Proc. of IEEE International Conference on Signal and Image Processing (ICSIP), 2006, Hubli, India.

VOICE QUALITY SYNTHESIS WITH THE BANDWIDTH ENHANCED SINUSOIDAL MODEL

Narsimh Kamath
NIT Karnataka
narsimh@gmail.com

Vishweshwara Rao
EE Dept, IIT-Bombay
vishu@ee.iitb.ac.in

Preeti Rao
EE Dept, IIT-Bombay
prao@ee.iitb.ac.in

ABSTRACT

Voice quality attributes have been found to play a significant role in the naturalness and perceived affect of synthesized speech. Yet, traditional synthesis techniques seem to offer inadequate control over voice quality in synthesized speech. In this paper, we investigate the use of the recently proposed bandwidth enhanced sinusoidal model for synthesis of the roughness attribute in spoken vowels. The vowels thus synthesized are compared with those synthesized using a traditional sinusoids+noise model, both, with respect to the extent of perceptual fusion achieved and the desired change in timbre towards roughness. The bandwidth enhanced sinusoidal model is observed to produce better fused sounds. Further, model parameter selection is investigated with a view to obtaining controlled variations in the perceived roughness of synthesized vowels.

KEY WORDS

Voice quality, synthesis, roughness, perceptual fusion.

1. Introduction

Voice quality is an important characteristic of any speech sound and refers to its overall perceived quality. Qualifiers such as 'rough', 'breathy', 'modal', etc. are commonly used to describe the voice quality attributes of a particular speech sound. In the context of synthetic speech, voice quality attributes are instrumental in determining naturalness, as well as the perceived affect. For example, the affect of anger in natural voices has been found to be characterized by a rough voice quality, while the affect of joy has been found to be associated with a breathy voice quality [1]. Variations in voice quality originate in the speech production mechanism. For instance, glottal pitch cycle perturbations (jitter and shimmer) are correlated with perceived roughness in the voice, as shown by numerous studies [2, 3]. The percept of breathiness has been found to be caused by aspiration noise [4].

The importance of voice quality attributes, such as roughness and breathiness, makes it imperative to search for methods of synthesis that can introduce, as well as control such voice quality attributes in synthetic speech. In speech synthesis and coding, sinusoidal models have been widely used due to the compactness of parameters they provide and the flexibility available for obtaining prosodic variation by easily implemented time- and pitch-scale modifications. The basic sinusoidal model [5] represents a signal as a sum of sinusoids with time-varying frequencies, amplitudes and phases. This model provides an inadequate representation of sounds having significant inharmonic content. More recent models such as the sines+noise models account for the inharmonic content of a sound by adding a stochastic component to the basic sinusoidal model. In the spectral modeling synthesis (SMS) method [6], spectrally shaped noise is added to the sum of sinusoids to account for any inharmonic content in the sound, and this represents an improvement over the basic sinusoidal method for synthesis of a wider range of natural sounds.

While the traditional sines+noise models provide a certain amount of flexibility in manipulating the individual sinusoids for desired time and pitch-scale modifications, there is considerably less flexibility where the modification of the corresponding noise component is concerned. Further, the loss of homogeneity arising from simple additive combination of distinct component types (periodic and noisy) has an important perceptual consequence, namely the lack of perceptual fusion between the components. This results in the audible presence of an unnatural background noise in the synthesized speech.

The bandwidth enhanced sinusoidal model, recently proposed by Kelly Fitz [7] as a variant of the sines+noise model, has been shown to give high fidelity synthesis for certain types of sounds such as transient sounds, and more crucially, breathy sounds. The homogeneity of the model promises greater control over voice quality modifications than that afforded by the traditional models. In particular, the association of noise with individual partials makes it easier to manipulate the sinusoidal and noise components together. During the synthesis of breathy sounds such as flute, Kelly reports that the synthesized noise is found to fuse with the sinusoids into a single sound, and this is attributed to the fact that in this model the energy in the noise spectrum exactly tracks the sinusoidal partial amplitudes.

In additive models of synthesis, the noise component and the sinusoidal component are the outputs of two different production mechanisms. Different techniques try to integrate these two components in different ways, resulting in more or less homogenous synthesis methods. A method of synthesis leading to a high level of perceptual fusion will result in an increased naturalness of the synthesized speech.

While there has been much research focus on the synthesis of breathy sounds, the synthesis of roughness in speech remains a relatively ignored, yet challenging, problem. The potential shown by the bandwidth enhanced sinusoidal model in the synthesis of breathy sounds due to its inherently fused structure motivate an exploration of the suitability of the model for the synthesis and control of perceived roughness in synthetic speech. In this paper, we present an experimental investigation of the Kelly model and its comparison with the spectral model synthesis (sines+noise) method for the synthesis of vowels at different fundamental frequencies characterized by controlled amounts of roughness. Reference sounds are provided by a source-filter model based synthesis of pitch jitter. The next section provides an introduction to the perceptual attributes of synthesized sounds.

2. Perceptual attributes of synthesized sound

While synthesizing the percept of roughness in vowels is of prime interest, the naturalness of such vowels depends to a large extent on the degree of perceptual fusion achieved between the synthesized periodic and aperiodic components. Very little attention has been focused on the issue of perceptual fusion in synthesized sounds. In fact, the very term perceptual fusion is found to elude a complete definition. In this context, notable is the work of Hermes on synthesis of breathy vowels [8]. A stream of high-pass filtered noise pulses are added to a glottal pulse train. This excitation signal is filtered using a formant filter in order to synthesize breathy vowels. Hermes notes that a decrease in loudness of the noise component in the synthesized sound is reflected in a change in the timbre of the vowel such that the vowel itself is perceived to contain high frequency content, in which case, he considers the noise and vowel to be perceptually fused. Generalising this observation, fusion refers to the extent of perceptual integration of two components A and B combined together in such a way so as to produce a third sound C of some desired timbre different from A or B. As the degree of fusion improves, it is expected that the change in timbre of the combined sound from either of its components increases while the perceived loudness of any unfused component decreases. When complete fusion is achieved, the sound C must appear to be produced from a single sounding object or source.

Fales and McAdams, in a study on perceptual fusion in African instrument sounds [9], subjectively evaluate the fusion of noise and tone using synthetic stimuli consisting of a single tone with added bandlimited noise. They consider three perceptual phenomena that are said to be possible from such a combination of noise and tone: fusion, layering and masking. The first phenomenon occurs when the two components of the sound are perceptually integrated into a new sound; the second occurs when the two components are perceptually segregated; and the third occurs when the noise masks the tone completely. The authors are prevented from evaluating the stimuli on a continuous scale of fusion due to a lack of clarity among the listeners on the definition of fusion, and conclude that the subjective judgments of 'tone not heard separately: not sure', and 'tone heard separately: not sure', best represent the state of perceptual fusion in the synthesized sound. A study of a naturally fused flute sound leads them to suggest that 'degrees of fusion' might be better suited towards representing the perceptual fusion in a complex tone plus noise.

Johansson [10] argues that the percept of fusion should not be considered in a categorical manner; rather characterizations such as 'layered', 'augmented', and 'fused' should be considered as different levels on a relative, and more or less continuous scale. He expresses reservations in using timbre change as an indicator of perceptual fusion, since timbre change might also be caused by artifacts introduced due to the test itself.

Thus while subtle differences exist in the previous work regarding the issue of perceptual fusion, there seems to be a broad consensus on at least two points:

1. A continuous scale of fusion, or at least a discrete scale representing degrees of fusion will be better suited for representing perceptual fusion, rather than a 'yes' or 'no' type decision.
2. Timbre change towards a certain 'target' percept can be construed as one of the indications of perceptual fusion, provided there arise no changes in timbre due to the test conditions themselves.

3. Vowel synthesis using the bandwidth enhanced sinusoidal model

One of the advantages of the bandwidth enhanced sinusoidal model over the traditional sine +noise models is its homogeneity. The reason for this is that Kelly [7] proposes a unified generator for both the noise component and the sinusoidal component of the synthesized sound. This unified generator employs amplitude modulation in which bandlimited noise can be considered to be the modulating signal, and a complex exponential corresponding to a particular partial can be considered to be the modulated signal. This unit of the model is called the bandwidth enhanced sinusoidal oscillator and is described in eq. (1).

$$y_n = (A + \beta[\varsigma_n * h_n]) \cdot e^{j\omega_c n} \qquad (1)$$

Here, y_n is the synthesized waveform, A is the sinusoidal carrier amplitude, β is the amplitude of the noise modulation, ς_n is the output of a random number generator, h_n is the impulse response of a low pass filter used on the random number sequence, and ω_c is the frequency of the complex exponential. A collection

of such bandwidth enhanced sinusoidal oscillators, each corresponding to a different partial, is used for synthesis using what we henceforth refer to as Kelly's model. If we now define the local average partial energy $\tilde{A}$ as

$$\tilde{A} = \sqrt{A^2 + \beta^2} \quad (2)$$

and the bandwidth coefficient κ as

$$\kappa = \frac{\beta^2}{A^2 + \beta^2} \quad (3)$$

then eq (1) can be rewritten as

$$y_n = \tilde{A}\left(\sqrt{1-\kappa} + \sqrt{2\kappa}\left[\zeta_n * h_n\right]\right) \cdot e^{j\omega_c n} \quad (4)$$

The model parameters are the bandwidth (BW) of the bandlimited noise and the bandwidth coefficient (κ). The stochastic modulation in eq. (4) leads to a spreading of the spectral energy around the partial center frequency, a phenomenon referred to as spectral line widening or bandwidth enhancement. An increase in the κ value leads to an increase in the line widening, and this appears as an increase in the partial bandwidth relative to the peak spectral amplitude, as shown in Fig. 1.

In his method of synthesis, Kelly assigns equal values of κ and BW to the partials in a certain frequency range in order to synthesize breathy sounds. The resulting synthesized sounds are reported to demonstrate a high level of fidelity.

For vowel synthesis, an FFT analysis of a single period of a reference vowel (synthesized using the source-filter speech production model) is done, and the complex FFT values so obtained at each harmonic are used in Kelly's model. This accounts for the harmonic component of the synthesized vowel. The model parameters κ and BW assigned to each partial can then be used in the synthesis of the inharmonic component associated with the individual partials of the synthesized vowel.

4. Experiments

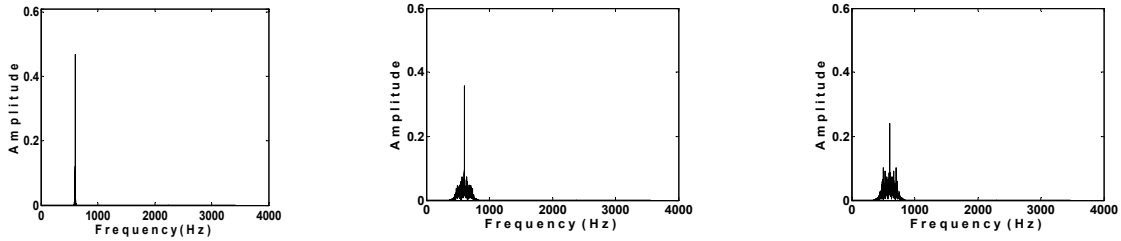
This section describes the experiments conducted for the synthesis of rough vowels and the listening tests performed in order to evaluate the same. In order to assist in identifying the percept of roughness, as well as to provide a quantitative measure of the amount of roughness in the synthesized vowels, reference vowels /a/, /i/, and /u/, were synthesized at fundamental frequencies of 100 Hz, 200 Hz, and 300 Hz with varying amounts of jitter, or pitch perturbation, using the unified glottal source model [11]. These vowels were synthesized by filtering an LF model glottal pulse train using the algorithm in [12], using acoustic parameters derived from natural vowels uttered by a low pitched male speaker. The sampling rate used was 8000 Hz. The speech production model facilitates the control of roughness by variation of the percentage of jitter parameter.

4.1 Synthesis of rough vowels using the spectral modeling synthesis (SMS) method:

In the spectral modeling synthesis method [6], the inharmonic component, or residual is spectrally shaped before being added to the harmonic component. This method represents an improvement over the basic sinusoidal model, and is a possible candidate for the synthesis of a percept of roughness. For the purpose of synthesis of vowels using this method, the sinusoidal component is synthesized using the basic sinusoidal model, while the noise component is generated by spectrally shaping white noise, in a particular frequency region. The spectral envelope of the sinusoidal component, obtained by interpolating the line spectrum is used for this purpose. This noise component is then added to the sinusoidal component to produce the synthesized vowel. The model parameters are the signal to noise energy ratio (SNR), defined in eq. 5, and

$$SNR = 10 \cdot \log_{10}\left(\frac{SinusoidalComponentEnergy}{NoiseComponentEnergy}\right) \quad (5)$$

the bandwidth (BW) and center frequency of the frequency region over which the residual is spectrally shaped. The synthesis of vowels /a/, /i/, and /u/, was attempted for fundamental frequencies of 100 Hz, 200 Hz, and 300 Hz. The BW was varied in steps of 200Hz (100Hz on either side of the center frequency), and SNR was varied in steps of 5dB. Spectrally shaped noise having a particular bandwidth was shifted around in frequency in steps of 100 Hz to select the ideal center frequency location. The SNR was increased till the tonal component with changed timbre became prominent and/or became similar to the reference sound. A rough estimate of the values of BW, SNR and frequency location of added noise, for synthesizing a percept of



(a) (b) (c)

Fig 1. A single partial having center frequency 500Hz synthesized using Kelly's model. The BW parameter value for all three cases was 200Hz. The k parameter values for (a), (b), and (c) were 0, 0.4 and 0.8 respectively. The increase in spectral line widening with an increase in k value can be observed.

roughness, was thus achieved by the above method, and further refined values of BW and SNR were found in a few cases. The best case vowels synthesized using the spectral modeling synthesis method and having a percept of roughness similar to that in the reference vowels with 1.5% jitter were thus obtained.

4.2 Synthesis of rough vowels using Kelly's method

Kelly's model was experimented with for synthesis of rough vowels /a/, /u/, and /i/, at fundamental frequencies 100Hz, 200 Hz, and 300 Hz. The frequency region between 0 to 4000Hz was divided into adjacent 600 Hz bands, and noise was added in each frequency band separately with varying SNR, here SNR is defined as in Eq. 5 for sinusoidal and noise components belonging to each individual partial. The noise was added in such relatively small bands of frequency to provide an idea of how the amplitude modulated noise in different frequency regions affects the timbre of the tonal component.

Taking a small BW value such as 10 Hz was found to give a quivering percept to the synthesized vowel. On the other hand taking a relatively larger BW value such as 100 Hz was found to lead to a timbre change in the tonal component of the synthesized vowel. Adding noise by Kelly's method to the partials in the frequency band 1200 Hz to 1800 Hz and in the higher frequency bands was found to change the timbre in the direction of breathiness. Hence noise was added only to the partials having center frequency less than 1200 Hz. This, however, did not ensure that adding noise in the lower frequency region would always lead to roughness. In particular, the amount of noise quantified by the SNR, which is determined by the parameter κ seems to play a role in determining the amount of timbre change, and hence the perceived quality. The SNR was increased from 0 dB in steps of 2.5 dB and noise was added in different frequency regions individually to achieve an optimum timbre change towards roughness.

Suitable frequency regions for adding noise ~~to~~ were thus determined. As a next step, adding noise together in the different frequency regions was attempted. As expected there was an overall timbre change in the tonal component towards roughness. The parameter values arrived at for the synthesis of such rough vowels indicate a consistent trend of lower SNR for the noise added to the higher partials as compared to that added to the lower partials. For the synthesized vowels so obtained, the level of perceptual fusion was found to be quite high.

4.3 Informal listening test A: comparison of perceptual fusion

Best case vowel samples, synthesized using the spectral modeling synthesis method, and Kelly's model, of the three vowel types at fundamental frequencies of 100Hz, 200 Hz, and 300Hz, with a percept of roughness similar to that in the reference vowels with 1.5% jitter were presented to a listener. The listener was asked to rank the two samples in each case, based on a decrease in the loudness of the noise source in the synthesized sound, and a corresponding change in the perceived timbre of the tonal component. The listener was allowed to listen to the sounds any number of times, and was given the option of not assigning a rank. Except for the cases of vowel /a/; pitch 100 Hz and 200 Hz, the listener in all other cases selected the samples synthesized using Kelly's model over those synthesized using the SMS method. The noise component in many of the samples synthesized using SMS method was reported to be loud, as compared to that in the samples synthesized using Kelly's model.

4.4 Determining Kelly model parameters for each partial

In contrast to the SMS method, Kelly's method enables a control over the noise added to each partial and this makes it more flexible than the SMS method. As such, in this section, we investigate assigning κ and BW values to each individual partial in order to synthesize rough vowels. Kelly uses constant values of κ and BW for all partials in the frequency region of interest towards synthesis of breathy sounds. In this context, it was found that assigning constant values of κ and BW to the harmonics did not always seem to synthesize the required percept of roughness and efforts were made to synthesize the same by assigning *frequency dependent* values of κ and BW to the partials. The partials were first assigned parameter values linearly varying with frequency. For a particular set of linearly varying values, if the synthesized sound was close to the reference sound having certain amount of jitter, then the values were varied slightly to match the reference sounds more closely. These linearly varying values were characterized by certain 'trends', such as increasing, decreasing as well as constant, with partial center frequency. One such trend, that of decreasing BW values with the partial center frequency and increasing κ values with the partial center frequency, was found to give a roughness percept similar to that in the reference synthesized vowels.

4.5 Informal listening test B: trends in Kelly model parameters

The vowels /a/, /u/, and /i/; pitch 100 Hz, 200 Hz, and 300Hz with a percept of roughness similar to that in the reference vowels with 1.5% jitter were synthesized using Kelly's model by assigning suitable parameter values with three trends: 1.decreasing BW values, and corresponding increasing κ values with the partial center frequencies; 2. equal BW and κ values for all partials; and 3. decreasing BW and κ values with the partial center frequencies. Two listeners were asked to rank the samples based on similarity with the reference sound, and based on naturalness of the synthesized

sound for the three vowels, and pitch values of 100 Hz, 200 Hz and 300 Hz. The aim of this listening test was to gauge the naturalness of the synthesized vowels, and if possible, to rank the samples in order to come up with one of the three trends as the most suitable one for assigning model parameter values towards the synthesis of rough vowels. The listeners were allowed to listen to the samples from a graphical user interface any number of times, and were given the choice of not ranking any of the samples, indicating either equal percept between that sound and some other sound, or a perceived unnaturalness in the sound. The samples presented to the listener were synthesized using parameter values corresponding to a target percept in the respective reference vowels having 1.5% jitter. The first listener almost always ranked the vowels synthesized using a trend of decreasing BW values and corresponding increasing κ values, as first choice, and the samples synthesized using constant values of k and BW as second choice. The second listener, made similar rankings for the vowels /u/ and /a/. This listener was unavailable for the vowel /i/, and a substitute listener was asked to continue the test. This listener too made similar rankings for the vowel /i/. The first listener remarked that though the percept of roughness was present in the best case synthesized vowels; on critical listening, he could make out the difference as being slightly different kinds of perturbation in the synthesized and reference vowels. The second listener and the substitute listener both remarked that the best case synthesized vowels were very similar to the reference vowels.

5. Results and Discussion

The results of the two listening tests conducted provide useful insights into the synthesis models used, as well as into the perceptual attributes of the synthesized sounds.

The results of the first listening test indicate a greater level of perceptual fusion in the rough vowels synthesized using Kelly's method as compared to the ones synthesized using the SMS method. That the listener found the noise component in the vowels synthesized using SMS particularly loud indicates that while the traditional sines+noise model provides for the inharmonic content of the synthesized sound, the model it adopts for the same is inadequate, especially for the synthesis of rough vowels. On the other hand, the better perceptual fusion achieved in the vowels synthesized using Kelly's model suggests that the technique of bandwidth enhancement, which has been shown by Kelly to give high fidelity synthesis of breathy sounds, is also capable of synthesis resulting in natural sounding rough vowels.

The second listening test further examined the rough vowels synthesized using Kelly's model with a finer set of associated parameter values. The results indicate that while assigning a constant set of parameter values to a particular group of partials seemed to work in the case of synthesis of breathy sounds as reported by Kelly, the same is not the case in the synthesis of rough vowels. In particular, listener preferences revealed that the rough vowels synthesized using a trend of decreasing bandwidth values of the noise, and that of increasing values of κ, both varying with increasing partial center frequency have a percept closest to that in the reference rough vowels. An example of these trends is illustrated in Fig. 2 and Fig. 3 for the synthesis of a percept of roughness similar to that in a reference rough vowel having 1.5% jitter for the vowel /a/.

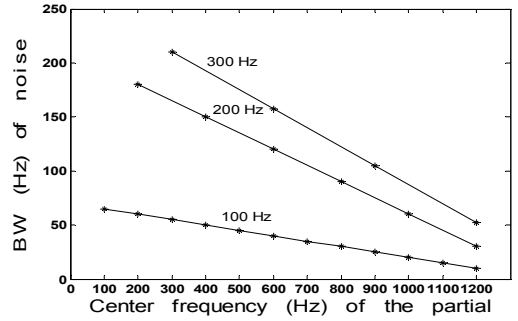


Fig 2. BW values assigned to the partials for synthesis of a percept of roughness corresponding to 1.5% jitter in the vowel /a/, using Kelly's model. Pitch values were 100Hz, 200Hz, and 300Hz. The decreasing trend in the parameter values can be observed.

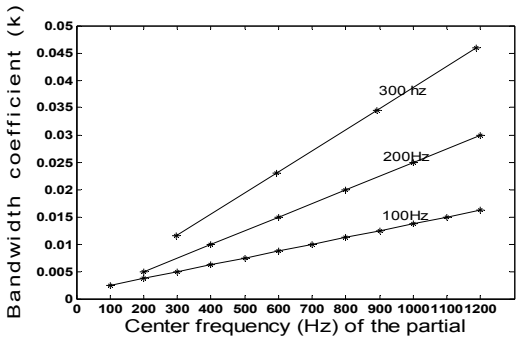


Fig 3. κ Values assigned to the partials for synthesis of a percept of roughness corresponding to 1.5% jitter in the vowel /a/, using Kelly's model. Pitch values were 100Hz, 200Hz, and 300Hz. The increasing trend in the parameter values can be observed.

Besides these overall trends followed by the κ and BW parameter values, the κ and BW values assigned to a particular vowel with a higher pitch were found to be greater than those for the vowel with a lower pitch. These trends in the model parameter values appear consistently for all three vowels that were synthesized. These trends suggest the possibility of a

much greater and easier control over voice quality modifications in the synthesized vowels than that possible using traditional sine+noise models. The vowels synthesized using the present set of values seem very similar to the rough vowels synthesized using the production model and removing the constraint of linear trends might improve the percept in the synthesized vowels slightly. However, it remains to be seen whether the loss of a possibly easy and predictable control over the voice quality attributes in the synthesized vowels, as suggested by the trends shown in Fig 2 and Fig. 3, by the removal of such a linear constraint will be worth the improvement. The trends in the model parameter values are also expected to prove useful in devising techniques for the analysis of natural rough vowels.

6. Conclusion

The bandwidth enhanced sinusoidal method of synthesis has been found to be suitable for the synthesis and control of voice quality attributes in vowels. Imposing the constraint of linear variation in model parameter values with partial center frequency leads to the emergence of certain preferred trends in model parameter values and these trends suggest easily controllable voice quality modifications in the synthesis of rough vowels. The vowels synthesized using Kelly's model are found to have a higher degree of perceptual fusion than those synthesized using the spectral modeling synthesis method and hence are perceived to be more natural. Future work will be directed towards explaining the correlation between perceived roughness and synthesis parameters based on available models of auditory perception. Finally, for the incorporation of the obtained results in a speech synthesis system, it is desirable to develop the corresponding analysis methods for the estimation of model parameters from natural speech.

7. References

[1] C. Drioli, G. Tisato, P. Cosi & F. Tesser, Emotions and Voice Quality: Experiments with Sinusoidal Modeling, *Proc. of the ISCA Tutorial and Research Workshop on Voice Quality: Functions, Analysis and Synthesis (Voqual'03)*, 2003, 133-138.

[2] J. Hillenbrand, Perception of aperiodicities in synthetically generated voices, *J. Acoust. Soc. Am, 83*(6), 1988.

[3] P. Murphy, Spectral characterization of jitter shimmer and additive noise in synthetically generated voice signals, *J. Acoust. Soc. Am*, *107*(2), 1999.

[4] D.G. Childers & C.K. Lee, Vocal quality factors: Analysis, synthesis, and perception. *J. Acoust. Soc. Am*, *90*, 1991, 2394 -2410.

[5] R.J. McAulay & T.F. Quatieri, Speech analysis/synthesis based on a sinusoidal representation, *IEEE Transactions on Acoustics, Speech and Signal Processing, 34*(4). 1986.

[6] X. Serra, Musical sound modeling with sinusoids plus noise, *Musical Signal Processing* (Swets & Zeitlinge, 1997).

[7] K.R. Fitz, The reassigned bandwidth enhanced method of additive synthesis, *PhD thesis, University of Illinois, Urbana-Champaign,* 1999.

[8] D. Hermes, Synthesis of breathy vowels: Some research methods, *Speech Communication (10),* 1991, 497-502.

[9] C. Fales & S. McAdams, The fusion and layering of noise and tone: implications for timbre in african instruments, *Leonardo Music Journal (4)*, 1994, 69-77.

[10] P. Johansson, Perceptual Fusion of Noise and Complex Tone by Means of Amplitude Modulation, *Masters thesis, Department of Speech, Music, and Hearing, KTH*, 2002.

[11] R. Veldhuis, A computationally efficient alternative for the Liljencrants-Fant model and its perceptual evaluation, *J. Acoust. Soc. Am*, *103*(1), 1998, 566–571.

[12] A.N. Lalwani & D.G. Childers, Modeling vocal disorders via formant synthesis, *Proc. of International Conference on Acoustics, Speech and Signal Processing*, 1991, 505–508.