ISSN 0968-0640

A Neurally Motivated Technique for Voicing Detection and F_0 Estimation for Speech.

Leslie S. Smith
CCCN/Department of Computer Science
University of Stirling
Stirling FK9 4LA
Scotland
email: lss@cs.stir.ac.uk

Centre for Cognitive and Computational Neuroscience
Technical Report CCCN-22
University of Stirling
Stirling FK9 4LA, Scotland
July 1996

Abstract

Speech consists of alternating voiced and unvoiced sections. Voiced speech consists of multiple harmonics of some fundamental (F_0); unvoiced speech consists of silence, or filtered noise. Here, speech is wideband bandpass filtered into many bands (modelling the cochlea). Each filter output is rectified (modelling the organ of Corti hair cell action), and bandpass filtered by convolution with the difference between two causal Gaussian averaging functions. This detects and emphasises the amplitude modulation resulting from unresolved harmonics (and models the combined effect of the auditory nerve and certain cochlear nucleus cell types). This output is compressed, summed across the bands, then used to discover glottal pulses. The presence of glottal pulses signals voicing, and the time between glottal pulses is used to find F_0. Results show good performance, particularly on male speakers. The system is reasonably resistant to background noise.

1 Background

Speech sounds may be voiced or unvoiced: that is, the vocal cords may be oscillating, resulting in the vocal tract being driven by an oscillating signal, or they may not, in which case the sound produced has a noise-like spectrum. The presence of voicing in speech is an important feature for interpretation. Determining whether speech sounds are voiced (and if so, what the fundamental frequency of the voicing is) is a problem with a long history going back to 1949 [17]. Gruentz [17] used the dominant frequency in the low-passed signal, and techniques which process the cepstrum are derivatives of this. Atal and Rabiner [3] use a pattern recognition approach based on the energy, the zero-crossing rate, the autocorrelation function, the first predictor coefficient of a 12-pole linear predictor output, and the prediction error of the linear predictor of the signal. This gives about 4% errors, though it requires classifier training. A more recent version [2] gives less than 0.5% errors. Siegel and Bessey [30] use additional features about the signal periodicity in

the more complex task of voiced/unvoiced/mixed excitation classification. Knorr [18] presents a hardware implementation of a technique based on the relative energy of the low frequency and high frequency parts of the signal, and claims $< 1\%$ errors.

The technique described here does not achieve this level of accuracy, achieving about 90% correct on continuous speech from the TIMIT database. However, it does so using a technique which (a) is simple, and easily incorporated into the other techniques, (b) has some basis in the biology, (c) does not require training, and (d) is reasonably resistant to noise. Further, we believe that considerable further tuning of the technique is possible (see section 5).

Knowledge of the fundamental frequency, F_0, of voiced speech can be useful in speaker identification, in detecting intonation, and in the monaural streaming of concurrent speech sounds. Estimation of F_0 in voiced speech has been attempted using many techniques: simple time-domain analysis of a low-passed signal [17, 13, 16], autocorrelation techniques applied to bandpass filtered speech [19, 31, 22], techniques based on Goldstein's theory of pitch detection (in essence, looking for an F_0 which best explains the partials present in the signal) [14, 10, 8, 6, 9], using derivatives of short-term power spectra [11], or most recently, using measures of instantaneous frequency [1, 27].

The aim of the work reported here is to show how an additional factor, namely the presence of amplitude modulation caused by unresolved harmonics in wideband bandpassed speech can be used for both voicing detection and F_0 estimation. Like the autocorrelation techniques, we use bandpass filtered speech from an auditory front end [26], but instead of autocorrelation, we seek amplitude modulation. This appears to fit with the neurobiology, in which certain cells (chop-S and onset-C) in cochlear nucleus are particularly responsive even to small amounts of amplitude modulation [15, 23, 25], and with work suggesting an amplitude modulation map in the inferior colliculus [20]. The use of amplitude modulation is thus neurally plausible. Indeed, Sachs et al [29] suggest that second and third formants cannot be found using only rate coding in the auditory nerve, but can be found using temporal information. Amplitude modulation has been used in computational models of early auditory processing [21, 34]. Although we do not model chop-S and onset-C cells directly, we model their effect by enhancing amplitude modulation present. Some aspects of this work have been filed as a patent [33].

2 Techniques Used

Figure 1 shows the stages of the algorithm used. Digitised sound (sampled at 16Ksamples/second or faster) was input to the AIM gammatone digital wideband bandpass filterbank [26]. This digitally filters the sound into a number of wideband bandpassed channels. Each channel follows the bandpass characteristic of the cochlea, with equivalent rectangular bandwidth (ERB)

$$\mathrm{ERB}_c = 24.7 + F_c/Q \tag{1}$$

for centre frequency F_c. Auditory nerve response is characterised by $Q = 9.265$ [12]. However, this value of Q is for pure tones at low sound pressure level (SPL), and the selectivity broadens (i.e. Q decreases) for higher SPLs, and for wideband sounds [24, 29]. Pure voiced sounds have the form

$$s(t) = \sum_{i=1}^{M} A_i \sin(2i\pi F_0 t + \phi_i) \tag{2}$$

so that each filter output will consist of zero or more adjacent harmonics of the fundamental frequency F_0. For speech F_0 generally lies between 100 and 250Hz. For harmonics to be unresolved, we require $\mathrm{ERB}_c \geq F_0$ so that

$$F_c \geq (F_0 - 24.7)Q \tag{3}$$

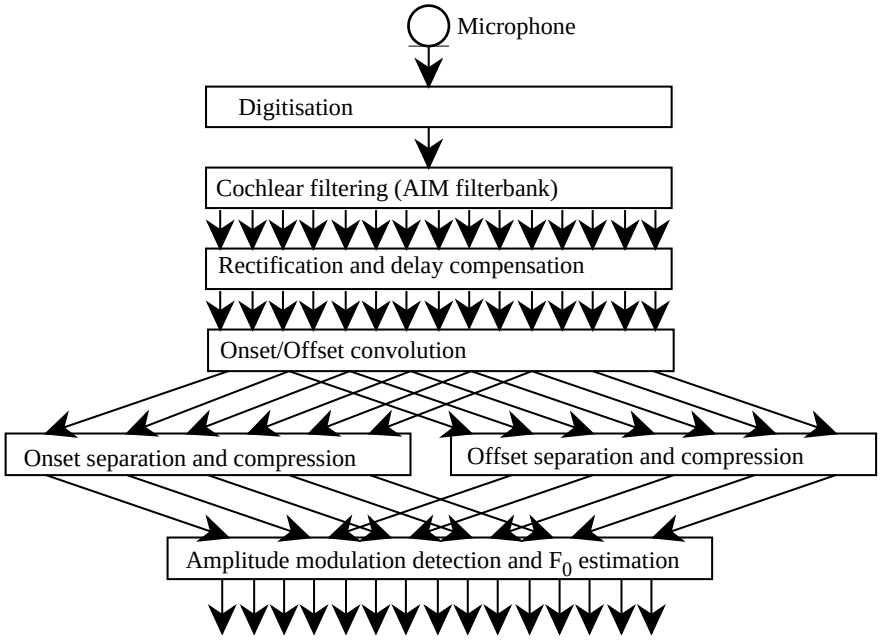

Figure 1: Outline of stages in algorithm.

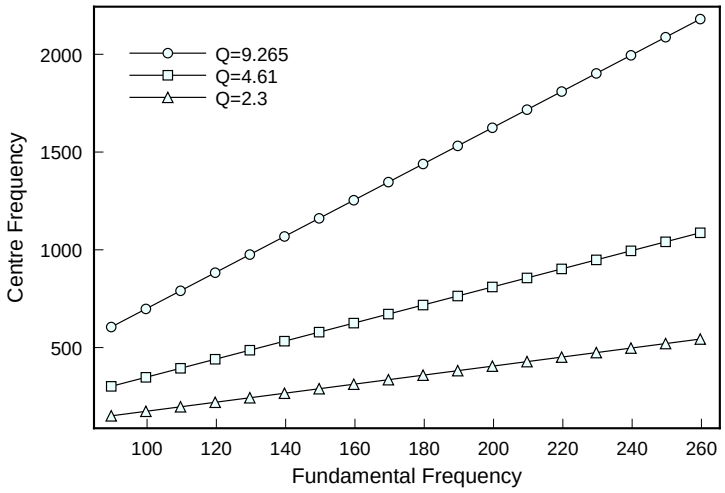

Figure 2: Minimum F_c for unresolved harmonics plotted against F_0 for varying Q.

and this is plotted in figure 2. Where F_c is in the middle of a number of reasonably strong adjacent harmonics, the (real) Q value will fall; however, the AIM software keeps Q constant throughout the frequency range. Because we seek amplitude modulation (AM) from unresolved harmonics, we adjusted the lowest F_c depending on Q and the speaker gender (i.e. approximately expected F_0). The highest F_c used was 3KHz (males) or 4KHz (females), and these were chosen to reflect the frequency at which the energy decreased considerably.

Channels were adjusted to compensate for the delay associated with each channel using the formula

$$\text{Delay}_{\text{ERB}} = \frac{n-1}{2\pi \text{ERB}} \tag{4}$$

where n is the filter order (here 4) [7]. This corresponds to the group delay, rather than the onset (or signal front) delay. The amplitude modulation information is contained in the envelope of the filter output. AM detection was achieved by rectification of each channel. This approximately models the effect of a population of inner hair cells [28]. To enhance amplitude modulation the rectified filter output was convolved with a difference of Gaussian averages convolving function. To reduce computational overhead, the data was first resampled to 4Ksamples/second. Rectification

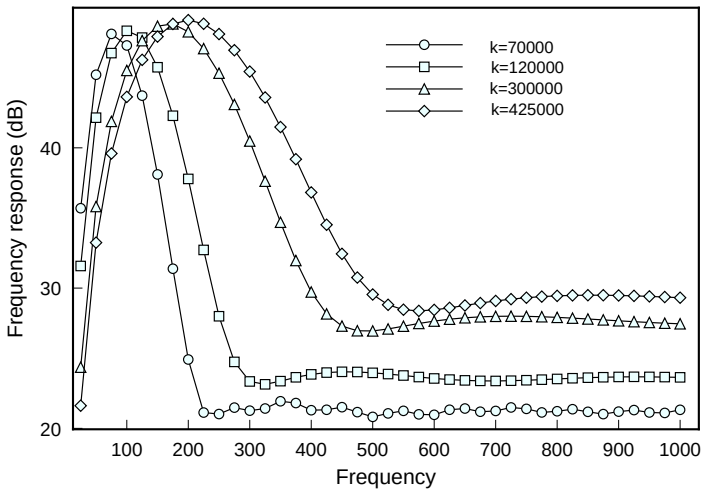

Figure 3: Bandpass filter characteristic for half-difference of Gaussian filter for varying values of k. $r = 1.2$ for all cases.

and convolution partly model effects found in the auditory nerve, but more importantly they model the amplitude modulation amplification provided by the onset-C cells of the cochlear nucleus [25]. At the same time, the convolution smooths the rapid signal fluctuations, reducing the iF_0 content. The effect of the convolution is to perform a digital bandpass filtering operation on the signal: the bandpass characteristic is shown in figure 3.

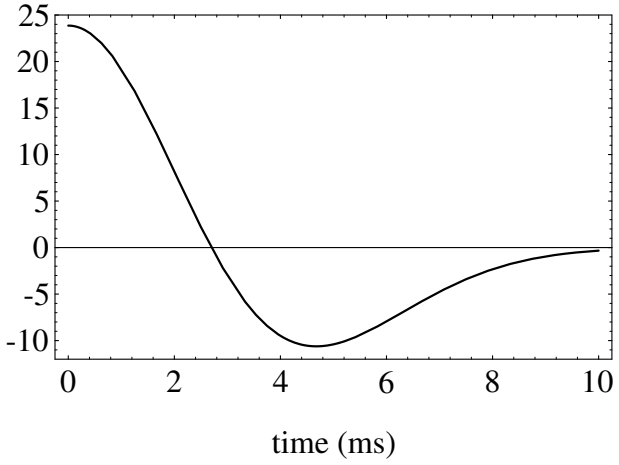

Figure 4: Onset/offset convolving function used ($r = 1.2$, $k = 150000$).

The convolving function, $g(t, k, r)$, (shown in figure 4) is

$$g(t, k, r) = f(t, k) - f(t, k/r) \tag{5}$$

where $k > 1$ and $f(x, y) = \sqrt{y}\exp(-yx^2)$. The convolution applied is

$$C_j(t) = \int_0^t s_j(t-\tau)g(\tau, k, r)d\tau \tag{6}$$

where $s_j(t)$ is the rectified output of the j'th filter. Since $\tau \geq 0$, this is a causal filter (and therefore suitable for real-time implementation), and we are applying the difference between a short-term and a longer term Gaussian average both with peak at t. Since $\int_0^\infty g(t, k, r)dt = 0$, $C_j(t) = 0$ for a signal of constant amplitude. When applied to a non-constant signal, $C_j(t)$ goes positive when the signal is increasing, and negative when the signal is decreasing. The choice of k and r determine the duration of pulses to which the transform is most sensitive, as can be seen in figure 3. For the work reported here, we used $r = 1.2$ and $k = 70000$ (males) or $k = 150000$ or $k = 300000$ (females).

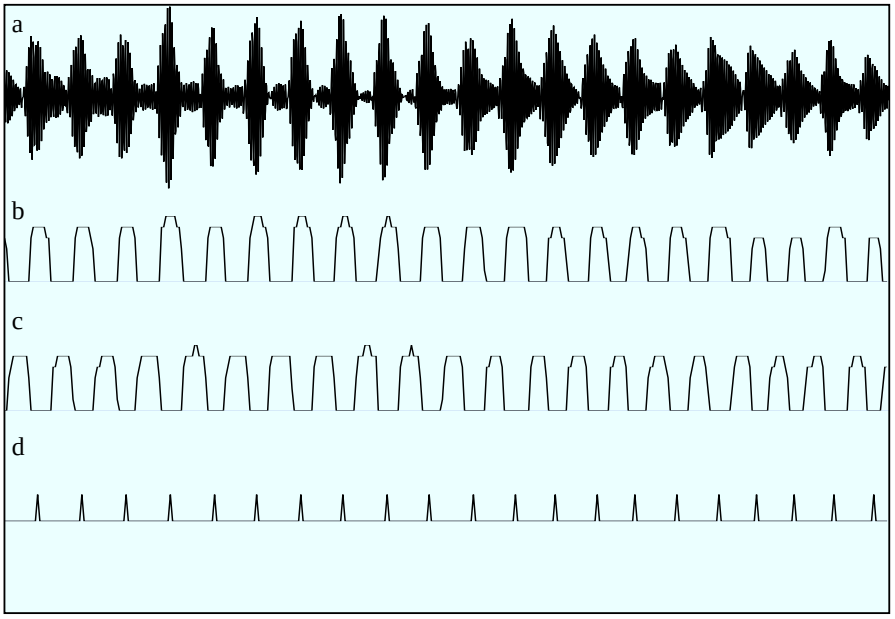

Figure 5: Response of one channel, with $F_c = 3200$Hz, $Q = 9.265$ to 100 ms of vowel sound /a/ (from TIMIT database, dr1/fsjk1/sa1). (a) shows the original filterbank output. (b) shows the compressed onset signal, (c) shows the compressed offset signal, and (d) shows the amplitude modulation pulses found.

The $C_j(t)$ were separated out into an onset signal (the positive-going part) and an offset signal (the inverted negative-going part). These were compressed logarithmically, but maintaining both as non-negative signals by taking $\log(x) = 0$ for $0 \leq x \leq 1$. This models compressive nonlinearities found in neurobiological systems. For the compressed onset signal the resulting output consists of a pulse due to the rapid rise in the envelope of an amplitude modulated signal; for the compressed offset signal, the pulse is caused by the rapid fall in the envelope. Thus, for an amplitude modulated signal, onset pulses and offset pulses alternate. These onset and offset signals were searched for pulses, allowing as pulses only those whose maximum value reached a threshold (generally 3 or 4), and whose duration was between 1.9ms and 8ms (females), or between 3ms and 11ms (males). The occurrence of a pulse in the onset signal followed within a short time by a pulse in the offset signal was taken to mark an amplitude modulation pulse: see figure 5. In this way, we produced an amplitude modulation (or voicing) map, with each pulse marking the occurrence of an amplitude modulation pulse at a particular time in a particular channel. The amplitude modulation caused by the summation of a number of unresolved adjacent harmonics with varying strengths and phases can occur at F_0, $2F_0$, $3F_0$, etc., depending on the exact phase and the relative strengths

of the partials [32]: however, we found that the F_0 component was much the strongest. Thus, the inter-pulse interval can be used to estimate F_0. Figure 6 shows the voicing map produced for 257 ms of one utterance.

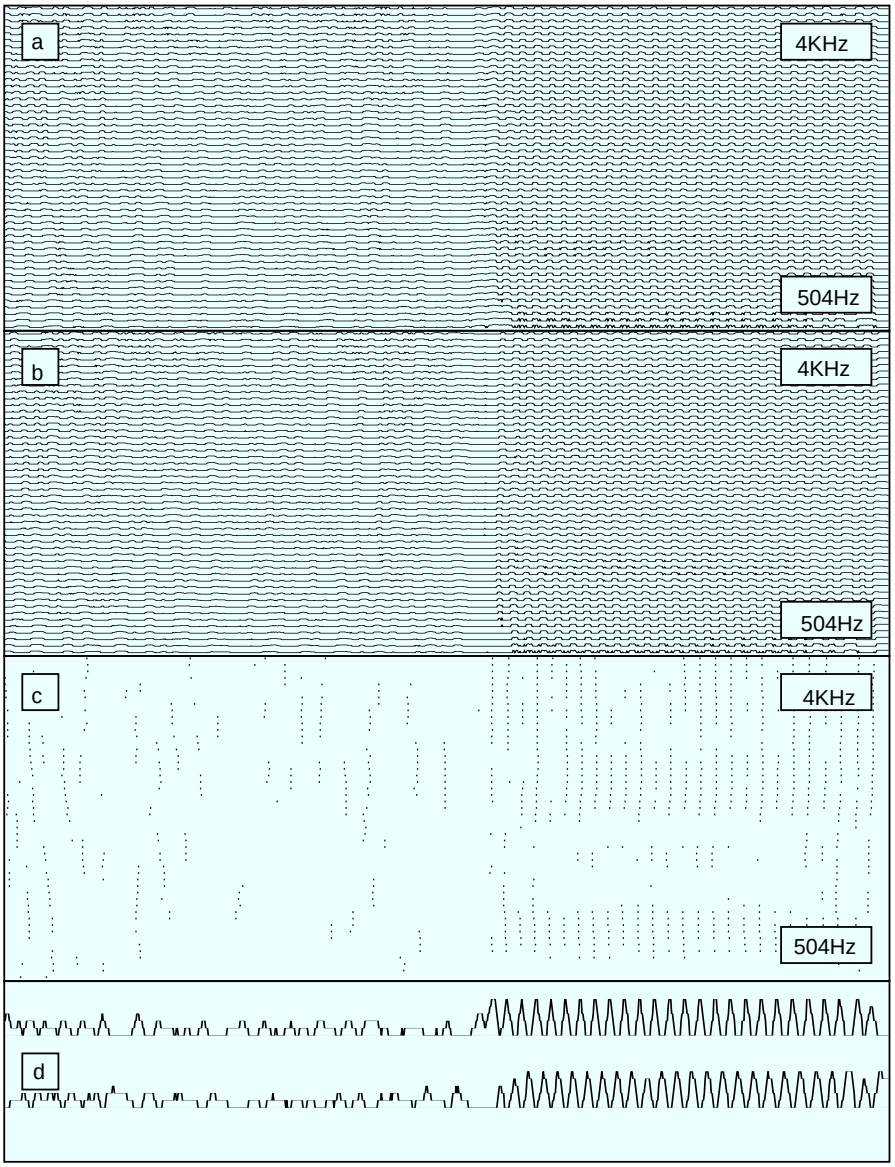

Figure 6: Sound /s ux/ from TIMIT utterance dr1/fsjk1/sa1, 1.013-1.270 seconds. Parameters used were $Q = 3.25$, and F_c from 504Hz (bottom) to 4000Hz (top), 50 channels, $k = 300000$ and $r = 1.2$. (a) shows compressed onset map, (b) shows compressed offset map, (c) shows the voicing map, (d) shows the summary compressed onset (top) and summary compressed offset (bottom) signals.

In addition to the voicing map, we also produced a summary voicing map. A summary compressed onset signal $D(t)$, and a summary compressed offset signal, $E(t)$ were produced by simple averaging of the compressed onset or offset signals from all the channels. The effectiveness of simply averaging the onset and offset signals relies heavily on the correction of the delays introduced by the cochlear filtering. Both the summary compressed onset and the summary compressed offset signals were searched for pulses, as above. The occurrence of a pulse in $D(t)$ followed within a short time by a

pulse in $E(t)$ was taken to mark a glottal pulse. The presence of a train of such pulses was taken to mean that the sound was voiced. Again, the inter-pulse interval can be used to estimate F_0.

3 Results

We present results for detecting voicing, and then for estimating F_0 in the sections of speech classified as voiced.

3.1 Detecting Voicing

To illustrate the techniques described above, we first applied them to one of the female TIMIT utterances, namely dr6/fsbk0/sa1. Figure 7 shows the voiced segments found using the summary onset/offset technique for a range of parameters. The lowest F_c used has been determined from

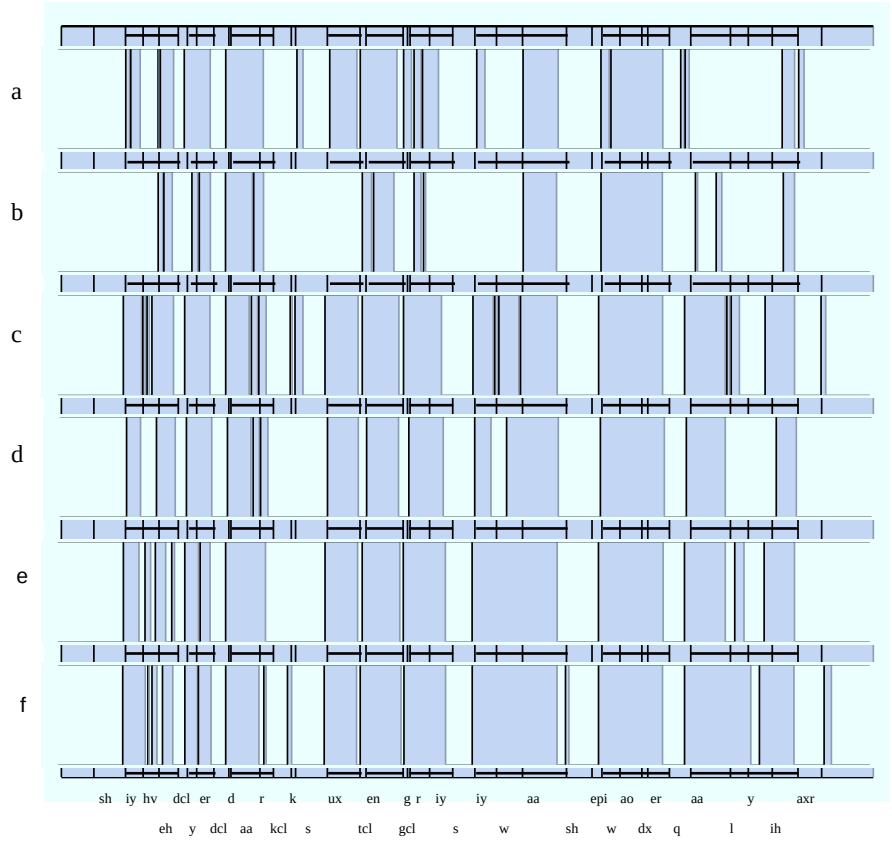

Figure 7: Voiced segments found in dr6/fsbk0/sa1 (*She had your dark suit in greasy wash water all year*). Background is the phoneme structure supplied with the TIMIT database, with lines indicating voiced sections (found by ear). Parameters were (a) 1434-3000Hz, $Q = 9.265$, density = 4 (24 bands) (b) as (a), but 500-3000Hz. (c), (d) 717-3000Hz, $Q = 4.6$, density = 8 (48 bands), (e) 504-3000Hz, $Q = 3.25$, density = 8 (43 bands), (f) 358-3000Hz, $Q = 2.3$, density = 8, 37 bands. Other parameters: all: Onset/offset filter $k = 300000$, $r = 1.2$, voicing pulse sizes 0.0019, 0.008, (a)-(c) pulse threshold 3, (d)-(f) pulse threshold 4.

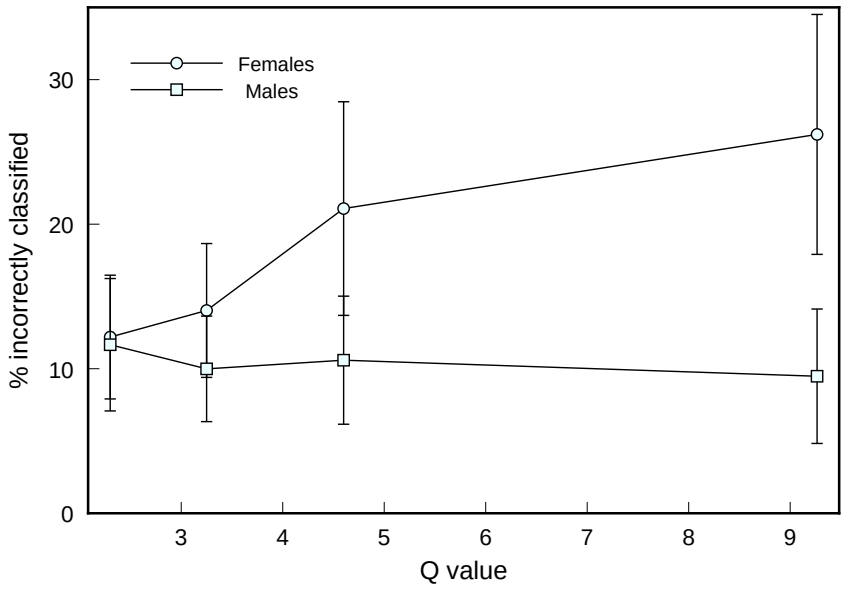

Figure 8: Classification errors for male and female utterances for varying Q. Y axis is percentage of utterance time.

equation 3 (excepting figure 7b). The density (i.e. number of channels per ERB) was chosen to keep the number of bands used manageable. With $Q = 9.265$, (figure 7a) many voiced sections are missed, and part of the /s/ in "suit" classed as voiced. This probably reflects the loss of formant information below 1434 Hz. However, although simply reducing the lowest F_c to 500Hz removes the misclassification of the /s/, it does not improve the missing voiced sections (figure 7b). Reducing Q to 4.6 (figure 7c) improves performance by correctly classifying more of the voiced sections, although the /s/ in "suit" remains misclassified unless the threshold is increased to 4 (figure 7d). Reducing Q further to 3.25 (figure 7e) with the threshold at 4 gives better results: reducing Q to 2.3 (figure 7f) results in some misclassification being reintroduced. Using $Q = 3.25$ has been found to be a good compromise in this case.

The techniques above were applied to 15 male and 15 female utterances from the TIMIT database. The utterances were also segmented by hand into voiced and unvoiced sections based on the phonemic classification, adjusted by listening. Most of the transitions between voiced and unvoiced were on the phoneme boundaries provided, although many phoneme boundaries were between pairs of voiced or pairs of unvoiced sounds. Thus there were fewer voiced and unvoiced segments than phonemes. The summary onset/offset technique was used to find the voiced sections, and these were compared with the hand-segmented versions.

The results are shown in figures 8. The computer segmentation was considered to be in error whenever the voiced/unvoiced boundary computed was greater than 10ms away from that found by hand. The Q value of the filter was varied from 2.3 to 9.265. The lowest value for F_c used was calculated using equation 3, and an assumed fundamental of 100Hz for males, and an assumed fundamental of 180Hz for females. For males, the Q value is much less critical than for the females. For all the utterances the concentration of energetic partials caused by the existence of a formant (and thus leading to amplitude modulation) occurs at only slightly lower frequency frequencies for males than for females since both male and female vocal tracts have similar dimensions (the difference is about 3 semitones [5], p480). However for high Q values, the lowest F_c used is so high for females that important concentrations of partials are missed.

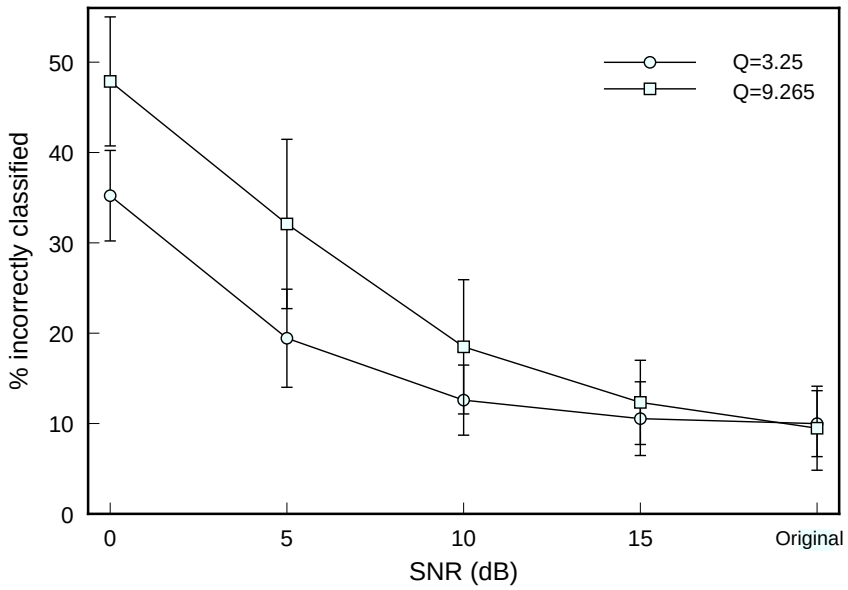

Figure 9: Classification errors for male utterances in noise for two values of Q.

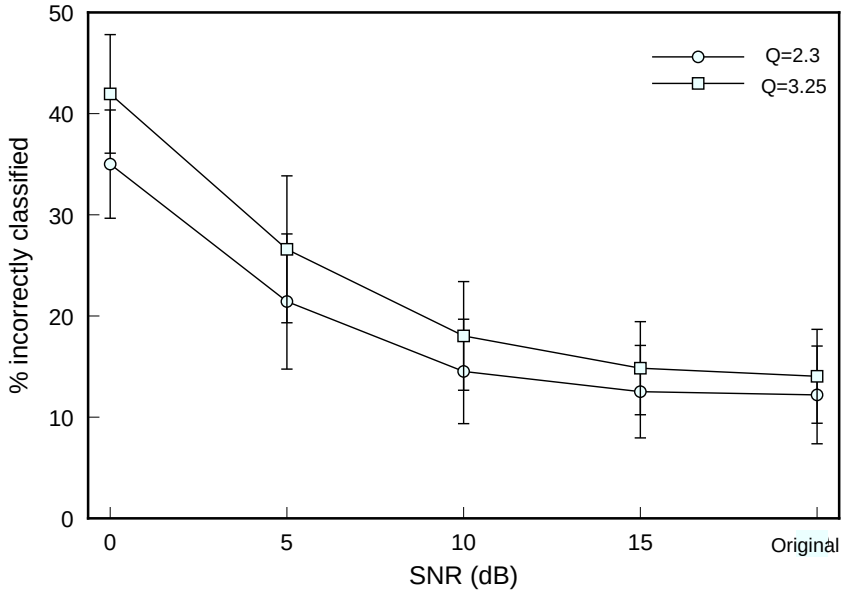

Figure 10: Classification errors for female utterances in noise for two values of Q.

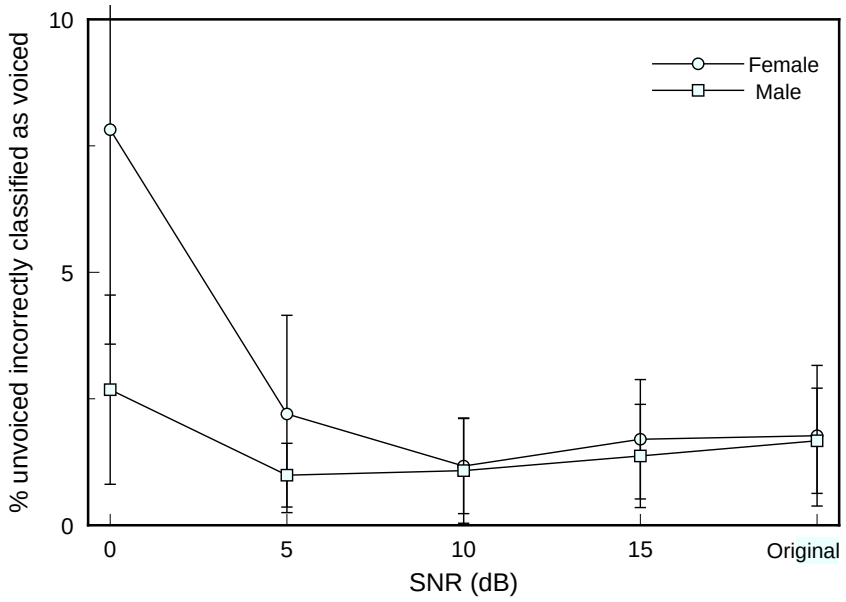

Figure 11: Unvoiced classed as voiced errors for female and male utterances in noise for varying SNR. Male utterances used $Q = 3.25$ and female utterances used $Q = 2.3$. Results for different Q values are similar.

The same techniques were used in sound with white noise added: the results for some different values of Q are shown in figure 9 for male speakers and in figure 10 for female speakers. In both cases, using a low value of Q gives better results when noise is present.

Both in the presence and absence of noise, errors due to voiced sections being classed unvoiced are much more frequent than errors due to unvoiced sections being classed voiced. Further, the addition of white noise hardly increases the likelihood of unvoiced sections being classed voiced at all until the SNR is very low: see figure 11. Thus, most of the errors consist of voiced sections failing to be identified as such, and the frequency of this type of error increases monotonically with decreasing SNR.

Many of the voiced classed as unvoiced errors occur entirely inside voiced sections. Closer examination of the amplitude modulation in different channels shows that precise pulse timing varies between channels, resulting in degradation of pulses formed by simple cross-channel summation. This can be seen in figure 12: the amount of jitter seems reasonably constant, but it causes more problems when F_0 is higher, accounting for the poorer performance on female voices.

3.2 Estimating F_0

Since the voicing detection technique uses the amplitude modulation pulses caused by unresolved adjacent harmonics, we can estimate the instantaneous F_0 using the time between these pulses. Reasonable results were obtained using the same summary technique as that used for detecting voicing. Figure 13 shows the estimates of the fundamental frequency, along with the original speech signal. The estimates follow the speech signal envelope periodicity.

Because the speech sounds were resampled at 4Ksamples/second between rectification and onset/offset convolution, the accuracy of F_0 estimation is reduced. The discrete jumps in F_0 estimation are clearly visible in both figures 13 and 14. This effect is stronger for higher F_0, so that

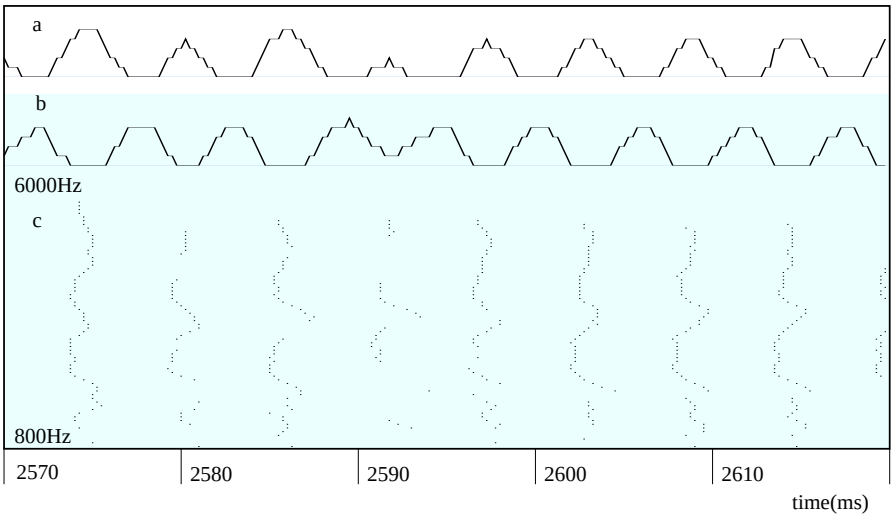

Figure 12: Response to 50ms of TIMIT utterance dr1/fsjk1/sa1 (female), 2570-2620ms. The phoneme /aa/ has a short section classified unvoiced caused by the small summary onset pulse at 2592ms. a: summary onset signal. b: summary offset signal. c: glottal pulses found in each channel (67 channels, 800-6000Hz, $Q = 9.265$).

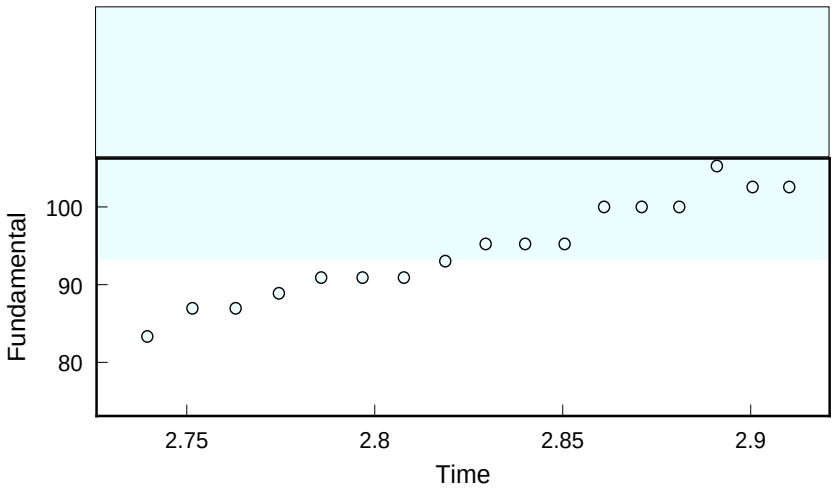

Figure 13: Fundamental frequency estimates for a short piece of voiced speech (/ao/ from dr2/mjhi0/sa1 in the TIMIT database).

better results are obtained for lower F_0. This, in addition to the jitter problem already described makes the system considerably better at F_0 estimation for male speech than for female speech. This can be seen by comparing figures 14A and C.

There are three different types of error present in figure 14: (i) errors at the start of a voiced segment, due to the way in which the amplitude modulation begins in different channels, (ii) errors due to jitter across channels resulting in the simple summation giving an incorrect estimate, and (iii) errors due to pulses not being strong enough to be detected, leading to an estimate of F_0 which is a fraction of the correct value. The type (iii) errors cause the four low outliers in figure 14A: all of these are half the correct value for F_0, and all occur at the ends of segments. The type

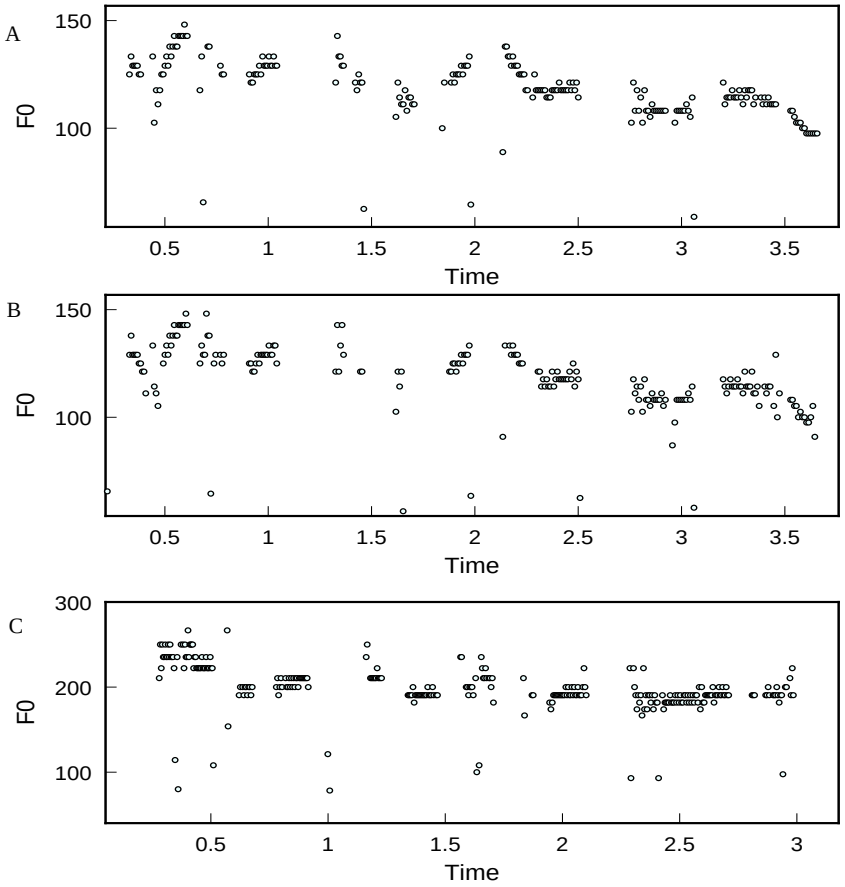

Figure 14: A: fundamental frequency estimates for TIMIT dataset dr3/mkls1/sa1 (male speaker) ("She had your dark suit in greasy wash water all year"). Parameters used were $Q = 3.25$, $k = 70000$, $r = 1.2$. B: as A, but with white noise added to give a SNR of 10dB. C: fundamental frequency estimates for TIMIT dataset dr2/feac0/sa1 (female speaker): parameters were $Q = 2.3$, $k = 300000$, and $r = 1.2$.

(ii) errors result in either a high estimate for F_0 being followed by a low one, or vice versa. This occurs 6 times in figure 14A, at times 0.441s, 1.326s, 1.618s, 1.662s, 2.280s, and 2.757s. All of these except the 1.662s one occur at the start of a segment, and so are errors of type (i) as well. In addition, there is one other type (i) error, at time 2.135s. When noise is added, some more errors occur although the basic movement of F_0 remains visible. The same types of errors occur in figure 14C: some of the type (iii) errors are one third of the correct F_0. Sections of speech mistakenly classified as voiced still give values for F_0, and this occurs ar time 0.999-1.018s here. Again, even with these errors, the movement of F_0 is quite visible.

In addition to these relatively major errors in F_0 estimation, there is an interaction between the jitter across channels and the resampling at 4Ksamples/second. The jitter across channels causes the amplitude modulation peak to move depending on the exact strength and time of the amplitude modulation pulse in each channel: the resampling rate forces movement in F_0 to be discrete. Thus, the estimated F_0 varies discretely and quickly between adjacent glottal pulses. This problem is worse for higher F_0, and results in the rather thick lines in figure 14C.

Methods for tackling all of these problems are discussed in section 5.

4 Conclusions

An effective biologically motivated technique for detecting voiced sections of speech has been demonstrated. A method for finding F_0 using ideas based on cochlear nucleus cell responses has been produced.

The system works better for male voices than for female voices. Best overall results were obtained using a higher value for Q for male voices than for female voices, in the absence of noise. However, the system is relatively insensitive to Q for male voices, as can be seen in figure 8. In the presence of white noise, better results were obtained using a lower value for Q. Using a high value for Q means ignoring much of the low to middle frequency content of the signal, particularly for speakers with a high F_0. Although the estimates for F_0 do contain some errors, the estimates are correct more than 90% of the time.

Although the technique does not provide such good results in voicing detection as, for example, [2], it does so without the need for training. The system retains its effectiveness in an SNR of 10dB, both for voicing detection and F_0 estimation: the effectiveness of earlier algorithms in noise is not stated. The voicing detection technique is relatively simple and (so far) unoptimised: it could be added to the set of techniques used in the pattern recognition approaches of [3, 2, 30].

The F_0 estimation technique is believed to be more immune to noise than techniques based on simple time-domain analysis of the low-passed signal [17, 13, 16] since extraneous low-frequency sounds will not affect this method, as it ignores low-frequencies altogether. Indeed, this technique and simple time-domain analysis of a low-passed signal can augment each other as they are independent. The technique here is simpler than autocorrelation based techniques since channels are analysed directly instead of seeking peaks in the autocorrelation function (ACF) in each channel as [31, 22] do. It is also more biologically plausible, since no evidence has been found for biological computation of the ACF. The technique described here has most similarity to techniques based on finding partials present in the signal [14, 10, 8, 6, 9]: precise computation of the frequency of partials is not biologically plausible, because it requires very high frequency resolution. We infer the difference in frequency between adjacent partials from the amplitude modulation it produces. By requiring that the same amplitude modulation frequency be present over a range of channels, we are in essence making an assertion about the set of partials present, namely that they are harmonically related to a fundamental at the amplitude modulation frequency. It is difficult to tell which technique is best: we believe that this technique represents a useful addition to the set of F_0 estimation techniques.

5 Further Work

Further development work planned includes improving voicing detection by using a more sophisticated technique for combining the information in different channels. This would entail combining information across a number of adjacent channels, rather than simply summing them all. Similarly, better F_0 estimates can be achieved by combining the information on the presence of amplitude modulation in the different channels more effectively. This should help resolve the type (ii) and (iii) errors discussed in section 3.2: estimating F_0 using a number of adjacent channels should overcome problems due to jitter, and working with subsets of the set of channels should reduce the likelihood of pulses being omitted. This would also supply information on exactly which bands AM is present in, and this should be useful for identifying particular voiced sounds.

To model the cochlea and organ of corti more accurately would require a filterbank whose Q was not fixed across the whole spectrum, but varied according to the input to the system. Thus, for

an input consisting of a spoken vowel, we might have $Q = 2.3$ in a channel whose F_c was near a number of energetic partials, rising to 9.265 where the SPL was low. The work here suggests that this could permit both accurate frequency location of the formants (lost when Q is low: see figure 6 in which voicing pulses are found across the whole spectrum), and sensitivity to amplitude modulation caused by the interaction of relatively low-numbered (i.e. low values of i in equation 2) harmonics which is lost when Q is high. If this was found to be effective, it could influence the design of the bandpassing for future cochlear implants.

The techniques described can be extended to permit sound streaming [4]. Cues such as common onset of amplitude modulation, common frequency of amplitude modulation, common onset of energy across a number of (not necessarily adjacent) channels could be used to group a number of channels. Thus, for example, if the energy in some channels shares a common onset time with co-frequency amplitude modulation, we would assume the energy in these channels came from the same source.

The system described is entirely data-driven. Clearly, it could be embedded in a larger system which included top-down information.

The current system, is a purely software implementation. Even although it uses downsampling after rectification it is slow. The technique described is suitable for parallel implementation, and this could be accomplished either using aVLSI or DSP technology. Either of these could be used to make the system work in real-time.

Acknowledgements

I wish to acknowledge useful discussions with other members of the CCCN, particularly Peter Hancock.

References

[1] T. Abe, T. Kobayashi, and S. Imai. Harmonics estimation based on instantaneous frequency and its application to pitch determination of speech. *IEICE Transactions on Information and Systems*, E78-D:1188–1194, 1995.

[2] B.A.R. Al-Hashemy and S.M.R. Taha. Voiced-unvoiced-silence classification of speech signals based on statistical approaches. *Applied Acoustics*, 25:169–179, 1988.

[3] B.S. Atal and L.R. Rabiner. A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-24(3):201–212, 1976.

[4] A.S. Bregman. *Auditory scene analysis.* MIT Press, 1990.

[5] M. Campbell and C. Greated. *The musicians guide to acoustics.* J.M. Dent and Sons, 1987.

[6] D. Chazan, Y. Stettiner, and D. Malah. Optimal multi-pitch estimation using the em algorithm for co-channel speech separation. In *International conference on acoustics, speech and signal processing*, pages II.728–II.731, 1993.

[7] M. Cooke. *Modelling Auditory Processing and Organisation.* Distinguished Dissertations in Computer Science. Cambridge University Press, 1993.

[8] B. Doval and X. Rodet. Estimation of fundamental frequency of musical sound signals. In *International conference on acoustics, speech and signal processing*, pages 3657–3660, 1991.

[9] B. Doval and X. Rodet. Fundamental frequency estimation and tracking using maximum likelihood harmonic matching and HMMs. In *International conference on acoustics, speech and signal processing*, pages I.221–I.224, 1993.

[10] H. Duifhuis, L.M. Willems, and R.J. Sluyter. Measurement of pitch in speech: an implementation of goldstein's theory of pitch perception. *Journal of the Acoustical Society of America*, 71(6):1568–1580, 1982.

[11] T. Funada. A method for the extraction of spectral peaks and its application to fundamental frequency estimation of speech signals. *Signal Processing*, 13:15–28, 1987.

[12] B.R. Glasberg and B.C.J. Moore. Derivation of filter shapes from notched-noise data. *Hearing Research*, 47:103–138, 1990.

[13] B. Gold and L. Rabiner. Parallel processing techniques for estimating pitch periods of speech in the time domain. *Journal of the Acoustical Society of America*, 46(2):442–448, 1969.

[14] J.L. Goldstein. An optimum processor for the central formation of pitch of complex tones. *Journal of the Acoustical Society of America*, 63:486–497, 1973.

[15] M.J. Hewitt, R. Meddis, and T.M. Shackleton. A computer model of a cochlear nucleus stellate cell: responses to amplitude-modulated and pure-tone stimuli. *Journal of the Acoustical Society of America*, 91(4):2096–2109, 1992.

[16] G.S. Jovanovic. A new algorithm for speech fundamental frequency estimation. *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-34(3):626–630, 1986.

[17] O.O. Gruentz Jr and L.O. Schott. Extraction and portrayal of pitch of speech sounds. *Journal of the Acoustical Society of America*, 21(5), September 1949.

[18] S.G. Knorr. Reliable voiced/unvoiced decision. *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-27(3):263–267, 1979.

[19] M. Lahat, R.J. Niederjohn, and D.A. Krubsack. A spectral autocorrelation method for measurement of the fundamental frequency of noise-corrupted speech. *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-35(6):741–750, 1987.

[20] G. Langner. Periodicity coding in the auditory system. *Hearing Research*, 60:115–142, 1992.

[21] C. Lorenzi and F. Berthommier. A computational model for amplitude modulation extraction and analysis of simultaneous amplitude modulated signals. *Journal de Physique IV*, 4:C5.379–C5.382, 1994.

[22] R. Meddis and M.J. Hewitt. Modelling the identification of concurrent vowels with different fundamental frequencies. *Journal of the Acoustical Society of America*, 91(1):233–245, 1992.

[23] G.F. Meyer and W.A. Ainsworth. Vowel pitch period extraction by models of neurones in the mammalian brain-stem. In *EUROSPEECH 93*, pages 2029–2032, 1993.

[24] A.R. Moller. *Auditory Physiology*. Academic Press, 1983.

[25] A.R. Palmer and I.M. Winter. Cochlear nerve and cochlear nucleus responses to the fundamental frequency of voiced speech sounds and harmonic complex tones. *Advances in the Biosciences*, 83:231–239, 1992.

[26] R.D. Patterson, M.H. Allerhand, and C. Giguere. Time-domain modelling of peripheral auditory processing: A modular architecture and a software platform. *Journal of the Acoustical Society of America*, 98:1890–1894, 1995.

[27] L. Qiu, H. Yang, and S-N Koh. Fundamental frequency determination based on instantaneous frequency estimation. *Signal Processing*, 44:233–241, 1995.

[28] M.A. Ruggero. Physiology and coding of sound in the auditory nerve. In A.N. Popper and R.R Fay, editors, *The Mammalian Auditory Pathway: Neurophysiology*. Springer-Verlag, 1992.

[29] M.B. Sachs, H.F. Voigt, and E.D. Young. Auditory nerve representation of vowels in background noise. *Journal of Neurophysiology*, 50(1):27–45, 1983.

[30] L.J. Siegel. Voiced/unvoiced/mixed excitation classification of speech. *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-30(3):451–460, 1982.

[31] M. Slaney and R.F. Lyon. A perceptual pitch detector. In *International conference on acoustics, speech and signal processing*, pages 357–360, 1990.

[32] L.S. Smith. Data-driven sound interpretation:its application to voiced sounds. In P.J.B. Hancock L.S. Smith, editor, *Neural Computing and Psychology, Stirling 1994*, pages 147–154. Springer, 1995.

[33] L.S. Smith. A detector and fundamental frequency estimator for voiced sounds. UK patent application GB 9605593.4, March 1996.

[34] N.P.McA Todd. Explorations of a model of pitch perception based on the recognition of coherent am spectra. British Journal of Audiology (in press).