A FRAMEWORK FOR PARAMETRIC SINGING VOICE ANALYSIS/SYNTHESIS

Youngmoo E. Kim

MIT Media Lab
20 Ames St., E15-492
Cambridge, MA 02139 USA
moo@media.mit.edu

ABSTRACT

The singing voice is the most variable and flexible of musical instruments. All voices are capable of producing the common phonemes necessary for language understanding and communication, yet each voice possesses distinctive qualities that are seemingly independent of phonemes and words. The unique acoustic qualities of an individual singer's voice arise from a combination of innate physical factors (e.g. vocal tract and vocal fold physiology) and time-varying characteristics of performance (e.g. pronunciation and musical expression). This research introduces a framework for singing voice analysis/synthesis that takes both physical and expressive factors into account by estimating source-filter voice model parameters (representing the physiology) and modeling the dynamic behavior of these features over time using a Hidden Markov Model (to represent aspects of expression). Historically, source and filter model features have been calculated independently, but here are estimated jointly to better model source-filter dependencies common in singing. Additionally, the vocal tract filter is estimated on a warped frequency scale, which more accurately reflects the frequency sensitivity of human perception. This framework has many possible applications, including singing voice analysis/synthesis and singer identification.

1. INTRODUCTION

The singing voice is acoustically one of the most complex instruments due to the flexibility of the vocal apparatus. The oscillatory opening and closing of the vocal folds (phonation) can occur over a wide range of fundamental frequencies. The vocal tract can be reconfigured into an infinite number of shapes through movement of the tongue, jaw, lips, etc., with each shape resulting in a different acoustic filter being applied to the vocal fold excitation. Sounds can also be formed without phonation (unvoiced phonemes), using turbulence caused by air being impeded by the vocal tract as a noise-like excitation. There are also mixed sounds that use both phonation and turbulence to create the overall sound.

The vast majority of singing, however, results from phonation ($>$90%). Furthermore, classically-trained singers are taught that vowel sounds in particular should be sustained as long as possible between consonants since they are the most efficient and audible sounds (consisting of several resonant peaks), which is especially important for being heard over other instruments. For this reason, we have chosen to focus primarily on the modeling of vowel sounds.

The framework presented in this paper draws upon a wide variety of work in voice-related research fields. The source-filter model, very effective for voice coding, has also been useful in establishing speaker (talker) and singer identity [1]. Glottal source features have proven to be important for talker identification [2]. Joint source-filter parameter estimation of the singing voice has been refined in [3]. And Hidden Markov Models have led to great advances in speech recognition research [4] and more recently have also been applied to systems for singing voice modeling as well [5].

This framework attempts to model both the physical and expressive factors that compose overall vocal quality and identity. A linear source-filter representation is used to model the physiology of the vocal folds and the vocal tract. In an extension of [3], parameters for both the source and filter models are jointly estimated for each oscillation period of the vocal folds using a warped frequency scale for linear prediction. The source-filter parameters are then modeled over time using a novel dynamic representation for the expressive characteristics of the voice in which a Hidden Markov Model (HMM) is trained for each primary vowel ([a], [e], [i], [o], and [u]) based on data from an individual singer.

2. OVERVIEW OF FRAMEWORK

The overall framework consists of the following key components: phonetic segmentation, glottal closure instant and period detection, source-filter parameter estimation, and HMM training. Figure 1 depicts the flow diagram linking each component of the framework. The following sections describe each component of the framework in detail.

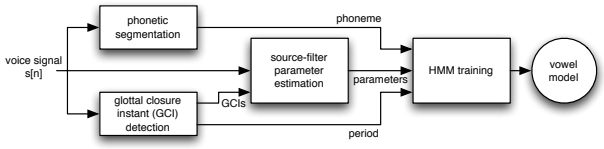

Figure 1: **Flow diagram of framework components**

2.1. Phonetic Segmentation

In order to automatically segment vowels from other phonemes, a straightforward approach is to train a segmentation system on recordings for which the individual vowel regions have been annotated. This type of phonetic transcription can be done by hand, but the task is quite tedious and error-prone. Fortunately, there exists an ample amount of accurate phonetically-segmented speech data used to train speech recognition systems. Transcribed speech data

from the TIMIT speech database was used to build average templates for each of 40 English phonemes. The templates were created by averaging the Mel-frequency cepstral coefficients (MFCCs) calculated from short-time (16 msec) frames of the segmented data. The Mel scale provides higher resolution at lower frequencies, roughly corresponding to the frequency sensitivity of the ear. MFCCs provide a general approximation of the spectral envelope and have been found to be good features for speech recognition systems.

MFCCs were calculated for each 16 msec frame of our sung training data, and the sum of squares distance to each phoneme template was calculated. By providing the order of phonemes, a phonetic transcription could be aligned to the singing data using a dynamic programming algorithm. Dynamic programming is an efficient technique for determining the lowest cost path traversing a distance matrix, given constraints such as the order of phonemes [6]. This technique was largely successful, although a small amount of hand editing was necessary to achieve phoneme alignment in some of the sung training data.

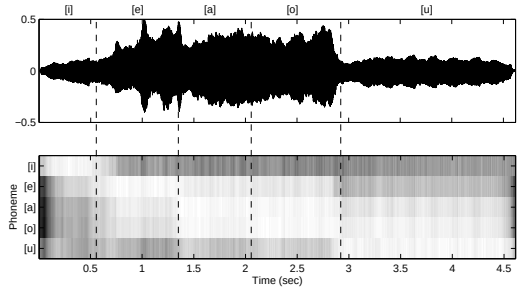

Figure 2: **Example of phonetic segmentation (lighter colors indicate shorter distances)**

2.2. Source Filter Parameter Estimation

The oscillation of the vocal folds is represented by a glottal excitation waveform, $g[n]$. Varying the rate of oscillation varies the pitch, and different modes of source articulation (e.g. breathy or pressed) will lead to different waveform shapes. The source waveform is then filtered by an appropriate vocal tract filter, $h[n]$ reflecting the current shape of the vocal tract. The convolution of the two produces the sung output, $s[n]$.

Traditionally, the calculation of parameters for the assumed source and filter models has been performed independently in order to reduce the complexity of the analysis. While this is a convenient approximation, we know that the movement of the vocal folds and shape of the vocal tract are not at all independent. In fact, singers frequently modify vowels (vocal tract shape) in order to more easily sing a high pitch. More importantly, these dependencies may be unique characteristics of an individual voice. Therefore, we will derive our model parameters jointly (simultaneously) from the acoustic data using a technique suggested in [3]. This technique models the vocal tract filter as an all-pole filter derived via linear prediction and uses the KLGLOTT88 model [7] for the glottal excitation. The technique is extended here by conducting the analysis on a warped frequency scale to more accurately reflect the frequency sensitivity of human perception.

2.2.1. Linear prediction and frequency warping

Linear prediction (LP) estimates a signal $s[n]$, using a linear combination of its p past samples. Here the signal of interest, $s[n]$, is the recorded singing voice. If we assume linear predictability, we obtain the following difference equation relating the glottal source $g[n]$ and the voice output.

$$s[n] = \sum_{k=1}^{p} \alpha_k s[n-k] + g[n] \tag{1}$$

From Equation (1), we derive the transfer function, $H(z)$ relating the voice output to the glottal source in the frequency domain:

$$H(z) = \frac{S(z)}{G(z)} = \frac{1}{1 - \sum_{k=1}^{p} \alpha_k z^{-k}} = \frac{1}{A(z)} \tag{2}$$

$H(z)$ is, of course, the z-transform of $h[n]$, defined as the vocal tract filter. Equation (2) shows that the transfer function is an all-pole filter. For convenience, we have defined the denominator polynomial separately as $A(z)$.

This analysis is continually performed over short contiguous windows of samples, resulting in a time-varying all-pole filter function of order p representing the vocal tract filter at a given time. The coefficients of $A(z)$ can be factored to determine the pole locations, which generally correspond to the vocal formants. Our data is sampled at 16 kHz, and we chose a value of $p = 16$.

One disadvantage of standard LP is that all frequencies are treated equally on a linear scale. The frequency sensitivity of the human ear, however, is close to logarithmic. As a result, standard LP analysis sometimes places poles at higher frequencies where the ear is less sensitive and misses closely spaced resonant peaks at lower frequencies where the ear is more sensitive. Using a higher order LPC is one way of compensating for this, but increasing the number of poles increases our feature dimensionality and makes it difficult to track correlations between analysis frames.

Instead, we use a warped Linear Prediction model by replacing each standard delay with an all-pass filter of the form:

$$z^{-1} \rightarrow D(z) = \frac{z^{-1} - \lambda}{1 - \lambda z^{-1}} \tag{3}$$

This has the effect of warping the power spectrum of each frame [8] and can be made to approximate the frequency sensitivity of the ear. A frequency ω is transformed to a warped frequency $\hat{\omega}$ via the following relation:

$$\hat{\omega} = \omega + 2\tan^{-1}\left(\frac{\lambda \sin\omega}{1 - \lambda\cos\omega}\right) \tag{4}$$

We use a value of $\lambda = 0.4$ in our analysis. While higher values of λ can be used to approximate the Bark scale, we found the higher warping more susceptible to instability in the filter estimates.

2.2.2. KLGLOTT88 model

The KLGLOTT88 model [7] is a relatively simple model for describing the derivative glottal wave. We choose to model the derivative glottal wave (rather than the glottal waveform itself) for two reasons: (1) to retain the simplicity of the model (a 2nd-order polynomial) and (2) to efficiently encapsulate the effects of lip radiation; Instead of differentiating the output, we equivalently apply the differentiation to the glottal wave in accordance with linear systems theory.

The KLGLOTT88 model $\hat{g}[n]$ for the derivative glottal wave is specified by the following equation:

$$\hat{g}[n] = \begin{cases} a2n - b3n^2, & 0 \le n < T \cdot OQ \\ 0, & T \cdot OQ \le n < T \end{cases} \quad (5)$$

T corresponds to the pitch period (in samples) and OQ is the open-quotient, or the fraction of the period for which the glottis is open. To maintain an appropriate waveshape, the parameters a and b are further related as follows:

$$a = b \cdot OQ \cdot T \quad (6)$$

The model has only two free parameters, a shape parameter a and the open-quotient OQ. Because of its relative simplicity, this model lends itself well to joint parameter estimation with the LP vocal tract model, as will be discussed below.

2.2.3. Joint parameter estimation

Simultaneous estimation of the KLGLOTT88 and all-pole filter parameters involves de-convolution of the source and filter functions from the recorded output.

$$G(z) = \frac{S(z)}{H(z)} = S(z)A(z) = S(z)\left(1 - \sum_{k=1}^{p} \alpha_k z^{-k}\right) \quad (7)$$

This gives a relation between the glottal source and linear combinations of the output, but the unit delays imply a linear frequency scale. To take advantage of frequency warping, we must replace each delay with the allpass filter $D(z)$ of Equation (3).

$$G(z) = S(z)\left(1 - \sum_{k=1}^{p} \alpha_k D(z)^k\right) \quad (8)$$

In the time domain, $D(z)^k$ represents a generalized shift operator, which is defined as a k-fold convolution of $\delta[n]$ with the signal the operator is applied to, where $\delta[n]$ is the impulse response of $D(z)$ [8].

$$\begin{aligned} d_1\{s[n]\} &\equiv \delta[n] * s[n] \\ d_2\{s[n]\} &\equiv \delta[n] * \delta[n] * s[n] \\ &\vdots \end{aligned} \quad (9)$$

Thus, in the time domain, we obtain the following relation between the glottal derivative and the recorded voice signal:

$$g[n] = s[n] - \sum_{k=1}^{p} \alpha_k d_k\{s[n]\} \quad (10)$$

We want to determine the parameter values that minimize the distance between the glottal derivative and our KLGLOTT88 model.

$$e[n] = \hat{g}[n] - g[n] = \begin{cases} 2an - 3bn^2 - s[n] + \sum_{k=1}^{p} \alpha_k d_k\{s[n]\}, & 0 \le n < T \cdot OQ \\ 0 - s[n] + \sum_{k=1}^{p} \alpha_k d_k\{s[n]\}, & T \cdot OQ \le n < T \end{cases} \quad (11)$$

This gives us the error at each sample, but we want to minimize the squared error over an entire period.

$$\min E = \min \sum_{n=0}^{T} (e[n])^2 \quad (12)$$

The relative simplicity of the glottal source waveform (a 2nd-order polynomial), guarantees that this minimization is a convex optimization problem and thus has a guaranteed optimal solution. As shown in [3], this optimization problem can be solved efficiently using quadratic programming [9]. The result is a simultaneous estimate of the excitation parameters a and b, and the warped LP filter coefficients (α_k) for each analysis frame. Minimizing Equation (11), however, assumes that T and OQ are known. T is known *a priori* from the glottal closure instant detection (described in the following section), but OQ is not known. Therefore, it is necessary to search for values of OQ that will minimize the overall error, E. The process is initialized by performing a linear search over all reasonable values of OQ (0.4 to 0.9). Since OQ will not vary greatly from one period to the next, we need only search a small range of values around the current OQ value for each successive frame. An example of parameter estimation from one period is shown in Figure 3.

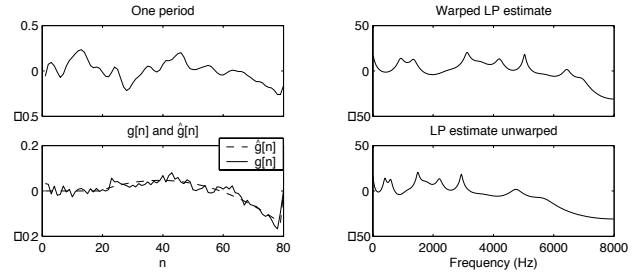

Figure 3: **Joint source-filter parameter estimation of vowel [e]. Top left is one period of the source waveform. Bottom left is the KLGLOTT88 model fit, $\hat{g}[n]$ and the inverse-filtered residual, $g[n]$. Right is the magnitude response of the jointly derived LP filter, plotted on warped (top) and linear (bottom) frequency scales.**

2.3. GCI detection

Source-filter parameter analysis is conducted pitch-synchronously, where each period corresponds to a single oscillation of the vocal folds. This requires each analysis frame to be time-aligned with the moment of closure of the vocal folds (the glottal closure instant, or GCI). In our framework, we select peaks of the residual signal formed by inverse filtering the voice signal with an LP filter calculated over large fixed-size frames (32 msec) using the autocorrelation method [6]. The peaks of the residual indicate the moments of least linear predictability, which correspond to the GCIs. We successively search for peaks within a small window surrounding an estimated pitch to find the GCIs, which demarcate the beginning and end of each pitch period for the source-filter parameter estimation step.

2.4. HMM Model Training

A Hidden Markov Model (HMM) assumes the process being modeled is represented by a fixed number of states and that at any given time the process will correspond to one of the states. These states do not necessarily (and most often don't) correspond to a tangible or easily described aspect of the modeled process, hence the "hidden" nature of the states. For each state in the model probabilities are calculated for remaining in the current state and transitioning from the current state to some or all of the other states. Training is performed using the Baum-Welch (Expectation-Maximization)

algorithm [4], which iteratively optimizes the state attributes and transition probabilities given a set or sets of training observations.

We train individual HMMs for each vowel ([a], [e], [i], [o], [u]) and each singer in our data set. Our observations consist of the following features from the previous sections: The line spectrum frequencies [6] (calculated from the filter coefficients α_k) representing the vocal tract filter, the parameters a and OQ for the glottal derivative wave, the period T, and the energy of each analysis frame. Since the observations are taken every pitch period, each state will correspond to one period and transitions will occur at this rate. The phonetic segmentation from Section 2.1 allows us to accurately segment voice data for each vowel model.

The framework currently uses 10 states for each vowel HMM, and each state is allowed to transition to any other state. The number of states was determined after some experimentation as a compromise between the divergent goals of encompassing the wide variability of each vowel's observed features and limiting the computational complexity of the model. Once the model is trained, we can calculate the state path of a new sequence of observations via the Viterbi algorithm [4], where each state is determined by the previous state, the transition probabilities, and latest vector of observed features. An example state path is shown in Figure 4. A

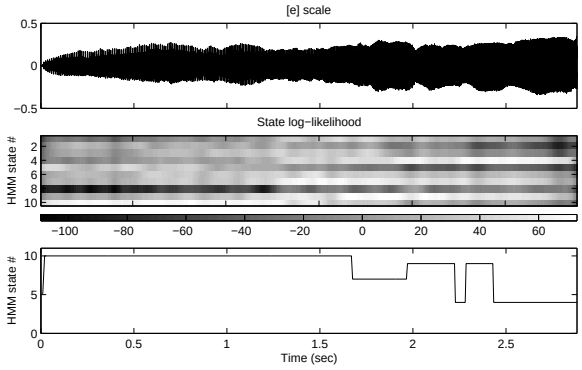

Figure 4: **HMM state path for a scale passage on the vowel [e].**

model that corresponds well to the observed process will have a higher overall likelihood than a poorly matching model, as calculated using the forward-backward algorithm [4], revealing a metric for comparing the relative performance of different HMMs.

3. PRELIMINARY APPLICATIONS

One application for this framework is for low-bitrate singing voice coding. Previous work has demonstrated the compression advantage of encoding vowels using pre-stored (static) templates [10]. Initial results with the current system demonstrate a much higher quality re-synthesis using the 10-state vowel models than with static vowel templates. Disregarding the overhead of transmitting the vowel models, each period can be represented using just a vowel identifier (3 bits), a state number (4 bits), the amplitude (6-16 bits), and the pitch period (8 bits). Given an average pitch period of 70 samples (at a sampling rate of 16 kHz) with 16-bit samples, we could achieve a possible compression of $>$30:1 during vowel segments. The framework could also be extended to include models for all voiced phonemes. Of course, significant modifications in the parameter estimation step would be needed to accommodate unvoiced phonemes.

Another application for this framework is singer identification. Our current data set consists of recordings from 4 conservatory-trained classical singers. Each singer performed a variety of vocal exercises (such as scales and arpeggios) emphasizing the 5 major vowels in addition to one entire piece. The exercise passages are used to train the vowel HMMs. Excerpts from the actual pieces are then segmented phonetically, and vowel segments are analyzed and evaluated against the trained HMMs. The identification system operates by determining which singer's HMM has the highest likelihood for a given vowel, collecting results over all of the vowels in the entire passage. The singer with the greatest number of vowel HMM matches is determined to be the singer of the excerpt. The system obviously requires a great deal of training data in order to build accurate HMMs for each vowel. Since the parameter estimation relies on very clean solo voice recordings, it is difficult to build models for many singers. On our very limited data set, preliminary results indicate a high degree of accuracy ($>$90%) in identifying the correct singer. Sound examples and updated experimental results are available via [11].

4. REFERENCES

[1] Y. E. Kim, "Singer identification in popular music recordings using voice coding features," in *Proc. International Symposium on Music Information Retrieval*, Paris, 2002.

[2] M. D. Plumpe, T. F. Quatieri, and D. A. Reynolds, "Modeling of the glottal flow derivative waveform with application to speaker identification," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 5, pp. 569–586, Sept. 1999.

[3] H.-L. Lu, "Toward a high-quality singing synthesizer with vocal texture control," Ph.D. dissertation, Stanford University, 2002.

[4] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, February 1989.

[5] Y. Meron, "High quality singing synthesis using the selection-based synthesis scheme," Ph.D. dissertation, University of Tokyo, 1999.

[6] D. O'Shaughnessy, *Speech Communication*. Addison-Wesley, 1987.

[7] D. H. Klatt and L. C. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *JASA*, vol. 82, no. 3, pp. 737–793, 1990.

[8] A. Härmä, "A comparison of warped and conventional linear predictive coding," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 5, pp. 579–588, 2001.

[9] R. Fletcher, *Practical Methods of Optimization, 2nd ed.* John Wiley & Sons, May 2000.

[10] Y. E. Kim, "Structured encoding of the singing voice using prior knowledge of the musical score," *Proc. IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics*, pp. 47–50, 1999.

[11] ——, "Singing voice analysis/synthesis," Ph.D. dissertation, Massachusetts Institute of Technology, 2003. [Online]. Available: http://sound.media.mit.edu/ moo/thesis