JOINT ESTIMATION OF GLOTTAL SOURCE AND VOCAL TRACT FOR VOCAL SYNTHESIS USING KALMAN SMOOTHING AND EM ALGORITHM

*Pamornpol Jinachitra**

Center for Computer Research in Music and Acoustics (CCRMA)
Stanford University, USA
pj97@ccrma.stanford.edu

Julius O. Smith III

Center for Computer Research in Music and Acoustics (CCRMA)
Stanford University, USA
jos@ccrma.stanford.edu

ABSTRACT

In this paper, a joint parameter estimation of the derivative glottal source waveform and the vocal tract filter is presented where aspiration noise and observation noise are taken into account within a state-space model. The Rosenberg-Klatt glottal model is used in conjunction with an all-pole filter to model voice production. The EM algorithm is employed to iteratively estimate the model parameters in a maximum-likelihood sense, utilizing a Kalman smoother in the expectation step. The model and estimator allow for improved estimates of model parameters for resynthesis, yielding an output which sounds natural and remains flexible for modification, a desirable property for expressive vocal synthesis.

1. INTRODUCTION

A source-filter model for voice production has been studied and used in a number of speech synthesizers, for example, the KLSYN88 [1]. A good parametric model of a glottal source offers a parsimonious representation of the sound for efficient coding and flexible resynthesis, as well as providing features for many identification applications. While most speech synthesis research has provided us with good voice production models, as well as methods for accurate estimations of model parameters [2][3][4], few have considered automatic estimation of model parameters explicitly when noise is present. In the speech enhancement research arena, however, statistical techniques for parameter estimation and speech enhancement in noise are well known. Examples of this type of speech enhancement based on various voice models are [5][6][7]. Algorithms that consider uncertainty due to noise and a good model of voice production should allow for more accurate identification of parameters. Its potential applications are in voice synthesis and coding of a vocal sound recorded in real life where noise may be present or the recording equipment is not entirely noise-free.

In this paper, we focus on the parameter estimation aspect for a reconstruction of a voice from recording, possibly under moderately noisy conditions, rather than general denoising. The state-space model is used to model the voice production and the observations obtained at the microphone. An Expectation-Maximization (EM) algorithm has been used in speech enhancement applications before [5]. Here, we extend beyond most Kalman filtering-based enhancement algorithms by modeling the source input using a parametric form of Rosenberg-Klatt (RK) [1], instead of just white noise or a pulse train as used in [5]. The RK glottal pulse model is a simplified derivative version of the more general and popular Liljencrants-Fant (LF) model [8]. It fits well into the linear state-space formulation and captures the characteristics of three different modes of voice: breathy, normal and pressed. This voice production model and its motivation follow closely the work by Lu [2] where convex optimization is used to jointly estimate the vocal tract filter and the glottal source. Here, we instead employ a statistical model to include uncertainty due to observation noise. In the remainder of this paper, we describe the model employed and analysis/resynthesis in terms of it, followed by some results and discussion.

*Supported by Toyota InfoTechnology Center, US.

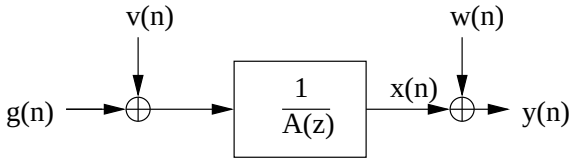

Figure 1: *Voice production model*

2. EM AND KALMAN SMOOTHING PARAMETER ESTIMATION

2.1. Model and Analysis

Voice production is modeled as a linearly separable source input cascaded with an all-pole vocal tract filter, as shown in Figure 1. The source consists of the derivative glottal waveform $g(n)$ summed with some aspiration noise $v(n)$. The lip radiation, modeled as a differentiator, has been folded into the glottal pulse waveform to give this derivative, assuming linearity. The model assumes there is no source and tract interaction or any form of non-linearity. It is also only applicable to non-nasal voice. The model for the derivative glottal waveform is the Rosenberg-Klatt model which can be expressed as follows:

$$g(n) = \begin{cases} 2a_g n/f_s - 2b_g(n/f_s)^2, & 0 \le n \le T_0 \cdot OQ \cdot f_s \\ 0, & T_0 \cdot OQ \cdot f_s \le n \le T_0 \cdot f_s \end{cases} \quad (1)$$

$$a_g = \frac{27 \cdot AV}{4 \cdot (OQ^2 \cdot T_0)} \quad (2)$$

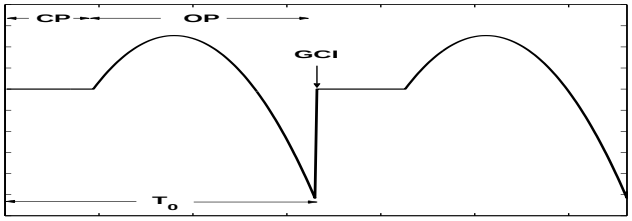

Figure 2: *Two periods of Rosenberg-Klatt derivative glottal waveform model showing the period (T_0), glottal closure instant (GCI), closed-phase (CP) and open-phase (OP)*

$$b_g = \frac{27 \cdot AV}{4 \cdot (OQ^3 \cdot T_0^2)} \tag{3}$$

where T_0 is the fundamental period, f_s is the sampling frequency, AV is the amplitude parameter, and OQ is the open-quotient of the glottal source. Note that spectral tilt is not explicitly modeled here, unlike in KLGLOTT88 [1] or [2]. An example of the waveform is shown in Figure 2.

We observe $y(n)$ at the microphone which is assumed to be the sum of the vocal sound and a stationary additive white Gaussian noise. Modeling the vocal tract as an all-pole filter of order P, the voice production model is similar to Lu's [2]. The system's state-space model representation is as follows:

State-Space Formulation of Lu's model

$$\mathbf{x}_m(n+1) = \mathbf{A}_m\mathbf{x}_m(n) + \mathbf{B}_m\mathbf{u}_m(n) + \mathbf{v}_m(n) \tag{4}$$

$$y_m(n) = \mathbf{C}\mathbf{x}_m(n) + w(n) \tag{5}$$

$$\mathbf{v}_m \sim \mathcal{N}(0, \mathbf{Q}_m), \qquad w \sim \mathcal{N}(0, R)$$

$$\mathbf{x}_m(n) = \begin{bmatrix} x_m(n) & x_m(n-1) & \cdots & x_m(n-P+1) \end{bmatrix}^T$$

$$\mathbf{A}_m = \begin{bmatrix} a_{1,m} & a_{2,m} & \cdots & & a_{P,m} \\ 1 & 0 & \cdots & & 0 \\ 0 & 1 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & & \\ 0 & \cdots & 0 & 1 & 0 \end{bmatrix}$$

$$\mathbf{B}_m = \begin{bmatrix} a_{g,m} & b_{g,m} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$$

$$\mathbf{u}_m(n) = \begin{cases} \begin{bmatrix} 2 \cdot (n - n_{c,m})/f_s \\ -3 \cdot ((n - n_{c,m})/f_s)^2 \end{bmatrix}, & n_{c,m} \le n \le N_m \\ \begin{bmatrix} 0 & 0 \end{bmatrix}^T, & \text{otherwise} \end{cases}$$

$$\mathbf{Q}_m = \begin{bmatrix} q_m & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$$

$$\mathbf{C} = \begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix}$$

where m is frame index, moving from one glottal period to another, and n is the sample index within a frame, ranging from 1 to N_m for frame m. We consider a period from one glottal closure instant (GCI) to the next so equation (1) is modified to have an integer offset n_c, which is the starting index of the glottal source open-phase. n_c therefore determines the open quotient (OQ) such that $OQ = (T_0 - n_c)/T_0$. The variances of process noise at frame m, v_m, and global observation noise w are q_m and R, respectively. Ideally, if the deterministic model of the voice production is accurate, $v_m(n)$ can be thought of as the aspiration noise, which is nothing more than a residual or model error here.

At each iteration, expectation of the likelihood followed by maximization (by finding ML estimates of the parameters) are performed. The E-step is achieved by computing the sufficient statistics of the posterior distribution, $P(X|Y, U, \theta)$ using Kalman smoothing, where X is the (hidden) clean speech, Y is the relevant set of observations, U is the input, and θ is the set of all model parameters. Let $\mathbf{D}_m = \begin{bmatrix} a_{1,m} & \cdots & a_{P,m} & a_{g,m} & b_{g,m} \end{bmatrix}^T$. During the M-step, the parameters are updated in turn as follows:

$$\mathbf{D}^{(new)} = \left[\textstyle\sum_{n=2}^{N_m} \mathbf{J}(n)\right]^{-1} \sum_{n=2}^{N_m} \begin{bmatrix} \mathbf{v}_1^1(n) \\ \hat{x}(n)\mathbf{u}(n) \end{bmatrix} \tag{6}$$

$$q^{(new)} = \frac{1}{N_m - 1} \sum_{n=2}^{N_m} \left(V_0^{(1,1)}(n) - 2\mathbf{D}^T \begin{bmatrix} \mathbf{V}_1(n) \\ \mathbf{u}(n)\hat{\mathbf{x}}^T(n-1) \end{bmatrix} + \mathbf{D}^T\mathbf{J}(n)\mathbf{D} \right) \tag{7}$$

$$R^{(new)} = \frac{1}{\sum_{m=1}^{M} N_m} \sum_{m=1}^{M} \sum_{n=1}^{N_m} \left[y_m^2(n) - \mathbf{C} \cdot \hat{\mathbf{x}}_m(n) y_m(n) \right] \tag{8}$$

where

$$\mathbf{J}(n) = \begin{bmatrix} \mathbf{V}_1(n) & \hat{\mathbf{x}}(n-1)\mathbf{u}^T(n) \\ \mathbf{u}(n)\hat{\mathbf{x}}^T(n-1) & \mathbf{u}(n)\mathbf{u}^T(n) \end{bmatrix} \tag{9}$$

The frame index is dropped when there is no confusion and all parameters on the right hand sides are current estimates. $\mathbf{V}_0(n)$ and $\mathbf{V}_1(n)$ are the covariances given by the Kalman smoother, $\mathbf{v}_1^1(n)$ is the first column of $\mathbf{V}_1^T(n)$ and $V_0^{(1,1)}(n)$ is the top-left-corner entry of $\mathbf{V}_0(n)$, i.e.,

$$\mathbf{V}_1(n) = \langle \mathbf{x}(n)\mathbf{x}^T(n-1) \rangle \tag{10}$$

$$\mathbf{V}_0(n) = \langle \mathbf{x}(n)\mathbf{x}^T(n) \rangle \tag{11}$$

where $\langle \cdot \rangle$ denotes the posterior averages from the Kalman smoother from the E-step. The iteration is repeated until convergence.

Unfortunately, the OQ-related parameter, n_c, is nonlinear with respect to the error minimized; we therefore need to do a grid search. Providing that its initialization value is close to the solution, only a few points in the vicinity of the current estimate are needed for calculation. However, it is important to say that for a given set of current estimates of other parameters at one iteration, even a grid search on n_c might not lead to an eventual globally optimal solution. A way to ensure optimality is to start by optimizing other parameters for all values of n_c and then pick the one that gives the highest likelihood at convergence. This method is,

however, computationally expensive. We therefore contend with trying to get good initialization and only make sure the likelihood function increases at each iteration doing grid search for the current ML estimates of other parameters. Keeping the likelihood function increasing, without actually maximizing it, still guarantees EM's monotonic convergence.

The M-step of AR filter coefficients estimation is equivalent to the covariance LPC method as a result of maximizing conditional likelihood [9]. The stability is therefore not guaranteed. Hence, we check for unstable poles at each iteration and reflect them back inside the unit circle. Instability, however, rarely happens unless, for example, the noise has very large spectral peaks. Another advantage in using EM is that constraining parameters can be done without sacrificing monotonic convergence behavior, as long as at each constraining, the likelihood is ensured to be increasing. We can therefore constrain $a \geq 0$ and $b \geq 0$, as well as the physical range of OQ and perform a stability check on the vocal tract filter.

At each iteration, the glottal closure instant (GCI) for the purpose of model fitting is also updated by searching for a minimum peak in the interested period of $g(n)$. While GCIs are assumed to be available, they need not be very accurate. The algorithm can refine all parameters, including the GCIs, to fit the observation.

After convergence, all parameters are smoothed using a narrow Hann window smoothing kernel, with AR coefficients converted to reflection coefficients before smoothing to preserve stability. Non-smoothed parameters otherwise cause audible artifacts. The resynthesis employs a lattice filter whose reflection coefficients change at each glottal closure instant.

2.2. Algorithm Initialization

Just like any other ML method, the EM algorithm has the risk of converging to a suboptimal local maximum. Therefore, good initialization is crucial for global convergence. Conventional LPC can be used for initialization of the AR coefficients under high SNR conditions. The OQ can be estimated from the following expression given by Fant [10]

$$H_1 - H_2 = -6 + 0.27 \cdot \exp(5.5\mathrm{OQ}) \qquad (12)$$

where H_1 and H_2 are the spectral amplitudes of the first and second harmonics. Alternatively, we can do exhaustive search for the best OQ in a few frames and use it as an initial value for adjacent frames. Figure 3 shows the prediction error surface which has an inverse relationship with the likelihood when all other parameters, except OQ, are at the true values for a synthetic input with $AV = 0.001$ and the AR coefficients are taken from LPC analysis of a frame of sound /a/. Where the range towards $OQ = 1$ is not shown in the top-left figure, the exponential trend continues. The figure shows that OQ should be over-initialized if in doubt. Given that physically from experimental studies [1], $0.4 \leq OQ \leq 1$, a grid search only in that range suffices to bring the likelihood up. Also note that the error curve is convex in that interval for these examples, implying a gradient method could be reliably used.

The presented ML estimation is not based on perception. Therefore, the internal LPC step will result in equal weighting across all frequencies, and high-frequency formants, which are important to perception, are not neccesarily modeled and could be missing. This is especially important in resynthesis since our derivative glottal source model has an inherent -6dB/octave roll-off. To remedy this, we obviously can do preemphasis before starting the iteration. However, this is not advisable when there is significant noise present. By doing so, observation noise is also no longer white, unless the noise can be assumed to have the same spectral roll-off characteristic. Alternatively, we can do estimation of the AR coefficients using more appropriate weighting, like warped frequency axis transform on the data, while making sure that the likelihood is still increasing. Also, some extrapolation or biasing will be needed to recover formants that are buried under noise. Incorporating these features within the presented iterative framework is the subject of further investigation.

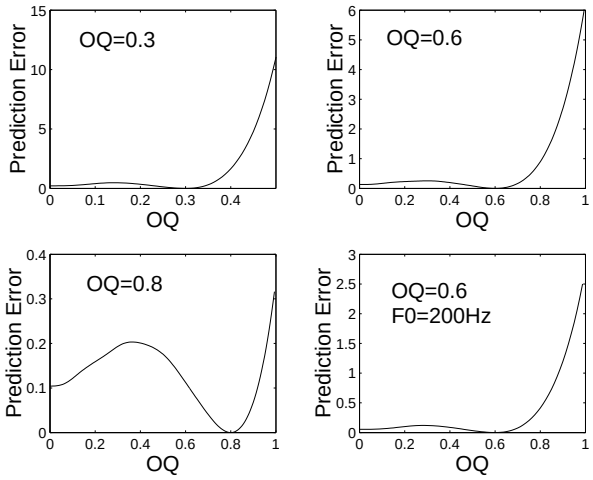

Figure 3: *Prediction error surface for different* OQ *values, all for F0=100Hz, except for the bottom right, where F0=200Hz*

3. RESULTS AND DISCUSSION

The algorithm has been applied to a male singing voice with vibrato and tremolo, singing /a/ at a fundamental frequency around 123 Hz. The sampling rate is 16 kHz. We first test on the clean signal applying preemphasis, using filter-order $P = 20$. The resynthesis sounds natural, and similar, though not exactly the same as the original. A slightly noisy sound, with white Gaussian noise added to give SNR=20dB, was also tried, using preemphasis and careful monitoring. The result is a similar sound that might sound less natural due to inaccuracy and smoothing of parameters. However, in contrast to general speech enhancement where the enhanced output is taken from state estimates given by Kalman filtering, it is completely free of musical noise. Wihtout preemphasis, the algorithm is more stable due to stable all-pole filter estimates. The output sound is still natural but does not sound as bright as the original. Using initial preemphasis does not seem to do as much harm when pink noise (roll-off -3dB/octave) is added instead at the same level of SNR=20dB and the resynthesized sound is excellent compared to the original. All results are available at *http://ccrma.stanford.edu/~pj97/WASPAA05/waspaa05_demo.html*

From listening, the sound generated using the vocal tract filter estimates after iteration sounds more accurate (closer to /a/) than using the initial noisy estimates. The median Itakura-Saito spectral distance between the all-pole filter coefficients obtained from clean signal and from iterations decreases significantly after a few iterations for all noisy cases above. Figure 4 shows that the deriva-

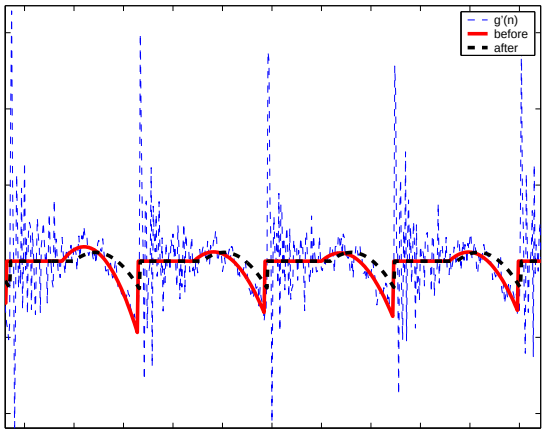

Figure 4: *The derivative glottal waveform, $g(n)$, the initial model (thick/dash) and the model after EM iterations (thick/solid)*

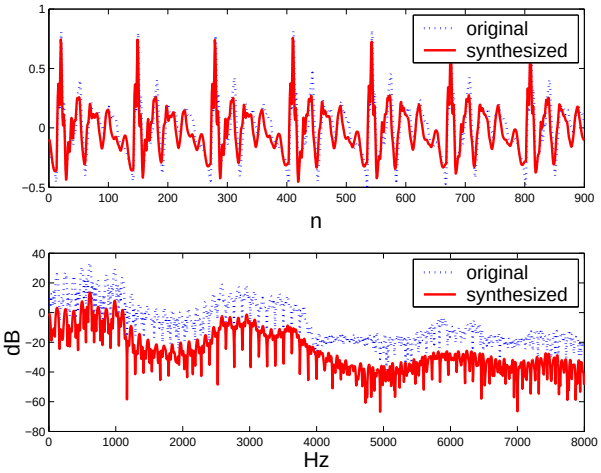

Figure 5: *The time-samples (top) and spectra (bottom) of the original (dots) and the synthesized signal (solid). The spectrum is offset for clarity.*

tive glottal waveform estimate is also closer to the original than initially. Figure 5 shows a close temporal and spectral comparison of the original and the resynthesized signal.

The resulting output may not sound exactly the same as the original, which is the price of using an economical parametric representation. What we gain, however, is the flexibility to modify the parameters to generate new sounds at will, which is valuable for speech synthesis and music applications. A breathy version of the original sound has been convincingly synthesized by increasing the OQ values and the variance of the synthesized aspiration noise. The generation of aspiration noise follows the model in [2]. It consists of a burst of noise right after the GCI, shaped by a Hann window, and additive background noise. One shortfall of our model is that, within a frame, the process noise variance estimate is attributed to both noises. To determine the proper variance, a post processing extraction is necessary. Alternatively, we can split the proportion of the estimated variance given by $\hat{q}_m$ between the two noises appropriately according to $\hat{q}_m \propto (1 - OQ) \cdot \sigma^2_{CP} + OQ \cdot \sigma^2_{OP}$, where σ^2_{CP} is the actual variance of the noise right after the GCI, mostly occupying the closed-phase, and σ^2_{OP} is that of the aspiration noise thereafter, mostly evident during the open-phase. Note that, however, $\hat{q}_m$ tends to be an over-estimation of the aspiration noise because of the inaccuracy of the deterministic model itself in modeling the sound. A switching state space model where the noise characteristic switches at some time instant could also be useful as a future improvement. A challenge also lies in female voice estimation due to generally more breathiness with possibly no closed-phase in the glottal waveform as well as the higher pitch which creates difficulty in the vocal tract filter estimation. An extension of the source model will be needed to deal with spectral zeros in nasal sounds and to capture small but still rather informative residue for a more faithful reconstruction.

4. CONCLUSION

In this paper, an iterative method to estimate voice production model parameters from a natural voice recording based on the EM algorithm and Kalman smoothing was presented. Under high SNR circumstances, the parameter estimates give a resynthesized sound that is natural while also allowing for modification and efficient coding.

5. REFERENCES

[1] D. H. Klatt and L. C. Klatt, "Analysis, synthesis and perception of voice quality variations among female and male talkers," *J.Acoust.Soc.Am.*, vol. 87, no. 2, pp. 820–857, 1990.

[2] H. L. Lu, "Towards a high quality singing synthesizer with vocal texture control," Ph.D. dissertation, Department of Electrical Engineering, Stanford University, 2001.

[3] M. Fröhilich, D. Michaelis, and H. W. Strube, "SIM-simultaneous inverse filtering and matching of a glottal flow model for acoustic speech signals," *J.Acousti.Soc.Am.*, vol. 110, no. 1, July 2001.

[4] P. R. Cook, "Identification of control parameters in an articulatory vocal tract model with applications to the synthesis of singing," Ph.D. dissertation, Department of Electrical Engineering, Stanford University, 1991.

[5] S. Gannot, D. Burshtein, and E. Weinstein, "Iterative and sequential Kalman filter-based speech enhancement algorithms," *IEEE Trans. on Speech and Audio Processing*, 1998.

[6] W. Ding and H. Kasuya, "A novel approach to the estimation of voice source and vocal tract parameters from speech signals," in *ICSLP*, 1996.

[7] K. Lee, B. Lee, I. Song, and S. Ann, "Robust estimation of AR parameters and its application for speech enhancement," in *ICASSP*, 1992.

[8] G. Fant, J. Liljencrants, and Q. Lin, "A four-parameter model of glottal flow," STL-QPSR, Tech. Rep., 1985.

[9] J. D. Markel and A. H. Gray, *Linear prediction of speech*. Springer-Verlag, 1976.

[10] G. Fant, "The voice source in connected speech," *Speech Communication*, no. 22, 1997.