TAILORING KALMAN FILTERING TOWARDS SPEAKER CHARACTERISATION

John McKenna and Stephen Isard

Centre for Speech Technology Research, University of Edinburgh,
80, South Bridge, Edinburgh, U.K. EH1 1HN
http://www.cstr.ed.ac.uk
email: {john, stepheni,}@cstr.ed.ac.uk

ABSTRACT

This paper describes a method for obtaining smoothed vocal tract parameters from analysis during the closed phase of the glottis. The method is based upon Expectation Maximisation (EM) and uses Kalman-Rauch forward-backward iterations through a voiced segment, in which the speech data during excitation and open phases are excluded by treating them as 'missing data'.

This approach exploits the non-independence of neighbouring spectra and compensates for small numbers of available points, while preserving speaker-characteristic information and tracking variations in it.

The vocal tract filter parameters are then used for inverse filtering the speech, thus obtaining estimates of the source excitation. The extracted excitation signal can be used to excite other sets of parameters to produce natural sounding speech.

1. INTRODUCTION

The ultimate goal of our current research is cross-language voice transfer: to separate the speaker's identity from the linguistic content in a way that facilitates speech synthesis in a language alien to the speaker whose voice quality we use in the synthesis. Applications for this goal can be found in areas such as automatic voice-to-voice translation and foreign language learning.

We would like to parametrise the movement of the vocal tract (VT) articulators so that when the speech is inverse-filtered with these parameters, we obtain an excitation function distinctive to the speaker. Of course, the excitation function alone will not completely characterise the speaker, as the speaker's articulatory habits (in the guise of the VT parameter values and trajectories) will also contribute somewhat to the speaker's individuality.

However, our aim here is to achieve a division of the glottal excitation function and the vocal tract filter in such a way as to facilitate modelling of both, which in turn should aid manipulation, in pursuit of our goal of voice transfer.

Our approach is based on the following assumptions:

- The supraglottal articulators move relatively smoothly.
- During the glottal closed phase, the VT is a linear filter, the parameters of which are a consequence of the positioning of the supraglottal articulators.
- During the glottal closed phase, the speech signal is representative of the impulse response of the VT filter and hence should be predictable from the VT filter with minimal error.

Given these assumptions, if we obtain smooth parameter trajectories, based on analysis during the closed phase, *and* a minimal residual signal (from inverse filtering) during the closed phase, then we should have characterised the supraglottal articulatory movements. It is important to insist on both criteria simultaneously. There are many ways of smoothing filter coefficients - as a reductio ad absurdum they can be set to arbitrary constant values - but at the expense of pushing important spectral properties of the speech into the residual. Conversely, if the filter coefficients are updated at every sample they can be made to predict the speech perfectly, but only by jumping in synchrony with the source excitation.

Since the smoothing algorithm we use provides interpolation for data outwith the closed phase, the residual during these times should be representative of the excitation, glottal and subglottal coupling, and nonlinearity that are present.

The real measure of success in our goals is the subjective impressions of the re-synthesised speech, but as objective criteria we adopt measures of the residual error during the closed phase, and smoothness of the filter parameters.

For concreteness, the discussion below will focus on linear predictor coefficients as VT filter parameters, although other representations are possible. In the waveform plots, an arbitrary subgroup of predictor coefficients are displayed and can be taken to be representative of the general behaviour of the coefficients. The *x*-axes represent sample numbers at 16kHz.

2. BACKGROUND & THEORY

Separation of the glottal excitation from the vocal 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) [4].

Conventional fixed frame pitch-asynchronous LPC [4] also builds upon the assumption that the VT articulators are slowly and smoothly varying. However, when the glottis opens, there tends to be sub-glottal interference, which affects the formants and their bandwidths [11]. 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.

In an effort to circumvent this problem, it is argued that if the analysis is performed only during the closed phase, when the speech is theoretically an excitation-free decaying oscillation, we can more accurately parametrise the VT resonances [9].

Comparative studies [2, 3] of such analyses highlight their relative merits and demerits. Closed-phase analysis relies on a limited number of sample points, assumes constant parameters during the closed phase, and fails to exploit the non-independence of neighbouring spectra. As previously stated, pitch-asynchronous analysis, while exploiting this non-independence, introduces spectral averaging distortions.

Our work exploits recent variants of, and alternatives to,

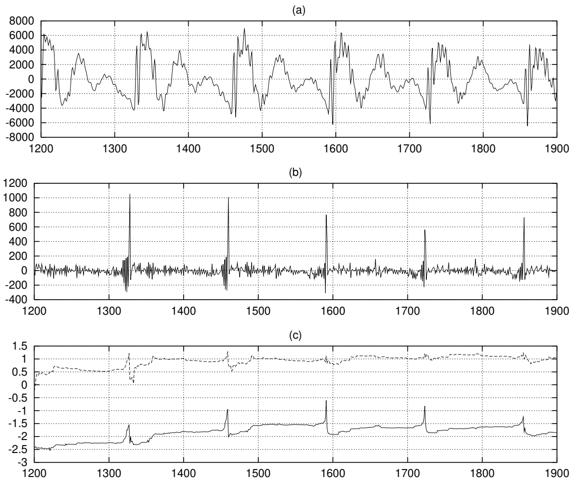

Figure 1: Results from ordinary Kalman filtering: (a) speech waveform; (b) residual waveform; (c) 2nd and 6th linear predictor coefficients

standard LPC, as in [5, 10]. It exploits the non-independence of neighbouring spectra and compensates for small numbers of available points: Kalman Filtering (KF) tailored to simultaneously produce a minimal closed-phase residual and smooth trajectories of the predictor coefficients.

2.1. Kalman Filtering

KF [1] 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:

$$s_n = \mathbf{H}_n \mathbf{x}_n + v_n \qquad n = 1, 2, \ldots, N \tag{1}$$

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

$$\mathbf{x}_n = \mathbf{\Phi} \mathbf{x}_{n-1} + \mathbf{w}_n \qquad n = 1, 2, \ldots, N \tag{2}$$

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 $p(\mathbf{w}) \sim N(\mathbf{0}, \mathbf{Q})$.

While we track $\mathbf{x}_n$ as our best estimate of the current state, we also maintain a confidence measure in the form of an error covariance matrix, $\mathbf{P}_n$, which is also updated at each stage.

The Kalman filter recursively bases the current prediction on all past measurements. In the 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 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.

Kalman filtering has been applied to speech analysis in the past but its value has often been underestimated due to its computational overhead [4]. This argument has weakened in recent years. [5] uses an extended Kalman filter (EKF) to directly track the formants. [10] tracks LPC coefficients. In both cases the tracking is allowed to vary during the closed and open phases, and also during the brief period of maximal excitation (usually just prior to the instant of glottal closure). This leads to a reasonable residual, but staggered parameter trajectories (Fig. 1). [10] introduces robustness to the algorithm to counteract the influence of the glottal closure on the parameter extraction (Fig. 2).

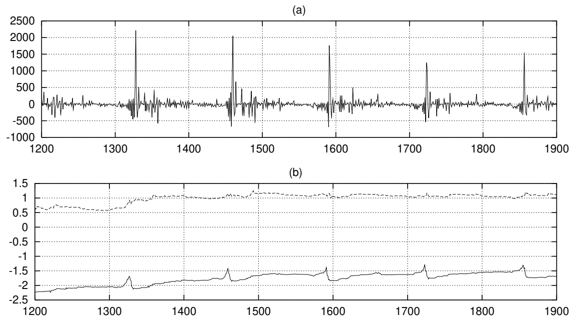

Figure 2: Results from robust Kalman filtering: (a) residual waveform; (b) 2nd and 6th linear predictor coefficients

These diagrams also illustrate the tradeoff between smooth coefficient trajectories and prediction error. Examination of the y-axes shows that the smoother coefficients are obtained at the cost of greater prediction error.

There is also the practical issue of choosing the initial values of the Kalman parameters. The studies we have cited do not discuss this issue, but presumably they use reasonable values based on experience and as much a priori knowledge of the speech as is available.

2.2. Tailoring the EM Algorithm

The method we propose has the following properties:

1. We add robustness to our KF by only considering data from the closed phase of the glottal cycle, thus eliminating the adverse effects of estimates during glottal excitation and the glottal open phase. The reasons for this were outlined in Section 1. For convenience, we obtain our estimate of the closed phase from an EGG (electroglottograph, or laryngograph) signal [8].

2. Unlike [5, 10], we predict movement of the predictor coefficients from point to point using a non-identity matrix for $\mathbf{\Phi}$. 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. We use an EM iterative technique [7] which having made a forward-backward pass through the all the data, presents appropriate initial filter parameter values for $\mathbf{\Phi}$, $\mathbf{Q}$ and R for use in the next pass. The technique is based on the Kalman forward equations [1] and the Rauch backward equations [6].

   The speech data points we believe to occur during the open phase and glottal closure are omitted from the analysis in the guise of 'missing data'.

On our first iteration through the speech, we must choose some initial filter parameter values. The LPC coefficients were set to zero; R was set approximate to the power of the silent segment prior to the onset of speech; $\mathbf{Q}$ was set to a diagonal matrix: $diag(10^{-5})$, an arbitrary small figure; $\mathbf{\Phi}$ was chosen as the identity matrix as we assume no prior knowledge of the

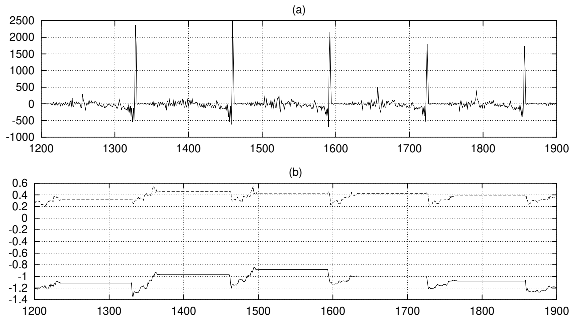

Figure 3: Results from Kalman filtering with a single forward pass, using analysis data from the closed phase only: (a) residual waveform; (b) 2nd and 6th linear predictor coefficients

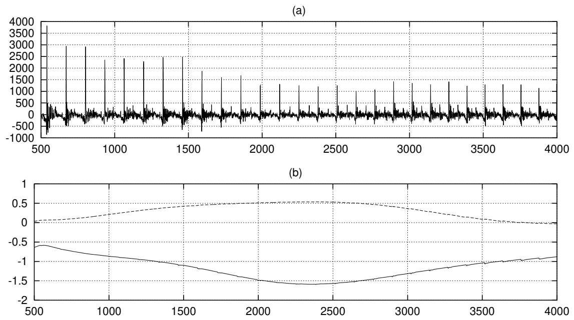

Figure 4: Results from Kalman filtering with a forward-backward-forward pass, using analysis data from the closed phase only: (a) residual waveform; (b) 2nd and 6th linear predictor coefficients

VT parameter trajectories, meaning we initially assume that they remain approximately the same from one sample to the next.

If we simply Kalman filter using a single forward iteration we achieve the results as illustrated in Fig. 3. Here, the closed-phase section is easily distinguished. It is characterised by a very small prediction error and coefficient estimates which waver considerably. Because we chose an identity matrix for $\mathbf{\Phi}$, the interpolated coefficient values during the open phase maintain a horizontal trajectory. Once we perform a forward backward iteration, we obtain a non-identity matrix for $\mathbf{\Phi}$, and our trajectories follow a smoother path.

Having backtracked through the data and taken into consideration the all the forward-pass parameters at each point, the algorithm arrives at a new set of initial parameters, with which a new forward pass is undertaken. During each forward pass, a log-likelihood score is calculated. We find that a second forward pass (i.e. after just one backward pass) produces satisfactory results which can be viewed in Fig. 4

3. COMPARISON WITH CLOSED-PHASE COVARIANCE ANALYSIS

Fig. 5 shows the result of a closed-phase covariance (CPC) analysis [9]. For this analysis, we took care to ensure that the coefficient values were obtained from a window in the closed phase, because, as reported in [3], smaller normalised errors can occur during the open phase.

[9] and other studies have often used sustained vowels. As there will be little variation in the parameters from one frame of analysis to the next, one set of parameter values can be used to

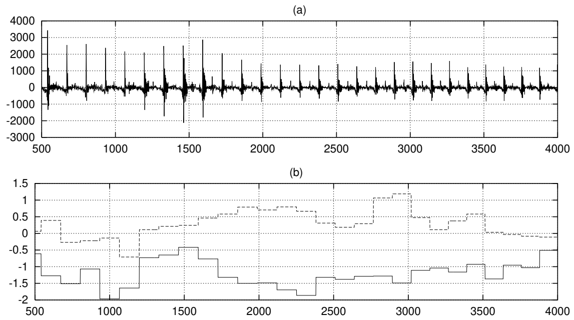

Figure 5: Results from the closed-phase covariance (CPC) method [9]: (a) residual waveform; (b) 2nd and 6th linear predictor coefficients

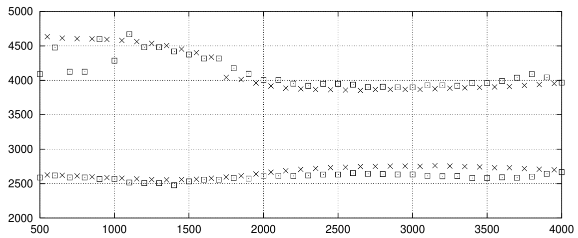

Figure 6: 3rd and 5th formants as estimated by the KF (×) and CPC (□) methods

inverse-filter a complete pitch period.

As in the preceding discussion and accompanying illustrations we take the voiced segment /əji/ as might be found in "a year", since we are interested in tracking time-varying VT parameters.

The speech was that of a north American male, aged 25, and considered to speak with normal phonation (as opposed to breathy, etc.). It was sampled at 16kHz, preemphasized, and an analysis order of 16 used.

The predictor coefficient trajectories produced by our method (Fig. 4) are far smoother than those derived using CPC analysis in Fig. 5, even after discounting the step effect due to using a single set of parameters for the entire pitch period.

Interestingly, we found that the corresponding formant trajectories are comparable in the lower formants, but the higher formant trajectories derived from the CPC analysis tend to be more variable, whereas our method produces steady formant trajectories, as illustrated in Fig. 6.

We also found that the power in the error signal as measured over all the closed phases in the segment, was approximately 3 times higher in the CPC method.

The covariance method also requires an analysis window at least the size of the analysis order, whereas our approach has no such limitations. This is promising 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. [11] also make use of the fact that the higher the fundamental frequency, the less variation in parameters from pitch period to pitch period.

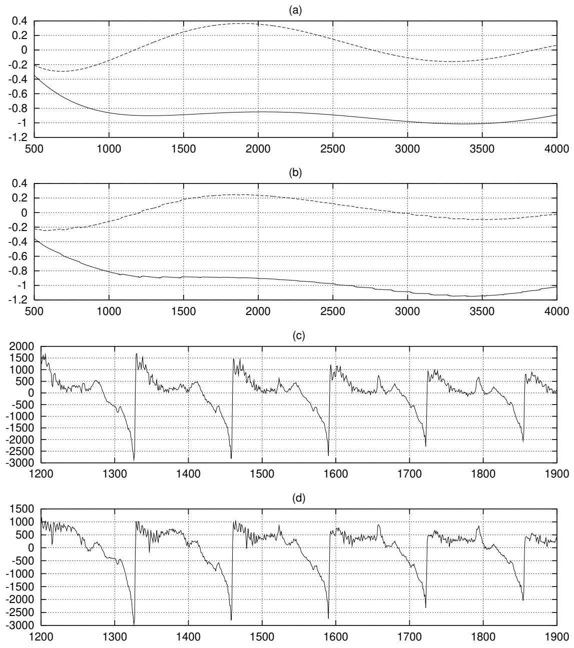

Figure 7: Comparison of synthesis (a) and (c), and analysis (b) and (d) components of the hybrid speech

4. PARAMETER SWAPPING

Smooth coefficient trajectories can be modelled using a low-order polynomial. We fitted a fifth order polynomial to each of the parameters exemplified in the Fig. 4, giving nearly a 700-fold data reduction (for the filter parameters).

We inverse-filtered the original unpreemphasised speech to obtain an estimate of the differentiated glottal flow (DGF) waveform (Fig. 7(c)). We found that resynthesis by exciting the polynomial-fitted coefficients with this estimated DGF produced speech perceptually indistinguishable from the original. A parametrisation of this kind of this kind is also easily manipulated in the time domain, either by stretching or time-warping.

In order to test our separation, we performed a ‘voice swapping’ experiment. We took the same utterance as spoken by a British speaker and analysed it as described in Section 2.2. The polynomial-fitted coefficients were time-stretched to match the length of the north American speaker’s speech and re-excited using his excitation.

The resulting hybrid speech was not clearly identifiable as having been uttered by the north American speaker, but was very natural sounding.

Furthermore, the hybrid speech provides us with a case where we have a ‘right answer’ for the source-filter separation to aim at. Fig. 7 shows the stretched form of the polynomial fitted to two of the predictor coefficients derived from the original ‘real’ speech (a) compared to the same coefficients extracted from the hybrid speech (b). Fig. 7(c) shows the DFG excitation of the hybrid speech. Fig. 7(d) shows the corresponding estimate from our analysis of the hybrid speech.

5. FUTURE WORK

As noted in [11] and elsewhere, vocal tract coupling makes the filter properties during the open phase different from those of the closed phase. In particular, the open phase is characterised by more damping, or higher formant bandwidths. At the moment, our form of analysis estimates the filter coefficients of the open phase as if it were another portion of the close phase, thereby transferring some filter properties to the residual. This in turn interferes with the pairing of the residual with new filter coefficients, either from the same or a different speaker. We are working on more accurate open phase analysis.

We will also incorporate vocal tract - or formant space - normalisation into our voice transfer methods to get more convincing speaker characterisation.

6. ACKNOWLEDGEMENTS

John McKenna is supported by UK Engineering and Physical Science Research Council Studentship Award Ref. No. 96307273.

REFERENCES

[1] R. E. Kalman. A new approach to linear filtering and prediction problems. *Trans. ASME J. of Basic Eng.*, 8:35–45, 1960.

[2] A. K. Krishnamurthy and D. G. Childers. Two-channel speech analysis. *IEEE Trans. ASSP*, 34(4):730–743, August 1886.

[3] J. N. Larar, Y. A. Alsaka, and D. G. Childers. Variability in closed phase analysis of speech. In *Proc. ICASSP*, volume 3, pages 1093–1096, 1985.

[4] J. D. Markel and A. H. Gray, Jr. *Linear Prediction of Speech.* Springer-Verlag, New York, 1976.

[5] M. Niranjan, I. J. Cox, and S. Hingorani. Recursive tracking of formants in speech signals. In *Proc. ICASSP*, volume 2, pages 205–208, 1994.

[6] H. E. Rauch. Solutions to the linear smoothing problem. *IEEE Trans. Automatic Control*, 8:371–372, 1963.

[7] R. H. Shumway and D. S. Stoffer. An approach to time series smoothing and forecasting using the EM algorithm. *J. of Time Series Analysis*, 3(4), 1982.

[8] D. E. Veeneman and S. L. BeMent. Automatic glottal inverse filtering from speech and electroglottographic signals. *IEEE Trans. ASSP*, 33(2):369–377, April 1985.

[9] D. Y. Wong, J. D. Markel, and A. H. Gray, Jr. Least squares glottal inverse filtering from the acoustic speech waveform. *IEEE Trans. ASSP*, 27(4):350–355, August 1970.

[10] 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*, 63:151–156, 1997.

[11] B. Yegnanarayana and R. N. J. Veldhuis. Extraction of vocal-tract system characteristics from speech signals. *IEEE Trans. Speech and Audio Processing*, 6(4):313–327, July 1998.