SEGMENTATION OF A SPEECH WAVEFORM ACCORDING TO GLOTTAL OPEN AND CLOSED PHASES USING AN AUTOREGRESSIVE-HMM

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

Gavin Smith, Tony Robinson

Cambridge University Engineering Department,
Trumpington Street, Cambridge, CB2 1PZ, United Kingdom.
{gas1003, ajr}@eng.cam.ac.uk

ABSTRACT

This paper presents an algorithm to segment speech according to glottal open and closed phases using the time waveform alone. Based on this, pitch, jitter and closed to open glottal ratios can be computed. Segmentation is achieved by identifying spectral changepoints at the sub-pitch period timescale. Changepoints are identified using a 3-state autoregressive hidden Markov model (AR-HMM) operating on the time waveform, with the Liljencrants-Fant (LF) glottal model as a theoretical basis. Model parameters and optimal state sequence are determined respectively using the expectation-maximisation (EM) algorithm and a bounded state duration (BSD) Viterbi algorithm. Experiments on synthetic speech give encouraging glottal segmentation for modal, fry and breathy voice types. Experiments on real speech obtained from TIMIT give meaningful segmentations also.

1. Introduction

During speech, air flows from the lungs, through the glottis and vocal tract, and is then radiated to the environment at the lips [1]. The speech production system can be simplified to three linear, decoupled subsystems in series: glottal source, vocal tract and lip-radiation. In this paper, these three subsystems are represented as an integrated Liljencrants-Fant pulse generator, switched autoregressive filters and a differencer respectively.

This paper presents an algorithm to segment the speech according to the glottal timing information using the speech waveform alone. From this segmentation, pitch period, pitch jitter and glottal open to closed ratios can be computed. The segmentation is determined by identifying spectral changepoints at the sub-pitch period level, which occur for example at glottal closure and glottal opening. These changepoints are identified using a 3-state hidden Markov model (HMM). The use of spectral changepoints and the HMM for glottal segmentation are the novel aspects in this paper.

The accurate estimation of glottal timing information has applications. Firstly, glottal information is characteristic of a speaker and is useful to assist in speaker identification or verification. Secondly, abnormal timing information is indicative of laryngeal disorders and is useful in medical diagnosis. Thirdly, the identification of open and closed phases is useful for glottal-excited linear predictive (GELP) coding, for studying source-tract interactions and for speech recognition.

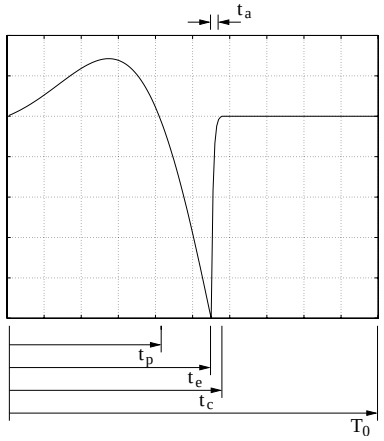

Figure 1: LF waveform for one pitch period for modal speech.

2. The Liljencrants-Fant (LF) Model

The Liljencrants-Fant (LF) model is a piecewise trigonometric and exponential model which represents the glottal waveform [2]. It is illustrated in Figure 1 for modal speech and is defined during a single pitch period as

$$g(t) = \begin{cases} E_0 e^{\alpha t} \sin(\omega_g t) & 0 \le t < t_e \\ -\dfrac{E_e}{\epsilon t_a}\{e^{-\epsilon(t-t_e)} - e^{-\epsilon(t_c-t_e)}\} & t_e \le t < t_c \\ 0 & t_c \le t < T_0 \end{cases} \quad (1)$$

where $\omega_g = \pi/t_p$ and T_0 is the pitch period. The following conditions for these equations hold:

$$\int_0^T g(t)dt = 0, \quad \epsilon t_a = 1 - e^{-\epsilon(t_c - t_e)}, \quad E_0 = \frac{-E_e}{e^{\alpha t_e}\sin(w_g t_e)}, \quad (2)$$

For convenience, these three piecewise segments are termed *open+*, *open-* and *closed* phases respectively and have distinct spectra. Changepoints between these three phases are evident as spectral changepoints in the glottal waveform. Because of the linear nature of the vocal tract and lip-radiation subsystems, these spectral changepoints are evident in the speech waveform also.

3. The Autoregressive Hidden Markov Model (AR-HMM)

In the linear prediction approach to speech synthesis and analysis, a single autoregressive model is used to represent the glottal source, vocal tract and lip-radiation subsystems combined. This has proven successful in speech coding applications. A progression from this is to adopt three autoregressive models, one for each glottal phase, and then control the interaction between these three models via a hidden Markov model. We therefore motivate the use of a cyclical three-state autoregressive hidden Markov model [3] for speech analysis represented in Figure 2. It is applied to voiced, steady-state speech sounds.

The observed signal $y_{1:T} = [y_1, y_2, \ldots, y_T]$ is T samples long. The state occupied at time t is r_t. The probability of being in state i at time $t = 1$ is $\mathbf{r}_1(i)$. The initial conditions for the speech waveform are $\mathbf{y}_0 = y_{-p+1:0}$. The ith state AR polynomial, excitation variance and order are termed $\mathbf{a}_i$, $\sigma_e^2(i)$ and p_i respectively. The autoregressive model for state i is defined as $\hat{y}_t^i = \sum_{j=1}^{p_i} \mathbf{a}_i(j) y_{t-j}$. The transition probability from state i to state j is $\pi_{i,j}$. The 3-state HMM is defined fully in terms of the following parameters:

$$\begin{array}{rclcrcl} \mathbf{y}_0 & = & y_{-p+1:0} & & \mathbf{r}_1 & = & \{\mathbf{r}_1(1), \mathbf{r}_1(2), \mathbf{r}_1(3)\} \\ \mathbf{A} & = & \{\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3\} & & \Pi & = & \{\pi_{1,1}, \ldots, \pi_{3,3}\} \\ \Sigma & = & \{\sigma_e^2(1), \sigma_e^2(2), \sigma_e^2(3)\} & & & & \end{array}$$

There are two problems: firstly, the estimation of model parameters given the observations, and secondly, the estimation of the optimal state sequence given the observations and parameter estimates, from which glottal timing information is obtained. The expectation-maximisation algorithm is applied to the first problem, and the bounded state duration Viterbi algorithm to the second. For convenience, $\Theta = \{\mathbf{y}_0, \mathbf{r}_1, \Pi, \mathbf{A}, \Sigma\}$.

4. The Expectation-Maximisation (EM) Algorithm

The expectation-maximisation (EM) algorithm is a general-purpose iterative algorithm to re-estimate model parameters given observations, a model and initial parameter estimates, such that the likelihood of the observations increases. The algorithm consists of two stages. In the first stage, the forward-backward algorithm, otherwise known as the Baum-Welch algorithm, is used. Standard computational procedures are employed [3] to compute the following variables:

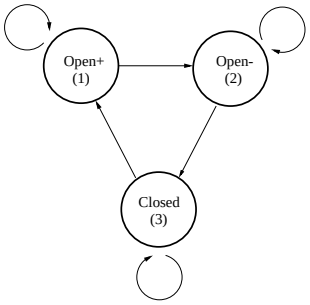

Figure 2: 3-state AR-HMM used for changepoint identification.

$$\begin{aligned} \alpha_t(i) &= p(y_{1:t}, r_t = i|\Theta) && (3) \\ \beta_t(i) &= p(y_{t+1:T}|r_t = i, \Theta) && (4) \\ \gamma_t(i) &= p(r_t = i|y_{1:T}, \Theta) && (5) \\ \xi_t(i,j) &= p(r_t = i, r_{t+1} = j|y_{1:T}, \Theta) && (6) \end{aligned}$$

In the second stage, parameters are re-estimated.

Initial state $1 \le i \le 3$	$\mathbf{r}_1(i) = \gamma_1(i)$
Trans matrix $1 \le i \le 3,\ 1 \le j \le 3$	$\pi_{i,j} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)}$
Excitation $1 \le i \le 3$	$\sigma_e(i)^2 = \frac{\sum_{t=1}^{T} \gamma_t(i)(\hat{y}_t^i - y_t)^2}{\sum_{t=1}^{T} \gamma_t(i)}$
Polynomial $1 \le i \le 3$	$\mathbf{a}_i = (\mathbf{C}_i\mathbf{Y})^\dagger \mathbf{C}_i\ \mathbf{y}_{1:T}$
Initial conditions	$\mathbf{y}_0 = \left\{\sum_{i=1}^{3} (\mathbf{C}_i\bar{\mathbf{A}}_i)'\mathbf{S}_i(\mathbf{C}_i\bar{\mathbf{A}}_i)\right\}^{-1} \left\{\sum_{i=1}^{3} (\mathbf{C}_i\bar{\mathbf{A}}_i)'\mathbf{S}_i (\mathbf{C}_i\mathbf{y}_{1:T} - \mathbf{C}_i\mathbf{A}_i\mathbf{y}_{1:T})\right\}$

$\mathbf{S}_i = \mathrm{diag}(\frac{1}{\sigma_e^2(i)})$ and † denotes the Moore-Penrose inverse such that $\mathbf{X}^\dagger = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. Relevant Toeplitz matrices and vectors are defined as follows

$$\mathbf{C}_i = \begin{bmatrix} \sqrt{\gamma_1(i)} & & & \\ & \sqrt{\gamma_2(i)} & & \\ & & \ddots & \\ & & & \sqrt{\gamma_T(i)} \end{bmatrix}$$

$$\mathbf{Y} = \begin{bmatrix} y_0 & y_{-1} & \cdots & y_{1-p} \\ y_1 & y_0 & \cdots & y_{2-p} \\ \vdots & \vdots & \vdots & \vdots \\ y_{T-1} & y_{T-2} & \cdots & y_{T-p} \end{bmatrix}$$

vowel	example word	modal			breathy			fry		
		open+	*open-*	*closed*	*open+*	*open-*	*closed*	*open+*	*open-*	*closed*
a	B<u>o</u>b	54(0.5)	4(0.3)	42(0.7)	65(0.3)	3(0.5)	32(0.6)	56(0.9)	4(0.6)	40(0.9)
ae	b<u>a</u>t	54(0.2)	4(0.6)	42(0.6)	65(0.3)	3(0.5)	32(0.7)	56(1.1)	4(1.2)	40(1.8)
e	b<u>e</u>t	54(0.2)	4(0.8)	42(0.9)	65(1.7)	5(2.4)	30(3.1)	58(1.1)	3(0.3)	39(1.1)
er	B<u>er</u>t	53(0.5)	4(0.3)	43(0.7)	61(6.3)	4(1.1)	35(5.8)	53(1.9)	3(0.7)	44(2.0)
i	b<u>i</u>t	54(0.3)	4(0.6)	42(0.7)	63(6.8)	4(2.0)	31(3.1)	56(0.7)	3(0.2)	41(0.8)
iy	b<u>ee</u>t	54(0.6)	4(0.8)	42(0.6)	64(0.5)	4(0.5)	32(1.0)	57(0.7)	4(0.6)	40(1.1)
oo	b<u>oo</u>t	53(0.5)	4(0.3)	43(0.6)	62(0.4)	4(0.6)	34(0.9)	58(2.6)	6(2.9)	36(2.8)
ow	b<u>ough</u>t	53(0.2)	4(0.2)	42(0.3)	65(0.3)	4(1.1)	31(1.2)	52(0.7)	4(1.2)	44(1.0)
u	b<u>oo</u>k	54(1.0)	4(0.3)	42(0.9)	65(1.2)	4(1.3)	32(1.7)	55(1.2)	3(0.4)	42(1.2)
uh	b<u>u</u>t	53(0.3)	4(0.4)	42(0.6)	65(0.3)	4(1.6)	31(1.7)	53(0.6)	3(0.4)	44(0.7)
mean value		54	4	42	64	4	32	55	4	41
true value		55	3	42	66	11	23	59	13	28

Table 1: State durations as percentages of a pitch period. Values are mean averages of the mean state duration across 100 realizations of a given synthetic vowel, with standard deviations in parentheses.

$$\bar{\mathbf{A}}_i = \begin{bmatrix} a_i(p) & a_i(p-1) & a_i(p-2) & \dots & a_i(1) \\ 0 & a_i(p) & a_i(p-1) & \dots & a_i(2) \\ \vdots & & & & \vdots \\ 0 & 0 & 0 & \dots & a_i(p) \\ 0 & 0 & 0 & \dots & 0 \\ \vdots & & & & \vdots \\ 0 & 0 & 0 & \dots & 0 \end{bmatrix}$$

$$\mathbf{A}_i = \begin{bmatrix} 0 & 0 & 0 & \dots & 0 \\ a_i(1) & 0 & 0 & \dots & 0 \\ a_i(2) & a_i(1) & 0 & \dots & 0 \\ \vdots & & & & \vdots \\ a_i(p) & a_i(p-1) & a_i(p-2) & \dots & 0 \\ 0 & a_i(p) & a_i(p-1) & \dots & 0 \\ \vdots & & & & \vdots \\ 0 & 0 & 0 & \dots & a_i(1) \end{bmatrix}$$

$$\mathbf{y} = \begin{bmatrix} y_1 & y_2 & y_3 & \dots & y_T \end{bmatrix}'$$

5. The Bounded State Duration (BSD) Viterbi Algorithm

To determine the optimal state sequence given model parameters, a bounded state duration (BSD) Viterbi algorithm similar to [4], rather than the standard Viterbi algorithm is used. This places lower and upper bounds, $d_1(i)$ and $d_2(i)$ respectively, on permissible state durations for the ith state and is necessary to prevent short-duration spurious state changes. The algorithm is based on the following recursions where bounds are specified *a priori*,

$$\phi_t(j,d) = \phi_{t-1}(j,d-1) + \log b_j(y_t) \qquad d \geq 2 \qquad (7)$$

$$\phi_t(j,1) = \max_{d_1(i)}^{d_2(i)} \left[\phi_{t-1}(i,d) + \log p_i(d)\right] + \log \pi_{i,j} + \log b_j(y_t) \qquad (8)$$

$p_i(d) = \pi_{i,i}^{d-1}(1-\pi_{i,i})$ is the probability of duration d in state i, and $b_j(y_t)$ is the jth state emission probability at time t. The likelihood score δ_T at time T is

$$\delta_T = \max_{i=1}^{3} \max_{d=1}^{d_2(i)} \left[\phi_T(i,d) + \log p_i(d)\right] \qquad (9)$$

For high signal-to-noise ratio (SNR) synthetic speech, the standard Viterbi algorithm is satisfactory.

6. Experiments

Firstly, experiments are conducted on synthetic speech and consist of synthesis, analysis and validation stages. Synthesis involves the generation of speech of length 2000 samples and pitch period 200 samples (no vocal jitter). Synthesis consists of an integrated-LF pulse generator producing modal, breathy and fry voice (glottal subsystem), a cascade of four 2nd order filters simulating 10 different vowels (vocal tract subsystem), and a differencer (lip-radiation subsystem), followed by white Gaussian noise addition (added at 80dB SNR for the purpose of preventing degeneracy in the EM algorithm). Analysis involves the use of the 3-state AR-HMM, 30 iterations of the EM algorithm where initial conditions and state vectors are kept fixed, and the BSD Viterbi algorithm to estimate the state sequence. Validation involves comparing the mean estimated state durations with their true values. Results are recorded in Table 1.

Secondly, experiments are conducted on real speech data. The central portions of long time-duration vowels are extracted from the TIMIT database (male speakers, dialect 1), and analysed as above. Reasonable state durations are given for all speakers as detailed in Table 2.

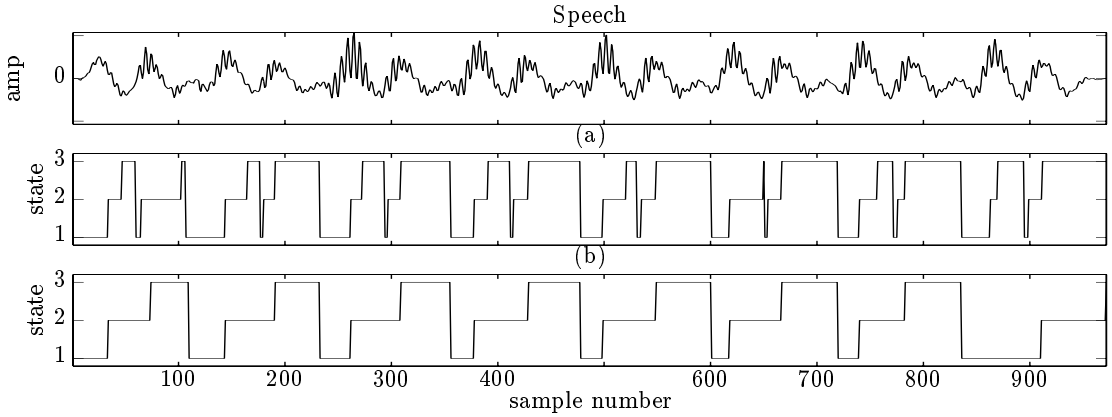

Figure 3: Example segmentation of real speech using the (a) standard and (b) the BSD Viterbi algorithms.

For both synthetic and real speech, changepoints are initialised at glottal closure instants estimated similarly to [5], and at 10 and 30 % of a pitch period lagging these. Also minimum durations are 30, 4 and 20 % of a pitch period for the *open+*, *open-* and *closed* phases respectively. All AR model orders are 16.

speaker	number of vowels	*open+*	*open-*	*closed*
mrcg0	90	49(14)	17(12)	37(12)
mrdd0	80	49(14)	13(9)	38(11)
mrso0	97	48(13)	21(11)	33(11)
mrws0	68	48(11)	15(11)	38(10)
mtjs0	78	48(11)	19(11)	36(11)
mtpf0	72	49(14)	14(11)	37(13)
mtrr0	93	45(11)	21(16)	44(14)
mwad0	58	50(14)	14(13)	39(11)
mwar0	85	47(12)	10(8)	43(12)

Table 2: State durations as percentages of a pitch period. Values are mean averages of the mean state duration across all long vowels for given TIMIT speakers.

7. Discussion

Experiments on synthetic speech show that segmentation depends on state sequence initialisation, a large *open+* phase model order is required to capture the low frequency resonance due to the glottis, segmentation deteriorates at low SNRs due to the absence of an explicit observation noise model, and the *open-* and *closed* changepoint is difficult to accurately estimate because the glottal signal often decays rapidly towards zero.

Figure 3 shows the segmentation of a vowel for the *mtrr0* TIMIT speaker, and its segmentation using the standard and BSD Viterbi algorithms. This shows the advantage of the BSD algorithm at preventing spurious state changes. An alternative is to introduce explicit state duration modelling into the HMM [6], but at increased computational cost.

An alternative to EM is Monte Carlo Markov chain (MCMC), which has been applied to the glottal segmentation problem with comparable results [7]. The use of informative priors may give more robust segmentation. Parameters are estimated using posterior distributions.

Acknowledgements. Gavin Smith is supported by the Schiff Foundation, Cambridge, and would like to thank M. Niranjan for help during initial research.

8. REFERENCES

1. J.R. Deller, J.G. Proakis, and J.H.L. Hansen. *Discrete-Time Processing of Speech Signals.* Macmillan Publishing Co., USA, 1993.
2. G. Fant, J. Liljencrants, and Q.G. Lin. A Four Parameter Model of Glottal Flow. *STL-QPSR*, 2-3:119–156, 1985.
3. L.R. Rabiner and B-H. Juang. *Fundamentals of Speech Recognition.* Prentice Hall, Englewood Cliffs, NJ, 1993.
4. H-Y. Gu, C-Y. Tseng, and L-S. Lee. Isolated-Utterance Speech Recognition Using Hidden Markov Models With Bounded State Durations. *IEEE Transactions on Signal Processing*, 39(8):1743–1752, 1991.
5. M. Hahn and D-G Kang. Precise Glottal Closure Instant Detector For Voiced Speech. *Electronics Letters*, 32(23):2117–2118, 1996.
6. G.A. Smith and A.J. Robinson. Segmentation of a Speech Waveform According to Glottal Timing Information using an Explicit Duration Autoregressive-HMM. Technical Report CUED/F-INFENG/TR.390, Cambridge University Engineering Dept., UK, 2000.
7. G.A. Smith and A.J. Robinson. Segmentation of a Speech Waveform According to Glottal Timing Information using a Standard Autoregressive-HMM. Technical Report CUED/F-INFENG/TR.389, Cambridge University Engineering Dept., UK, 2000.