ROBUST DETECTION OF VOICED SEGMENTS IN SAMPLES OF EVERYDAY CONVERSATIONS USING UNSUPERVISED HMMS

Meysam Asgari, Izhak Shafran and Alireza Bayestehtash

ABSTRACT

We investigate methods for detecting voiced segments in everyday conversations from ambient recordings. Such recordings contain high diversity of background noise, making it difficult or infeasible to collect representative labelled samples for estimating noise-specific HMM models. The popular utility get-f0 and its derivatives compute normalized cross-correlation for detecting voiced segments, which unfortunately is sensitive to different types of noise. Exploiting the fact that voiced speech is not just periodic but also rich in harmonic, we model voiced segments by adopting harmonic models, which have recently gained considerable attention. In previous work, the parameters of the model were estimated independently for each frame using maximum likelihood criterion. However, since the distribution of harmonic coefficients depend on articulators of speakers, we estimate the model parameters more robustly using a maximum a posteriori criterion. We use the likelihood of voicing, computed from the harmonic model, as an observation probability of an HMM and detect speech using this unsupervised HMM. The one caveat of the harmonic model is that they fail to distinguish speech from other stationary harmonic noise. We rectify this weakness by taking advantage of the non-stationary property of speech. We evaluate our models empirically on a task of detecting speech on a large corpora of everyday speech and demonstrate that these models perform significantly better than standard voice detection algorithm employed in popular tools.

Index Terms — voice detection, speech detection, life log

1. INTRODUCTION

With the widespread deployment of speech interface in smartphones and robots, there is a growing demand for robust algorithms to detect speech amidst a variety of noise conditions. Most previous algorithms reported in literature were developed and evaluated on task with at most a handful of noise types [1, 2]. The performance of these algorithms cannot be easily extrapolated to diverse noise backgrounds encountered in everyday life. As a result, there has been a resurgence of interest in methods based on statistical signal processing.
A simplified version assumes the amplitudes of the harmonic are constant over the duration of a frame, and will be referred to as the harmonic model with constant amplitude (HM-CA). Intuitively, the amplitudes can be viewed as the spectral characteristics of oral cavity (channel). The oral cavity cannot vary arbitrarily across frames from the superposition of small number of basis functions \( \psi_i \) as in Equation 2 [5].

\[
a_h(t) = \sum_{i=1}^{I} \alpha_{i,h} \psi_i(t) \\
b_h(t) = \sum_{i=1}^{I} \beta_{i,h} \psi_i(t)
\]  

(2)

The estimation of the parameters \( m \) is simplified by assuming that the observed frame \( y \) is corrupted by uncorrelated independent identically distributed additive zero-mean Gaussian noise \( n \) with unknown variance \( \sigma_n^2 \), thus \( y = A \ m + n \). Then, the maximum likelihood (ML) estimate is computed.

\[
\hat{m}_{ML} = (A^T A)^{-1} A^T y
\]

(4)

Under the harmonic model, the reconstructed signal is \( \hat{s} = \hat{A} \hat{m} \). So far, we assumed that the pitch \( \omega_0 \) was given. However, in practice, the pitch needs to be estimated. It may be estimated independently [7] or computed within this framework by maximizing the energy of the reconstructed signal.

\[
\hat{\omega}_{ML} = \arg \max_\omega \hat{s}^T \hat{s}
\]

(5)

Once the model parameters, \( \Theta = \{ m, \omega \} \), are estimated, we can readily compute the likelihood of observing voiced (v) and unvoiced (u) frames. The residual noise component is \( n = y - \hat{s} \) which is assumed to be Gaussian.

\[
\log P(y|v) = - (y^T y - \hat{s}^T \hat{s}) + C \\
\log P(y|u) = - y^T y + C
\]

(6)

By comparing the likelihoods, a frame is classified as voiced or unvoiced. The constant factor \( C \) effects both likelihoods equally and is dropped.
the same speaker. The ML estimate ignores these articulatory constraints and can potentially over-fit the data. The maximum a posteriori estimate of the model parameters is factored.

\[ \hat{\Theta}_{MAP} = \arg\max_{\Theta} p(\Theta|y) = \arg\max_{\Theta} p(y|\Theta)p(\Theta) \]  

(7)

The likelihood of a voiced frame, \( p(y|\Theta) \), is estimated from Equation 6. For simplicity the prior term, \( p(\Theta) \), is factored as \( p(\Theta) = p(m)p(\omega) \) where \( p(\omega) \) is a uniform distribution from 50 to 500Hz. In our data, we observe that the coefficients estimated independently per frame using Equation 4 is approximately Gaussian, as illustrated in Figure 2. Hence, we model the prior \( p(m) \) as multivariate Gaussian distribution \( \mathcal{N}(\mu_m, \Sigma_m) \). Since the likelihood and the prior are Gaussian, we obtain a closed form MAP estimate by differentiating the Equation 7.

\[ \frac{\partial}{\partial m} \log p(\Theta|y) = \frac{2A^T}{\sigma_n^2}(y - A m) + 2\Sigma_m^{-1}(\mu_m - m) \]

The derivative is set to zero and the closed form analytical expression for the MAP estimate can be computed.

\[ \hat{m}_{MAP} = (A^T A + \sigma_n^2 \Sigma_m^{-1})^{-1}(A^T y + \sigma_n^2 \Sigma_m^{-1} \mu_m) \]  

(8)

Note, Bayesian estimate of all the model parameters is significantly more complex and requires expensive numerical approximations compared to our simpler MAP estimate with its closed form analytical solution [5]. The MAP estimate derived in a related previous work smooths the likelihood using a first order HMM transition model and hence differs from our approach [6].

2.3. Detecting Voiced Segments

Now that we can compute probability of observing a voiced or unvoiced frame, the frame-level scores is smoothed to obtain a segment-level decision using a hidden Markov model (HMM), as in [6]. Specifically, this is achieved using an HMM with two states, the voiced (v) and the unvoiced (u) states, whose observation probabilities are modeled using Equation 6. The transition probabilities, probabilities of staying in the same state and transition across states, represented by two parameters can be tuned for a task. With this HMM, the voiced segments of any utterance is computed using a Viterbi search. Unlike HMMs trained on cepstral features for speech recognition tasks, the parameters of the observation probability are estimated for each frame from the observations themselves.

One additional concern that needs to be addressed is the possibility that the background noise (e.g., fan noise) may also be rich in harmonics. We utilize the non-stationary property of speech to distinguish it from stationary harmonic noise. We condition the observation probability \( p(y|v) \) with additional indicator variable \( t \), which is then factored.

\[ p(y|v, t) = p(y|v)p(\delta m|t, v) \]

(9)

Here \( \delta m \) is the difference between the harmonic coefficients of the current frame and its neighbors. This difference will be low for stationary harmonic noise and is modeled as a univariate Gaussian with just two variables. For each utterance, we compute the per-frame estimate of the model parameter and from those estimate the priors distribution \( \mathcal{N}(\mu_m, \Sigma_m) \). Important to note, our algorithm for detecting voiced segments contains only a handful of variables unlike the hundreds or thousands of variables in the standard HMM models with GMM observations of MFCC features. These few variables are tuned to a new task with very few examples.

3. EMPIRICAL EVALUATION

3.1. Keele dataset

We characterize the performance of our harmonic model on a task of detecting voiced frames using controlled experiments on the well-studied Keele dataset [8]. The data set contains 10 audio files from 10 speakers, 5 males and 5 females, along with labels for voicing obtained from a laryngograph. The speech was recorded in noise free conditions and for testing the robustness of our algorithm we introduce additive white Gaussian noise at different SNRs. For evaluation, we exclude frames for which the voicing label in the corpus is uncertain. We compare the performance of our algorithm under both ML and MAP estimation with that of getf0, an algorithm employed in many popular tools (wavesurfer, praat, etc). For completeness, we include the average 10-fold cross-validation error rate using a standard GMM with MFCC features.
features. The GMMs were evaluated in a realistic settings were the noise in the test setup is unavailable for training. They were estimated using clean speech and tested on speech with different levels of noise corruption. The \textit{get-f0} employs hand-tuned preprocessing followed by normalized cross correlation (NCC). For our algorithm, the pitch estimate was performed by searching over the frequency range of 50-500Hz with a resolution of 1Hz. The performance of \textit{openSMILE}, another popular tool, was significantly worse and we decided not to pursue it further. For all the experiments reported below, in Table 1, we chose an operating point that maximized equal error rate (EER) for each SNR condition. The results show that our algorithm performs consistently better than \textit{get-f0} at all SNRs. Further, the MAP version outperforms the ML version.

<table>
<thead>
<tr>
<th>SNR</th>
<th>Sup-GMM</th>
<th>\textit{get-f0}</th>
<th>ML</th>
<th>MAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>clean</td>
<td>3.2</td>
<td>6.5</td>
<td>3.4</td>
<td>3.3</td>
</tr>
<tr>
<td>20 dB</td>
<td>3.3</td>
<td>6.7</td>
<td>3.9</td>
<td>3.7</td>
</tr>
<tr>
<td>15 dB</td>
<td>4.3</td>
<td>7.2</td>
<td>4.1</td>
<td>3.6</td>
</tr>
<tr>
<td>10 dB</td>
<td>4.6</td>
<td>7.1</td>
<td>4.2</td>
<td>3.6</td>
</tr>
<tr>
<td>5 dB</td>
<td>5.3</td>
<td>7.9</td>
<td>4.7</td>
<td>4.5</td>
</tr>
<tr>
<td>0 dB</td>
<td>6.0</td>
<td>8.6</td>
<td>5.3</td>
<td>4.9</td>
</tr>
</tbody>
</table>

Table 1. Comparison of performance (equal error rate) between \textit{get-f0} and the two versions, ML and MAP, of harmonic model on Keele dataset. For completeness, we also include supervised GMMs with MFCC features.

As expected, the error in detecting voiced segments increases with noise irrespective of the method employed. The optimal thresholds for ML and MAP were the same and were found to increase with SNR and was -1.7, -1.2, -0.8, -0.6, and -0.3 for 0 dB, 5 dB, 10 dB, 15 dB and clean conditions respectively. Potentially this could be automated from the estimate of the SNR in the input signal. The detection trade-off is illustrated further in Figure 3, where false accept is the percentage of voiced frames incorrectly classified as unvoiced frames and false alarm is the percentage of unvoiced frames incorrectly identified as voiced frames. The plot for 0dB SNR shows that harmonic model outperforms \textit{get-f0} and the MAP version is better than the ML version.

3.2. Corpus of Everyday Conversations

Next, we evaluate the performance of our algorithm on a large corpus of samples of everyday conversations [9]. These recordings were collected 97 students using a lavaliere microphone and a digital recorder. The recorder was timed to record 30-second clips every 12 minutes during the participant’s day. The recordings have been transcribed by research assistants. We created a corpus of non-speech and mostly speech utterances. Utterances with no reference transcripts were treated as non-speech. We identified utterances with very few silences by estimating the distribution of number of words in the utterances and picking the top (more than 18 words) quintile. This gave us an evaluation set of 4620 recordings containing no words and 1106 utterances containing mostly speech.

The performance of MAP-version of harmonic models, the \textit{get-f0} and the \textit{openSMILE} are shown in the Figure 4. The total number of voiced and unvoiced frames were computed in each utterance and they were classified as speech or non-speech utterances using a threshold. By varying the threshold, we obtained the DET curve. The results show that the MAP-version of our algorithm outperforms both the baseline methods by a substantial margin.

4. CONCLUSIONS

In summary, this paper develops a speech detection algorithm that exploits harmonic rich nature of voiced speech. For this purpose, we adopt the harmonic model with time-varying amplitudes. We show how the parameters of the model can be computed for each frame using maximum \textit{a posteriori} estimate. Thus the likelihood of voicing under the harmonic model is estimated in an unsupervised manner. We use this likelihood as the observation probability of our HMMs for detecting speech. We also overcome the weakness of harmonic model in differentiating speech from stationary harmonic noise using the non-stationary property of speech. An alternative approach that is well-suited for removing stationary harmonic noise is the non-parameteric estimation of background noise using order (minimum) statistics [10]. This is left for future work. We demonstrate the advantage of our
Fig. 4. Comparing performance on detecting utterance with only noise (or speech) using MAP-version of harmonic model and two popular tools, get-f0 and openSMILE.

algorithm on two data sets. On Keele data set, for all levels of additive noise, the equal error rate of our algorithm is substantially better than the popular, hand-tuned tool, get-f0. More importantly, on a large collection of 30-second samples of ambient recordings of everyday life, we show that our algorithm performs significantly better than other popular alternatives.

5. ACKNOWLEDGEMENTS

We are grateful to Matthias Mehl for making his corpus of naturalistic samples of everyday conversations available to us for evaluating our algorithms. This research was supported in part by NIH awards 5K25AG033723, 5R01AG027481, and 5R01AG033581 and NSF awards 0905095 and 1027834. Any opinions, findings, conclusions or recommendations expressed in this publication are those of the authors and do not reflect the views of the NIH or NSF.

6. REFERENCES


