SPECTRAL MODIFICATION FOR DIGITAL SINGING VOICE SYNTHESIS USING ASYMMETRIC GENERALIZED GAUSSIANS

Matthew E. Lee and Mark J. T. Smith

Georgia Institute of Technology
Center for Signal and Image Processing
School of Electrical and Computer Engineering
Atlanta, GA 30332-0250 USA
mattlee@ece.gatech.edu, mjts@ece.gatech.edu

ABSTRACT

This paper examines the problem of modelling and resynthesis of voiced song with the goal of improving the subjective performance quality. A set of methods is introduced based on the sinusoidal model for speech which enables precise modification of spectral characteristics as well as vibrato structure while maintaining the original speech quality and naturalness of the voice. Spectral characteristics are modified by modelling the formant structure with a set of asymmetric generalized Gaussians. Subjective tests were conducted which show that the proposed methods are effective in providing high quality modifications to vocal characteristics.

1. INTRODUCTION

For countless years, society has had an appreciation for good singing voices and trained performing artists. While we, as individuals, are able to recognize a good singing voice, or a bad one for that matter, digital characterization of subjective singing quality is quite challenging. A number of models have been advanced in the past for synthesizing voiced song [1, 2]. In addition, several methods have been considered for modifying the fundamental frequency (pitch-scale modification) and the duration (time-scale modification) of singing [3, 4, 5]. However, little has been done in the way of parameterizing the characteristics associated with singing in a way that allows one to digitally transform a poor singer into a good one. In this paper, we take some basic steps in this direction.

2. BACKGROUND

Singing typically contains a much higher ratio of voiced sounds to unvoiced sounds than normal speech. One characteristic of sung voiced sounds is vibrato. Vibrato can be described as a nearly sinusoidal modulation of the fundamental frequency during voiced segments. Studies have shown that the voices of trained singers exhibit vibrato with greater depth and regularity than for those of untrained singers [6]. The pitch contour for the vowel /o/ sung by both a trained singer and an untrained singer is shown in Figure 1. Both signals clearly show vibrato-like fluctuations, but the depth and consistency are much greater in the contour of the trained singer.

Trained singers often create a resonance in the range of 2000 to 3500 Hz by clustering the third, fourth, and sometimes fifth formants. This resonance, referred to as the *singer's formant*, adds a perceptual loudness which allows the singer's voice to be heard over a background accompaniment [7]. Figure 2, which presents the spectral envelope of the two aforementioned singers averaged over time, clearly illustrates the singer's formant in the region of 2500 to 3500 Hz.

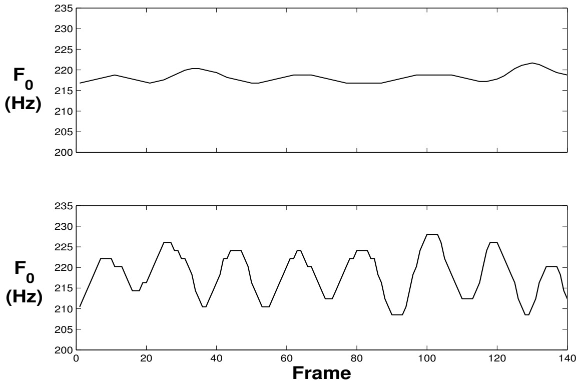

Fig. 1. Fundamental frequency tracks of the vowel /o/ for (a) an untrained singer and (b) a trained singer.

Trained singers also modify their formant structure in other ways to add certain desirable characteristics to their voice. For example, a lowered second formant results in a "darker" voice while a raised second formant produces a "brighter" voice [8]. Female singers often shift their first formant to match the fundamental frequency when the fundamental rises above the first formant. This has the effect of increasing the intelligibility of a vowel sound for a performing artist.

The glottal source can also play an important role in vocal quality. The open quotient (OQ) is the ratio between the open phase of the glottis and the fundamental period and has been shown to be proportional to the ratio of the first two harmonics [9]. We have observed that the open quotient is markedly higher for trained singers than for untrained singers.

In this paper, we present a set of methods enabling control over important vocal characteristics in the singing voice as well as the overall vocal quality. First, we discuss a method for spectral modification based on the Analysis-by-Synthesis/Overlap-Add (AbS/OLA) sinusoidal model for speech. Our method charac-

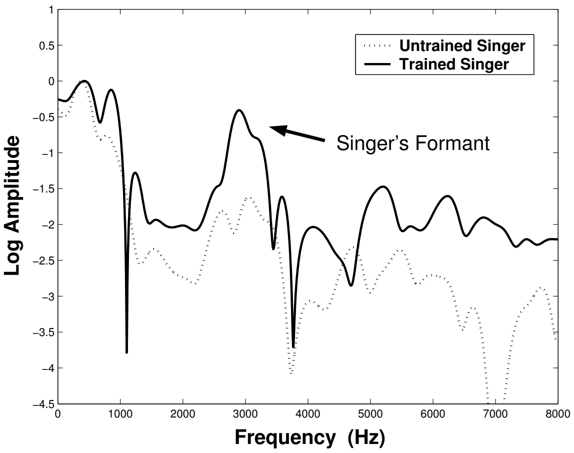

Fig. 2. Average spectral envelopes for the vowel */o/* for a trained singer and an untrained singer. The arrow indicates the region of the singer's formant.

terizes the formant structure as a sum of weighted asymmetric generalized Gaussian functions. This allows flexible and artifact-free modifications to be made by altering the Gaussian parameters. Then we present a method for modifying the vocal vibrato by modelling and parameterizing vibrato within the fundamental frequency track of a voice.

3. SPECTRAL MODIFICATION

Most current methods of spectral modification are based on the source-filter model for speech. These methods model the vocal tract as a slowly varying filter and determine the source by inverse filtering the original speech waveform. In techniques based on linear prediction, pole locations are assigned to formants and then modified to correspond with desired formant modifications. Modification of formants in the line spectrum domain have also been investigated [10].

Other methods use second-order resonant filters to model the formants individually [11]. Fourier or LPC analysis can be used to automatically extract formant frequencies, bandwidths, and source parameters from recorded speech. These parameters are then used to determine the coefficients of the filters.

Formant modification is often difficult with all-pole filters because the assignment of poles to formants is not always obvious. With formant filter models, this is not an issue since the filter specification process requires the characteristics of the formants *a priori*. Formant modification can be performed by scaling the pole angles and/or amplitudes. However, the main drawback to this type of modification is the lack of control over the formant shape. Given the conjugate pole pair $z = r_0 e^{\pm j\theta_0}$ and sampling frequency f_s, the formant frequency F and 3-dB bandwidth B are determined by:

$$F = \frac{f_s}{2\pi}\theta_0 \text{ Hz}, \tag{1}$$

$$B = -\frac{f_s}{\pi}\ln r_0 \text{ Hz}. \tag{2}$$

Thus, the amplitude, r_0, and bandwidth, B, of a formant can not be controlled independently of one another.

3.1. Formant Modelling using Asymmetric Generalized Gaussians

Our approach is based on the AbS/OLA model for speech [4]. This model represents a signal by a sum of equal-length, overlapping, short-time synthesis frames where each frame is approximated by a sum of sinusoids given by

$$s[n] = \sum_{k=0}^{K} a_k \cos\left(\omega_k n + \phi_k\right). \tag{3}$$

Several efficient methods exist for estimating the sinusoidal parameters [3, 2]. For our spectral modification technique, we employ an iterative analysis-by-synthesis procedure developed by George and Smith designed to minimize the mean-squared error [12].

Once the sinusoidal parameters have been obtained, the vocal tract response is estimated. Instead of using LPC or formant filters to determine the response, we interpolate the sinusoidal amplitudes in the power spectrum across frequency using cubic splines. This spectrum is then fitted with asymmetric generalized Gaussian functions. The discrete vocal tract response $V[k]$ is approximated as

$$V[k] = \sum_{m=0}^{M} A_m G_m[k], \tag{4}$$

where

$$G[k] = \begin{cases} \exp\left(-\left(\frac{|k-\mu|}{\beta^l}\right)^{\alpha^l}\right), & k \le \mu, \\ \exp\left(-\left(\frac{|k-\mu|}{\beta^r}\right)^{\alpha^r}\right), & k > \mu, \end{cases} \tag{5}$$

and k is the discrete frequency index.

Initial amplitudes and center frequencies of the formants are identified by peak picking the interpolated frequency response. The parameters of the generalized Gaussians are then determined using an iterative analysis-by-synthesis method similar to the implementation in [13]. The flexibility of the asymmetric generalized Gaussians enable accurate modelling of the formant structure. The left and right spectral widths of each generalized Gaussian are specified by β^l and β^r, while α^l and α^r dictate the decay of the function. These properties are illustrated in Figure 3.

Before formant modification can be performed, each formant must be mapped to a particular Gaussian. Errors can often occur when trying to assign smooth formant trajectories to continuously varying spectral shapes. Formants can merge, split, and sometimes disappear. Since formant changes occur relatively slowly over time, we developed a formant tracking system to perform the mapping within each frame as well as form tracks across frames. The process is based on McAulay and Quatieri's peak matching algorithm for tracking harmonics [14]. A cost function is employed which is based on proximity in frequency and difference in amplitude. Formant tracks are derived such that the cost function is minimized. "Births" and "deaths" of formant tracks are allowed to account for the possibility of the number of distinguishable formants changing from frame to frame.

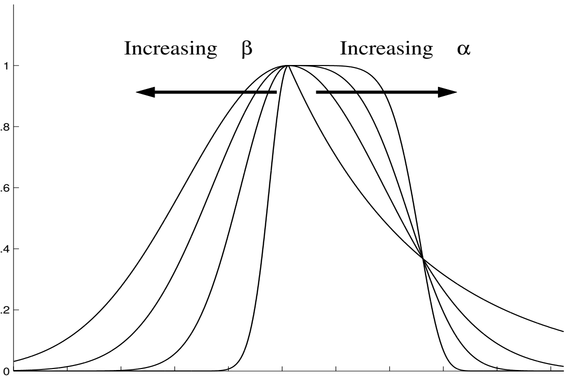

Fig. 3. Asymmetric generalized Gaussians with increasing width parameters (β) to the left and increasing shape parameters (α) to the right.

The parameter structure of the generalized Gaussians enable independent modification of each formant's frequency, amplitude, bandwidth, and shape. This provides the flexibility to improve the performance quality of the voice while enabling the identity of the singer to be maintained to some extent.

The final vocal tract response is obtained by estimating the generalized Gaussian formant structure with a high-order cepstral approximation. The purpose of this is to couple phase characteristics with the magnitude spectrum. When no spectral modifications are applied, the final vocal tract response should closely fit the sinusoidal parameters. This is illustrated in Figure 4 for the vowel /o/.

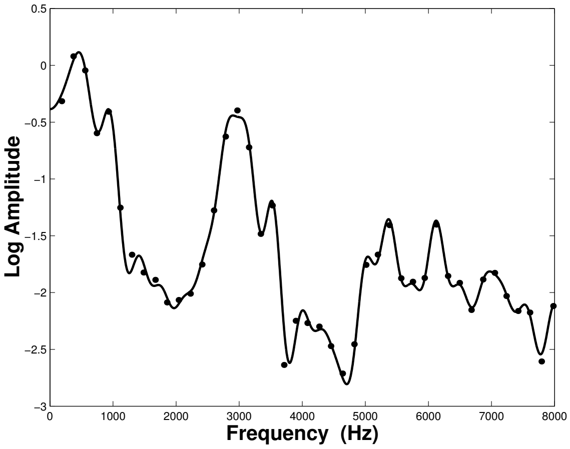

Fig. 4. Log amplitude of the asymmetric generalized Gaussian model (solid line) fit to the sinusoidal amplitudes (dots).

4. VIBRATO MODIFICATION

The AbS/OLA model employs a phasor interpolation process on the excitation signal to perform frame-by-frame pitch scaling of a singing voice while maintaining a certain degree of naturalness and minimizing the presence of artifacts [4]. However, as pitch-scaling factors increase, distortions are introduced, and a tonal quality is imparted on the synthesized signal. Therefore, it is desirable to minimize the amount of pitch-scaling when modifying vibrato characteristics.

As mentioned earlier, vibrato is manifested as a sinusoidal oscillation of the fundamental frequency. Often it is desirable to increase the depth and consistency of an untrained singer's vibrato. We have developed a simple method for detecting regions where vibrato exists and defining minimum frame-by-frame pitch scaling factors for modifying the characteristics of the vibrato.

During voiced segments of the singing waveform, local maxima and minima of the fundamental frequency track are detected. Vibrato is determined to be present in portions where the spacing between adjacent maxima or minima falls within the natural range of vibrato (4-8 Hz). These points are then linearly interpolated to form a *vibrato envelope*. The mean pitch is calculated by averaging the upper and lower boundaries, as illustrated in Figure 5.

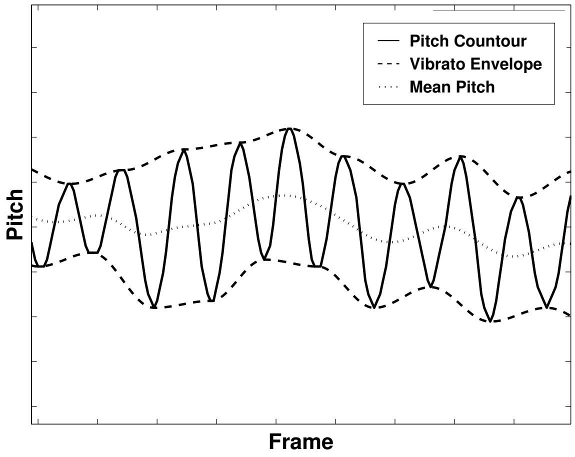

Fig. 5. Vibrato envelope and mean pitch for the pitch curve of a singing voice waveform.

Vibrato depth can be modified at this point by increasing the spacing between the upper and lower boundaries of the vibrato envelope. In addition, the contours of the vibrato and mean pitch can be altered by applying a scaling function. Once the desired vibrato envelope is obtained, the new pitch curve can be calculated by interpolating between points on the desired envelope at the locations of the maxima and minima. A modified cosine interpolation scheme is used such that half a period of a cosine wave is fitted between two points, guaranteeing zero slope at the boundaries. The final pitch-scaling factors can be determined by dividing the modified pitch contour by the original contour.

5. RESULTS

Subjective tests were conducted to evaluate the effectiveness of our modification methods. We first recorded a segment performed by both a male professional singer and an untrained singer at a sampling frequency of 16 kHz. Parameters for the generalized Gaussian model were then calculated for the spectra of each of the segments. Vibrato characteristics were extracted as well. The following modifications were applied to the voiced portions of the untrained singer's voice to assimilate the vocal quality of the trained singer's voice:

- **Singer's Formant:** Cluster 3rd and 4th formants to 2500-3000 Hz and increase amplitude.
- **"Darken" Voice:** Lower 2nd formant and increase amplitude.
- **Modify F1:** Widen the left side of 1st Formant to increase the open quotient (OQ).
- **Vibrato:** Increase depth and regularity.

Thirteen subjects were asked to rate the "overall vocal quality" of the original segments, segments by the untrained singer with each of the modifications performed, and a segment with all of the modifications applied. The subjects rated the segments on a scale of 1 to 100. Figure 6 shows that the greatest improvement to the untrained singer's voice resulted from increasing the depth and regularity of the vibrato using our vibrato modification method. Widening the left side of the first formant also led to a substantial improvement in vocal quality. Each of the other modifications also improved the overall vocal quality but to lesser degrees. While several modification combinations also produced an improved singing voice, the interaction between modifications is yet to be fully understood.

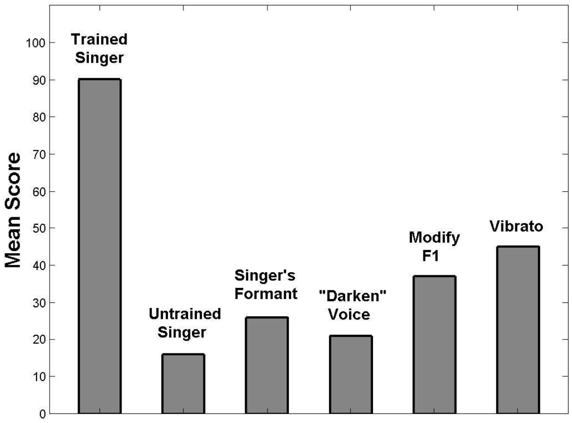

Fig. 6. Thirteen subjects rated the "overall vocal quality" of the synthesized waveforms.

6. CONCLUSION

A set of tools for modification of spectral and vibrato parameters has been presented. The use of asymmetric generalized Gaussians enables spectral modifications to be performed in a flexible yet simple manner while maintaining the original formant characteristics. Vibrato characteristics can also be altered using a simple process that minimizes the amount of pitch-scaling required for the desired modifications. Our experiments show that these methods can be used to enhance vocal quality to a certain extent without degrading naturalness or altering the identity of the original voice. While the goal of transforming the voice of a poor singer into a good one has yet to be fully reached, the proposed models are capable of significantly improving subjective quality.

7. REFERENCES

[1] M. Macon, L. Jensen-Link, J. Oliverio, M. A. Clements, and E. B. George, "A singing voice synthesis system based on sinusoidal modeling," in *ICASSP*, May 1997, vol. 1, pp. 435–438.

[2] Y. Stylianou, J. Laroche, and E. Moulines, "High quality speech modification based on a harmonic + noise model," in *EUROSPEECH*, Sept. 1995, pp. 451–454.

[3] T. F. Quatieri and R. J. McAulay, "Shape invariant time-scale and pitch modification of speech," *IEEE Transactions on Signal Processing*, vol. 40, pp. 497–510, Mar. 1992.

[4] E. B. George and M. J. T. Smith, "Speech analysis/synthesis and modification using an analysis-by-synthesis/overlap-add sinusoidal model," *IEEE Transactions of Speech and Audio Processing*, vol. 5, pp. 389–406, Sept. 1997.

[5] M. Lee and M. J. T. Smith, "Digital singing voice synthesis using a new alternating reflection model," in *ISCAS*, May 2002, vol. 2, pp. 341–344.

[6] P. Cook, "Pitch, periodicity, and noise in the voice," in *Music, Cognition, and Computerized Sound*, pp. 195–208. M.I.T. Press, 1999.

[7] J. Sundberg, "Perception of singing," in *The psychology of music*, D. Deutsch, Ed., pp. 171–214. Academic Press, 1982.

[8] J. Sundberg, "The acoustics of the singing voice," *Scientific American*, vol. 236, pp. 82–91, 1977.

[9] C. d'Alessandro and B. Doval, "Experiments in voice quality modificaion of natural speech signals: the spectral approach," *Third ESCA/COCOSDA Workshop on Speech Synthesis*, Nov. 1998.

[10] R.W. Morris and M.A. Clements, "Modification of formants in the line spectrum domain," *Signal Procesing Letters*, vol. 9, pp. 19–21, Jan. 2002.

[11] T. Styger and E. Keller, "Formant synthesis," in *Fundamentals of Speech Synthesis and Speech Recognition: Basic Concepts, State of the Art, and Future Challenges*, pp. 109–128. John Wiley, 1994.

[12] E. B. George and M. J. T. Smith, "An analysis-by-synthesis approach to sinusoidal modeling applied to the analysis and synthesis of musical tones," *Journal of the Audio Engineering Society*, vol. 40, pp. 497–516, June 1992.

[13] D. Britton and M. J. T. Smith, "Generalized Gaussian decompositions for SAR enhancement using analysis-by-synthesis," *to appear in Digital Signal Processing Journal*, 2003.

[14] T. F. Quatieri and R. J. McAulay, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, pp. 744–754, Aug. 1986.