



Voice processing by dynamic glottal models with applications to speech enhancement

Carlo Drioli, Andrea Calanca

Department of Computer Science, University of Verona, I-37134 Verona, Italy

carlo.drioli@univr.it, andrea.calanca@univr.it

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 speech encoding and enhancement, achievable by using a state space version of the dynamical model in a Extended Kalman filtering framework. The proposed method is shown to provide better SNR improvement if compared to a standard AR Kalman filtering scheme.

Index Terms: Speech enhancement, 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 autoregressive moving average (ARMA) process [1] and linear prediction coding (LPC) [2] techniques for speech coding, are only loosely inspired by voice acoustics and they are properly signal models more than a physical one, even if LPC coefficients are related to the shape of the vocal tract. Recent investigations that have addressed the representation of speech through physically-inspired source-tract models include the use of analytical glottal source models for joint source-filter model optimization [3, 4], 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. Experiments targeted at demonstrating the feasibility of deriving the control parameters of dynamical physical models have been reported for example in [5]. A well known application of AR and ARMA modeling, in combination with the Kalman filtering framework, is the enhancement of speech corrupted by noise[6]. The linear nature of the ARMA modeling makes them suitable to be used in state-space linear Gaussian modeling and filtering. Recently, however, interest has been demonstrated toward speech enhancement processing through nonlinear models of speech productions and general state-space models [7, 8].

In this paper, we discuss the use of a class of low dimensional glottis models for applications in the framework of speech processing, glottal pulse estimation, and speech enhancement. 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. With respect to previous investigations, in which we mainly discussed the dynamical properties of this class of low-

dimensional physically constrained models and the possibility of fitting real voice samples [9], we focus here on the possibility and the effectiveness of using them in combination with the Kalman filtering framework.

The paper is organized as follows. In Sec. 2, the model is illustrated and its state space form is derived. Also, the procedure for parameters and states estimation is sketched; in Sec. 3, the method proposed is applied to speech data and its performance is compared with a standard AR Kalman filtering scheme; 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

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

where a_1, \dots, a_N are the AR coefficient of an all-pole model of the vocal tract, $u_g(t)$ is the excitation glottal pulse waveform, and $n(t)$ represents additive background noise. The voice source model used to represent u_g relies on the mass-spring paradigm adopted, among others, in 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 [10], and here we only briefly recall the essential components. The lower edge of the folds is represented by a single

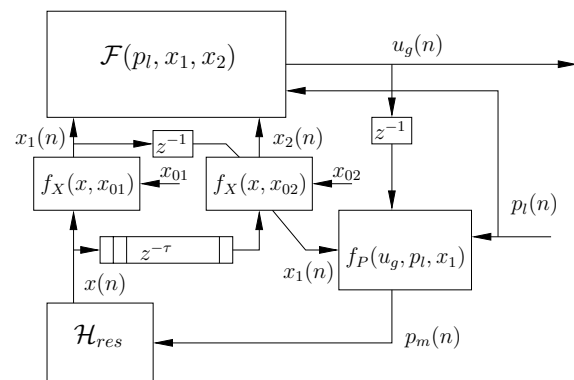


Figure 1: Schematic of the low-dimensional voice source used as glottal waveform generator (note that the vocal tract model is not represented here).

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 us call x_1 the displacement of the fold at

10.21437/Interspeech.2011-502

the entrance of the glottis, and x_2 the displacement at the exit. An impact model f_X 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 (f_P in the scheme). 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. $\mathcal{F}(x_1, x_2) = k_g \min(x_1, x_2)$ (where the lung pressure term p_l is included in the parameter k_g). The propagation line of length τ reproduces the vertical phase difference of the vibration of the cord edges, which is an essential element for the production of self-sustained oscillations without a vocal tract load. The pressure lung, p_l , has its principal role in 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. A refined flow model has been also explored in past investigations, in which a kernel machine is used to adapt the flow model to real flow waveforms obtained by inverse filtering from recorded speech. However, we decided not to include this component in the present study, to avoid overfitting problems that could arise when dealing with in noisy speech.

To use the phonation model sketched above in a recursive state estimation framework, we derive the state-space representation as follows. Let $X = [\mathbf{x}_d, \mathbf{x}_{res}, x_1, u_g, \mathbf{x}_{ar}]^T$ be the state of the whole system in (1), in which $\mathbf{x}_d = [x_{d,1}, \dots, x_{d,\tau}]$ is the state of the delay line $z^{-\tau}$, \mathbf{x}_{res} , u_g and x_1 are respectively the state of the mass-spring system, the output flow and the displacement of the vocal fold lower edge, and $\mathbf{x}_{ar} = [x_{ar,1} \dots x_{ar,N}]$ is the state of the N -order AR model of the vocal tract. The state transition at each discrete time instant n can be now written as

$$X^{n+1} = LX^n + NL(X^n), \quad (2)$$

where the linear and nonlinear dynamics have been explicitly separated. The linear state transition matrix L and the nonlinear state transition map $NL(X^n)$ can be written respectively as

$$L = \begin{bmatrix} \mathbf{I}_d & \mathbf{S}_C & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_{res} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{B}_{ar} & \mathbf{A}_{ar} \end{bmatrix}, \quad (3)$$

and

$$NL(X) = \begin{bmatrix} \mathbf{S}_D \cdot p_m \\ \mathbf{B}_{res} \cdot p_m \\ f_X(x, x_{01}) \\ \alpha \min(\gamma h_1, h_2) \\ 0 \end{bmatrix}, \quad (4)$$

with

$$p_m = f_P(u_g, x_1) \quad (5)$$

$$x = \mathbf{C}_{res} \mathbf{x}_{res} + \mathbf{B}_{res} p_m \quad (6)$$

$$h_1 = f_X(y_{res}, x_{01}) \quad (7)$$

$$h_2 = f_X(x_{d,\tau}, x_{02}) \quad (8)$$

and

$$\mathbf{S}_C = \begin{bmatrix} \mathbf{C}_{res} \\ 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \end{bmatrix}, \quad \mathbf{S}_D = \begin{bmatrix} \mathbf{D}_{res} \\ 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \end{bmatrix} \quad (9)$$

In the above matrices, \mathbf{I}_d represents the shift of the delay line of length τ , \mathbf{A}_{res} , \mathbf{B}_{res} , \mathbf{C}_{res} , and \mathbf{D}_{res} are the state update and output matrices of the \mathbf{H}_{res} second order resonant filter, and \mathbf{A}_{ar} , \mathbf{B}_{ar} are the state update matrices of the AR N -th order filter modeling the vocal tract. Finally, the output speech pressure is given by the last element of the state vector, i.e. $y(n) = [0 \dots 1]X$. We note that the linear matrix contains not only the linear model of the vocal tract, but also the part of the glottal model dynamical system that can be represented with linear state update relations.

In the following of the paper, the model is fitted to time-varying recorded speech data. To this aim, a pitch-synchronous parameter identification procedure is used, which performs 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 is estimated through an autocorrelation based pitch detector, and is used to tune the mass-spring system representing the folds. Then, the glottal model is used to generate a glottal pulse tuned with the speech signal. This pulse is aligned to the voice source obtained by inverse filtering through the LPC filter.
- the LPC parameters are further refined by considering the glottal pulse and the speech signal as input and output of a AR process, and solving it as a least squares problem. Finally, only complex conjugate AR coefficients are considered and mapped into formant frequencies and bandwidths. The AR order is set to 8, thus at most 4 formants can be represented.

At this time, other parameters that are considered to be less relevant to the effectiveness of the modeling, such as the length τ of the delay line and the displacement of the fold x_0 , are selected empirically and kept constant.

3. Application to speech enhancement and glottal source estimation

Linear auto-regressive models and Kalman state estimation algorithms are among the most effective and appreciated tools for speech enhancement [11]. With the aim of exploring the opportunity of using a nonlinear source model, the speech production model proposed was applied to a speech enhancement task, by performing a Kalman filtering along the pitch synchronous parametric training illustrated above. Due to the nonlinear nature of (part of) the state update equations, an Extended Kalman Filter (EKF) procedure was applied on a linearization of the system, obtained by computing the Jacobian of the map $NL(X)$ with respect to the state. Actually, it was observed that by letting the Kalman filtering process act mainly on a subset of the states, the performance improved. This is due to the fact that the model is a non linear loop and some states are more robust to perturbation than others. It was also noticed that the observer can go out of phase with respect to target speech signal because of the pitch error which is cumulative. To help the Kalman state estimation to keep the synchronization with the speech signal, an augmented state was used which included the fold frequency resonance, considered as a simple random walk processes. Also we

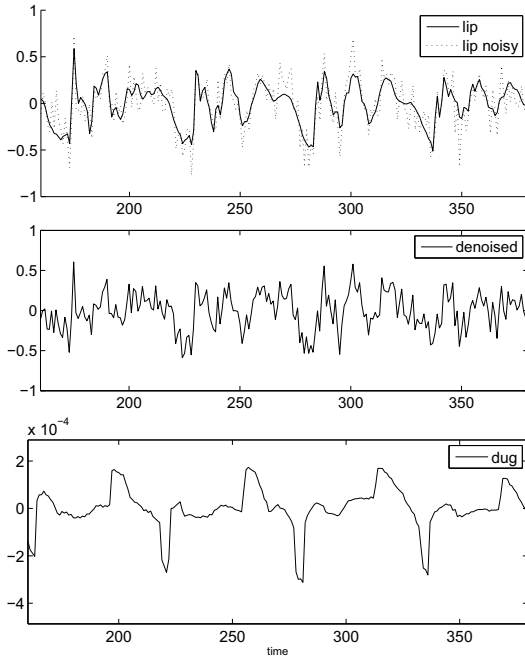


Figure 2: Example of filtering a speech waveform with additive stationary noise. Noisy speech (upper panel, dashed line) has a SNR of 0 dB with respect to clean speech (upper panel, cont. line). Enhanced speech (center panel) has a SNR of 2.4 dB. Lower panel shows the estimate of the glottal flow derivative, provided as part of the system state estimation process.

introduced an additive noise state to consider the gap between the simplified glottal model and the real glottal pulse.

The algorithm illustrated has been applied on a male voice sample recorded at 16KHz, 16 bit mono, and reproducing the Italian word *aiuola*. Stationary white gaussian noise was generated and added to the signal to obtain noisy speech. The result of the Kalman filtering on a speech frame with a three period length is illustrated in Fig. 2, for two different input SNR values. We note that the Kalman filtering provides an estimate of the speech frame as well as an estimate of a source glottal pulse which best fits the constraints given by the glottal model, and given the data. The use of this scheme as an enhanced source-vocal tract deconvolution algorithm for true glottal flow estimation and for improved formant frequency tracking will be the subject of further investigations.

The enhancement procedure was then evaluated through SNR and Itakura-Saito (IS) spectral distance measures. In each working condition the experiment was repeated one hundred times and the mean SNR and IS distance are reported (additive noise was randomly generated for each experiment, and used to corrupt the same speech signal). The two plots in Fig. 3 and 4, show the resulting improvements due to the enhancement, compared to a standard Kalman enhancement procedure based on a linear AR modeling of order 8. In Fig. 5, we represent each experiment by SNR and IS distance as two-dimensional coordinates. This representation shows that standard AR method performances are affected by background noise: the more the noise increases, the more the filtering process lead to variable results in terms of IS distance. In other words, not only the mean of the IS distance increases with noise but also its variance, leading to more unpredictable results. On the other hand, it seems that

	model	ref	data
SNR	1.40	2.42	-0.0059
IS Distance	60.05	20.87	8.15

Table 1: Effect on SNR and IS distance, of varying the confidence weight from the model to the data. The label *model* represents a configurations of variances that determines a high model confidence. Conversely, the label *data* represents a configurations of variances that determines high confidence to the data. The *ref* label correspond to a balanced confidence configuration, and is the one used for all the plots presented here.

model based filtering isn't affected by such problem.

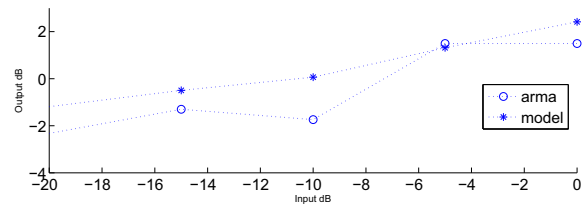


Figure 3: Comparison of SNR between standard linear AR and the proposed model based enhancement.

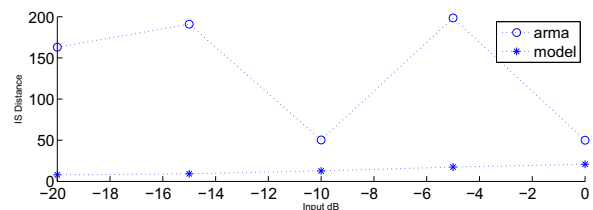


Figure 4: Comparison of IS distance between standard linear AR and the proposed model based enhancement.

It is known that Kalman filtering results are sensible to the initial tuning, and acting on the variance matrices, it is possible to tune the behavior of the algorithm. In particular, we decided to use a higher variance for the states of the source model that are more stable, and we assume the vocal tract as more accurate than the source model. As we also refine the estimate of the vocal cords oscillation frequency by the Kalman recursions, we tuned the variance in order to obtain a reasonable behavior, according to the pitch detection algorithm. However, the variance ratio between the whole model and the data is a free variable that allows to vary the confidence degree continuously from the model to the data. The effect of this variable on the measures is reported in Table 1. Note that, the more the confidence is given to the model, the more the noise is reduced but also the IS distance increases. This effect can explain the positive trend of model IS distance in Fig. 4. This is probably due to a frequency shaping characteristic of the model that we will investigate in future work.

4. Conclusions

We discussed the use of low-dimensional physically based speech models in the frameworks of speech enhancement and speech processing. The model scheme 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 and enhanced speech, as well as for effective encoding of the speech signal. The phonation source modeling based on the mass-spring system coupled to the simple flow model was shown to be suitable to be put in state space form and to be effectively coupled to an Extended Kalman filtering scheme. In our Kalman based algorithm, tracking of the formant frequencies and random walk modelling of the glottal pulse frequency are also implemented. Improvements over standard linear AR modeling was shown by experimental result both in terms of SNR improvement and IS distance reduction. Also we shows that glottal model based filtering leads to more predictable behaviour in terms of IS distance which variance results in a rather independent behaviour in respect with the level of background noise.

5. Acknowledgements

This study has been supported by the European Social Fund (ESF, project n. 1695/1/1/2215/2009 "Integration of advanced speech based interaction functionalities in PBX systems.").

6. References

- [1] H. Morikawa and H. Fujisaki, "System identification of the speech production process based on a state-space representation," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 32, no. 2, pp. 252 – 262, Apr. 1984.
- [2] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561 – 580, 1975.
- [3] 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.
- [4] P. Jinachitra and J. O. Smith, "Generative model of voice in noise for structured coding applications," in *ICASSP (1)*, 2007, pp. 281–284.
- [5] 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.
- [6] K. K. Paliwal and A. Basu, "A speech enhancement method based on Kalman filtering," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 12, Apr. 1987, pp. 177–180.
- [7] S. Windmann and R. Haeb-Umbach, "Iterative speech enhancement using a non-linear dynamic state model of speech and its parameters," in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, vol. 1, May 2006, p. 1.
- [8] J. Vermaak, C. Andrieu, A. Doucet, and S. Godsill, "Particle methods for bayesian modeling and enhancement of speech signals," *Speech and Audio Processing, IEEE Transactions on*, vol. 10, no. 3, pp. 173 –185, Mar. 2002.
- [9] 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.
- [10] —, "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.
- [11] S. Gannot, D. Burshtein, and E. Weinstein, "Iterative and sequential kalman filter-based speech enhancement algorithms," *Speech and Audio Processing, IEEE Transactions on*, vol. 6, no. 4, pp. 373 –385, jul. 1998.

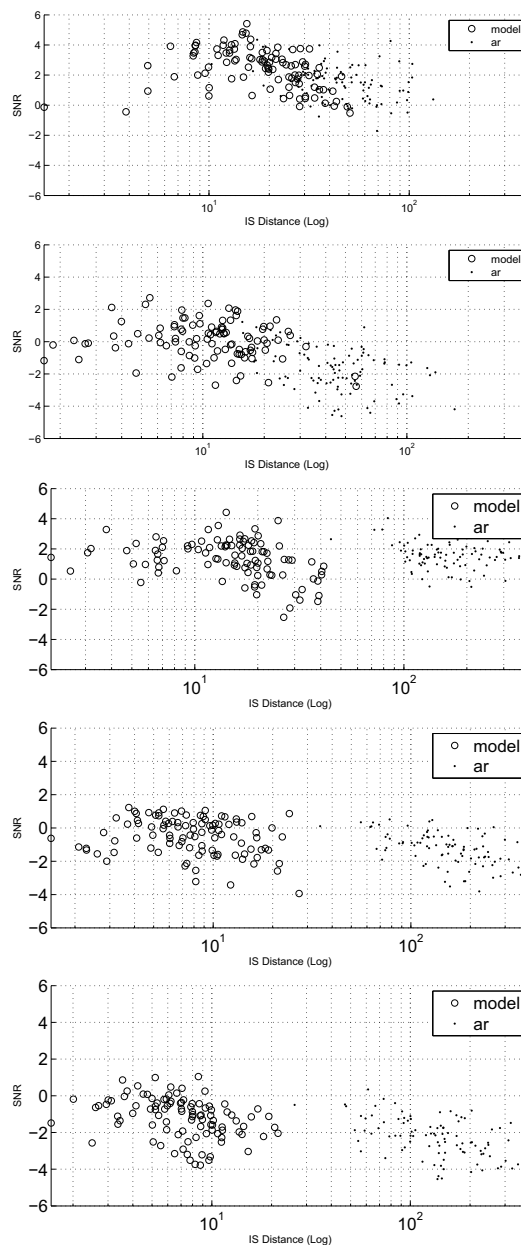


Figure 5: Results of the enhancement procedure applied to a signal with SNR=0 dB (upper plot), with SNR=-5, with SNR=-10, with SNR=-15, and with SNR=-20 dB (lower plot). Plots show the comparison between the proposed model vs a standard linear AR model.