ACOUSTICS 2012

Analysis-synthesis of vocal sounds based on a voice production model driven by the glottal area

T. Hezard, T. Hélie, R. Caussé and B. Doval

Institut de Recherche et Coordination Acoustique/Musique, 1, place Igor Stravinsky 75004 Paris
thomas.hezard@ircam.fr

The source-filter paradigm has been widely used to model, synthesize and analyse vocal sounds. In spite of their efficiency, most of such models neglect some physical phenomena which could significantly improve naturalness. In this paper, we consider a modified source-filter model which includes a simplified aeroacoustic coupling between the glottal airflow and the vocal tract while keeping low-cost computation and efficient analysis methods.
We introduce a glottal area waveform model derived from observations on high-speed video-endoscopic recordings and based on the so-called Liljencrants-Fant (LF) model. The vocal tract is modeled by a concatenation of straight pipes with lossless plane waves propagation. The coupling is ensured by the standard Bernoulli equation, a flow-separation model and continuity constraints for acoustic pressure and flow at the inner end of the vocal tract. This voice production model is driven by the sub-glottal pressure, the glottal area and the vocal tract geometry. Moreover, we introduce a sound analysis method for this model.
In conclusion, we present some synthesis, analysis and transformation examples to evaluate the performances of this model and compare it with a classic source-filter model.

1 Introduction

In this paper, we present a voice production model based on simple well-known elements. With this model, we aim to highlight the importance of the coupling between the source and the filter.

This paper is organised as follows: in section 2 we present a glottal area waveform model and a method to estimate the model parameters from measured glottal area waveforms, in section 3 we present the complete voice production model and we discuss the relevance of this model in section 4.

2 Glottal area waveform model

2.1 Observation of high-speed videoendoscopic recordings

Among all the techniques of glottal activity exploration, high-speed videoendoscopy is one of the most accurate. Therefore, during former works at IRCAM, Gilles Degottex and Erkki Bianco have built a database of synchronised biometric signals measured during phonation including high-speed video-endoscopic recordings of glottal folds. These signals also include electroglottographic recordings (EGG) and voice recordings. Gilles Degottex established a method in order to extract glottal area waveform (GAW) from a high-speed video-endoscopic recording of the glottis [1].

In our work, we began by observing the glottal area waveforms extracted from the database. Figure 1 presents an image coming from a video presenting simultaneously high-speed video-enscopy, extracted GAW, recorded voice and some spectral features of GAW and sound.

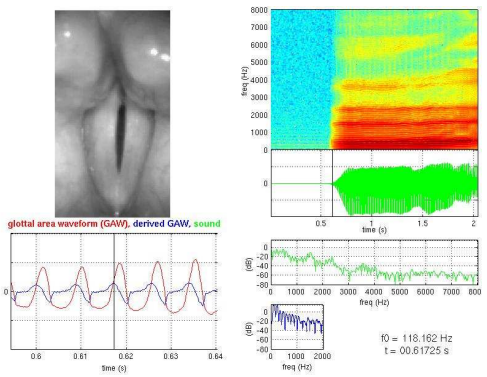

Figure 1: Image coming from a video presenting high-speed video-enscopy of the glottal folds, extracted GAW, recorded sound and some spectral features of GAW and sound.

Despite the huge variability of the extracted GAWs, we observed that most of the signals have a quite similar periodic pulse shape. A very common pulse model used in voice production model is the Liljencrants-Fant model (LF model) [2]-[3]. Initially conceived as a glottal flow model, this model reveals to be relevant as a GAW model. The general shape of the LF model, driven by one temporal parameter (the period), one amplitude parameter and four shape parameters is presented on Figure 2.

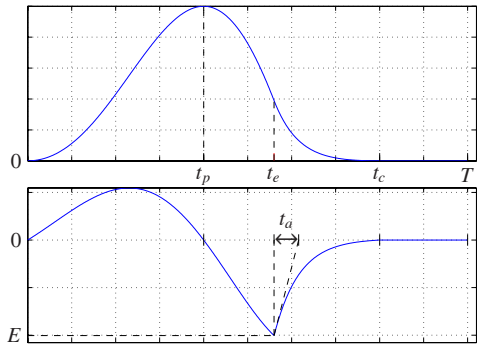

Figure 2: One period of LF model. This model has one temporal parameter T, one amplitude parameter E and four shape parameters $\{t_p, t_e, t_a, t_c\}$ (top: function, bottom: derived function).

2.2 Glottal area waveform estimation method

We have developped an estimation method of the LF parameters for measured glottal area waveforms. The goal of this algorithm is to estimate the LF parameters that give the best estimation of a target derived GAW signal extracted from the database.

Supposing that the target signal contains N periods, the estimation algorithm consists of N steps:

- rough estimation of the parameters by detecting maxima, minima and zeros on the target signal,
- $1 \leq n \leq N-1$: optimization of the LF parameters for periods n and $n+1$ with the simplex method [4], the criterion is the quadratic distance between the model and the target.

Figure 3 presents some results of this estimation method. One can see on this figure that, despite their variability, LF model is relevant to model GAWs. Nevetherless, this model is not able to capture tiny riples and other phenomena like

bouncing during the closing phase. Thus, new diveristy models for glottal area waveforms are presently under consideration.

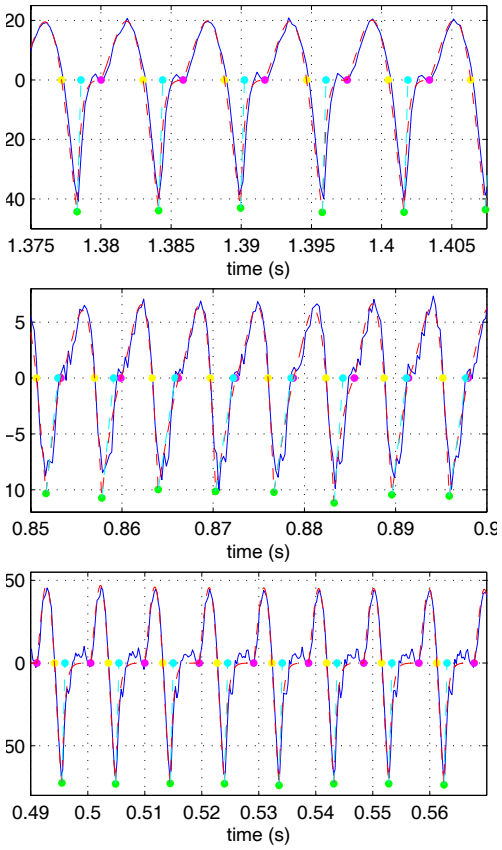

Figure 3: LF estimation examples.
(–) derived GAW, (– • –) LF estimation.

3 Voice production model

3.1 Vocal tract and radiation models

The source-filter model, in its most common version, considers an autoregressive filter. One can interpret this filter as an acoustic tube model. J.D. Markel and A.H. Gray pointed out the link between an autoregressive filter and the transmittance of a straight pipes concatenation with lossless propagation of plane waves [5]. In our model, we consider a similar vocal tract model, presented in Figure 4, under lossless plane waves propagation hypothesis. This vocal tract is entirely described by the area of its M sections $\{A_0 \cdots A_{M-1}\}$.

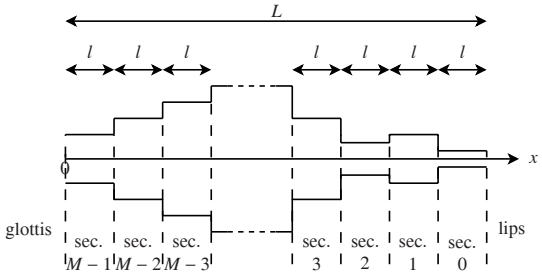

Figure 4: Vocal tract model

We introduce the following notations:

- $p(x,t)$ stands for the acoustic pressure inside the vocal tract at the abscissa x and at the time t and $P(x,s)$ for its Laplace tranform,
- $u(x,t)$ stands for the acoustic flow inside the vocal tract at the abscissa x and at the time t and $U(x,s)$ for its Laplace tranform,
- $o(t)$ stands for the output sound.

Additionally, we consider an ideal radiation model:

$$\left\{ \begin{array}{rcl} p(L,t) & = & 0\,, \\ o(t) & \propto & \dfrac{\partial u}{\partial t}(L,t)\,. \end{array} \right. \tag{1}$$

Under these hypothesis, the vocal tract presented in Figure 4 is modeled by the following set of equations in the Laplace domain:

$$\left\{ \begin{array}{rcl} U(L,s) & = & H_t(s)U(0,s) \\ P(0,s) & = & Z_{in}(s)U(0,s) \end{array} \right. , \tag{2}$$

$$\left\{ \begin{array}{rcl} H_t(s) & = & \dfrac{2c_{M-1}\,\mathrm{e}^{-2M\tau s}}{H^+_{M-1}(s)+H^-_{M-1}(s)} \\ Z_{in}(s) & = & \dfrac{\rho c}{A_{M-1}}\,\dfrac{H^+_{M-1}(s)-H^-_{M-1}(s)}{H^+_{M-1}(s)+H^-_{M-1}(s)} \end{array} \right. , \tag{3}$$

$$\left[\begin{array}{c} H^+_m(s) \\ H^-_m(s) \end{array} \right] = \prod_{i=m}^{1} \left[\begin{array}{cc} 1 & -\mu_i \\ -\mu_i\,\mathrm{e}^{-4\tau s} & \mathrm{e}^{-4\tau s} \end{array} \right] \left[\begin{array}{c} 1 \\ \mathrm{e}^{-4\tau s} \end{array} \right], \tag{4}$$

$$c_m = \prod_{i=1}^{m}(1-\mu_i)\,, \tag{5}$$

$$\mu_i = \frac{A_i - A_{i-1}}{A_i + A_{i-1}}\,, \tag{6}$$

where ρ stands for the air density, c for the celerity of sound and $\tau = \frac{l}{2c}$.

3.2 Remarks about the immitances

The digitized versions of the transmittance H_t and the input impedance Z_{in} obtained for the sampling period $T_s = 4\tau = 2l/c$ are

$$\left\{ \begin{array}{rcl} H_t(z) & = & \dfrac{2c_{M-1}z^{-\frac{M}{2}}}{H^+_{M-1}(z)+H^-_{M-1}(z)}\,, \\ Z_{in}(z) & = & \dfrac{\rho c}{A_{M-1}}\,\dfrac{H^+_{M-1}(z)-H^-_{M-1}(z)}{H^+_{M-1}(z)+H^-_{M-1}(z)}\,, \\ \left[\begin{array}{c} H^+_m(z) \\ H^-_m(z) \end{array} \right] & = & \prod_{i=m}^{1} \left[\begin{array}{cc} 1 & -\mu_i \\ -\mu_i z^{-1} & z^{-1} \end{array} \right] \left[\begin{array}{c} 1 \\ z^{-1} \end{array} \right]. \end{array} \right. \tag{7}$$

We can also write $H_t(z) = \dfrac{2c_{M-1}z^{-\frac{M}{2}}}{\sum_{k=0}^{M} h_k z^{-k}}$ and we can then show the following property

$$\left\{ \begin{array}{rcl} h_0 & = & h_M = 1\,, \\ h_k & = & h_{M-k} \quad 1 \le k \le M-1\,. \end{array} \right. \tag{8}$$

This property is a very strong constraint on the autoregressive filter H_t and have various consequences on its behavior. However, this property is entirely due to the ideal radiation model, i.e the perfect reflection at the outer end of the vocal tract, and will disappear as soon as the radiation model is modified.

3.3 Forced glottal folds model and aeroacoustic coupling

In our voice production model, we consider that the glottis is entirely described by its area $A_G(t)$, which means that it has no thickness. We consider the lungs and the trachea as an ideal pressure source, i.e. sub-glottal acoustic speed and flow are null. According to the Bernoulli's principle, considering an incompressible flow under quasi-stationarity hypothesis inside the glottis, we have

$$\left\{ \begin{array}{rcl} p_s(t) & = & p_G(t) + \frac{1}{2}\rho\, v_G(t)^2 \\ u_G(t) & = & A_G(t)\, v_G(t) \end{array} \right. , \tag{9}$$

where $p_s(t)$ stands for the sub-glottal pressure, $p_G(t)$ for the glottal pressure, $v_G(t)$ for the glottal acoustic speed and $u_G(t)$ for the glottal air flow.

The coupling between the glottis model (9) and the vocal tract model (2-6) is ensured by the continuity of the acoustic state at the inner end of the vocal tract

$$\left\{ \begin{array}{rcl} p_G(t) & = & p(0,t) \\ u_G(t) & = & u(0,t) \end{array} \right. . \tag{10}$$

The inputs of this source model are the sub-glottal pressure and the glottal area. Thus, contrary to the classic source-filter model, the acoustic state $[p_G(t), u_G(t)]$ is not forced but is induced by the model. In other words, while the classic source-filter model is driven by forced acoustics, our model is driven by forced geometry.

3.4 Complete voice production model

The complete voice production model is written

$$\left\{ \begin{array}{rcl} O(s) & = & sU(L,s)\,, \\ U(L,s) & = & H_t(s)U_G(s)\,, \\ P_G(s) & = & Z_{in}(s)U_G(s)\,, \\ U_G(t) & = & A_G(t)\,\xi(t)\sqrt{\frac{2}{\rho}|p_s(t)-p_G(t)|}\,, \\ \xi(t) & = & \mathrm{sgn}\Big(p_s(t)-p_G(t)\Big)\,. \end{array} \right. \tag{11}$$

where $H_t(s)$ and $Z_{in}(s)$ are defined in (3).

We can show that this system is not realizable and that a realizable version of this model is

$$\left\{ \begin{array}{rcl} O(s) & = & sU(L,s)\,, \\ U(L,s) & = & H_t(s)U_G(s)\,, \\ \widetilde{P}_G(s) & = & \mathrm{e}^{-4\tau s}\widetilde{Z}_{in}(s)U_G(s)\,, \\ U_G(t) & = & A_G(t)\,F\Big(p_s(t)-\widetilde{p}_G(t)\Big)\,, \\ F(x) & = & \mathrm{sgn}(x)\left(\sqrt{c^2\frac{A_G(t)^2}{A_{M-1}{}^2}+\frac{2}{\rho}|x|}-c\frac{A_G(t)}{A_{M-1}}\right) \end{array} \right. \tag{12}$$

where

$$\widetilde{Z}_{in}(s) = -\frac{\rho c}{A_{M-1}}\frac{2\,\mathrm{e}^{4\tau s}H^-_{M-1}(s)}{H^+_{M-1}(s)+H^-_{M-1}(s)}\,. \tag{13}$$

Finally, the digitized version of the model is immediately obtained by choosing the sampling period $T_s = 4\tau = \frac{2l}{c}$.

4 Examples and validation

4.1 Influence of the coupling

Figure 5 presents the influence of the coupling on the glottal airflow. Three sounds have been synthesized with the same subglottal pressure and the same GAW. The vocal tracts are differents for the three sounds and correspond respectively to the vowels \u\, \i\ and \o\ (from top to bottom). These vocal tract geometries come from [6]. As one can see, the influence of the coupling on the glottal airflow can be drastic.

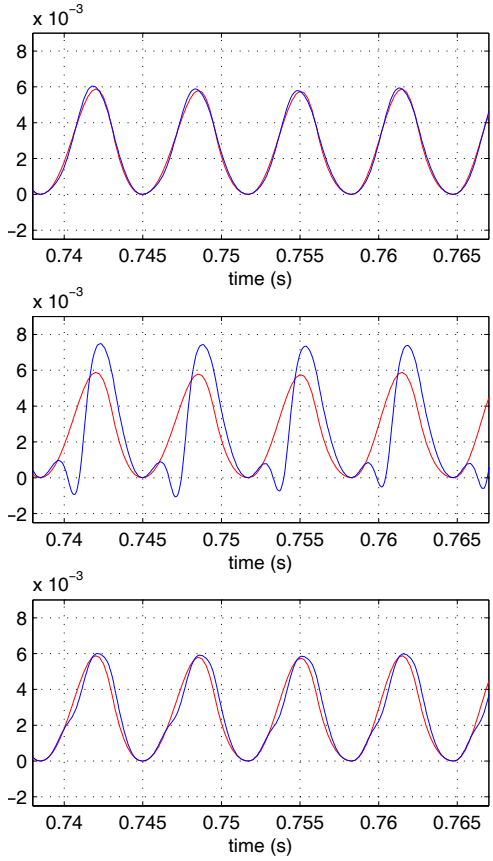

Figure 5: Comparison of $A_G(t)$ (-) and $u_G(t)$(-)
From top to bottom: \u\, \i\, \o\

4.2 Influence of the non-linearity

Figure 6 presents the influence of the non-linearity introduced by the coupling. 5 sounds have been synthesized with the same GAW (extracted from the database) and the same vocal tract (corresponding to the voyel \a\ [6]). We can observe that increasing the subglottal pressure induces an important spectral enrichment. This is only due to the simple non-linearity introduced by the Bernoulli's equation and corresponds to a natural behavior of voice production.

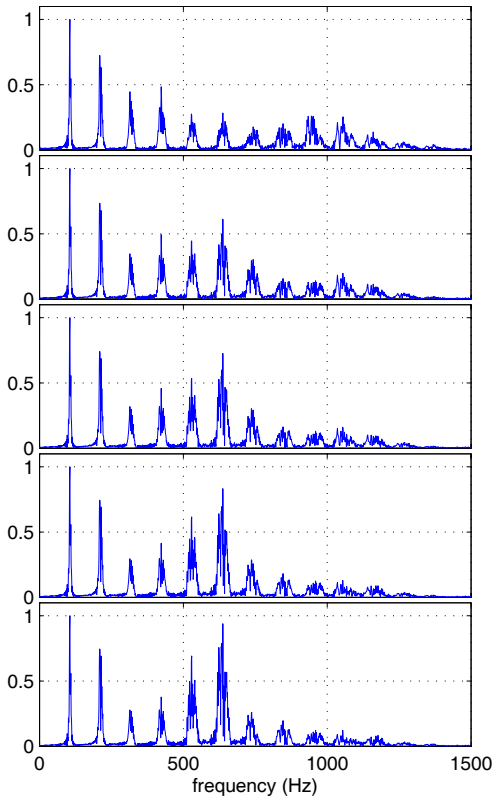

Figure 6: Evolution of the synthesized voice's spectrum when p_s inscreases.
From top to bottom: $p_s = 1cmH_2O$, $p_s = 5cmH_2O$, $p_s = 10cmH_2O$, $p_s = 20cmH_2O$, $p_s = 40cmH_2O$

5 Conclusion

In this paper, we presented a voice production model based on a modified source-filter model which includes a simplified aeroacoustic coupling between the glottal airflow and the vocal tract. We also presented an analysis method for LF parameters estimation on measured glottal area waveforms. We observed the relevance of the LF model as a glottal area waveform model and we saw that the coupling between the glottal airflow and vocal tract is crucial for the naturalness of the sound and the improvement of the model driving. A global analysis method, estimating all the model's parameters to re-synthesize a measure from the database is currently under development.

Furthermore, the perspectives of this work are a global improvement of the model. In our work, we start from the source-filter model and we try to introduce step by step some physical phenomena which could significantly improve naturalness by building at each step a new model and the corresponding analysis method.

Finally, we are also working on an electronic device for in vivo and non invasive exploration of the glottal activity. This device should bring much more information than actual technologies like electroglottography and should help us to build more accurante glottal source models.

References

[1] G. Degotex, "Glottal source and vocal-tract separation.", *PhD Thesis*, UPMC-Ircam (2010)

[2] G. Fant, J. Liljencrants and Q. Lin, "A four-parameter model of glottal flow.", *STL-QPSR* **26-4**, 1-13 (1985)

[3] G. Fant, "The LF-model revisited. Transformations and frequency domain analysis.", *STL-QPSR* **2-3**, 119-155 (1995)

[4] J.C. Lagarias, J.A. Reeds, M.H. Wright, P.E. Wright, "Convergence Properties of the Nelder-Mead Simplex Method in Low Dimensions.", *SIAM Journal of Optimization* **9-1**, 112-147 (1998)

[5] J.D. Markel, A.H. Gray, *Linear prediction of speech*, Springer-Verlag, Berlin, New York (1976)

[6] B.H. Story, I.R. Titze, "Parametrization of vocal tract area functions bu empirical orthogonal modes.", *Journal of Phonetics* **26**, 223-260 (1998)

[7] A. Roebel, X. Rodet, "Efficient Spectral Envelope Estimation and its application to pitch shifting and envelope preservation.", *International Conference on Digital Audio Effects*, 30-35 (2005)