ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol4, Issue 6, June 2017

Speech File Detection by a Rule Based System

[1] Punnoose A K, [2] Ravishanker M
Applied Researcher Flare Speech Systems Bangalore

Abstract— **This paper discuss about detecting speech files in a real world speech recognition task. Detecting files with small background speech or noise, changes the overall behaviour of the Interactive Voice Response System. We experiment with neural networks trained to recognize phonemes, and outline a very simple yet effective approach to discriminate files that contains speech from that of noisy files. We use some popular publically available dataset, to validate our approach.**

Keywords—Noise Robustness, neural networks, Interactive Voice Response Systems

I. INTRODUCTION

In modern speech based information access, the process follows essentially a finite state machine approach. At any stage in the dialogue, we are in a particular state and in that state a question will be asked and based on the user answer, decoded, system will move to the next state. In speech recognition, noise robustness is a key aspect that significantly changes the behaviour of the system and impacts the user ex- perience. There are two issues associated with noise robustness in an IVR based system. One is speech recognition in noisy environment, which is dubbed as noisy speech recognition. Another is recognizing wave files with only slight noise or background speech and to determine whether such a file should be passed on to the recognizer at all. The former case is the classic case of noise robust speech recognition. There are different ways of tackling it. A recent approach is Noise Aware Training(NAT), which include different type of noises in the training data, and training the system with the noisy speech to make it more robust to noise.

In this paper we address the later case, where the wave file recorded may or may not contain speech. A decision has to be made whether the wave file should be be send to the speech recognizer or not. If the noisy wave file is send to the recognizer, there is always a chance of misrecognition. Moreover it depends on the type of recognizer in use. If it's a full blown sentence recognizer, in conjunction with a dialogue manager, an undesirable sentence coming out of the recognizer, drives the dialogue in a wrong path. In this case, based on some language model score and thresholding, to an extend, recovery is possible.

On the other hand if the recognizer is a isolated word based one, it further aggravates the situation, which will allow the system to proceed in a wrong path. Confidence scoring should come from the acoustic model, which may not be consistent all the time. This paper discuss an approach, which uses a neural network trained to recognize phonemes, to detect irrelevant.

II. PROBLEM DEFINITION

Given a wave file, we need to determine whether it has to be send to the speech recognizer to recognize it or not. A decision mechanism is needed in the preprocessing stage, which in a sense, tells how good the wave file is, so that the chances of a correct recognition by the speech recognizer is more likely.

III. PRIOR WORK

In [1], author talks about an approach where deep neural nets are used for noise robust speech recognition. Noise is added into the training data and posterior classifiers are trained. This can be used further in 2 ways. One is that posteriors probability f every phone can be treated as a feature and can be used to train a GMM-HMM based speech recognizer. Another way is to train neural network to predict the likelihood of states of context dependant triphones, directly, and do Viterbi using the likelihood, thus bypassing GMM at all. In both of the approaches, noise is added in the training data, thus making it more immune to noise in the testing conditions. In [3], author discusses about multiple stream of features. it can be like features derived from different frequency bands. Typically feature would be the posterior output of neural network trained to recognize different phonemes, from a specific band of frequency. Author discuss a way of combining different features based on some similarity measure. The crux of this approach is that, even if one stream of features are corrupted by noise, the other streams will cover for it.

In [2], author discusses 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 [4], authors discusses 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 [5], 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 frames in time, is treated as a feature.

Our simple approach is explained in the following steps.
1) Train a neural network classifier to predict phones, from frame as input. Here the input is [x1x2:::x9]. That is 9 frames are concatenated together to form a single vector. Each xi is a Perceptual Linear Prediction(PLP) coefficients. Each xi corresponds to time
sample of 25ms.
2) For a given wave file, run across all of the frames to get the phoneme output. Let's say P = [P1; P2; P3:::PN] output labels are there, each corresponding to a frame. N is the total number of frames in the wave file.
3) Calculate the following tentative statistics

a) Total Number of Different Phones
$$S_P = \sum_i \mathbb{1}(P_i \neq P_j) \; \forall j \neq i$$
b) $$Sil_{perc} = \frac{\sum_i \mathbb{1}(P_i = sil)}{N}$$
c) $$NonSil_{perc} = \frac{\sum_i \mathbb{1}(P_i \neq sil)}{N}$$

4) Make a set of tentative rules like
a) If(SP < a1 and Silperc > b2) then noise
b) If SP > a2 and Silperc < b2) then speech
where a1; a2; b1; b2 are thresholds which has to be experimentally determined.

V. EXPERIMENTS

A. Rationale for Voxforge as Training Data

For Experiments we used Voxforge data, which is available free for public use. The reason for selecting Voxforge data is multi-fold. First is that it's telephonic narrow band data. Sec- ond and foremost reason is that it's recorded in an uncontrolled way by different people with different accent, with different mother tongue, etc. This will give the necessary variability in the data, which is very much crucial for making a speaker independent telephonic information access system. This is very much against the popular notion of using a very well known data base like TIMIT. as the focus here is on real world telephonic IVRS, where the user response is simply silence or background speech, or just murmuring, or traffic noise, or noise of any other kind. A rough approximation of analysing a real world speech based information access system will show that roughly a 20% of the user utterance is of any significant speech content. This heavily bias us to use a speech database which is uncontrolled and with wide variability.

B. Training Details

For training a neural network classifier, we used approximately 27000 wave files. We first forced aligned the wave files with the transcript, with HTK toolkit, and got the labels. We used ICSI feacalc tool for generating perceptual linear prediction(plp) plp coefficients. Finally the plp and the labels are combined to create a training set. ICSI Quicknet is used for training the neural network. 23000 wavefiles are used for training and 4000 files for cross validation. Quicknet stops training when the cross validation accuracy doesn't increase above a threshold in two successive epochs. Mini batch gradi- ent descent is used as the training mechanism. Cross Entropy is used as the criterion for backpropogation training.

C. Testing Data

We use ChiME dataset which is publically available for testing the background noise. We use only the background noise section of the CHiME dataset, for testing how well our approach work. This dataset is divided into 7 parts with Signal to Noise Ratio(SNR) of 0dB, 3dB, 6dB, 9dB, 12dB, 15dB,18dB. We use a subset of Voxforge data, approximately 19000 wave files for testing our approach against speech files. We report results against these 2 type of data and experimentally validate the effectiveness our simple approach.

VI. RESULTS & ANALYSIS

Results are reported for speech and background noise for various measures such as number of distinct phonemes and phoneme coverage percentage.

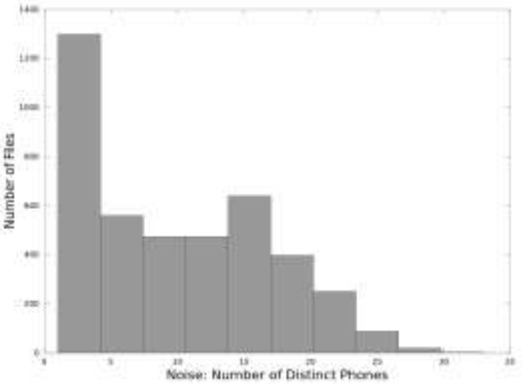


Fig. 1. Background Noise - Number of unique phones

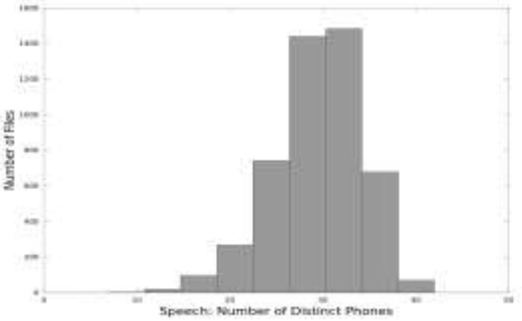


Fig. 2. Speech Data - Number of unique phones

1) Number of Unique Phones: Figure 1 and Figure 2 reports the histogram of number of unique phones in background noise and in speech data. We can see that number of phones in background noise is much lesser than the files with speech content. Note that the size of the file in background noise is around 1.2 seconds, and around 2 seconds in speech data files. Despite this difference the noisy background files individually are consistent spectrally. What we mean by that is the presence of similar kind of noise throughout in a single file. This assumption holds true in real world environment because, a wave file containing traffic noise is more likely to contain the similar kind of traffic noise through out the file.

The difference between Figure 1 and Figure 2 is clear. If it's background noise, the number of unique phones detected by the trained classifier is heavily skewed towards very less number of phones. Note that these phonemes excludes the sil phoneme. On the other hand the total number of unique phonemes are high and follow a somewhat Gaussian distribu- tion for the speech files. Figure 1 is the aggregate of all the unique phones for all the different Signal to Noise Ratios.

2) Number of Unique Phone vs SNR: It's interesting to note the histogram of number of unique phones for different Signal to Noise Ratios for background noise. This plots helps in understand, how a trained classifier can be used as a pre- processing tool to identify wave files of background noises. Note that the unique phones does not include the silence phone. The Figure 1 is actually a combination of Figure 3,4,5 and 6. Each of these histograms are plotted from 600 files.

For SNR of 0dB, which is most noisy of all the background noises, where the noise characteristics may vary across the file, or can be treated as quazi stationary noise, the number of phones are wide spread. This is due to the tendency of the neural network classifier to output more unique phones for the noisy background files. If the noise spectrogram is truly time varying, then number of distinct phonemes would be more, thus approaching that of a speech file, rendering it unable to differentiate from speech file. It's worth to note that the distribution tends to be more of Gaussian, which is reasonable. For 6dB, classification of frames are more or less same across the number of phones. It follows a similar trend of 0dB, but more flattened, where more phones are recognized. 6dB is a transition point between 0dB and the higher SNR's.

As it moves to 12dB, its clearly noticeable that the classification of the noisy files tend to be distributed across very less number of phones. As the SNR goes high, the stationarity of signal remains more or less the same. A skewness in the number of distinct phonemes can be clearly

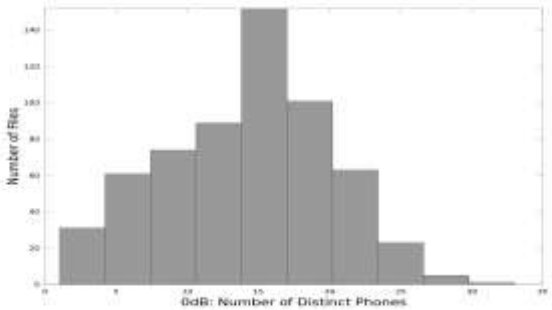


Fig. 3. Background Noise - Different Phones - 0dB

noticed. Around 170 files out of 600 have less than 5 phonemes recognized, which clearly shows the relationship between Signal To Noise ration of the background noise files and the frame classifier's accuracy.For 18dB, it's very much skewed towards very few phonemes, as the signal is stationary across the wave files. Around 420 files are have less than 5 phones, which clearly establishes the trend.

From the above observations it's obvious that the number of distinct phonemes detected in a wave file can be used as a reliable measure of the background noise. Even if we compare between the speech files and background noise at 0dB, ie between Figure 1 and Figure 3, the distribution of number of distinct phonemes for speech files are around 30, while that of the noise files at 0dB is around 15 phones. Note that this holds true for any classifier trained from decent amount of data, to predict the phones. Later in this paper, we discuss about the thresholds on the number of distinct phones vs false positives. Another way of approaching this problem is to include a class label called noise, and get sufficient noise data to train the noise class. We presume that if such a label is available, exclusively for the noise data, most of the distinct phonemes detected in the background noise files, would have been classified as noise class, thus increasing the precision of other phonemes. But for this approach to work effectively noise data has to be manually labelled with high precision, which is a daunting task in itself. Moreover different type of noise warrants different type of noise phonemes, which makes it a much more involved task. Below is a table which can be used to determine the rule for speech vs background noise.

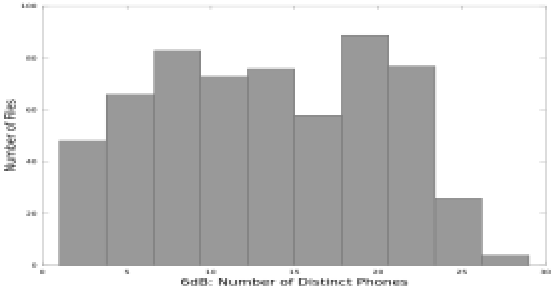


Fig. 4. Background Noise - Different Phones - 6dB

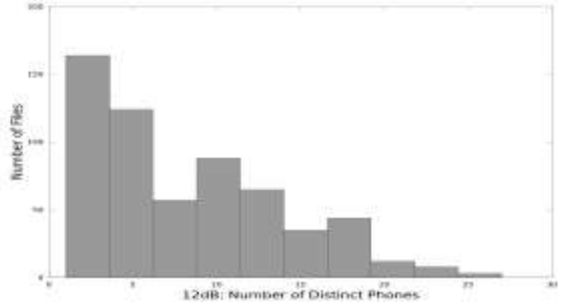


Fig. 5. Background Noise - Different Phones - 12dB

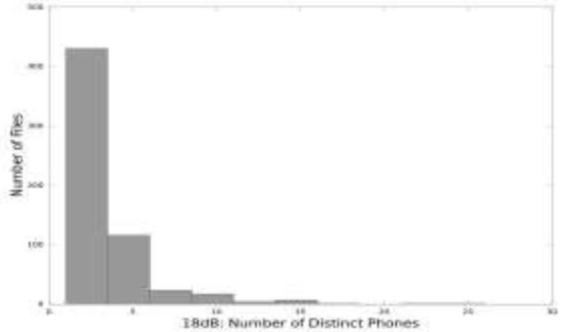
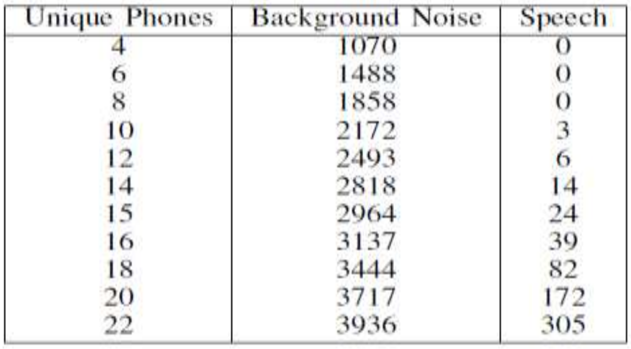


Fig. 6. Background Noise - Different Phones - 18dB



Unique Phones	Background Noise	Speech
4	1070	0
6	1488	0
8	1858	0
10	2172	3
12	2493	6
14	2818	14
15	2964	24
16	3137	39
18	3444	82
20	3717	172
22	3936	305

Table1: Unique Phones vs Noise and Speech

This table shows the at various number of distinct phones how many files are classified as noise and speech. Based on this a simple rule for background noise detection can be,

- If the number of distinct phones < 10, label the wave file as background noise

This simple rule alone is sufficient for a reasonable separation between background noise and speech files.

3) Coverage of Number of Distinct Phones for Noise: An interesting observation, for the noisy files is the positive correlation between the number of distinct phonemes and

the percentage coverage of those phonemes in the wave file. Note that silence phoneme is excluded from this analysis.

This plot shows the number of distinct phones vs coverage of those phones in percentage, for noisy background files of 0dB. A strong correlation can be observed in this plot. The correlation between the variables is calculated to be 0.7, which is significant. It means that as the number of distinct phonemes increases, so is the percentage coverage of those phonemes in the wave file. An implicit point is that,though more frames

are getting classified into a single phone, frames may or may not be contiguous. If a set of contiguous frames are classified into single phone, and the number of distinct phones to which different chunks of wave file, getting classified is more, then it signifies the non stationarity of the noise. In some sense, it also signifies the fact that the underlying neural network is consistent to noisy frames of same type, in the sense that they are classified to same phoneme. On the other hand, if the correlation between, number of distinct phonemes detected, and the percentage coverage of those phoneme, is negative, it means that the underlying neural network classifier doesn't classify the frames with similar noise characteristics, into same phoneme, thus casting doubt over the consistency of the classifier.

4) Coverage of Number of Distinct Phones for Speech: The following plot shows the coverage of distinct phonemes in the wave files. It's clearly noticeable, the difference between the coverage of different phones in the case of noise and speech. In the speech files the correlation is 0.5 only as opposed to a 0.7 in background noise files

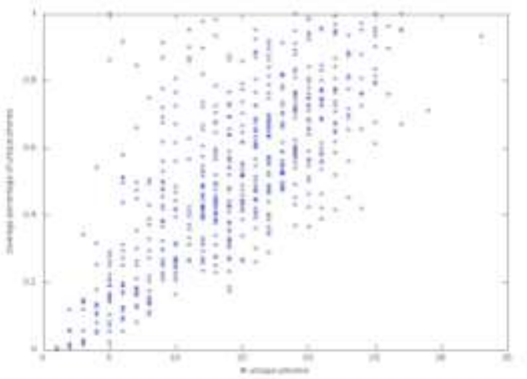


Fig. 7. Background Noise - Number of Distinct Phones vs Coverage Percentage

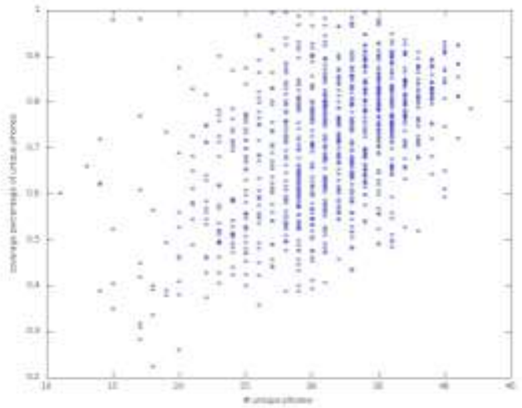


Fig. 8. Speech - Number of Distinct Phones vs Coverage Percentage

With this insight that, there is a significant difference in correlation of number of distinct phones and the coverage of those phones in the wave file, for noisy and speech files, we create a simple model for predicting the phoneme coverage percentage from the number of distinct phones. For that a straight line is fit between the independent variable, the number of distinct phones, and the dependant variable, the coverage percentage of those phones, for both of the noisy files as well as speech files. For each value of unique number of phones, the phoneme coverage percentage of multiple files are averaged, to get a composite phoneme coverage percentage. The equation of straight line is of the form y = mx + b, where y is the number is the percentage coverage of the distinct phonemes and x is the number of distinct phones. Least Squares fit gives the following equations for m and b

$$m = \frac{N\sum_{i=1}^{N} x_i y_i - \sum_{i=1}^{N} x_i \sum_{i=1}^{N} y_i}{(N\sum_{i=1}^{N} x_i^2) - (\sum_{i=1}^{N} x_i)^2}$$

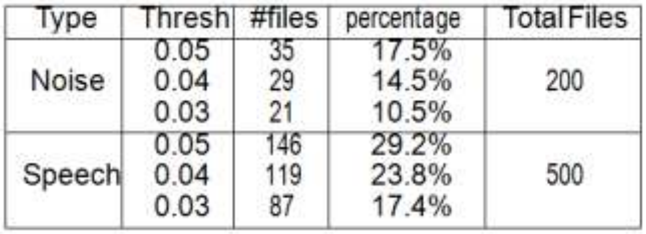
$$b = \frac{\sum_{i=1}^{N} y_i - m\sum_{i=1}^{N} x_i}{N}$$

Using Least squares linear regression on both speech and noise data, we got

Speech Files : $y = 0.015x + 0.19$
Noise Files : $y = 0.03x + 0.06$

For noisy data only the files of 0dB SNR, which are 400 files, are considered, and for speech data 4500 files are used, to fit the straight line. For noise files, all the files which are of SnR other than of 0dB are discarded because any results derived from 0dB data will subsume all the other SnR data.

5) Analysis of Straight Line Fit:



Type	Thresh	#files	percentage	Total Files
Noise	0.05	35	17.5%	200
	0.04	29	14.5%	
	0.03	21	10.5%	
Speech	0.05	146	29.2%	500
	0.04	119	23.8%	
	0.03	87	17.4%	

Table2: Linear Regression Fit - Testing; Noise vs Speech

Above table shows the result of the rule based speech vs noise classifier system. 200 files are used for noise testing and

500 files for speech testing. All of the testing data for speech has number of unique phones greater than 15. It is apparent from the results that speech classification is more robust than that of noisy file detection, which allows it to be used more to detect speech files.

VII. CONCLUSION AND FUTURE WORK

We described an extremely simple and effective way of determining whether a wave file has to be sent to the speech recognizer or not, depending upon the amount of speech/noise in the file. A neural network is trained to detect phonemes, is used as a tool to get statistics of certain measures, which is instructive in differentiating speech files and noise files. A set of plots which shows the distribution of number of distinct phones in the background noises vs speech files is discussed. We further shows the distribution of number of unique phonemes for background noise for different Signal to Noise Ratios. A simple rule based mechanism is derived which predicts the amount of speech content in the file. Another rule derived from the characteristics of the noisy data is also discussed. The rule mechanism is designed in such a manner which abstains from predicting speech/noise in certain conditions. The proposed technique works because of the consis- tency of the multi layer perceptron in discriminating different phonemes in speech. The precision and recall of different phonemes is ignored in the present context. This information could be used in fine tuning the rule based mechanism. In fact threshold could be assigned for each of the phonemes to make it more robust in real world scenarios. Another approach is using multiple MLPs for phoneme detection and combining the detectors in a data driven manner to better predict speech vs noise. Multiple MLPs can be as simple as using different architecture for different MLPs but trained on same data. This approach is promising because different architectures have different strengths in phoneme detection.

This approach can be further extended by finding a more parameters that changes between speech and noise, and incorporate them into the model. More spectral based features, which are discriminatory could be used for a better discrimination. The distribution of length of the specific phoneme chunks for specific noises could be used as another robust speech detecting mechanism, which we aim to exploit further

REFERENCES

[1] Michael L. Seltzer, Dong Yu,Yongqiang Wang, “An Investigation of
Deep Neural Networks for Noise Robust Speech Recognition”

[2] Ananya Misra, “ NonSpeech Segmentation in Web Videos”,

[3] Nima Mesgarani, Samuel Thomas, Hynek Hermansky, “Adaptive
Stream Fusion in Multistream Recognition of Speech”

[4] E. Verteletskaya, K. Sakhnov, “Voice Activity Detection for Speech
Enhancement Application”,

[5] Reinhard Sonnleitner, Bernhard Niedermayer, Gerhard Widmer, Jan Schluter, “A Simple and Effective Spectral Feature for Speech Detection in Mixed Audio Signal”,