3rd International Workshop MAVEBA 2003, 151-154

NON-MODAL VOICE SYNTHESIS BY LOW-DIMENSIONAL PHYSICAL MODELS

C. Drioli[1], F. Avanzini[2]
[1]Institute of Phonetics and Dialectology, ISTC-CNR, Padova, Italy
[2]Department of Information Engineering, University of Padova, Padova, Italy

***Abstrac*t: The synthesis of different voice qualities by means of a low-dimensional glottal model is discussed. The glottal model is based on a one-mass model provided with a number of enhancements that make it suitable to the aim of the study. The simulation of modal and non-modal phonatory regimes is discussed. Both symmetric and non-symmetric configurations are explored. The class of models under consideration is shown to be able to reproduce a broad range of phonation styles and to provide interesting control properties.**
Keywords: physical models of vocal emission; non-modal phonation types.

I. INTRODUCTION

The possibility of reproducing different voice qualities by means of a voice synthesis tool has been explored for different applications such as emotive and natural-sounding speech synthesis [1], pathologic voice assessment [2], analysis of voice quality [3], [4], [5]. Many of the acoustic and perceptual features of an individual's voice are believed to be due to specific characteristics of the quasi-periodic excitation signal (glottal flow waveform) provided by the vocal folds. Accordingly, source models have received considerable attention and they come today in a number of versions, the most important ones being the parameterization by analytical functions, such as the LF-model [6], and the physiological modeling of the glottis, such as the multi-mass models [7], [8].

Most source models come with a set of controls to manipulate the pulse shape. The LF-model is provided with parameters for the control of the glottal pulse open phase, return phase, and closed phase durations, with parameters for the control of spectral tilt and the high-frequency content of the spectrum, and with parameters to control the diplophonia observed, for example, in laryngalized or harsh voice [3]. As for physical models, the direct control of the pulse shape is usually less simple, due to the large amount of parameters which are physically motivated but not always connected in a clear way to the characteristics of the glottal pulse. On the other hand, many authors have explored the effect of asymmetries in the mechanical components with respect to non-modal and pathological phonation types (e.g., [9]).

In this paper we explore the use of a class of low-complexity physical models loosely based on the Ishizaka&Flanagan's one- and two-mass models, and on Titze's mucosal wave model, with the specific aim of reproducing non-modal phonation modalities. The use of simplified physical models is justified by the interest raised recently in the field of natural-sounding speech synthesis, in which the possibility of generating a wide range of phonatory styles and voice qualities is highly desirable.

The paper is organized as follows. Section II gives an overview of the voice production model under investigation. In Section III the experimental setting is introduced and results from the simulation of the model are presented for both balanced and imbalanced configurations. In Section IV the conclusions are given.

II. VOICE PRODUCTION MODEL

The voice production model assumed is a source-filter scheme in which the volume velocity at the glottis is produced by a physical model and the vocal tract is represented by a parallel of four formant filters. The glottis model adopted here is a low-dimensional body-cover model in which the lower edge of the folds is represented by a single mass-spring system k; r; m and the propagation of the displacement is represented by a delay line of length T [10], [11]. The coronal cross-section of the model is illustrated in Fig. 1. The equations of the aerodynamics of the model can be found in the referenced papers and will be not repeated here. Briefly, the structure is a one-mass model with a propagation line aimed at simulating the propagation of the motion along the thickness of the fold. A second-order resonant filter represents the oscillating folds, an impact model reproduces the impact distortions on the fold displacement and adds an offset x_0 (the resting position of the folds). The driving pressure Pm acting on the folds is computed from the flow and the fold displacement using Bernoulli's law. A flow model converts the glottis area given by the fold displacement into the airflow at the entrance of the vocal tract. The glottis area is computed as the minimum cross-sectional area between the areas at lower and the upper vocal fold edge, and the flow is assumed proportional to the glottal

area. The propagation line is an approximation of the vocal cord along the thickness (vertical) direction and reproduces the vertical phase difference of the vibration of the cord edges, and it is an essential element for the production of self-sustained oscillations without a vocal tract load.

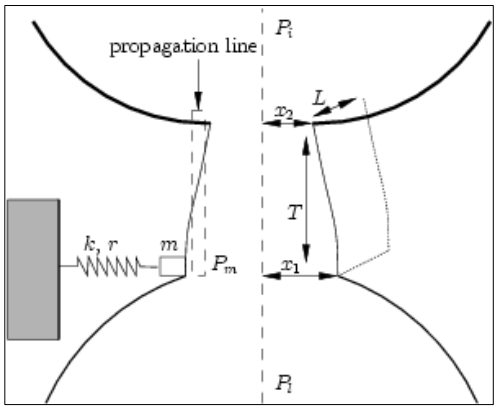

Fig. 1: Low-dimensional body-cover model of the vocal folds. From bottom to top, P_l is the lung pressure, P_m is the driving pressure acting on the vocal folds, m, k, and r represent respectively the mass, stiffness, and damping of the fold, T represents its thickness, x_1 and x_2 are the fold displacements at entrance and exit of the glottis, and P_i is the pressure at entrance of the vocal tract.

III. SYNTHESIS OF VOICED SOUNDS WITH DIFFERENT VOICE QUALITIES

The model adopted here has demonstrated to be successful in reproducing the essential dynamics of voice source and has shown to be able to reproduce real glottal flow waveforms, when extended with an opportune data-driven parametric component [10], [12]. Here we focus on the control of the phonation quality offered by this class of physical models. In particular, we look at the possibility of reproducing convincing 1) breathy, 2) pressed or creaky, and 3) bifurcated phonation types.

The differentiated glottal volume velocity produced by the model is convolved with a vocal tract filter to provide a lip pressure signal for perceptual evaluation of the synthesis.

A. Symmetric structure

A bilaterally symmetric one-delayed mass model is assumed for this section. Model refinements and strategies to produce the target voice quality modifications are described in the following.

Breathy phonation is characterized by the presence of a turbulent aspiration noise combined with the periodic component. The rendering of this phonation type is not always trivial due to the fact that the noise component has a precise phase relation with the periodic voiced component, and a white noise source added to the airflow can sometimes lead to the perception of two distinct sources. n improvement to this aspiration noise model is that of reproducing the amplitude modulation given to the noise by the opening and closing of the glottis. A noise component modulated by the airflow amplitude is thus added to the airflow. Figs. 2 b) and c) shows the result of the simulation for increasing noise component. Fig. 2 d) shows a typical situation of breathy phonation in which the glottis is never completely closed at back, and a DC component is summed to the periodic flow.

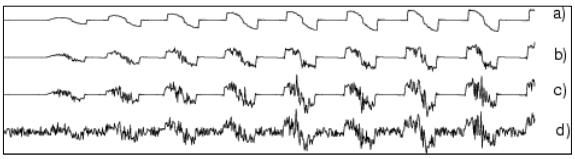

Fig. 2: Differentiated glottal flow waveforms generated by the symmetric model: a) normal; b) and c) increasing breathiness; d) breathy with dc component.

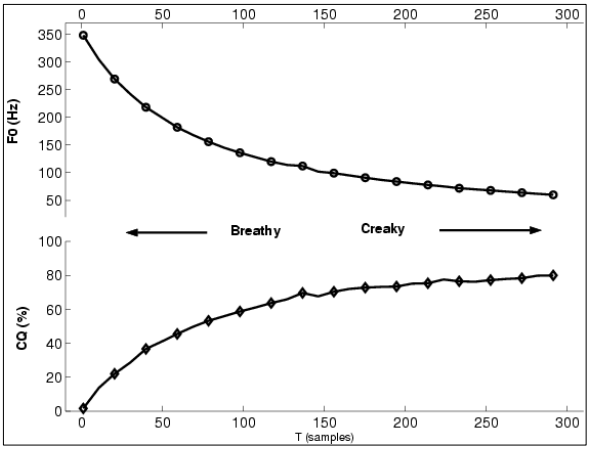

Fig. 3: Relation between the propagation line length (in samples), the resulting pitch, and the resulting closed quotient CQ.

Pressed or creaky phonation is characterized by a narrow airflow pulse (small open quotient) and by a low fundamental frequency. An action on the propagation line length showed to be an effective mean to control the pulse closed phase duration. A physiological interpretation of this parameter can be easier if we look at the propagation line length as the part of the fold actually involved in the oscillation, instead of as the thickness of the whole vocal fold. It is also to note that for all the model configurations tested, the parameter T affects the closed phase duration of the pulse as well as the pitch of the glottal pulse. Fig. 3 shows the relation between the propagation line length (in samples), the resultant pitch, and the resultant closed quotient CQ, defined as the ratio of the closed phase duration to the period length. The fold mass, damping and tension were respectively m = 0.1 g, r = 0.085 Nsm^{-1}, k = 40 Nm^{-1}, resulting in a resonance frequency f_c = 100 Hz and selectivity factor Q = 0.7441 at sampling rate 22050 Hz.

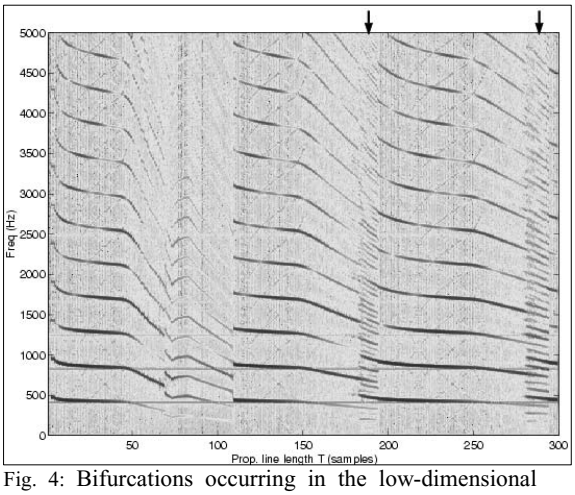

Fig. 4: Bifurcations occurring in the low-dimensional model with symmetric configuration when the length T of the delay line representing the propagation of the fold displacement is slowly varied over time.

Bifurcated phonation is characterized by the presence of period-doubling or sub-harmonics which result in large irregularities in the time domain, usually perceived as a ”rough” voice quality. Bifurcated phonation and irregularities appear occasionally in normal phonation and speech, but is often symptomatic in voice pathology. Often instabilities and sub-harmonic components are the result of tension and mass imbalance of the left and right vocal fold. Asymmetric configurations of the glottal model are explored in the next section. Even with a symmetric configuration, however, we observed the presence of such dynamic behavior when the length of the propagating delay-line was set to values extremely high with respect to the pulse duration. The fold mass, damping and tension were respectively m = 0.05 g, r = 0.002 Nsm^{-1} , k = 80 Nm^{-1} , resulting in a resonance frequency f_c = 200 Hz and selectivity factor Q = 31 at sampling rate 22050 Hz. An empirical rule for the production of bifurcations in the balanced configuration turned out to be the adoption of a considerably high Q factor. Fig. 4 shows the spectrogram of the voiced sound generated by continuously rising the propagation line length. Two clear bifurcation regions can be observed for values of T around 200 and 300.

B. *Asymmetric structure*

In this section, asymmetries are included in the low - dimensional model described in Section II. Earlier studies have already observed the phenomena arising in multi-mass models when the mechanical properties of the folds are made asymmetric [13], [14]. In particular, imbalance in bilateral tension and mass, a configuration usually related to unilateral paralysis, has been extensively explored. Typically, non-stationary regimes are observed when the mass and tension of the two folds are imbalanced. An asymmetry parameter Q_i ? (0, 1] is introduced, which is used to compute the right-hand fold mass and stiffness as $k_r = Q_i\, k_l$ and $m_r = m_l / Q_i$. Fig. 5 shows the simulated fold displacements for different values of Q_i. The values used for the mass-spring system for this examples were m = 0.17 g, r = 0.02 Nsm^{-1}, k = 64 Nm^{-1}, corresponding to a fundamental period of 100 Hz for the balanced configuration.

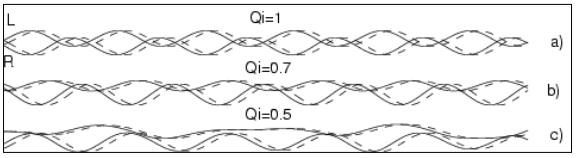

Fig. 5: Fold displacements (left and right; lower edge: solid line, upper edge: dashed line) for various imbalanced configurations. a) symmetrical; b) and c) asymmetrical.

Fig. 6 shows the spectrogram of the synthesis when the asymmetry parameter Q_i is slowly varied over time. Two bifurcation regions are clearly visible as Q_i approaches the values 0.51 and 0.55.

IV. Conclusions

The dynamic behavior of a low-dimensional one-mass model with delayed mass has been investigated, both for balanced and imbalanced configurations. In the balanced configuration normal, pressed, breathy

phonation styles were obtained, as well as bifurcation phenomena in some regions of the parametric space. In general the synthesis results were judged convincing on the basis of informal perceptual evaluations. In the imbalanced configuration, typical non-stationary and bifurcated regimes were observed. The class of low-complexity models presented is characterized by a wide variety of dynamical behaviors and offers in some cases a simple control interface to switch from modal phonation to non-modal phonatory regimes. The computational efficiency of the model suggests that this could be useful in real-time speech synthesis application.

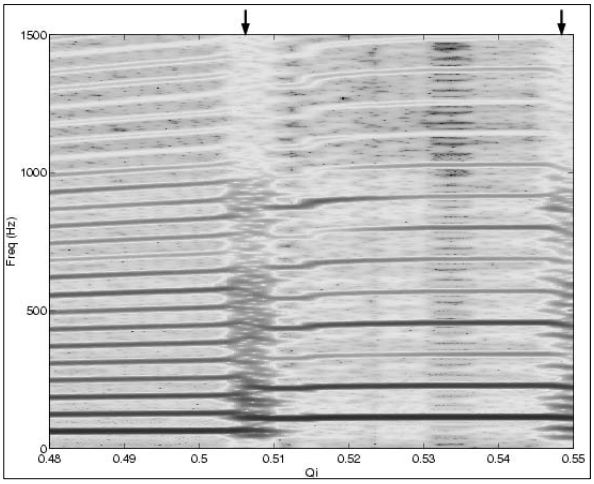

Fig. 6: Bifurcations occurring in the low-dimensional model with asymmetric configuration when the asymmetry parameter Q_i is slowly varied over time.

REFERENCES

[1] C. Gobl and A. N. Chasaide, "The role of the voice quality in communicating emotions, mood and attitude," Speech Communication, vol. 40, pp. 189–212, 2003.

[2] P. Bangayan, C. Long, A. Alwan, J. Kreiman, and B. Gerratt, "Analysis by synthesis of pathological voices using the klatt synthesizer," Speech Communication, vol. 22, pp. 343–368, 1997.

[3] D. H. Klatt and L. C. Klatt, "Analysis, synthesis and perception of voice quality variation among female and male talkers," Journal of the Acoustical Society of America, vol. 87, no. 2, pp. 820–857, February 1990.

[4] D.G. Childers and C.K. Lee, "Vocal quality factors: analysis, synthesis, and perception," J. Acoust. Soc. Am., vol. 90, no. 5, pp. 2394–2410, November 1991.

[5] D.G. Childers and C. Ahn, "Modeling the glottal volume-velocity waveform for three voice types," J. Acoust. Soc. Am., vol. 97, no. 1, pp. 505–519, January 1995.

[6] G. Fant, J. Liljencrants, and Q. Lin, "A four-parameter model of glottal flow," STL-QPS, pp. 1–13, 1985.

[7] K. Ishizaka and J. L. Flanagan, "Synthesis of voiced sounds from a two-mass model of the vocal cords," The Bell Syst. Tech. J., vol. 51, no. 6, pp. 1233–1268, July-August 1972.

[8] I. R. Titze, "The physics of small-amplitude oscillations of the vocal folds," J. Acoust. Soc. Am., vol. 83, no. 4, pp. 1536–1552, April 1988.

[9] K. Ishizaka and N. Isshiki, "Computer simulation of pathological vocal-cord vibration," The Bell Syst. Tech. J., vol. 60, pp. 1193–1198, 1976.

[10] F. Avanzini, C. Drioli, and P. Alku, "Synthesis of the voice source using a physically-informed model of the glottis," in Proc. of the Int. Symposium on Musical Acoustics (ISMA), pp. 31–34, 2001, available at http://www.dei.unipd.it/~avanzini/papers.html.

[11] F. Avanzini, P. Alku, and M. Karjalainen, "One-delayed-mass modelfor efficient synthesis of glottal flow," Proc. of Eurospeech Conf, pp. 51–54, September 2001, available at http://www.dei.unipd.it/~avanzini/papers.html.

[12] C. Drioli, "A flow waveform adaptive mechanical glottal model," TMH-QPSR, vol. 43, pp. 69–79, 2002, available at http://www.speech.kth.se/qpsr/tmh/.

[13] K. Ishizaka and N. Ishiki, "Computer simulation of pathological vocal-cord vibrations," J. Acoust. Soc. Am., vol. 60, no. 5, pp. 1193–1198, 1976.

[14] H. Herzel, I. Titze, and I. Steinecke, "Nonlinear dynamics of the voice: signal analysis and biomechanical modeling," CHAOS, vol. 5, no. 1, pp. 30–34, 1995.