ISSN (Online) 2394-6849

International Journal of Engineering Research in Electronics and Communication Engineering (IJERECE)
Vol 5, Issue 6, June 2018

A Novel Approach for Speech File Detection

[1] Punnoose A K

***Abstract*— This paper discuss a novel approach to detect speech files using a frame classifier. The speech files tends to have the subphones, corresponding to a phone, recognized in sequence, while run through a frame classifier. Duration of subphone sequence corresponding to a phone also tends to differ in speech and noise. Distributions are used to capture the count statistics of recognized subphone sequence, along with the phone duration. A probabilistic framework is formulated to score a wave file for the presence of speech. Relevant speech and noise datasets are used to benchmark the approach.**

Index Terms:—Cobb angle, Sum of Squared Difference (SSD), polynomial curve fitting method.

INTRODUCTION

In most speech recognition based interactive voice response system(IVRS), a pre-processing step is needed which tells whether a file contains speech or not. A misrecognition in one of the steps could prompt the dialogue manager, which directs the dialogue, to take undesirable paths through the dialog tree. Mostly signal processing based approaches are used to detect the level of noise or speech in a wave file. A major drawback with signal processing based approaches is that, it often makes assumptions about the noise, which is generally not practical. One such assumption is the stationarity of noise, which assumes that the spectrum of noise is relatively same across time. This allows spectral subtraction to be employed. But in reality, real-world noise conditions seldom follow stationarity in spectrum. In fact real-world noise will be anything but being stationary. Moreover many phones has a lot of similarity with noise, spectrum wise, which will make spectral subtraction difficult. Another approach is model the speech, rather than noise. As the spectral variations in speech will be limited and more contained as compared to that of noise which could be very broad, it will be easy to model the aspects of speech such as harmonicity, pitch, etc so that differentiation between speech and noise is easier. But a lot of noise types are also harmonic, which will cause difficulties in discriminating speech and noise eventually.

In terms of application, a dialogue manager will have the information regarding what type of confidence scoring for speech, to be employed, depending upon the node. A node in a dialog path is a system prompt followed by a user utterance. If the dialogue nodes corresponds to a confirmation, where a false positive will be too expensive, the wave file can only be passed to the speech recognition engine, once the there is enough confidence that the file contains speech.

On the other hand if the dialogue node involves the recognition of a word from a list, then skipping the preprocessing step may be prefered, thus allowing the speech recognition engine to output a hypothesis, either frame wise or phone wise or word wise, depending upon the engine. Now using a mathematical model to suggest how a phone might get affected by the presence of noise, some recovery is possible.

In critical applications such as banking, not even a single false positives can be afforded, even at the expense of missing some of the genuine speech files. In such cases, a pre-processing step before passing the wave file to a speech recognition engine is very much necessary. This paper captures the biases of frame classifiers, for noise and speech, and presents a probabilistic model to score the presence of speech in a wave file.

II. PROBLEM DEFINITION

Given a wave file, derive a mechanism to find out whether the file is speech or not.

III. PRIOR WORK

In [1], author discuss an approach using a set of temporal and spectral features to segment the videos into speech and non speech. Author uses features like Low short-time energy ratio,high zero-crossing rate ratio, Line Spectral Pairs, Spectral centroid, Spectral Roll-off, Spectral Flux, etc. Classifiers are trained to predict whether a segment is speech or non-speech. In [2], authors use a neural network for learning the phone durations. The input features are derived from the phone identities of a context window of phones, along with the durations of preceding phones within that window.

In [3], authors discuss about a noise robust Voice Activity Detection(VAD) system, utilizing periodicity of signal, full band energy and ratio of high to low band signal energy. Voice regions of speech are identified and then proceeds to differentiate unvoiced regions from silence and background noise using energy ratio and energy of total signal. In [4], authors present spectral feature for detecting the presence of spoken speech in presence of mixed signal. The feature is based on the presence of a trajectory of harmonics, in speech signal. The property that, speech harmonics cover multiple

ISSN (Online) 2394-6849

International Journal of Engineering Research in Electronics and Communication Engineering (IJERECE)
Vol 5, Issue 6, June 2018

frames in time, is treated as a feature.

In [5], authors use harmonics, pitch and subband energy to locate the speech and track the time-varying noise. Pitch measurements are used to detect the vowel segments. Subbands are divided based on energy and frequency and based on predetermined thresholds from determinate noise, voiced parts of potential voice regions, are identified. In [6], author propose a new feature named Mean-Delta feature. This feature is the mean of the absolute values of the delta spectral autocorrelation function of the power spectrum.

IV. APPROACH

The fundamental insight of this paper is that, speech files when run through a frame classifier trained to detect phones, tends to detect subphones in the right sequence, compared to noise files. Or in another words, for noise files, the subphones are less likely to be recognized in a sequence. The approach is summarized in the following steps

1) Train a frame classifier with subphone labels, with speech data

2) Run the classifier across speech and the noise data to produce the data for training the variables, which is used eventually to discriminate between speech and noise.

3) Make distributions for speech and noise data, on identified variables. Count of phones where the sub phones are detected sequentially, as well as the phone duration detected are used as the variables.

4) Formulate a probabilistic framework to score for speech files.

Three datasets are used in this approach. td1 is used to create the frame classifer. td2 is used to fit the distributions. td3 is used for testing. For td1, a subset of Voxforge dataset is used. For td2, a subset of Voxforge data as well as a subset of CHiME background data is used. For td3 , another subset of Voxforge data as well as a subset of CHiME background data is used.

A. Frame Classifier Details

A Multi Layer Perceptron(MLP) is trained with plp coefficients as the features and subphone units as labels. Each wave file while testing, will produce a sequence of subphone labels, each corresponding to a frame. If a sequence of labels recognized are subphones in a sequence corresponding to a phone, then it is treated as the recognized phone with all constituent subphones. Define ps2-s4 as a phone p recognized with the all the subphones ps_2; ps_3; ps_4, occurring monotonically. Let l be the length of the recognized phone which is the number of the frames in that phone.

B. Measures for Speech vs Noise File Detection

Two measures are explored to separate speech files from noise files. The count of phones where subphone sequence is observed and the difference in length of the phones recognized, between speech and noise files.

1) Subphone Sequence Count: To understand which all subphone sequence has to be used for speech vs noise separation, the count of different subphone sequence detected in speech is compared to that of noise. Fig 1 and 2 shows the count of phones with all possible subphone sequences detected for speech and noise respectively.

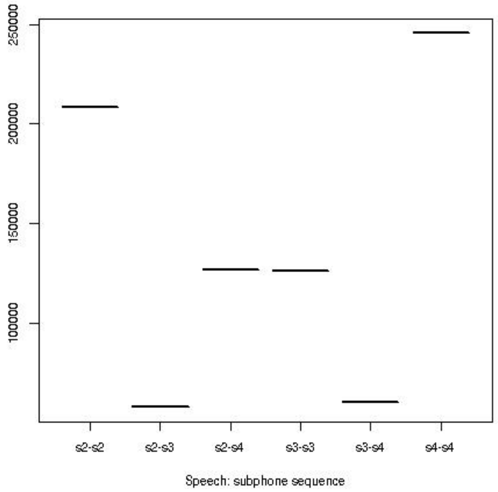


Fig. 1. Count statistics in speech

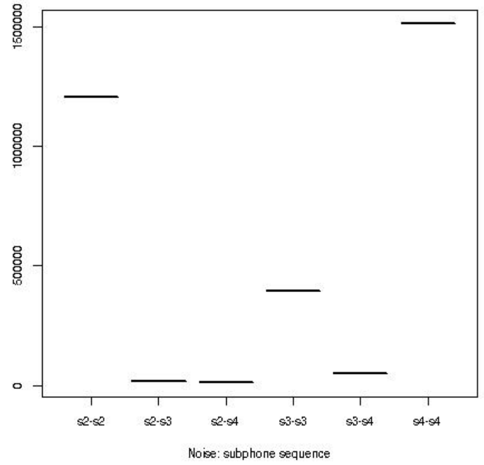


Fig. 2. Count statistics in noise

It is clear from both the figures that the maximum difference in count is for the state transitions s2-s4 and for s2-s3. Note that this figures are plotted from td2, which has roughly equal amount of speech and noise data. Silence phones are also excluded from this analysis. So any count statistics is the reflection of speech and noise characteristics.
For noise data, s2-s2 and s4-s4, has the maximum count, irrespective of the phone, which itself is an indication of noise. Thus suggest the presence of two type of noise. One is noise which is stationary in spectrum, which itself will get manifested in a long chunk of one subphone. Another could be totally non stationary noise, which will be evident in the presence of large number of small chunks of single subphones.

The above 2 plots gives the justification of using only 2 subphone sequence ie, s2-s3 and s2-s4, for further analysis and rejecting all other subphone sequence, as it doesn't provide sufficient discrimination in terms of counts, for separation between speech and noise.

2) Phone Weightage: The count of phones detected for speech data and noise data is very different. Fig 3 and 4 plots the count of various phones detected in speech and noise respectively. Phones detected from the speech data tends to be distributed across the phones. On the other hand very few distinct phones are detected for noise data.

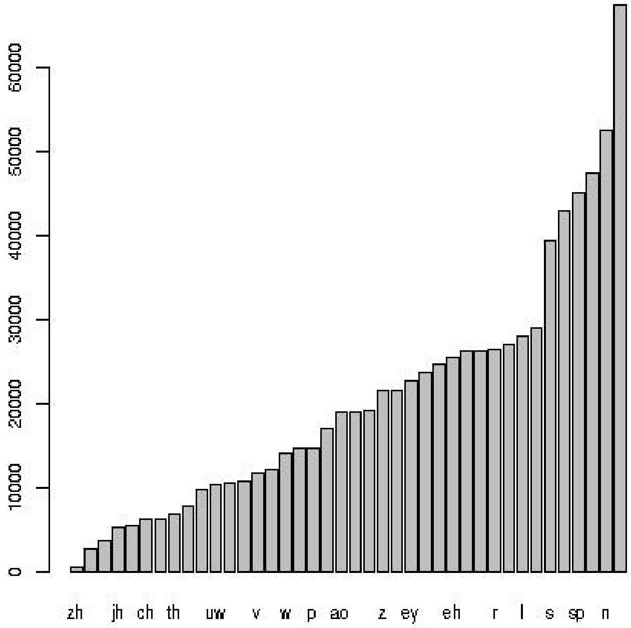


Fig. 3. Phone count in speech

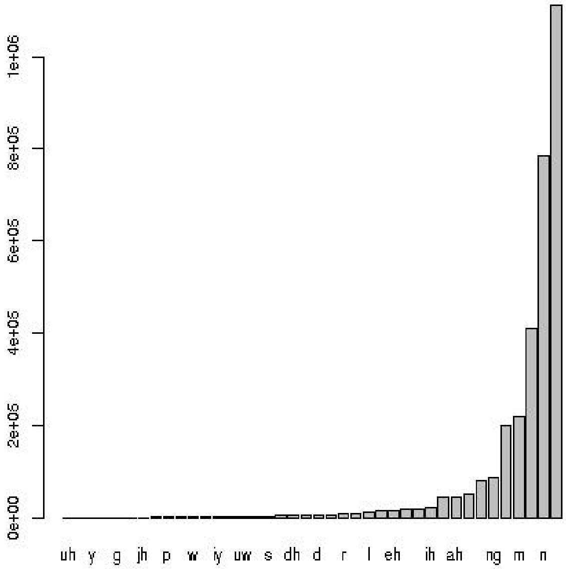


Fig. 4. Phone count in noise

For any mechanism which is aimed at detecting speech, where the counts of phones plays a role, a phone weighting mechanism is needed, which weighs phones based on the count of occurence in speech and noise data. Note the phone n which has high count in speech as well as in noise data. In the absence of a phone weightage mechanism skewed at detecting phones in speech, phones like n will result in more false positives. Denote Cs(p) and Cn(p) as the count of phone p detected in speech and noise data respectively. Define a weighting mechanism,

$$\begin{aligned} q_s(p) &= \frac{\frac{C_s(p)}{\sum_p C_s(p)}}{\frac{C_n(p)+1}{\sum_p C_n(p)}} \\ W_s(p) &= \frac{q_s(p)}{\sum p q_s(p)} \end{aligned} \qquad (1)$$

This effectively normalizes the count of a phone in speech, with the count in noise, so that phones with maximum count in speech and minimum count in noise gets the highest weightage.
3) Phone Length Distribution: Length of a phone is the number of frames corresponding to that phone. Rather than taking length of every phones separately, distributions of phone length, where there is a subphone sequence recognized and the case where no subphone sequence is recognized, is taken.

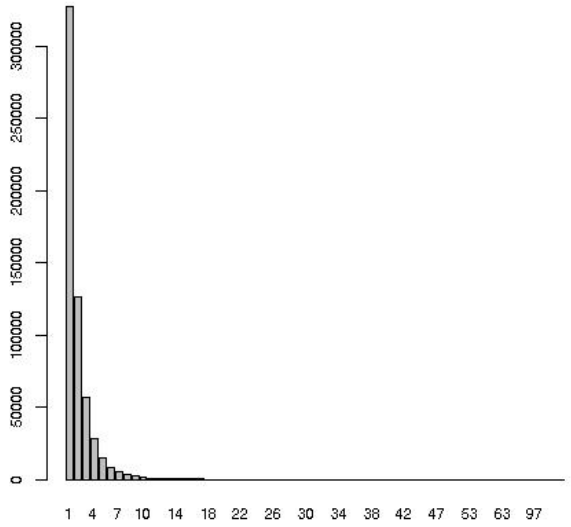


Fig. 5. Speech: Phone duration counts where subphones not in sequence

Fig 5 plots the distribution of phone lengths of all phones combined, for speech data, where no subphone sequence is recognized. Note that most phones recognized are of single frames, which could have been the case of a misrecognition.

Fig 6 plots the distribution of phone lengths of all phones combined, for speech data, where subphone sequence is recognized. It's evident that the count increases at about 4 frames then becomes trails off gradually.

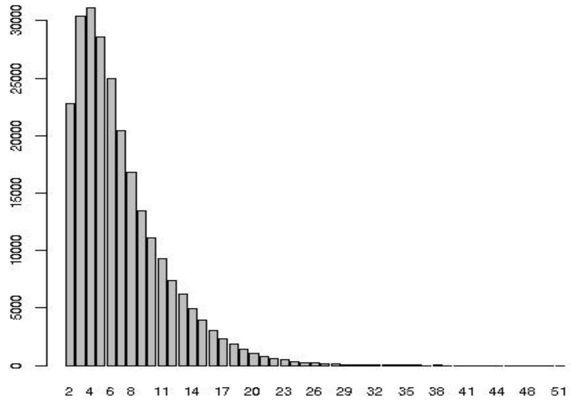


Fig. 6. Speech: Phone duration counts where subphones in sequence

Fig 7 plots the distribution of phone length of all phones for noise data, where there is no sequence of subphones detected. Fig 8 plots for the noise data, where there is subphone sequence detected. The shape of both the plots are almost the same, and is virtually very similar to the Fig 5.

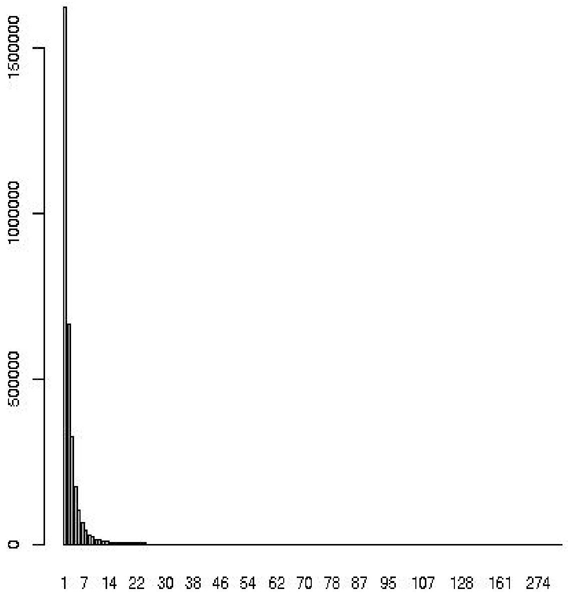


Fig. 7. Noise: Phone duration counts where subphones not in sequence

The similarity of plots 5 and 7, indicates that, for cases where the subphone sequence is absent, then it is difficult to differentiate speech and noise. This adds to the point that the subphone sequence is a key factor, if detected, could differentiate speech and noise robustly.

Converting Fig 5 to a discrete distribution on phone length.

$$P(l;s) = \frac{C(l)}{\sum_l C(l)} \tag{2}$$

This distribution is independent of phone. It's only dependant on the phone length.

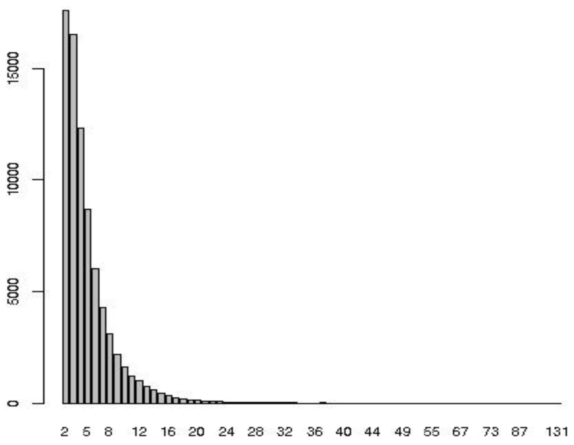


Fig. 8. Noise: Phone duration counts where subphones in sequence

An issue with the using the distribution in Equation (2) is that, a noise file is more likely to give more average probability from (2) than a speech file. This is because a noise file with subphone sequence detected, has the subphone chunk length very small, as is shown in Fig 8, compared to subphone chunks in a speech file. Subphone chunks in a speech file tends to be distributed more widely in length, as shown in Fig 6, than that of a noise file, which is very narrowly distributed to very small lengths. A solution to this problem is to use the phone chunk, if the chunk size is less than a threshold length.

V. EXPERIMENTAL DETAILS AND RESULTS

For a phone chunk p detected with length l, the posterior probability of that chunk belonging to speech can be written as,

$$\begin{aligned} P(c_s|p,l) &= \frac{P(l|p,c_s)P(p|c_s)P(c_s)}{P(p,l)} \\ &= \frac{P(l|c_s)P(p|c_s)P(c_s)}{P(p,l)} \qquad (3) \\ &\propto P(l|c_s)P(p|c_s)P(c_s) \end{aligned}$$

where c_s is a chunk of wave file, where the phone p is recognized, which is deemed to be speech.
With $P(p|c_s) = W_s(p)$ from (1) and $P(l|c_s) = P(l;s)$ from equation (2), and with the prior $P(c_s) = 1$, the chunk level posterior can be rewritten as

$$P(c_s|p,l) \propto P(l;s)W_s(p)$$

Extending to a file,

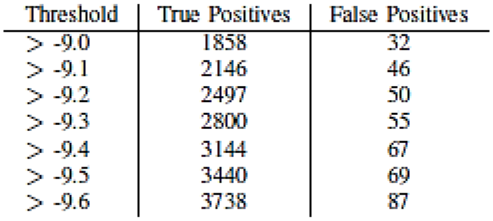
$$P(f = speech|(p_1,l_1)...(p_n,l_n)) \quad = \prod_{i=1}^{n} P(c_{s_i}|p_i,l_i)$$

This is under the assumption that each phone chunk is independant. To avoid the underflow at the file level, the above equation can be rewritten as,

$$P(f = speech|(p_1,l_1)...(p_n,l_n)) \propto \frac{1}{N} ln(P(c_{s_i}|p_i,l_i)) \quad (4)$$

A. Results

From the testing dataset td3, speech and noise files are run against the speech model given in Equation (2). The results are shown in Table 1. True Positive are the cases, where a speech file is detected as speech, and false positives are where when a noise file is detected as speech. As the threshold increases, the true positives and false positives increases. Precision is shown in the Fig 9.
Threshold True Positives False Positives



Threshold	True Positives	False Positives
> -9.0	1858	32
> -9.1	2146	46
> -9.2	2497	50
> -9.3	2800	55
> -9.4	3144	67
> -9.5	3440	69
> -9.6	3738	87

TABLE I. RESULTS: TRUE POSITIVES VS FALSE POSITIVES

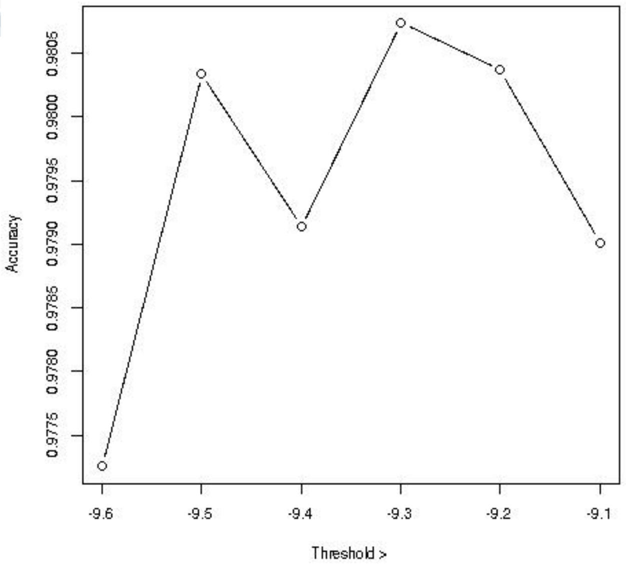


Fig. 9. Precision in Detecting Speech

VI. CONCLUSION AND FUTURE WORK

A new approach is presented to detect the files which contains speech. A frame classifier is trained to detect the subphones. The inherent bias of frame classifier in terms of difference in detecting subphone sequence, corresponding to phones, for speech and noise data, are codified in terms of distributions. Another feature used is the length of the phone chunk irrespective of the phone. Using a probabilistic framework, a decision mechanism based on thresholding is presented. The results are shown for precision of the detector. Any new feature can be incorporated into this framework with proper conditional independence assumptions. In future we aim to explore features specific to particular phones or broad phone classes, to make the approach more robust.

REFERENCES

[1] Ananya Misra, " NonSpeech Segmentation in Web Videos",
[2] Hossein Hadian, Daniel Povey, Hossein Sameti, Sanjeev Khudanpur, "Phone duration modeling for LVCSR using neural networks"
[3] E. Verteletskaya, K. Sakhnov, "Voice Activity Detection for Speech Enhancement Application",
[4] Reinhard Sonnleitner, Bernhard Niedermayer, Gerhard Widmer, Jan Schluter, "A Simple and Effective Spectral Feature for Speech Detection in Mixed Audio Signal",
[5] Zhihao Ahang and Jinlong Lin,"Robust Voice Activity detection Based on Pitch and Subband Energy"
[6] Atanas Ouzouniv,"A Robust Feature for Speech Recognition"