GENERATIVE MODEL OF VOICE IN NOISE FOR STRUCTURED CODING APPLICATIONS

Pamornpol Jinachitra, Julius O. Smith

Center for Computer Research in Music and Acoustics
Stanford University
USA

ABSTRACT

A generative model of a human voice is presented, based on many pseudo-physical considerations. For robustness, observation noise is also included in the model. An EM-algorithm framework for inference and learning is then described. An instance of approximate inference and subsequent learning presented allows an extraction of voice parameter which can be used for structured coding application. This set of parameters allows a great amount of compression as well as the flexibility in making modification to pitch, duration and breathiness, noise-free synthesis compared to other non-parametric approaches.

Index Terms— Structured coding, parametric voice modeling, speech enhancement, generative model of voice

1. INTRODUCTION

The term structured audio coding has been used to refer to a way of describing a sound by its semantic model and parameters as opposed to direct information-theoretic compression of the sound signal [1]. The concept broadly covers many sound synthesis models from spectral modeling method such as the sinusoid model, to code-excited source-filter model for speech. However, the most benefit is captured, perhaps, when the model is physically intuitive and its parameters are easily compressed and easy to modify. The structured nature means we can often re-render the sound in different ways subject to applications at the receiver's end. Because of this flexibility, its applications include singing voice coding, expressive speech synthesis and intelligibility-enhanced speech coding.

In order for structured audio coding to work, besides having a good model, parameter extraction needs to be robust and accurate. In this paper, we consider the scenario of a human voice recording, from speech or singing, contaminated by generic colored noise. While many techniques applicable to clean voice have been shown in the past [2], having generic noise in observation voids procedures such as pre-emphasis or assumption of white Gaussian noise. Instead, a probabilistic framework is proposed and techniques to solve for the model parameters, based on Expectation-Maximization (EM) are presented. The EM framework allows many useful constraints, especially physically-motivated ones, to be applied during iteration while keeping the monotonic convergence property. Additionally, it provides a single framework for joint glottal segmentation and joint source-filter estimation of a voice sound. Previously, a similar idea of sound source modeling has been used for dehissing application of a string instrument [3] and an example of structured audio coding for a two-voice guitar has shown an excellent result in [4]. In this paper, we show a similar application where the voice parameters are extracted which can be used to resynthesize the sound noise-free. Comparing to filtering method of noise suppression [5] which usually suffers from musical noise or distortion, this approach has no such problem thanks to the resynthesis nature, at a price of missing fine details due to the use of a crude model. After the model and learning algorithm are presented, variations of applications, exploiting the structured nature of the sound, will be briefly demonstrated.

P. Jinachitra is supported by Toyota InfoTechnology Center, US.

2. VOICE PRODUCTION MODEL

Voice production is modeled as a linearly separable source input cascaded with an all-pole vocal tract filter, as shown in Equation (1).

$$x(n) = \sum_{p=1}^{P} a_p\, x(n-p) + g(n) + v(n) \tag{1}$$

P is the order of the autoregressive (AR) filter and v represents small modeling error, which may include aspiration noise. The main excitation source is the derivative glottal waveform, $g(n)$, which will be modeled by the Rosenberg's parametric model [6], expressed for one glottal period as follows:

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

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

where OQ is the open-quotient of the glottal pulse period, T_0 is the fundamental period, and AV represents the amplitude.

3. GENERATIVE MODEL OF VOICE IN NOISE

A generative model helps determine the relationship among different random variables. In this case, we need to estimate the glottal source parameters and the vocal tract filter coefficients. The model output is a clean voice that, when combined with noise, results in the observation. However, due to the glottal-synchronous segmental model shown previously, the glottal periods also need to be identified simultaneously.

Let $\mathbf{b} = \{b_0, b_1, ..., b_K\}$ represent a random variable indicating glottal period segmentation indices. Given the segmentation estimate, $\hat{\mathbf{b}}$, the state-space model of *each* glottal period with length N is represented by

$$\begin{aligned} \mathbf{x}_{n+1} &= \mathbf{A}_s \mathbf{x}_n + \mathbf{B}\mathbf{u}_n + \mathbf{v}_n \\ \mathbf{w}_{n+1} &= \mathbf{A}_n \mathbf{w}_n + \epsilon_n \\ y_n &= \mathbf{C}\mathbf{z}_n + r_n \end{aligned} \tag{4}$$

where

$$\mathbf{z}_n = \begin{bmatrix}\mathbf{x}_n^T & \mathbf{w}_n^T\end{bmatrix}^T \tag{5}$$

$$\mathbf{A}_s = \begin{bmatrix}\alpha_s^T \\ \mathbf{I}_{P_s-1} \quad \mathbf{0}\end{bmatrix}, \qquad \mathbf{C} = \begin{bmatrix}1 & \mathbf{0}_{P_s-1}^T & 1 & \mathbf{0}_{P_n-1}^T\end{bmatrix} \tag{6}$$

$$\mathbf{B} = \begin{bmatrix}a_g & b_g \\ \mathbf{0} & \mathbf{0}\end{bmatrix}, \mathbf{u}_n = \begin{cases}\begin{bmatrix}2\cdot(n-n_o)/f_s \\ -3\cdot((n-n_o)/f_s)^2\end{bmatrix}, & n_o \le n \le N \\ \begin{bmatrix}0 & 0\end{bmatrix}^T, & \text{otherwise}\end{cases} \tag{7}$$

$$\mathbf{v} \sim \mathcal{N}(0, \mathbf{Q}_s), \qquad \epsilon \sim \mathcal{N}(0, \mathbf{Q}_n), \qquad r \sim \mathcal{N}(0, R), \tag{8}$$

where n is the sample index within one glottal period. $\mathbf{A}_s$ contains AR coefficients of the vocal tract filter, α_s, on the top row and $\mathbf{A}_n$ contains those of the colored noise, α_n. The product $\mathbf{B}\mathbf{u}_n$ gives the glottal excitation, g_n, as shown in (2). $\mathbf{z}$ is a state variable consisting of length-P_s clean speech signal concatenated with length-P_n noise to be inferred. The noisy observation y_n is then the sum of the two instantaneous samples, added by small observation error, r_n. $\mathbf{Q}_s$ and $\mathbf{Q}_n$ only have a non-zero element q_s and q_n respectively at top-left. They are forced to have such form during estimation for stability. The variance R of r_n is fixed to a small number. We consider a period from one glottal closure instant (GCI) to the next so Equation (2) is modified to have an integer offset n_o, which is the starting index of the glottal source open-phase. n_o therefore determines the open quotient (OQ) such that $OQ = (T_0 - n_o)/T_0$. Parameters to be estimated are referred to collectively as $\theta = \{\alpha_s, a_g, b_g, n_o, \alpha_n, q_s, q_n\}$.

4. INFERENCE AND LEARNING

A framework of EM algorithm can be used for iterative inference and learning. The segmentation variable, $\mathbf{b}$, the clean speech and the colored noise, Z, are hidden variables which will be inferred during E-step. The model parameters, θ will be learned in the following M-step until convergence is reached. During E-step, we need to find the sufficient statistics of $p(Z, b|Y, \theta)$. Due to the combinatorial explosion of candidate points for segmentation, this inference is not tractable. In this work, we make approximation by taking the distribution of $\mathbf{b}$ to concentrate as a delta function at the MAP estimate (to be discussed in Section 4.3). Therefore, when $\mathbf{b}$ is marginalized,

$$\int_b p(Z, \mathbf{b}|Y, \theta) = \int_b p(Z|\mathbf{b}, Y, \theta)\cdot p(\mathbf{b}|Y, \theta) = p(Z|\hat{\mathbf{b}}_{MAP}, Y, \theta) \tag{9}$$

The sufficient statistics of the latter's can now be derived using Kalman smoothing and the model in (4) [7]. During M-step, a frame-by-frame maximum likelihood estimate of the state-space model parameters can be derived using the statistics from E-step. While standard expressions of these estimates can be found elsewhere, we present two methods, as an extension to [7], that encourage smoothness in parameters estimation which is very important for good synthesis. While most speech enhancement techniques which involve some smoothness constraints report better filtered speech results [8], the requirement for resynthesis is much more stringent since a few sample jitter and over smoothing can be heard very easily.

4.1. Penalized Maximum Likelihood

A penalized maximum likelihood adds a penalty term to the original likelihood expression to be maximized. In this case, we chose a Gaussian error between the estimate and some measure of parameters' mean derived from neighboring values. This can also be viewed as probabilistic prior model of the parameters. The mean of this prior is taken to be the half Hann window-weighted average of previous frames' estimates. Since averaging AR coefficients does not guarantee stability, it is done instead using line spectral frequencies (LSF). The covariance matrix of the LSF could be converted back to the AR domain using Unscented Transform to retain accuracy.

The penalized log-likelihood term pertaining to the voice parameters becomes

$$L(\theta_\mathbf{s}) \propto \frac{1}{q_s}\sum_{n=2}^{N}(x_n - \theta_s^T\mathbf{d}_n) + \lambda\cdot(\theta_\mathbf{s} - \bar{\theta_\mathbf{s}})^T\Sigma_{\theta_\mathbf{s}}^{-1}(\theta_\mathbf{s} - \bar{\theta_\mathbf{s}}) \tag{10}$$

where $\theta_\mathbf{s} = \begin{bmatrix}\alpha_s^T & a_g & b_g\end{bmatrix}^T$ and $\mathbf{d}_n = \begin{bmatrix}\mathbf{x}_{n-1}^T & \mathbf{u}_n^T\end{bmatrix}^T$. λ is the normalizing constant or penalty weight, including the frame length factor. The estimated mean and covariance of the prior are $\theta_\mathbf{s}$ and $\Sigma_{\theta_\mathbf{s}}$ respectively. The constrained estimate of $\theta_\mathbf{s}$ is then given by

$$\hat{\theta}_{\mathbf{s}PML} = \left[\mathbf{J} + \lambda\cdot q_s\cdot\Sigma_{\theta_\mathbf{s}}^{-1}\right]^{-1}\left[\mathbf{D} + \lambda\cdot q_s\cdot\Sigma_{\theta_\mathbf{s}}^{-1}\bar{\theta_\mathbf{s}}\right] \tag{11}$$

where

$$\mathbf{J} = \sum_{n=2}^{N}\begin{bmatrix}\mathbf{V}_0(n) & \hat{\mathbf{x}}(n-1)\mathbf{u}^T(n) \\ \mathbf{u}(n)\hat{\mathbf{x}}^T(n-1) & \mathbf{u}(n)\mathbf{u}^T(n)\end{bmatrix} \tag{12}$$

$$\mathbf{D} = \sum_{n=2}^{N}\begin{bmatrix}\mathbf{v}_1^1(n) \\ \hat{x}(n)\mathbf{u}(n)\end{bmatrix} \tag{13}$$

Both $\mathbf{J}$ and $\mathbf{D}$ can be derived from basic Kalman smoothing (see [7]). The contribution of the prior to the estimation is controlled by λ, q_s and the prior covariance. q_s decreases with iteration, meaning the contribution of prior is less and less, once the evidence is more reliable. From experiments, this constraint is found to help only during the first few iterations. During the last iterations, its contribution becomes smaller and the normal likelihood term dominates. This results in a more robust convergence, especially at low SNRs, but not smooth enough for re-synthesis purpose. Another smoothing mechanism is still needed as described next.

4.2. Post Kalman Smoothing

Assuming slowly varying parameters, a state-space model can be constructed as

$$\begin{aligned}\tilde{\theta}_{n+1} &= F\cdot\tilde{\theta}_n + e_n \\ \theta_n &= \tilde{\theta}_n + \eta_n\end{aligned} \tag{14}$$

where θ is the parameter's ML estimate, obtained from (11) or otherwise. $\tilde{\theta}$ is the smoothed estimate. Given the state-space model parameters, smoothed vocal parameters can be found using Kalman smoothing over a sequence of raw ML estimates.

An EM-algorithm can also be performed here to determined appropriate matrix F and the noise variances for each period where parameter dynamics can be assumed stationary. However, from experiments, a simple drift model, where F is an identity matrix, is enough and seems more robust. The process covariance determines the inertia or degree of smoothness in the inference estimates: the smaller, the smoother (strong prior belief), and is fixed in experiments. Kalman smoothing is applied on every 0.2-second segment

where the dynamics of the parameters are expected to be stationary. This is performed only on a_g, b_g and the AR coefficients. The latter, also requires a conversion to LSF, before performing smoothing to retain stability.

Physical waveshape constraints such as $a_g > 0$, $b_g > 0$ and $2a_g f_s/3b_g < N - n_o$, where N is glottal period length, are applied at every iteration. Filter stability is also checked and unstable poles get flipped inside the unit circle, although in all experiments, no instability has been encountered. Other types of constraints can also be applied, for example, a codebook constraint, where a codebook of highband LSF is searched for using Euclidean distance of the current lowband LSF estimates. However, we leave this as future work.

4.3. MAP Estimate of Boundary Variable

In this section, we discuss a possible probabilistic formulation of glottal segmental model and refer to two methods that fit the bill, as used in experiments. As mentioned earlier, the boundary variable is assumed to concentrate at the MAP solution. Assuming a Markovian relationship, the posterior can be expressed as

$$p(\mathbf{b}|X,\theta) \propto p(X|\mathbf{b},\theta)\cdot p(\mathbf{b}) = p(x_{b_0}^{b_1}) \prod_{k=2}^{K} p(x_{b_k}^{b_{k-1}}|x_{b_{k-1}}^{b_{k-2}},\theta)p(\mathbf{b}) \quad (15)$$

The conditional probability, $p(\mathbf{b}|X,\theta)$ is used to approximate the original posterior probability shown in Equation (9) where X represents clean speech estimates. Many forms of probability function can be used for $p(X|\mathbf{b},\theta)$. One possibility is a simple harmonic model and spectral voice template used in [9] to calculate each conditional probability and the initial probability respectively. Alternatively, due to conditioning on θ, we can also perform approximate inference on the derivative glottal waveform obtained from inverse-filtering the current estimate of clean speech by the current estimate of AR parameters. The algorithm in [10] can then be used to find MAP estimates of $\mathbf{b}$. Instead of using $p(X|\mathbf{b},\theta)$, we then use $p(G|\mathbf{b})$ where G represents the glottal derivative waveform. The prior, $p(\mathbf{b})$, can be chosen wisely to limit the number of initial candidates and bias more probable points. For example in [9], they are set to be uniform over zero-crossing points, whereas in [10], they are a combination of zero-crossings and a group-delay indicator.

5. EXPERIMENTS

The algorithms have been applied to a male modal singing voice (/aa/) corrupted by noise. The voice has vibrato and tremolo which can expose bad estimation very easily. Figure 1 shows a comparison of spectral estimates from a frame of singing voice recorded with pink noise mixed in for SNR=20 dB. The reference is calculated using close-phase covariance LPC on a pre-emphasized clean speech and a result from the proposed algorithm is shown in comparison to the original noisy sound's autocorrelation LPC. The main advantage of joint estimation of the source and the filter parameters can be seen in the spectral tilt compensation by the source model. While the proposed smoothness constraint does not significantly improve the estimate's spectral distance measure, compared to frame-independent iteration in [7], the synthesized sound is much more superior because of the natural smoothness. In this example, the segmentation performs almost perfectly in the first pass, except for a few cycles at the beginning and the end, so iteration on segmentation did not improve anything.

Once the parameters $\mathbf{b}, \alpha_s, a_g, b_g$ and n_o have been found, a resynthesis can be done according to Equation (1)-(2). The parameters above can also be converted to OQ, T_0 and AV for more intuitive coding and modification.

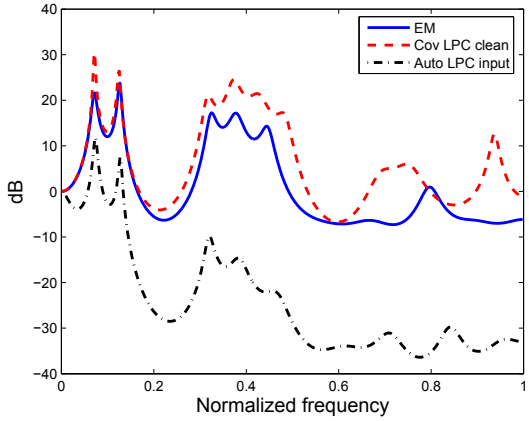

Fig. 1. Spectral envelopes of reference pre-emphasized clean speech close-phase covariance LPC, noisy autocorrelation LPC and the result of joint source-filter estimation from EM algorithm.

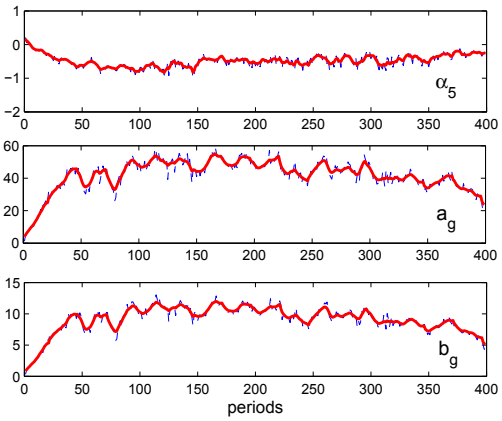

Fig. 2. Raw (dash) and smoothed (solid) estimates of α_5, a_g and b_g at the last iteration of EM and post Kalman smoothing.

Figure 3 shows examples of segmentation using methods in [9] and [10]. The former can be applied directly to the noisy sound. The latter must be applied to LPC residual of clean speech estimates only. Figure 4 shows the canonical parameters T_0, OQ and AV obtained from the algorithm with no smoothing in n_o.

5.1. Applications

Equipped with a physically intuitive set of parameters representing the original voice, many high-level applications are possible, as shown as examples in the list below.

Denoising: The denoising effect is achieved here by resynthesis. With good parameter estimation and good smoothing heuristic, we can get a noise-free reconstruction of the original voice. This is

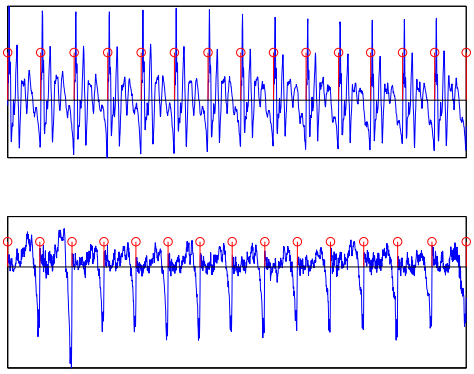

Fig. 3. Examples of glottal period segmentation using [9] on noisy speech and [10] on estimated LPC residual.

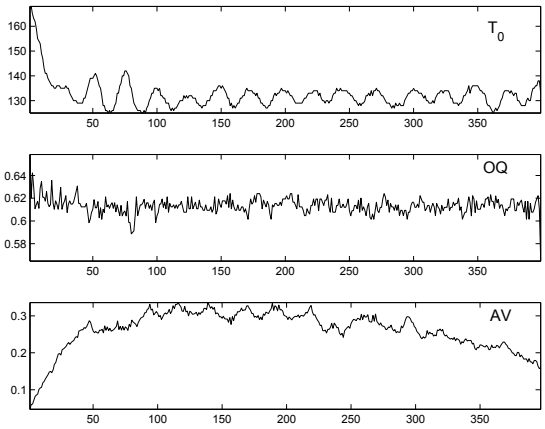

Fig. 4. Smoothed estimates of fundamental period (T_0), open-quotient (OQ) and amplitude (AV).

in contrast to a conventional filtering approach to noise suppression which often suffers from distortion (over suppression) or musical-noise artifacts (incorrect filter estimates). An informal listening test shows that half of the six test subjects prefer the reconstructed version over the filtered one. The rest still prefer the basic Kalman-filtered noise suppression output due to the remaining naturalness in the voice, versus the buzzyness that still manifests in our reconstruction. This seems to be a matter of preferences, however. The buzzyness in the resynthesis is especially revealing in spectral valley regions at high frequencies where the vocal tract filter is not well estimated. The parametric reconstruction receives the Mean Opinion Score (MOS) of 3.5 versus 3.3 for the filtered version while the noise-corrupted one gets 1.8. The clean voice reference is assigned the score of 5. Better models and estimates of the vocal tract filter in noise are still desirable as well as a large-scale listening test.

Voice Modification: Pitch modification is simply changing the value of T_0. To keep the duration the same, interpolation is required. Duration changing is also done easily by dropping or adding, with interpolation, the parameters. The modifications in parameters are notably more seamless than time-domain methods such as the PSOLA technique. In contrast to spectral modeling, our model allows the open-quotient to be modified independently and artificial noise can be added pitch-synchronously for different breathiness levels [2].

Bandwidth Extension: Upsampling the glottal source excitation is simple since none of the parameters are sampling rate dependent. However, the tract filter high-frequency characteristics have to be derived. There are many ways to do this, for example, a simple codebook table look-up of LSF coefficients, akin [11].

All sound samples can be found at *http://ccrma.stanford.edu/~pj97/icassp07_demo.html*.

6. CONCLUSIONS

A generative model has been shown for a parametric glottal source-filter voice production as observed in generic colored noise. An EM framework has been described and the parameters of the voice production are extracted after a few iterations of glottal segmentation and joint estimation. These parameters have been used to resynthesize the voice as demonstrated in a number of applications. While the idea of structured coding of a human voice is still in its infancy, the paper has demonstrated some encouraging results and a flexible estimation framework for more complete models in the future.

7. REFERENCES

[1] B. Vercoe, W. Gardner, and E. Scheirer, "Structured audio: Creation, transmission, and rendering of parametric sound representations," *Proc. IEEE*, vol. 86, no. 5, pp. 922–940, May 1998.

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

[3] P. A. A. Esquef, V. Välimäki, and M. Karjalainen, "Restoration and enhancement of solo guitar recordings based on sound source modeling," *Journal of the Audio Engineering Society*, vol. 50, no. 4, pp. 227–236, 2002.

[4] T. Tolonen, "Object-based sound source modeling for musical signals," in *AES 109th Convention*, September 2000.

[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] A. E. Rosenberg, "Effect of glottal pulse shape on the quality of natural vowels," *J. Acoust. Soc. of Am.*, vol. 49, no. 2, pp. 583–590, 1971.

[7] P. Jinachitra and J. O. Smith III, "Joint estimation of glottal source and vocal tract for vocal synthesis using kalman smoothing and EM algorithm," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, October 2005.

[8] J. H. Hansen and M. A. Clements, "Iterative speech enhancement with spectral constraints," in *ICASSP*, April 1987, vol. 12, pp. 189–192.

[9] K. Achan, S. Roweis, A. Hertzmann, and B. Frey, "A segment-based probabilistic generative model of speech," in *ICASSP*, 2005.

[10] P. Jinachitra, "Glottal closure and opening detection for flexible parametric voice coding," in *Interspeech*, September 2006.

[11] Y. Yoshida and M. Abe, "An algorithm to reconstruct wideband speech from narrow-band speech based on codebook mapping," in *ICSLP*, 1994.