Proceedings of the Stockholm Music Acoustics Conference, August 6-9, 2003 (SMAC 03), Stockholm, Sweden

SINGER IDENTIFICATION AND TRANSFORMATION THROUGH DYNAMIC MODELING OF VOCAL FOLD AND VOCAL TRACT PARAMETERS

Youngmoo E. Kim

MIT Media Lab
Cambridge, MA USA
moo@media.mit.edu

ABSTRACT

Oftentimes when we listen to a familiar singer, the unique qualities of that performer's voice allow us to establish the singer's identity with relative ease. It is believed that 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 individual characteristics of performance and expression (e.g. pronunciation and accent). In this research, we jointly estimate pole-zero filter and LF glottal waveform model parameters to model the shape of the vocal tract and the glottal excitation, respectively over short time periods. These time-varying parameters, corresponding to the physical characteristics of the singer, are used to train a Hidden Markov Model (HMM). The HMM is used to model the dynamic behavior of the source-filter parameters, corresponding to some of the expressive characteristics of the singer. We propose a system that is able to identify singers based upon the model of greatest likelihood among the individually trained HMMs. We also explore the use of individual HMM states as a method of mapping from the parameters of one singer to another in a preliminary attempt at singing voice transformation. The data used in this analysis was recorded from four conservatory-level classically trained singers.

1. INTRODUCTION

Individual voices tend to be easily distinguished and thus reflect the identity of an individual. We are undeniably sensitive to the unique qualities of voices and can perceive those qualities after listening to a voice for just a short time. This is because the vocal apparatus, while being extremely complex and flexible, is also highly self-consistent. The distinctive properties of the voice are believed to be a combination of physiological factors (e.g. vocal fold stiffness and vocal tract size) and expressive factors (e.g. pronunciation and accent). In this research, we attempt to model both types of factors in order to establish and simulate singer identity.

A good deal of research has been performed on features for speaker (talker) identification. Much of this work has focused on spectral features used in speech recognition systems, such as Mel-frequency cepstral coefficients (MFCCs), which correlate to the shape of the vocal tract [1]. Some research has also investigated glottal excitation features and their potential usefulness in the determination of speaker identity. Both approaches have proven to be moderately successful, and performance improves when they are used together [2].

The system presented in this paper is specific to the classically-trained singing voice and takes advantage of certain assumptions that discriminate classical singing from speaking. For example, in classical singing there is a much higher degree of voicing and longer vowel durations than in speech. The ultimate goal of this research is not only the identification of a particular singer's voice, but also the parameterization of features unique to that voice. Our system also has possible applications in singing voice transmission (coding), evaluating voice similarity, and singing voice synthesis.

2. OVERVIEW OF SYSTEM

This section contains a summary of the major components of our analysis system. Figure 1 shows a block diagram of the framework.

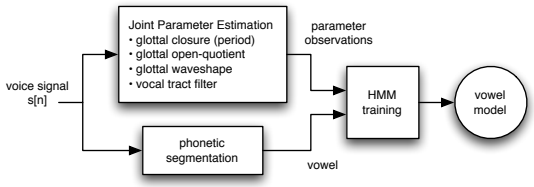

Figure 1: **Flow diagram of framework components**

2.1. Source-Filter Parameter Estimation

The source-filter model reflects the physical mechanism of vocal production and is commonly used for voice analysis/synthesis applications. For the source component, rather than modeling the glottal flow directly it is common to model the glottal derivative wave, which conveniently accounts for the effect of lip radiation (a differentiation). In our representation, the glottal derivative wave, $g[n]$ is filtered by the vocal tract impulse response, $h[n]$, to produce the voice ouput, $s[n]$.

$$s[n] = g[n] * h[n] \tag{1}$$

Historically, the estimation of source and filter parameters has been performed independently in order to simplify the analysis. Though convenient, this approximation is also inaccurate since vocal fold oscillation and vocal tract shape are actually somewhat dependent. For example, singers frequently modify their vowels (vocal tract shape) in order to more easily sing a high pitch. These dependencies may also be distinctive features of an individual voice. To account for these variables, we derive our source-filter model parameters jointly (simultaneously) from the acoustic data.

In our analysis, we initially use the KLGLOTT88 model [3] to represent the glottal derivative source, and a fixed-order filter

(derived via linear prediction) to model the vocal tract filter. These models lend themselves to a particularly efficient solution for joint parameter estimation, via convex optimization. Using the jointly-derived filter estimates, the excitation is then re-parameterized using the more complex LF model, which more accurately reflects the waveshape of the glottal derivative. Effects not represented by these models (such as turbulence) result in a residual noise signal, which is modeled seperately. The following estimation procedure was first suggested in [4] and a brief summary follows.

The KLGLOTT88 glottal derivative model is defined as follows:

$$\hat{g}[n] = \begin{cases} 2an - 3bn^2, & 0 \leq n < T \cdot OQ \\ 0, & T \cdot OQ \leq n < T \end{cases} \quad (2)$$

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. The parameters a and b are further related as follows:

$$a = b \cdot OQ \cdot T \quad (3)$$

Simultaneous estimation of the glottal derivative and vocal tract parameters involves de-convolution of the source and filter functions from the voice signal. To accomplish this, we attempt to minimize the distance between the KLGLOTT88 source model, $\hat{g}[n]$ and linear prediction residual, $g[n]$, over one period given values of the period T and open-quotient OQ. This error function is convex and therefore has a guaranteed optimal solution. The solution is calculated using quadratic programming, resulting in simultaneous estimates for the KLGLOTT88 parameters and polynomial filter coefficients of the LP filter.

Our representation differs from [4] in the following ways: 1) Joint parameter estimation is performed on a warped frequency scale, to more accurately model the frequency sensitivity of human perception, 2) Glottal closure instants are not calculated *a priori*, but are optimized from the data given the assumed models for source and filter, and 3) The residual noise is modeled using a stochastic codebook, individually trained for each singer. These extensions are described in greater detail in the sections that follow.

2.1.1. Parameter Estimation with Warped Linear Prediction

Standard *Linear Prediction* (LP) estimates a signal from a linear combination of previous samples [5].

$$s[n] = \sum_{k=1}^{p} \alpha_k s[n-k] + g[n] \quad (4)$$

From the source-filter relation (Eq. 1), we can derive the transfer function, $H(z)$, which is an all-pole filter.

$$H(z) = \frac{S(z)}{G(z)} = \frac{1}{1 - \sum_{k=1}^{p} \alpha_k z^{-k}} = \frac{1}{A(z)} \quad (5)$$

A disadvantage of standard LP is that all frequencies are treated equally on a linear scale while the frequency sensitivity of the human ear is closer to logarithmic. As a result, LP analysis sometimes places additional poles at higher frequencies while neglecting closely spaced formants at lower frequencies where the ear is more sensitive. Instead, we use *Warped Linear Prediction* (WLP) to nonlinearly warp the spectrum of a signal to increase resolution at lower frequencies. This can be accomplished by replacing each standard delay with the following all-pass filter [6].

$$z^{-1} \rightarrow D(z) = \frac{z^{-1} - \lambda}{1 - \lambda z^{-1}} \quad (6)$$

We use a parameter value of λ=0.4, which provides a moderate degree of frequency warping. Higher values were found to cause more instability in the filter estimates.

To use WLP in the joint source-filter parameter estimation, we must reformulate the LP residual as a WLP residual. From Equation (5), we obtain the z-transform of the LP residual, $G(z)$:

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

To take advantage of frequency warping, we must replace each delay with the allpass filter $D(z)$ of Equation (6).

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

In the time domain, $D(z)^k$ is the k-fold convolution of $\delta[n]$ (the impulse response of $D(z)$) with the original non-delayed signal. We denote this using the generalized shift operator $d_k\{\cdot\}$. [6].

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

Thus, we obtain the following relation between the WLP residual and the 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 $g[n]$ and our KLGLOTT88 model, $\hat{g}[n]$:

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

This gives us the error at each sample, but we want to minimize the L_2-norm (squared error) over an entire period.

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

As in [4], this modified constrained optimization problem can be solved efficiently using quadratic programming. 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. An example of parameter estimation from one period is shown in Figure 2. Minimizing Equation (11), however, assumes that T and OQ are known. The following section (2.1.2) describes the estimation of these parameters.

The warped analysis results in an all-pole filter in the warped frequency domain, but when transformed to the linear frequency

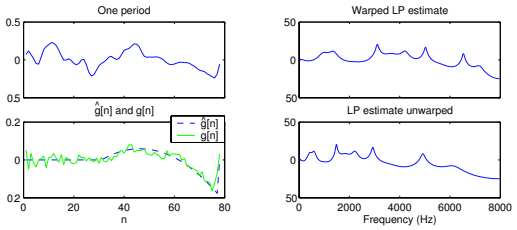

Figure 2: **Joint source-filter parameter estimation of vowel [e]**

domain it is actually a pole-zero filter. Fortunately, the analysis is conducted entirely in the warped domain, which maintains the simplicity of the all-pole representation (requiring fewer coefficients).

Once the filter parameters α_k are estimated, the WLP residual $g[n]$ can then be fitted to the more sophisticated LF model, $\tilde{g}[n]$, which more accurately describes the shape of the wave.

$$\tilde{g}[n] = \begin{cases} E_o e^{\alpha n T_s} \sin(\omega_g n T_s), & 0 \le nT_s < T_e \\ -\frac{E_o}{\epsilon T_a}\left[e^{-\epsilon(nT_s - T_e)} - e^{-\epsilon(T_c - T_e)}\right], & T_e \le nT_s < T_c \\ 0, & T_c \le nT_s < T_o \end{cases} \quad (13)$$

$E_o, \alpha, \omega_g, \epsilon, T_a, T_e$, and T_c are all free parameters, though there are dependencies between some of them. T_s is the sampling period and T_o is the pitch period. The parameters are estimated using constrained nonlinear minimization, which is described in detail in [4].

2.1.2. Glottal Closure Instant and Period Detection

The parameter estimation from the previous section is pitch-synchronous, where each analysis frame is time-aligned with each period of the output waveform. This requires the estimation of period boundaries which are aligned to the glottal closure instants (GCIs). Several techniques have been proposed for GCI detection (e.g. [7]), but each technique has some deficiencies which can result in GCI errors, leading to poor model fitting and poor source-filter parameter estimates. Instead of attempting to calculate the GCIs *a priori*, we perform a search for the period resulting in the best model fit.

Likewise, it is necessary to search for the appropriate value of the open quotient, OQ. Therefore, we simultaneously search for the optimal values of T and OQ that will minimize the overall error (Eq. 12). We obtain an initial estimate of the first several GCIs from a fixed-frame size LP residual. The residual has large peaks at moments of least linear predictability, which are usually close to the instants of glottal closure. The search is initialized by performing a linear search over all reasonable values of OQ (0.4 to 0.9) and values of T close to (within a few samples) the initial GCI estimate. Since neither T nor OQ will not vary greatly from one period to the next, we need only search a small range of values around the current values for each successive frame.

2.1.3. Stochastic Component Estimation

There are several stochastic sources that contribute to the overall vocal output, such as glottal aspiration noise and air turbulence in the vocal tract. These other noise-like sources prevent a perfect match to the vocal wave using the LF and WLP parameterization. Glottal aspiration noise is strongly correlated to the instants of glottal opening and closure, and certain vocal tract shapes are also more susceptible to air turbulence. Thus, the stochastic components must be modeled in a very specific way to reflect these dependencies. Glottal aspiration noise has been previously modeled statistically using wavelet de-noising [4].

We use a stochastic codebook approach, where the codebook is determined via *Principle Components Analysis* (PCA) of the glottal derivative residuals. PCA involves calculating the eigenvectors and eigenvalues from a matrix of input vectors. The eigenvectors corresponding to the highest eigenvalues capture the greatest amount of variance, statistically, of the input data. We want to find the eigenvectors of our set of residual noise vectors, $r[n]$, where

$$r[n] = \tilde{g}[n] - g[n] \quad (14)$$

PCA requires that all input vectors be the same length for statistical analysis, but in our case the length of each residual noise vector $r[n]$ varies according to the period. So we transform each residual to the frequency domain using equal-sized FFTs of N points to obtain:

$$R[\omega_k] = \mathcal{F}\{r[n]\}, \text{where } \omega_k = \frac{2\pi k}{N}, k = 0, \ldots, N-1 \quad (15)$$

Since $r[n]$ is real, we need only the first $\frac{N}{2} + 1$ values of $R[\omega_k]$, and PCA is then performed on these vectors. The analysis is performed on the complex FFT values in order to preserve the phase information, which is crucial to our noise model. From this, we obtain $\frac{N}{2}$ eigenvectors, which comprise the codebook.

For each noise vector, $\mathbf{r} = r[n]$, we obtain a weighting vector $\mathbf{w}$ corresponding to the contribution of each of the eigenvectors in codebook $\mathbf{C}$:

$$\mathbf{w} = \mathbf{C}\mathbf{r} \quad (16)$$

The n_c highest weighted codebook vectors, corresponding to the highest n_c values of $\mathbf{w}$ are then used to estimate $r[n]$ for each period.

2.2. Phonetic Segmentation

The analysis framework used in the singer identification system operates on individual vowel segments, requiring the input to be phonetically segmented. Using labeled data from the TIMIT speech database, simple templates were built for each of the 42 English phonemes. The templates were created by averaging the MFCCs calculated from short-time frames of the labeled data. The phonetic segmentation system assumes that a phonetic transcript of the singing input is available *a priori*, a restriction currently necessary for accurate segmentation. The system calculates MFCCs for each short-time frame of input singing data and determines the L_2 norm from each MFCC frame to each phoneme template. The phonemes are aligned to the singing data using dynamic programming to find the shortest-distance path over the matrix of MFCC-to-phoneme template distances, given the constraint of phoneme order.

2.3. HMM Model Training

We trained individual HMMs for each vowel ([a], [e], [i], [o], [u]) and each singer in our data set. Our observations consisted of the source-filter features defined in the previous sections. Each HMM state corresponds to one period, and transitions occur at each period boundary. The identification system operates by determining which singer's HMM has the highest likelihood for a given vowel, as established by the phonetic segmentation. The system currently

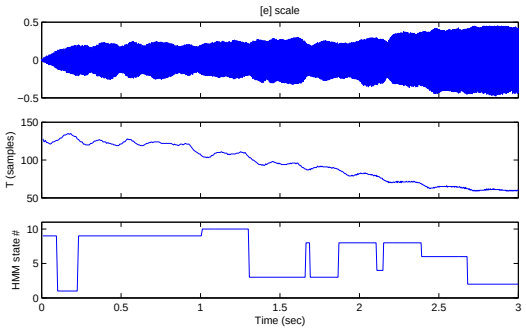

Figure 3: **HMM state path for a scale passage on the vowel [e].**

uses 10 states for each HMM. This value was determined experimentally to cover the wide variability of each vowel's observed features while limiting the computational complexity of the model. An example state path is shown in Figure 3.

3. SINGER IDENTIFICATION

The current data set consists of recordings from 4 conservatory-trained classical singers (two sopranos, one tenor, and one bass-baritone). Each singer performed a variety of vocal exercises (such as scales and arpeggios) emphasizing the 5 major vowels in addition to one entire piece from the classical repertoire. The exercises were segmented by vowel, and each vowel segment was used to train the vowel HMMs specific to each singer.

Vowel segments were extracted from the actual pieces and evaluated against the trained HMMs. The identification system operates by determining which singer's HMM has the highest likelihood for a given vowel, and the singer with the most vowel HMM matches in the excerpt is identified as the source performer. Five 5-7 second excerpts from each piece were used from each singer. On this admittedly small data set, the identification system performed with an overall accuracy of $>$90% when operating over entire excerpts and $\sim$70% over the individual vowel segments. More thorough results are presented in [8].

4. VOICE TRANSFORMATION

The system described in this paper has also been used for analysis/synthesis, where the HMM state paths are used to reconstruct the voice output. Because the various parameters for each pitch period can be represented by a single state value, a great deal of compression can be achieved. The sound quality, however, can be inconsistent, depending on how well the HMM states represent the input signal. Informal listening has shown that the sound quality over the vowel segments is mostly preserved. Higher sound quality is likely to be achieved with larger amounts of training data and a higher number of HMM states per model.

The HMM states provide a common point of reference between differing voice models, allowing us to map parameters from one to another. The states themselves represent clusters of significant concentration in each singer's parameter space. Since there is no consistency in the labeling of states between singers, we reorder each HMM's state labels according to their frequency of occurrence (state 1 becomes the most often occurring state, followed by state 2, etc.), enforcing a semi-statistical relationship between the state numbers (Figure 4). We then used the revised state path from one singer to drive another singer's HMM for re-synthesis. Informal listening does demonstrate a definite transformation in voice quality, though the accuracy of the effect is difficult to quantify. This is a simple mapping and we are currently investigating alternative methods for mapping between the states of different singers.

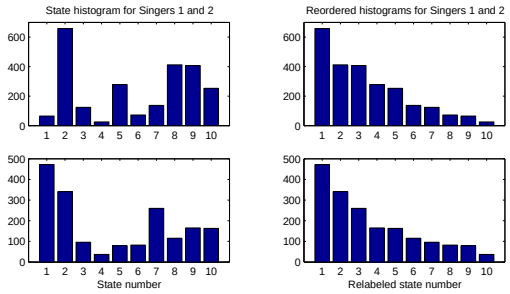

Figure 4: **HMM state histogram for two singers for vowel [e].**

We did not utilize the stochastic codebooks in the these preliminary voice transformation experiments. There is no simple way to map from one codebook to another. An initial attempt at simply applying the code vector weights calculated from one singer's codebook to another resulted in a more distorted sound, similar to phase distortion. A different parameterization may be needed to provide a more suitable mapping of the stochastic component of the excitation.

5. REFERENCES

[1] R. Mammone, X. Zhang, and R. P. Ramachandran, "Robust speaker recognition: A feature-based approach," *IEEE Signal Processing Magazine*, vol. 13, no. 5, pp. 58–71, 1996.

[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] 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.

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

[5] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. Englewood Cliffs, NJ: Prentice-Hall, 1978.

[6] 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.

[7] R. Smits and B. Yegnanayarana, "Determination of instants of significant excitation in speech using group delay function," *IEEE Transactions on Speech and Audio Processing*, vol. 3, pp. 325–333, 1995.

[8] Y. E. Kim, "Singing voice identification via dynamic parameter modeling," *submitted to the 2003 International Symposium on Music Information Retrieval*, October 2003.