ISCA Archive
http://www.isca-speech.org/archive

INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

Speech modeling and processing by low-dimensional dynamic glottal models

Carlo Drioli[1], *Andrea Calanca*[2]

[1]Department of Mathematics and Computer Science, University of Udine, Udine, Italy
[2]Department of Computer Science, University of Verona, Verona, Italy
carlo.drioli@uniud.it, andrea.calanca@univr.it

10.21437/Interspeech.2012-442

Abstract

We discuss the use of low-dimensional physical models of the voice source for speech coding and processing applications. A class of waveform-adaptive dynamic glottal models and parameter tracking procedures are illustrated. The model and analysis procedures are assessed by addressing signal transformations on recorded speech, achievable by fitting the model to the data, and then acting on the physically-oriented parameters of the voice source. The class of models proposed provides in principle a tool for both the estimation of glottal source signals, and the encoding of the speech signal for transformation purposes. The application of this model to time stretching and to frequency control (pitch shifting) is also illustrated. The experiments show that copy synthesis is perceptually almost indistinguishable form the target, and that time stretching and "pitch extrapolation" effects can be obtained by simple control strategies.

Index Terms: speech synthesis, glottal modeling, speech coding, physical modeling

1. Introduction

Despite the fact that physical models of the speech production including the glottal source have nowadays reached a high degree of accuracy, it is remarkable that they are rarely found in common applications like speech processing and speech synthesis. The widespread linear prediction coding (LPC) technique for speech coding, is only loosely inspired by voice acoustics and it is properly a signal model more than an physical one, even if LPC coefficients are related to the shape of the vocal tract. If we look at the computer graphics community, we note how much effort is being devoted recently to the development of effective physically based models and model-based parameter tracking algorithms, and how this has led to highly realistic and natural animations. The reason of such a difference is probably the lack of compact articulatory speech models and robust model inversion techniques that can provide a framework to accurately represent recorded data and allow robust parameters control at the same time. One of first papers addressing this issue is [1], in which first experiments targeted at demonstrating the feasibility of deriving the control parameters of dynamical physical models are reported. Recent investigations that have also addressed the representation of speech through physically-inspired source-tract models include the use of analytical models of the glottal flow for joint source and vocal tract filter optimization [2, 3], with effective results. This class of source models however cannot reproduce the dynamical properties of the glottis, which is a desirable characteristic for a speech model. Growing interest toward accurate glottal source coding for speech synthesis applications has also been recently demonstrated [4].

In this paper, we discuss the use of a class of glottis models characterized by a low dimensional dynamics for applications in the framework of model-based speech coding, glottal source estimation and voice transformations. The voice source model proposed is a source-filter scheme in which the vocal tract is represented by an all-pole filter and the voice source model relies on a lumped mechano aerodynamic scheme inspired by the mass-spring paradigm. In previous investigations, we discussed the possibility of fitting this class of low-dimensional physically constrained models to real voice samples [5], and illustrated its stability and control characteristics [6]. Here, we focus on the fitting of time varying voice samples (i.e., short speech utterances), and provide the basis for voice transformations.

The paper is organized as follows. In Sec. 2, the model is illustrated and the inversion procedure used to fit voice data is discussed; in Sec. 3, the method proposed is applied to speech data and its performance is evaluated with respect to speech encoding. Results on simple time stretching and pitch shifting transformations are reported; in Sec. 4, the conclusions are drawn.

2. Speech model and parameter tracking

Let the lip pressure signal measured by the microphone be given by

$$y(t) = -\sum_{k=1}^{N} a_k y(t-k) + u_g(t) \qquad (1)$$

where $a_1, \ldots, a_N$ are the auto regressive (AR) coefficient of an all-pole model of the vocal tract, and $u_g(t)$ is the excitation glottal pulse waveform. The voice source model used to represent u_g relies on the mass-spring paradigm adopted, among others, by the well known Ishizaka-Flanagan one-mass and two-mass models. The details of the glottal excitation model, illustrated in Fig. 1, can be found elsewhere [7], and here we only briefly recall the essential components. The lower edge of the folds is represented by a single mass-spring system $\mathcal{H}_{res}$ and the propagation of the displacement x along the thickness of the fold is represented by a delay line of length τ. Let x_1 be the displacement of the fold at glottis entrance, and x_2 the displacement at the exit. 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 p_m acting on the folds is computed from the flow u_g and the lower glottal area a_1 using Bernoulli's law. A flow model $\mathcal{F}$ converts the glottis area given by the fold displacements into the airflow at the entrance of the vocal tract. In its simplest form, the glottis area is computed as the minimum cross-sectional area between the area a_1 at lower vocal fold edge and the area a_2 at upper vocal fold edge, and the flow is assumed proportional to the glottal area, i.e. $\mathcal{F}_0(x1, x2) = k_g \min(x_1, x_2)$ (where the lung pressure p_l

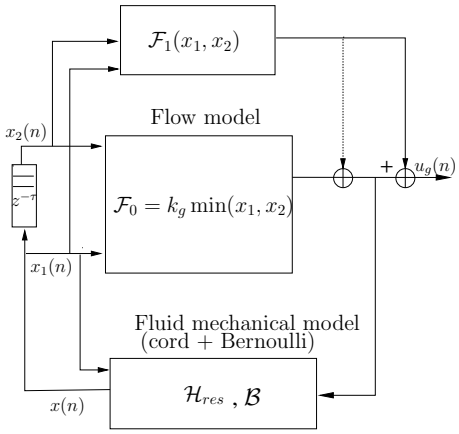

Figure 1: Scheme of the low-dimensional voice source used as glottal waveform generator (note that the vocal tract model is not represented here). The flow model is splitted into a basic component ($\mathcal{F}_0$), and a refinement component based on a kernel machine ($\mathcal{F}_1$).

is included in k_g). The propagation line of length τ reproduces the vertical phase difference of the vibration of the cord edges, which is essential for the production of self-sustained oscillations without a vocal tract load. The pressure lung, p_l, has a role in determining the onset and offset of the oscillation. In our simulations, it is kept constant during the system evolution and is omitted for simplicity in what follows. The mass-spring system $\mathcal{H}_{res}$ is modeled as a second-order resonant filter, characterized by a resonance frequency f_0.

A refined flow model in which a kernel machine component $\mathcal{F}_1(x1, x2)$ is aimed at improving the flow waveform matching properties of the basic model, was introduced in [7] (Fig. 1, dashed signal path). We further explore this solution by using a modified scheme (illustrated in Fig. 1 by a continuous path in place of the dotted path), in which the kernel machine component do not interfere with the dynamics of the main iterated map. This solution ensures that the stability of the system is not affected by the flow modifications introduced by kernel training. The details of the training procedure to match real glottal flow data have been discussed in [7]. In brief, the kernel component reshapes the surface on which the trajectory of the system, at the flow model inputs and output, lies during the periodic motion, thus refining the rough flow model. We are interested in changing the frequency of oscillations while keeping the waveform characteristics learned from the data and stored in the kernel machine component. Changes of the frequency of oscillations of the glottal pulse, while preserving waveform characteristics and formants, will ideally provide a physiologically-constrained pitch shift transformation. The result of pitch control with the in-loop scheme, however, often results in performance degradation, due to the influence of the kernel component in the iterated map dynamics. This situation is illustrated in Fig. 2, in which the glottal waveforms as well as the trajectories in the (x_1,x_2,u_g) space are depicted. In the alternative solution, since the contribution of the kernel machine does not affect the iterated map dynamics of the glottal model, the resulting glottal waveforms retains the same characteristics of the original period shape at the new oscillation frequency, as shown in Fig. 3. For this reason, this solution seems to be a better choice for the modeling finalized at speech transformations, although the in-loop scheme is more correct from a modeling point of view. In the following, we will refer to the stable out-of-loop scheme when talking of refined glottal flow waveform.

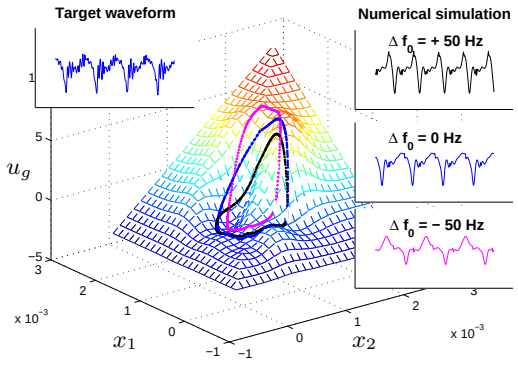

Figure 2: Result of pitch control using a scheme in which the flow refinement affects the closed loop dynamics. The target waveform is a glottal flow estimate obtained by inverse filtering the pressure waveform through the all-pole model of the vocal tract. The trajectories describing the dynamics of the flow part of the system ($\mathcal{F}_0 + \mathcal{F}_1$) is also shown.

2.1. Model inversion

In the following, the model is fitted to time-varying recorded speech data, as summarized in figure 4. To this aim, a pitch-synchronous parameter identification procedure is used, which performs a joint source-vocal tract identification through the following steps:

- a fixed length running analysis window is shifted by a variable hop size equal to the period length.
- for the analysis frame under investigation, whose length corresponds to around three periods of speech, a traditional LPC analysis is used to obtain a rough estimate of the formants.
- the fundamental frequency estimated through a pitch detector is used to tune the mass-spring system representing the folds, and the glottal model is used to generate a glottal pulse.
- a least-square fitting procedure, based on QR factorization, is used to solve the estimation problem which provides the parameters of the vocal tract filter given its time aligned input (the glottal source) and output (the target speech signal) time series.

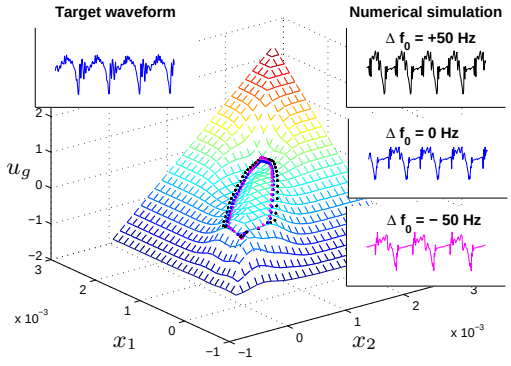

Figure 3: Result of pitch control using the solution in which the flow refinement does not affects the closed loop dynamics.

- the glottal pulse is finally refined by training the kernel machine component $\mathcal{F}_1(x1, x2)$ to match the target glottal pulse, obtained by inverse filtering the target speech waveform through the vocal tract filter obtained in the previous step.

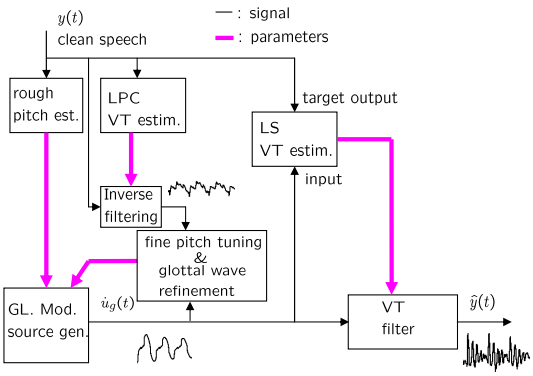

Figure 4: Parameter tracking scheme.

During the parameter tracking process, the tuning of the vocal fold model is performed in two steps: first, a rough estimate of the mass-spring system parameters is given by a pitch detection algorithm; in the second step, this estimate is refined through an iterative procedure which at each iteration j attempts at reducing the time lag $\xi(j-1)$, evaluated through the autocorrelation between the glottal pulse and the target signal, inverse filtered using the LPC filter. The fine tuning of the pitch is achieved using the adaptation formula $f_0(j) = f_0(j-1) - \epsilon\xi(j-1)$.

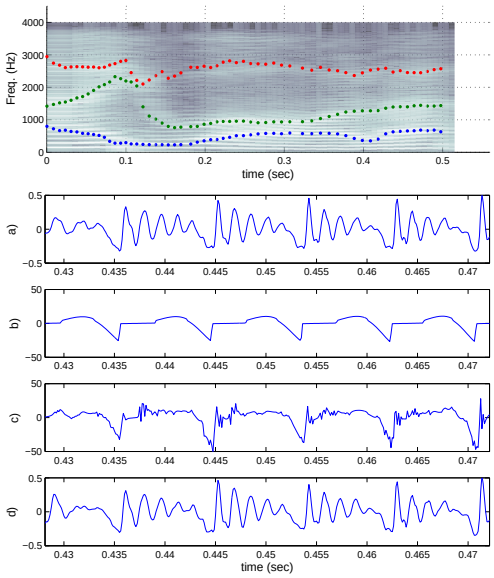

Figure 5: Example of matching a speech signal with the proposed model. Upper panel: spectrogram of the target signal and tracks of identified formant frequencies; panel a): target signal frame within sec 0.41 and sec 0.47; panel b): rough glottal waveform; panel c): refined glottal waveform; panel d): reconstructed speech waveform.

This two-steps procedure is required to ensure that the glottal pulse signal and the target waveform are kept synchronized. Due to the nonlinear nature of the dynamical glottal model, the resulting glottal pulse may in fact be characterized by an actual fundamental frequency f_0' which can deviate from the control frequency f_0.

Table 1: Segmental SNR and IS distance values for distinct modeling settings. Label *LPC no sync* refers to an LPC analysis procedure excited by a non-synchronized glottal pulse; *LPC* refers to an LPC analysis procedure excited by a synchronized glottal pulse; *Gl+LS* refers to a least squares identification problem with rough flow model; *Gl(ref)+LS* refers to a least squares identification problem with refined flow model; Segmental SNR and IS values are averaged over 60 frames, each one corresponding to a waveform period.

	LPC no sync	LPC	Gl+LS	Gl(ref)+LS
SNR	-3.10	2.67	2.95	19.28
IS	0.91	0.40	0.33	0.19

3. Application to speech transformations

The illustrated tracking procedure was used to encode a male voice sample recorded at 8KHz, 16 bit, mono, reproducing the Italian word *aiuola*. The LPC order was set to 8, allowing to detect up to 4 formants, however only the first 3, the more stable ones, were retained at the end of the procedure. Figure 5 shows qualitative results in terms of reconstructed glottal waveform. In Table 1, coding performances of the proposed model are compared in terms of segmental signal to noise ratio (SNR) and Itakura-Saito distance (IS) (average of SNR and IS values calculated for segments of data). First column refers to a reference LPC analysis procedure in which a standard LPC of order 8 is used to compute the AR filter, an autocorrelation pitch detection algorithm is used to control the generation of a glottal pulse (rough version), and no fine pitch tuning is performed. Thus, the target and reconstructed speech may loose synchronization, and this explains the very poor SNR performance; the second column differs from the first in that the fine tuning step is introduced, thus the two waveforms are guaranteed to be in phase; third column refers to the situation in which the vocal tract AR estimate is performed by solving a least squares problem considering the generated glottal pulse as input, and the target speech signal as output. Input and output are kept in phase by the fine tuning procedure. The last column differs from the preceding one, in that the refinement step of the glottal pulse is introduced. The kernel machine used in these experiments was a radial basis function (RBF) with gaussian kernels. A constant number of 30 kernels was used for each analysis frame. It can be seen how the fine tuning and the glottal pulse refinements steps improves considerably the matching performances.

The analysis performed with the tracking procedure was then used to synthesize transformed versions of the original sample. The following transformations are evaluated:

- time stretch, obtained by allowing the glottal model to run with a given set of analysis parameters for longer or shorter time intervals with respect to the corresponding analysis hop time intervals.
- pitch shift, obtained by multiplying by a scale factor the analysis mass-spring system parameters responsible for the oscillation frequency.

Other feasible transformations, to be investigated in the future, include concatenation of variable length speech units (e.g.,

phonemes, diphones, words, etc.), and transformations of voice quality obtained by acting on other parameters of the source model.

Figure 6 shows the spectrograms of the result of time compression (upper panels) and expansion (lower panel). Figure 7 shows the details of two pitch shifting transformations, produced by $\Delta f_0 = +30\%$ (middle plots) and by $\Delta f_0 = -30\%$ (lower plots). vspace-0.5cm

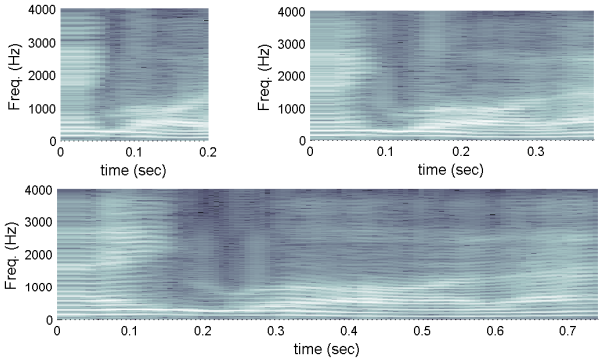

Figure 6: Spectrograms showing time stretching transformations of the original sample of Fig.5: upper left and upper right, a time compression of a factor of 0.4 and 0.8 respectively (the original sample has a total duration of 0.52 sec); lower panel: a time expansion of a factor 1.4.

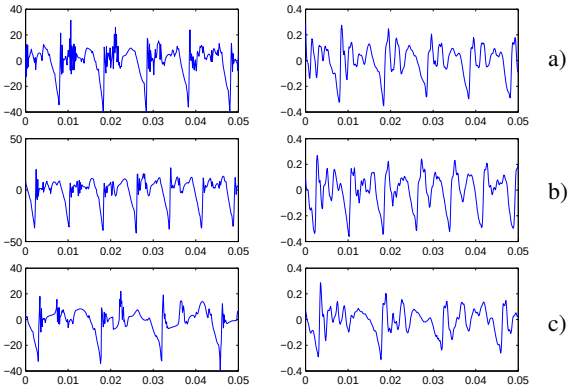

Figure 7: Pitch transformations: a) a 20 msec frame of the signal reconstructed from analysis (glottal flow derivative on the left, lip pressure on the right); b) the reconstruction when $\Delta f_0 = +30\%$ is used to tune the cord model; c) same as b) with $\Delta f_0 = -30\%$

From a perceptual point of view, the copy synthesis version of the target speech samples obtained by using the rough version of the glottal pulse model are characterized by a low-pass average quality. The characteristics of the original speech and speaker are however retained, and the uttered word as well as the speaker are perfectly recognizable. When the refined glottal pulse is used in the copy synthesis, the reconstructed speech can be made almost indistinguishable from the target, by using an adequate number of kernel terms (30 where sufficient in the the experiment reported here). When transformations are performed by acting on the analysis parameters as described, the resulting speech signal retains the characteristics of the copy synthesis, although a slight degradation of the sound quality is audible. This is less pronounced in time stretching transformations, where few artifacts are noticed, probably due to signal discontinuities arising from inaccuracies in the formant tracking (no formant track smoothing procedures were used at this time). In any case, the resulting signal is not affected by the well known phase-related degradation typical of processing based on phase vocoder or sinusoidal modeling. In pitch shifting transformations, a slight loss of speech timbre naturalness is observed as the distance from original pitch increases. We believe that this is due to the fact that the analysis parameters of the formants are not modified when rising or lowering the pitch of the glottal source, which is actually not too realistic since the original interaction relating source and vocal tract is not retained. Further investigation is foreseen with respect to this point.

4. Conclusions

We discussed the use of low-dimensional physically based speech models in the frameworks of speech coding and speech processing. The voice model used provides self-sustained oscillations and data fitting capability that can be used to adapt the model to recorded speech. The class of models proposed provides in principle a tool for both estimating glottal source signals, as well as for encoding the speech signal for transformation purposes. The application to time stretching and to frequency control (pitch shifting) of these schemes was also reported. The experimental results showed effective results as for time stretching, and that in principle pitch extrapolation is effectively obtained with the class of models proposed. Improvements in the tracking and smoothing of parameters identification and in the control of formants for pitch shifting, might further improve the final quality of the signal.

Future work will investigate the model versatility with respect to different transformations, especially voice quality, and if the inherent dynamical structure can improve the perceived naturalness of transformed speech, if compared to dynamic-less glottal models [8].

5. References

[1] J. Schroeter and M. Sondhi, "Speech coding based on physiological models of speech production," *In: Sondhi, MM, Furui, S. (Eds.), Advances in Speech Processing. Marcel Dekker, New York*, pp. 231 – 268, 1991.

[2] Q. Fu and P. Murphy, "Robust glottal source estimation based on joint source-filter model optimization," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 2, pp. 492 – 501, Mar. 2006.

[3] P. Jinachitra and J. O. Smith, "Generative model of voice in noise for structured coding applications," in *ICASSP (1)*, 2007, pp. 281–284.

[4] T. Raitio, A. Suni, J. Yamagishi, H. Pulakka, J. Nurminen, M. Vainio, and P. Alku, "Hmm-based speech synthesis utilizing glottal inverse filtering," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 1, pp. 153 –165, Jan. 2011.

[5] C. Drioli, "Synthesis of voiced sounds by means of waveform adaptive physical models," *in Proc. of Stockholm Music Acoustics Conference (SMAC)*, pp. 377–380, 2003.

[6] C. Drioli and F. Avanzini, "Non-modal voice synthesis by low-dimensional physical models," *in Proc. 3rd Int. Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications (MAVEBA)*, 2003.

[7] C. Drioli, "A flow waveform-matched low-dimensional glottal model based on physical knowledge," *J. Acoust. Soc. Am.*, vol. 117, no. 5, pp. 3184–3195, May 2005.

[8] Y. Agiomyrgiannakis and O. Rosec, "ARX-LF-based source-filter methods for voice modification and transformation," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, April 2009, pp. 3589 –3592.