REAL-TIME ESTIMATION OF THE VOCAL TRACT SHAPE FOR MUSICAL CONTROL

Adam P. Kestian and Tamara Smyth
School of Computing Science, Simon Fraser University
apk6@sfu.ca, tamaras@cs.sfu.ca

ABSTRACT

Voiced vowel production in human speech depends both on oscillation of the vocal folds and on the vocal tract shape, the latter contributing to the appearance of formants in the spectrum of the speech signal. Many speech synthesis models use a feed-forward source-filter model, where the magnitude frequency response of the vocal tract is approximated with sufficient accuracy by the spectral envelope of the speech signal. In this research, a method is presented for real-time estimation of the vocal tract area function from the recorded voice by matching spectral formants to those in the output spectra of a piecewise cylindrical waveguide model having various configurations of cross-sectional area. When a match is found, the formants are placed into streams so their movement may be tracked over time and unintended action such as dropped formants or the wavering of an untrained voice may be accounted for. A parameter is made available to adjust the algorithm's sensitivity to change in the produced sound: sensitivity can be reduced for novice users and later increased for estimation of more subtle nuances.

1 INTRODUCTION

Estimation of the vocal tract area function from an incoming voice signal is a task that has numerous proposed solutions, as several applications would greatly benefit from accurate depictions of the vocal tract shape during speech. For example, studies have found that displaying the vocal tract to the hearing impaired can improve their overall speech performance [11, 15] as well as being an effective instructional tool for singers [14]. Recent studies have aimed to identify useful features from the voice signal for musical control [7, 6]. Some have identified the vocal tract shape as a potential feature for musical control and have used vision-based methods for its estimation [5, 10]. Here, we propose using the vocal tract shape as input data for control, though leaving the actual application and mapping strategy to the user, and present a method for extracting this shape directly from recorded voice by calibrating to the output of a waveguide model.

Human speech depends both on the oscillation of the vocal folds in the glottis and on the shape of the vocal tract. Producing a particular vowel sound requires changing the effective cross-sectional area function of the vocal tract that contributes to the appearance of formant peaks in the frequency spectrum of the speech signal. While several algorithms have been proposed to identify the vocal tract area function from recorded speech, the most common involve either computing reflection coefficients from linear predictive coding (LPC) or autoregressive (AR) models using, for example, Yule-Walker or Levinson-Durbin methods [2, 18], or by directly tracking formant peaks in the spectral envelope of recorded speech [9, 3]. Both approaches assume a simplified feed-forward source-filter model of the voice, and produce best results on vowel sounds, as the articulation of consonants is more complicated than purely changing the vocal tract shape.

For many applications, the feed-forward filter model is considered to be sufficiently accurate as there is weak coupling between the massy vocal folds and the vocal tract. It should be mentioned however, that a stronger influence of the vibrating source is observed on the spectral envelope—and thus the formant peaks—for feed-back models that more strongly couple the glottal excitation and vocal tract [17]. Likely due to this and other simplifications in the filter model, such as inaccurate estimates of unknown propagation losses and termination reflection/transmission functions, the detection of vocal tract shape from reflection coefficients has produced inconsistent and inaccurate results, both in this work and in that of other authors [9, 12, 19]. Furthermore, accuracy deteriorates rapidly with an increase in sampling rate, making its use limited for the singing voice which typically requires a greater bandwidth than its spoken counterpart.

Formant-based analysis-by-synthesis methods may produce better results as they overcome filter inaccuracies by only requiring a most likely fit to a synthesis model's output. Depending on the method used, calibrating a recorded signal to the output of a very detailed and accurate model could introduce potentially restrictive computational cost, thus making it inappropriate for real-time use. In this particular case, there is also the possibility of estimating one of several vocal tract shapes that produce indistinguishable

spectra. Care must therefore be taken to constrain possible shapes so they are physiologically reasonable.

In this work, we present an efficient and accurate formant-based vocal tract area function estimation algorithm, specifically designed for real-time musical control. As described in Section 3, the algorithm identifies formants on a sample frame basis, and places them into streams, enabling their movement to be tracked over time.

In Section 4, minimum action is applied to improve usability, algorithm performance, and the visual feedback to the user, by smoothing formant streams to account for unintended action such as dropped formants or the wavering of an untrained voice (see Section 4). It considers that some users will have greater control of their voice than others, and provides a parameter for adjusting the algorithm's sensitivity to a change in the produced sound.

Section 5 describes how estimated formants are compared and matched to a database of formants collected offline from the output of the vocal tract model (described in Section 2) having various configurations of cross-sectional area.

2 A SIMPLIFIED VOCAL TRACT MODEL

It is well known that digital waveguide modeling may be used to simulate plane and spherical waves propagating in cylindrical and conical acoustic tubes [16]. More complex shapes, such as those produced by the vocal tract when using the velum, jaw, tongue and lips to voice different vowel sounds, can be approximated using a sequence of cylindrical waveguide elements separated by scattering junctions accounting for the reflection and transmission that occurs when a change in cross-sectional area creates a corresponding change of impedance.

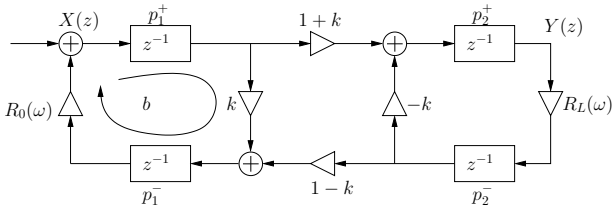

Figure 1. A waveguide model of an acoustic tube closed at one end with reflection $R_0(\omega)$ and open at the other with reflection $R_L(\omega)$. The tube has a single change in cross-sectional area at the center, creating a two-port scattering junction with reflection and transmission defined by k, a coefficient set according to the relative areas of the two sections.

The way in which a shape departing from the purely cylindrical contributes to formant peaks in the magnitude spectrum may be seen by considering a simple model with only two cylindrical waveguide segments of the same length but with different cross-sectional areas, A_1 and A_2 respectively (see Figure 1). The two-port scattering junction models the reflection and transmission that occurs at the change in cross-sectional area, where the reflection coefficient is given by

$$k = (A_1 - A_2)/(A_1 + A_2). \tag{1}$$

The response at $Y(z)$ (corresponding to the mouth) to input $X(z)$ (corresponding to the glottis) is given by

$$\begin{aligned} Y(z) &= X(z)(1+k)z^{-2}[1+b+b^2+\ldots] + \\ &\quad Y(z)R_L(z)R_0(z)(1-k^2)z^{-4}[1+b+\ldots] + \\ &\quad Y(z)R_L(-k)z^{-2}, \end{aligned} \tag{2}$$

where

$$b = R_0(z)kz^{-2}. \tag{3}$$

Equation (2) is the sum of three terms corresponding to the possible signal flow paths to $Y(z)$, with the first two terms including the infinite series generated by the circulating path in the first waveguide segment. Using the closed form resprepresentation and substituting (3) into (2) yields all-pole filter transfer function

$$\begin{aligned} H(z) &= Y(z)/X(z) \\ &= \frac{(1+k)z^{-2}}{1+k(R_L(z)-R_0(z))z^{-2}-R_L(z)R_0(z)z^{-4}}. \end{aligned} \tag{4}$$

Setting end reflection functions

$$R_0(z) = 1, \quad \text{and} \quad R_L(z) = -1$$

makes the system lossless with transfer function

$$\hat{H}(z) = \frac{(1+k)z^{-2}}{1-2kz^{-2}+z^{-4}}, \tag{5}$$

preserves the harmonic structure of an open end tube, and allows for observation of the effects of the junction.

Factoring the denominator in (5) yields the intermediate complex conjugate pair,

$$\rho = k + j\sqrt{1-k^2} \qquad \rho^* = k - j\sqrt{1-k^2}, \tag{6}$$

and ultimately the four roots/poles of the filter given by

$$p_1 = \sqrt{\rho}, \quad p_2 = -\sqrt{\rho}, \quad p_3 = p_1^*, \quad p_4 = p_2^*. \tag{7}$$

Figure 2 shows how the poles shift as a function of k and thus in response to a change in cross-sectional area. Shifting poles corresponds to a shift of harmonic peaks in the magnitude which, when more sections with varying cross-section are considered, leads to regions in the spectrum with increased and decreased energy, and the appearance of formant peaks during vowel production.

Increasing the number of segments increases the number of poles in the vocal tract transfer function. As shown in (6) and (7), filter poles are a function of the reflection

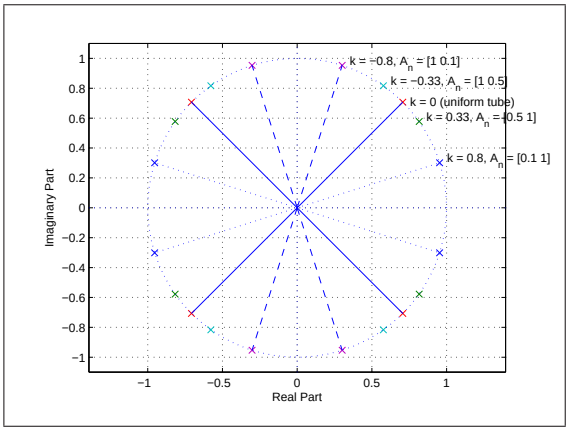

Figure 2. The four poles of transfer function (5) plotted for five values of reflection coefficient k. The uniform cylindrical tube has a reflection coefficient of $k = 0$ and corresponds to a uniform/harmonic spacing of poles (or peaks in the magnitude spectrum). A change in k corresponds to a change in the cross-sectional area to the tube and the observed shifting of poles in the vocal tract transfer function.

coefficient, allowing a change in cross-sectional area to be inferred directly from filter poles. The complexity involved in this recursive problem however, is unnecessarily expensive (though not prohibitively so) for real-time applications, and yields far more data than is required to identify the vocal tract shape with the accuracy desired here. Rather, it was found that a vocal tract shape could be sufficiently characterized using only up four formant peaks in the magnitude of its frequency response.

3 TRACKING FORMANTS IN VOCAL OUTPUT

An incoming sample frame of recorded voice is windowed and processed to extract its spectral envelope—a curve assumed to approximate the magnitude of the vocal tract frequency response—with formants peaks in the spectra being identified by tracking curve local maxima.

Notice from the log spectrum in Figure 3 (upper curve) that the position of weaker formants is sometimes obscured by the presence of more pronounced formants having greater amplitude and bandwidth. As shown by the broken-line curve in Figure 3 (lower curve), the second derivative of the log magnitude spectrum may be taken to produce more prominent bends in the curve contour at peak locations, effectively reducing the formant bandwidth and accentuating the position of "merged" formants [13]. Though is also possible to take the third derivative of the phase spectrum [8] to yield further improvement in bandwidth attenuation and peak accentuation, this method was found to be less successful in tracking merged formants with significantly different amplitudes, and thus produced less consistent results.

Once the most prominent formant peaks are detected,

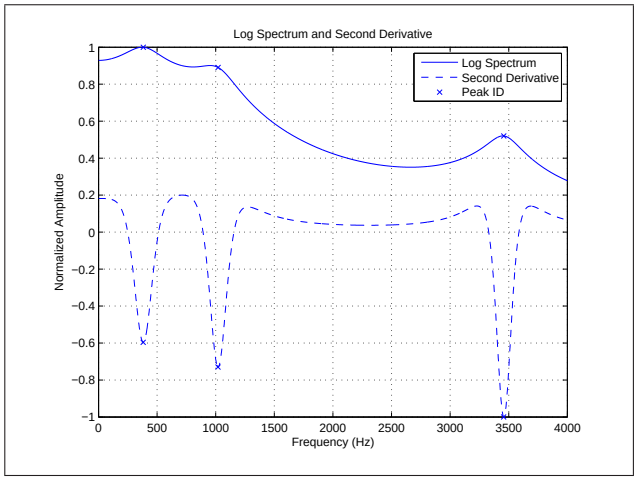

Figure 3. The log magnitude spectrum of an input sample frame (upper solid line) and its second derivative (lower broken line), with the latter accentuating the position of "merged" formants.

they are placed into formant streams that track the movement of a formant number from frame to frame. These formant streams are necessary to account for dropped formants and improve usability and performance as discussed in Section 4. Limiting the streams to a maximum of four was sufficient to uniquely identify a corresponding vocal tract model.

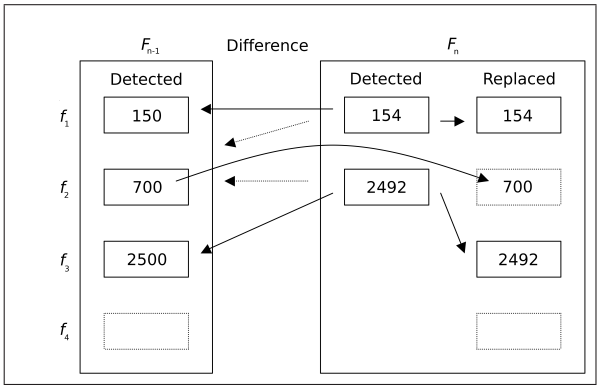

Figure 4. Example of formant stream assignment: Analysis of current frame F_n yields two detected formant peaks at 154 Hz and 2492 Hz. The first peak at 154 Hz is closest to the first formant of the previous frame F_{n-1} and is thus assigned to the first formant stream $f_1(n)$. Similarly, the peak at 2492 Hz is assigned to the third formant stream $f_3(n)$. To accommodate for the "dropped" formant in the second stream, the last value assigned from the previous frame is held over to the current frame.

To determine to which stream a formant peak should be assigned, a distance measure is taken between the estimated formant peak of a current frame F_n and the stream-assigned neighbouring formants of the previous frame, F_{n-1}, with

the formant being assigned to the stream of its neighbour closest in frequency. If the difference between formant frequencies exceeds a threshold, an additional formant stream becomes active. Though it is possible to have up to four streams, it is most common to use only three.

The process is illustrated by the example in Figure 4, where only two peaks have been detected with three active streams, flagging the possibility that a formant was unintentionally "dropped". Accounting for such absent formant peaks, as described in Section 4, further improves usability of the system.

4 MINIMUM ACTION FOR IMPROVED USABILITY

There are the two situations in which a formant may unintentionally temporarily disappear from one sample frame to the next: 1) when the algorithm fails to detect it for a particular frame or more likely 2) an untrained voice is unable to consistently sustain the quality of the produced sound. As shown in Figure 5 (top), either instance generates a sharp null in the formant tracking curve.

To accommodate for this, minimum action is assumed, and such significant temporary drops in the curve are considered unlikely or unintentional (with minimum action suggesting too much effort would be required to intentionally produce such a drastic change). Consider again the example shown in Figure 4, which shows only two peaks being detected in the presence of three active formant curves. Since the detected peak is placed in the third stream, it is the second formant that was dropped. In this case, the last value in the second stream is held over to the current frame, $f_2(n) = f_2(n-1)$. In this way, when/if the formant returns in subsequent frame analysis, it will be placed in the correct stream and the sharp nulls in the curves will be avoided (see top and middle of Figure 5). The repetition of formants within a stream can occur up to four times before the stream is deemed inactive.

Algorithm performance and visual feedback to the user is further improved by applying a smoothing filter to the formant tracking curves, effectively stabilizing the movement of the formants and compensating for unintentional wavering of the less-trained voice. An amplitude envelope follower given by,

$$\hat{f}_m(n) = (1-\nu)|f_m(n)| + \nu\hat{f}_m(n-1), \qquad (8)$$

is applied to the formant streams $f_m(n)$, where ν determines how quickly changes in $f_m(n)$ are tracked. If ν is close to one, changes are tracked slowly, making the smoothed curve $\hat{f}_m(n)$ less sensitive to change. If ν is close to zero, $f_m(n)$ has an immediate influence on $\hat{f}_m(n)$. A higher ν, therefore may be appropriate for untrained voices, but with practice, the parameter value may be decreased to allow for better detection of subtle nuances.

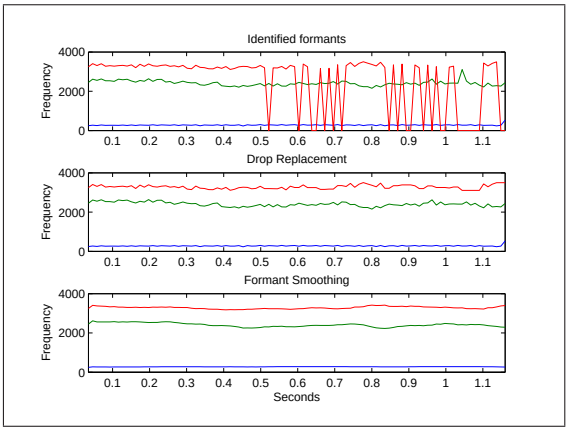

Figure 5. Three active formant streams with the third stream having sharp nulls due to temporarily dropped formants (top). Nulls are avoided by holding values from the previous frame when a formant is flagged as being dropped (middle). Further smoothing is applied to compensate for unintentional wavering in the less-trained voice (bottom).

Regardless of ability, formants shift rapidly during an onset of (or change in) the vocal/vowel sound, and thus detected formants are added to streams only once the formants have settled and the speech waveform is more sustained. (This creates latency in the visual feedback to the user equal to the onset duration). Extensive methods for detecting attacks in sounds from musical instruments are not necessary here, particularly since they are considered to be less effective when applied to the voice [4]. Figure 6 shows the waveform recorded when a speaker produces the vowel sounds /ee/ to /oo/ and back to /ee/. In spite of the speaker's attempt to keep the amplitude constant, the waveform envelope clearly shows a change in energy at the locations of the changing events. This result is expected when considering that the waxing and waning in the frequency spectrum due to shifting formants will have an equivalent effect on the signal's energy in both time and frequency domains (Parseval's theorem).

The onset of an event is therefore identified solely by tracking steep slopes in the amplitude envelope of the speech signal. At an event onset, the formant peaks are still detected, but with their rate of change being recorded from frame to frame. Once this value is sufficiently reduced and the position of the speaker/singer's formants settle, the onset region is considered to have passed and the algorithm may resume with the process of formant stream assignment and vocal tract shape estimation.

5 ESTIMATION OF THE VOCAL TRACT SHAPE

With the estimated formant streams in place, the final step is to search the database produced by the output of the model configured to various shapes, and find the most likely can-

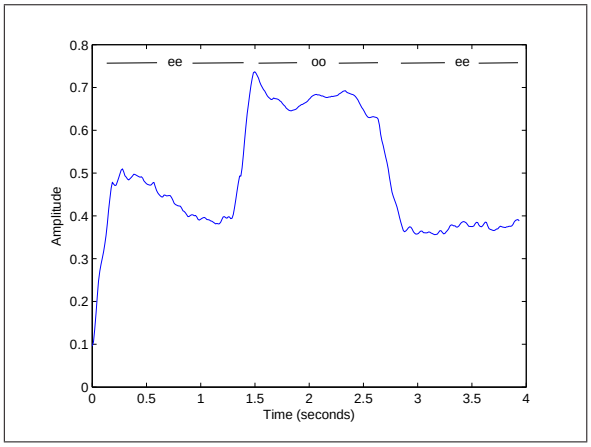

Figure 6. Amplitude envelope produced when voicing vowels /ee/ then /oo/ then back to /ee/ while attempting to keep the amplitude (loudness) constant. The amplitude envelope clearly shows the locations of these event changes.

didate.

A piecewise cylindrical waveguide model, similar to that shown in Figure 1, was developed having N_s sections, each with N_A possible cross-sectional area values, yielding $N_A^{N_s}$ possible combinations. Considering all possible combinations yields duplicate shapes however, fewer corresponding vocal tract area functions are produced by considering only the change in cross section. Here, $N_s = 7$ and $N_A = 5$ seemed to produce the best results when considering performance, usability and accuracy, and the intended application of real-time musical control.

A table maps the model's vocal tract area function to the formant frequencies in its output spectra. The table is sorted by grouping the shapes based on the number of detected formants. The formants detected from the incoming speech signal (as described in Section 3) are compared to the entire portion of the database having the same number of formants. The Euclidean distance $d(fU, fM)$ is used to measure which set of formants generated by the model fM_m best match those generated by the user fU_m:

$$d(fU, fM) = \sqrt{\sum_{m=1}^{M} (fU_m - fM_m)^2}, \qquad (9)$$

where M is the number of detected formants. The estimated vocal tract shape displayed to the user remains static until another event onset is detected. Once a new shape is identified, linear interpolation is performed over the next frame to smooth the transition, and thus the visual display, between the two shapes.

6 RESULTS AND CONCLUSIONS

Figures 7 and 8 provide vocal tract estimation examples of sung vowels /au/ and /ee/. The lower plots of the two fig-

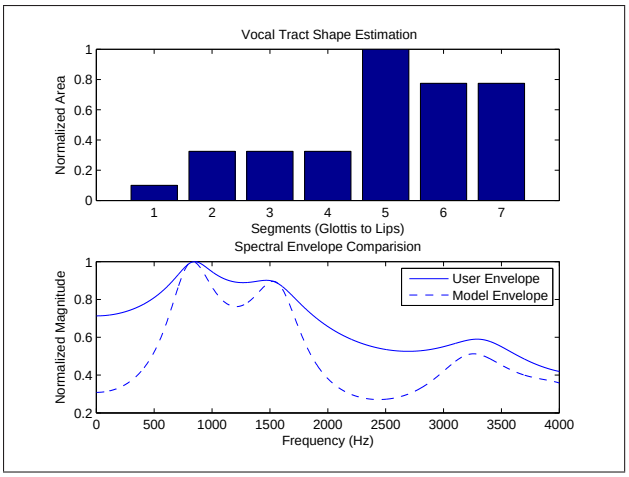

Figure 7. Estimation of vocal tract shape for vowel sound /au/.

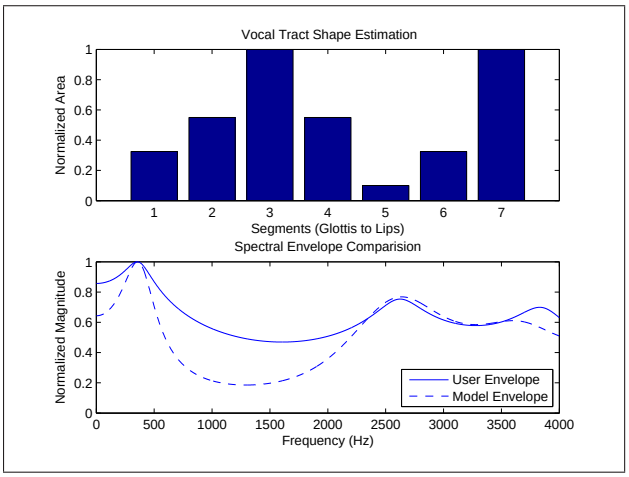

Figure 8. Estimation of vocal tract shape for vowel sound /ee/.

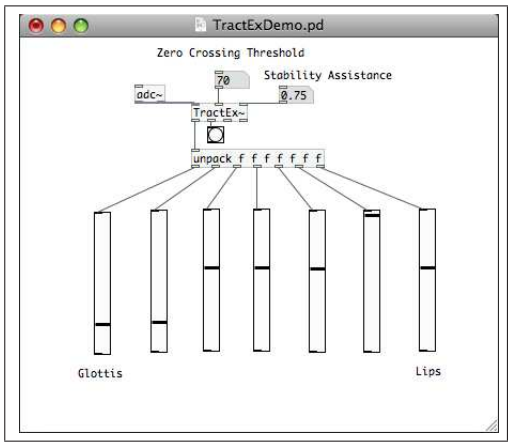

Figure 9. Pure Data object displaying /au/ example

ures compare the spectral envelope of the voice signal to the spectral envelope generated by the vocal tract model of the estimated shape. The differences between the spectral envelope of the voice and the model can be attributed to simplifications of the source-filter model, the wave propagation losses and unknown termination reflection/transmission functions (as discussed in Section 1). Also a factor is the limited resolution of the vocal tract model look-up table. Though increasing the resolution would produce more spectral envelopes for comparison, it would increase the number of permutations and possibly adversely affect real-time performance without necessarily improving estimate accuracy.

With this method, we provide a vocal tract area function estimation algorithm that offers a suitable level of sensitivity for users having varying abilities. The use of formant streams enable a formant's movement to be tracked over time so that the vocal tract shape may be stabilized by accounting for unintended action, thereby improving its use for musical control.

Strategies for mapping the vocal tract area function data to control other music applications are left to the user. The authors are currently interfacing this work to the control of parametric synthesis models and polyphonic virtual instruments.

The algorithm is implemented in PureData (Pd) [1], a real-time audio programming environment popular among musicians. This facilitates control data acquisition, visual and audio feedback display, as well as mappings to other music and sound synthesis applications. The object will be made available to the public upon request.

7 REFERENCES

[1] "Pd," http://www.pure-data.org.

[2] P. Broersen, "Finite sample criteria for autoregressive order selection," in *IEEE Trans. Signal Processing*, vol. 48, 2000, pp. 3550–3558.

[3] J. Dang and K. Honda, "Estimation of vocal tract shapes from speech sounds with a physiological articulatory model," in *Journal of Phonetics*, vol. 30, 2002, pp. 511–532.

[4] S. Dixon, "Onset detection revisited," in *Proceedings of the 9th International Conference on Digital Audio Effects*, 2006, pp. 133–137.

[5] V. Florian, G. McCaig, M. Ali, and S. Feis, "Tongue 'n' groove: an ultrasound based music controller," in *Proc. New Interfaces for Musical Expression*, 2002.

[6] A. Hazan, "Performing expressive rhythms with billaboop voice-driven drum generator," in *Proc. of the 8th Int. Conference on Digital Audio Effects*, 2005.

[7] J. Janer and M. de Boer, "Extending voice-driven synthesis to audio mosaicing," in *5th Sound and Music Computing Conference*, 2008.

[8] C. Jinhai, J. Gangji, and Z. Lihe, "A new method for extracting speech formants using lpc phase spectrum," in *IEEE Electronics Letters*, vol. 29, 1993, pp. 2081–2082.

[9] P. Ladefoged, R. Harshman, L. Goldstein, and L. Rice, "Generating vocal tract shapes from formant frequencies," in *J. Acoust. Soc. Am*, vol. 64, 1978, pp. 1027–1035.

[10] M. Lyons, M. Haehnel, and N. Tetsutani, "Designing, playing, and performing with a vision-based mouth interface," in *Proc. New Interfaces for Musical Expression*, 2003.

[11] S. Park, D. Kim, J. Lee, and T. Yoon, "Integrated speech training systems for hearing impaired," in *IEEE Transactions on Rehabilitation Engineering*, vol. 2, 1994, pp. 189–196.

[12] D. Paul, "Estimation of the vocal tract shape from the acoustic waveform," in *Ph.D. Diss., Massachusetts Inst. Technol.*, 1976.

[13] P. P. R. Christensen, W. Strong, "A comparison of three methods of extracting resonance information from predictor-coefficient coded speech," in *IEEE Trans. Acoust. Speech Signal Process.*, vol. 24, no. 1, 1976, p. 814.

[14] D. Rossiter, D. Howard, and M. Downes, "A real-time lpc-based vocal tract area display for voice development," in *Journal of Voice*, vol. 8, 1994, pp. 314–319.

[15] M. Shah and P. Pandley, "Areagram display for investigating the estimation of vocal tract shape for a speech training aid," in *Proc. Symposium on Frontiers of Research on Speech and Music*, 2003, pp. 121–124.

[16] J. O. Smith, *Digital Waveguide Modeling of Musical Instruments*. ccrma.stanford.edu/˜jos/waveguide/, 2003, last viewed 12/4/08.

[17] T. Smyth and A. Fathi, "Voice synthesis using the generalized pressure controlled-valve," in *Proceedings of ICMC 2008*, Belfast, Ireland, August 2008, pp. 57–60.

[18] H. Wakita, "Direct estimation of the vocal-tract shape by inverse filtering of acoustic speech waveforms," in *IEEE Trans. Audio Electroacout.*, vol. AU-21, 1973, pp. 417–427.

[19] ——, "Estimation of vocal-tract shapes from acoustical analysis of the speech wave: the state of the art," in *IEEE Trans. Acoust. Speech Signal Process*, vol. ASSP-27, 1979, pp. 281–285.