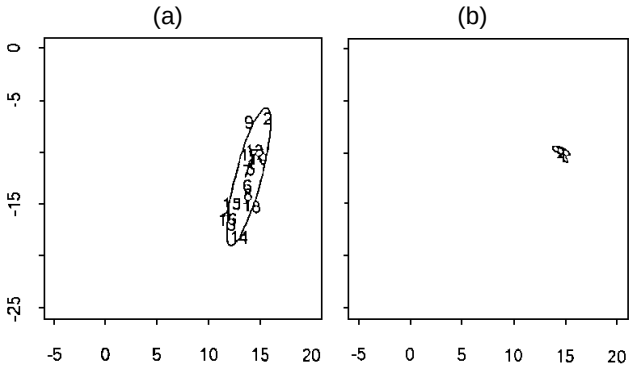

Figure 5. Lower lip position during /k/ productions in various contexts with many repetitions shown in (a), and during /k/ followed by /u/ (in "school") with many repetitions in (b). The x-y coordinates are in mm. (These figures were kindly prepared by Jim S. Dembowski at University of Wisconsin-Madison, using "Wisconsin X-ray microbeam speech production database".)

6. CONCLUDING REMARKS

If the lip rounding during [k] were auditorily required when it precedes the rounded vowel [u] the "anticipatory" rounding maneuver, intentionally or not, must be controlled and articulated. If this is the case, we should expect in articulatory data that the variability of the "articulated" rounded lip position during [k] is small when it precedes the vowel [u]. In contrast, the dispersion of non-controlled lip positions before others vowels should be large, because there is no acoustic and auditory reason to precisely articulate the lips during [k]. The data shown in Figure 5 seem to support our assertion: The lower lip position during [k] production in various contexts, shown in Figure 5a, exhibits a large dispersion. Whereas the dispersion of lip positions for [k] extracted form many tokens of the English word "school" is vary small as shown in Figure 5b. The same contrastive dispersions were observed for the upper lip position. We interpret this small dispersion as an indication that the speaker articulated the rounded lips, because it is perceptually required.

In conclusion, the acoustical considerations, perception test, and articulatory data indicate that the labialization of [k] followed by [u] is not a "simple" anticipatory coarticulation, but an auditorily required explicit articulation.

REFERENCES

[1] Bouabana, S. and Maeda, S. 1998. Multi-pulse LPC modeling of articulatory movements. Speech Communication, 24, 227-248.

[2] Carré R., Chennoukh S., Jospa, P. and Maeda S. 1996. The ears are not sensitive to certain coarticulatory variations: Results from VCV synthesis/perceptual experiments. In *Proceedings of the 1st ESCA Tutorial and Research Workshop on Speech Production Modeling: From control strategies to acoustics and the 4th Speech Production Seminar: Models and data*, Autrans (France), 13-16.

[3] Fant, G. 1960. *Acoustic theory of speech production.* Mouton.

[4] Fant, G. 1979. Glottal source and excitation analysis. *STL-QPSR*, No.1, 85-107.

[5] Flanagan, J.L. 1972. *Speech analysis, synthesis, and perception.* Springer, New York.

[6] Ishizaka, K. and Flanagan, J.L. 1972. Synthesis of voiced sounds from a two-mass model of the vocal cords. *Bell System Technical Journal*, **50**, No. 6, 1233-1268.

[7] Laver, J. 1994. *Principles of phonetics*. Cambridge University Press.

[8] Maeda S. 1982. A digital simulation method of the vocal-tract system. *Speech Communication*, **1**, 199-229.

[9] Maeda, S.(1991. On articulatory and acoustic variabilities. *Journal of Phonetics*, **19**, 321-331.

[10] Perkell, J.S., Matties, M.L., Svirsky, M.A., and Jordan, M.I. 1993. Trading relations between tongue-body raising and lip rounding in production of the vowel /u/: A pilot motor equivalence study. *Journal of the Acoustical Society of America*, **93**, 2948-2961.

[11] Rothenberg, M. 1968. The breath-stream dynamics of simple-released-plosive production. *Bibliotheca Phonetica*, No.6, Karger, Basel.

[12] Scully, C. 1986. Speech production simulated with a functional model of the larynx and the vocal tract. *Journal of Phonetics*, **14**, 407-414.

[13] Shadle, C.H. 1985. *The acoustics of fricative consonants*. M.I.T. PhD thesis, Department of Electrical Engineering and Computer Science; R.L.E. Technical Report 506.

[14] Stevens, K.N. 1971. Airflow and Turbulence noise for fricative and stop consonants: Static considerations. *Journal of the Acoustical Society of America*, **50**, No.4, 1180-1192.

[15] Stevens, K.N. 1989. On the quantal nature of speech. *Journal of Phonetics*, **17**, 3-45.

[16] Stevens, K.N. 1993. Modelling affricate consonants. *Speech Communication*, **13**, 33-43.

[17] Stevens, K.N. and Blumstein, S.E. 1978. Invariant cues for place of articulation in stop consonants. *Journal of the Acoustical Society of America*, **64**, 1358-1368.

[18] Woods, S.A.J. 1997. A cinefluorographic study of the temporal organization of articulator gestures: Examples from Greenlandic. *Speech Communication*, **22**, Nos. 2-2, 207-225.