A LINKED-HMM MODEL FOR ROBUST VOICING AND SPEECH DETECTION

Sumit Basu

Microsoft Research
sumitb@microsoft.com

ABSTRACT

We present a novel method for simultaneous voicing and speech detection based on a linked-HMM architecture, with robust features that are independent of the signal energy. Because this approach models the change in *dynamics* between speech and non-speech regions, it is robust to low sampling rates, significant levels of additive noise, and large distances from the microphone. We demonstrate the performance of our method in a variety of testing conditions and also compare it to other methods reported in the literature.

1. INTRODUCTION

As we move towards advanced applications in ubiquitous environments, speech detection becomes an increasingly important problem. Being able to detect the presence of speech and its boundaries is a key capability of an interactive system: unless it knows when to listen, it will attempt to decode environmental sounds. Furthermore, finding the voiced segments within the speech can give us a variety of additional information, such as speaking rate and the boundaries of pitch contours.

Our goal is to robustly identify the voiced and unvoiced regions, as well as to group them into chunks of speech to separate them from non-speech regions. Furthermore, we want to do this in a way that is robust to low sampling rates, far-field microphones, and ambient noise. Clearly, to work in such broad conditions, we cannot depend on the spectral visibility of the unvoiced regions. There has been a variety of work on trying to find the boundaries of speech, a task known in the speech community as "endpoint detection." Most of the earlier work on this topic has been very simplistic as the speech recognition community tends to depend on a close-talking, noise-free microphone situation. More recently, there has been some interest in robustness to noise, due to the advent of cellular phones and hands-free headsets. For instance, there is the work of Junqua et al. [1] which presents a number of adaptive energy-based techniques, the work of Huang and Yang [2], which uses a spectral entropy measure to pick out voiced regions, and later the work of Wu and Lin [3], which extends the work of Junqua et. al by looking at multiple bands and using a neural network to learn the appropriate thresholds. Recently, there is also the work of Ahmadi and Spanias [4], in which a combination of energy and cepstral peaks are used to identify voiced frames, the noisy results of which are smoothed with median filtering. The basic approach of these methods is to find features for the detection of voiced segments (i.e., vowels) and then to group them together into utterances. We found this compelling, but noted that many of the features suggested by the authors above could be easily fooled by environmental noises, especially those depending on energy.

We thus set out to develop a new method for voicing and speech detection which was different from the previous work in two ways. First, we wanted to make our low-level features independent of energy, in order to be truly robust to different microphone and noise conditions. Second, we wished to take advantage of the *multi-scale dynamics* of the voiced and unvoiced segments. Looking again at the spectrograms, there is a clear pattern that distinguishes the speech regions from silence. It is not in the low-level features, certainly – the unvoiced regions often look precisely like the silence regions. In speech regions, though, we see that voicing state is transitioning rapidly between voiced (state value 1) and unvoiced/silence (state value 0), whereas in the non-speech regions, the signal simply stays in the unvoiced state. The dynamics of the transitions, then, are different for the speech and non-speech regions. In probabilistic terms, we can represent this as follows:

$$P(V_t = 1|V_{t-1} = 1, S_t = 1) \neq P(V_t = 1|V_{t-1} = 1, S_t = 0) \quad (1)$$

This is clearly more than the simple HMM can model, for in it the current state can depend only on the previous state, not on an additional parent as well. We must turn instead to the more general world of dynamic Bayesian nets and use the "linked HMM" model proposed by Saul and Jordan [5]. The graphical model for the linked HMM is shown in figure 1. The lowest level states are the continous observations from our features, the next level up (V_t) are the voicing states,

This work was done while the author was at the MIT Media Laboratory

and the highest level (S_t) are the speech states.

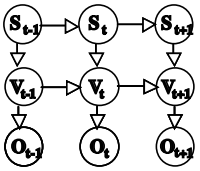

Fig. 1. Graphical model for the linked HMM of Saul and Jordan.

This model gives us precisely the dependencies we needed from equation 1. Note that as in the simple HMM, excepting the initial timestep 0, all of the states in each layer have tied parameters. In a rough sense, the states of the lower level, V_t, can then model the voicing state like a simple HMM, while the value of the higher level S_t will change the transition matrices used by that HMM. This is in fact the same model used by the vision community for modeling multi-level dynamics, there referred to as switching linear dynamics systems (as in [6]). Our case is nominally different in that both hidden layers are discrete, but the philosophy is the same. If we can afford exact inference on this model, this can be very powerful indeed: if there are some places where the low-level observations $P(O_t|V_t)$ give good evidence for voicing, the higher level state will be biased towards being in a speech state. Since the speech state will have much slower dynamics than the voicing state, this will in turn bias other nearby frames to be seen as voiced, as the probability of voicing under the speech state will be much higher than in the non-speech state. We will see this phenomenon later on in the results.

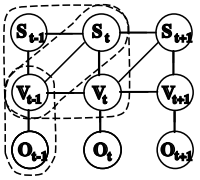

Fig. 2. The clique structure for the moralized graph for the linked HMM.

In our case, exact inference and thus learning are fairly simple applications of the Junction Tree Algorithm [7], as both of our sets of hidden states are discrete and binary. The clique structure for the moralized, triangulated graph of the model is shown in figure 2. The maximal clique size is three, with all binary states, so the maximum table size is $2^3 = 8$. This is quite tractable, even though we will have an average of two of these cliques per timestep. Doing inference on the resulting junction tree is analagous to the forward-backward algorithm HMM except for an additional clique for each timestep. Thus we still only need a single forward and backward pass to do exact inference. Overall, we end up with 36 operations per timestep vs. 24 for a single-layer binary HMM.

2. FEATURES

We are using three features for the observations: the non-initial maximum of the normalized "noisy" autocorrelation, the number of autocorrelation peaks, and the normalized spectral entropy. These are all computed on a per-frame basis – in our case, we are always working with 8 kHz speech, with a framesize of 256 samples (32 milliseconds) and an overlap of 128 samples (16 milliseconds) between frames.

2.1. Noisy Autocorrelation

The standard short-time normalized autocorrelation of the signal $s[n]$ of length N is defined as follows:

$$a[k] = \frac{\sum_{n=k}^{N} s[n]s[n-k]}{(\sum_{n=0}^{N-k} s[n]^2)^{\frac{1}{2}}(\sum_{n=k}^{N} s[n]^2)^{\frac{1}{2}}} \quad (2)$$

We define the set of autocorrelation peaks as the set of points greater than zero that are the maxima between the nearest zero-crossings, discounting the initial peak at zero ($a[0]$ is guaranteed to be 1 by the definition). Given this definition, we see a small number of strong peaks for voiced frames because of their periodic component. Unvoiced frames, on the other hand, are more random in nature, and thus result in a large number of small peaks. We thus use both the maximum peak value and the number of peaks as our first two features.

There is one significant problem to the standard normalized autocorrelation, though – very small-valued and noisy periodic signals will still result in strong peaks. We could deal with this by simply cutting out frames that were below a certain energy, but this would make us very sensitive to the energy of the signal. We instead devised a much softer solution, which is to add a very low-power Gaussian noise signal to each frame before taking the autocorrelation. In the regions of the signal that have a strong periodic component, this has practically no effect on the autocorrelation. In lower power regions, though, it greatly disrupts the structure of a low-power, periodic noise source. To estimate the amount of noise to use, we use a two-pass approach – we first run the linked-HMM to get a rough segmentation of voicing and use the resulting non-speech regions to estimate the signal variance during silence. We then add a Gaussian noise signal of this variance to the entire signal and run the segmentation again.

2.2. Spectral Entropy

Another key feature distinguishing voiced frames from unvoiced is the nature of the FFT magnitudes. Voiced frames

have a series of very strong peaks resulting from the pitch period's Fourier transform $P[w]$ multiplying the spectral envelope $V[w]$. This results in the banded regions we have seen in the spectrograms and in a highly structured set of peaks in the FFT. In unvoiced frames, on the other hand, we see a fairly noisy spectrum, be it silence (with low magnitudes) or a plosive sound (higher magnitudes). We thus expect the entropy of a distribution taking this form to be relatively high. This leads us the notion of spectral entropy, as introduced by Huang and Yang [2].

We take this concept one step further and compute the *relative* spectral entropy with respect to the mean spectrum. This can be very useful in situations where there is a constant voicing source, such as a loud fan or a wind blowing across a microphone aperture. The relative spectral entropy is simply the KL divergence between the current spectrum and the local mean spectrum, computed over the neighboring 500 frames where $m[w]$ is the mean spectrum:

$$H_r = -\sum_w p[w] \log \frac{p[w]}{m[w]}, \tag{3}$$

3. TRAINING

With our features selected, we are now ready to parametrize and train the model. We choose to model the observations with single Gaussians having diagonal covariances. It would be a simple extension to use mixtures of Gaussians here, but since the features appear well separated we expected this would not be necessary. Furthermore, reducing the number of parameters in the model greatly reduces the amount of training data necessary to train the model. We trained the model with the Expectation-Maximization (EM) algorithm [7] using several minutes of speech data from two speakers in the callhome database (8000 frames of 8 kHz, 8-bit mulaw data) with speech and voicing states labeled in each frame. Since all states were labeled, it was only necessary to run EM for one complete iteration.

4. PERFORMANCE

To illustrate the strengths of our two-layer approach, we begin by showing the results of applying an ordinary HMM versus our linked HMM in noisy conditions in figure 3. Notice how our model is able to more reliably find the voicing states. We can understand why by thinking about the flow of information in the inference process: the "strong" voicing states (chunks 1, 3, and 4), which are captured by both models, are feeding information into the upper (speech state) level, biasing it towards a speech state. This then flows back down to the voicing level, since the probability of a voiced state is much higher when the upper level is in a speech state.

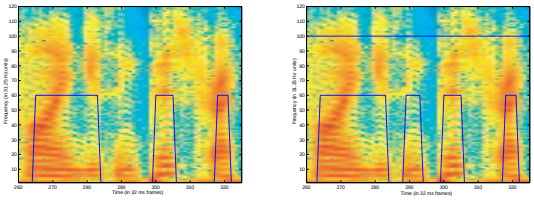

Fig. 3. Comparison of an ordinary HMM (left) versus our linked HMM model (right) on a chunk of noisy data. Notice how our model more reliably finds the voiced segments.

4.1. Robustness to Noise

In this set of experiments, we show the robustness of our algorithm to noise and compare our results with some related work. We will use the measure of *segmental signal-to-noise ratio*, or SSNR, to evaluate the noise condition. We can compute the SSNR for K frames of a zero-mean signal $s[n]$ with added noise $w[n]$ as follows:

$$SSNR = \frac{1}{K} \sum_{i=1}^{K} 20 \log \frac{\sigma^2_{s_i[n]}}{\sigma^2_{w_i[n]}}. \tag{4}$$

The best reported results in the literature for voicing detection in noise are from Spanias and Ahmadi [4]. Their approach was to use the logical "and" of two features: an adaptive threshold test for energy and one for the cepstral peak. The threshold for each feature is chosen as the median of that signal over the entire file. They employ no time dynamics, but use a 5-frame median filter to smooth the results of their detection. As in their work, we hand labeled a small set of speech (2000 frames in our case). Each frame was labeled as being voice/unvoiced and speech/non-speech by both examining the clean spectrogram and listening to the corresponding audio. We then added Gaussian noise of varying power to simulate various noise conditions. As their signals themselves were not available, we implemented their technique so that we could compare its results directly on our data. The results for are shown in figure 4. In the interests of space, we have shown only the total voicing error, which includes voiced frames classified as unvoiced (V-UV) and unvoiced frames classified as voiced (UV-V). Note however that for both algorithms, the UV-V error was an order of magnitude smaller than the V-UV error.

It is interesting to note that the results of the Ahmadi and Spanias method do not worsen monotonically with increasing noise. This is due to their heuristic for choosing feature thresholds – under just the right amount of noise, the thresholds achieve their best values. As a result, adding noise can actually improve the performance by shifting the thresholds in the right way. In addition, because their method requires the energy *and* the cepstral peak to be above a threshold, it tends to clip off the beginning and end of many voiced seg-

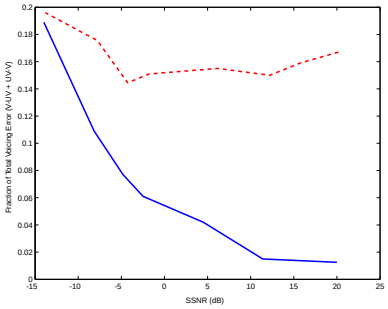

Fig. 4. Comparison of the voicing segmentation error (V-UV + UV-V) in various noise conditions using our method (solid line) against our implementation of the Ahmadi and Spanias algorithm [4] (dashed line).

ments, which tend to be lower energy though still clearly voiced.

Before we leave this experiment, we would like to show one last performance figure – the performance of the speech segmentation with respect to noise (figure 5). The performance is quite robust. Even at -14dB, we are only misclassifying 17% of the frames. As with the voicing segmentation, this error is almost entirely made up of S-US errors.

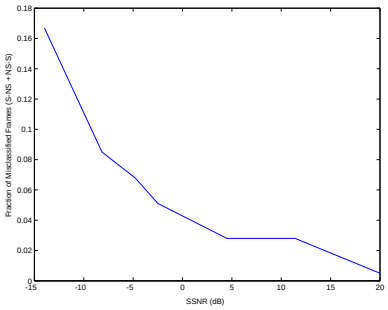

Fig. 5. Speech Segmentation Error (S-NS + NS-S) in various noise conditions vs. SSNR.

4.2. Robustness to Microphone Distance

Another important goal for our method is to be robust to microphone distance. In any real environment, distance adds more than Gaussian noise – now the sounds of fans, doors, chairs and the like all become comparable in power to the speech signal as the distance increases. We tested this condition by putting a far-field condenser microphone (an AKG C1000s) on a table in an office environment, then moving successively further away from the mic. The total voicing error and speech error for this experiment are shown in figure 6.

We estimate the SSNR of the signal at 24 feet to be -18 dB. However, since the noise no longer has a white spectrum, it is potentially more difficult to contend with. However, our method is still robust to this difficult condition. By 21 feet (about -10 dB of SSNR), we still have less than 10% error in both voicing and speech segmentation.

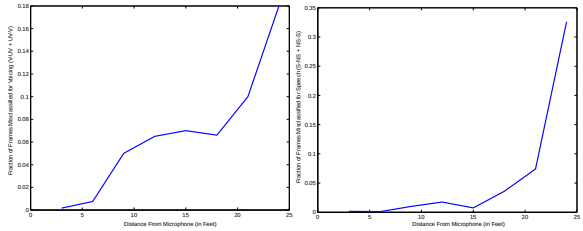

Fig. 6. Performance of the voicing and speech segmentation with distance from mic (in feet). The first plot shows the total voicing error; the second shows the total speech error.

5. REFERENCES

[1] Jean-Claude Junqua, Brian Mak, and Ben Reaves, "A robust algorithm for word boundary detection in the presence of noise," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 3, pp. 406–412, 1994.

[2] Liang-sheng Huang and Chung-ho Yang, "A novel approach to robust speech endpoint detection in car environments," in *Proceedings of ICASSP'00*, pp. 1751–1754. IEEE Signal Processing Society, 2000.

[3] Gin-Der Wu and Chin-Teng Lin, "Word boundary detection with mel-scale frequency bank in noisy environment," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 5, pp. 541–553, 2000.

[4] Sassan Ahmadi and Andreas S. Spanias, "Cepstrum-based pitch detection using a new statistical v/uv classification algorithm (correspondence)," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 3, pp. 333–338, 1999.

[5] L. K. Saul and M. I. Jordan, "Boltzmann chains and hidden markov models," in *Neural Information Processing Systems 7 (NIPS 7)*, 1995.

[6] Vladimir Pavlovic, James Rehg, and John MacCormick, "Learning switching linear models of human motion," in *Neural Information Processing Systems (NIPS)*, 2000.

[7] Michael Jordan and Chris Bishop, *An Introduction to Graphical Models*, 2002 (in press).