SEGMENTATION OF SPEECH WAVEFORMS ACCORDING TO OPEN AND CLOSED PHASES USING DURATION MODELLING

G.A. Smith Engineering Dept, University of Cambridge, Cambridge. CB2 1PZ.
A.J. Robinson Engineering Dept, University of Cambridge, Cambridge, CB2 1PZ.
Now at SoltSound Ltd, Autonomy Systems Ltd, Cambridge. CB4 0WZ.

ABSTRACT

This paper presents a model-based technique to segment speech according to its glottal phase using the waveform alone, and is achieved by identifying spectral changepoints during the pitch period. Spectral changes include formant modulation and glottal resonances during the open phase. This paper assesses the technique through experiments on two separate speech databases, and determines whether explicit state duration control improves segmentation. The technique differs from conventional glottal inverse filtering and joint source-tract estimation methods.

The analysis model employed is a 3-state autoregressive hidden semi-Markov model (AR-HSMM), with the standard hidden Markov model (HMM) as a special case. States correspond to the glottal closed, open and return phases. Expectation-maximisation is used for parameter estimation, and Viterbi-type algorithms for waveform segmentation into glottal phases. Three methodologies are applied to the segmentation problem, each with a greater degree of state duration control. Duration control is helpful to prevent pitch halving and spurious state cycles, to penalise overly short open phase durations, and to cause more consistent phase durations from one pitch period to the next. Because of the computational load of the HSMM and the lack of duration control of the standard HMM, a compromise methodology is proposed as a compromise.

Experiments on the first database compare estimated glottal phase segmentations with electroglot- tography (EGG) waveforms. Experiments on the second ISOLET database use closed and open phase speech MFCC feature vectors in a 6 vowel phoneme classification task. Classification rates for the closed phase are greater than for the open phase. However it is difficult to determine whether closed phase classification improves with greater state duration control.

1 INTRODUCTION

During voiced speech, air flows from the lungs through the glottis and vocal tract and is radiated to the environment at the lips. The glottis opens and closes at a quasi-periodic rate equal to the fundamental frequency. The main excitation per pitch period occurs at glottal closure, although lesser excitations may occur elsewhere. This paper addresses the problem of determining the time instants of glottal
Proceedings of the Institute of Acoustics

closure and opening from the speech waveform alone using a model-based technique. Section 1 motivates the technique from a physiological perspective, section 2 introduces the analysis model, and section 3 presents several parameter estimation and segmentation algorithms. Section 4 then applies the technique to real speech data. Finally, conclusions are presented. Measurement of glottal behaviour has applications in speaker and voice type identification, medical diagnosis of laryngeal disorders, the study of source-tract interactions and speech coding.

Voiced speech production can be modelled as three subsystems in series: glottal source (a volume velocity waveform), vocal tract and lip-radiation (Fant [9]). For simplicity these subsystems are often assumed linear and separable. The glottal source, vocal tract and lip-radiation can be modelled as a pulse generator, all-pole (autoregressive) filter and differencer respectively. Because of the linearity of the model, these three subsystems are often reduced to two subsystems: the vocal tract excited by the glottal source derivative (Plumpe et al. [15]), which is the approach adopted in this paper. There are two common general techniques to determine this two subsystems model from the speech waveform directly: glottal inverse filtering and joint source-tract estimation.

In glottal inverse filtering, the vocal tract transfer function is estimated during the glottal closed phase, and its inverse is then applied to the speech waveform to obtain a residue which approximates to the glottal source derivative. There are problems with this technique. Firstly, the closed phase may be difficult to estimate accurately. Secondly, closed phase durations may be short or non-existent for breathy or female speech causing non-robust vocal tract estimates. Thirdly, phase distortion introduced by the microphone or tape recorder may cause the inverse filter residue to be a poor match with expected glottal waveform shapes.

In joint source-tract estimation, the glottal source is represented in parametric form, and the source and vocal tract parameters are estimated jointly (Fujsaki and Ljungqvist [10], Milenkovic [14]). There are several problems associated with this. Firstly, joint source-tract estimation is an underdetermined problem, often requiring non-linear iterative optimisation procedures operating on multimodal objective functions. Secondly, phase distortion during recording may cause a poor match between the inverse filter residue and parametric glottal waveform. Thirdly, source-tract interaction introduces non-linear coupling effects such as ripples on the inverse filter residue, which are not catered for by most glottal waveform models.

This paper proposes an alternative technique to determine glottal timing information which, like glottal inverse filtering, does not require a parametric representation of the glottal source. This alternative technique works on the principle that spectral changepoints in the time waveform occur at glottal closure and glottal opening, and that these can be identified using an autoregressive hidden semi-Markov model (AR-HSMM); different states correspond to regions of different spectrum over a given pitch period. This paper extends previous research [16] in that explicit duration modelling is investigated, more comprehensive experiments on real speech are conducted, and the speech is bandpass filtered prior to analysis. Two spectral changes during the pitch period highlighted here are formant modulation and glottal resonances, which are discussed by Ananthapadmanabha and Fant [4], Plumpe et al. [15] and Stevens [17]. Relevant details are presented below.
Amplitude and frequency modulation of formants occurs due to non-stationary and non-linear coupling between the source and vocal tract. Modulation is greatest for the first formant. At glottal opening, both formant frequency and bandwidth tend to increase. Ananthapadmanabha and Fant [4] show that changes in formant bandwidth follow changes in glottal area, whereas changes in formant frequency follow changes in the glottal area derivative, meaning changes in formant frequency are more instantaneous. In the time-domain, this formant modulation is manifest as ripples on the inverse filter residue at a ripple frequency in the region of the first formant, and a truncation of the speech waveform at glottal opening due to the increase in bandwidth or damping. Therefore the open phase is characterised by non-stationary formants and ripples in the inverse filter waveform, whereas the closed phase is identified by relatively (although not completely) stationary formants, and this fact can be used to identify glottal opening instants (Plumpe et al. [15]). Although coupling effects may extend back into the closed phase region when complete glottal closure does not occur, effects remain relatively small compared to glottal opening.

The acoustic impedance looking into the trachea from just below the glottis is characterised by a series of poles and zeros. The lowest three poles (sub-glottal resonances) are typically 600, 1550 and 2200 Hz for adult males with bandwidths in the range 200 to 400 Hz, and are essentially speaker-dependent and fixed. Only when the glottis is open does the sub-glottal system introduce pole-zero pairs into the transfer function between the glottis and the lips, thus modifying the speech spectrum relative to its closed phase representation (Stevens [17], sections 3.6.4, 6.8). The perturbation in the speech spectrum depends on the size of glottal opening, the supraglottal configuration and the spectral frequency. Perturbation is common near the second subglottal resonant frequency and is often manifest as an extra spectral peak in the speech spectrum between 1400 and 1800 Hz. Sometimes there is prominence in the vicinity of the third subglottal resonance. The effect is less obvious at the first subglottal resonant frequency because acoustic losses on glottal opening at such low frequencies reduce the degree of prominence.

The closed glottal phase is marked by relatively stationary formants, whereas the open phase is often characterised by perturbations in the spectrum in the vicinity of the sub-glottal resonances and formant modulation (usually frequency and bandwidth increases). The closed and open phase spectra thus differ. This paper presents a technique which models the closed and open phases using separate all-pole models, and estimates glottal closure and opening as the switching instants between these models. A three-state autoregressive hidden semi-Markov model is used for analysis, because this allows explicit state duration modelling. The glottal return, closed and open phases correspond to states 1, 2 and 3 respectively.

2 THE MODEL

The general model is a three-state autoregressive hidden semi-Markov model (AR-HSMM). The HSMM model enables explicit control over state durations via non-parametric or parametric probability distributions, and is therefore also known as the explicit state duration hidden Markov model. A special case of the HSMM when state durations are exponential is the standard HMM. HSMMs have been applied to speech recognition, but often in the context of feature vectors rather than the time
waveform (Hochberg [12], Levinson, [13] and Vaseghi [18]). A 3-state model is illustrated in figure 1.

![Diagram of 3-state model](image)

Figure 1: The 3-state model used for glottal phase modelling

The speech sample or observation at time $t$ is $y_t \in \mathbb{R}$, and the observations sequence is $y_1:T$. The initial conditions for the speech waveform are $y_0 = y_{-p^*+1:0} \in \mathbb{R}^{p^*}$, where $p^*$ is of appropriate dimension to initialise all state AR polynomials. The AR polynomial, excitation variance and autoregressive model for state $i$ are respectively $a_i \in \mathbb{R}^{p_i}$, $\sigma_i^2(t) \in \mathbb{R}$, and $y_t = \sum_{i=1}^{p_i} a_i(j)y_{t-j}$. The probability of transition from state $i$ to state $j$ is $\pi_{ij} \in \mathbb{R}$. The probability of duration $\tau$ in state $i$ is a Gamma distribution $\pi_i(\tau) = \text{Gam}(\alpha_i, \beta_i)$. The state occupied at time $t$ is $r_t \in \mathbb{R}$. The probability of being in state $i$ at time $t = 1$ is $r_1(i)$, where $r_1 \in \mathbb{R}^3$. The 3-state AR-HSMM is defined fully in terms of the following parameters:

- initial conditions: $y_0 = y_{-p^*+1:0}$
- AR polynomials: $A = \{a_1, a_2, a_3\}$
- AR excitation variances: $\Sigma = \{\sigma_1^2(1), \sigma_1^2(2), \sigma_3^2(3)\}$
- transition matrix: $\Pi = \{\pi_{1,1}, \pi_{1,2}, \pi_{1,3}, \pi_{2,1}, \pi_{2,2}, \pi_{2,3}, \pi_{3,1}, \pi_{3,2}, \pi_{3,3}\}$
- state durations: $d = \{\alpha_1, \alpha_2, \alpha_3, \beta_1, \beta_2, \beta_3\}$
- initial state probability: $r_1 = \{r_1(1), r_1(2), r_1(3)\}$

In this paper, $y_0 = 0$ and $r_1 = (1, 0, 0)$ and these are kept fixed during model estimation for simplicity. The duration parameters $d$ are set a priori as a function of the pitch period and are also kept fixed. All other parameters are estimated given the speech waveform. Estimation and segmentation strategies are presented in the next section.

### 3 SEGMENTATION METHODOLOGIES

This paper compares three different methodologies to segment speech waveforms according to glottal phase, to determine whether explicit state duration control improves segmentation. Each methodology consists of two stages. Firstly, parameters are estimated. Secondly, the speech waveform is
segmented according to glottal phase using a Viterbi-type algorithm. The three methodologies differ regarding the degree of duration control.

1. The **standard** methodology uses the standard HMM model and standard Viterbi algorithm (no explicit control over state durations).

2. The **BSD** methodology employs the standard HMM model and the bounded state duration (BSD) Viterbi algorithm (explicit control over state durations during segmentation only).

3. The **explicit** approach uses a HSMM model and HSMM Viterbi algorithm (explicit control over state durations during both parameter estimation and segmentation).

All three approaches use the expectation-maximisation (EM) algorithm for parameter estimation. EM is a general-purpose iterative algorithm to reestimate model parameters given observations, a model and initial parameter estimates, such that the likelihood is guaranteed to increase during each iteration (Dempster et al. [8]). EM implements a local search of the likelihood surface only so its success depends on initial parameter estimates and the shape of the likelihood surface. The Baum-Welch algorithm is the same as EM except with a modified objective function. In this section the Baum-Welch algorithm is applied to a general N-state AR-HSMM.

### 3.1 The Standard Methodology

The Baum-Welch algorithm is applied to the estimation of the state transition matrix, AR polynomials and excitation variances $\Theta = \{\Pi, A, \Sigma\}$ for a standard HMM. The initial conditions $y_0$ and initial state distribution $r_0$ are kept fixed throughout and are not reestimated. The implicit state duration for a HMM is an exponential. The algorithm consists of forward and backward procedures, which involve the estimation of $\alpha_t(i) = p(y_{1:t}, r_t = i|\Theta)$ and $\beta_t(i) = p(y_{t+1:T}|r_t = i, \Theta)$ respectively. Both procedures require on the order of $NT^2$ computations. Backward procedure initialisation arbitrarily sets $\beta_0(i)$ to unity for all states. $b_t(y_t) = p(y_t|r_t = i, y_{1:t-1}) = p(c_t(i))$ where $c_t(i)$ is the residue at time $t$ from inverse filtering the speech with the state $i$ AR polynomial.

\[
\alpha_{t+1}(i) = \begin{cases} r_t(i) b_t(y_t) & t = 0, \quad 1 \leq i \leq N \\ \left[ \sum_{j=1}^{N} \alpha_t(j) \pi_{j,i} \right] b_t(y_{t+1}) & 1 \leq t < T, \quad 1 \leq i \leq N \end{cases} \tag{1}
\]

\[
\beta_t(i) = \begin{cases} 1 & t = T, \quad 1 \leq i \leq N \\ \sum_{j=1}^{N} \pi_{i,j} \beta_{t+1}(j) & 1 \leq t < T, \quad 1 \leq i \leq N \end{cases} \tag{2}
\]

The $\gamma$ and $\xi$ variables are next computed, where $\gamma_t(i) = p(r_t = i|y_{1:T}, \Theta)$ is the state occupation probability and $\xi_t(i,j) = p(r_t = i, r_{t+1} = j|y_{1:T}, \Theta)$.
Consider the diagonal matrix $\mathbf{\Gamma}_i = \text{diag}(\gamma_i(i)) \in \mathbb{R}^{T \times T}$, the Toeplitz matrix $\mathbf{Y} \in \mathbb{R}^{T \times p}$ with first column $[y_0, \ldots, y_{T-1}]'$ and first row $[y_T, \ldots, y_{1-p}]$, and the column vector $\mathbf{y} = [y_1 \ldots y_T]'$. With these definitions, parameter estimates are

$$\gamma_i(i) = \frac{\alpha_i(i)\beta_i(i)}{\sum_{i=1}^{N} \alpha_i(i)\beta_i(i)} \quad 1 \leq t \leq T \quad 1 \leq j \leq N$$

$$\xi_{t}(i,j) = \frac{\alpha_i(i)\pi_{i,j}b_j(y_{t+1})\beta_{t+1}(j)}{\sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i(i)\pi_{i,j}b_j(y_{t+1})\beta_{t+1}(j)} \quad 1 \leq i \leq N \quad 1 \leq j \leq N$$

Once parameters are estimated, the standard Viterbi algorithm is applied to the speech to determine the minimum cost (maximum likelihood) path. It relies on the following recursion where $\phi_i(1)$ is the cost for state $i$ at time $t$.

$$\phi_i(t) = \max_{j=1}^{N} \left[ \phi_{t-1}(j) + \log \pi_{j,i} \right] + \log b_i(y_t)$$

### 3.2 The BSD Methodology

Model parameters are estimated as for the standard approach above, but the bounded state duration (BSD) rather than the standard Viterbi algorithm is used for segmentation (Gu et al. [11]). This algorithm enforces minimum and maximum bounds on state duration $\tau_i(i)$ and $\tau_{j(i)}$ respectively for the $i$th state. The implicit state duration density is therefore exponential $d_i(\tau) = \pi_{i,i}^{-1}(1 - \pi_{i,i})$, but with lower and upper limits on duration. The BSD algorithm is based on two recursions. It requires the computation of the minimum score path through a three-dimensional time-state-duration trellis. $\phi_i(i, \tau)$ is the score for state $i$ at time $t$ with duration $\tau$ in that state.

$$\phi_i(i, \tau) = \phi_{i-1}(i, \tau - 1) + \log b_i(y_t) \quad \tau \geq 2$$
\[ \phi_t(i, 1) = \max_{j=1}^{N} \left[ \max_{\tau_{T}(j)} \left[ \phi_{t-1}(j, \tau) + \log d_j(\tau) \right] + \log \pi_{j,i} + \log b_j(y_t) \right] \] (10)

### 3.3 The Explicit Methodology

Because the HSMM state evolves according to a semi-Markov process rather than the Markov process of the standard HMM, different equations are required for the estimation of the \( \{\alpha, \beta\} \) variables. During these experiments, state durations are modelled as gamma distributions which are kept fixed at some fraction of the pitch period and are not reestimated.

\[ \alpha_*^t(i) = \begin{cases} r_1(i) & t = 0 \quad 1 \leq i \leq N \\ \sum_{j=1}^{N} \alpha_t(j) \pi_{j,i} & 1 \leq t < T \quad 1 \leq i \leq N \\ \end{cases} \] (11)

\[ \beta_*^t(i) = \begin{cases} \sum_{j=1}^{N} \beta_{T,j}(j) \pi_{i,j} & t = T \quad 1 \leq i \leq N \\ \sum_{j=1}^{N} \beta_{T,j}(j) \pi_{i,j} & 1 \leq t < T \quad 1 \leq i \leq N \\ \end{cases} \] (12)

\[ \alpha_t(i) = \sum_{r=1}^{T-t} \alpha_{t-r}(i) d_r(i) \prod_{r=0}^{T-t} b_r(y_{t-r}) \quad 1 \leq t \leq T \quad 1 \leq i \leq N \] (13)

\[ \beta_t^r(i) = \sum_{T-t}^{T-r} \beta_{t-r}(i) d_r(i) \prod_{r=1}^{T-r} b_r(y_{t-r}) \quad 0 \leq t < T \quad 1 \leq i \leq N \] (14)

\[ \pi_{i,j} = \frac{\sum_{t=1}^{T-1} \beta_{t,r}(j) \pi_{i,j} \alpha_t(i)}{\sum_{t=1}^{N} \sum_{t=1}^{T-1} \beta_{t,r}(j) \pi_{i,j} \alpha_t(i)} \quad 1 \leq j \leq N \quad 1 \leq i \leq N \] (15)

Here we set \( r_T = \{0, 0, 1\} \). The HSMM requires considerably greater computation and memory requirements as evident from the summation over \( r \) and this is the main disadvantage of such a model. The estimation of all other variables is the same as for the standard HMM. Once parameters are estimated, then the optimal state sequence is determined using a HSMM Viterbi algorithm (Ljolje and Levinson [3]).

\[ \phi_t(i) = \max_{\tau=1}^{T} \left[ \max_{j=1}^{N} \left[ \phi_{t-1}(j) + \log \pi_{j,i} + \log d_r(i) + \sum_{r=0}^{T-t} \log b_r(y_{t-r}) \right] \right] \] (16)

### 4 EXPERIMENTAL DETAILS

Two sets of experiments are conducted. In the first set, a small database containing simultaneous speech and EGG recordings is used to compare estimated and EGG-derived glottal segmentations.
for the three different methodologies. In the second set, a larger database containing speech only is
analysed, closed and open phase spectral information measured according to the three methodolo-
gies, and then used in a vowel phoneme classification task. In both sets of experiments, speech is
preprocessed as follows.

Firstly, for each waveform the ESPS toolkit [1] is used to determine fundamental frequency and
epoch locations. Fundamental frequency is determined using a normalised cross-correlation and
dynamic programming method with minimum and maximum bounds on frequency of 80 and 160
Hz respectively, and epochs are measured by peak-picking the inverse filter residue and dynamic
programming. To minimise end effects, epochs are first located for the entire phoneme waveform
before a central quasi-stationary section is isolated. From these epochs an initial state sequence
segmentation is derived; open-return, return-closed and closed-open phase changepoints are set
to \{0, 0.05, 0.4\} of a pitch period offset from the epoch. From this initial segmentation, all other
parameters are initialised.

Secondly, each speech waveform is filtered twice with an order 30 bandpass finite impulse response
(FIR) filter, first in the time-forward direction and second in the time-reverse direction, to preserve
timing instants. The filter has bandpass cut-off frequencies at 1kHz and 2kHz. This emphasises the
spectral region where the second glottal resonance is expected. This was found to produce better
results. The filtered waveform is then preemphasised to reduce spectral roll-off, and windowed etc.

Segmentation of speech into return, closed and open states operates on this bandpass filtered pre-
emphasised speech waveform, using AR model orders \{6, 8, 8\} for return, closed and open states
respectively. Once a segmentation is determined, AR models for return, closed and open are recom-
puted using non-bandpass-filtered preemphasised speech, the derived segmentation, and model
orders \{8,16,16\}. This second set of AR models is used to compute feature vectors used in the
phoneme classification task.

Minimum and maximum state durations for the BSD Viterbi decoder as a fraction of the pitch period
are set respectively to \{0.05, 0.2, 0.3\} and \{0.3, 0.8, 0.8\} of a pitch period, where the pitch period is
estimated using the ESPS toolkit as above. Duration probability distributions for the explicit methodo-
logy are set to Gamma distributions with means given by the initial segmentation durations and
variances set to 50 % of the ESPS-estimated pitch period. These duration probability distributions
are not reestimated during the experiments.

4.1 Experiment Set 1: Comparison with EGG Data

The aim is to test the validity of the three methodologies by comparing their segmentations with EGG-
derived glottal segmentations. EGG (electroglottography) waveforms (Baken [5]) record the variation
in electrical impedance across the neck during phonation and measure the degree of contact between
the vocal folds. It is generally accepted that the positive and negative peaks in the derivative of the
impedance waveform correspond to glottal opening and closure instants respectively, and vice versa
for the conductance waveform.
Proceedings of the Institute of Acoustics

The normalLm folder of a CD-ROM database due to Childers [6] is used. This contains simultaneous speech and EGG recordings, each sampled at 10kHz. Recordings for 25 males with normal larynges speaking 12 sustained vowels are analysed, giving 300 utterance in total. These vowels are /IY, IH, EY, EH, AE, UW, UH, OW, AO, AA, AH, ER/ using standard ARPAbet notation. For each utterance, a segment between samples 7000 and 8023 is extracted, assumed quasi-stationary, and then analysed using the three different methodologies. Final segmentations are compared with true segmentations measured directly from the EGG waveforms.

The following are inferred from qualitative assessment of experimental results. The standard methodology is susceptible to pitch doubling, allows short duration spurious state cycles in the segmentation, and tends to overestimate closed phase durations (perhaps due to difficulties in identifying the closed-open phase changepoint). The BSD methodology is robust against pitch doubling (provided lower and upper duration limits are chosen carefully), but similarly shows inconsistency and overestimation of closed phase durations. The explicit methodology produces more consistent segmentations across pitch periods, and tends to penalise overly short closed or open phase durations. Though rarer than for the standard methodology, the explicit approach does experience some pitch doubling, because the Gamma distributions used to model durations assign non-zero probability to short durations.

Figures 2 to 4 are example segmentations compared with their corresponding EGG derivative impedance waveforms (offset by 7 samples to account for acoustic delay). States 1, 2 and 3 correspond to return, closed and open phases respectively. Figure 2 shows close agreement between all three methods and the EGG waveform. Figure 3 shows pitch doubling for the standard approach, and inconsistent phase durations from one pitch period to the next for the BSD methodology. Finally figure 4 demonstrates when the algorithms give different results, and where the explicit approach favours shorter durations for the closed phase.

Figure 5 shows formant tracks and the perturbation in these formants for some bandpass filtered preemphasised speech. It is based on an order 8 sliding window covariance analysis with window size 24, shifted one sample at a time, where only poles with bandwidths less than 500 Hz are shown. Finally, the histograms in figure 6 show duty cycles. For the three methodologies, our duty cycle is the duration of return and closed phases combined as a fraction of the EGG-derived pitch period. For the EGG, it is the duration between glottal closure and opening as a fraction of this pitch period. The explicit methodology gives a distribution similar to the EGG but shifted to the left, whereas the standard and BSD methodologies shifts and skews the distribution to the right, which means they tend to overestimate the closed phase duration. Histograms are also affected by pitch-doubling effects. The standard deviation difference between estimated and EGG-derived duty cycles are \{0.22, 0.17, 0.12\} for standard, BSD and explicit respectively.

4.2 Experiment Set 2: Vowel Classification

The aim is to segment speech phonemes according to glottal phase using the three methodologies, and then use feature vectors derived from the closed and open phases in a vowel phoneme classification task. The hypothesis is whether increased state duration control yields better closed phase
Proceedings of the Institute of Acoustics

Figure 2: Agreement between all three methodologies and the EGG data (utterance m0606)

Figure 3: Pitch-doubling for the standard methodology (utterance m0606)
Figure 4: Closed phase duration estimate differs for all three (utterance m0507)

Figure 5: Formant tracks during a sliding covariance analysis, vertical dotted lines mark the epochs (utterance m0606)
models, which should be evident as greater classification rates. Real speech is obtained from ISO-LET [7], which is a database of letters of the American English alphabet spoken in isolation. The database consists of 5 subsets termed ISOLET-1 to ISOLET-5. Each subset contains each letter spoken twice by each of 15 male and 15 female speakers.

Firstly, each ISOLET utterance is segmented into constituent phonemes. Because phoneme start and end times are not given in the database, they are estimated as part of the experiment. This is achieved by modelling each phoneme as a three-state left-to-right HMM, and then using these models with a Viterbi decoder to segment each utterance into its constituent phonemes. HMMs are trained using ISOLET-1 through ISOLET-4 subsets so as to approximately minimise the classification error for ISOLET-5 phonemes based on a single state 10-component mixture Gaussian classifier. The HTK toolkit is used throughout [2].

Once each ISOLET utterance is represented using its phoneme transcription, all phonemes /Y, EY, EH, UW, OW, AA/ from ISOLET-1, ISOLET-2 and ISOLET-5 subsets exceeding 2024 samples in duration are extracted, and the central 1024 samples for each retained. Each 1024 segment is then analysed using the three different methodologies, segmentations derived, and then order 16 LPC coefficients for closed and open glottal phases computed. These LPC coefficients are then converted into MFCCs via a 20 filterbank analysis, and the first 12 MFCCs used to construct a single-state diagonal covariance Gaussian phoneme classifier. The classifier is trained using phonemes from ISOLET-1 and ISOLET-2, and tested using ISOLET-5. The numbers of training utterances for phonemes /Y, EY, EH, UW, OW, AA/ are {458, 225, 248, 107, 60, 59} respectively, and the numbers of test utterances are {224, 107, 125, 54, 29, 26} respectively.
The vowel classification rates for closed and open phase feature vectors for the three methodologies are presented in table 1. The “benchmark” is where closed and open phase LPC coefficients are derived from the initialisation state segmentation. The “both” column denotes classification where a single AR model is used across both the closed and open phases (but not the return phase).

<table>
<thead>
<tr>
<th>approach</th>
<th>closed phase</th>
<th>open phase</th>
<th>both</th>
</tr>
</thead>
<tbody>
<tr>
<td>benchmark</td>
<td>87.61</td>
<td>82.65</td>
<td>91.50</td>
</tr>
<tr>
<td>standard</td>
<td>89.91</td>
<td>72.39</td>
<td>91.50</td>
</tr>
<tr>
<td>BSD</td>
<td>89.03</td>
<td>79.12</td>
<td>91.50</td>
</tr>
<tr>
<td>explicit</td>
<td>89.91</td>
<td>64.78</td>
<td>91.50</td>
</tr>
</tbody>
</table>

Table 1: Vowel phoneme classification rates for benchmark and three methodologies

These results show that classification rates for the closed glottal phase are always greater than for the open glottal phase. Possible reasons include the following three. Firstly, formants during the closed phase are more stationary than during the open phase, giving more consistent closed phase spectral estimates. Secondly, the closed phase relates to the vocal tract only, whereas the open phase relates to the vocal tract and glottal source which is speaker dependent. Thirdly, results from experiment set 1 suggest that the algorithm sometimes underestimates open phase durations, which may lead to non-robust open phase spectral estimates due to the smaller amount of data.

Classification rates for the closed phase for all three methodologies are competitive and greater than the benchmark, which suggests that the three methodologies do tend to move state boundaries from their initial positions to a better segmentation. But it is difficult to say whether closed phase classification rates are increased by greater state duration control. However, open phase classification rates are increased with duration control to above the benchmark. Classification rates using a single model for both closed and open phase are greater, due probably to the fact that more data is available for spectral estimates giving greater robustness and consistency within a phoneme class.

In conclusion, because of the computational load of the HSMM relative to the standard HMM, the BSD methodology is proposed as the optimal methodology. But modifications may be required to penalise overly short open phase durations. Finally, we note that the algorithm is sensitive to the initial segmentation, particularly the location of the epochs.

5 Conclusions

Glottal closure and opening instants can be identified by changes in spectrum across a pitch period, such as formant modulation and glottal resonances during the open phase. These spectral change-points can be identified using a three-state autoregressive hidden semi-Markov model, of which the
hidden Markov model is a special case. Duration control is helpful to prevent pitch halving and spurious state cycles, to penalise overly short open phase durations and to cause more consistent phase durations from one pitch period to the next. Experiments on a vowel phoneme classification task show that closed phase feature vectors give greater classification rates than open phase feature vectors. However it is difficult to determine whether closed phase classification improves with greater state duration control. Because of the computational load of the HSMM relative to the standard HMM, the BSD methodology is proposed as the optimal methodology, but modifications may be required to penalise overly short open phase durations.

Acknowledgements. Gavin Smith is grateful for funding from the Schiff Foundation, Cambridge University, and would like to thank Professor Niranjan (Sheffield University), and his current supervisor Dr. Simon Godsill (Cambridge University).

References

Proceedings of the Institute of Acoustics


