DECOMPOSITION OF SPEECH INTO VOICED AND UNVOICED COMPONENTS BASED ON A STATE-SPACE SIGNAL MODEL

Mark Thomson[1], Simon Boland[2], Mike Wu[1], Julien Epps[1], Michael Smithers[3]

[1]Motorola Labs, Botany, NSW, Australia
[2]Cross Avaya R & D, North Ryde, NSW, Australia
[3]Dolby Laboratories, San Francisco, CA, USA

ABSTRACT

We present a novel method for decomposing speech into voiced and unvoiced components. After demodulating variations in spectral envelope, energy and pitch, the method involves applying a bank of Kalman filters to separate the harmonic and non-harmonic components of the signal. This approach relies on a state-space representation of the composite signal, and provides a way to accurately estimate the harmonic component without the large delay required by a linear phase comb filter. However it also requires prior knowledge of the variance of the unvoiced component and the state transition parameters. We present a novel method to accurately determine these parameters based on a variant of the Expectation-Maximization algorithm. Modifications for dealing with unvoiced segments and voicing onset are also described.

1. INTRODUCTION

The distinction between voiced and unvoiced sounds is important in many areas of speech technology. This is particularly true in speech coding, where different mechanisms are often used to encode the voiced and unvoiced parts of speech [6].

In early low bit rate vocoders, it was common to classify any particular segment of speech as being either purely voiced or purely unvoiced [1]. In reality, however, many speech segments contain significant amounts of both quasiperiodic and noise-like energy. For this reason, in many more recent parametric coders, a measure of voicing strength is used to control the relative amount of periodic and non-periodic energy in the excitation of a linear prediction filter [4].

In other cases, an attempt is made to explicitly separate the voiced and unvoiced components. In codebook-excited linear predictive coders, for example, analysis-by-synthesis is used to identify optimal contributions to the vocal tract excitation from adaptive and fixed codebooks, representing voiced and unvoiced energy respectively [3].

In contrast, in interpolation-based coding a linear low pass filter is used to separate slowly evolving and rapidly evolving components of the pitch cycle [6], corresponding to voiced and unvoiced speech. This is similar to the use of a linear comb filter to isolate the voiced component of speech based on its harmonic structure. A limitation of this approach, however, is that its effectiveness depends on having a filter with a sharp roll-off. This requires a long impulse response and an undesirably large delay, and also creates difficulties in dealing with rapid transitions.

Achieving good decomposition without a large delay requires the use of more prior knowledge about signal behavior. One approach is to impose a deterministic parametric model on the evolution of the harmonic coefficients [8]. However, the signal model is then highly non-linear, and parameter estimation becomes very complex. Stochastic models of signal evolution have been suggested in both [5] and [7]. However both of those approaches also involve very complex estimation processes.

In this paper we present a new method of decomposition that is also based on a stochastic model, but which is much simpler to implement, and also permits more control over the behavior of the decomposition.

2. SIGNAL MODELING AND ESTIMATION

In keeping with common practice, we represent speech as the response of an autoregressive (AR) system, representing the vocal tract filter:

$$z_k = -\sum_{i=1}^{M} a_i z_{k-i} + g.y_k \qquad (1)$$

where

$$y_k = x_k + v_k \, . \qquad (2)$$

g is a gain factor, x_k is a quasiperiodic signal, and v_k is an uncorrelated Gaussian random variable with variance σ_v^2. The response of the vocal tract filter to each of the two components, x_k and v_k, constitute the voiced and unvoiced components of speech respectively.

Fundamental to our approach to decomposing z_k is the assumption that x_k evolves according to

$$x_k = \alpha x_{k-T} + w_{k-T} \tag{3}$$

where T is the period of x_k, w_k is an uncorrelated Gaussian random variable with variance σ_w^2, and α is a gain value.

Based on this model, the decomposition process is as follows. In our implementation, processing is carried out on a frame-by-frame basis with frames of 20ms duration. We begin by demodulating the variation in the energy, spectral envelope, and pitch of the signal. Energy is estimated on a subframe basis (4 subframes/frame). Demodulation of the spectral envelope variation is achieved by applying an inverse filter estimated once per frame by linear prediction. The linear prediction residual is used to estimate the pitch period, and the period is used to time-warp the signal to a fixed period.

The demodulated signal is an approximation, $\hat{y}_k$, of y_k defined on a warped time scale. Equations (2) and (3) together constitute a state space representation of this signal, with w_k representing the process noise and v_k the observation noise. Based on this, a Kalman filter can be used to estimate the state variable x_k, using the standard recursion,

$$\hat{x}_{k|k} = \alpha_k \hat{x}_{k-T|k-T} + K(\hat{y}_k - \alpha_k \hat{x}_{k-T|k-T}) \tag{4}$$

with

$$K = \Sigma_{k|k-T} \Big/ (\Sigma_{k|k-T} + \sigma_v^2)\,, \tag{5}$$

where

$$\Sigma_{k|k-T} = \alpha^2 \Sigma_{k-T|k-T} + \sigma_w^2 \tag{6}$$

is the variance of the error in the predicted state estimate, $\hat{x}_{k|k-T}$, and

$$\Sigma_{k|k} = (1-K)\Sigma_{k|k-T} \tag{7}$$

is the variance in the error in the filtered state estimate, $\hat{x}_{k|k}$. σ_w^2 may be chosen to control the rate at which the estimated quasiperiodic component evolves. However α and σ_v^2 must be estimated from the input data. We describe a new method to do this in the next section.

Since the state variable is different for each sample in the period, estimation of the entire period effectively constitutes a bank of multiple scalar Kalman filters. The smoothing form of the Kalman filter may also be used to take advantage of future pitch cycles to estimate each current sample. The observation noise is estimated as $\hat{v}_k = (\hat{y}_k - \hat{x}_k)$. The estimated quasiperiodic and noisy components can then be remodulated using the estimated period, LPC filter and energy to produce the voiced and unvoiced components of the speech.

The model on which our method is based has some similarity to those in [5]. However the use here of a time domain state representation makes it possible to use only scalar Kalman filter estimators, resulting in significantly lower complexity. In addition, the methods in [5] explicitly assumed that σ_v^2 is known in advance, and made no allowance for an explicit state transition gain α. The latter point is particularly important in decomposing speech, because the overall amplitude of consecutive cycles can change more rapidly than their shape.

The model developed in [7] is almost identical to that described by (2) and (3), but again there was no allowance there for a variable transition gain, and also no provision for explicitly controlling σ_w^2. In addition, because σ_v^2 was not known or determined prior to decomposition, it was not possible to use a Kalman filter for signal estimation. Instead a much more complex algorithm was proposed based on singular value decomposition.

3. ESTIMATION OF DYNAMICAL SYSTEM PARAMETERS

3.1 Estimation Based on Expectation Maximization

Good estimates of α and σ_v^2 are essential in order for the decomposition described above to be effective. An iterative method for determining parameters, θ, of a general linear dynamic system from observations of its output was developed by Digalakis et al [2] based on the Expectation Maximization (EM) algorithm. Each iteration involves maximizing the expected joint log likelihood of the observed data sequence and the unknown state sequence conditioned on the observed data and the previous estimate of θ.

Using this procedure, the estimated values of α and σ_v^2 are:

$$\alpha = \sum_{k=1}^{N} E\{x_k x_{k-T}\} \Big/ \sum_{k=1}^{N} \left[\hat{x}_{k-T|N}^2 + \Sigma_{k-T|N}\right] \tag{8}$$

$$\sigma_v^2 = \frac{1}{N} \sum_{k=1}^{N} \left(y_k^2 - y_k \hat{x}_{k|N}\right) \tag{9}$$

where N represents a fixed interval over which α and σ_v^2 are assumed constant, and $\Sigma_{k-T|N}$ is the covariance of $(\hat{x}_{k-T|N} - x_{k-T})$. The expectation in (8) should be understood to be conditioned on both the observed data up to N and initial estimates of α and σ_v^2.

However, the effectiveness of the recursion defined by (8) and (9) depends significantly on the accuracy of the initial estimates of α and σ_v^2. Inaccurate starting values will lead to slow convergence, and may cause the algorithm to converge to a local optimum. Although no method to obtain initial estimates was suggested in [2], in that case this was not a significant problem since the application of interest there was training of acoustic models for speech recognition. In that situation, estimation occurs off-line. However, for our application, α and σ_v^2 vary throughout the speech waveform, and must be estimated on-line.

We present here a method to obtain these values using only the observed data and *past* values of the estimated state sequence. The method is derived from (8) and (9) above, but relies on an assumption that α and σ_v^2 are approximately constant over intervals of no more that one period. Although, in principal these estimates may be used as initial values for subsequent EM iterations, in our experience they are generally sufficiently accurate themselves, without resorting to further recursion.

To estimate α, we first note that since v_k is uncorrelated with x_{k-T} the expectation in the numerator of (8) can be written as $E\{(y_k - v_k)x_{k-T}\} = y_k \hat{x}_{k-T|N}$. The smoothed state estimate $\hat{x}_{k-T|N}$, which also appears in the denominator, is not known a priori. However, in the mean over the interval from $1 \ldots N$, $\hat{x}_{k-T|N}$ is well approximated by the filtered estimate, $\hat{x}_{k-T|k-T}$. In addition, the error variance, $\Sigma_{k-T|N}$ can be expected to be small compared with $\hat{x}_{k-T|N}^2$. Thus α can be approximated by

$$\alpha \approx \sum_{k=1}^{N} y_k \hat{x}_{k-T|k-T} \Bigg/ \sum_{k=1}^{N} \hat{x}_{k-T|k-T}^2 \qquad (10)$$

Provided the summation interval in (10) is no more than one period, all terms in the right hand side are known.

Using α computed from (10) can be found as follows. Again assuming that $k \le N$, then $\hat{x}_{k|N}$ in (9) is equivalent to $\hat{x}_{k|k}$. Using (4) to compute this value results in

$$\sigma_v^2 = \frac{1}{N} \sum_{k=1}^{N} \left[(y_k^2 - \alpha y_k \hat{x}_{k-T|k-T}) \left(\frac{\sigma_v^2}{\Sigma_{k|k-T} + \sigma_v^2} \right) \right] \qquad (11)$$

$\Sigma_{k|k-T}$ is determined from (6). Equation (11) can be manipulated to produce a quadratic in σ_v^2. Assuming that the signal is not noise-free ($\sigma_v^2 = 0$), the value of σ_v^2 that satisfies this is

$$\sigma_v^2 = \frac{1}{N} \sum_{k=1}^{N} \left[y_k^2 - \alpha y_k \hat{x}_{k-T|k-T} \right] - \Sigma_{k|k-T} \qquad (12)$$

3.2 Modifications for Unvoiced Speech and Voicing Onsets

Despite the theoretical optimality of the procedure described in the preceding section, in practice two problems can arise. First, as speech moves from a voiced to unvoiced segment, the value of α will decrease, and the amplitude of the estimates $\hat{x}_{k|k}$ will drop accordingly.

However during the unvoiced segment, spurious correlations between the harmonic component and $\hat{y}_k$ may cause large values of α to be calculated, resulting in an undesired increase in the amplitude of $\hat{x}_{k|k}$.

To overcome this problem, we have found it useful to test the ratio of the energies of the predicted harmonic component and the noise component. If the signal to noise ratio is less than zero dB and the estimated predictor gain is large, we limit the gain to 1.0.

However at voiced onsets it is important to allow α to be large. To detect onsets we compare $\hat{y}_k$ within the current period with values predicted based on both $\hat{x}_{k-T|k-T}$ and $\hat{y}_{k-T}$, from the previous period. If the residual in the previous period is a better predictor of the current observed data than the previous harmonic component, we assume that an onset has occurred. The predictor gain is calculated based on the residual in the previous period, and the harmonic component is reset to the observed residual data. This procedure is also important in preventing the propagation of decomposition errors due to inaccurate pitch estimates.

4. RESULTS AND DISCUSSION

Figure 1 illustrates the application of our algorithm to a segment of speech consisting of a dominant unvoiced component followed by a dominant voiced component. The smoothing form of the Kalman filter was used with a

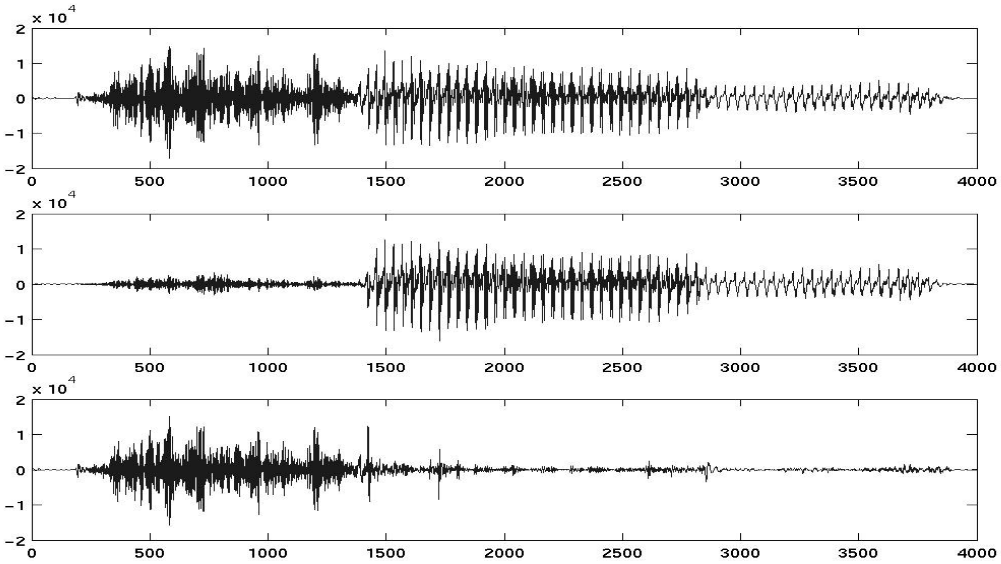

Figure 1: (top to bottom) speech waveform, estimated voiced component, estimated unvoiced component

two period look-ahead. The results show that the algorithm successfully decomposes the speech, with strong attenuation of noisy energy in the voiced component, and no visible harmonic energy in the unvoiced component. The presence of unvoiced signal energy during segments that would generally be classified as voiced is significant. Listening tests indicate that the unvoiced component retains the intelligibility of the original speech, but with a whispered quality.

5. CONCLUSIONS

We have presented a novel method for decomposing speech into voiced and unvoiced components in the time domain. The algorithm is distinctive in its use of a Kalman filterbank, based on dynamical system parameters estimated on-line using a form of the Expectation-Maximisation algorithm.

6. ACKNOWLEDGEMENTS

This work was performed while the authors were all with Motorola. Simon Boland is now with Cross Avaya Research and Development and Michael Smithers is with Dolby Laboratories.

7. REFERENCES

[1] J. P. Campbell Jr & T. R. Tremain, “Voiced/unvoiced classification of speech with applications to the US Government LPC10E Algorithm,” *Proceedings of the International Conference on Acoustic Speech and Signal Processing*, pp 473-476, 1986.

[2] V. Digalakis, J. R. Rohlicek & M. Ostendorf, “ML estimation of a stochastic linear system with the EM algorithm and its application to speech recognition,” *IEEE Transactions on Speech and Audio Processing*, Vol. 1 No. 4, pp. 431-442, 1993.

[3] I. A. Gerson & M. A. Jasiuk, “Techniques for improving the performance of CELP-type speech coders,” *IEEE Journal on Selected Areas in Communications*, Vol. 10, No. 5, pp 858-865, 1992.

[4] D. W. Griffin & J. S. Lim, “Multiband excitation vocoder,” *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol 36, No. 8, pp 1223 –1235, 1988.

[5] P. Gruber & J. Tödtli, “Estimation of quasiperiodic signal parameters by means of dynamic signal models,” *IEEE Transactions on Signal Processing*, Vol. 42, No. 3, pp. 552-562, 1994.

[6] W. B. Kleijn & J. Haagen, “Transformation and decomposition of the speech signal for coding,” *IEEE Signal Processing Letters*, Vol. 1, No. 9, pp 136-138, 1994.

[7] J. Stachurski, “A Pitch Pulse Evolution Model for Linear Predictive Coding of Speech,” Ph.D. Thesis, McGill University, Montreal, Canada, 1997.

[8] Y. Stylianou, “Efficient decomposition of speech signals into a deterministic and a stochastic part,” *Proceedings of the International Symposium on Signal Processing and its Applications*, Vol. 1, pp. 5-8, 1996.