A SEGMENT-BASED PROBABILISTIC GENERATIVE MODEL OF SPEECH

Kannan Achan, Sam Roweis, Aaron Hertzmann, Brendan Frey

Departments of Computer Science and ECE, 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 of 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.

1. INTRODUCTION

Processing of speech signals directly in the time domain is commonly regarded to be difficult and unstable, due to the fact that perceptually very similar utterances exhibit very large variability in their raw waveforms. As a result, the majority of speech processing systems preprocess the raw waveform into a time-frequency representation, using a variety of spectral analysis and filterbank techniques. These methods often discard phase, and employ an arbitrary uniform windowing in time. In contrast, the time domain is appealing because it does not discard any information from the input signal. 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").

While many competitive algorithms [1, 2, 3, 4] exist for solving various individual speech processing tasks in isolation, *the appeal of our model is that it enables a wide range of applications in a single framework.* Our model does not require any training data and all operations are performed using only the input signal. 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 vs. 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. In fact, we show preliminary evidence that challenging operations, such as denoising or replacing missing waveform samples will be possible with this approach. More sophisticated operations, such as pitch modification, gender and voice conversion, and companding (volume equalization) should be possible by operating on waveform segments one by one without the need for a cepstral or other such representation.

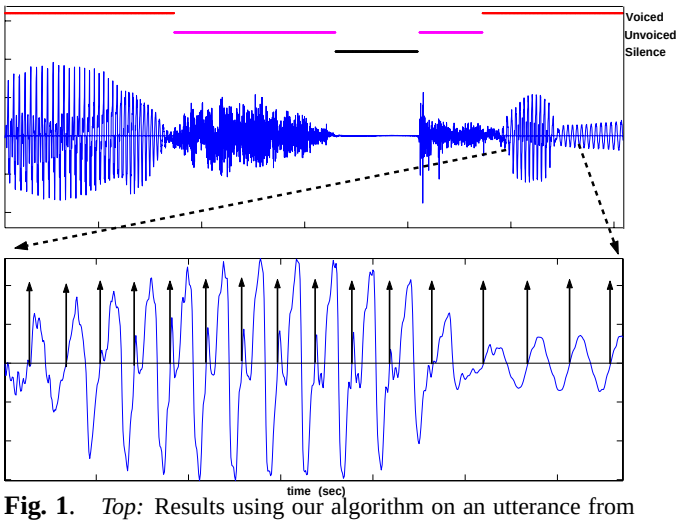

Fig. 1. *Top:* Results using our algorithm on an utterance from the WSJ dataset; voicing/unvoicing decision is indicated by the bars above the signal *Bottom:* Inferred segments. Upwards arrows are used to mark the segment boundaries.

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 several 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 either a glottal pulse period in voiced regions or a segment of unvoiced colored noise (see Fig. 1). Let $\boldsymbol{b} = b_0, \ldots, b_K$ denote the time indices of the segment boundaries so that the left and right boundaries of the k^{th} segment would be b_{k-1} and $b_k - 1$. Let $\boldsymbol{x}_b^{b'}$ be the vector $[x_b, \ldots, x_{b'}]$. The binary hidden variable controlling the type of segment k is denoted 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 $v_1, \ldots, v_k$. We detect silent segments beforehand by finding runs of contiguous samples all below a certain amplitude threshold. We then process each remaining waveform section between two silence periods in-

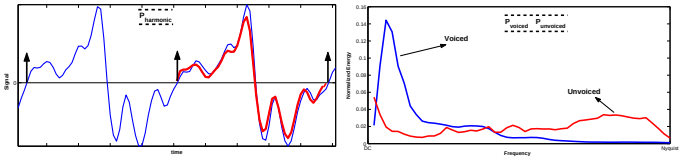

Fig. 2. *Left:* Two glottal pulse periods corresponding to a voiced region is shown. The prediction for the second segment, shown as a thick overlay is a time warped, amplitude scaled and shifted version of the previous segment. *Right:* Typical learned spectrum of voiced and unvoiced region. These are used to model a segment when it cannot be reliably predicted using its previous segment

dependently, breaking it into voiced and unvoiced segments. 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*
- *unvoiced to unvoiced*

Given the segment boundaries, $b_0, b_1, \ldots, b_K$, our model assumes that, when there are two successive voiced segments, the second segment is 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 $\boldsymbol{t}_k = (\beta_k, \gamma_k)$.

In cases where either of 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 $\boldsymbol{v}$ (voiced or unvoiced), and the transformation variables $\boldsymbol{t}$ (relevant only when two successive frames are voiced), the generative model is a conditional Markov model

$$P(\boldsymbol{x}|\boldsymbol{b},\boldsymbol{v},\boldsymbol{t}) = P(\boldsymbol{x}_{b_0}^{b_1}|v_1)\prod_{k=2}^{K} P(\boldsymbol{x}_{b_{k-1}}^{b_k}|\boldsymbol{x}_{b_{k-2}}^{b_{k-1}}, v_k, v_{k-1}, \boldsymbol{t}_k).$$

As the boundary condition for the Markov chain, we assume the segment $\boldsymbol{x}_{b_0}^{b_1}$ before the first sample is all zeros and is unvoiced. Depending on the hidden variables $\boldsymbol{v}$, each transition distribution $P(\boldsymbol{y}|\boldsymbol{y}', v_{k-1}, v_k, \boldsymbol{t}_k)$ in the Markov model takes one of three forms:

- $P_{unvoiced}(\boldsymbol{y})$ if $v_{k-1} = 0$
- $P_{voiced}(\boldsymbol{y})$ if $v_{k-1} = 1$ and $v_k = 0$
- $P_{harmonic}(\boldsymbol{y}|\boldsymbol{y}', \boldsymbol{t}_k)$ if $v_{k-1} = 1$ and $v_k = 1$

where,

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

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

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. The vector function $\mathbf{f}(\boldsymbol{y})$ computes the normalized power spectrum of $\boldsymbol{y}$ given by $\mathrm{abs}(\mathcal{F}(\boldsymbol{y}))/\parallel \mathrm{abs}(\mathcal{F}(\boldsymbol{y}))\parallel$, and resamples it to match the size of $\boldsymbol{\lambda}_0$ or $\boldsymbol{\lambda}_1$; here $\mathcal{F}$ is the DFT matrix.

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

$$P_{harmonic}(\boldsymbol{y}|\boldsymbol{y}', \boldsymbol{t}) \propto \exp(-\frac{1}{2}\mathbf{T_y}^\top \boldsymbol{\Sigma}^{-1}\mathbf{T_y}) \qquad (1)$$

where $\mathbf{T}_{\boldsymbol{y}} = (\boldsymbol{y} - (\beta_k(\mathbf{r}(\boldsymbol{y}')) + \gamma_k))$. 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 distribution over the boundaries, voiced/unvoiced switches and transformations has a product form given by $P(\boldsymbol{b},\boldsymbol{v},\boldsymbol{t}) = P(\boldsymbol{b})P(\boldsymbol{v})P(\boldsymbol{t})$

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[1] 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 $\Delta_{\min} \leq b_k - b_{k-1} \leq \Delta_{\max}$, where $\Delta_{\min}$ and $\Delta_{\max}$ are the minimum and maximum segment lengths, chosen by hand to reflect the expected range of pitch periods and the sampling frequency. We assume the probability $P(\boldsymbol{b})$ is otherwise uniform, subject to the above constraints. The distribution $P(\boldsymbol{v})$ over the voiced/unvoiced switch is uniform. The scale variables β_k are assumed to be independent and normally distributed with mean 1 and a variance that penalizes large deviations from the mean. The joint distribution over the signal, segment boundaries $\boldsymbol{b}$, segment types $\boldsymbol{v}$and transformation parameters can be written as

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

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.

2.1. 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 which finds the MAP estimates of the hidden variables (i.e., the single most likely segmentation, voiced/unvoiced labeling and transformation parameters).

[1] Suggested by John Hopfield in 1998

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 [5]. 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 $\boldsymbol{x}_{b_{k-1}}^{b_k}$, For a given configuration of b_{k-2}, b_{k-1}, b_k, we denote the optimal transformation obtained using linear regression 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 Viterbi or max-product algorithm. In practice, we can implement this inference algorithm by populating two dynamic programming grids in the space of valid configurations of the boundary variables. In this paper, we constraint the boundary variables to lie only on upward zero crossings.

Let $\mathbf{z}$ denote the set of all valid upward zero crossings. The grids, $\mathcal{C}_0$ and $\mathcal{C}_1$ are two dimensional square arrays with edge size given by the number of upward zero crossings. $\mathcal{C}_0(i,j)$ represents the probability of the best segmentation of $\boldsymbol{x}_{z_0}^{z_j}$ in which the last segment is unvoiced and bounded by z_i and z_j; Similarly, $\mathcal{C}_1(i,j)$ represents the probability of the best segmentation of $\boldsymbol{x}_{z_0}^{z_j}$ in which the last segment is voiced and bounded by z_i and z_j.

There is now a simple recursion for filling in the table:

$$\begin{aligned}\mathcal{C}_s(i,j) &= \max P(\boldsymbol{x}_{z_0}^{z_i}, \boldsymbol{x}_{z_i}^{z_j}, v_{ij} = s), \quad s \in \{0,1\} \\ &= \max_{k<i, q\in\{0,1\}} \mathcal{C}_q(k,i) P(\boldsymbol{x}_{z_i}^{z_j} | \boldsymbol{x}_{z_k}^{z_i}, v_{ij} = s)\end{aligned}$$

where we have introduced a simplifying notation for segment type, v_{ij} (of segment $\boldsymbol{x}_{z_i}^{z_j}$). The optimal value of z_k 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 $\boldsymbol{b}$.

The parameters of the model are estimated by maximizing the log likelihood of the observed waveform given the best possible segmentation inferred by dynamic programming. The new values of $\boldsymbol{\lambda}_0$ and $\boldsymbol{\lambda}_1$ correspond to the normalized average of the spectrum of unvoiced and voiced segments respectively, while $\boldsymbol{\Phi}_0$ and $\boldsymbol{\Phi}_1$ become the diagonal variances of those spectra respectively. We iterate between the inference and learning steps until the estimated parameters stabilize (usually 3-4 iterations). We have also tried learning using the full posterior distribution over segmentations, using a method isomorphic to the forward-backward algorithm and found the results were essentially the same as those produced by our Viterbi-based algorithm.

We initialize $\boldsymbol{\lambda}_0$ to have uniform weight in the interval $\omega = [.9f, f]$ and $\boldsymbol{\lambda}_1$ in $\omega = [0, .1f]$, where f is the Nyquist frequency. All the other frequency bins of both the spectra are initially set to zero. We preset the number of frequency bins to 100. Typical converged estimates of $\boldsymbol{\lambda}_0$ and $\boldsymbol{\lambda}_1$ are shown in Fig. 2. The threshold $\Delta_{\min}$ on the minimum pitch period was set at to be $2ms$ and $\Delta_{\max}$ on the maximum pitch period was set to $20ms$.

3. EXPERIMENTS

Since our algorithm learns a segment-based madel of the time-domain signal, the model can be used for a wide range of speech processing tasks. We emphasize that the appeal of our model is that this single framework applies to a wide range of applications. Below, we present results on pitch tracking, timescale modification, and voiced/unvoiced discrimination. 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, see

http://www.psi.toronto.edu/∼kannan/Segmental

Voicing detection and pitch tracking: In voiced regions, we can directly estimate the pitch by taking the reciprocal of the segment length. We evaluated the estimates obtained using our algorithm using the Keele dataset[2] [6] and compared our results with some of the other well known pitch tracking algorithms. The evaluation framework is similar to the one used in [7]. Unvoiced/voiced error reports the percentage of frames that were misclassified. Gross error denotes the percentage of voiced frames where the error in f_0 estimates exceed 20%. The fourth row reports the average error (in Hz) in f_0 estimates for frames without gross errors. It is worth pointing out that despite its generality, results from our algorithm are comparable with other state of the art techniques.

Results for a single utterance in the Keele dataset spoken by a female speaker is shown in Fig.3. It is well known that excitation for voiced speech manifests as sharp peaks 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.

Time Scale Modification: By replicating or deleting some or all of the inferred segments, we can easily perform timescale modification without changing the perceived pitch or formant structure of the utterance. 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 deleting every other segment and concatenating the remaining ones; similarly by replicating each segment we can achieve the effect of playing at half the speed (two times slower); this is further illustrated in Fig.4. In fact, we can speed up or slowdown at different rates by stochastically copying or deleting

[2]The Keele data has utterances spoken by both male and female speakers and includes a reference f_0 estimate at a resolution of 10ms.

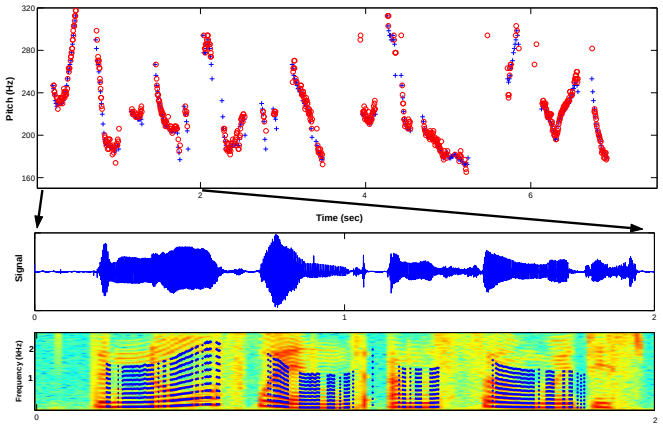

Fig. 3. *Top:*Pitch estimates using our approach for a female speaker in the Keele dataset. Notice that the inferred pitch (circle) consistently agrees with the reference provided (plus mark). *Middle:* section of the input time domain signal. *Bottom:* Spectrogram is marked at the integer multiples of fundamental frequency.

Table 1. Pitch Tracking Error rates (Keele dataset)

Error type	Seg HMM	MLS [3][7]	MLS+ [7]	get_f0 [4]	YIN [8]
Unvoiced(%)	8.89	8.60	7.90	6.83	-
Voiced(%)	8.49	8.87	7.03	3.24	-
Gross(%)	2.28	1.68	1.5	2.29	3.28
rms(Hz)	4.48	4.68	4.54	4.5	3.62

segments at appropriate intervals. By doing all of our operations directly in the time domain we never need to worry about inconsistent phase estimates.

4. FUTURE WORK AND CONCLUSION

In summary, we have presented a simple segmental Hidden Markov Model for analyzing speech waveforms directly in the time domain and derived an efficient algorithm for MAP inference in this model. The proposed method 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).

Fig. 5 demonstrates a preliminary experiment on cleaning severely corrupted signals. This suggests that our method holds promise for challenging problems in speech restoration, such as clipped speech restoration, denoising, and filling in missing regions of speech. For example, denoising involves modeling the clean signal $\boldsymbol{x}$ as a latent variable and associating it to the noisy observation $\boldsymbol{y}$ using a Gaussian likelihood function. In this generative model, clean segments are modeled as before and the observations are modeled as a noisy copy of the corresponding clean segment. Inference in this model amounts to simultaneously performing denoising and segmentation on the denoised waveform.

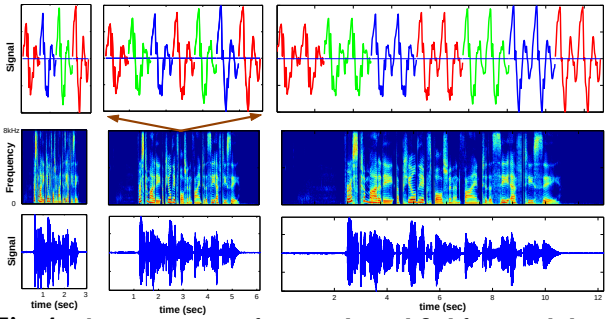

Fig. 4. The spectrogram of time scale modified faster and slower versions of a signal are shown. The actual time domain operation is shown on top for a particular time instant in the spectrogram.

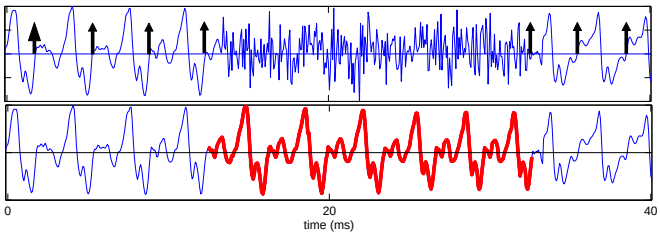

Fig. 5. *Top:* A signal severely corrupted by noise in part of a voiced region. Estimated glottal pulse boundaries are marked by vertical arrows. Our segmentation algorithm treats the corrupted region as unvoiced. *Bottom:* Reconstructed signal. We filled in the corrupted region by generating new segments with periods between the two bounding voiced regions. The scale factor for the filled-in regions was computed by matching the two bounding segments and interpolating.

Given an estimate for the clean signal, we use the inference algorithm detailed earlier to perform segmentation and given a segmentation, we update $\boldsymbol{x}$ to maximize the likelihood $p(\boldsymbol{y}|\boldsymbol{x})$. Further, this optimization is done segment-wise, due to the Markovian nature of the factorization. This alternating optimization is performed until the estimate for the signal stabilizes.

We are also investigating other possible applications with the same model, including voice conversion, volume equalization, and addition/removal of reverberant filtering.

5. REFERENCES

[1] H. Attias, J. C. Platt, A. Acero, and L. Deng, "Speech denoising and dereverberation using probabilistic models," in *Advances in Neural Information Processing Systems 13*. MIT Press, Cambridge MA., 2001.

[2] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Communication*, vol. 9, pp. 453–467, 1990.

[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.," in *Advances in Neural Information Processing Systems 15*. MIT Press, MA, 2003.

[4] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," in *Speech coding and Synthesis*, W.B.Klejin and K.K.Paliwal, Eds. Elseiver Science, 1995.

[5] K. Achan, S. T. Roweis, A. Hertzmann, and B. J. Frey, "A segmental HMM for speech waveforms," University of Toronto Technical Report UTML-TR-2004-001, 2004 (revised).

[6] F. Plante, W. A. Ainsworth, and G. F. Meyer, "A pitch extraction reference database," in *Eurospeech*, 1995.

[7] F. Sha, A. Burgoyne, and L. K. Saul, "Multiband statistical learning for f_0 estimation in speeh," in *ICASSP*. 2004, IEEE.

[8] A. de Cheveigné, "Yin, a fundamental frequency estimator for speech and music," *Journal of the Acoustical Society of America*, vol. 111, 2002.