International Journal of Electronics Communication and Computer Engineering
Volume 10, Issue 2, ISSN (Online): 2249–071X

A Novel Method for Microphone Independent Speech Activity Detection

Punnoose A.K.
Flare Speech Systems.
Corresponding author email id: punnoose@flarespeech.com

Date of publication (dd/mm/yyyy): 23/04/2019

***Abstract* – Pre-processing stages of a speech based information access system are often adversely impacted by the end user microphone quality. This may result in the rejection of a speech recording as a silence recording and vice versa. Parameters of the pre-processing stages need to be calibrated, independent of the end user microphone parameters. This paper discusses an approach for detecting whether a recording contains speech or not, independent of the microphone used to record the file. A pitch detection algorithm is used as the baseline and two new recording level pitch based features, pitch chunk partition ratio and average pitch chunk dynamic range are proposed. Experimental results are shown which proves the speech activity detection ability of the proposed features.**

***Keywords* – Microphone Sensitivity, Speech Vs Silence Detection, Pitch Based Features.**

I. INTRODUCTION

Modern automatic speech recognition is done in a multilevel manner. Bottom level corresponds to frame recognition. Next levels, would be phoneme recognition, word recognition, sentence recognition, respectively. In sentence level, the language model plays a key role in the overall word error. Roughly the bottom levels constitute a 50% and the top levels another 50%, in the overall accuracy. If frame level accuracies are less than a certain lower limit, then it is not possible to improve the word level accuracies, with any type of language models.

A crucial requirement of the recognition engine is the robustness. i.e., the ability to recognize under noisy conditions. To deal with noisy recordings, there are broadly two approaches. One is noise aware training (NAT) and the other is dealing with noise in the early pre-processing stage. Noise aware training is advantageous because, the spectral properties of the noisy speech signal will get learned by the recognition model, be it discriminative or generative. There is no need of any assumptions about the nature of noise. In most of the cases where noise is widespread throughout the recording, not even a noise tag is needed in the training phase. But this assumption free noisy training may not yield good recognition while testing in a totally different noise environment.

On the other hand a pre-processing stage is often used, based primarily on the end user conditions. Most noise detection and signal enhancement algorithms are employed at this pre-processing stage. If the end user noise environment is known in prior, the appropriate noise reduction algorithms can be used in the pre-processing stage. For eg, if the speech recognition engine is to be used in an automobile, then the pre-processing stage involves the vehicle sound detection, horn detection, etc. This approach enables to have a modular approach where the early pre-processing stage is heavily biased towards the final end user environment, insulating the core recognition engine.

Another prime consideration at the pre-processing stage is to decide whether the recording contains speech or noise or silence. This decision has a tremendous effect on the recognition engine afterward. If a noise file is wrongly identified as a speech file and forced to the recognition engine, a possible low acoustic score obtained due to noisy input can be overridden by a good language model score and may result in a detecting a sequence of words which may make sense.

In the pre-processing stage, another crucial issue is the characteristics of the microphone used. The parameters like sensitivity, impedance and even the type of PC/mobile device has such a crucial impact on the pre-processing state. A simple energy based speech/silence detection algorithm can be completely misled by the microphone characteristics alone.

Silence levels of one microphone may be completely different from the silence levels of another microphone. Even if the preprocessing steps involves spectral level processing, the ambient noise of the microphone demands additional noise reduction steps be performed, which also increases the computational demands.

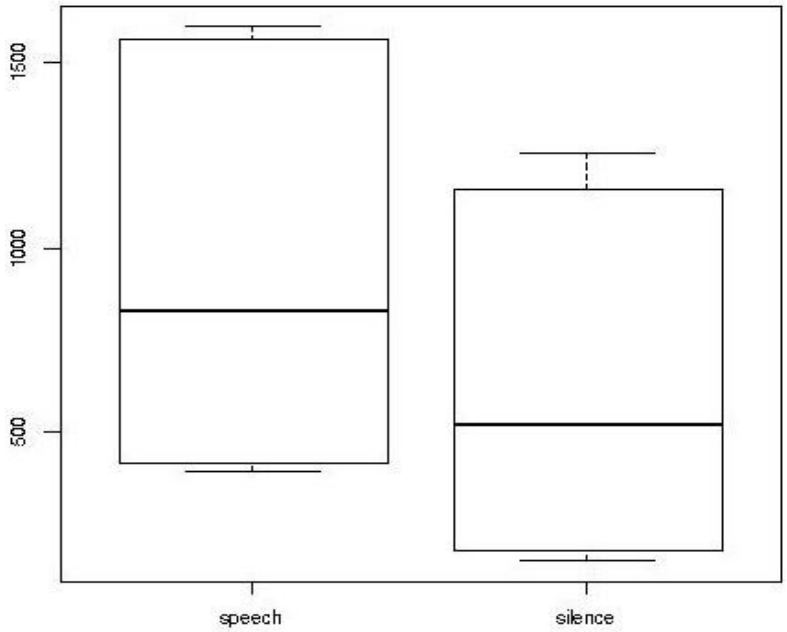

Fig. 1. Average energy of speech and silence files.

Fig.1 plots the average energy of a set of speech and silence recordings. All silence files are recorded with a single microphone and all speech files are recorded using another microphone. It is clear from the box plots that speech and silence recordings have overlapping average energy. Thus any energy based speech vs silence file detector will be confused by the overlap between speech and silence energy values.

This paper aims to develop features that aid an energy based speech activity detector, which works well in case of high and low energy irrespective of microphone used, but unable to discriminate speech and silence in a mid level overlapping energy case as shown in fig 1.

II. Related Work

Speech activity detection has a rich literature. One common approach is to use sparse coding to learn a combined dictionary of speech and noise and then removing noise part, to get the pure speech representation [1, 2]. The correspondence between the features derived from the clean speech dictionary and the speech/ non-speech labels can be learned using discriminative models like conditional random fields [3].

Autocorrelation functions and its various derivatives have been used extensively for voice activity detection.

Sub band decomposition and suppressing certain subbands based on stationarity assumptions on autocorrelation function is used for robust VAD [4]. Autocorrelation derived features like harmonicity, clarity, periodicity provides more speech like characteristics. Pitch continuity in speech has also been exploited for robust speech

activity detection [5]. For highly degraded channels, GABOR features along with autocorrelation derived features are also used [6]. Modulation frequency has also been used in conjunction with harmonicity for VAD [7].

Another approach is to model the whole acoustic space using a universal background model (UBM). Gaussian mixture models are used as universal background models, which needs only unlabelled data to train. Using a small set of labeled speech and non-speech data, summary such as Baum-Welch statistics can be calculated using a universal background model and stored as prototype vectors representing speech and non-speech classes. A simple thresholding mechanism can be used to determine whether a recording is speech or non-speech [8].

Another very common method is to use mel frequency cepstral features with classifiers like SVMs to predict speech regions [9]. Derived spectral features like low short-time energy ratio, high zero-crossing rate ratio, line spectral pairs, spectral flux, spectral centroid, spectral roll off, ratio of magnitude in speech band, top peaks, ratio of magnitude under top peaks are also used to predict speech/ non-speech regions [10].

In [11], authors propose a full front end for speech activity detection, based on multiple derived acoustic features. Frequency domain linear prediction (FDLP) is used to derive a set of short-term spectral features and long-term modulation features. Another set of spectro-temporal features is derived from passing spectrogram through a bank of spectro-temporal modulation selective filters, which reflects the cortical multiscale representation of speech. Yet another set of feature is the posteriors from multi-layer perceptrons trained to differentiate speech/ non-speech using the above mentioned derived features. Dimensionality reduction is employed on these features sets and a Gaussian mixture model is trained to classify speech and non-speech.

The rest of the paper is organized as follows. Two file level pitch based features, pitch chunk partition ratio and average pitch chunk dynamic range are explained. Experimental results are provided to prove the discriminative power of the proposed features.

III. Approach and Analysis

An autocorrelation based pitch detector with an upper and lower cut-off is used as the baseline pitch detector. Although there is a difference in the pitch range of men and women, it is not considered in this analysis. Fig 2 and 3 plots the count of the pitch values of each pitch frame, for a speech recording and silence recording, recorded using 2 different microphones, respectively. Both recordings are made in the same silent environment with very minimal background noise. Each pitch frame is of duration 50ms.

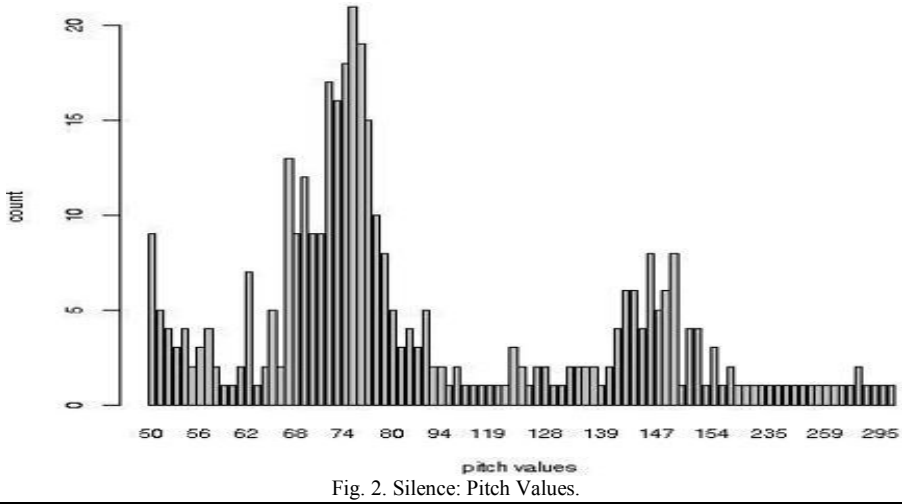

Fig. 2. Silence: Pitch Values.

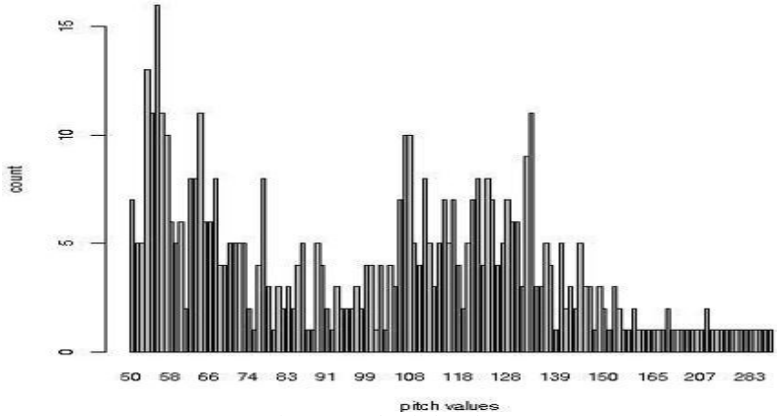

Fig. 3. Speech: Pitch Values.

There could be a vast difference between microphones in the parameters like sensitivity, impedance, etc. In a consumer facing speech application often speaker tends to tinker with the microphone volume. For a low quality microphone, if the recording volume is high, internal system noise tends to be recorded, which may have some harmonic nature. This harmonic nature will get picked by the pitch detector, as seen in fig 2. A sharp peak at around 75Hz in fig 2 could be the noise signature of the microphone used for recording the silence. In fact, it is clearly difficult to distinguish between speech and silence recordings if the pitch values are considered in isolation.

Motivated by the insight that some of the system internal noise captured by a low quality microphone may have harmonic nature, we explore whether this harmonic nature, if captured, can be used for voice activity detection. As individual pitch frames don't seem to offer the needed differentiation, chunks of frames have to be considered. A pitch chunk is defined as the consecutive pitch frames with pitch values greater than a threshold. Pitch chunk width is the number of pitch frames in the chunk.

1. *Pitch Chunk Partition Ratio:*

Pitch chunk partition ratio Prr, of a recording is defined as

$$P_{rr} = \frac{\sum_{i=3}^{K} C(k)}{\sum_{i=1}^{2} C(k)}$$

where $C(k)$ is the count of pitch chunk of size k in the recording. K is the maximum pitch chunk size. For speech recordings, the number of large pitch chunks will be more than small pitch chunks. This is due to the concept of pitch continuity. For silence recordings with minimal background noise, the number of large pitch chunks will also be more than small pitch chunks, but relatively less compared to that of speech recordings. Even though there could be harmonicity in the silence recordings with minimal background noise or background speech, the pitch continuity will be relatively less.

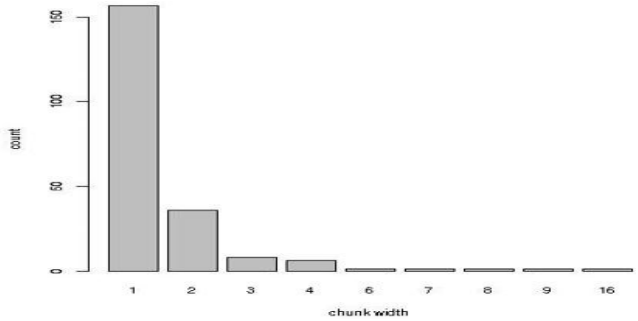

Fig. 4. Silence pitch chunk width.

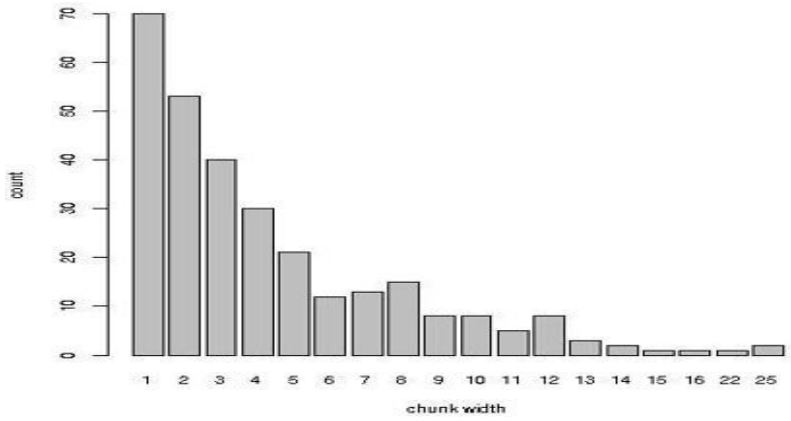

Fig. 5. Speech pitch chunk width.

Fig 4 and 5 plots the count of pitch chunks of various width for a speech file and silence file, again recorded using different microphones. It is apparent the difference in the chunk counts.

2. *Average Pitch Chunk Dynamic Range:*

Average pitch chunk dynamic range Pv, of a file is defined as,

$$P_v = \frac{1}{C}\sum_{c=0}^{C}(max\ (P_c) - \min(P_c))$$

where a file C is the number of pitch chunks in the recording. min (Pc) and max(Pc) is the minimum and maximum pitch values in the pitch chunk c. Fig 6 and 7 shows the count of difference in max and min pitch of pitch chunk of width 4, for silence and speech recordings. It is evident from the plots that for silence the pitch variation chunk wise is more constrained than that of speech recordings.

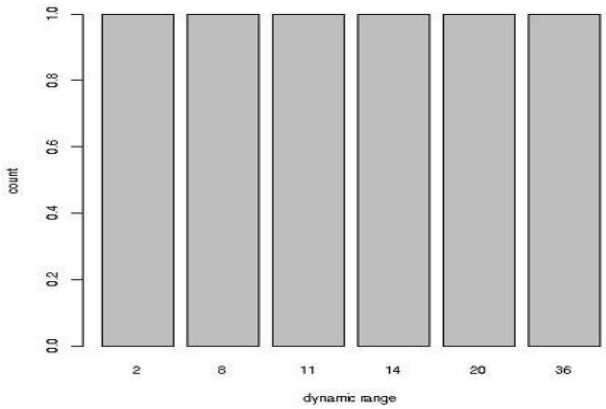

Fig. 6. Silence: Average dynamic range.

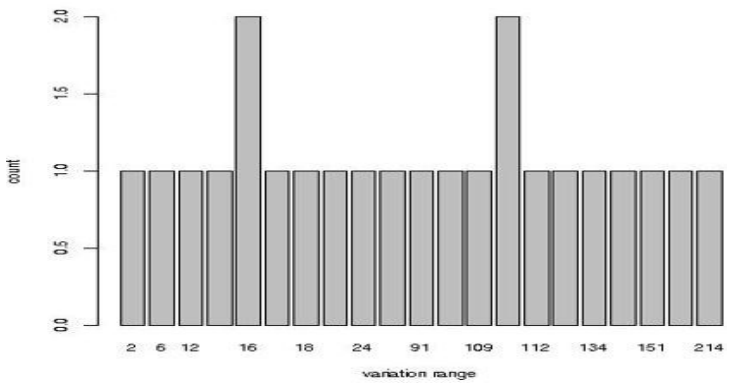

Fig. 7. Speech: Average dynamic range.

IV. Experimental Results

For proving the discriminative power of the proposed features, NOIZEUS corpus is used for speech data and background noise subset of the CHiME dataset is used as the silence data. Noise in the background in speech as well as in silence data will account for the real-world conditions. Note that the CHiME background noise contains even background speech in a lot of instances, which allow us to interpret the results, as the discriminatory power of the proposed features to even discriminate between foreground and background speech.

The Receiver Operating Characteristic (ROC) curve for the features are plotted for various values of Prr and Pv in fig 8 and 9 respectively. It is clear from the plots that both of the features can aid as an add-on to any speech activity detection already in place.

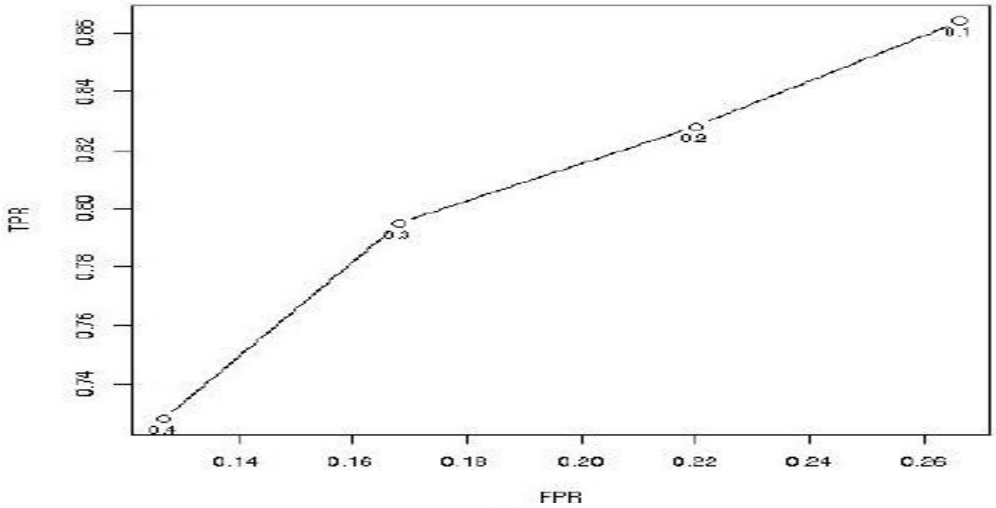

Fig. 8. ROC for pitch chunk partition ratio.

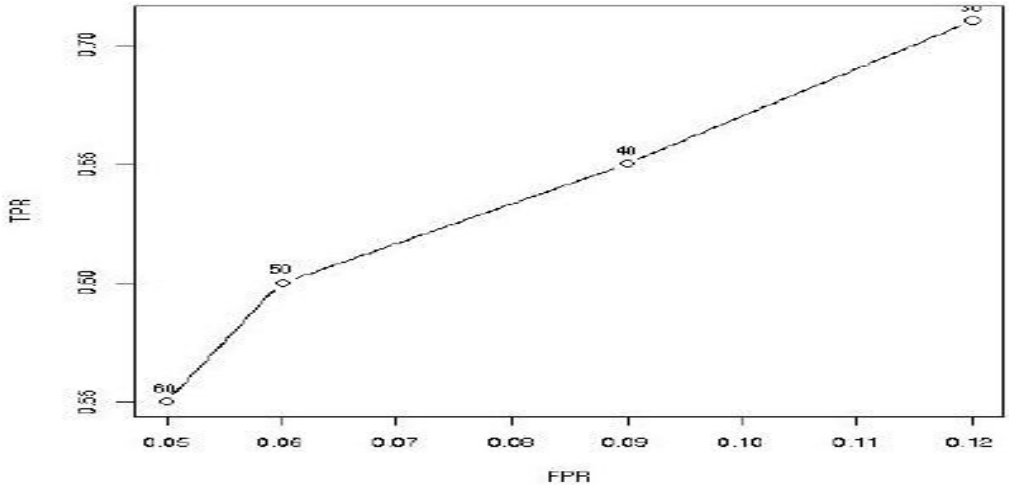

Fig. 9. ROC for average pitch chunk dynamic range.

V. Conclusion and Future Work

In this paper, a novel approach to detect speech activity, which aids an energy based speech activity detector, in a file level is proposed. Two new recording level features, pitch chunk partition ratio and average pitch chunk dynamic range, are proposed. The discriminating capability of the proposed features is demonstrated. Two noisy datasets, one for speech and the other for pure background noise, are used for experimental validation of the robustness of the proposed features.

Subband based frequency analysis capture various dynamic characteristics of the speech, which is more robust to noise. We intend to incorporate more robust spectral features into the speech activity detection framework. Rather than analyzing features in isolation, combining features will improve predictive capability.

REFERENCES

[1] Deng, Shi-wen & Han, Jiqing. (2013). "Statistical voice activity detection based on sparse representation over learned dictionary." Digital Signal Processing. 23. 12281232. 10.1016/j.dsp.2013.03.005.
[2] Ahmadi, Parvin & Joneidi, Mohsen. (2014). "A New Method for Voice Activity Detection Based on Sparse Representation." Proceedings – 2014 7th International Congress on Image and Signal Processing, CISP 2014. 10.1109/CISP.2014.7003901.
[3] Peng Teng and Yunde Jia "Voice Activity Detection via Noise Reducing Using Non-Negative Sparse Coding" IEEE Signal Processing Letters, Volume: 20, Issue: 5, May 2013, Page (s): 475-478.
[4] Keansub Lee and Daniel P. W. Ellis "Voice Activity Detection in Personal Audio Recordings Using Autocorrelogram Compensation" INTERSPEECH 2006: ICSLP: Proceedings of the Ninth International Conference on Spoken Language Processing: September 17 - 21, 2006, Pittsburgh, Pennsylvania, USA.
[5] Shao, Yiwen & Lin, Qiguang. (2018). "Use of Pitch Continuity for Robust Speech Activity Detection." 5534-5538. 10.1109/ ICASSP. 2018.8462482.
[6] M. Graciarena, A. Alwan, D. Ellis, H. Franco, L. Ferrer, J.H. Hansen, A. Janin, B.S. Lee, Y. Lei, V. Mitra, et al., "All for one: Feature combination for highly channel degraded speech activity detection" in INTERSPEECH, pp. 709713, 2013.
[7] E. Chuangsuwanich and J. Glass, "Robust voice activity detector for real world applications using harmonicity and modulation frequency," in Twelfth Annual Conference of the International Speech Communication Association, 2011.
[8] Omid Ghahabi, Wei Zhou, Volker Fischer "A robust voice activity detection for real-time automatic speech recognition system".
[9] Kinnunen, Tomi & Chernenko, Evgenia & Tuononen, Marko & Frnti, Pasi & Li, Haizhou. (2012). "Voice Activity Detection using MFCC Features and Support Vector Machine." 2.
[10] A. Misra, "Speech/non-speech segmentation in web videos," in Proceedings of INTERSPEECH, 2012.
[11] Thomas, S & Mallidi, S.H. & Janu, T & Hermansky, Hynek & Mesgarani, N & Zhou, X & Shamma, Shihab & Ng, Tim & Zhang, B & Nguyen, Long & Matsoukas, S. (2012). "Acoustic and data driven features for robust speech activity detection." 13th Annual Conference of the International Speech Communication Association 2012, INTERSPEECH 2012. 3. 1983-1986.