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

Meysam Asgari, Izhak Shafran and Alireza Bayestehtashk

Center for Spoken Language Understanding, OHSU, Portland, OR, USA

{asgari, shafrani, bayesteh}@ohsu.edu

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

Our goal in this work is to develop a robust algorithm to detect speech from background noise in audio recorded, for example, using a lavalier microphone and a digital recorder. Privacy concerns in many applications preclude the possibility of annotating representative noise samples for training or adapting parametric models such as hidden Markov models (HMMs). Moreover, the diversity of background noise (open set) precludes the possibility of collecting and labeling representative samples for every type of noise.

Harmonic model of voiced speech is an approach that has nice theoretical guarantees on residual errors [3] and has recently gained significant attention. It focuses the modeling effort on innate properties of voiced speech – it's rich harmonic structure. We review the model briefly in Section 2 and then describe the estimation of model parameters. In previous work, the model parameters for each frame were estimated using maximum likelihood. In this work, we apply a prior on the model parameters and compute the maximum *a posteriori* estimate. We use the likelihoods of voicing, computed from the harmonic model, for each frame as the observation probability of an HMM. Thus, these HMMs are different from the conventional HMMs in that the observation probabilities are estimated in an unsupervised manner. In Section 3, we empirically evaluate the efficacy of the model in detecting voiced segments under different amounts of additive noise and then measure performance on a large collection of utterances recorded from several speakers over the course of their everyday lives. Finally, we summarize the contribution of this paper and performance of our speech detection algorithm.

2. MODELING HARMONIC-RICH VOICED SPEECH

The most widely used model of voiced speech is the source-channel model where speech is a convolution of a source of periodic glottal pulses and a channel corresponding to the response of the oral cavity. The glottal pulses are approximately sawtooth in shape and contain rich harmonics of its periodic frequency (pitch). Thus, the speech frame $s(t)$ can be modeled using a discrete Fourier series expansions with sinusoidal harmonics of pitch period ω whose amplitudes, $a_h(t)$ and $b_h(t)$, vary slowly. The harmonic model with time-varying amplitude (HM-VA) is represented as shown in Equa-

978-1-4673-5126-3/12/$31.00 ©2012 IEEE

tion 1 [4, 5].

$$s(t) = a_0(t) + \sum_{h=1}^{H} [a_h(t)\cos(h\omega t)] + \sum_{h=1}^{H} [b_h(t)\sin(h\omega t)] \quad (1)$$

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 and is constrained by the smooth and continuous movement of the articulators. The associated coefficients $a_h(t)$ and $b_h(t)$ may be expressed as a smooth function obtained from the superposition of small number of basis functions ψ_i as in Equation 2 [5].

$$\begin{aligned} a_h(t) &= \sum_{i=1}^{I} \alpha_{i,h}\psi_i(t) \\ b_h(t) &= \sum_{i=1}^{I} \beta_{i,h}\psi_i(t) \end{aligned} \quad (2)$$

We represent this smoothness constraints within a frame using four ($I = 4$) Hanning windows as basis functions. For a frame of length M, the windows are centered at 0, $M/3$, $2M/3$, and M. Each basis function is $2M/3$ samples long and has an overlap of $M/3$ with immediate adjacent window. Figure 1 illustrates the amplitude of a harmonic component obtained by combination of four basis functions.

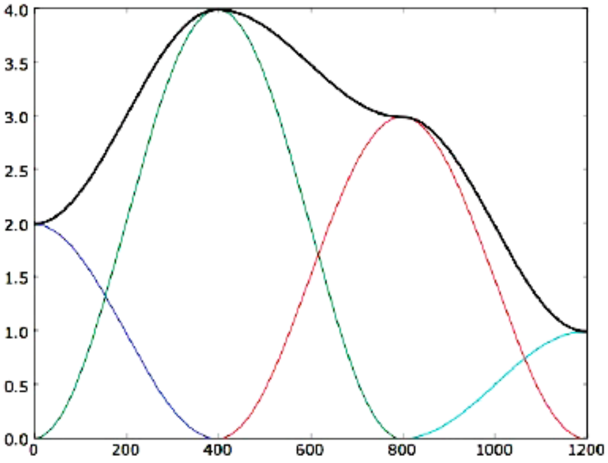

Fig. 1. An illustration of time-varying amplitude of a harmonic component modeled as a superposition of four bases functions spanning the duration of the frame.

2.1. ML Estimation of the Model Parameters

When the pitch is given, the harmonic model with time-varying amplitude is parameterized only by the coefficients, $\alpha_{i,h}$ and $\beta_{i,h}$, which are constant over the duration of the frame. Furthermore, the model is linear in these coefficients. This is obvious when the frame vector s containing T samples is written as a product of a matrix-vector multiplication [6].

$$\begin{aligned} \mathbf{s} &= \mathbf{A}\,\mathbf{m} = [\,\bar{1}\ \ \mathbf{A}_c\ \ \mathbf{A}_s\,]\,[\,a_0\ \ \mathbf{a}\ \ \mathbf{b}\,]^T \quad (3) \\ \mathbf{A}_c &= \begin{bmatrix} \psi_1(1)\cos(\omega) & \cdots & \psi_I(1)\cos(H\omega) \\ \cdots & \cdots & \cdots \\ \psi_1(T)\cos(\omega T) & \cdots & \psi_I(T)\cos(H\omega T) \end{bmatrix} \\ \mathbf{A}_s &= \begin{bmatrix} \psi_1(1)\sin(\omega) & \cdots & \psi_I(1)\sin(H\omega) \\ \cdots & \cdots & \cdots \\ \psi_1(T)\sin(\omega T) & \cdots & \psi_I(T)\sin(H\omega T) \end{bmatrix} \\ \mathbf{a} &= [\alpha_{1,1}\cdots\alpha_{1,H}\cdots\alpha_{I,1}\cdots\alpha_{I,H}] \\ \mathbf{b} &= [\beta_{1,1}\cdots\beta_{1,H}\cdots\beta_{I,1}\cdots\beta_{I,H}] \end{aligned}$$

The coefficients of the cosines are represented as a stacked vector, **a**, consisting of components that span the I basis vectors and the H harmonics. The sinusoidal components, $\mathbf{A}_s$ **b**, are represented likewise. In the above equations, $\bar{1}$ denotes a vector of ones.

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 σ_n^2, thus $\mathbf{y} = \mathbf{A}\ \mathbf{m} + \mathbf{n}$. Then, the maximum likelihood (ML) estimate is computed.

$$\hat{\mathbf{m}}_{ML} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{y} \quad (4)$$

Under the harmonic model, the reconstructed signal is $\hat{\mathbf{s}} = \mathbf{A}\ \mathbf{m}$. So far, we assumed that the pitch ω_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.

$$\hat{\omega}_{ML} = \arg\max_{\omega}\ \hat{\mathbf{s}}^T\hat{\mathbf{s}} \quad (5)$$

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

$$\begin{aligned} \log P(\mathbf{y}|v) &= -(\mathbf{y}^T\mathbf{y} - \hat{\mathbf{s}}^T\hat{\mathbf{s}}) + C \\ \log P(\mathbf{y}|u) &= -\mathbf{y}^T\mathbf{y} + C \end{aligned} \quad (6)$$

By comparing the likelihoods, a frame is classified as voiced or unvoiced. The constant factor C effects both likelihoods equally and is dropped.

2.2. MAP Estimation of the Model Parameters

We propose to improve the robustness of the model by exploiting the fact that model parameters depend on the articulators and hence cannot vary arbitrarily across frames from 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.

$$\hat{\Theta}_{MAP} = \arg\max_{\Theta} p(\Theta|\mathbf{y}) = \arg\max_{\Theta} p(\mathbf{y}|\Theta)p(\Theta) \qquad (7)$$

The likelihood of a voiced frame, $p(\mathbf{y}|\Theta)$, is estimated from Equation 6. For simplicity the prior term, $p(\Theta)$, is factored as $p(\Theta) = p(\mathbf{m})p(\omega)$ 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 $p(\mathbf{m})$ as multivariate Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)$. Since the likelihood and the prior are Gaussian, we obtain a closed form MAP estimate by differentiating the Equation 7.

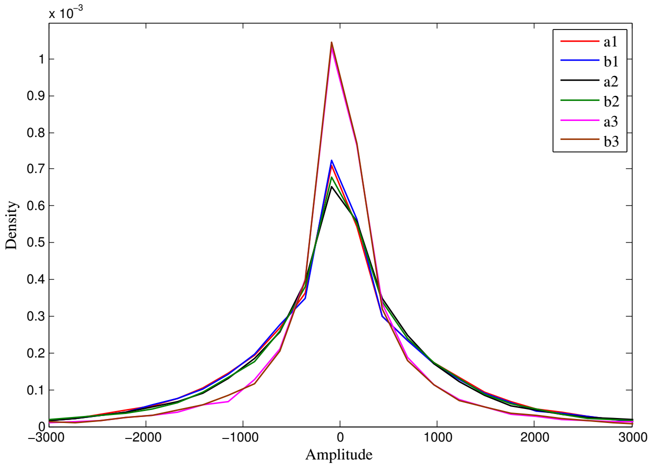

Fig. 2. The empirical distribution of the first three coefficients of harmonic sinusoids and cosines in the voiced frames of Keele dataset.

$$\frac{\partial}{\partial \mathbf{m}} \log p(\Theta|\mathbf{y}) = \frac{2\mathbf{A}^T}{\sigma_n^2}(\mathbf{y} - \mathbf{A}\,\mathbf{m}) + 2\boldsymbol{\Sigma}_m^{-1}(\boldsymbol{\mu}_m - \mathbf{m})$$

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

$$\hat{\mathbf{m}}_{MAP} = (\mathbf{A}^T\mathbf{A} + \sigma_n^2\boldsymbol{\Sigma}_m^{-1})^{-1}(\mathbf{A}^T\mathbf{y} + \sigma_n^2\boldsymbol{\Sigma}_m^{-1}\boldsymbol{\mu}_m) \qquad (8)$$

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 $p(\mathbf{y}|v)$ with additional indicator variable t, which is then factored.

$$p(\mathbf{y}|v,t) = p(\mathbf{y}|v)p(\delta\mathbf{m}|t,v) \qquad (9)$$

Here $\delta\mathbf{m}$ 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 $\mathcal{N}(\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)$. 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 *get-f0*, 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. 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 performance of the GMM system with spherical covariance was higher than that with diagonal, full or tied covariance. The *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 *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 *get-f0* at all SNRs. Further, the MAP version outperforms the ML version.

SNR	Sup-GMM	*get-f0*	ML	MAP
clean	**3.2**	6.5	3.4	3.3
20 dB	**3.3**	6.7	3.9	3.7
15 dB	4.3	7.2	4.1	**3.6**
10 dB	4.6	7.1	4.2	**3.6**
5 dB	5.3	7.9	4.7	**4.5**
0 dB	6.0	8.6	5.3	**4.9**

Table 1. Comparison of performance (equal error rate) between *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 *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.

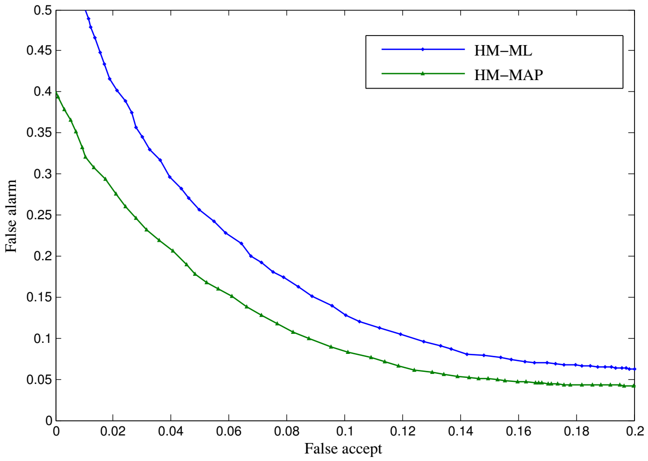

Fig. 3. DET curve for detecting voiced frames with ML and MAP version of harmonic models in Keele dataset with additive noise at 0dB SNR.

The performance of MAP-version of harmonic models, the *get-f0* and the *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 *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

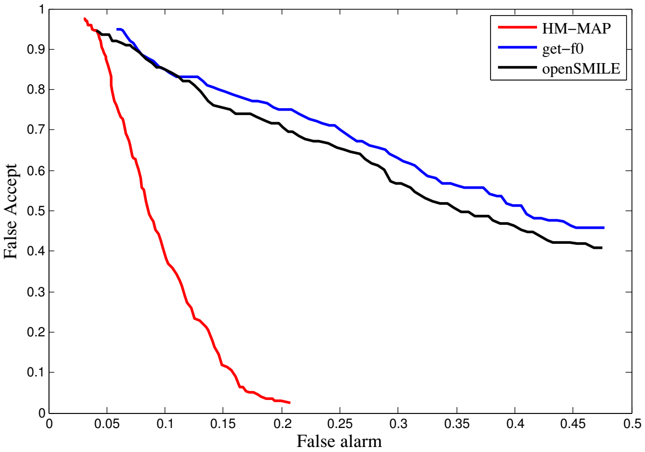

Fig. 4. Comparing performance on detecting utterance with only noise (or speech) using MAP-version of harmonic model and two popular tools, *get-f0* and *openSMILE*.

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

[1] David Pearce and Hans-Gnter Hirsch, “The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions.,” in *INTERSPEECH*. 2000, pp. 29–32, ISCA.

[2] Thomas H. Crystal, Astrid Schmidt-Nielsen, and Elaine Marsh, “Speech in noisy environments (SPINE) adds new dimension to speech recognition R&D,” in *Proceedings of the second international conference on Human Language Technology Research (HLT)*, 2002, pp. 212–216.

[3] E. Fisher, J. Tabrikian, and S. Dubnov, “Generalized likelihood ratio test for voiced-unvoiced decision in noisy speech using the harmonic model,” *IEEE Trans. Audio, Speech Lang. Process.*, vol. 14, no. 2, pp. 502–510, 2004.

[4] Y. Stylianou, “Harmonic plus noise models for speech, combined with statistical methods, for speech and speaker modification,” in *Ph.D. dissertation, Ecole Nationale des Tlcomminications*, 1996.

[5] S. Godsill and M. Davy, “Bayesian harmonic models for musical pitch estimation and analysis,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2002, vol. 2, pp. 1769–72.

[6] J. Tabrikian, S. Dubnov, and Y. Dickalov, “Maximum a-posteriori probability pitch tracking in noisy environments using harmonic model,” *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 1, pp. 76 – 87, 2004.

[7] J. Droppo and A. Acero, “Maximum a posteriori pitch tracking,” in *Proc. ICSLP*, 1998, p. 943946.

[8] F. Plante, Georg F. Meyer, and William A. Ainsworth, “A pitch extraction reference database,” in *Proc. EUROSPEECH*, 1995, pp. 837–840.

[9] Matthias R Mehl, Samuel D Gosling, and James W Pennebaker, “Personality in its natural habitat: manifestations and implicit folk theories of personality in daily life,” *Journal of Personality and Social Psychology*, vol. 90, no. 5, pp. 862–877, 2006.

[10] Izhak Shafran and Richard Rose, “Robust speech detection and segmentation for real-time ASR applications,” in *Proc. ICASSP*, 2003, vol. 1, pp. 432–445.