EVALUATION OF A COMPUTATIONALLY EFFICIENT METHOD FOR GENERATING A VOICED-SOURCE SYNCHRONISED TIMING SIGNAL

David R.L. Davies and J. Bruce Millar

Computer Sciences Laboratory, Research School of Information Sciences and Engineering, Australian National University

ABSTRACT - This paper evaluates the performance of a system that comprises a low-pass filter and a feature detecting post-processor to generate a voice-source synchronised timing signal. The evaluation of the output signal is described in terms of its phase relationship to the electro-glottograph signal. The paper discusses the choice of reference phase within the electro-glottograph signal, the degree of synchronisation achieved for data containing a wide range of vowel qualities and excitation fundamentals and the control of phase in dynamically iterative architectures. We have chosen a simple but extensible architecture and have discussed its failure modes over a wide range of signal conditions.

INTRODUCTION

The generation of a voiced-source synchronous timing signal (commonly, but erroneously referred to in the literature as pitch epoch detection) is perhaps the most basic of all speech feature extraction processes. It is an algorithmic process that raises some interesting issues of definition that are not simply resolved, as well as representing the non-trivial deconvolution of a complex physical process. A thorough evaluation of such systems will relate to their performance when presented with a range of phonation types, phonation rates, and vowel qualities in all combinations, together with dynamic variation in each of these input dimensions.

We have chosen to evaluate an approach referred to by Hess (1983) as Fundamental Harmonic Extraction (FHE). It is one of the simplest and oldest techniques used in the estimation of the fundamental period of speech signals. The speech signal is filtered with a low-pass or band-pass filter to extract the dominant low frequency component as a near-sinusoidal waveform.

The issue addressed in the current study is the degree to which the phase of the extracted sinusoid can be reliably related to the instant of glottal closure (IGC) derived from the electrical resistance of the glottis (EG).

We have focussed on the design and performance of the filter and use a simple peak picking algorithm to generate a pulse train to represent the phase of the sinusoid. We have started with a simple filter that is capable of expansion through added stages (iterations) or, ultimately, adaptation of parameters to particular signal categories and have explored a range of signal conditions looking for significant failure modes. From a software engineering perspective it is important to be able to detect signal conditions under which extra computational effort is required or, if an algorithm is likely to fail, another substituted.

Results are presented from application to the first one minute reading passage of the Australian corpus collected by Millar, O'Kane, and Bryant (1989). This corpus was used because it provides a significant amount of coincident speech and EG data. Approximately ten minutes of voiced speech was selected from seventeen speakers (nine female and eight male) chosen for the quality of their EG signals.

THE REFERENCE SIGNAL

The IGC was estimated from the EG signal by detecting the largest negative peaks of the second derivative. An exponentially decaying threshold with a preset minimum was used to avoid smaller peaks within the source period. Of the peaks at the IGC, only those exceeding a fixed fraction (0.2) of the maximum peak height were used, thus restricting our analysis to the strongly voiced speech that displayed a distinctive knee in the EG signal (fig. 1).

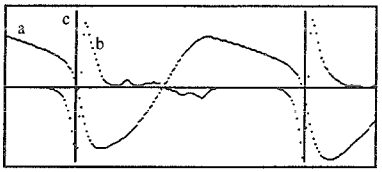

Figure 1. Detecting the knee in the EG signal
$a = EG$, $b = \partial^2 EG/\partial t^2$, c = IGC marker

MULTI-STAGE LOW-PASS FILTERING

When extracting the lowest frequency component of a signal, maximum frequency discrimination can be attained, for a given filter order, by operating with the signal spanning the monotonic roll-off of a low-pass filter and using floating-point arithmetic to cope with the increased dynamic range of the filter output. The disadvantage of operating in this regime is that short term fluctuations in F0 appear as low frequency amplitude fluctuations. The significance of this as a limit to filter order is discussed below.

Dologlou and Caratannis (1989) have investigated this approach using a second order FIR low-pass filter. The low order of their filter required a large number of iterations (>400) which makes their approach computationally intensive and led to a focus on the determination of an effective halting condition for the iterations in the subsequent debate (Dologlou and Caratannis, 1991).

Reducing the number of iterations to a more practical level requires a higher order filter. These are still, generally, computationally intensive and not readily made adaptive when that is necessary to track changes in F0 for optimal performance. A rectangularly windowed moving average of any order can be implemented recursively with the computational complexity of a second order filter but suffers from strong side-lobes in its frequency response. Another window that can be implemented with twice second order complexity is the symmetric exponential window which has side lobes that decrease with increasing order and decay rate.

In this work an approximation to a symmetric exponential filter with maximal order was implemented in recursive IIR form using a leaky integrator filter. Time reversal was used to produce time symmetry and, thus, zero phase characteristics. The breakdown of this approximation near the ends of the filtered speech segment was investigated and reported below.

The difference equation for a simple leaky integrator for a signal x(n) and filtered signal y(n) at time n is

$$y(n) = ay(n-1) + (1-a)x(n).$$

The magnitude frequency response is

$$|H(\omega)| = (1-a)(1 + a^2 - 2a\mathrm{Cos}\omega)^{-1/2}.$$

The Time Reversed Leaky Integrator (TRLI) used here has two functional parameters, the memory constant (a) and the number of iterations (l). Its frequency response is

$$|H(\omega)|_l = (1-a)^{2l}(1 + a^2 - 2a\mathrm{Cos}\omega)^{-l}$$

which has no side-lobes in the frequency range of interest and thus has a monotonic roll-off. Increasing 'l' steepens the roll-off while decreasing 'a' pushes it higher in frequency.

PREPROCESSING

All signals used in this study were pre-filtered with the same low-cutoff, high-pass filter to reduce low frequency noise. Since the presence of low frequency noise arising from frequency demodulation is confused by issues of signal recording quality the question of choosing an appropriate level of pre-filtering will not be discussed here.

Voicing segmentation markers were generated by filtering the squared EG signal with a 10ms moving average filter then generating segmentation onset/offset pulses when this smoothed energy crossed a threshold set at 20% of the maximum value. These markers were used to select the portions of speech for subsequent processing.

TEST PROCEDURES

Two series of preliminary tests were conducted to determine a suitable combination of filter parameters for processing speech from a range of speakers and voicing quality without filter adaptation. The test procedure used was to generate a histogram of the time differences between the EG signal and filtered speech and to maximise the height/area ratio of the histograms (fig. 2).

The first series used approximately 3 seconds of voiced speech from one speaker and a broad range of parameters spanning a = 0.5 to 0.99 and l = 1 to 50. The second series was conducted on approximately 20 seconds of voiced speech from two speakers chosen to have a large inter-speaker F0 difference and high internal variance in F0 over a narrower filter parameter space indicated as optimal in the first tests. A 2D raster scan was performed over the parameter space spanned by a = 0.86 to 0.94 and l = 2 to 16 giving optimum values of a = 0.88 and l = 8.

FEATURE DETECTION AND PHASE ALIGNMENT

The speech signal was processed by the fixed low-pass filter optimised above and a peak detection algorithm. An alignment algorithm paired the EG and speech signal timing pulses rejecting marks bounded by intervals differing by greater than a factor of two or outside the instantaneous frequency range of 50 to 500 Hz to cope with momentary breakdown of voicing segmentation.

PHASE VARIANCE TESTS

The time separation of these pulse pairs (positive for signal leading EG) constituted the primary data for the following analysis. A histogram was generated of the temporal distribution of this phase shift (fig. 2).

An attempt was made to characterise the speech signals that were associated with phase variance relative to the peak of the histogram. In this study we have not attempted to analyse in detail the differences in absolute peak position found to occur between speakers but offer discussion below.

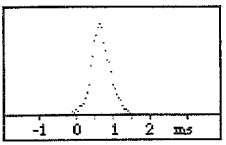

Fig. 2 Phase Histogram for ~37 seconds of voiced speech from speaker 6.

Five hypotheses as to the source of the relative phase variance were tested: (1) residual formant energy in the filtered signal, (2) signal energy level, (3) the sign of the time derivative of signal energy, (4) local variations in F0, (5) the sign of the time derivative of F0.

Hypothesis (1) is based on the breakdown of the basic assumption of the FHE method and a linear source-filter model of speech production in which the harmonics of the glottal source are coloured by the formants. Under this assumption, removing high frequency components should leave a signal in

phase with the source. The rationale for hypothesis (2) was that the strength of voicing or glottal closure and related variations in glottal acoustic impedance might be a significant factor in the phase deviations. Hypothesis (3) was included after visual inspection of time-domain plots of the raw phase-marked data showed an increased phase shift during rising signal energy in some instances. Local variations in F0 could produce phase deviations through F0 smoothing effects of the low-pass filter or frequency demodulation in the filter skirt, leading to the inclusion of hypotheses (4) and (5) which address the magnitude and direction of local F0 changes.

Testing of these hypotheses was conducted by visual inspection of scatter plots such as figure 3. Each set comprises plots representing the distribution of data about the phase histogram peak with (a) all voice-source detections, (b) detections which lag the histogram peak, (c) detections within 100μs of the peak, (d) detections which lead the peak, where lagging and leading is by more than 100μs. Density distribution histograms can be seen on the left and lower borders of each plot and (e), (f), (g) are plots of the ratio between the densities of (b), (c), (d) and the totals (a). Estimates of F1 were made using the ESPS 'formant' routine. F0 is taken as the sampling rate (Fs = 20000s/s) divided by the current source period (T0).

DISCUSSION

For 'l' greater that 8 the primary limiting factor was the dynamic amplitude range of the filter output compounding the impact of low frequency noise. Errors due to end effects arising from the assumption of filter window symmetry were found to drop to less than signal quantisation errors within the 40 samples of segment ends. Data in this region was not used.

Figure 2 shows the phase histogram for approximately 37s of voiced speech from speaker 6. It has a standard deviation of 7.7% of the average source period. Over the 17 speakers the SD ranged from 4.0% to 19.2% with a mean of 7.4%. The broadest histogram came from the female speaker with the highest mean F0 (226Hz). Since the broader histograms clearly deviated from a normal distribution we also estimated a peak-adjusted inter-quartile range spanning 50% of the points. These varied from 1.6% to 5.2% of the source period with a mean of 3.4%.

These phase distributions can be compared with the results reported by Dologlou and Caratannis (1989) of 2.4%. Using their FIR filter with 400 iterations on our data gave a SD of 5.0%. The difference can be attributed to the fact that they used a variable, and possibly higher (>400), number of iterations but also to differences in speaker characteristics.

In addition to the phase variance about the histogram peak for each speaker, there was an inter-speaker variation in histogram peak positions, or absolute phase, of over 1ms. This absolute phase difference showed little variation with the degree of filtering (l). Male and female speakers differed, on average, by 0.45ms or 2 standard deviations. Assuming an average vocal tract length of 17.5 cm (Fant, 1960) and a mouth to microphone distance of 27cm, the total time delay from source to microphone is 1.4ms. The assumption of inter-speaker variations of up to 10cm (including variations of speaker size and head position) can account for approximately 0.3ms of the 1ms inter-speaker variation.

Further analysis of the absolute phase would require consideration of inter-speaker variations in such factors as nasalisation, source acoustic impedance, and relative acoustic radiation characteristics of the mouth, nose and possibly the chest since the microphone was hung on the speaker's chest. Nasalisation, if associated with a low frequency (250Hz) zero could have significant impact on radiated phase for F0 over the 169-226Hz range of the female speakers. Formant bandwidth is another possible candidate since high bandwidth formants have most of their energy in an asymmetric waveform early in the source period and low bandwidth formants have significant residual energy at the end of the source period and are thus likely to be involved in some degree of mode locking with the source. These complexities are beyond the scope of this paper.

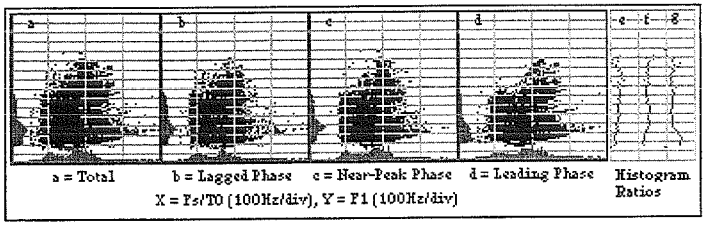

Figure 3. Phase analysis for all speakers displayed as F1 vs F0

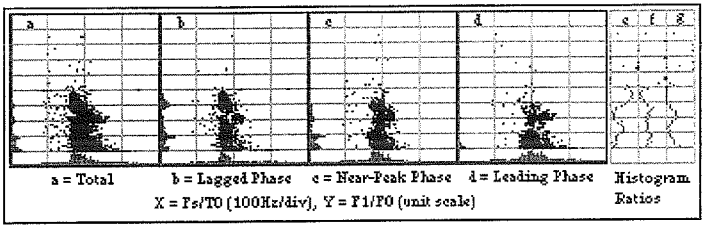

Figure 4. Phase analysis for speaker 6 displayed as F1/F0 vs F0

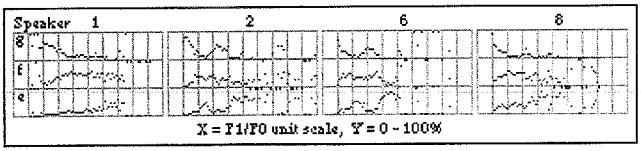

Figure 5. Individual histogram ratios for four speakers

Addressing the first hypothesis, the influence of residual F1, we see the reduction in the number of near-phase marks in figure 3f below 400Hz, with a corresponding increase in 3g, as evidence for residual F1 causing phase variance over all 17 speakers. In figure 4, plotting the ratio F1/F0 against F0 for a single speaker shows more detailed structure related to the F1/F0 resonances in the histogram ratios (e,f,g). This structure is seen in other typical plots for individual speakers in figure 5. While it may be expected that any persistent contribution to phase variance from F1 would be related to integer F1/F0 ratios, this structure should be viewed with some caution since it showed some dependence on the order of the LPC analysis used in determining F1.

Plots such as figure 3 and 4 were calculated for high/low RMS, rising/falling RMS, rising/falling F0 and steady/changing F0 showed similar but less distinct patterns giving a negative result for hypotheses 2 to 5. Calculations for steady/changing F0 over a range of filter iterations (8, 12, 16) showed no clear influence of frequency demodulation effects as a limit to filter iterations.

CONCLUSIONS

No single factor has thus far emerged as the primary determinant of phase variance. Over short segments of speech (~0.1 to 1s) patterns can be observed in both time and frequency domains that disappear on aggregation across time and speakers. Quantitative analysis of short term patterns might prove productive if analysed against a phonetic annotation of the speech. Over longer segments (~30s) some structure in the F1/F0 histogram ratios is discernible in single speaker plots. This is reduced on further aggregation over the 17 speakers to a distinct drop in in-phase marks below 400Hz or for F1/F0 < 2.

Our analysis suggests that nasal and other supraglottal or subglottal zeros, along with acoustic radiation factors are likely to play an important role and possibly dominate both the absolute inter-speaker phase variance and the variance about the mean phase of the detected acoustic radiation relative to the physical vibration of the glottis.

In assessing the overall performance of the FHE approach it is necessary to consider the application domain. For application to high performance LPC coding (Kroon & Atal, 1991) where source timing variations below 20μs were found to be perceptually significant, this technique is unlikely to provide more than an initial estimate for more sophisticated techniques. For more general, and less stringent, applications such as positioning a window for source-synchronous FFT or cepstral analysis, this approach is likely to be adequate for the phonation range explored in this study.

Preliminary tests have shown that considerable advantage can be gained from using an adaptive filter to track F0. We coupled the high-pass prefilter with the low-pass filter and adapted them in step. Since the procedure produces unimodal phase histograms over a wide range of signal conditions, this adaptation need only be a one step process.

The TRLI/FHE technique we have investigated can provide a very fast and reliable method of generating an F0 signal where positioning of markers is not required. It can also provide, for strongly phonated speech, a fast, reliable and moderately accurate estimate of IGC where this is required.

REFERENCES

Dologlou, I & Caratannis, G. (1989) *Pitch Detection Based on Zero Phase Filtering*, Speech Communication 8, 309-318.

Dologlou, I & Caratannis, G. (1991) *A Reply to "Some remarks on the halting criterion for iterative low-pass filtering in a recent proposed pitch detection algorithm" by G. Hult*, Speech Communication 10, 227-228.

Fant, G. (1960) *Acoustic Theory of Speech Production*, (Mouton and Co.: The Hague).

Hess, W. (1983) *Pitch determination of Speech Signals*, (Springer Verlag: Berlin).

Kroon, P. & Atal, B.S. (1991) *On the Use of Pitch Predictors with High Temporal Resolution*, IEEE Transactions on Signal Processing 39, 733-735.

Millar, J.B. O'Kane, M. & Bryant, P. (1989) *Design, Collection and Description of a Database of Spoken Australian English*, Australian Journal of Linguistics 9, 165-189.