

Automatic Glottal Closed-Phase Location and Analysis by Kalman Filtering

John G. McKenna

Centre for Speech Technology Research, University of Edinburgh,
2 Buccleuch Place, Edinburgh, U.K. EH1 1HN, <http://www.cstr.ed.ac.uk>.

& School of Computer Applications, Dublin City University, Dublin 9,

<http://www.compapp.dcu.ie>; john@compapp.dcu.ie

Abstract

In an effort to develop techniques that enhance data-driven techniques in speaker characterisation for speech synthesis, this paper describes a method for automatically determining the location of the closed phase (CP) of the glottal cycle, with subsequent linear predictive (LP) analysis on the CP speech data. Our approach to detecting the CP is designed with the intention of excluding intervals that are *not* within the CP rather than accurately locating the instants of glottal closure and opening. The indicator used is the log determinant of the Kalman filter (KF) estimate error covariance matrix. The CP LP analysis applies a Kalman filter to the CP data only by treating the open-phase data as “missing” and harnessing the non-independence of neighbouring CP spectra. The Kalman filtering process in both techniques is refined to accommodate smoothing, Kalman parameter re-estimation, handling of missing data, and estimation robustification.

1. Introduction

This work forms an important part of our current research in automatic speaker characterisation which is initially based on achieving an automatic division of the glottal excitation function and the vocal tract (VT) filter. The division should facilitate subsequent modelling of both, which in turn should aid manipulation, in pursuit of our goal of speaker characterisation.

Speaker characterisation has important implications for speech synthesis, and speech technology in general. As an example, consider an automatic interpreting system with a speaker characterisation module capable of separating the linguistic information in the speech signal from that which is characteristic of the speaker. By allowing speaker-specific information to input to the synthesis end, we will enjoy the benefit of translated speech which is characteristic of the source speaker. This allows the speaker to maintain their individual identity across the translation medium. Secondly, by removing this speaker-specific information and considering only the linguistic-related information as input to the speech recognition module, we might expect a higher recognition rate.

[1] identify multilingual, multi-speaker, and multistyle speech synthesis as important trends in text-to-speech (TTS) applications. With recent advances in data-driven learning, they point to the need for “at least semi-automatic techniques” in order to collect the necessary data for these applications. [2] also bemoans the “lack of satisfactory methods for continuous and automatic extraction of voice source parameters”. Current automatic techniques offer limited success in estimates of pitch,

glottal events and vocal tract shape. Improvements are found in using pitch-synchronous analysis, but while this type of analysis generally relies on manual intervention, the potential of automation is undeniably immense. [3] also claim that where automatic techniques have been used for source-filter separation, they have been found to work well with modal male voices only and they suggest that more reliable algorithms should be developed for female and pathological voices. We hope that our work here is a major step towards addressing these complaints.

The outline of the paper is as follows. First we outline the topic background. Then we briefly review the principles of the Kalman filter and how we apply it to speech analysis as first reported in [4]. We then step through the method for automatically locating closed-phase data. We illustrate results for both synthetic and real speech.

For concreteness, the discussion below will focus on linear predictor coefficients as VT filter parameters, although other representations are possible. In the plots which we use to illustrate our results, the x -axes represent sample numbers at 16kHz, and rather than plotting LP coefficient trajectories, we plot the formants as obtained from the roots of the characteristic polynomials.

2. Background

2.1. Linear Prediction and Inverse Filtering

Separation of the glottal excitation from the VT tract parameters is quite a common goal and choice of method will often depend on the purpose of the separation. However, it is typically performed using a form of Linear Predictive Coding (LPC) [5].

Conventional fixed-frame pitch-asynchronous LPC [5], typically using the autocorrelation method, builds upon the assumption that the VT articulators are slowly and smoothly varying, and so performs analysis over a number of pitch periods.

However, because fixed-frame analysis is performed during excitation and open phases of the glottal cycle, there are two adverse effects on the estimation of the VT filter parameters when the glottis is open.

Firstly, the vocal tract tube is no longer open at one end – invalidating the LP model. So when the glottis is open, coupling takes place with the subglottal cavity introducing subglottal resonances and antiresonances to the spectrum. These are superimposed on the supraglottal spectrum. The typical effects of this sub-glottal interference are to reduce formant frequencies while increasing formant bandwidths [6]. Thus, if the period of analysis is over both closed and open glottal phases, there will be a smearing or averaging of the parameters, and consequent loss of speaker-characteristic information when we inverse filter with these parameters.

Secondly, the speech is no longer excitation-free. LP autoregressive (AR) analysis techniques assume zero-mean input to the VT filter. This assumption is no longer valid while the glottis is open.

2.2. Glottal Closed Phase Analysis

In an effort to circumvent these problems, it is argued that if the analysis is performed only during the closed phase, when the speech is theoretically an excitation-free decaying oscillation, and the resonances of only the supraglottal VT are responsible for these oscillations, we can more accurately parametrise the VT resonances [7].

However, closed-phase covariance analysis relies on a limited number of sample points; specifically, it requires an analysis window at least the size of the analysis order, which often makes it unsuitable for analysis of female voices.

Closed-phase covariance analysis also assumes constant parameters during the closed phase, and fails to exploit the non-independence of neighbouring spectra. Fixed-frame pitch-asynchronous analysis exploits this non-independence by using overlapping frames, but, as we have already claimed, introduces spectral averaging distortions.

2.3. Stationarity and the Non-independence of Neighbouring Analysis Intervals

During the analysis intervals of the autocorrelation and covariance methods, the signal is assumed to be stationary, i.e. the LP coefficients do not change. This is a reasonable assumption during the steady-state portion of a phone. However, during transitions the stationarity assumption becomes less valid. The typical autocorrelation frame size is 20-40ms. During this time considerable changes in the filter spectrum may occur, for which the autocorrelation method will simply present an “average” spectrum.

Applying the covariance method pitch-synchronously during the glottal closed phase (CP) should produce more accurate estimates during non-stationary parts of the speech signal. However, because the estimates are based on a relatively small number of samples, they have a larger error covariance and the estimated parameters can vary widely from CP to CP.

Efforts have been made to address this issue. [6] use a “multicycle covariance method” which averages covariance estimates over a number of consecutive periods. [8] and [9] apply linear modelling to the dynamics of the formants.

2.4. Glottal Closed Phase Detection

When a closed phase of the glottal cycle is assumed to exist, attempts have been made to locate the CP in order to perform covariance LP analysis. These approaches can be classed as single channel analysis *or* dual channel analysis.

Single channel analysis uses only the speech signal to locate the closed phase. However, because of the difficulty in locating the glottal opening, many of these techniques, e.g. [10, 11], rely on simply estimating the instant of glottal closure (IGC) and assuming that an ad-hoc choice of post-IGC interval length will lie within the closed phase. These lengths are generally chosen to be *either*: a fixed constant length e.g. 2ms; *or* a percentage of the pitch period, e.g. 30%. Other methods, like that of [7], rely on appropriate thresholds being applied.

The methods that rely on using the speech signal alone have proved unreliable in locating the closed phase. Consequently, it has been fairly common for studies and analyses to use a dual-

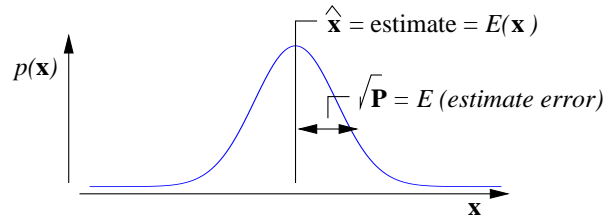


Figure 1: 1-dimensional probability distribution $p(\mathbf{x})$ of coefficient set \mathbf{x} .

channel approach [12, 13], where a laryngograph is used to locate the closed phase. However, this will not be appropriate for speech analysis outside laboratory conditions.

2.5. Conclusion

Conventional LP analysis methods carry many limitations. Our work as presented in [4] overcomes these shortcomings by harnessing the non-independence of neighbouring closed-phase spectra and consequently compensating for small numbers of available closed-phase sample points. This makes it suitable for the analysis of higher-pitched female speech where the smaller number of closed-phase data points available in a single pitch period is compensated by shorter accompanying open phases and a greater number of closed phases per unit time. This is because the rate of movement of the articulators is independent of the fundamental frequency of excitation. The method is also dynamic in that it does not assume stationarity over an interval. We review the technique in Section 3.

In [4], we relied on an laryngograph signal to determine the glottal closed phase, however this is not considered appropriate for automation. It is desirable to be able to determine the closed phase directly from the speech signal. Our automatic approach to this problem is outlined in Section 4.

3. Closed-Phase Kalman Filtering of Speech

3.1. Kalman Filtering

KF [14] permits use of past measurements to produce a priori estimates for prediction and corresponding confidence gauges of the subsequent a posteriori estimates. The state-space equations are given as:

$$s_n = \mathbf{H}_n \mathbf{x}_n + v_n \quad n = 1, 2, \dots, N \quad (1)$$

where s_n , the *measurement*, is the speech at time n ; \mathbf{x}_n , the *state*, is the set of p LPC predictor coefficients, $[a_1 \dots a_p]^T$, which are linearly related to s_n by \mathbf{H}_n a number of preceding points, $[s_{n-1} \dots s_{n-p}]$; v_n is the measurement noise, assumed Gaussian with probability distribution $p(v) \sim N(0, R)$.

$$\mathbf{x}_n = \mathbf{\Phi} \mathbf{x}_{n-1} + \mathbf{w}_n \quad n = 1, 2, \dots, N \quad (2)$$

where $\mathbf{\Phi}$ directs the current a posteriori state estimate to the a priori estimate of the state at the next time step; \mathbf{w}_n is the process noise, with probability distribution $p(\mathbf{w}) \sim N(0, \mathbf{Q})$.

While we track \mathbf{x}_n , we also maintain a confidence measure in the form of an error covariance matrix, \mathbf{P}_n , which is also updated at each stage (see Figures 1 and 2).

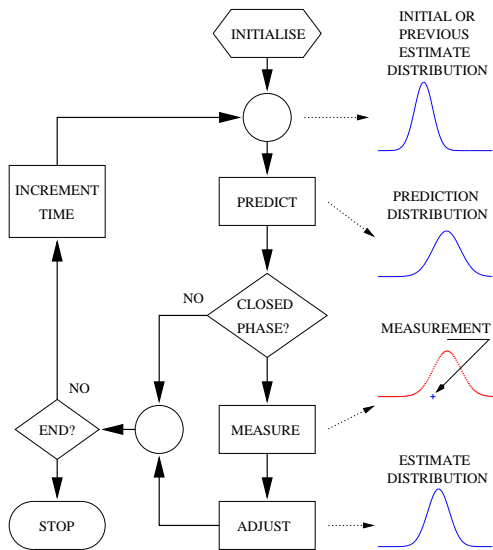


Figure 2: Kalman filtering with robustification scheme for using only closed-phase data.

The Kalman filter recursively bases the current prediction on all past measurements. In updating the state estimate, $\hat{\mathbf{x}}_n$, the smaller the measurement error variance R , the more trust is placed in the actual measurement s_n . Conversely, as the measurement error variance R outweighs the a priori measurement estimate error variance $\mathbf{H}_n \mathbf{P}_n \mathbf{H}_n^T$, more trust is placed in the a priori predicted measurement $\mathbf{H}_n \hat{\mathbf{x}}_n$ than in the actual measurement.

3.2. Kalman Parameter Reestimation

There is also the practical issue of choosing the initial values of the Kalman parameters. We use an EM iterative technique [15] which having made a forward-backward iteration through all the data, presents appropriate initial filter parameter values for Φ , \mathbf{Q} , R (the three Kalman parameters whose values are most important), and \mathbf{x}_0 , for use in the next iteration. The technique is based on the Kalman forward equations [14] and the Rauch-Tung-Streifel backward equations [16]. During the forward part of each iteration, a log-likelihood score can be calculated and is guaranteed to increase.

While convergence is guaranteed using this technique, careful choice of the initial parameters on the first iteration can greatly reduce the number of further iterations necessary for convergence. The initial values of Φ , \mathbf{Q} , R used in the closed phase analysis are derived from the first pass that is used to locate the closed phases. This is discussed in Section 4.2.

Unlike [17, 18], reestimation of Φ allows us to predict movement of the predictor coefficients from point to point using a non-identity matrix. In other words, rather than attributing any change in the coefficients solely to noise or error, we are able to reduce the uncertainty by capturing a certain amount of predictable movement in a non-identity matrix.

3.3. Robustification and Missing Data

We can robustify our estimates by excluding “undesirable” sections of data. In CP analysis we wish to exclude non-CP data. Reasonable estimates can be made through sections of missing

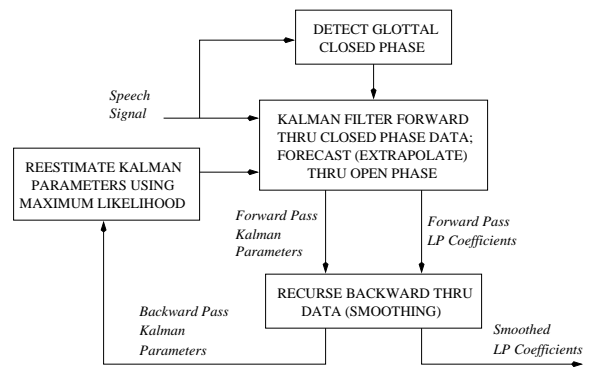


Figure 3: Architecture of closed-phase Kalman filter linear prediction system.

data as long as there are no significant changes of direction in the underlying process during the interval where the data is missing.

For example, when we choose to use only closed-phase data, we can exclude other data points by using the system as in the flow chart of Figure 2. The estimates for excluded-data intervals are simply $\Phi \mathbf{x}_{n-1}$, the a priori state estimates without measurement update; uncertainty is added to each such estimate by adding \mathbf{Q} to $\Phi \mathbf{P}_{n-1} \Phi^T$ – i.e. the a priori estimate error covariance.

The architecture of our closed-phase Kalman filter linear prediction system is sketched in Figure 3.

4. Glottal Closed Phase Location

We shall now show how Kalman filtering can be applied to the problem of locating closed phase samples. We begin by discussing the preprocessing of the speech signal.

4.1. Preprocessing

Firstly, fixed-frame linear prediction analysis using the autocorrelation method is performed on the preemphasised speech signal. We then inverse filter to obtain a fixed-frame residual.

The residual is then rectified and then moving-median filtered to exclude the large impulses which occur at points of excitation. We then calculate the power of the median-filtered signal. This power value will serve as an initial estimate for the Kalman parameter R_n – the variance of the measurement noise. In other words, we have initially guessed the noise, or error, element of our AR modelled speech to be that of the fixed-frame residual with the excitatory impulses filtered out.

We would like the analysis to be robust against the excitatory spikes that tend to throw the estimation process out of step. This was a weakness in previous approaches [17, 19] which produced staggered parameter trajectories. [18] introduces some robustness to the algorithm to counteract the influence of the glottal closure on the parameter extraction.

As explored in [20], we choose to use a 3-sigma hard rejection robustness criterion – i.e. we ignore data at sample points where the a priori measurement error exceeds 3 times the expected error (i.e. $3\sqrt{R}$). These data points are treated as missing (see Section 3.3).

4.2. Initial Kalman Parameters

Φ was chosen as the identity matrix as we assume no prior knowledge of the VT parameter trajectories, meaning we initially assume that they remain approximately the same from one sample to the next.

\mathbf{Q} was empirically set to a diagonal matrix: $\text{diag}(5 \times 10^{-5})$, which is large enough to allow significant variation in the LP parameters.

The LPC coefficients, \mathbf{x}_0 , were set to zero; the initial estimate error covariance, \mathbf{P}_0 , is fixed throughout the iterations at a reasonable baseline level which we set at $\text{diag}(5 \times 10^{-5})^1$.

R is most dependent on the particular speech being analysed in that it will depend greatly on the intensity of the signal. Therefore, this is derived from the power in the median-filtered rectified fixed-frame residual as discussed in Section 4.1.

We mentioned in Section 3.2 that careful choice of initial Kalman parameter values can help speed up convergence. For our purposes of closed phase determination, we found that our initial values required only two forward-backward iterations to provide satisfactory results – and which did not improve significantly on subsequent iterations.

For closed phase analysis, we used three iterations. The initial values of Φ , \mathbf{Q} and \mathbf{x}_0 used in the CP analysis pass were obtained from reestimation after the last iteration of the CP location pass. R is taken to be the power in the residual (as obtained from the the last iteration of the CP location pass) over all the CPs as determined by our method.

4.3. Discussion and Results

Initially, due to the ability of the Kalman filter to track dynamics, we expected to find variation in the formants (obtained from root-solving the predictor polynomial) consistent with the glottal open and closed phases. However, we found that the variation, while existent, was inconsistent across the formants (see Figure 4).

We then, as [21] did, looked to the covariance of the estimate error, where again we found variation. In an attempt to gauge the magnitude to the error covariance, we calculated the determinant of the a posteriori error covariance matrix at each sample time. While we found significant variations synchronous with the open and closed phases, the magnitude of the variations required us to apply a log operation.

We also found that there tended to be considerable low-frequency drift on the log-determinant function. To eliminate this and preserve the local variations, we applied a high-pass filter whose cutoff frequency was a function of the local pitch period as estimated from the method of [22].

We then apply a “ $< \mu - \sigma$ ” thresholding criterion, where μ is a local mean and σ is a local standard deviation from a window which is made equal to the local pitch period. In previous studies, e.g. [12], a “ $< 50\%$ ” threshold is used on the laryngograph signal in deciding the boundaries of the closed phase. We opt here for a more conservative “ $< \mu - \sigma$ ” which proved to be a practical-yet-safe criterion.

Examples of the results we obtain are found in Figures 5 and 6.

Examples of results of the subsequent closed-phase analysis are plotted in Figures 7 – 10. It should be noted that for the duration of a segment, Φ_n , \mathbf{Q}_n , R_n are kept constant. This is reasonable for a short segment of speech - like a monophthong

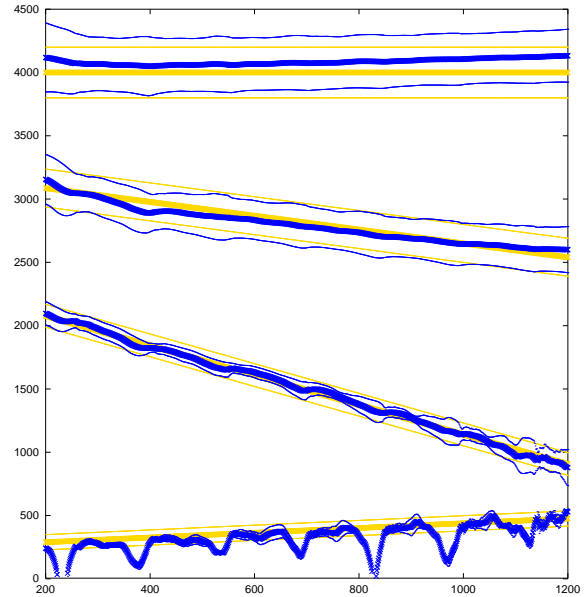


Figure 4: Formant estimates of synthetic speech from Kalman filtering through all data. Bandwidth delimiters are shown with thin lines. Lighter lines represent true formants; darker represent estimates.

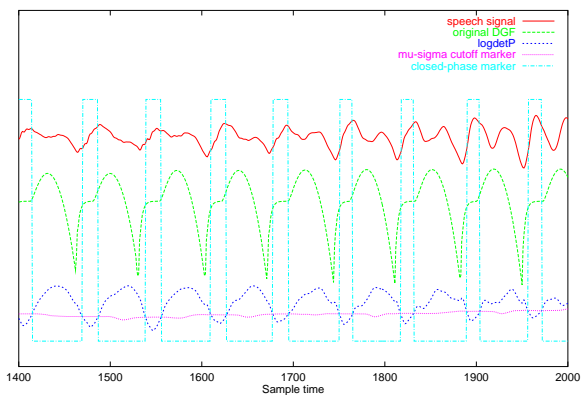


Figure 5: Closed phase location in synthetic female speech; $F_0 \approx 230$ Hz.

¹We plan to carry out further studies on more robust choices of these baseline values.

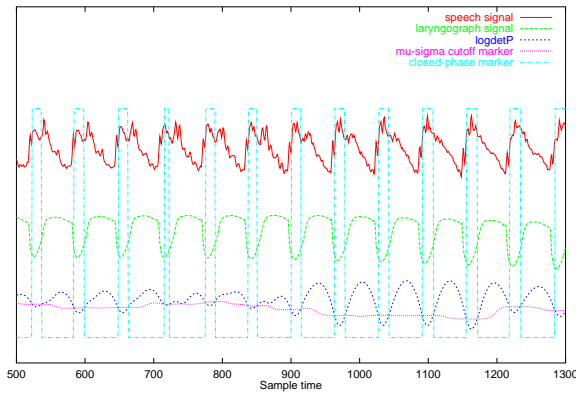


Figure 6: Closed phase location in real female speech; $F_0 \approx 250$ Hz.

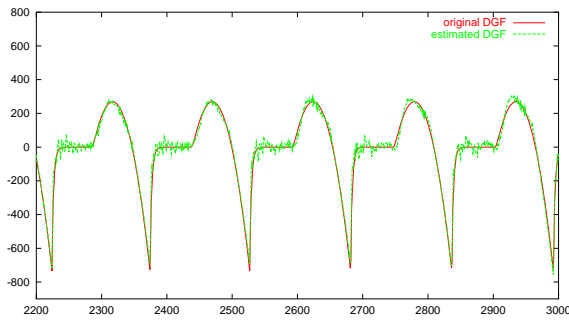


Figure 7: DGF estimation from synthetic male speech.

or diphthong. However, time-varying values of the Kalman parameters should ideally be used over longer segments of continuous voiced speech. This is highlighted in Figure 8 where Φ causes a deterioration in tracking at the beginning of the segment and during the open phases where it is responsible for interpolating estimates. Parameter trajectories with sharp turning points or unnaturally straight trajectories may also pose difficulties for Φ . Fortunately, we can expect smoother trajectories in real speech (see Figure 9).

5. Conclusion

5.1. CP Location

It is clear that an approach that is automatic, uses only the speech signal, and defines an appropriate beginning *and* end to the closed phase will be an important advance on the current state of affairs. Our novel technique has these qualities.

5.2. CP Analysis

We have highlighted the flaws associated with conventional methods of LP analysis. Fixed-frame (autocorrelation method) analysis averages over several successive glottal cycles, averages over closed and open phases of the glottal cycle, and does not handle non-stationarity well. Conventional CP (covariance method) analysis makes independent estimates for each CP, requires a certain number of CP data samples in each CP, and is often unsuitable for female analysis.

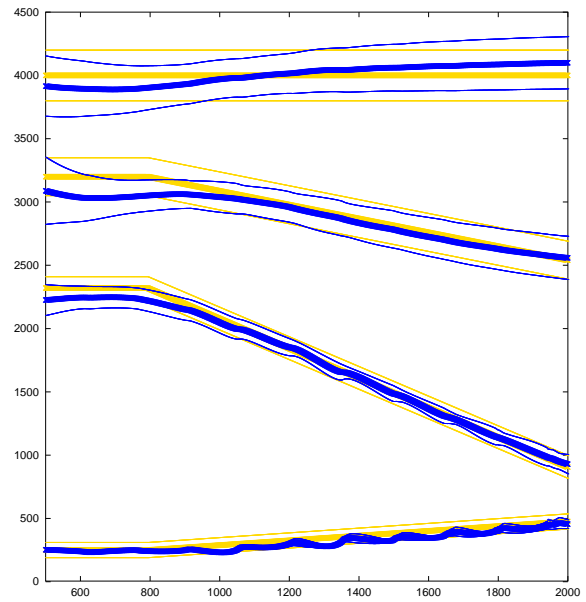


Figure 8: Formant estimation from synthetic male speech. Bandwidth delimiters are shown with thin lines. Lighter lines represent true formants; darker represent estimates.

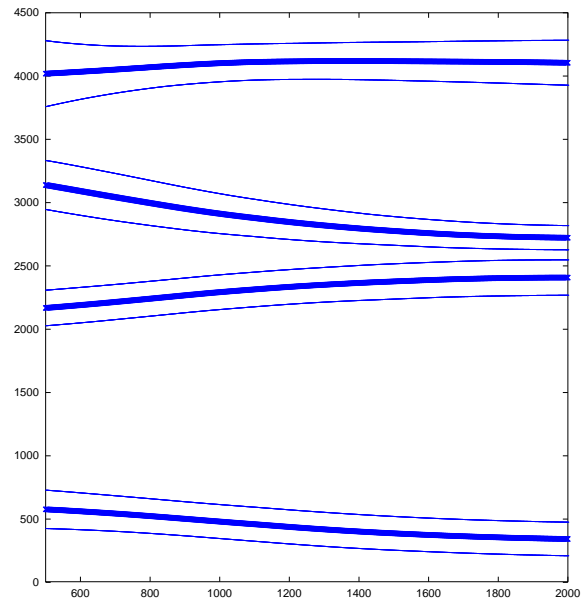


Figure 9: Formant estimation from real female speech: diphthong /ai/. Bandwidth delimiters are shown with thin lines.

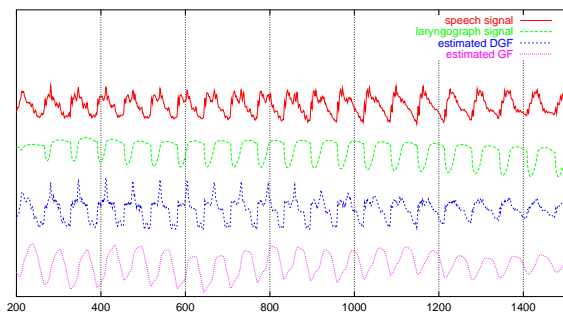


Figure 10: DGF and GF estimation from real female speech.

Our method overcomes these and offers accurate separation of source and filter, smooth trajectories that ease modelling, and sets a solid foundation for tackling speaker characterisation for speech synthesis.

6. Future Work

In CP location, the determinant of the estimate error covariance is influenced by the magnitude of the speech signal. We would like to remove this dependence using some form of normalisation. Our initial attempts, like those of [21], have not produced results of any significance. Further investigation is desirable.

The research to date has been primarily on vowels. We would like to extend our investigations to other sounds – particularly those that require ARMA analysis such as nasals.

7. Acknowledgements

Many thanks to Steve Isard for his advice throughout this project.

John McKenna was supported by UK Engineering and Physical Science Research Council Studentship Award Ref. No. 96307273 while this work was carried out.

8. References

- [1] R. Carlson and B. Grantström, “Speech synthesis,” in *The Handbook of Phonetic Sciences* (W. H. Hardcastle and J. Laver, eds.), ch. 26, pp. 768–788, Blackwell, 1997.
- [2] G. Fant, “Some problems in voice source analysis,” *Speech Communication*, vol. 13, pp. 7–22, 1993.
- [3] M. Lee and D. G. Childers, “Manual glottal inverse filtering algorithm,” in *Proceedings of the IASTED International Conference on Signal and Image Processing (SIP '96)*, (Orlando, Florida), pp. 34–37, November 1996.
- [4] J. McKenna and S. Isard, “Tailoring Kalman filtering towards speaker characterisation,” in *Proceedings of Eurospeech 99*, vol. 6, (Budapest), pp. 2793–2796, 1999.
- [5] J. D. Markel and A. H. Gray, Jr., *Linear Prediction of Speech*. New York: Springer-Verlag, 1976.
- [6] B. Yegnanarayana and R. N. Veldhuis, “Extraction of vocal-tract system characteristics from speech signals,” *IEEE Transactions on Speech and Audio Processing*, vol. 6, pp. 313–327, July 1998.
- [7] D. Y. Wong, J. D. Markel, and A. H. Gray, Jr., “Least squares glottal inverse filtering from the acoustic speech

waveform,” *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 27, pp. 350–355, August 1970.

- [8] Y.-T. Lee and H. F. Silverman, “A model for nonstationary analysis of speech,” in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, (Tokyo), pp. 1617–1620, 1986.
- [9] K. Nathan, Y.-T. Lee, and H. F. Silverman, “A time-varying analysis method for rapid transitions in speech,” *IEEE Transactions on Signal Processing*, vol. 39, no. 4, pp. 815–824, 1991.
- [10] D. H. Deterding, “Pitch-synchronous linear prediction,” *Cambridge Papers in Phonetics and Experimental Linguistics*, vol. 5, pp. 1–13, 1986.
- [11] D. Childers and C. K. Lee, “Vocal quality factors: Analysis, synthesis and perception,” *Journal of the Acoustical Society of America*, vol. 90, pp. 2394–2410, November 1991.
- [12] D. E. Veeneman and S. L. BeMent, “Automatic glottal inverse filtering from speech and electroglottographic signals,” *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 33, pp. 369–377, April 1985.
- [13] A. K. Krishnamurthy and D. G. Childers, “Two-channel speech analysis,” *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 34, no. 4, pp. 730–743, 1986.
- [14] R. E. Kalman, “A new approach to linear filtering and prediction problems,” *Transactions of the ASME Journal of Basic Engineering*, vol. 8, pp. 35–45, 1960.
- [15] R. H. Shumway and D. S. Stoffer, “An approach to time series smoothing and forecasting using the EM algorithm,” *Journal of Time Series Analysis*, vol. 3, no. 4, 1982.
- [16] H. E. Rauch, F. Tung, and C. T. Streibel, “Maximum likelihood estimates of linear dynamic systems,” *AIAA Journal*, vol. 3, pp. 1445–1450, 1965.
- [17] M. Niranjana, I. J. Cox, and S. Hingorani, “Recursive tracking of formants in speech signals,” in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pp. 205–208, 1994.
- [18] T. Yang, J. H. Lee, K. Y. Lee, and K. M. Sung, “On robust Kalman filtering with forgetting factor for sequential speech analysis,” *Signal Processing*, vol. 63, pp. 151–156, 1997.
- [19] G. Rigoll, “A new algorithm for estimation of formant trajectories directly from the speech signal based on an extended Kalman filter,” in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, (Tokyo), pp. 1229–1232, 1986.
- [20] B. D. Kovačević, M. M. Milosavljević, and M. D. Veinović, “Robust recursive AR speech analysis,” *Signal Processing*, vol. 44, pp. 125–138, 1995.
- [21] H. W. Strube, “Determination of the instant of glottal closure from the speech wave,” *Journal of the Acoustical Society of America*, vol. 56, no. 5, pp. 1625–1629, 1974.
- [22] D. Talkin, “A robust algorithm for pitch tracking (RAPT),” in *Speech Coding and Synthesis* (W. B. Kleijn and K. K. Paliwal, eds.), ch. 14, pp. 495–518, Elsevier, 1995.