DETECTING NONLINEARITIES IN SPEECH SOUNDS USING THE BICOHERENCE[1]

J W A Fackrell[2], S McLaughlin

Signals and Systems Group, Dept of Electrical Engineering,
University of Edinburgh, Edinburgh, EH9 3JL, UK. email: sml@ee.ed.ac.uk

A signal property called quadratic phase coupling (QPC) is the signature of a quadratically nonlinear signal production mechanism, and tools from the field of higher-order statistics (HOS) can be used to detect this property. This paper describes an investigation into the use of a HOS tool called the bicoherence for detecting QPC in continuant speech signals. A detector for QPC is described, and applied to a number of speech sounds from a specially-recorded database. The results are consistent with the hypotheses that unvoiced fricatives are Gaussian, and that voiced continuants (vowels, voiced fricatives and nasals) do not exhibit QPC. Therefore it seems that quadratic models of speech signals are unlikely to offer improvements over existing, linear, models.

1. INTRODUCTION

Most speech synthesizers are based on the linear model of speech production, and this model, although mathematically simple, is capable of describing a great many of the properties of human speech. Indeed much of the speech generated by linear synthesizers is *intelligible*, which for many applications is all that is required. However, synthesized speech does not sound very *natural*, and this problem has led several researchers to investigate the possibilities of using *nonlinear* models for speech production. Kubin [1] provides a comprehensive review of these.

One of the problems encountered in attempts to develop nonlinear models for speech production is that of choosing what type of nonlinear model is appropriate. A viable first step in such work is to try to identify the signatures of certain types of nonlinearity in human speech signals. If these signatures are found, then they provide clues to what type of nonlinear speech production model is appropriate.

This paper is concerned with the identification of the signature *quadratic nonlinearity*, which is just one particular type of nonlinearity. Signals which are produced by models containing *quadratic* nonlinearities exhibit a property called quadratic phase coupling (QPC). The idea is that if QPC is detected in real speech, then models including quadratic nonlinearities may be good ones for speech signals.

This QPC phenomenon can be detected using measures from the field of higher-order statistics (HOS). HOS techniques can provide information about signal phase, nonlinearities and deviations

[1]During the period in which this work was carried out, Justin Fackrell was supported by a CASE studentship from EPSRC in collaboration with BT Laboratories, Martlesham Heath. Steve McLaughlin is supported by the Royal Society.

[2]Now at Lernout & Hauspie Speech Products NV, Sint-Krispijnstraat 7, 8900-Ieper, Belgium.

from Gaussian probability densities which are *hidden* from conventional techniques such as spectral analysis and autocorrelation analysis [2]. In particular, two frequency-domain HOS measures based on third-order statistics, the *bispectrum* and *bicoherence*, can provide information about QPC[3]. It is stressed that this information is simply unavailable using conventional techniques.

There have been several previous attempts to use these tools in speech analysis. Many HOS applications in speech processing attempt to exploit the "fact" that speech has very different HOS properties to background noise in order to obtain robust estimates of measures usually calculated by conventional techniques (e.g. [4, 5]). This "fact" is supported mostly by evidence from the literature that speech signals have non-zero, and therefore interesting, third- and fourth-order HOS measures. Some of this evidence is relevant to the current study, since it involves the use of bicoherence and bispectra:

Wells [6] reported preliminary results into a study of the pitch-synchronous bispectrum/bicoherence properties of vowels and unvoiced fricatives. It was reported that unvoiced sounds have zero bispectra/bicoherences, but that voiced sounds have non-zero bispectra/bicoherences. These measures were then proposed as detectors for voicing, an application that was later extended by Navarro [7] in a more speech endpoint detector. In related work, Boianov et al [8] reported that the bicoherence appeared to be suitable for the detection of laryngeal pathologies.

Recently, in a more detailed study, Thyssen [9] concluded that the ensemble-averaged bicoherences, computed *pitch-asynchronously*, of vowel sounds showed evidence of coupling primarily at harmonics of the speech fundamental frequency. At high frequencies, the extent of coupling was found to fall if the pitch varied over the record duration, an effect which has also been observed (and explained) in other signals [10].

In some of these applications [7, 8] the bicoherence has been estimated from a single record of data, a practice which can, as recent evidence shows [11, 12, 9], lead to *ambiguous* interpretations. This paper shows how to overcome these ambiguities. In addition, this is the first paper to describe some of the bicoherence properties of sounds other than unvoiced fricatives and vowels.

The paper is structured as follows; Section 2 provides a simple example of a model for generating a signal which exhibits QPC, and illustrates how the bicoherence can be used to detect it. Section 3 then describes a database assembled especially for QPC analysis, and Section 4 describes the results of the application of the QPC-detection tools to the database. Finally Section 5 interprets these results from a nonlinearity detection viewpoint, before suggesting areas of further work.

2. QPC

Figure 1a) shows a simple combination of a linear and a quadratically-nonlinear filter. If the input to this system contains sinusoidal components $\cos(2\pi f_1 + \phi_1)$ and $\cos(2\pi f_2 + \phi_2)$ then it is easy to show [13] that the output will contain the additional frequency components[3] $2f_1$, $2f_2$, $f_1 + f_2$ and $f_1 - f_2$, , with phases $2\phi_1$, $2\phi_2$ $\phi_1 + \phi_2$ and $\phi_1 - \phi_2$. Now, for a stationary signal, the *bispectrum* $B(k,l) = X(k)X(l)X^*(k+l)$ (where $X(k) = |X(k)|e^{j\phi(k,l)}$ is the M-pointdiscrete Fourier transform)

[3] as well as a DC component.

has a magnitude $|B(k,l)| = |X(k)X(l)X^*(k+l)|$ and phase (or *biphase*) $\angle B(k,l) = \Phi(k,l) = \phi(k) + \phi(l) - \phi(k+l)$ and it is easy to see that for this sinusoidal signal, the bispectra $B(f_1, f_1)$, $B(f_2, f_2)$, $B(f_1, f_2)$, $B(f_1, -f_2)$ will have large magnitude, and zero biphase Φ. This property forms the basis of the use of the bispectrum as a detector for QPC, and is described in greater detail elsewhere ([13, 3]). The bispectrum can be estimated by calculating the raw bispectrum $B_i(k,l)$

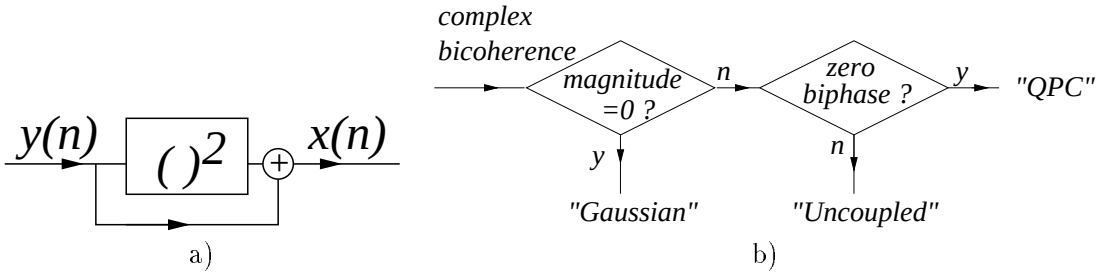

Figure 1: a) Simple system which generates signals with QPC. b) Two part test for QPC.

from some frame i of data (where $i \in \mathcal{K}$, a set of frames with K members), and then averaging over the i raw estimates. The frames can originate from multiple realisations of the signal (*ensemble* averaging), or from a single record divided into equal length frames (*segment* averaging). (However under certain circumstances, there can be a difference between these approaches, and this will be commented on below.)

From an interpretative stance, it is helpful to normalise the bispectrum to form the complex bicoherence $b_c(k,l)$, estimated as [14]

$$b_c(k,l) = \frac{\frac{1}{K}\sum X_i(k)X_i(l)X_i^*(k+l)}{\sqrt{\frac{1}{K}\sum |X_i(k)X_i(l)|^2 \frac{1}{K}\sum |X_i(k+l)|^2}}. \quad (1)$$

where $\sum \equiv \sum_{i\in\mathcal{K}}$. $b_c(k,l)$ is a complex quantity, so is slightly different to previously proposed bicoherence measures [3, 9]). However, its square $b_c^2(k,l)$ is identical to the "squared bicoherence" used elsewere, and its phase $\angle b_c(k,l)$ is the same as the biphase. Thus b_c contains both bicoherence and bispectrum information.

If b_c is estimated by ensemble averaging, then $b_c^2(k,l)$ can be interpreted as the proportion of coupled energy at frequencies k, l and $k+l$. An explicit demonstration of this was recently reported [14]. The same interpretation holds for segment-averaged b_c^2's, but only if the phases of the components concerned are re-randomised in each frame (i.e. there is "phase randomisation" PR [11]).

If the phases are not re-randomised (no PR) , then b_c^2 measures the propotion of *coherent* energy at each bifrequency [4]. In this case a peak in b_c^2 does not necessarily indicate that there is QPC, a result that has been reported also by [12]. In such cases b_c^2 can still be useful in determining which bifrequencies contain 'signal' and which contain 'noise', but the biphase $\Phi(k,l)$ must now be used to test for QPC. If there is QPC then the biphase will be zero. However, if *only* the biphase is tested, there is the risk of detecting QPC at bifrequencies which are dominated by noise rather than signal.

[4] Coherent in the sense that the phase relations between the components remain the same in all frames.

A two-part test for QPC which attempts to solve this problem has been proposed, as shown in Figure 1b). The test is carried out at each bifrequency. Firstly a test is done to see if b_c^2 is statistically different from zero. If it is not, then that bifrequency probably just contains Gaussian noise, and so is not of interest from a nonlinearity detection viewpoint. If b_c^2 is (statistically) nonzero then a test is carried out on the biphase $\Phi(k,l)$ to see if it is (statistically) zero. This time, a zero result indicates that *there is* coupling, and a non-zero result indicates no coupling.

This detector works whether or not the PR assumption is valid, and its properties have been reported in [11, 15]. It has been found from simulations [15] that the detector works well at high SNRs, but at low SNRs ($\approx -6dB$, if $M = 64$) the number of Type II errors (i.e. detections of zero biphase when the biphase is in fact non-zero) can become unacceptably high. The $P(\text{Type II})$ is shown to be [15, 14]

$$P(\text{Type II}) = \text{angular width of biphase acceptance region}/2\pi = \frac{c_\alpha}{2K}\left(\frac{1}{b_c^2(k,l)} - 1\right)/2\pi \qquad (2)$$

where c_α is the (two-sided) critical value for a standardised Normal distribution $N(0,1)$ at significance α, K is the number of data segments, and $b_c^2(k,l)$ is the squared complex bicoherence defined above.

Two useful summary measures will be used to measure the extent of QPC in the speech signals in the database. The first is $\overline{b_c^2}$ - which is the *average* value of b_c^2 over the L bifrequency bins in the region of the bifrequency plane called the Inner Triangle (IT) [13]. $\overline{b_c^2}$ can be interpreted [14] as the average proportion of coupled energy in the IT region. It can be shown [14] that, for ensemble averages, or if PR is valid, the null hypothesis H_0, that the signal is Gaussian, (i.e. noise only, and contains no significant bispectral properties) corresponds to $2KL\overline{b_c^2}$ being approximately centrally-χ^2 distributed with $2L$ degrees of freedom. Thus $\overline{b_c^2}$ can be interpreted in two ways - as a proportion of coupled energy, or a Gaussian test statistic.

The second measure is $\overline{q}$, defined as the proportion of bifrequency bins in the IT at which QPC is detected by the two-part detector described above. Signals which involve a lot of interactions due to quadratic nonlinearity will thus have high $\overline{b_c^2}$ *and* high $\overline{q}$, signals which are deterministic-dominated but uncouopled have high $\overline{b_c^2}$ but low $\overline{q}$ and signals which are noise-dominated will have low $\overline{b_c^2}$.

3. EXPERIMENTAL METHOD

This section describes a database assembled especially for QPC detection analysis. HOS quantities such as the bicoherence generally have much higher variances than estimates of conventional measures such as the power spectrum. In an effort to control this variance, data lengths must be as long as possible. As this paper describes exploratory work, in which the interest lies in reliable identification of quadratic nonlinearities, rather than the production of a new synthesis system, the database was constructed of speech sounds deliberately lengthened by the speakers.

The speech was recorded using an Audio-Technica ATM73a headset microphone and a Laryngograph in a low noise environment. The data was sampled at 22.05kHz using the "Clinassist" package[5],

[5]Developed by Dr Alan Wrench of Queen Margaret's College, Edinburgh.

but the voiced sounds were subsequently subsampled to 11.025 kHz. The speech of 16 volunteers was recorded, and each target phoneme of interest, placed in a word context, was repeated 5 times. The sounds recorded include fricatives (voiced and unvoiced), vowels and nasals. Only the fricative results will be shown in Section 4, for the phonemes (in SAM-PA notation [16], followed by context word) [s "bus"], [T "both"] (unvoiced), [z "buzz"] and [D "loathe"](voiced).

For the voiced sounds, the laryngograph was used to extract 64 samples (after subsampling) after each glottal closure (allowing for an appropriate acoustic propagation delay). Each analysis was performed over 64 frames of data - for the voiced sounds this corresponds to 64 consecutive pitch periods (with $M = 64$ samples extracted from each pitch period) and for the unvoiced sounds this corresponds to a simple segment-averaging estimation over 64 consecutive frames. More details of the choices of estimation parameters can be found in [14].

4. RESULTS

Figure 2 shows scatter diagrams of the $\overline{b_c^2}$ values the unvoiced fricative sounds for all 16 speakers, indexed by sex (M or F) along the bottom. The $\alpha = 0.001$ threshold for signficance is shown, which means that with confidence 99.9%, Gaussian signals should have $\overline{b_c^2}$ below this line, and non-Gaussian signals should have $\overline{b_c^2}$ above it. Clearly most speech sounds/speaker combinations (apart from speaker 6 saying T marked on Figure 2) give results consistent with the unvoiced fricatives being Gaussian. Figure 3 (top) shows the values of $\overline{b_c^2}$ for the voiced fricatives spoken by the same

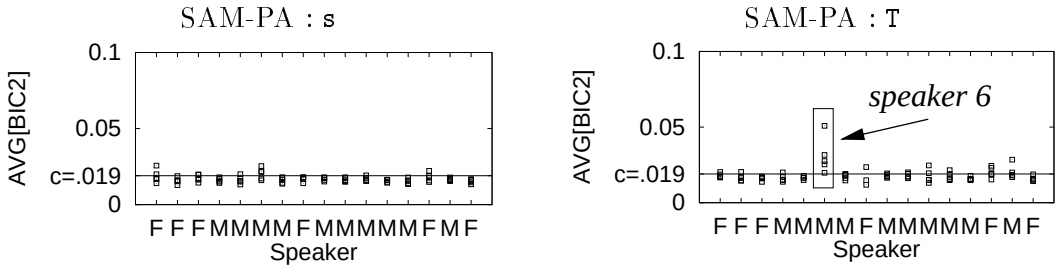

Figure 2: Scatter diagrams showing $\overline{b_c^2}$ for the unvoiced fricatives f (left), T (right) by all 16 speakers. The null hypothesis H_0 that the signal is Gaussian is rejected at the $\alpha = 0.001$ level if $\overline{b_c^2} > c$ the critical level.

16 speakers. Since the PR assumption discussed in Section 2 is not valid here (see [14]), the high level of $\overline{b_c^2}$ **cannot** be taken as evidence of QPC. However, it is interesting to note that the female speakers generally all have higher levels of $\overline{b_c^2}$ than the male speakers. The figure also shows the values of $\overline{q}$ for the same data, in which this trend is reversed - the females all have *low* levels of QPC detection (consider for example the results for speaker 3 shown in the figure). Because of limitations of space, the results for the vowels and nasals are not shown here, but they show very similar patterns to the voiced fricatives - females having higher levels of $\overline{b_c^2}$ than males, but much lower levels of $\overline{q}$ than males.

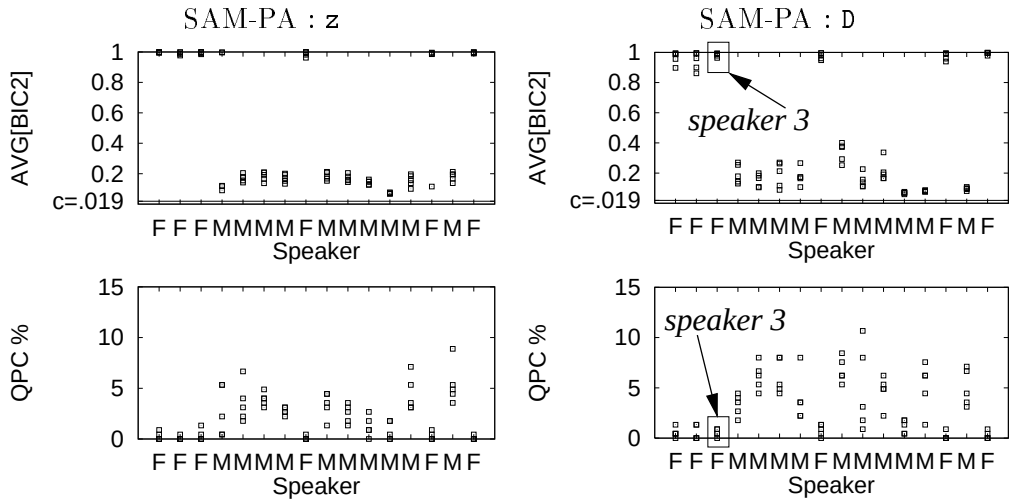

Figure 3: Scatter diagrams showing $\overline{b_c^2}$ (top) and $\overline{q}$ (bottom) for voiced fricatives z (left), D (right) by all 16 speakers. For $\overline{b_c^2}$ (top) c is the critical value at the $\alpha = 0.001$ for the Gaussian hypothesis test.

Further insight can be gained from calculation of the Pearson correlation coefficient r between the two measures $\overline{b_c^2}$ and $\overline{q}$ for each sound class. This is calculated from the $16 \times 5 \times n$ recordings of each class of speech sound. For the voiced fricatives ($n = 4$), vowels ($n = 12$) and nasals ($n = 3$) $r = -.67, -.59$ and $-.74$ respectively. Each of these coefficients is significant at the .05 level, indicating that high levels of $\overline{b_c^2}$ result in *low* levels of $\overline{q}$.

5. DISCUSSION AND CONCLUSION

The results from the unvoiced fricatives indicate that the values of b_c^2 for these sounds are not statistically different from zero. This suggests that there is unlikely to be any advantage in trying to model these sounds with quadratically nonlinear filters, and also (though not conclusively) that unvoiced speech can be satisfactorily modelled using Gaussian (i.e. second-order) models, and

The results from the voiced sounds require more careful interpretation. The overall trend of negative correlation between $\overline{b_c^2}$ and $\overline{q}$ can be interpreted in terms of the false-alarm properties of the QPC detector (see [11]). It can be seen from Equation 2, and it is shown schematically in Figure 4a), that signals with low $|b_c|$ will have a wide acceptance region, and hence a high false detection rate, whereas signals with high $|b_c|$ will have a narrow acceptance region, and so a low false detection rate.

If voiced speech signals *do* exhibit QPC, then it would be expected that those that have more bifrequencies passing the first stage of the two-part test in Figure 1 would have more passing the second stage : this would result in a positive correlation between $\overline{b_c^2}$ and $\overline{q}$, such as shown by line $A - A$ in Figure 4b). Alternatively, if the signals *do not* exhibit QPC, then the detection rate at the second part of the test would, from Equation 2, depend *inversely* on the bicoherence level, as shown

by the line $B - B$ in Figure 4. This would result in a *negative correlation* between $\overline{b_c^2}$ and $\overline{q}$.

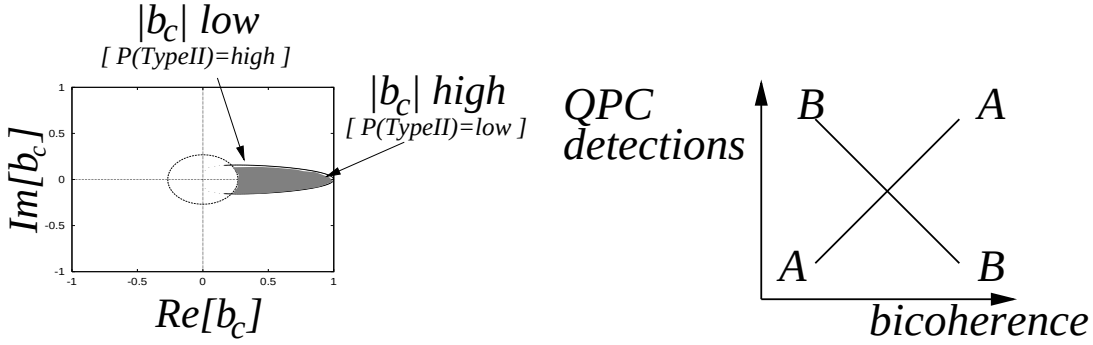

Figure 4: Schematic diagrams illustrating a) how P(Type II) errors varies with magnitude of b_c, b) types of correlations between $\overline{b_c^2}$ and $\overline{q}$.

On the basis of the observed results, it appears that the second scenario is the more plausible one, which means that the QPC detections made for the voiced sounds are in fact just false alarms, and do not correspond to QPC. Thus the conclusion is drawn that continuant voiced speech sounds do not exhibit significant levels of QPC, and so quadratically nonlinear models of these sounds are unlikely to be able to provide information not already encapsulated by linear models.

Further work is warranted in the extension of these results to other classes of speech sounds. It is extremely difficult to obtain reliable bicoherence estimates of short-lived signals, and this has to date ruled out similar research on stop sounds, and on transitions between speech sounds. The situation for analogous coherence functions at fourth-order (tricoherence) and even higher-orders is even worse, since the estimates of those quantities have higher variances, and so investigation into the properties of speech using these properties is impossible.

Of course, when better estimators for these measures have been developed, these goals may become reachable, but the current evidence is that speech does not possess any important third-order properties.

6. REFERENCES

[1] G. Kubin, "Nonlinear processing of speech," in *Speech Coding and Synthesis* (W. B. Kleijn and K. K. Paliwal, eds.), pp. 557–610, New York: Elsevier Science, 1995.

[2] C. L. Nikias and A. P. Petropulu, *Higher-Order Spectra analysis.* PTR Prentice Hall, New Jersey, 1st ed., 1993.

[3] Y. C. Kim and E. J. Powers, "Digital bispectral analysis and its applications to nonlinear wave interactions," *IEEE Transactions on Plasma Science*, vol. PS-7, no. 2, pp. 120–131, 1979.

[4] A. Moreno, J. A. R. Fonollosa, and J. Vidal, "Vocoder design based on HOS," in *Eurospeech '93*, (Berlin, Germany), pp. 519–522, September 1993.

[5] M. C. Dogan and J. M. Mendel, "Real time robust pitch detector," in *International Conference on Acoustics, Speech and Signal Processing*, (San Francisco, USA), pp. I129–I132, 1992.

[6] B. B. Wells, "Voiced/unvoiced decision based on the bispectrum," in *International Conference on Acoustics, Speech and Signal Processing*, pp. 1589–1592, 1985.

[7] J. L. Navarro, A. Moreno, and E. LLeida, "Bispectral-based statistics applied to speech endpoint detection," in *IEEE Signal Processing ATHOS Workshop on Higher-Order Statistics*, (Begur, Girona, Spain), pp. 280–283, June 1995.

[8] B. Boyanov, S. Hadjitodorov, and T. Ivanov, "Analysis of voiced speech by means of bispectrum," *Electronics Letters*, vol. 27, no. 24, 1991.

[9] J. Thyssen, *Non-linear analysis, prediction, and coding of speech*. PhD thesis, Tele Danmark Research, Hørsholm, Denmark, July 1995.

[10] J. W. A. Fackrell, P. R. White, J. K. Hammond, R. J. Pinnington, and A. T. Parsons, "The interpretation of the bispectra of vibration signals: part 1 - theory," *Mechanical Systems and Signal Processing*, vol. 9, no. 3, pp. 257–266, 1995.

[11] J. W. A. Fackrell, S. McLaughlin, and P. R. White, "Practical issues in the application of the bicoherence for the detection of quadratic phase coupling," in *IEEE Signal Processing ATHOS Workshop on Higher-Order Statistics*, (Begur, Girona, Spain), pp. 310–314, June 1995.

[12] G. Zhou, G. B. Giannakis, and A. Swami, "HOS for processes with mixed spectra," in *IEEE Signal Processing ATHOS Workshop on Higher-Order Statistics*, (Begur, Girona, Spain), pp. 352–356, IEEE, June 1995.

[13] J. W. A. Fackrell and S. McLaughlin, "Quadratic phase coupling detection using Higher Order Statistics," in *Proceedings of the IEE Colloquium on Higher Order Statistics*, no. 1995/111, (London, UK), pp. 9/1–9/8, IEE, May 1995.

[14] J. W. A. Fackrell, *The Bispectral Analysis of Speech Signals*. PhD thesis, University of Edinburgh, Edinburgh, UK, 1996. (submitted September 1996).

[15] J. W. A. Fackrell and S. McLaughlin, "Determining the false-alarm performance of HOS-based quadratic phase coupling detectors," in *Proceedings of EUSIPCO-96, Eighth European Signal Processing Conference*, (Trieste, Italy), September 1996. (to be presented).

[16] A. Breen, "Speech synthesis models : a review," *Electronics and Communication Engineering Journal*, pp. 19–31, February 1992.