Articulatory VCV Synthesis from EMA Data

Asterios Toutios, Shinji Maeda

CNRS LTCI; TELECOM ParisTech, Paris, France

{toutios,maeda}@telecom-paristech.fr

Abstract

This paper reports experiments in synthesizing VCV sequences with French unvoiced stop or fricative consonants, using a time-domain simulation of the vocal-tract system. The necessary dynamics of the vocal-tract shape are derived in two steps: first, time-varying parameters of an articulatory model are calculated automatically from electromagnetic articulography (EMA) data, using a method previously published by the first author; second, semi-automatic corrections are applied to properly account for consonantal events. Time-varying characteristics of the glottis are set using empirical rules. Friction noise is generated along the length of the whole vocal-tract instead of locally at the narrowest constriction. Spectrogams of satisfactory synthesis results are presented alongside those of real speech recorded simultaneously with the articulatory data. Corresponding audio files are available online.

Index Terms: speech production, speech synthesis, articulatory model, electromagnetic articulography

1. Introduction

Articulatory synthesis, i.e. the generation of speech by simulating the vocal-tract system, is interesting from both the scientific and technological viewpoints. From the scientific viewpoint, it is an indispensable tool for the study of the articulatory-acoustic relationship toward a better understanding of human speech production [1]. From the technological viewpoint, it promises an alternative approach to text-to-speech synthesis that would overcome certain shortcomings of current approaches [2].

In order to synthesize a speech sound this way, two elements are needed. The first one is an articulatory-to-acoustic simulation method of the vocal-tract system. Such a simulation has been previously developed by the second author [3]. The second element needed is a description of the vocal-tract shape, glottal and noise sources. In broad terms, for most isolated static speech sounds of major languages, such descriptions are today part of standard phonetic knowledge.

The problem is more complicated for continuous speech. It is well-known that continuous speech cannot be considered as the result of a straightforward concatenation of static speech sounds, neither at the acoustic nor at the articulatory level [4]. Furthermore, an important class of sounds, stop consonants, are strictly dynamic events that have no specific static equivalents. Ideally, for articulatory synthesis of continuous speech, data on the corresponding vocal-tract dynamics should be available. But such data have been scarce and/or not directly usable.

One technique that records information on the dynamics of speech production is electromagnetic articulography (EMA) [5]. However, this information concerns only the positions of a few sensors attached on articulators and does not provide *per se* a complete information for speech synthesis. This paper focuses on how to complement the missing information in order to synthesize speech. As described, some information is extrapolated from EMA data and other by rules. We build upon two previous works of ours: the first [6] presented a setup for VCV synthesis and a series of experiments where the vocal-tract dynamics were derived by interpolation between target shapes for the three phonemes; the second [7] presented a method for estimating the dynamics of an articulatory model [8] from EMA. To illustrate our ideas, we synthesize French V_1CV_2 sequences where V_1 and V_2 are one of /a/,/i/,/u/ and C is one of the unvoiced stops /p/,/k/,/t/ or fricatives /f/,/s/,/ʃ/. This is a step toward VCV synthesis covering a complete inventory of French consonants and vowels which we view, in turn, as a prerequisite to building a full text-to-speech system based on the articulatory synthesis paradigm.

In what follows, we will present our articulatory synthesis method with some emphasis on elements that are new or different with respect to the aforementioned previous works. We will show and discuss spectrograms of our synthesized sequences presented alongside spectrograms of the real speech recorded simultaneously with EMA. Audio files of our results can be found online at http://sites.google.com/site/toutios/vcvis2012/.

2. Data and Methods

In order to synthesize a speech sequence with our method, we begin with corresponding EMA data. From these, we calculate the dynamics of our articulatory model using an automatic method. These are then converted into area-function dynamics, which are subsequently corrected to better account for consonantal events. The corrected area function dynamics are supplemented by contours of the time-varying characteristics of the glottis, which we determine using empirical rules. Finally, friction noise is appropriately added using simple aerodynamic rules.

2.1. EMA dataset

EMA data corresponding to 54 V_1CV_2 sequences covering all the combinations of vowels /a/,/i/,/u/ and consonants /p/,/k/,/t/,/f/,/s/,/ʃ/ were recorded in LORIA, Nancy, using the AG500 articulograph [9], as part of a larger EMA dataset. The subject was a phonetically aware native French male speaker. The data concerned the three-dimensional dynamics of four sensors attached along the surface of the tongue on the mid-sagittal plane, from the apex to the vicinity of the velum, and sensors on the lower incisor, lower lip, upper lip and lip corners. Additional sensors on the bridge of the nose and behind the ears were used for head movement correction. The sample rate of these data was 200 Hz, however for the work described herein they were down-sampled at 100 Hz. The audio signal was recorded simultaneously and synchronized automatically using the articulograph's internal software.

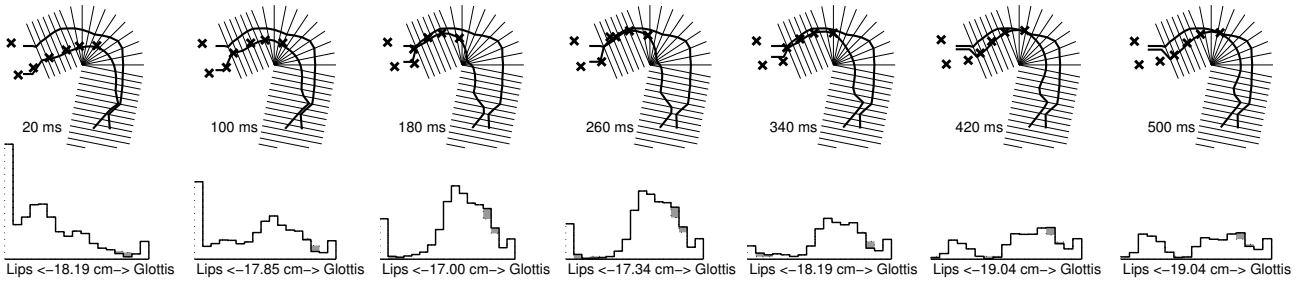

Figure 1: *[Upper part] Samples of vocal-tract model shapes during* /atu/ *derived automatically from EMA, super-posed on the semi-polar grid. Crosses indicate the positions of EMA sensors on tongue, lower incisor, upper and lower lip. [Lower part] Corresponding area functions (lines). Gray areas represent the adjustments described in Section 2.2.*

2.2. Area functions

Our articulatory model [8] describes the vocal-tract shape by means of seven parameters: jaw opening; lip opening; lip protrusion; tongue dorsum position; tongue dorsum shape; tongue apex position; and larynx height. The 7-tuples of articulatory parameters specify mid-sagittal profiles of the vocal-tract, plotted over a pre-defined semi-polar grid.

We used an updated version of a previously presented method [7] to calculate the dynamics of the jaw, lip, and tongue parameters from EMA dynamics (the exception was the larynx parameter which was considered fixed at its mean value). The updates consisted in: (i) a new formulation of the optimization problem, using the relationships between the articulatory parameters and the tongue contour at the exact positions of the tongue sensors, determined on the basis of the relationships at the grid-lines (already defined in the model), instead of drawing a spline curve over the sensors and finding its intersections with the grid-lines; and (ii) a new definition of lip protrusion on the basis of the lower lip sensor only instead of the mid-point between upper and lower lip sensors, giving more relevant results for labiodental consonants.

As a representative example of this process, the upper part of Fig. 1 shows snapshots of derived mid-sagittal vocal-tract profiles for the sequence /atu/. These EMA-derived profiles were converted to area functions, represented by 17 uniform tubes (sections). This number of sections was a good compromise between spatial precision and simplicity. The sections had equal length, however this length was subject to dynamic change. Such area functions are drawn with thick lines at the lower part of Fig. 1.

These automatically derived area functions were subjected to two additional sets of corrections, indicated by the gray areas in lower Fig. 1. The first set concerned the tongue root: The articulatory model, with a limited number of parameters, tends to give an unrealistic vocal-tract contour at the tongue root region for some extreme articulations, in that this part of the contour comes often in contact with (or very close to) the rear pharyngeal wall, thus impeding airflow (see e.g. the left-most example in Fig. 1). To avoid this problem, we discarded the area information of the corresponding sections (3rd and 4th sections from the glottis) and replaced them by interpolating linearly the areas of the adjacent sections (2nd and 5th).

The second set of corrections concerned the narrowest constriction along the vocal tract. Our automatically derived area functions were not accurate enough to distinguish between consonantal events. In broad terms, a constriction of about 0.3 cm^2 would let the airflow pass relatively freely, an area of about 0.1 cm^2 would raise friction noise, and a zero area would completely block the airflow. This lack of accuracy was not surprising since (i) the articulatory model has not been developed with such level of accuracy in mind, and (ii) EMA measurements are accurate only to a certain degree [10].

consonant	section	consonant	section
/f/	17	/p/	17
/s/	15, 16	/t/	15, 16
/ʃ/	14	/k/	10, 11, 12

Table 1: *Locations of narrowest constriction for consonants, represented by vocal-tract sections. Counting of the 17 sections begins at the glottis and ends at the lips. When more than one sections are mentioned, the constriction is formed at one of them, or at two (adjacent) of them combined, depending on vocalic context. Note that for /f/, lip protrusion is zero, and so section 17 corresponds to the area of the teeth, rather than the lips.*

Based on *a priori* phonetic knowledge and through experimentation we formulated rules about the location of the narrowest constriction of the vocal tract for every consonant. The sections, or combination of sections, where we located the constrictions are presented in Table 1. During fricatives the area at the constriction sections was set to 0.1 cm^2. During stops the area was set to 0.01 cm^2 (lowest value that did not raise problems in the numerical simulation) for the closed phase and 0.1 cm^2 for the release phase. For all consonants, and to avoid discontinuities likely to be raised by the previous adjustments, the trajectory of the area at the constriction was replaced by interpolated values during 50 msec before V_1 offset and after V_2 onset. A low threshold of 0.3 cm^2 was applied to the areas of all sections during vowels, and to all sections but the constriction section(s) during consonants.

2.3. Glottis

Glottal area is modeled as the sum of a slow and a fast-varying component [6, 11]. Very briefly, the fast-varying component is a triangular glottal pulse with amplitude A_p and fundamental frequency $F0$ which is added to a non-vibrating (slow-varying) area component A_{g0}. Amplitude A_p is non-zero only for voiced speech segments while A_{g0} should be significantly larger than the area of the narrowest vocal-tract constriction during unvoiced segments, to avoid aspiration noise [12]. Such considerations gave rise to empirical determinations of A_p and A_{g0} contours for synthesis. An example is shown in Fig. 2, which

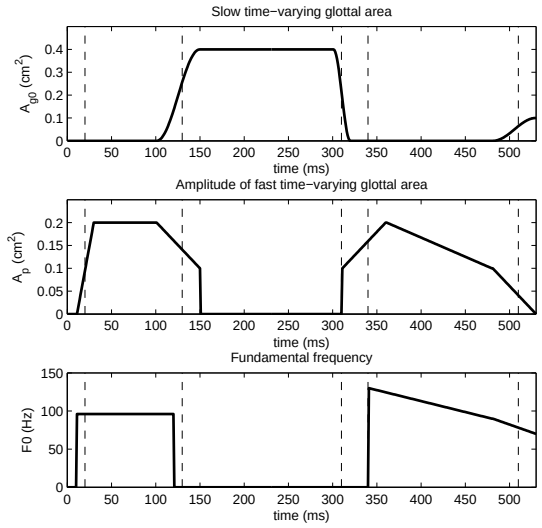

Figure 2: *Slow time-varying glottal area, amplitude of glottal pulse and F0 contours for synthesis of* /atu/. *The five vertical lines denote, from left to right:* /a/ *onset,* /t/ *closure,* /t/ *release,* /u/ *onset,* /u/ *offset.*

also shows the general shape of the F0 contour we adopted to synthesize all our VCV sequences.

2.4. Friction noise

In previous versions of the vocal-tract simulation a turbulent noise source was inserted at the exit of the narrowest constriction of the vocal tract. The gain N_{amp} of the noise magnitude was determined by a cubic law:

$$N_{amp} \propto U_{dc}^3 / A^{2.5} \quad (1)$$

where U_{dc} is a DC airflow and A is the cross-sectional area at the constriction.

We found that inserting noise locally in the vocal tract raised a very specific problem: Since we worked with real data it could happen that, during the production of a vowel, the place of the narrowest constriction moved along the vocal tract. With it, the noise source moved discontinuously, which had the audible effect of introducing a click noise in synthesis. As a solution to this problem, we introduced noise sources according to Eq. 1 in all sections of the vocal tract. U_{dc} was still defined on the basis of the area at the narrowest oral constriction (other than the glottis). The total turbulent noise was thus the sum of the noises generated at each section. Obviously, the section with the narrowest constriction still had the most important contribution.

3. Results

We synthesized all VCV sequences corresponding to the previously described EMA dataset. We did not perform any formal perceptual assessment but, in casual listening, our synthesis results sounded quite natural and intelligible. As already mentioned, these synthesis results are available online. Fig. 3 shows spectrograms of some of these results, together with spectrograms of the speech that was recorded simultaneously with EMA. In broad terms, the synthesis results seem to agree well with the recorded speech.

The formants of vowels are in a general agreement with the literature, e.g. with Lonchamp [13]. The only problem we observed was that in some cases our /u/'s, especially at V_1 position, were not very clear and could be potentially perceived as /o/'s. It is a issue we will investigate further. The formant transitions into and out of the consonants also agree with the literature, e.g. with Delattre [14], but, sometimes, they are *too* pronounced, as in the case of the /atu/ shown in Fig. 3. A probable source of this problem is the replacement of the area-function trajectory by interpolated values before V_1 offset and after V_2 onset, as already described. We used an interval with a duration of 50 msec which we should perhaps re-consider.

The spectrum of noise during fricatives is consistent with the literature, e.g. with Shadle [15]: /f/ is spread almost evenly along the frequency range; /s/ rises as we move from zero to 5kHz; and /ʃ/ presents a peak at around 3 kHz. On the other hand the spectrum at the release of stops presents a peak at very low frequencies, which is not consistent with the literature, e.g. with Blumstein [16]. If we were to remove this peak, the release spectrums of /p/ and /t/ would agree at large with Blumstein, i.e. they would be diffuse falling and diffuse rising, respectively. On the other hand, /k/ would present a peak at around 2.5 kHz which would be correct, but, still, it would not be as compact as expected.

4. Conclusion

We presented a method for synthesis of VCV sequences using a time-domain simulation of the vocal-tract system driven by EMA data. Our first results on sequences including French unvoiced stop and fricative consonants indicate the potential of this approach toward the goal of constructing a complete system that will generate speech from text by simulating the dynamics of the vocal-tract system.

In the future, we want to work toward synthesis of VCV sequences that would ideally cover the complete inventory of vowels and consonants in, at least, the French language. Oral vowels are already well covered by the articulatory model and the method to estimate its parameters. Synthesis of the voiced counterparts of the consonants presented herein should be straightforward, since all that has to change is the trajectories of the time-varying glottal characteristics. Nasal vowels and consonants are well addressed by the vocal-tract simulation, however simulating the dynamics of the velum is an open issue: one solution is to gather data on it; another solution is to use empirical rules. Synthesis of liquids is more challenging, as they have not been well covered by the articulatory model and vocal-tract simulation due to particularities in their production, i.e. the presence of two extra lateral channels for /l/ and an extra cavity for /r/. Extension to the case of VCCV sequences should pose additional challenges.

5. Acknowledgements

This work was supported, in part, by the French National Research Agency (ANR) under contract number ANR-08-EMER-001-02 (ARTIS project). We thank Yves Laprie and Slim Ouni, LORIA, Nancy, for the acquisition of EMA data.

6. References

[1] J. Vaissiere, "Area functions and articulatory modeling as a tool for investigating the articulatory, acoustic and perceptual properties of sounds across languages," in *Experimental Approaches to*

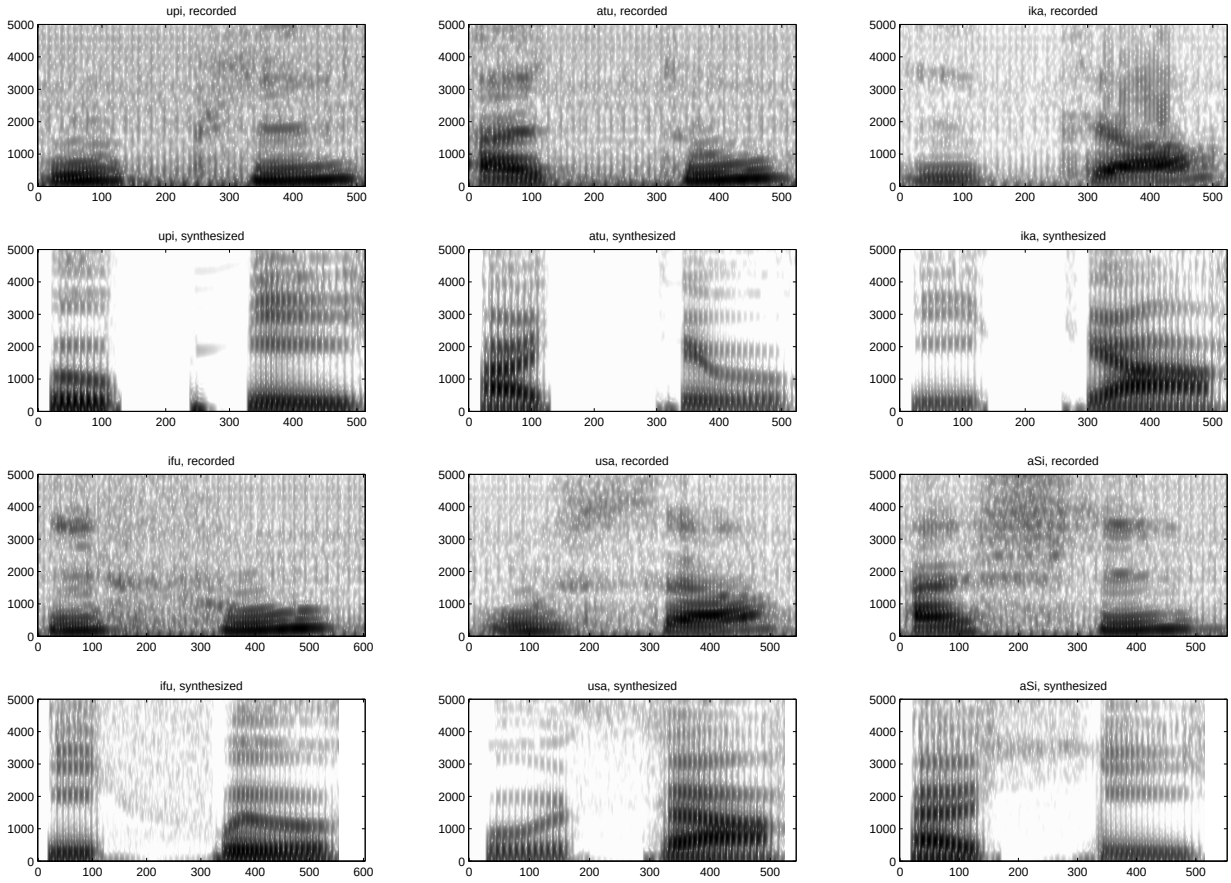

Figure 3: *Spectrograms of real speech (odd rows) and of speech synthesized on the basis of simultaneously recorded EMA data (even rows) for /upi/, /atu/, /ika/, /ifu/, /usa/ and /aʃi/. Note that there was some noise in the real speech recordings.*

Phonology, M. Solé, P. Beddor, and M. Ohala, Eds. Oxford: OUP, 2007, pp. 54–71.

[2] D. Klatt, "Review of text-to-speech conversion for English," *The Journal of the Acoustical Society of America*, vol. 82, no. 3, pp. 737–793, 1987.

[3] S. Maeda, "A digital simulation method of the vocal-tract system," *Speech Communication*, vol. 1, no. 3-4, pp. 199–229, 1982.

[4] S. Öhman, "Coarticulation in VCV utterances: spectrographic measurements." *Journal of the Acoustical Society of America*, vol. 39, no. 1, pp. 151–168, 1966.

[5] P. Schonle, K. Grabe, P. Wenig, J. Hohne, J. Schrader, and B. Conrad, "Electromagnetic articulography: use of alternating magnetic fields for tracking movements of multiple points inside and outside the vocal tract." *Brain and Language*, vol. 31, no. 1, pp. 26–35, 1987.

[6] S. Maeda, "Phonemes as concatenable units: VCV synthesis using a vocal-tract synthesizer," in *Sound Patterns of Connected Speech: Description, Models and Explanation*, A. Simpson and M. Pätzold, Eds., 1996, pp. 145–164.

[7] A. Toutios, S. Ouni, and Y. Laprie, "Estimating the parameters of an articulatory model from electromagnetic articulograph data," *Journal of the Acoustical Society of America*, vol. 129, no. 5, pp. 3245–3257, 2011.

[8] S. Maeda, "Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal tract shapes using an articulatory model," in *Speech production and speech modelling*, W. Hardcastle and A. Marchal, Eds. Amsterdam: Kluwer Academic Publisher, 1990, pp. 131–149.

[9] A. Zierdt, P. Hoole, M. Honda, T. Kaburagi, and H. Tillmann, "Extracting tongues from moving heads," in *5th Speech Production Seminar*, Kloster Seeon, Germany, 2000, pp. 313–316.

[10] C. Kroos, "Measurement accuracy in 3D electromagnetic articulography (Carstens AG500)," in *International Seminar on Speech Production*, Strasbourg, France, 2008, pp. 61–64.

[11] G. Fant, *Acoustic Theory of Speech Production*. The Hague: Mouton, 1960.

[12] C. Scully, "Speech production simulated with a functional model of the larynx and the vocal tract," *Journal of Phonetics*, vol. 14, pp. 407–413, 1986.

[13] F. Lonchamp, "Description acoustique," in *La parole et son traitement automatic, par Calliope*, J. Tubach, Ed. Paris: Masson, 1989, pp. 79–130.

[14] P. Delattre, A. Liberman, and F. Cooper, "Acoustic loci and transitional cues for consonants," *The Journal of the Acoustical Society of America*, vol. 27, no. 4, pp. 769–773, 1955.

[15] C. Shadle, "The Acoustics of Fricative Consonants," Ph.D. dissertation, Massachusetts Institute of Technology, Cambridge, MA, 1985.

[16] S. Blumstein and K. Stevens, "Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants," *The Journal of the Acoustical Society of America*, vol. 66, no. 4, pp. 1001–1017, 1979.