Department of Computer Science
University of Toronto
http://learning.cs.toronto.edu

6 King's College Rd, Toronto
M5S 3G4, Canada
fax: +1 416 978 1455

Funded in part by NSERC

Revised May 20, 2004

UTML TR 2004–001

A Segmental HMM for Speech Waveforms

Kannan Achan, Sam Roweis, Aaron Hertzmann, Brendan Frey
Machine Learning Group, University of Toronto

Abstract

We present a purely time domain approach to speech processing which identifies waveform samples at the boundaries between glottal pulse periods (in voiced speech) or at the boundaries between unvoiced segments. An efficient algorithm for inferring these boundaries and estimating the average spectra of voiced and unvoiced regions is derived from a simple probabilistic generative model. Competitive results are presented on pitch tracking, voiced/unvoiced detection and timescale modification; all these tasks and several others can be performed using the single segmentation provided by inference in the model.

A Segmental HMM for Speech Waveforms

Kannan Achan, Sam Roweis, Aaron Hertzmann, Brendan Frey
Machine Learning Group, University of Toronto

1 Speech Segments in the Time Domain

Processing of speech signals directly in the time domain is commonly regarded to be difficult and unstable, due to fact that perceptually very similar utterances exhibit very large variability in their raw waveforms. As a result, by far the most common preprocessing step for most speech systems is to convert the raw waveform into a time-frequency representation, using a variety of spectral analysis and filterbank techniques. These methods often discard phase, employ an arbitrary uniform windowing in time and can be computationally demanding. In this paper we pursue a purely time domain approach to speech processing in which we identify the samples at the boundaries between glottal pulse periods (in voiced speech) or at the boundaries between unvoiced segments of similar spectral shape ("colour"). As we show, working directly with the raw speech data can be extremely practical, produce excellent results on a wide variety of speech tasks, allows operations not possible in the spectral domain and is often computationally as attractive (or more) than short-time Fourier methods.

Having identified segment boundaries, we can perform a variety of important low level speech analysis operations directly and conveniently. For example, we make a voiced/unvoiced decision on each segment by examining the periodicity of the waveform in that segment only. In voiced segments we can estimate the pitch as the reciprocal of the segment length. Timescale modification without pitch or format distortion can be achieved by stochastically eliminating or replicating segments in the time domain directly. More sophisticated operations, such as pitch modification, gender and voice conversion, and companding (volume equalization) are also naturally performed by operating on waveform segments one by one without the need for a cepstral or other such representation.

The computational challenge with this approach is in efficiently and robustly identifying the segment boundaries, across silence, unvoiced and voiced segments. In this paper we introduce a segmental Hidden Markov Model, defined on variable length sections of the time domain waveform, and show that performing inference in this model allows us to identify segment boundaries and achieve excellent results on the speech processing tasks described above.

2 A Probabilistic Generative Model of Time-Domain Speech Segments

The goal of our algorithm is to break the time domain speech signal $\boldsymbol{x}= x_1, \ldots, x_N$ into a set of segments, each of which corresponds to a glottal pulse period or a segment of unvoiced colored noise. Let $\boldsymbol{b} = b_0, \ldots, b_K$ denote the time index of the segment boundaries, then the left and right boundaries of the k^{th} segment would be b_{k-1} and $b_k - 1$. For notational ease, let $\boldsymbol{x}_b^{b'}$ be the vector $[x_b, \ldots, x_{b'}]$. The binary hidden variable controlling the mode of segment k is denoted

† Thanks to John Hopfield.

by v_k, where $v_k = 0$ means that segment k is unvoiced and $v_k = 1$ means it is voiced and let $\boldsymbol{v}$ be the set $v_1, \ldots, v_k$.

Assuming that the segments are generated by a first order Markov chain, we have four possible type of transitions in the generative process - voiced to voiced, voiced to unvoiced, unvoiced to voiced and unvoiced to unvoiced. Given the segment boundaries, $b_0, b_2, \ldots, b_K$, our model assumes that when there are two successive voiced segments, the second segment is accurately modeled as a time-warped, amplitude-scaled and amplitude-shifted version of the previous segment. This is motivated by the strong phase coherence in harmonic regions of the speech wave. We denote the 2-vector containing the amplitude-scale and amplitude-shift used to map segment $k-1$ to segment k by $\boldsymbol{t}_k$.

In cases where the two successive frames are not voiced, our model assumes that the phase information present in the second segment cannot be predicted from the previous segment. So, in these cases, only the power spectrum of the second segment is modeled, as described below.

Given the segment boundaries $\boldsymbol{b}$, the segment types (voiced or unvoiced) $\boldsymbol{v}$, and the transformation variables $\boldsymbol{t} = [\boldsymbol{t}_1, \ldots, \boldsymbol{t}_K]$ (which are only relevant when two successive frames are voiced), the generative model is a conditional Markov model

$$P(\boldsymbol{x}|\boldsymbol{b}, \boldsymbol{v}, \boldsymbol{t}) = \prod_{k=2}^{K} P(\boldsymbol{x}_{b_{k-1}}^{b_k} | \boldsymbol{x}_{b_{k-2}}^{b_{k-1}}, v_k, v_{k-1}, \boldsymbol{t}_k). \tag{1}$$

Depending on the hidden variables $\boldsymbol{v}$, each transition distribution in the Markov model takes one of three forms:

$$P(\boldsymbol{y}|\boldsymbol{y}', v_{k-1}, v_k, \boldsymbol{t}_k) = \begin{cases} P_{unvoiced}(\boldsymbol{y}) & \text{if } v_{k-1} = 0 \\ P_{voiced}(\boldsymbol{y}) & \text{if } v_{k-1} = 1 \text{ and } v_k = 0 \\ P_{harmonic}(\boldsymbol{y}|\boldsymbol{y}', \boldsymbol{t}_k) & \text{if } v =_{k-1} 1 \text{ and } v_k = 1 \end{cases}$$

where,

$$P_{unvoiced}(\boldsymbol{y}) \propto \exp(-\frac{1}{2}(\mathbf{f}(\boldsymbol{y}) - \boldsymbol{\lambda}_0)^\top \boldsymbol{\Phi}_0^{-1}(\mathbf{f}(\boldsymbol{y}) - \boldsymbol{\lambda}_0)) \tag{2}$$

$$P_{voiced}(\boldsymbol{y}) \propto \exp(-\frac{1}{2}(\mathbf{f}(\boldsymbol{y}) - \boldsymbol{\lambda}_1)^\top \boldsymbol{\Phi}_1^{-1}(\mathbf{f}(\boldsymbol{y}) - \boldsymbol{\lambda}_1)). \tag{3}$$

Here, the vector function $\mathbf{f}(\boldsymbol{y}) = abs(\mathcal{F}(\boldsymbol{y}))/ \parallel abs(\mathcal{F}(\boldsymbol{y})) \parallel$ computes the normalized power spectrum of its argument, where $\mathcal{F}$ is the DFT matrix. We define $\boldsymbol{\lambda}_0$ and $\boldsymbol{\Phi}_0$ to be the normalized mean and covariance (assumed to be diagonal) of the power spectrum for unvoiced regions and $\boldsymbol{\lambda}_1$ and $\boldsymbol{\Phi}_1$ to be the same for the voiced regions. These parameters are set to reasonable prior values initially, but are then adapted to the data using an EM algorithm, as described later so that the final segmentation does not depend on their initial values, only on their converged estimates.

The harmonic model which predicts a voiced segment from a previous voiced segment is given by

$$P_{harmonic}(\boldsymbol{y}|\boldsymbol{y}', \boldsymbol{t}) \propto \exp(-\frac{1}{2}(\boldsymbol{y} - [\mathbf{r}(\boldsymbol{y}') \; \mathbf{1}]\boldsymbol{t}_k)^\top \boldsymbol{\Sigma}^{-1}(\mathbf{r}(\boldsymbol{y}) - [\boldsymbol{y}' \; \mathbf{1}]\boldsymbol{t}_k)), \tag{4}$$

where the vector function $\mathbf{r}(\boldsymbol{y}')$ performs linear interpolation and resampling on $\boldsymbol{y}'$ to produce a vector with the same dimension as $\boldsymbol{y}$. The expression $[\mathbf{r}(\boldsymbol{y}')\mathbf{1}]$ is a 2-column matrix with 1's in the second column. The product $[\mathbf{r}(\boldsymbol{y}')\mathbf{1}]\boldsymbol{t}_k$ scales each element of $\mathbf{r}(\boldsymbol{y}')$ by the first element of $\boldsymbol{t}_k$ and then shifts each element of the result by the second element of $\boldsymbol{t}_k$.

The distribution over the boundaries, voiced/unvoiced switches and transformations has a product form:

$$P(\boldsymbol{b}, \boldsymbol{v}, \boldsymbol{t}) = P(\boldsymbol{b})P(\boldsymbol{v})P(\boldsymbol{t}) \tag{5}$$

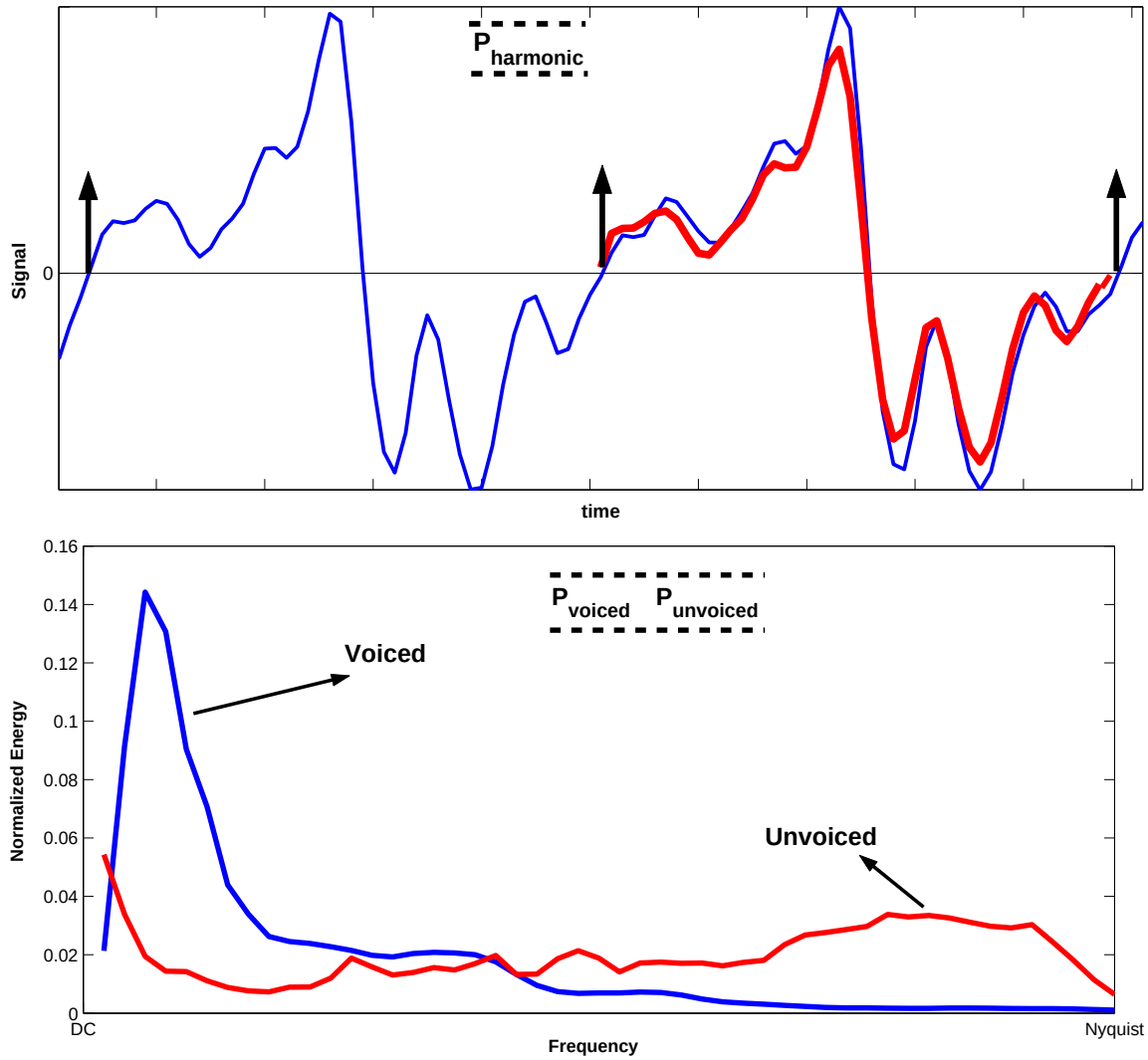

Figure 1: Generative model: *(top)*Transformation model $\boldsymbol{t}$: Two glottal pule periods corresponding to a voiced region is shown. The thicker overlay on the second pulse period is the prediction based on the previous glottal pulse period using the transformation model $\boldsymbol{t}$. *(bottom)*Spectrum of voiced and unvoiced region: The voiced model has all its energy spread along the low frequency components; the unvoiced model has bulk of its energy in the high frequency regions and some along mid and low frequency

Generally the joint prior probability mass function on segment boundaries $P(\boldsymbol{b})$ can be quite complex. Since the computational complexity of the inference algorithm will depend on the number of allowed configurations of segment boundaries, we use a prior that is non-zero only on an appropriate subset of configurations. In particular, we exploit a very simple heuristic (first suggested by John Hopfield in 1998) by *restricting segments to begin and end only on zero crossings of the signal* (or possibly only on upward or downward going zero crossings). This restriction also allows arbitrary segments to be relocated beside each other and still preserve waveform continuity, which will be important in our later applications. To further restrict the range of inferred segment lengths, we require that $\Delta_{\min} \leq b_k - b_{k-1}$, where $\Delta_{\min}$ is the minimum segment length, satisfying $\Delta_{\min} > 0$. This minimum length is selected by hand and is determined by the expected range of pitch periods and the sampling frequency, in a straightforward fashion. We assume the probability $P(\boldsymbol{b})$ is otherwise uniform, subject to the above constraints. The distribution over the voiced/unvoiced switch is uniform. The scale and

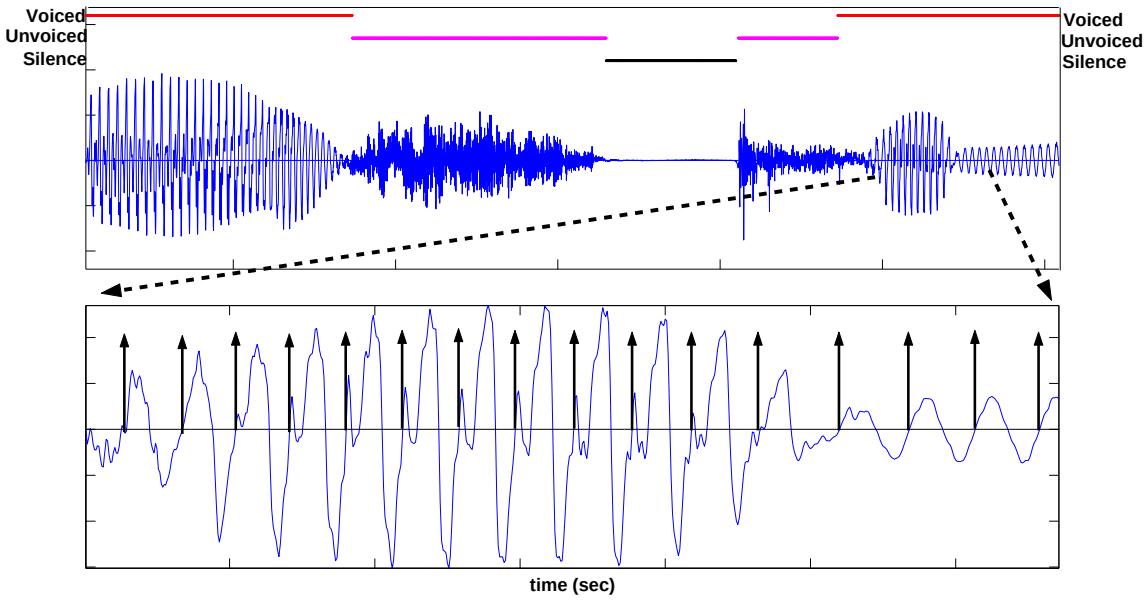

Figure 2: *(top)*Results of inference on an utterance from WSJ dataset. voicing/unvoicing decision is indicated by the bars above the signal *(bottom)*Inferred segments - upwards arrows are used to mark the segment boundaries.

shift variables are assumed to be independent and normally distributed with a large variance. The joint distribution over the signal, segment boundaries $\boldsymbol{b}$, segment types $\boldsymbol{v}$and transformation parameters can be written as,

$$P(\boldsymbol{x}, \boldsymbol{b}, \boldsymbol{v}, \boldsymbol{t} | \boldsymbol{\lambda}_0, \boldsymbol{\lambda}_1, \boldsymbol{\Phi}_0, \boldsymbol{\Phi}_1) \propto P(\boldsymbol{x} | \boldsymbol{b}, \boldsymbol{v}, \boldsymbol{t}) P(\boldsymbol{b}) P(\boldsymbol{v}) P(\boldsymbol{t}) \tag{6}$$

Each segment is either modeled as a noisy copy of the transformed version of the previous segment or is generated using the parameters $\boldsymbol{\lambda}_0$ and $\boldsymbol{\lambda}_1$. These assumptions simplify the inference and estimation algorithm described below. Of course, the segment boundaries are unknown and must be inferred from the speech wave: this inference is the main computation performed by our algorithm.

3 Inference and Learning

Given a time-domain signal, the computational task now at hand is to determine the segment boundaries, segment types and transformation parameters (where needed). We present an iterative algorithm to efficiently infer the hidden variables and learn the parameters of the model. Of course, the number of valid configurations of the boundary variables is exponential in the length of the waveform and this makes computing the full posterior distribution over segmentations intractable. We outline an approach (similar to the max-product algorithm for inference in graphical models) which finds the MAP estimates of the hidden variables (i.e. the single most likely segmentation, voiced/unvoiced labeling and transformation parameters). In fact, we use this estimation process as an approximate inference step for an EM-like algorithm in which the approximate E-step corresponds to using a variational distribution that is a delta function with peaks at the MAP estimate.

To simplify the inference algorithm, we make use of the fact that given boundary variables

and their segment types, the MAP estimate of the transformations can be computed locally.

$$\begin{aligned}&\underset{\boldsymbol{t}_k}{\operatorname{argmax}}\ P(\boldsymbol{x}_{b_0}^{b_k}, b_0, \ldots, b_k, v_1, \ldots, v_k, \boldsymbol{t}_1, \ldots, \boldsymbol{t}_k)\\ &= \arg\max_{\boldsymbol{t}_k} P(\boldsymbol{t}_k) P(\boldsymbol{x}_{b_{k-1}}^{b_k} | \boldsymbol{x}_{b_{k-2}}^{b_{k-1}}, v_{k-1} = v_k = 1, \boldsymbol{t}_k). \end{aligned} \tag{7}$$

In particular, the time-warping is unique and is given by $(b_k - b_{k-1})/(b_{k-1} - b_{k-2})$. The warped version of $\boldsymbol{x}_{b_{k-2}}^{b_{k-1}}$ is denoted by $\hat{\boldsymbol{x}}_{b_{k-2}}^{b_{k-1}}$ and can be obtained using linear interpolation. Note that whereas $\boldsymbol{x}_{b_{k-2}}^{b_k - 1}$ contains $(b_{k-1} - b_{k-2})$ samples, $\hat{\boldsymbol{x}}_{b_{k-2}}^{b_{k-1}}$ contains $b_k - b_{k-1}$ samples. The amplitude-domain scale β_k and shift γ_k are obtained by performing a least-squares regression of $\hat{\boldsymbol{x}}_{b_{k-2}}^{b_{k-1}}$ onto $\boldsymbol{x}_{b_{k-1}}^{b_k}$, *i.e.* by solving

$$\beta_k^*, \gamma_k^* = \arg\min_{\beta_k, \gamma_k} (\beta_k \hat{\boldsymbol{x}}_{b_{k-2}}^{b_{k-1}} + \gamma_k - \boldsymbol{x}_{b_{k-1}}^{b_k})^\top (\beta_k \hat{\boldsymbol{x}}_{b_{k-2}}^{b_{k-1}} + \gamma_k - \boldsymbol{x}_{b_{k-1}}^{b_k}), \tag{8}$$

For a given configuration of b_{k-2}, b_{k-1}, b_k, we denote the optimal transformation obtained in the above fashion by $\mathbf{t}_k^* = [\beta_k^*, \gamma_k^*]$. This optimization is performed at each step of the search over the boundary variables, when the adjacent segments happen to be voiced.

Estimating the MAP setting for the boundary variables and the corresponding segment types involves running a recursion isomorphic to max-product or Viterbi. In practice we can implement this inference algorithm by populating two dynamic programming grids in the space of valid configurations of the boundary variables. The grids, $\mathcal{C}_0$ and $\mathcal{C}_1$ are two dimensional arrays with dimension given by the cardinality of the set of upward zero crossings. $\mathcal{C}_0(b_a, b_b)$ represents the (log) probability of the best segmentation of $\boldsymbol{x}_{b_0}^{b_b}$ in which the last segment is unvoiced and bounded by b_a and b_b; Similarly, $\mathcal{C}_1(b_a, b_b)$ represents the (log) probability of the best segmentation of $\boldsymbol{x}_{b_0}^{b_b}$ in which the last segment is voiced and bounded by b_a and b_b.

There is now a simple recursion for filling in the table, which corresponds exactly to message passing:

$$\mathcal{C}_s(b_a, b_b) = \max_{b_c, q \in \{0,1\}} P(\boldsymbol{x}_{b_0}^{b_b}, v(x_{b_a}^{b_b}) = s) \tag{9}$$

$$= \max_{b_c, q \in \{0,1\}} \mathcal{C}_q(b_c, b_a) P(\boldsymbol{x}_{b_a}^{b_b} | \boldsymbol{x}_{b_c}^{b_a}, v(x_{b_a}^{b_b}) = s), \quad s \in \{0, 1\} \tag{10}$$

where we have introduced a simplifying notation for segment type, $v(\boldsymbol{x}_{b_a}^{b_b})$ (of segment $\boldsymbol{x}_{b_a}^{b_b}$). The optimal value of b_c in the above optimization should also be stored in a table.

Once the dynamic programming grid and the associated data structures are filled in, we use a Viterbi-like algorithm to backtrack and find the single best configuration (MAP estimate) of the boundary variables and the corresponding segment types. We highlight the fact that inference can be done tractably due to the sparsity induced by the prior on $\boldsymbol{b}$which sets the minimum and maximum segment lengths.

The M-step of the EM algorithm involves updating the parameters of the model by optimizing the expected value of the complete log likelihood under the (approximate delta-function) posterior distribution inferred in the E-step. The updates for the parameters $\boldsymbol{\lambda}_0$, $\boldsymbol{\lambda}_1$, $\boldsymbol{\Phi}_0$, $\boldsymbol{\Phi}_1$ are

$$\boldsymbol{\lambda}_g = \frac{1}{\sum_{l=1}^K \delta(v_l = g)} \sum_{k=1}^K \delta(v_k = g) \mathbf{f}(\boldsymbol{x}_{b_{k-1}}^{b_k}) \tag{11}$$

$$\boldsymbol{\Phi}_g = \sum_{k=1}^K \delta(v_k = g)(\mathbf{f}(\boldsymbol{x}_{b_{k-1}}^{b_k}) - \boldsymbol{\lambda}_g)(\mathbf{f}(\boldsymbol{x}_{b_{k-1}}^{b_k}) - \boldsymbol{\lambda}_g)^\top, \quad g = \{0, 1\} \tag{12}$$

Notice that the update for $\boldsymbol{\lambda}_0$ and $\boldsymbol{\lambda}_1$ correspond to the normalized average of the spectrum of unvoiced and voiced segments found in the inference step respectively, while $\boldsymbol{\Phi}$ are the diagonal variances of those spectra.

We initialize $\boldsymbol{\lambda}_0$ to have uniform weight at the top 10% of Nyquist frequency and $\boldsymbol{\lambda}_1$ to uniform weights at the bottom 10% of Nyquist frequency. All the other frequency bins of both the spectra are initially set to zero. This initialization is driven by the prior knowledge that voiced segments are rich in low frequency content and unvoiced signals carry more high frequency information. Typical converged estimates of the means $\boldsymbol{\lambda}_0$ and $\boldsymbol{\lambda}_1$ can be seen in Fig.1.

4 Experiments with Pitch Tracking, Voiced-Unvoiced Detection and Timescale Modification

We can apply the results of our segment inference algorithm to a wide range of speech processing tasks. By replicating or deleting some or all of the inferred segments, we can easily achieve high quality timescale modification without changing the perceived pitch or formant structure of the utterance. In voiced regions, we can directly estimate the pitch by taking the reciprocal of the segment length. Below, we present results on timescale modification, voiced/unvoiced discrimination, and pitch tracking. Other applications such as gender and voice conversion, companding and concert hall effects are also possible. We emphasize that all the experiments were performed in *time domain* using the inferred pitch periods. For audio demonstrations and samples, please check http://www.psi.toronto.edu/~kannan/Segmental

As initial experiments, we applied our segmental inference procedure to clean, wideband recordings of single-talker speech, from both males and females taken from the the Keele pitch reference dataset [4] and from the Wall Street Journal (WSJ) corpus.

The threshold $\Delta_{\min}$ on the minimum pitch period was set at to be $2ms$ (corresponding to a maximum pitch of 500Hz) and $\Delta_{\max}$ on the maximum pitch period was set to $20ms$ (corresponding to a minimum pitch of 50Hz). The set of upward going zero crossings followed by at least two positive samples were used to define the space of feasible segment boundary points.

Voicing Detection and Pitch Tracking

For voicing detection and pitch tracking, we evaluated the estimates obtained using our algorithm using the Keele dataset, since it has ground truth values for these quantities. (In particular, the Keele data has utterances spoken by both male and female speakers and includes a reference estimate for the fundamental frequency at a resolution of 10ms. Each utterance is approximately 30 seconds long and the sampling frequency is 20kHz.)

Our method was able to correctly identify 89.03 % of the voiced segments averaged over all the 10 utterances of males and females in the Keele dataset.

Pitch tracking is trivially achieved by taking the reciprocal of the segment lengths in the voiced regions. Results for a single utterance in the Keele dataset spoken by a female speaker is shown in Fig.3. Pitch estimates obtained using our approach are very consistent with the reference estimates; similar performance was obtained on other utterances in the dataset as well. Averaged over 10 utterances the median absolute pitch error was 7.8Hz.

It is well known that excitation for voiced speech manifests as sharp bursts at integer multiples of fundamental frequency. In Fig.3, we have shown a few integer multiples of the fundamental frequency of a signal on its spectrogram using pitch estimates obtained from the application of our algorithm.

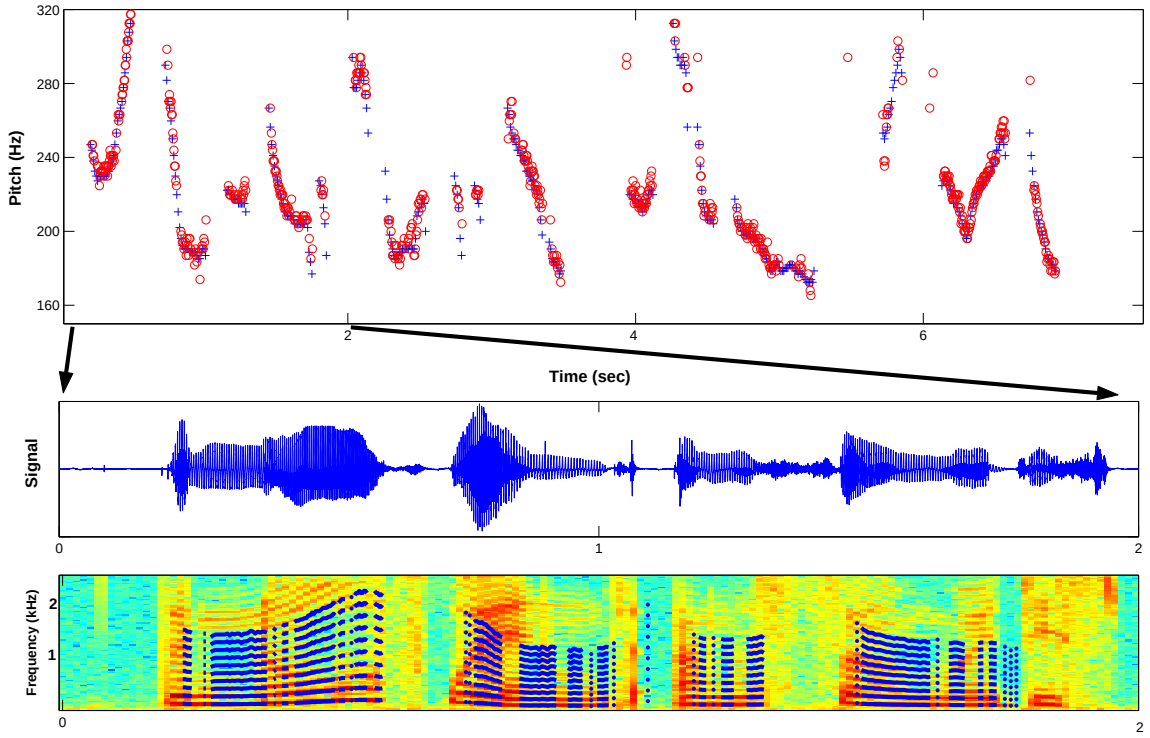

Figure 3: *(top)*Pitch estimates using our approach for a female speaker in the Keele dataset. Notice that the inferred pitch(red circle) consistently agrees with the reference provided (blue plus mark). *(center)* section of the input time domain signal corresponding to the marked region. *(bottom)* Spectrogram of the section of the time domain signal marked at the integer multiples of fundamental frequency. For clarity only low frequencies are shown.

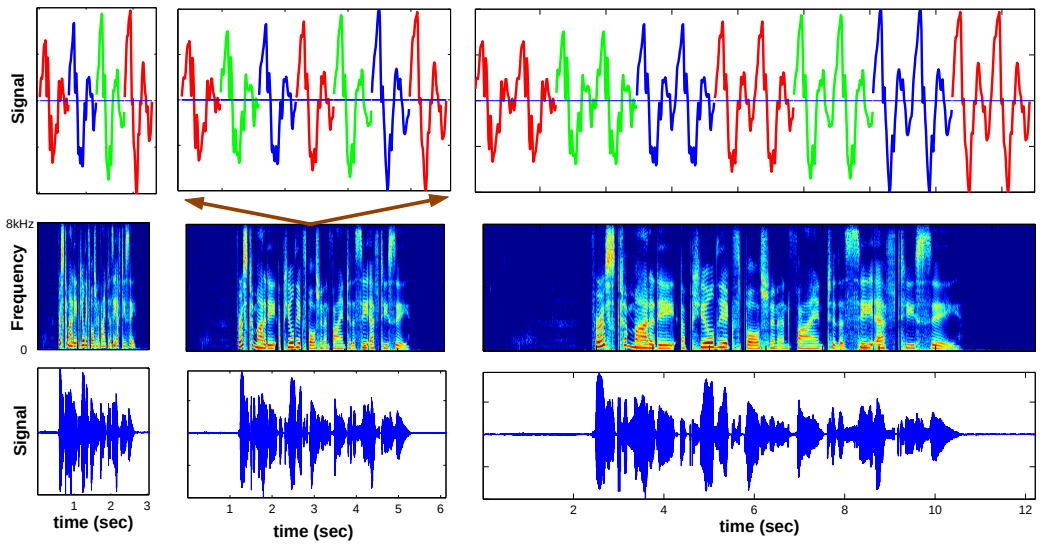

Figure 4: The spectrogram of time scale modified faster and slower versions of a signal are shown. The actual time domain operation is shown on top for a particular time instant in the spectrogram.

Time Scale Modification

For timescale modification experiments, we have used utterances from the WSJ corpus. Once the segments are identified by our algorithm, we can play the signal twice as fast by delet-

ing every other segment and concatenating the remaining ones; similarly by replicating each segment we can achieve the effect of playing the at half the speed (two times slower); this is further illustrated in Fig.4. This approach is substantially different from methods such as [1] that manipulate spectrograms. By doing all of our operations directly in the time domain we never need to worry about inconsistent phase estimates.

5 Conclusions

We have presented a simple segmental Hidden Markov Model for analyzing speech waveforms directly in the time domain and derived an efficient algorithm for approximate inference in this model. Applied to an observed signal, this inference algorithm is capable of automatically identifying the boundaries of glottal pulse periods in voiced speech and of unvoiced segments; as well, it estimates the average power spectra of voiced and unvoiced speech. Using these inferred boundaries we are able to easily and accurately detect voicing, track pitch and modify the timescales. We are investigating other possible applications with the same basic model, including voice conversion, volume equalization and reverberant filtering.

One limitation of our model is that it can only be applied to clean, single speaker recordings. We are currently formulating an extended model in which the clean signal is represented by latent variables and the observed signal is modeled as a noisy realization of those underlying quantities. Inference in this extended model will simultaneously perform denoising and segmentation on the input waveform.

While many competitive algorithms exist for solving each of these speech processing tasks in isolation, the appeal of our model is that it directly analyzes the speech wave in an unsupervised fashion and decomposes it into fundamental atomic blocks. After this segmentation, many disparate speech processing tasks are quite naturally performed, indicating that we have managed to extract some fundamental structure from the signal. The algorithm is extremely simple and efficient and builds on the most basic facts about speech production, namely that there is a voiced mode (in which phase is coherent and signal energy is concentrated in the lower part of the spectrum) and an unvoiced mode (in which phase is random).

Acknowledgments

STR is supported in part by the LEARN project of IRIS and by NSERC Canada.

References

[1] Roucos, S. and A. M. Wilgus. High Quality Time-Scale Modification for Speech. *In Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, IEEE, 1985, 493-496.*

[2] D. Talkin, A robust algorithm for pitch tracking (RAPT), *in Speech Coding and Synthesis, W. B. Kleijn and K. K. Paliwal, Eds., pp. 497 518. Elsevier Science, 1995.*

[3] L. K. Saul, D. D. Lee, C. L. Isbell, and Y. LeCun Real time voice processing with audiovisual feedback: toward autonomous agents with perfect pitch. *Advances in Neural Information Processing Systems 15. MIT Press: Cambridge, MA, 2003*

[4] Plante, F and Ainsworth, W.A. and Meyer, G.F A Pitch Extraction Reference Database *In Proceedings of Eurospeech, 1995*

[5] F. Sha, J. A. Burgoyne, and L. K. Saul Multiband statistical learning for f0 estimation in speech *In Proceedings of the International Conference of Speech, Acoustics, and Signal Processing, 2004),*

[6] K. Achan, S. T. Roweis, B. J. Frey A Segmental HMM for Speech Waveforms *Technical Report UTML-TR-2004-001, University of Toronto, January 2004*