A Simple Hybrid Acoustic / Morphologically-Constrained Technique for the Synthesis of Stop Consonants in Various Vocalic Contexts

Frédéric Berthommier, Laurent Girin, Louis-Jean Boë

GIPSA-lab (Speech and Cognition Dept),
CNRS / Grenoble Institute of Technology, Grenoble, France
{Frederic.Berthommier, Laurent.Girin, Louis-Jean.Boe}@gipsa-lab.grenoble-inp.fr

Abstract

The predominant way to synthesize stop consonants is currently to use an articulatory model controlled by vocal tract parameters. We propose a new method to make this synthesis in various vocalic contexts. To generate the formant transitions, the basic principle is to apply an opening function on the (equal-length section) area function derived from the linear predictive (LP) model of speech signals. The definition of this opening function is empirically based on morphological considerations, and the main parameter is the place of articulation. Syllabic sounds with /b d g/ in /a i u/ vowel contexts are generated using LP synthesis with reflections coefficients corresponding to the interpolated area function. We show that the general structure of the formant transitions can be well represented using this model, and provide intelligible sound examples.

Index Terms: syllable synthesis, co-articulation, stop consonants, place of articulation, acoustic tube model, formant transitions, reflection coefficients.

1. Introduction

The linear prediction (LP) model of speech signals [1] has been extensively used for more than 40 years in speech processing, synthesis, coding and more, due to its high versatility. For example, the area function calculated from the reflection coefficients can be interpreted and used as an acoustic representation for speech sounds. The *p* poles of a speech sound spectrum correspond to a unique setting of the area of each section of a *p*-order equal-length tube model. In contrast, the relationship between the formant space and the articulatory parameters of an articulatory model mimicking the vocal tract configurations is non-bijective and non-linear [2]. There is a many-to-one relationship with different articulatory configurations able to produce the same three first formants of a speech spectrum. In return, a drawback of the "acoustical LP model" is that the unique solution provided for the acoustical area function is more or less closely related to the corresponding underlying articulatory (vocal tract) shape area function. Wakita [3] found that for stationary vowels, the degree of similarity is quite satisfactory. This qualitative fitting is however generally lost for consonants and non-stationary speech segments, and the frame-by-frame estimate is a poor solution for the inversion problem, i.e. recovering of articulatory configurations from speech sounds. Practically, the linear predictive model has been very poorly used for tackling the inversion problem. In essence, this is because it has been shown that the relationship with the realistic articulatory shapes cannot be achieved well without the adding of *morphological* constraints [2].

We propose to apply this concept of morphological constraint for plosive-vowel synthesis in a quite simple way, directly in the acoustical LP domain, hence without using a sophisticated articulatory model. Considering that the coarse acoustical LP estimate preserves the global morphology of the vocal tract for the target vocalic part, and that consonants are defined by their locus of articulation [4], we define an opening function that fixes the realization of the constriction for the production of stop consonants directly on the LP area function of the target vowel and provides interpolated area function values for the plosive-vowel transition. In the present paper, we apply this paradigm for the synthesis of /b d g/ in /a i u/ vowel contexts, shown as "canonical" by Schwartz *et al.* [5]. Note that, despite the simplicity of the proposed method, and the wide use of the LP model over so many years, to our knowledge, it is the first time that such hybrid morphologically-constrained synthesis is achieved.

The following of the paper is organized as follows. In Section 2 we present the synthesis method in three steps: setting of the vowel target, definition of the opening function applied to the area function, and finally sound synthesis from an area function sequence. Section 3 provides some results obtained with our implementation.

2. The consonant-vowel synthesis method

2.1. Setting the vowel target

The first step of the proposed synthesis process consists of setting the vowel configuration, i.e. a typical vowel LP area function. A vowel spectrum and LP area function can be generated directly by LP analysis on a stationary portion of vowel signal [1]. In the present study, we rather start from an articulatory model in order to compare the articulatory-based area function with the derivated LP area function and check for the coarse correspondence, in the spirit of [3], hence justifying the articulatory sense of the LP area function in the vocalic context. Therefore, similarly to [3], the synthetic vowel is first produced by the articulatory VLAM (Variable Length Articulatory Model) model, as used by, e.g., Ménard *et al.* for studying the growth of the vocal tract in [6]. Each vowel production is set manually with 5 control parameters (lips aperture and protrusion, jaw, tongue body and dorsum). The VLAM area function is determined from the vocal tract shape of this model and has 29 tubes of variable length (Figure 1, bottom). This area function allows to synthesize a spectrum with the method of Badin and Fant [7], and then to use this spectrum to infer the acoustical LP area function corresponding to an equal-length tube model. This is done by inverse Fourier

transform of the power spectrum and selection of the p+1 first resulting autocorrelation coefficients. Those coefficients can be bijectively transformed in a set of either p+1 prediction coefficients, p reflection coefficients, or as targeted here, p area values $A_v(x)$ corresponding to the sections of the acoustical p-tube model, using the classic routines described in [1]. Note that, for a given vowel spectrum, after setting the LP model order p, this analysis produces a *unique* acoustical solution.

In our framework, the comparison between articulatory and LP area functions is easy because on the one hand the articulatory area function roughly fulfils the equal-length tube condition, and on the other hand the adding of supplementary poles to the acoustical model simply adds up a queue of constant tubes. Figure 1 displays the (VLAM and corresponding LP) area functions that we obtained for the set /a i u/. The LP order was set to 30 to be compliant with the number of tubes in VLAM. Qualitatively, the distance between the two area functions is larger than described in [3]. The number of tubes is larger here (8 tubes only in [3]) and the acoustical model has the tendency to produce large steps in the lips region for /a/ and /i/ and smaller steps for /u/. Otherwise, the overall shape of the two area functions is grossly preserved and we can see that the constriction point (located by arrows in Figure 1) is back for /a/, front for /i/ and halfway for /u/.

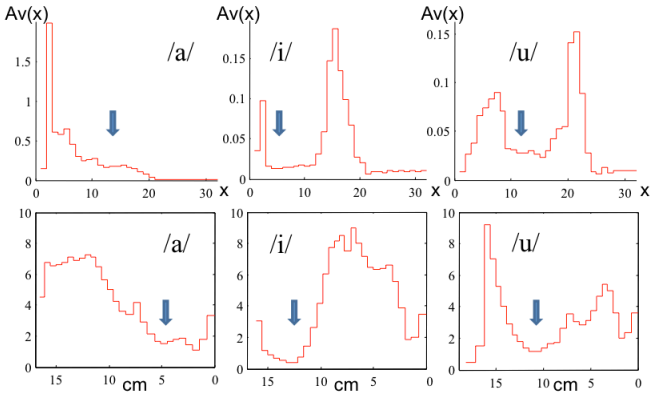

Figure 1: *Area function $A_v(x)$ from lips to glottis (left to right on x axis) of the 3 cardinal vowels* /a i u/ *for the LP acoustic tube model (top) and for VLAM (bottom). The vowels are initially synthesized by VLAM. Note that for the* /a/, $A_v(x)$ *remains constant after* $x = 22$.

2.2. Opening function

Once the morphological characteristic is assumed to be fairly recovered in the LP model for the target vowel, the first crucial point is to identify along the LP area function $A_v(x)$ the equivalent constriction point L as the place of articulation of consonants. This point can vary depending on the vowel because of the co-articulation phenomenon. The bilabial /b/ is the easiest to find because it is always the first tube $x = 1$.

Then we must define an opening function that controls the dynamic aperture of the acoustic area function, in order to control the formant trajectories. The locus of constriction L is assumed to be the main parameter of this opening function and is assumed to be sufficient to reproduce the consonant-vowel transition, which is characterized by the temporal dynamic of the three first formants. This phenomenon is reproduced (at least for /ba da ga/) with articulatory models (e.g., CASY [8]) in the control parameters space by interpolating the initial closed vocal tract configuration and the steady-state vowel configuration. In the present study, this function has been derived empirically by hearing the output synthesis signals and observing the formant trajectories. The function is defined around the constriction point L, in the front and the back neighborhood, and is relative to the target vowel area function $A_v(x)$. For each of the p-equal length tubes, the variation in time (denoted by time-frame index t) of the area function is a linear interpolation from $t = 0$ to $t = T$:

$$A(x,t) = A(x,0) + \left(A_v(x) - A(x,0)\right)\frac{t}{T} \quad (2)$$

$A(x,0)$ is the initial aperture defined below and the vowel area function $A_v(x)$ is reached at the end of the transition $t = T$. The setup of the initial configuration $A(x,0)$ is different for the front cavity ($x = 1$ to $L - 1$) and the back cavity ($x = L + 1$ to p) so that the front cavity starts less opened than the back cavity, and this effect is greater when L is large:

$$A(x,0) = A_v(x)/\max\left(1, x - L + p/2\right), x = 1,\ldots,L-1 \quad (3)$$

$$A(L,0) = A_v(L)/500 \quad (4)$$

$$A(x,0) = A_v(x)/\max\left(1, L - x + p/(L+4)\right), x = L+1,..,p \quad (5)$$

The value of 500 is used in Eq. (4) to produce a very small constriction whatever the value of $A_v(L)$. Let us remark that the present model is defined *ad hoc*, directly within an acoustic representation and without reference to the precise morphology of the articulatory space. An example of opening function is shown in Figure 2 with an *ad hoc* sine profile. For our speech area function, the co-articulation between the consonant and the vowel is effective because the initial position $A(x,0)$ is bound to the vowel area function $A_v(x)$, and the setting of the locus L depends on the vowel.

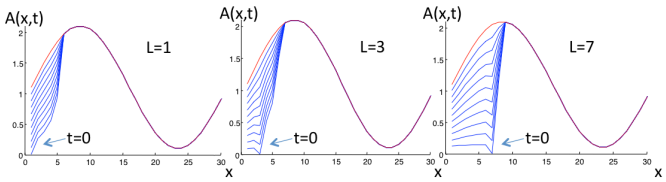

Figure 2: *Opening function applied on a sine area function (p=30).*

2.3. Syllable signal synthesis

2.3.1. General principle

The signal synthesis is processed with classical "source-filter" speech synthesis based on the LP model. A glottal source signal is defined (see Section 2.3.3) and filtered by the infinite impulse response (IIR) "vocal tract filter" that uniquely corresponds to the LP area function (see the routines for LP representation transcoding in [1]). The filter is updated at every new area function $A(x,t)$, $t = 0$ to T. In the present study, the time frame is set to 5ms, and $T = 50$ms (all signals are sampled at 20kHz). To ensure smooth time evolution of the synthesis signal, the filtering process in implemented using the lattice structure [1] that enables smooth transitions between successive filters: filtering equations are expressed in terms of input source signal,

reflection coefficients $k_i(t)$, $i = 1$ to p, and corresponding buffers. Interpolation between successive filters is thus carried out by linear interpolation of the successive reflection coefficients sets within the lattice filtering process (i.e., linear interpolation from $k_i(t)$ to $k_i(t+1)$, for each $i = 1$ to p). This is a classical updating routine in LP model based speech coders. The duration of the interpolation can be set to any value between 1ms and 5ms without affecting much the signal quality.

2.3.2. *Pre-voicing and steady-state regions*

To complete the sound synthesis, a pre-voicing region is added at the beginning of the syllable (Region P1 in Figure 3) and a steady-state region is added at the end (Region P3 in Figure 3; the transition is denoted P2). A unique filter is defined for the whole pre-voicing region. This filter is derived from the first filter of the transition region (corresponding to the "closed" area function $A(x,0)$). More precisely, the power spectrum corresponding to $A(x,0)$ is "lowpass-filtered" above 800Hz by a multiplication with a decreasing exponential, so that, only the first formant is grossly preserved. The pre-voicing filter is derived from the resulting low frequency power spectrum the same way as for the transition region (inverse Fourier transform followed by direct LP modeling using the $p+1$ first autocorrelation coefficients). The junction between P1 and P2 is made by linear interpolation of the reflection coefficients of the pre-voicing filter and the "$A(x,0)$ filter" in the lattice filtering implementation, as what is done within the P2 section.

The steady state region is obtained very simply by extending on several time frames the filtering of the glottal source signal with the last filter of the transition region (i.e. the filter corresponding to the target vowel).

2.3.3. *Glottal source signal*

The modeled glottal source signal is generated by convolution of a unitary pulse train with the glottal pulse model of Rosenberg [9], which is basically the concatenation of an order-3 polynomial (for the rising part) and order-2 polynomial (for the decreasing part; see Figure 3c). The frequency contour of the pulse train follows an arbitrary f_0 pattern inspired from measures obtained on actual acoustic realization of the considered syllables (Figure 3b). A small amount of random noise is added to the f_0 pattern to grossly simulate jitter and add some naturalness to the synthesized signal. Finally a gain is also defined to appropriately modulate the modeled excitation signal (Figure 3a). Of course, the gain contour is correlated to the LP spectrum/filter dynamics[1]. The amplitude during P1 is a small proportion of the vowel amplitude (usually 0.003). It evolves during the transition (Region P2 in Figure 3) according either to a linear function (for $L<6$) or the first part of an increasing sigmoid (for $L\geq 6$). It stays quasi-stationary during P3 before declining (this latter part of the gain pattern is arbitrary and correspond to some "fade out" of the speech signal).

[1] The reflection coefficients used in the lattice filter implementation do not consider the absolute gain of the frequency response, since they only encode the normalized LP spectrum / frequency response (which is centered on a log-modulus scale [1]). This is compliant with the fact that the area function does not provide gain information. The gain contour applied on the excitation signal can be seen as a denormalization of the filter sequence, which is particularly crucial at the pre-voicing / $A(x,0)$ transition.

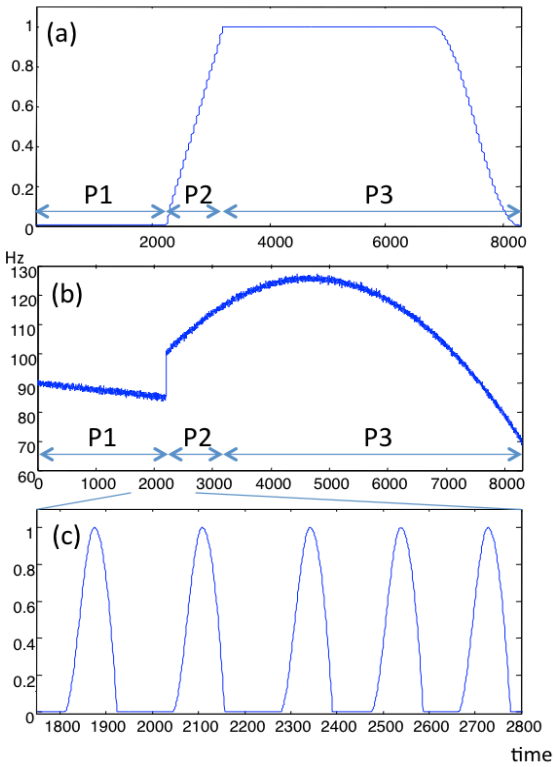

Figure 3: *Characteristics of synthesized syllables: (a) Excitation gain, (b) f0 contour, and (c) glottal pulse train (zoom on about 50ms). P1, P2 and P3 denote the pre-voicing, transition, and steady-state regions respectively. Abscissa is given in sample index (the sampling frequency is 20kHz).*

3. Simulations

The three plosives /b d g/ were produced in /a i u/ vocalic context. We assume that all syllables are intelligible. Examples of sounds that were generated using the presented technique can be found in the accompanying file or at http://www.gipsa-lab.fr/~laurent.girin/demo/stop-cv-synthesis.zip. The corresponding area functions $A(x,t)$ and spectrograms are displayed in Figure 5. The comparison between the general classical patterns of formant transitions for /ba da ga/ (Figure 4) and the transitions observed on the spectrograms (Figure 5) are globally satisfactory.

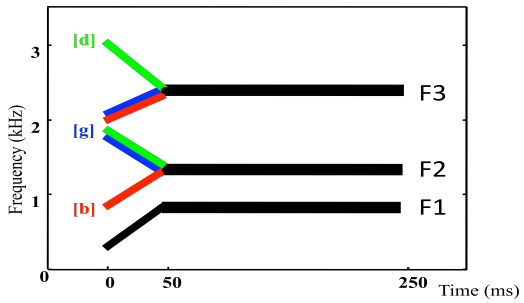

Figure 4: *Classical formant transitions for* /ba da ga/.

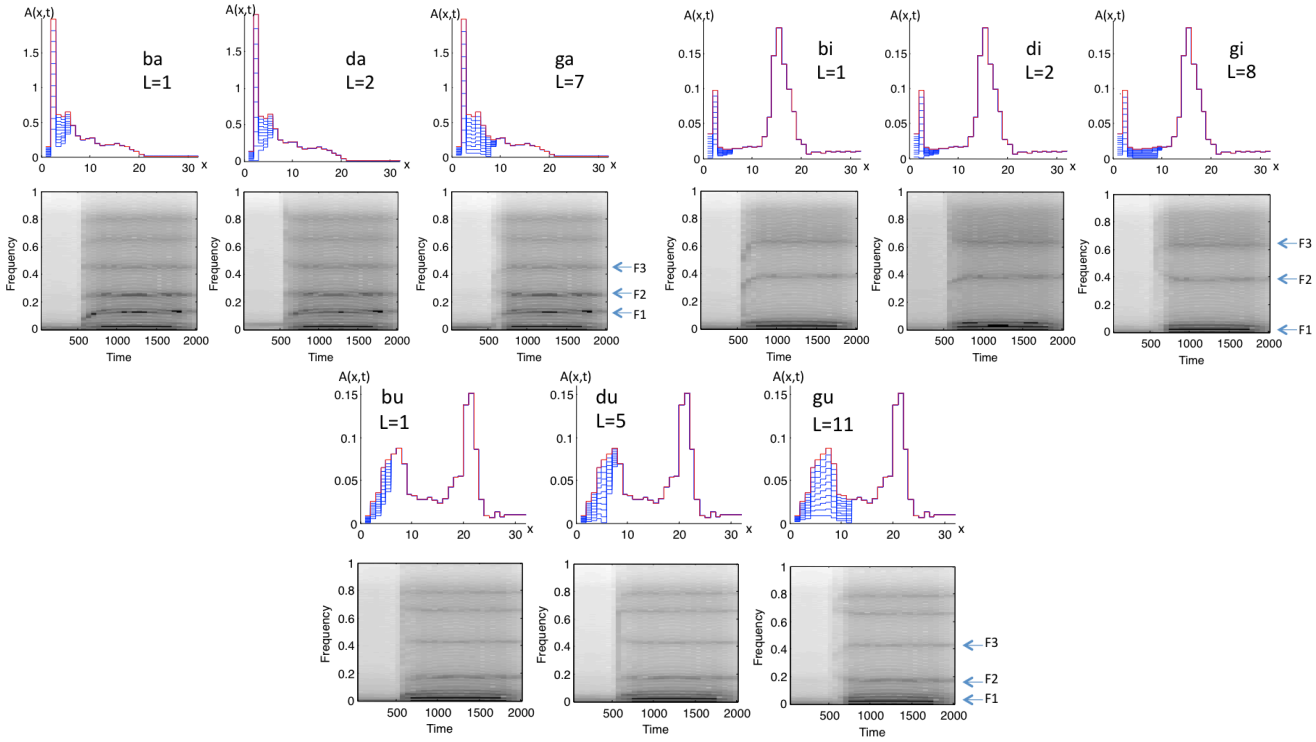

Figure 5: *Simulations with the "canonical" syllables* /b d g/ *to* /a i u/.

The locus of formant transitions and the place of articulation are closely related [4] but there is no general model of such relationship. Whereas synthesis is generally not easy to achieve with an articulatory model, the hybrid synthesis model that we propose offers a remarkably simple framework for acoustic speech synthesis motivated and constrained by articulatory (morphological) considerations. Using only a single free parameter L with clear articulatory meaning within a simple interpolation scheme (inverse of a linear function of x and linear interpolation of time t), this model supports the major role of the place of articulation for the production of stop consonants in various vocalic contexts, and fairly simulates the co-articulation phenomenon.

A discussion on the interest of such a simple model for advanced studies on articulatory-acoustic relationships in speech production and perception is left out of the scope of the present paper, which focuses on the technical implementation and basic results in terms of formant trajectories and sound samples.

4. Acknowledgements

This work is supported by the ANR (French National Research Agency) as part of the Skullspeech project. The authors thank Pascal Perrier for useful discussion on the content of this paper.

5. References

[1] Markel, J.D. and Gray Jr, A.H., "Linear prediction of speech", Springer-Verlag, 1976.

[2] Yehia, H. and Itakura, F., "A method to combine acoustic and morphological constraints in the speech production inverse problem", Speech Com., 18:151-174, 1996.

[3] Wakita, H., "Direct estimation of the vocal-tract shape by inverse filtering of the acoustic speech waveforms", IEEE Trans. Audio and Electroacoustics, 5(AU-21):417-427, 1973.

[4] Delattre, C.P., Liberman, A. M. and Cooper, F.S., "Acoustic loci and transitional cues for consonants", J. Acoust. Soc. Am., 27(4):769-773, 1955.

[5] Schwartz, J-L., Boë, L-J., Badin, P. and Sawalis, T., "Grounding stop place systems in the perceptuo-motor substance of speech: On the universality of the labial-coronal-velar stop series", J. of Phonetics, 40 :20-36, 2012.

[6] Ménard, L., Schwartz, J.-L., Boë, L.-J. and Aubin, J., "Articulatory-acoustic relationships during vocal tract growth for French vowels: Analysis of real data and simulations with an articulatory model", J. of Phonetics, 35:1-19, 2007.

[7] Badin, P. and Fant, G., "Notes on vocal tract computation", Speech Transmission Laboratory - Quarterly Progress Status Report - Stockholm, 2-3/1984, 53-108, 1984.

[8] Iskarous, K., Goldstein, L.M., Whalen, D.H., Tiede, M.K. and Rubin, P.E., "CASY: The Haskins Configurable Articulatory Synthesizer", Proc. of the 15th Int. Conf. on Phonetics (ICPhS), Barcelona, Spain, pp. 185-188, 2003.

[9] Rosenberg, A. E., "Effect of glottal pulse shape on the quality of natural vowels", J. Acoust. Soc. Am., 49(2):583-590, 1971.