VOICE ACTIVITY DETECTION IN MOBILE APPLICATIONS

Nezar Assawiel, Daniel Di Matteo

The Edward S. Rogers Sr. Department of Electrical and Computer Engineering,
University of Toronto
Toronto, Ontario, Canada
nezar.assawiel@mail.utoronto.ca, dandm@ece.utoronto.ca

ABSTRACT

Motivated by the desire to accurately perform Voice Activity Detection (VAD) in the noisy environments encountered by users of mobile devices, this work applies biologically-inspired models of human auditory filtering and hearing to a statistical derivation of the VAD technique, applying VAD to a newly-created dataset representative of realistic, everyday scenarios. Two different models, with varying degrees of complexity, are used as a front-end to a statistical voice-detection backend, and accuracies are compared. The more biologically-accurate system achieves an accuracy of 67%, outperforming the simpler model.

1. INTRODUCTION

Voice activity detection (VAD) systems, which detect the presence or absence of speech in an audio stream, are a mature technology used in a variety of applications. Common uses are in VoIP and teleconferencing systems, where they are used to determine when to send packets over the network, as audio not containing speech does not need to be transmitted. In low-noise environments, such as business meetings and conferences, VAD systems have very high levels of accuracy, but this accuracy is degraded by the presence of environmental noise.

The ability to accurately apply VAD in a high-noise environment can be of value, however. This work is motivated by an application to psychiatry, in which patients undergoing treatment have data from their mobile devices mined, processed, and presented to their doctors to provide supplementary information to aid in diagnosis and treatment. Consider, for example, a patient undergoing treatment for social anxiety disorder, where it is known that patients who fail to respond to treatment will continue to avoid social interaction. If the presence of speech can be used as a proxy for social interaction, an objective measure of time spent in the presence of speech can provide clinicians with useful information pertinent to the patient, such as response to treatment.

This work takes steps towards the development of a VAD system which has enough noise resilience to process audio recorded from a single microphone, in noisy environments, and where the microphone may be muffled (e.g., the device is within a pocket). The novelty of this work is that it uses biologically-inspired models for speech and human hearing and applies them to VAD in a mobile setting. Existing work that has explored biologically-inspired algorithms has typically been validated against synthetic benchmarks (with noise/sounds artificially added to otherwise clean speech), and not against audio representative of the more realistic settings.

The rest of the paper is structured as follows. Section 2 gives a description of the VAD system we have adapted from [1]. Section 3 describes the experimental setup we have used to quantify the accuracy of the system and also presents those results. Finally we conclude the report in Section 4.

2. SYSTEM ARCHITECTURE

The system can be divided into two parts: a front-end that performs digital filtering and processing of the incoming audio in a manner inspired by the operation of the human hearing system, producing a representation known as a cochleagram, and a back end that takes as input the time-series data of the cochleagram and classifies it as speech or non-speech.

2.1. Cochleagram Front-End

The cochleagram is a time-frequency representation of audio, similar to a spectrogram [2]. It depicts the distribution of spectral components of audio by computing the firing rates of auditory nerve cells across the basilar membrane of the ear. Since the basilar membrane is naturally tuned to different frequencies along its length, the spatial distribution of nerve cells along its length produces patterns of activities that are essentially tuned to different frequency bands, just as a filterbank can be used to compute a spectrogram.

The cochleagram in this work was computed as follows. Digitized audio, in PCM format, is passed through a bank of gammatone filters. Each gammatone filter models the resonance of a particular location of the basilar membrane, where the impulse response of the gammatone filter is as follows:

$$g(t) = t^{N-1} e^{-2\pi bt} \cos(2\pi f_c t + \phi) u(t)$$

where N is the order of the filter, b is the bandwidth of the filter, f_c is the center frequency, ϕ is the phase offset, and $u(t)$ is the unit step function. A fourth-order gammatone filter was shown to match experimentally-derived data for the auditory filters in human hearing [3]. The bandwidth of the gammatone filter is set equal to empirically-derived values for the human hearing critical-bands at that particular center frequency [4].

Multiple gammatone filters (this number is parameterizable) were used to form a filterbank over a parameterizable range, where the center frequencies of the filters were chosen such that they are evenly spaced on the ERB-rate scale [4]. Spacing the center frequencies in such a manner ensures that each filter in the bank has roughly an equivalent bandwidth to human auditory filters.

Finally, the output of each gammatone filter was fed through a Meddis hair-cell model to capture how the inner ear cells transduce vibrations in the basilar membrane into action potentials [5]. The model used does not produce individual action potentials/spikes, but instead computes the firing rate that would be observed. A simpler cochleagram model was also produced, which instead of using the Meddis hair cell model for transduction simply takes the cube-root of the absolute value of the output of each gammatone filter. This simpler model captures how auditory nerve cells both perform rectification and loudness/amplitude compression of the signal produced by the basilar membrane (if one imagines the vibration of the basilar membrane as a signal).

Finally, for both versions of the cochleagram, the frame rate of the cochleagram was decimated by a factor of x (by keeping only every x-th frame) to ease the computational load in the back-end. This is commonly applied in speech-processing algorithms without a loss in quality as long as the value of x is reasonable.

2.2. Statistical VAD Back-End

A statistical approach is used to perform the voice activity detection, in the form of a hypothesis test, with hypothesis H_1 representing speech is present in the frame, and hypothesis H_0 representing speech is absent from the frame. Both the speech and non-speech cochleagram frames are modeled as N-dimensional Gaussian random processes (where N is the number of channels cochleagram/filterbank).

Three important assumptions are made here. Firstly, that the dimensions are independent of one another, which is not true but a reasonable approximation since the filters in the filterbank are tuned to separate center frequencies with not much spectral overlap. Secondly, that frames of the cochleagram are independent from one another in time - this also not true but done to simplify the analysis. Finally, that the means of these random variables are zero. This appears to be inherited from a Gaussian statistical model applied to the same problem formulated for the use of Fourier expansion coefficients of audio (i.e., the spectrogram) [6].

Using this statistical formulation, the likelihood functions are the following:

$$p(G|H_0) = \prod_{c=1}^{N} \frac{1}{\pi \lambda_N(c)} \exp\left(-\frac{G_c^2}{\lambda_n(c)}\right)$$

$$p(G|H_1) = \prod_{c=1}^{N} \frac{1}{\pi [\lambda_N(c) + \lambda_S(c)]} \exp\left(-\frac{G_c^2}{\lambda_N(c) + \lambda_S(c)}\right)$$

where G is the entire frame (all channels) of the cochleagram, N is the number of channels in the cochleagram, G_c is the value of the cochleagram in filter channel c, and $\lambda_N(c)$ and $\lambda_S(c)$ are the noise and speech variances in channel c, respectively. Note that λ_N and λ_S are the *a-priori* variances, and therefore need to be trained or otherwise estimated from a sample of the population of noise and speech audio.

For each frame of the cochleagram, the log-likelihood ratio is computed as follows:

$$\log \Lambda = \frac{1}{N} \sum_{c=1}^{N} \log\left(\frac{p(G|H_1)}{p(G|H_0)}\right)$$

A likelihood ratio test is then performed to classify the frame as containing speech or non-speech by applying a threshold θ to the log-likelihood ratio as follows:

$$\text{frame} = \begin{cases} \text{speech}, & \log \Lambda > \theta \\ \text{non-speech}, & \log \Lambda \leq \theta \end{cases}$$

It is worth noting that a simple enhancement was applied here. After computing the log-likelihood ratios of every frame (but before classification via application of a threshold), these values were smoothed by applying an n-point moving average as in [1]. This is motivated by the fact that likelihood ratio values will likely jump rapidly relative to their neighboring points, yet in reality the audio is not rapidly transitioning between speech and non-speech classes at a rate of milliseconds. This helps to address the loss in accuracy introduced by the second assumption made in the modeling (i.e., that the frames are independent in time).

3. EVALUATION AND RESULTS

3.1. Experimental Setup

The system in Section 2 was implemented using MATLAB and was used to perform VAD upon a newly-made corpus of audio. The following two subsections describe the audio that was recorded for use as a dataset and the various values of parameters and settings used in the implementation of the system.

3.1.1. Dataset

Audio was recording using a Nexus 5 smartphone in the following format: WAV, 16 bit (signed) samples, little-endian, 16kHz sample rate. Three hours of audio was recorded, with roughly 47% of that containing speech. After recording, the audio was hand-transcribed to identify which times contained speech. Table 1 contains a description of the recorded audio used in the dataset. Note that recordings 5, 7, and 8 did not contain speech, while all others contained a mix of both speech and non-speech.

Table 1. Description of audio in the dataset.

No.	Length (mm:ss)	Setting	Phone location
1	24:31	Meeting	On a table
2	23:50	Meeting	On a table
3	27:30	Meeting	On a table
4	11:21	Meeting	On a table
5	11:52	Walking outside	In a jacket
6	21:56	Meeting	In pants pocket
7	14:45	In office	On desk
8	48:32	In office	On desk

Due to the heavy computational load of this processing and the inefficiency of MATLAB, it was difficult to process more than any one single file using a significant number of channels. Therefore, the results that follow will be based on the most difficult to classify recording that contained a mix of speech and non-speech - recording 6. The first half (roughly 12 minutes) of recording 6 was partitioned into training data and the second half into testing data.

3.1.2. Parameters and Settings

The following parameters and settings were used in the implementation of the VAD system and in all following experiments:

- Gammatone filterbank: N fourth-order filters evenly spaced on the ERB-rate scale from 20Hz to 5000Hz were used (where the value of N will be swept in the following experiments).
- Meddis hair-cell model: Parameters for simulating medium spontaneous-rate fibers from [5] were used.
- Frame rate of the cochleagram: The frame rate was decimated by a factor of 160 to yield 100 frames of the cochleagram per second.
- Statistical model parameters: Estimates of the noise and speech variances were produced using the Maximum Likelihood Estimate of the cochleagram frames produced from the training data.
- Thresholds for likelihood-ratio testing: 100 threshold values were swept, linearly spaced from the minimum log-likelihood ratio value of any test frame encountered to the maximum.
- Log-likelihood ratio smoothing: Values were smoothed by applying a moving average over an 11-frame window (0.1 second window).

3.2. Experiment 1: VAD using Meddis-based cochleagram

In this experiment, the cochleagram was computed using the Meddis hair cell model. The number of filterbank channels was swept in powers of two, from 4 to 32. VAD was performed upon each frame of the test data, and using the true classes of each frame from the transcripts as reference (i.e., speech or non-speech), true positives and false positives were counted. Figure 1 shows the family of ROC curves for this experiment.

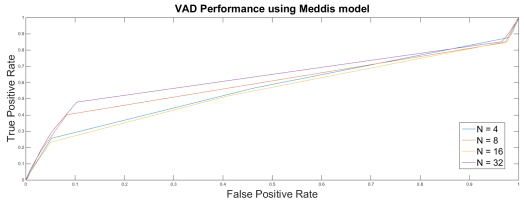

Fig. 1. ROC curve for VAD performed using a Meddis hair cell-based cochleagram.

The maximum accuracy achieved by this system is presented in Table 2, where accuracy is defined here as the sum true positives and true negatives divided by the total number of observances (i.e., frames).

3.3. Experiment 2: VAD using simple Cochleagram

This second experiment was designed to assess what the impact of using a simpler, non-biologically inspired model would be. To do so, the Meddis hair cell model for transduction was replaced with the simpler absolute-value and

Table 2. VAD Accuracy using the Meddis hair cell-based cochleagram.

No. of Channels	Accuracy
4	57%
8	64%
16	56%
32	67%

cube root method described in Section 2. As in the last experiment, the number of filterbank channels was swept in powers of two, from 4 to 32.

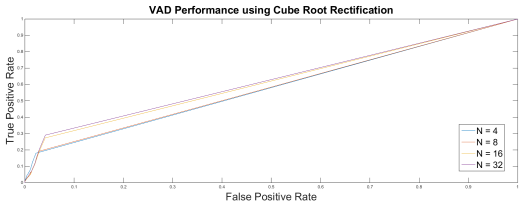

Fig. 2. ROC curve for VAD performed using cube root-based cochleagram.

Table 3. VAD Accuracy using the Meddis hair cell-based cochleagram.

No. of Channels	Accuracy
4	55%
8	55%
16	59%
32	60%

The maximum accuracy achieved by this system is presented in Table 3. It is clear that this system performs much worse that the system which uses the Meddis hair cell model in the computation of the cochleagram. One possible reason for this is due to the fact that the cube root-based model of transduction only captures the rectification and amplitude compression capabilities of auditory nerve cells, but not the other properties. Properties such as adaptation might be particularly helpful in VAD since adaptation serves to greatly accentuate spikes in neural activities when "new" signals are present, and the sudden onset of a voice might be one such cause of a spike.

4. CONCLUSION AND FUTURE WORK

Motivated by the desire to accurately perform VAD in the noisy environments encountered by users of mobile devices, this work applied biologically-inspired models of human auditory filtering and hearing to statistical derivation of the VAD technique. It was shown that the more biologically-accurate computation of the cochleagram, performed using the Meddis hair cell model, outperforms more simple cochleagram model based upon a cube-root approximation.

A more thorough analysis on a more significant dataset is the first improvement that the authors would like to pursue. Due to the computational limits of MATLAB, a more efficient, multithreaded and natively-coded prototype could enable such work. Specifically, it would be interesting to explore how the performance of the two systems compare when using a large number of filter channels. Furthermore, the application of the system in a wide range of environments, which was a guiding motivation for this work, was not fully explored. It would be interesting to see how the estimates of the statistical parameters used in the back end generalize to a broad range of noise sources. Lack of performance here might motivate the use of a more complex statistical model.

5. REFERENCES

[1] M. Tu, X. Xie, and X. Na, "Computational auditory scene analysis based voice activity detection," in *Pattern Recognition (ICPR), 2014 22nd International Conference on*, Aug 2014, pp. 797–802.

[2] D. Wang and G. Brown, *Fundamentals of Computational Auditory Scene Analysis*. Wiley-IEEE Press, 2006, pp. 1–44.

[3] R. Patterson, I. Nimmo-Smith, J. Holdsworth, and P. Rice, "An efficient auditory filterbank based on the gammatone function," MRC Applied Psychology Unit, Cambridge, Tech. Rep., 1987.

[4] B. R. Glasberg and B. C. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing Research*, vol. 47, no. 1, pp. 103 – 138, 1990. [Online]. Available: http://www.sciencedirect.com/science/article/pii/037859559090170T

[5] R. Meddis, M. J. Hewitt, and T. M. Shackleton, "Implementation details of a computation model of the inner hair cell auditory nerve synapse," *The Journal of the Acoustical Society of America*, vol. 87, no. 4, 1990.

[6] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 33, no. 2, pp. 443–445, Apr 1985.