FLEXIBLE AND EFFICIENT HARMONIC RESYNTHESIS BY MODULATED SINUSOIDS

Jonathan Teutenberg[1], *Catherine I. Watson*[2]

[1]Department of Computer Science
[2]Department of Electrical and Computer Engineering
University of Auckland

ABSTRACT

An harmonic plus noise system is presented that uses interpolation across smooth harmonic and noise spectral surfaces as an alternative to copying the spectral envelope from analysis instants. This combines the benefits of isolating the harmonic and noise parts of speech, and simplifying the application of complex transformations to the vocal tract filter. This has been coupled with a method of resynthesising the harmonic part of speech based on modulated sinusoids that more accurately models the harmonic component, and is half the time complexity of traditional overlap and add synthesis. Both the general case of synthesising across arbitrary voiced segments, and a simpler case for synthesising between pitch synchronous instants are detailed. The necessity of two dimensional phase unwrapping is discussed, and a solution presented. This approach is shown to result in high quality modified speech in a perceptual test comparison with the STRAIGHT system.

1. INTRODUCTION

Modification of acoustic properties of speech is required in such tasks as voice conversion, accent modification and speech synthesis. Our interest is in the area of accent modification. As in voice conversion this involves transforming speaker characteristics, however in accent modification the transformations are restricted to pronunciation and prosodic aspects. Age, sex, vocal tract length and articulatory flexibility are expected to remain unchanged after accent modification, in comparison to voice conversion where wholesale changes to a voice are made. As such, accent modification requires greater flexibility in the representation of the source and filter than broad voice conversion approaches.

Previous work into accent modification and the modification of pronunciation have transformed the filter of the source/ filter model by methods that include LPC pole rotation [13] and frequency warping of smoothed spectral surfaces [1]. The smooth spectral surface representation of the vocal tract used by the STRAIGHT system [5] is appealing for the ease with which complex transformations can be applied such as in [1]. Since for continuous surfaces, modifications to duration, formant frequencies, and pitch can be simply described by affine transforms. One drawback to this representation however, is that it does not explicitly separate the voiced and noise parts of speech. The transformations we are considering in regards to accent modification focus on the voiced part of speech and it can be advantageous to isolate the noise part from these transformations. For example, when managing the coarticulation between vowels and fricatives, being able to isolate modifications to the voiced part can eliminate unwanted distortion to the noise component.

Harmonic/stochastic methods of speech analysis [6, 10] divide the speech into harmonic (voiced) and noise parts, allowing each to be modified independently at any given frame of an utterance. Existing implementations of the Harmonic Plus Noise Model (HNM) synthesise speech based on instantaneous spectral envelopes. These allow smooth interpolation across frequencies, but across time these implementations take copies of the envelope at the nearest analysis instant so no interpolation is performed. Without this interpolation, a surface based on the sequence of spectral envelopes will not be continuous, and the general approach to modifications based on affine transforms is no longer possible. It was noted [11] that choosing to not interpolate across time when modifying prosody was merely an implementation decision. This paper seeks to determine whether a combination of an harmonic/stochastic system with a smooth surface spectral representation will provide high quality speech.

In this study a system is implemented that combines the harmonic/stochastic approach with interpolation over smoothed surfaces as in STRAIGHT. For evaluation purposes this can be viewed as either an extension of an harmonic/stochastic model by the inclusion of smoothed surface representation, or as a modification of a STRAIGHT-like model to separate voiced and noise components. As at the time of writing the harmonic/stochastic implementation described in [11] was not freely available, we have chosen to take the latter view and compare the perceived quality of speech based on our approach with that of STRAIGHT. Through a perceptual test we show that the perceived quality of speech produced by a system combining both harmonic/stochastic and smoothed surfaces is comparable to that of STRAIGHT. We also introduce a synthesis method for the harmonic component based on modulated sinusoids that more closely models the voiced part of the original signal and has less than half the time complexity of direct overlap and add synthesis.

In this paper 'traditional sinusoidal models' refers to those of McAulay and Quatieri [8] and their derivatives, including STRAIGHT. We use this term to contrast with harmonic/ stochastic models that explicitly separate the noise from the harmonic parts of speech.

2. METHOD

Harmonic/stochastic representations assume that speech has an voiced part made up of harmonically related sinusoids with amplitude and frequency that vary slowly over time, and a noise part. Analysis provides snapshots of the harmonics' parameters by examining short frames of speech. The sys-

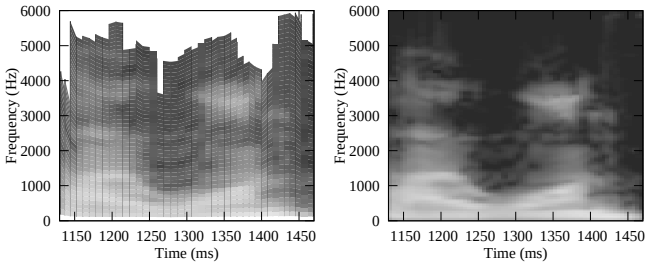

Figure 1: The harmonic envelopes of speech from HNM_1 analysis (left), and the resulting smooth spectral surface (right). Note that higher magnitude is shown with lighter shades.

tem described in this paper performs analysis as in HNM_1 [10] and will not be further discussed here. After analysis the harmonic envelopes are used to create a magnitude surface $A(t, f)$ and a phase surface $\Phi(t, f)$ by cubic interpolation. An example of a magnitude surface is shown in Figure 1 on the right hand side. Here t is time in seconds and f is frequency in Hertz. It should be noted that these figures are not spectrograms, but rather visualisations of the models of speech. Thus the apparent coarse resolution is due to the fact that there is a single sample per pitch period in time, and one sample per f_0 step in frequency.

For the harmonic component, existing harmonic/stochastic methods have followed traditional sinusoidal models in inverting the analysis process at synthesis time. The overlapping speech frames are reconstructed, and added together to generate a resynthesised signal. It is our claim that as the harmonic part of the signal has been isolated, it is both more accurate and efficient to resynthesise it in line with the model assumptions: as continuous sinusoids with slowly varying amplitudes and frequencies rather than short, overlapping frames of harmonic sinusoids.

2.1 Overlap and Add Synthesis

Both STRAIGHT and HNM rely on an overlap and add procedure to synthesise speech. Overlap and add synthesis is simply the inverse of the analysis process. Each short frame is reproduced separately as the sum of stationary, harmonically related sinusoids. In this work, magnitudes and phase offsets of each sinusoid are taken by sampling the smooth magnitude and phase surfaces A and Φ at harmonic intervals of F_0 at each pitch-synchronous analysis instant. In contrast, the original HNM implementation resamples the magnitude and phase envelope of the nearest analysis instant - that is, it does not interpolate across the time axis. As depicted in Figure 2(a), the sum of the overlapping frames produces the resynthesised signal.

2.2 Modulated Sinusoids Synthesis

The method of synthesis by continuous sinusoids modulated in amplitude and frequency (that is, both amplitude and frequency are continuous functions of time) used here and shown in Figure 2(b) is similar to that used by the original phase vocoder [2]. However, the phase vocoder's component sinusoids were at fixed frequencies, and were not harmonically related. Extending the phase vocoder, harmonic coders based on the short-time Fourier transform such as [9, 7] have

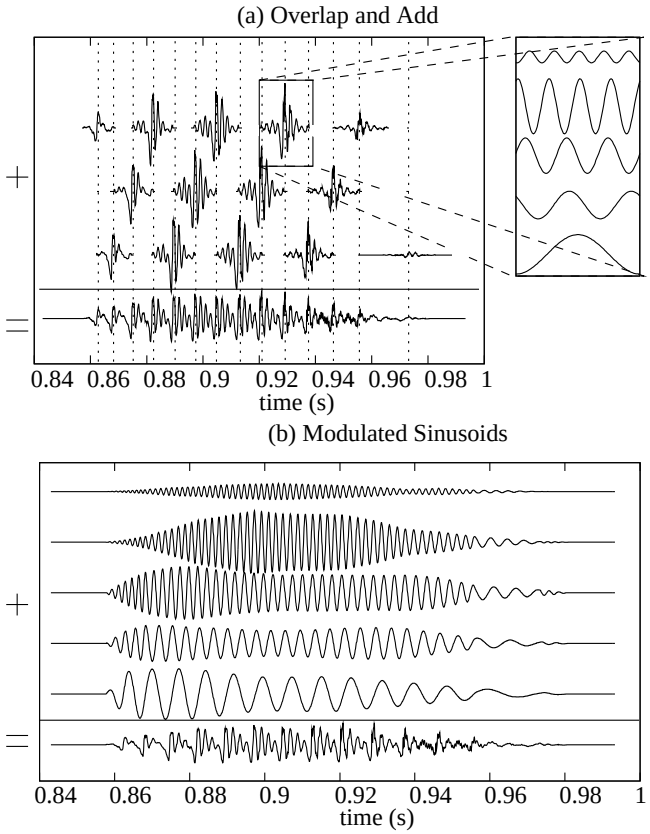

Figure 2: Two methods of synthesising the harmonic part of speech. Diagram (a) shows the overlap and add method, where 14 overlapping frames make up the speech segment. The right panel shows the first five harmonics of one frame, with fixed amplitude and phase offset. Diagram (b) shows the first five harmonics when synthesising using modulated sinusoids.

constrained the sinusoids to harmonic frequencies at the expense of continuity across time. In this work we relax the harmonic constraint, synthesising the underlying near-harmonic sinusoids independently of one another, and allowing perfect continuity across time.

Values of the harmonic part $h(t)$ are computed between two analysis points, given as time instants τ_i, τ_{i+1}. The fundamental frequency contour is represented by a continuous function $f_0(t)$, typically providing linearly or cubicly interpolated values between analysis points.

In this general case, the value of the harmonic part $h(t)$ is given by

$$h(t) = \sum_{k=1}^{L} a_k(t) cos(2\pi k \varepsilon(t) + \phi_k(t)) \quad (1)$$

where L is the highest harmonic in the voiced segment and the instantaneous amplitudes $a_k(t)$ and phase offsets $\phi_k(t)$ are taken directly from the surfaces A and Φ as

$$a_k(t) = A(x, f_0(t)k) \quad (2)$$
$$\phi_k(t) = \Phi(t, f_0(t)k) \quad (3)$$

The function ε provides the number of pitch periods between

the first analysis instant τ_i and t, given by

$$\varepsilon(t) = \int_{\tau_i}^{t} f_0(y).dy \quad (4)$$

When the number of harmonics differs between analysis instants, the missing harmonics up to L are treated as having zero magnitude.

In this work, τ_i and τ_{i+1} are placed at adjacent excitation instants. As these span a single pitch period, the fundamental frequency can be fixed and simple linear interpolation used to calculate amplitude and phase. The function ε in Equation 4 then provides the location of t as a fraction of the distance between the two analysis points τ_i,τ_{i+1}, thereby simplifying to

$$\varepsilon(t) = \frac{t - \tau_i}{\tau_{i+1} - \tau_i} \quad (5)$$

and the instantaneous amplitudes $a_k(t)$ and phase offsets $\phi_k(t)$ can be calculated by

$$a_k(t) = \varepsilon(t)A(\tau_{i+1}, f_0(\tau_{i+1})k) + (1-\varepsilon(t))A(\tau_i, f_0(\tau_i)k) \quad (6)$$
$$\phi_k(t) = \varepsilon(t)\Phi(\tau_{i+1}, f_0(\tau_{i+1})k) + (1-\varepsilon(t))\Phi(\tau_i, f_0(\tau_i)k) \quad (7)$$

As all n lie within the same pitch period, the instantaneous fundamental frequency f_0 is simply taken from the average of the two adjacent pitch periods.

The rate of change in phase is equivalent to frequency. Thus when the offset based on the phase surface varies over time, the sinusoidal components in Equation 1 are no longer at integer multiples of the fundamental frequency. In practice phase changes slowly over time, and the sinusoidal components are therefore nearly harmonically related, in the same way that the source speech is nearly periodic.

2.2.1 Phase Unwrapping

Phase unwrapping is a method for smoothing the phase envelope at an analysis instant. By adding or subtracting integer multiples of 2π to phase values that are adjacent on the frequency axis, the slope, or group delay, of the envelope is minimised without losing information. The minimising of slope becomes significant when resampling the phase envelope, as interpolating between highly disparate phase values produces results that appear random.

When synthesising by modulated sinusoids, the phase surface is being sampled in time as well as frequency. To ensure a smooth variation in phase offset, the phase surface must be unwrapped in two dimensions. In Figure 3, the discontinuities over frequency in the phases (top) are removed by phase unwrapping. However when this is only performed across the frequency axis, discontinuities still occur across time, particularly at around 1400ms in the example. Our approach is to process phase envelopes in time order, and select the multiple of 2π to adjust phase at a frequency f that minimises a weighted sum of the difference between the previous frequency entry in the same envelope (as in standard phase unwrapping) and the difference from the value of the previous envelope in time, at frequency f. The results of this process are shown in Figure 3 (bottom), where the discontinuities in time can be seen to be eliminated. As in Figure 1, the apparent coarse spectrograms are due to the single sample per pitch period approach of HNM$_1$.

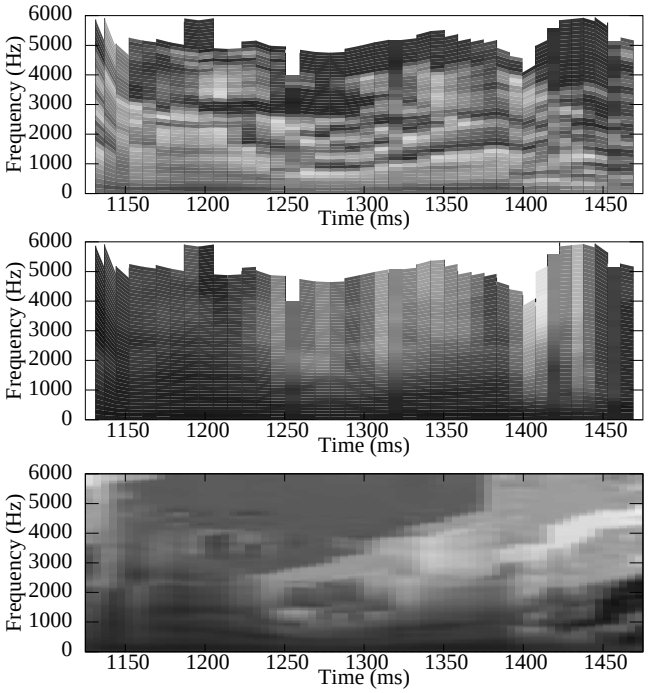

Figure 3: Approaches to phase unwrapping. Original phase envelopes (top), unwrapped phase envelopes (middle), and phase surface unwrapped in both time and frequency (bottom).

The two dimensional phase unwrapping proposed here is not required by HNM, as its overlap and add approach does not interpolate across time. In STRAIGHT, the source phase information is discarded, and minimum group delay is assumed during resynthesis.

3. EVALUATION

3.1 Experimental Setup

Five utterances from five speakers of the Kluwer dataset [3] were used for evaluation. Each utterance consists of a read sentence, and all speakers were males with Australian English accents.

The efficiency of the two synthesis methods were compared by taking the mean time to synthesise the utterances 100 times. The two methods were implemented in Java, running on a single 2GHz processor on a GNU/Linux system. The methods were run completely separately to avoid interference by the Just-In-Time compiler. This was repeated ten times to provide a mean and standard deviation.

To assess how well the two methods model the voiced part of speech, the resynthesised harmonic (voiced) part was removed from the original signal to give a residual (noise component). The signal to noise ratio of the harmonic to residual for each approach over the utterances was compared. It is assumed that as speech is not truly periodic some of the voiced part is contained in the residual signal. Thus a higher signal to noise ratio would indicate a more accurate modelling of the utterance.

Method	Running Time (s)	SNR (dB)
Overlap and Add	29.9 +/- 0.5	5.75
Modulated Sinusoids	12.7 +/- 0.4	6.44

Table 1: Signal to noise ratio of harmonic parts of speech and mean running times of synthesis methods.

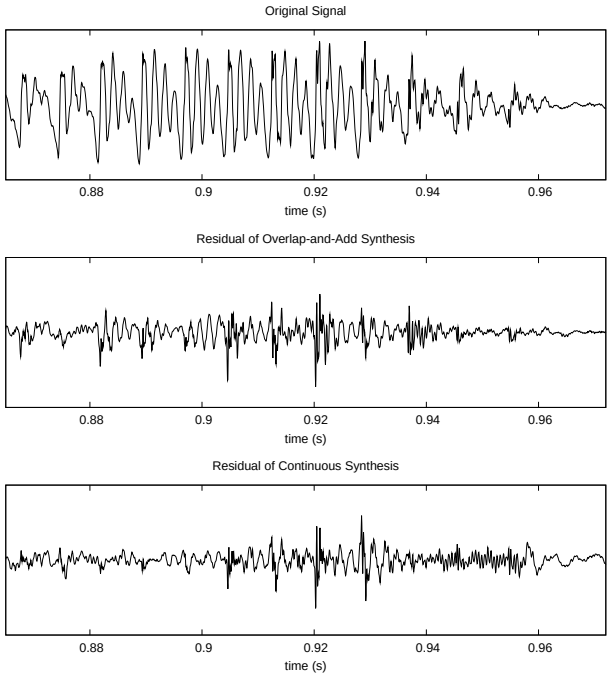

Figure 4: Original speech segment (top) and residuals after removing overlap and add (middle) and modulated sinusoid harmonic part (bottom). Note that the amplitude axis has the same scale in all three plots.

3.2 Results and Discussion

As can be seen in Figure 2(a) when synthesising by the overlap and add method, each sample appears in approximately two windows. So for each sample around $2L$ cosine evaluations need to be made, whereas synthesising with modulated sinusoids use approximately L such evaluations. Table 1 shows the results of running the two methods (see column 2). Synthesising with modulated sinusoids is slightly faster than the doubling of speed that was expected, most likely due to the elimination of the overheads of managing multiple frames and applying windowing functions.

It should be noted that methods of speeding up the synthesis of frames based on harmonically related sinusoids have shown much greater efficiency gains. Some of these, such as those based on recurrence relations, are no longer applicable to modulated sinusoids. However the most promising of these, Delayed Multi-Resampled Cosines [12] which replaces each expensive cosine evaluation with an array lookup, can be adapted to work in conjunction with the modulated sinusoids.

Table 1 also shows the signal-to-noise ratio between the harmonic and noise parts for both methods (see column 3). The modulated sinusoids produce an harmonic component with significantly higher SNR. The significance was tested using a pairwise Student's T-test over each sample. This was found to be significant with $p < 10^{-7}$. Figure 4 shows an example of the residuals of the two methods. The residual of the modulated sinusoids can be seen to have lower amplitude than that of overlap and add across the majority of the speech segment.

Method	A (MOS)	B (MOS)	C (MOS)
Original	4.47*	-	-
Overlap and Add	3.33	3.13	2.31
Modulated Sinusoids	3.42	3.22	2.07*
STRAIGHT	(4.47)	3.04	2.4

Table 2: Mean opinion scores of speech quality resulting from experiments A,B and C. A has unmodified speech rate and pitch, B is faster and lower pitch, and C is slower and higher pitch. Entries marked * are statistically significantly different from other entries in the same column. STRAIGHT's quality for experiment A was assumed to be the same as that of the original.

4. PERCEPTUAL TEST

A perceptual test was performed investigate the perceived quality of resynthesised speech using magnitude and phase surfaces by modulated sinusoids and overlap and add. At this time the original HNM implementation is not available, so the two methods of synthesis presented here are compared to STRAIGHT as a high quality bench mark.

4.1 Experimental Setup

Three versions of five utterances were resynthesised using the two approaches and STRAIGHT for comparison. The first version (experiment A) is copy synthesis where the two methods were compared to the original speech. The second (experiment B) had 20% faster speech rate and 30% higher pitch, with the two methods and STRAIGHT being compared. The third version was 50% slower and had 50% higher pitch, also comparing with STRAIGHT. Thus experiment B compresses the surfaces in both frequency and time, and experiment C expands the surface in both dimensions. These were then presented to 12 listeners who were asked to rate the quality of the resynthesised utterances on a scale from 1 to 5.

4.2 Results and Discussion

The results of the surveys are shown in Table 2 as mean opinion scores. The STRAIGHT system was not tested on copy synthesis as previous studies [4] have reported no perceptible difference between the results of copy synthesis and the original speech. Experiment A showed that there was no significant difference between the quality of the overlap and add and modulated sinusoids approaches. This also showed that the resynthesised speech is of statistically significantly lower quality than the original according to a pairwise Student's T-test. So the presented system introduces some perceptible degradation in quality during copy synthesis.

However, after pitch and time-scale modifications the system described here produces speech of a similar quality to that of STRAIGHT. The exception in the experimental results is that of modulated sinusoids on experiment C (raising the pitch and extending the duration). This was due to an error in the pitch marking used for analysis in one of the example sentences that resulted in relatively large changes in

phase offset between the more closely spaced pitch marks. With this example excluded, the MOS for the experiment was 2.28, and not significantly different from either of the other approaches. A further test was run that showed that this effect can be eliminated by constraining the change in phase offset over time.

In experiment C, the speech with higher pitch and lower rate of speech was an unnatural method of speech production - somewhat falsetto-esque. As such, the lower MOS scores in comparison to the unmodified resynthesised speech are a conflation of both the quality *and* the unnaturalness of the resynthesised speech. The speech in experiment B, with lower pitch and higher rate of speech was less unusual, and the MOS scores for this experiment are felt to be a more accurate measure of the quality of the modified speech.

Examples of speech segments used in the perceptual tests can be found at http://www.cs.auckland.ac.nz/~jteu004/EUSIPCO/

5. CONCLUSION

We have introduced a method for resynthesising the harmonic component of speech by continuous modulated sinusoids. Both the general case of synthesising across arbitrary voiced segments, and a more efficient case where synthesis occurs between pitch synchronous instants have been detailed, and a method for two dimensional phase unwrapping presented. This approach has been shown to halve the time complexity in comparison to overlap and add synthesis. The signal-to-noise ratio of the harmonic part to its residual was also shown to be significantly reduced, indicating that the voiced component of speech is being more accurately modelled.

An analysis/synthesis system using this approach, combined with harmonic/stochastic analysis and the flexibility of a STRAIGHT-like representation that interpolates over not only frequency, but also time. Our perceptual tests have shown that after pitch and time modifications this approach produces speech that is of comparable quality to that of STRAIGHT, with the advantage that the noise and harmonic parts are explicitly separated in the model.

REFERENCES

[1] P. F. Assmann and T. M. Nearey. Frequency shifts and vowel identification. In *International Conference of Phonetic Sciences*, pages 1397–1400, 2003.

[2] J. Flanagan and R. Golden. Phase vocoder. In *Bell System Technical Journal*, volume 45, pages 1493–1509, 1966.

[3] J. Harrington and S. Cassidy. *Techniques in Speech Acoustics*. Kluwer Academic Publishers, Dordrecht, 1999.

[4] H. Kawahara, I. Masuda-Kasuse, and A. de Cheveigne. Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a reptitive structure in sounds. In *Speech Communication*, volume 27, pages 187–207, 1999.

[5] H. Kawahara, M. Morise, T. Takahashi, R. Nisimura, T. Irino, and H. Banno. Tandem-straight: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, f0 and aperiodicity estimation. In *International Conference on Acoustics, Speech and Signal Processing*, pages 3933–3936, 2008.

[6] J. Laroche, Y. Stylianou, and E. Moulines. HNS: Speech modification based on a harmonic+noise model. In *International Conference on Acoustics, Speech and Signal Processing*, volume 2, pages 550–553, 1993.

[7] J. Marques, L. Almeida, and J. Tribolet. Harmonic coding at 4.8 kb/s. In *International Conference on Acoustics, Speech and Signal Processing*, volume 1, pages 17–20, 1990.

[8] R. McAulay and T. Quatieri. Speech analysis/synthesis based on a sinusoidal representation. In *International Conference on Acoustics, Speech and Signal Processing*, volume 34, pages 744–754, 1986.

[9] R. McAulay and T. Quatieri. Sinusoidal coding. In W. Kleijn and K. Paliwal, editors, *Speech Coding and Synthesis*, 1995.

[10] Y. Stylianou. Decomposition of speech signals into a deterministic and stochastic part. In *Fourth International Conference on Speech and Language Processing*, volume 2, pages 1213–1216, 1996.

[11] Y. Stylianou. Concatenative speech synthesis using a harmonic plus noise model. In *3rd ESCA/COCOSDA Workshop on Speech Synthesis*, pages 261–266, 1998.

[12] Y. Stylianou. A simple and fast way for generating a harmonic signal. In *Signal Processing Letters, IEEE*, volume 7, pages 111–113, 2002.

[13] Q. Yan, S. Vaseghi, D. Rentzos, and C. Ho. Analysis by synthesis of acoustic correlates of british, australian and american accents. In *International Conference on Acoustics, Speech and Signal Processing Proceedings*, volume 1, pages 637–640, 2004.