**Department of Electrical Engineering - UCLA**

**EE 214A – Digital Speech Processing**

**Project 2014 : Speech Activity Detection**

Speech activity detection (SAD) is used to detect the active and inactive regions of speech in an audio signal. SAD plays an important role in several signal processing applications. In speech communication systems, only active regions of speech are fully coded and transmitted, thus reducing the required bandwidth of the system [1]. In automatic speech recognition, endpoint detection(as it is referred to in this field) is used to extract words from the continuous acoustic signal prior to recognition [2]. Finally, in surveillance and monitoring systems, SAD can be used to detect the presence

of speakers [3]. In each of these applications, noise robustness is an important aspect.

**Task :** Design a noise robust SAD system to detect active regions within speech signals.

**Files :** You are provided with 10 utterances; 5 spoken by males and 5 by females. Each utterance is a phone number and labeled as follows: XYZ\_BCDEFGHA.08.

X indicates the gender (F or M), YZ is the speaker id, following that is a 7 digit number (if there is a ’Z’ in the phone number, it means that the person said ’zero’), followed by an A (which stands for 1st repetition,) and 08 refers to the sampling frequency (8 kHz).

Additionally, you have been provided with noisy versions of these speech signals at 0dB and 10dB SNR. The type of noise is babble noise. These are stored in the “Noise\_0dB” and “Noise\_10dB” folders

All the files are placed under ’project files’ in the ’Handouts’ section of the class webpage. When unzipped, you will find a folder for each noise condition and there are also annotation files.

**Processing and Evaluation**

1) First, convert your sphere files to .wav format, you can use the following Matlab script:

1

fid = fopen(filename, ’r’, ’b’);

y = fread(fid,inf, ’int16’);

fclose(fid);  
x = y/max(abs(y));

wavwrite(x,fs,filename)

Alternatively, the utterances can be opened use ’readsph’ command from the voicebox toolkit:

http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html

2) Any window length can be used to frame the signal, but the shift rate must be 10 ms per frame. For each utterance, you will made a decision every 10 msec as to whether or not a segment is speech (1) or non-speech (0). You will then generate a vector of ’0’ and ’1’ for each file.

3) If you need to train you algorithm, use the function [out] = read labels(filename, window length) to obtain ground truth labels for the textfile specified in filename. The second parameter window length is the size of your window in ms.

4) The output of your algorithm must be a column vector with decision values (1=speech, 0=non-speech) for each frame. A Matlab function will be provided to calculate the performance (per file) of your algorithm:

Function [Pfa, Pmiss, aveError] = evalSAD(decision,label,window length)

**Inputs**:

**decision**: A Nx1 column vector with labels 1 (speech) and 0 (non-speech) calculated from your algorithm.

**label:** String path indicating the location of your labeled text file for the given file.

**window length:** Window size that you use to frame the speech signal (in ms).

**Outputs:**

Pfa = Probability of False Alarm  
Pmiss = Probability of Missing a speech frame aveError = Average error

**Example:**

decision=[000000000011111111000000]’;label=’FBA\_79776O5OA.08.txt ’;

window leevangth = 30;

5) Calculate the average Pfa, Pmiss, and Average error across files for each evaluation folder (10dB, 0dB and CLEAN).

6) Since your decision vector with speech/non-speech frames could be sensitive to some threshold, you might want to test different thresholds to see how your algorithm performs. In that case, plot a ROC curve (Probability of false alarm vs. probability of missing) using the pairs of (Pfa, Pmiss) obtained with different thresholds. Try some thresholds that give you (Pfa, Pmiss) pairs close to the origin. Select the threshold that gives minimum average error for your final algorithm.

You cannot use a different algorithm for each gender, utterance or noise level. The same algorithm needs to be used in all cases. If you do use a statistical approach, you can only train on the clean samples.

**Code**:  
You are to write the algorithm in Matlab or C or C++.

**Groups and Presentations:**

Please work in groups of 2 or 3. Email me the list of students in your group as soon as possible. During the last lecture, each group will present their work, in class, for about 12-15 minutes (depending on the number of groups). Every member of each group should participate in these presentations.

**Report**:

The report should not exceed 6 pages and be in the style of an ICASSP paper (double columns; single spacing.) For more information on formatting, please go to :

<http://www.icassp2014.org/PaperKit.html>

and click on Part II Templates (note that ICASSP has a 5-page limit while we will allow up to 6 pages). The report should include:

Introduction and Background (what is the problem/why is it important; liter- 3

ature survey).

1. Project Description (methodology, implementation, results, etc.). Include a clear justification of the design choices, and highlight simulation results that show the performance of your system. Figures and flowcharts generally help clarify the text.
2. Summary and Discussion (also, ideas for future work).  
   Note that ’Error Analysis’ is crucial! Where and why did the algorithm fail and how to improve it.
3. The code that you used. Please document the code carefully and email it to me along with a README file showing how to run the code. Note that if you use someone else’s algorithm, please acknowledge that reference.

**Hints**: Many successful SAD methods rely on a number of features (measures) which provide complementary information since active speech can be recognized in vari- ous ways. For example, speech tends to involve higher energy levels, and thus the short-time energy is an accurate feature for speech classification [8]. The presence of voicing is also an accurate method to predict active speech, and voicing can be recognized by a number of distinctive patterns in both the time and frequency do- main. Furthermore, spectral entropy or variance can be used to predict active speech [6]. Many other successful features have been shown to work for SAD. Most modern VAD algorithms use a combination of features to classify speech as active or inactive [4],[7]. The numerous features can then be integrated using a voting-based method.

**References**:

[1] A. M. Kondoz, ”Digital Speech: Coding for Low Bitrate Communication Sys- tems,” Wiley, 2004.

[2] L. R. Rabiner and M. R.. Sambur, ”An algorithm for determining the end- points of isolated utterances,” Journal of the Acoustical Society of America, Vol. 56, Issue S1, pp. 31, 1974.

[3] J. Kacur, J. Frank, and G. Rozinaj, ”Speech Detection in the Noisy Environ- ment Using Wavelet Transforms,” EURASIP Conf. on Multimedia Communications, 2003.

[4] W.-H. Shin, B.-S. Lee, Y.-K. Lee, and J.-S. Lee, ”Speech/Non-Speech Classi- fication Using Multiple Features for Robust Endpoint Detection,” Proc. of ICASSP, pp. 1399-1402, 2000.

4

[5] R. V. Prasad, R. Muralishankar, V. S. Shankar, P. Pawelczak, and I. Niemegeers, ”Voice Activity Detection for VoIP- An Information Theoretic Approach,” GLOBE- COM, pp. 1-6, 2006.

[6] A. Benyassine, E. Shlomot, H.-Y. Su, and E. Yeun, ”A Robust Low Complex- ity Voice Activity Detection Algorithm for Speech Communication Systems,” IEEE Workshop on Speech Coding for Telecommunications, pp. 97-98, 1997.

[7] R. V. Prasad, A. Sangwan, H. S. Jamadagni, M. C. Chiranth, R. Sah, and V. Gaurav, ”Comparison of Voice Activity Detection Algorithms for VoIP,” Proc. of the International Symp. Of Computers and Communications, 2002.

[8] D. Enqing, L. Guizhong, Z. Yatong, and C. Yu, ”Voice Activity Detection Based on Short-Time Energy and Noise Spectrum Adaptation,” Proc. of ICSP, 2002.

[9] S. G. Gokhun and H. Ozer, ”Voice Activity in Nonstationary Noise”, IEEE Trans. on Speech and Audio Processing, Vol. 8, No 4. 2000.

[10] L. N. Tan, B. J. Borgstrom and A. Alwan, ”Voice Activity Detection us- ing Harmonic Frequency Components in Likelihood Ratio Test,” ICASSP 2010, pp. 4466-4469.

[11] J. Sohn, N. S. Kim, and W. Sung, A statistical model-based voice activity detection, IEEE Signal Processing Letters, vol. 6, no. 1, pp. 1-3, Jan. 1999.