A PHYSICAL MODEL FOR ARTICULATORY SPEECH SYNTHESIS. THEORETICAL AND NUMERICAL PRINCIPLES

N. Ruty[1], J. Cisonni[1], X. Pelorson[1], P. Perrier[1], P. Badin[1], A. Van Hirtum[1]

[1]Institut de la Communication Parlée, UMR5009-CNRS, Institut National Polytechnique de Grenoble, France

I. Introduction

The objective of this study is to develop a simple physical model able to produce articulatory synthesis of continuous vowel transitions. The proposed model considers simplified physical approximations for the glottal source, acoustical propagation in the vocal tract and lip radiation. The articulatory quality of the synthesized results is interpreted in terms of the simplifications applied in each model element.

II. Theoretical description

A. Source models

Two types of glottal flow sources are considered. Both were chosen because they are well-known and representative for two different modeling approaches.

The first one is the Liljencrants-Fant (LF) model described in [1]. This analytical ad-hoc model is easily controllable because only four wave shape parameters are needed. Fig. 1 shows a typical period of glottal flow derivative synthesized by this model during a period t_0, corresponding to a fundamental frequency $F_0 = 1/t_0$. The derivative of the flow is described by these two equations:

$$E(t) = E_0 e^{\alpha t} \sin(\pi t / t_p) \qquad if \qquad t < t_e \tag{1}$$

$$E(t) = E(t_e)\frac{e^{\varepsilon(t-t_e)} - e^{\varepsilon(t_0-t_e)}}{1-e^{\varepsilon(t_0-t_e)}} \qquad if \qquad t_e < t < t_0 \tag{2}$$

where t_0, t_e, t_a and E_0 are the control parameters of the model, $\varepsilon = 1/t_a$ and α are calculated using the method described in [2], as a function of t_0, t_e, t_a and t_p. Such a parametric description of the glottal flow is very popular due to its simplicity and its computational efficiency. Although it involves almost no physics it will be used in the following as a reference.

As a more physical alternative, we propose the use of the symmetrical two mass model [3] schematically depicted in Fig. 2. The two mass model is capable to simulate flow-induced vibrations of the vocal folds when coupled to an airflow model [4]. This model is controlled by a set of mechanical parameters: the masses (m), stiffness (K, K_c) and damping (r). The applied parameter values determine the mechanical response, i.e. the resonance frequencies and the quality factors [5]. The applied flow model accounts for an unsteady flow separation point (using Liljencrants 'ad-hoc' criterion) allowing to predict the pressure distribution within the glottis. The influence of viscous losses and inertia of air on the main flow is also considered. This way, the pressure forces exerted by the flow on the vocal folds walls can be estimated from the calculated pressure distribution. The mechanical differential equations are numerically solved in order to estimate the vocal folds displacement and the glottal flow velocity. An example of the source signal (derivative of the volume flow) obtained with this model is depicted in Fig. 3.

B. Acoustical propagation

Acoustical wave propagation inside the vocal tract is described in the time domain [6]. First the geometry of the vocal tract is concatenated into a finite number of tubes of area A_i, where (i) is the index of the tube in the flow direction. In each tube (i) the pressure distribution at a location x along the vocal tract axis can be written as:

$$p_i(x,t) = \left[p_i^+(t - x/c) + p_i^-(t + x/c)\right] \tag{3}$$

with p_i pressure, x position, p_i^+ and p_i^- respectively the travelling and regressive pressure waves in tube (i).

Assuming continuity at the junction between the tubes (i) and (i+1), one obtains:

$$\begin{aligned} \frac{1}{A_i}\left(p_i^+(t - l_i/c) - p_i^-(t + l_i/c)\right) &= \frac{1}{A_{i+1}}\left(p_{i+1}^+(t) - p_{i+1}^-(t)\right) \\ p_i^+(t - l_i/c) + p_i^-(t + l_i/c) &= p_{i+1}^+(t) + p_{i+1}^-(t) \end{aligned} \tag{4}$$

where l_i is the length of the tube (i), A_{i+1} the area, and p_{i+1} is the pressure in the tube (i+1).

The travelling and regressive component of the pressure in the tube (i+1) are given by:

$$\begin{aligned} p_{i+1}^+(t) &= \beta_i p_i^+(t - li/c) + r_i p_{i+1}^-(t) \\ p_i^-(t + li/c) &= -r_i p_i^+(t - li/c) + \phi_i p_{i+1}^-(t) \end{aligned} \tag{5}$$

where $r_i = \dfrac{A_{i+1} - A_i}{A_{i+1} + A_i}$ is the reflection coefficient at the junction (i), $\beta_i = 1 - r_i$ and $\phi_i = 1 + r_i$ are the propagation coefficients.

Models and analysis of vocal emissions for biomedical applications. 4th international workshop. October 29-31, 2005 – Firenze, Italy.
Edited by C. Manfredi. ISBN 88-8453-320-1 (online)

In first approximation, the reflection coefficient at the glottis is considered to be $r_0 = 1$. The reflection coefficient at the lips is discussed in the following subsection.

C. Reflection coefficient at the lips

The acoustical behavior of the lips is approximated by the radiation from a piston set in an infinite rigid baffle. This radiation impedance is calculated using the analytic formula described in [7]:

$$Z_r = \pi a^2 \rho c\left(1 - J_1(2ka)/(ka) + S_1(2ka)\right) \quad (6)$$

where a is the radius of the equivalent piston (i.e. the lips aperture), ρ the air density, c the sound velocity, k=2πf/c, f is the frequency, J_1 is the first kind Bessel [7] function of order 1, and S_1 the Struve function of order 1.

This analytic form is used for calculations, only the Bessel and Struve functions are approximated. Given Zr, the normalized reflection coefficient is calculated as following:

$$R = \frac{1 - Z_r/(\pi a^2 \rho c)}{1 + Z_r/(\pi a^2 \rho c)} \quad (7)$$

D. Glottis/vocal tract coupling

Using the LF model for the glottal source, the source signal is simply injected at the entrance of the vocal tract, i.e. the tube with index i = 0. Next, the propagation is calculated for each time step.

The use of the two mass vocal fold model allows to account for the interaction between the source and the vocal tract. Indeed, the acoustical pressure past the glottis (i=0) is constantly varying with time due to the multiple reflections. These variations modify the pressure drop between the entrance and the outlet of the glottis. Since the pressure distribution within the glottis depends on the total pressure drop across the glottis, the forces and thus the vocal fold movements can be affected by the acoustical pressure fluctuations.

III. From theoretical domain to numerical domain

A. Vocal Tract and articulatory synthesis

The concatenation of the vocal tract into elementary tubes imposes, for calculation efficiency, a sample frequency $F_e = c / L_x$; where $L_x = L / N$ is the length of a vocal tract tube retrieved as the ratio between the total vocal tract length L and the number of tubes N. Thus, a modification of the vocal tract shape (a variation of length) implies a variation of the sample frequency. In order to keep a sound signal sampled at a constant frequency and considering a continuous vocal tract variation, the generated signal is resampled following the method proposed in [8]. Firstly, the numeric signal x(n) is converted in an analogical signal. Then, this continuous signal is sampled at a fixed frequency. The resampled signal y(m) is defined by:

$$y(m) = x(mT_S) = \sum_{n=N1}^{N2} x(n) \cdot \frac{\sin(\pi \cdot \frac{mT_S - nT_e}{T_e})}{\pi \cdot \frac{mT_S - nT_e}{T_e}} \cdot W(mT_S - nT_e) \quad (8)$$

Where m is the new sample index, T_s and T_e correspond to the periods associated with respectively F_s and F_e as $T_S = 1/F_S$, $T_e = 1/F_e$, and W is the applied time (Hamming) window, with N1 et N2 its limits.

This way, the continuous variation between two vowels can be computed.

B. Reflection function: from theory to filter design

We want to design a filter with a frequency response as close as possible to the reflection coefficient. As observed in [9], numerical treatment and more precisely inverse FFT of the reflection coefficient is likely to cause disturbances, due to sampling and windowing effects. Therefore some of the approximations can be erroneous. The applied digitisation of the vocal tract yields a sampling frequency of around 80000Hz. The exact value depends on the vocal tract length corresponding to the vowels we want to synthesize. The reflection coefficient at the lips is calculated for frequencies ranging between 0 and 40000Hz, for a total of 2^{15} values. Then the inverse FFT of this signal is computed. From this reflection function, the first thirty coefficients are kept as filter coefficients.

IV. Results and discussion

A. Validity of the approximated transfer function

The validity of the mentioned hypothesis concerning vocal tract propagation and lip radiation needs to be tested. Therefore the theoretical transfer function of the acoustical model is computed following the method described in [10]. The comparison with the FFT of the impulse response described in III.B is presented in Fig. 4 for a sampling frequency of 80000Hz, and for two extreme equivalent piston radii (0.001m and 0.02m). From this figure it can be seen that the mean error concerning the reflection coefficient is less than 10% at the most. Note that the commonly used low frequency

approximation: $Z_r = \pi a^2 \rho c \left[\frac{1}{2}(ka)^2 + i\frac{8ka}{3\pi} \right]$ [11] can lead to strong departures even at moderate frequencies (of order of 4 kHz). In figure 5 a comparison between the computed transfer function a 32 sections approximation of the vowel [i] and the theoretical prediction is shown. A very good agreement can be observed.

B. Variation of vocal tract length

Fig. 6 shows the spectrogram of the transfert function of a uniform vocal tract. The length of the vocal tract is continuously varying from 10cm up to 17cm by steps of 1mm. The method to simulate this lengthening is detailed in subsection III.A. As expected, the spectrogram exhibits no discontinuities and hence the method seems appropriate to simulate vowel transitions.

C. Some examples of synthesis

As a typical example, Fig. 7 and Fig. 8 present the results of a synthesised vowel [a]. This vowel is chosen as an example because it is an extreme part of the vocalic triangle Fig. 7 shows a time simulation of 1s and Fig.8 shows the frequency representation. The example shown is obtained with the LF source model, the acoustics is simulated as explained previously and using a resample frequency of 16kHz.. Further simulations using the two mass model for the source lead to an increase in perceptual quality since this approach accounts for vocal tract/source interaction and hence results in high quality articulatory synthesis.

V. CONCLUSION

The current study presents a simplified physical model for articulatory synthesis of continuous vowel transitions. The relevance of a high frequency radiation description has been shown. Further, the numerical implementation – including variable sampling frequency- has been developed and validated. Typical synthetic examples – including vowel transition- will be played during the conference [12].

REFERENCES

[1] G. Fant, J. Liljencrants, Q. Lin,"A four-parameter model of glottal flow," *STL-QPSR*, vol. 4. , pp. 001-013, 1985.
[2] R. Veldhuis, "A computationally efficient alternative for the Liljencrants-Fant model and its perceptual evaluation," *J. Acoust. Soc. Am. 103 (1)*. pp 566-71. 1998
[3] N. J. C. Lous, G. C. J. Hofmans, R. N. J. Veldhuis, A. Hirschberg, "A Symmetrical Two-Mass Vocal-Fold Model Coupled to Vocal Tract and Trachea, with Application to Prosthesis Design". *Acta Acustica United with Acustica*.Vol. 84 (6). pp 1135-50. 1998.
[4] X. Pelorson, A. Hirschberg, Van Hassel, R.R. A.P.J. Wijnands, Y. Auregan, "Theoretical and experimental study of quasisteady-flow separation within the glottis during the phonation, Application to a modified two-mass model," *J. Acoust. Soc. Am. 96 (6)*. pp 3416-31. 1996
[5] I. Lopez, A. Van Hirtum, M.H. Schellekens, N.M. Driessen, A. Hirschberg, X. Pelorson, "Buzzing lips and vocal folds: the effect of acoustical feedback", in Flow Induced Vibrations; de Lanfre & Axisa, Paris, France. 2004
[6] D. O'Shaughnessy, *Human and machine,* Speech Communication, Addison-Wesley Publishing Company. 1997, pp 41-127.
[7] P.M. Morse, K. Uno Ingard, *Theoretical acoustics*, Mc Graw-Hil Book Companylvol. New York. 1968, pp.383-388.
[8] H.Y. WU, P. Badin, Y.M. Cheng, B. Guérin, "Simulation du conduit vocal: réalisation de la variation continue de longueur dans un modèle de Kelly-Lochbaum," *Bulletin du Laboratoire de la Communication Parlée,* vol. 1, pp. 01-27, 1987.
[9] J.D. Polack, X. Meynial, J. Kergomard, C. Cosnard, M. Bruneau, "Reflection function of a plane sound wave in a cylindrical tube," *Revue. Phys. Appl.,* vol. 22, pp. 331-337, 1987.
[10] X. Pelorson, R. Laboissière, S. El Masri, "Vocal tract acoustics at high frequencies," *Proc. 4ème Congrès Français d'Acoustique*, vol. 1, pp. 401-404, 1997.
[11] G. Fant, *The acoustic theory of speech production.* Mouton, The Hague, 1960.
[12] http://www.icp.inpg.fr/~ruty

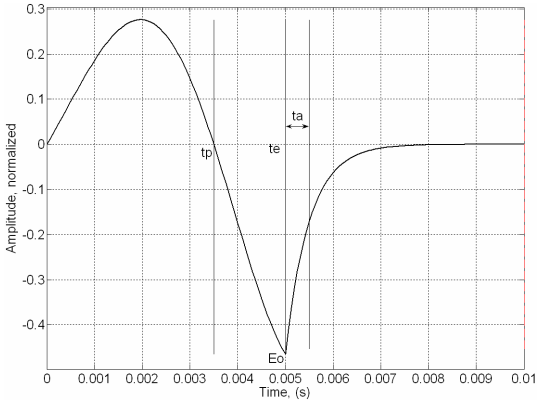

Figure 1: Derivative of the glottal airflow, generated with the LF model

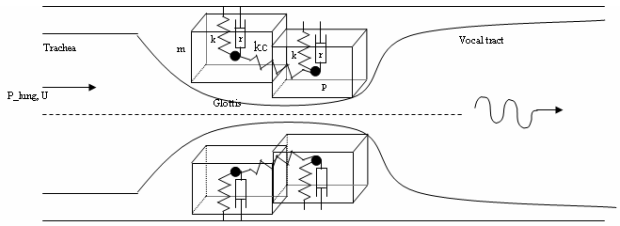

Figure 2 : two mass model of vocal folds, P_lung is the subglottal pressure, U the volume airflow, m the masses, k and kc the stiffness of the spring, and r the damping.

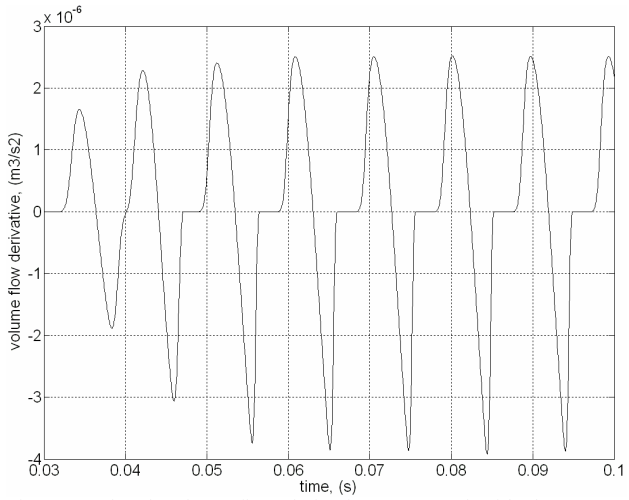

Figure 3: glottal volume flow derivative, generated with the 2 mass model for a subglottal pressure increasing from 0 to 600 Pa.

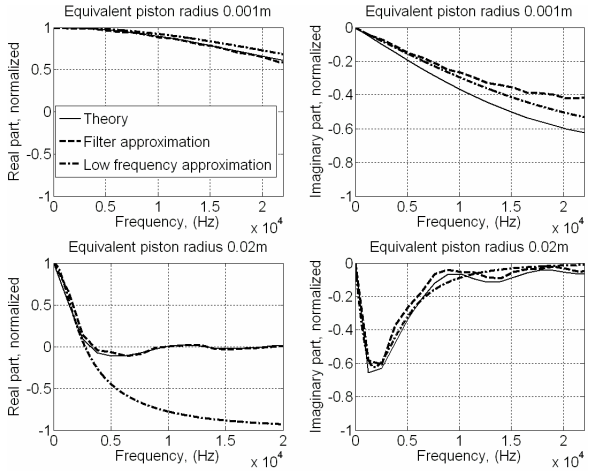

Figure 4: comparison of the theoretical reflection function and its filter approximation, 30 coefficients for the FIR filter, for a sample frequency of 80kHz. Two extreme values of the equivalent piston radius are considered (0.001m and 0.02m).

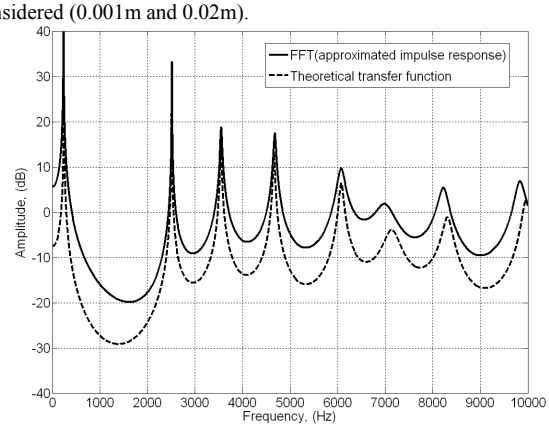

Figure 5: comparison of the theoretical transfer function and the inverse FFT of the impulse response, for the vowel [i], resampled at 44kHz.

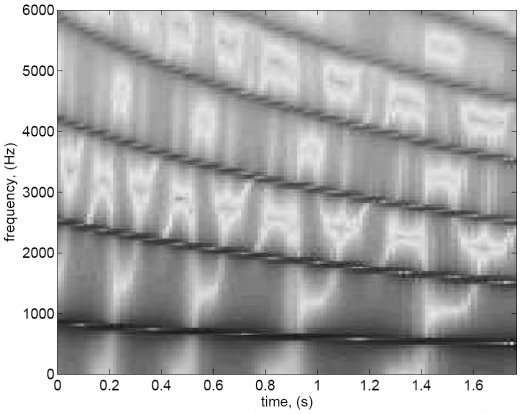

Figure 6: Spectrogram for a continuous variation of the tube length, from 0.1m to 0.17m, reference of sampling frequency Fs=20000Hz, tube section dimension 0.04*0.03m².

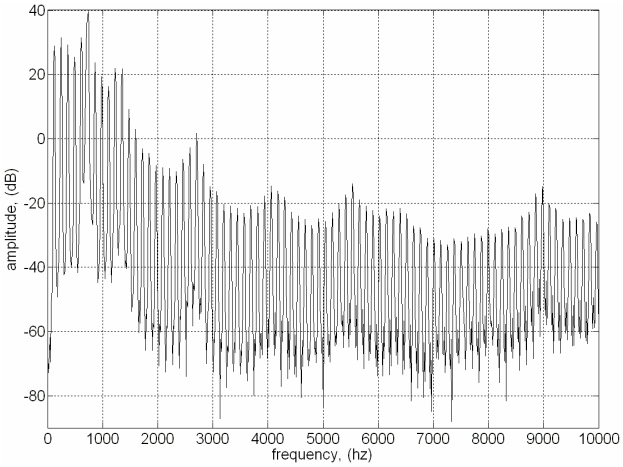

Figure 7: Frequency representation of the synthesized voiced sound [a]. Resample frequency 44kHhz, duration 1s, fundamental frequency 123Hz.

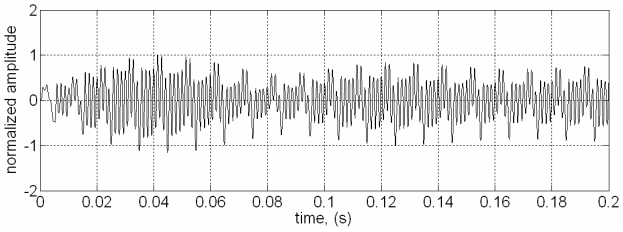

Figure 8: Time representation of the synthesized voiced sound [a]. Resample frequency 44kHz, duration 1s, fundamental frequency 123Hz.