A PROBABILISTIC APPROACH FOR DETECTING SPEECH FILE

Punnoose A K

Flare Speech Systems, India

ABSTRACT

This paper discuss an approach to detect whether a wave file contains speech or not. A frame classifier is trained to classify frames to phones. The inherent biases of the frame classifier, in terms of various aspects of recognition, is captured in terms of probability distributions. Using the distributions of speech and noise, an approach is presented, which scores wave file for the presence or absence of speech. Relevant databases are used to test the detection rate of this approach.

KEYWORDS

Noise Robustness, Neural Networks, Interactive Voice Response Systems, Confidence Scoring

1. 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

Natarajan Meghanathan et al. (Eds) : SAI, ICAITA, CSITA, ISPR, Signal - 2018
pp. 07–18, 2018. © CS & IT-CSCP 2018 DOI : 10.5121/csit.2018.81002

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 a frame classifier, for noise and speech, and presents a couple of probabilistic models to score the presence of speech in a wave file.

2. PROBLEM DEFINITION

Given the frame classifier output of a wave file, which is a sequence of phones, each corresponding to a frame, derive a confidence score which can indicate whether a file contains speech or not.

3. 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 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.

4. APPROACH OUTLINE

First a neural network is trained to classify frames to phones. Frames correspond to the usual 25ms of time with a 15ms overlap between successive frames. Context independent phones are used as the labels. Phones are preferred as labels as opposed to subphones. This is because a subphone based forced aligner doesn't align the boundaries well, thus affecting the quality of frame classifier. Assuming a decent level of accuracy, we capture the inherent classification biases of the frame classifier, in terms of phone duration, and in the distribution of softmax probabilities, for noise and speech separately in probability distributions.

Probability Distributions on phone chunk durations and softmax probabilities are defined, for noise and speech. Simple rules are derived from these distributions, to classify files into speech/noise. The rules are made to decrease the false positives as much as possible at the expense of false negatives.

5. DETAILED APPROACH & ANALYSIS

A multi layer perceptron(MLP) is trained to predict phones, with softmax layer at the ouput. For a given feature frame at the input, the MLP outputs a probability vector. The phone which has the maximum value in the probability vector is treated as the detected phone. The classified phone for a frame is also termed the top phone for that frame. A set of continuous frames with the same phone detected is regarded as a phone chunk. Size of the phone chunk is the number of frames in the phone chunk.

Common Notations:

- q : A phone in the phone set, Q.
- q_j : Phone chunk q of size j
- *N*: denotes noise.
- *S*: denotes speech.
- $C_N(q_j)$: Count of phone chunk q, of size j, in noise.
- $C_S(q_j)$: Count of phone chunk q, of size j, in speech.
- *M*: maximum chunk size.

Distribution on Phone Chunk Sizes: Fig 1 and Fig 2 plots the count of the phone /b/ for speech and noise respectively. It is clear from the plots that for noise data, chunks with higher duration are totally absent. This means that phone /b/ is resilient to the presence of noise. This motivates us to make a probability distribution on the phone chunk width, to discriminate between speech and noise.

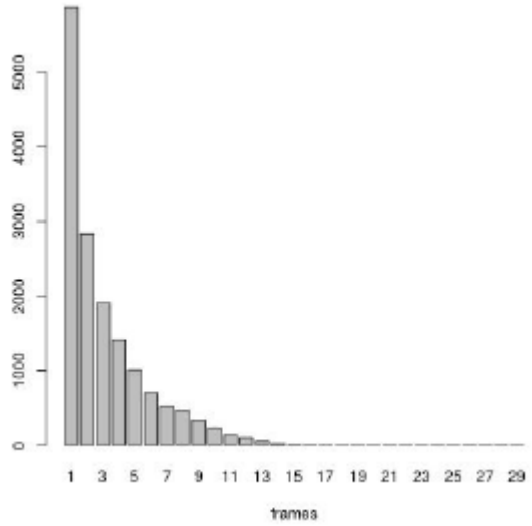

Fig. 1. Phone /b/ : Chunk width vs Count: Speech Data

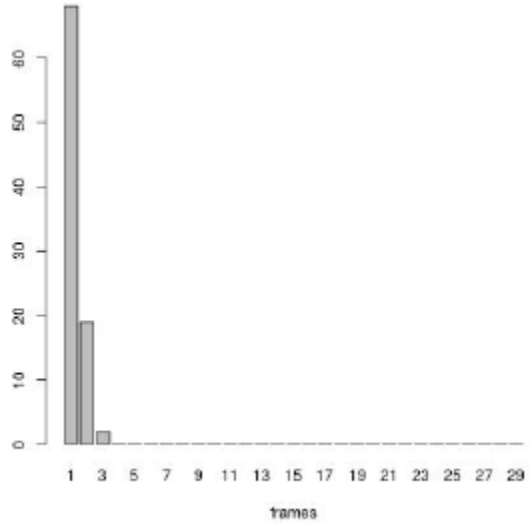

Fig. 2. Phone /b/ : Chunk width vs Count: Noise Data

Define $P_1(q_j \mid N)$ and $P_1(q_j \mid S)$ which is a probability distribution on phone chunk size, for noise and speech respectively. $P_1(q_j \mid N)$ is the probability of phone q of chunk size j, occuring in noise data.

$$P_1(q_j|N) = \frac{\sum_{j=1}^{M} C_N(q_j)}{\sum_q \sum_{j=1}^{M} C_N(q_j)} \qquad (1)$$

where, $P_1(q|N)$, is the probability of finding chunks, be whatever size, of phone q, given N. $P_1(q_j \mid q;N)$, is the probability of finding a chunk of size j, given the phone is q, in the noise data N.

Distribution on Softmax Probabilities: Fig 3 and Fig 4 plots the histogram for the phone /f/, for noise and speech data. Note that these are the instances where the frame is classified as /f/ phone. ie, /f/ is the top phone for that frame. So the data plotted here is the maximum probability of all the phones. It is clear the difference between the probabilities for noise and speech. For speech the probability is concentrated at the right end, while for noise, it is focused more around the middle.

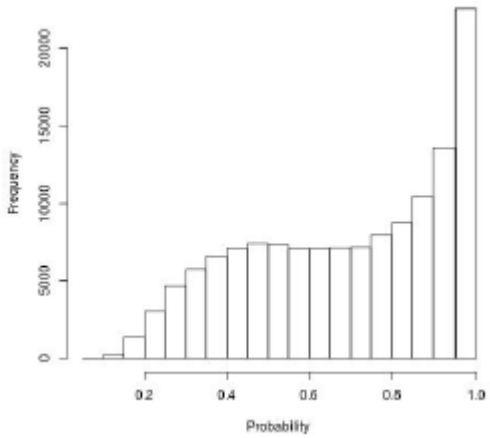

Fig. 3. Histogram of softmax probabilities of /f/, for clean data

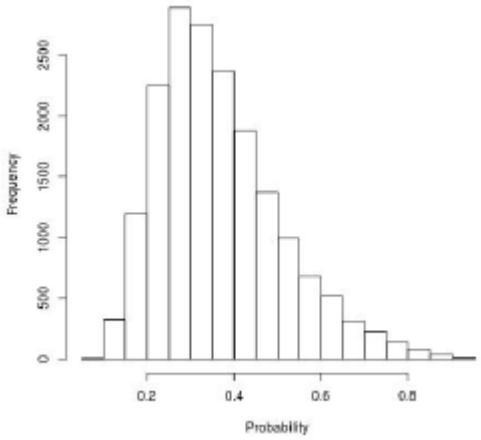

Fig. 4. Histogram of softmax probabilities of /f/, for noisy data

This serves as a valid feature to discriminate frames of speech from noise. We construct the second probability distribution on this data. Denoting p as the softmax probability of the phone, $b(p)$ gives the probability bin of p, and $C(b(p))$ is the count of instances in that probability bin.

$$P(p|N;q) = \frac{C_N(b(p))}{\sum_b C_N(b)}$$

Denoting the probability of noise, given the softmax probability of the top phone q, as $P_2(N|(p,q))$, and by using Bayes theorem,

$$P_2(N|\langle p,q\rangle) = \frac{P(p|N;q)P(N)}{P(p|N;q)P(N)+P(p|S;q)P(S)} \quad (2)$$

A. *Using the Distributions*

In equation (1) and (2), distributions are defined on phone chunk level. To make predictions in a file level, we need $P_1(N|\text{wavefile})$ and $P_2(N|\text{wavefile})$. ie, distributions defined at the file level.

1) File Level Phone Chunk Distribution :

Let $[q_j^i]$ be the phone chunk sequence for a wave file, where the superscript i is the index and j is the chunk length. Each of $q^i \in Q$, where $1 \leq i \leq X$, where X is the number of phone chunks in the wave file. Assuming each phone chunk to be independent, the probability of the wave file being noise, can be interpreted as the probability of each chunk in the chunk sequence being noise. The posterior probability can be written as,

$$P_1(N|\text{wavefile}) = P_1(N|q_{j_1}^1 \wedge N|q_{j_2}^2 \wedge ..N|q_{j_X}^X) \quad (3)$$

where $q_{j_i}^i$ is the i th chunk in the chunk sequence with the length j_i. By the independence of phone chunks

$$P_1(N|q_{j_1}^1 \wedge N|q_{j_2}^2 \wedge ..N|q_{j_X}^X) = \prod_{i=1}^{X} P_1(N|q_{j_i}^i) \quad (4)$$

where

$$P_1(N|q_{j_i}^i) = \frac{P_1(q_{j_i}^i|N)P(N)}{P_1(q_{j_i}^i|N)P(N)+P_1(q_{j_i}^i|S)P(S)}$$

by Bayes theorem. $P(S)$ and $P(N)$ are the prior probability of speech and noise respectively.

2) *File Level Softmax Probability Distribution:*

Denote the softmax probability and the associated top phone by $[\langle p^i, q^i\rangle]$, where $1 \leq i \leq Z$. Note that top phones can occur intermittently or continuously. Z is the total number of the top phones in the wave file, which is the same as the number of frames in the file. Assuming Z top phones are seen, the probability that the file is noisy is given by,

$$P_2(N|\langle p^1 q^1\rangle \wedge ... \wedge N|\langle p^Z q^Z\rangle) = \prod_{i=1}^{Z} P_2(N|\langle p^i q^i\rangle) \quad (5)$$

6. EXPERIMENTAL DETAILS & RESULTS

Experimentation is broadly divided into three stages.

1) Train a frame classifier to predict a frame into one of the phones.

2) Using the frame classifier, model the conditional distributions on phone chunk size and softmax probability of top phone, for speech and noise data.

3) Use the distributions for testing speech and noise files to see whether they can be discriminated.

A. Dataset Details

Voxforge dataset is used as the speech data and CHiME dataset is used as the noise data.

Rationale for Voxforge Data: The foremost reason for using Voxforge data is that, it is recorded in an uncontrolled environment 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 speech recognition system. This is very much against the popular notion of using a very well known database like TIMIT, which is recorded in a controlled environment, as the focus here is on real world 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 analyzing a real world speech based information access system will show that roughly only 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.

Frame Classifier Details: A MLP is trained to predict phones from speech features. Perceptual Linear PredictionCoefficients(plp) is used as feature. plp along with delta and double delta coefficients are used as the feature. Standard 41 phone set of English is used as the labels. Mini batch gradient descent is used as the training mechanism. Cross Entropy Error is used as the measure for backpropagation training. 3 hidden layers are used and weights of MLP are initialized randomly between -1 and +1. Softmax layer is used in the output layer which outputs a probability vector, given a plp frame as input.

Noise Data Details: Pure background noise from CHiME4 Dataset is used as noise data. Background noise in various environment like street, bus, etc are used. Unlike older CHiME datasets, CHiME4 is not segregated based on SNR. CHiME data is divided into 2 subsets and used in the second and third stages.

We present the results for both distributions, independently, to figure out how speech files can be separated from the noisy ones. It is to be noted that for all the three stages discussed above, three different dataset is used. For all the stages for speech, 3 different subset of Voxforge data is employed. For stage 2 and 3, for noise data, different subset of CHiME data is used.

Conditional distributions $P_1(N|q_j)$, $P_1(S|q_j)$, $P_2(N|(p^i q^i))$ and $P_2(S|(p^i q^i))$ are learned in the second stage and the posterior probabilities $P_1(N|\text{wavefile})$ and $P_1(S|\text{wavefile})$ are calculated in the testing stage. With a focus on precision results are given for true positives and false positives, for both approaches.

B. Phone Chunk Size Distribution Results

As our aim is to discriminate speech and noise files, Equation (4) can be rewritten as,

$$P_1(N|q_{j_1}^1 \wedge N|q_{j_2}^2 \wedge ..N|q_{j_X}^X) \propto \frac{1}{X}\sum_{i=1}^{X} ln(P_1(N|q_{j_i}^i))$$

By (3), the posterior can be written as

$$P_1(N|\text{wavefile}) \propto \frac{1}{X}\sum_{i=1}^{X} ln(P_1(N|q_{j_i}^i)) \qquad (6)$$

This is mainly done to avoid the underflow, while using equation (2). And the results are averaged, to make sure the same scale for otherwise longer files. Phones whose counts falls below a threshold, in the calculation of the conditional densities are excluded from the analysis.

Fig 5, plots the results of posterior probabilities, given the speech data. The posteriors are approximated using (7). The posteriors from each speech wavefile is plotted as histograms. Green histogram represents the $P_1(S|\text{speech})$ and blue histogram represents the $P_1(N|\text{speech})$. It is evident from the plot that both the posteriors are clearly separated, given the input speech data.

As the plots are in log scale, the values closer to 0 means more probable. For speech data, it is seen from the plot that the green histogram which is the speech posterior is closer to the 0, than the noise posterior. Also it is evident from the plot that the posterior of noise for speech data is very wide spread than compared to that of speech posterior, which is narrowly concentrated.

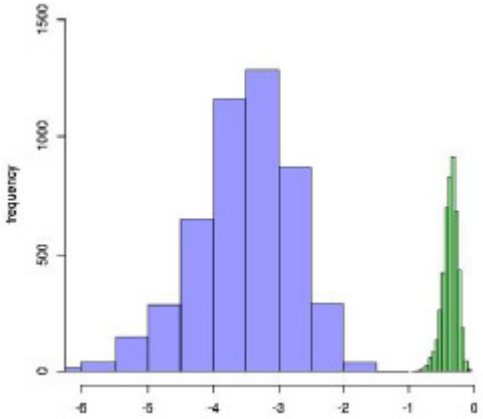

Fig. 5. Speech and Noise Posterior for Speech Data

Fig 6, plots histogram of speech and noise posteriors, given the noise data as input. Green represents the $P_1(S|\text{noise})$, and blue the $P_1(N|\text{noise})$. Both of the histograms are evenly spread in the log domain.

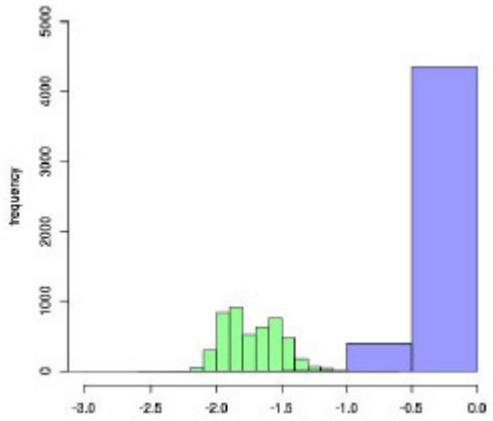

Fig. 6. Speech and Noise Posterior for Noise Data

As the focus is more on speech file detection, it is worth looking at the false positives and true positives. Fig 7, plots the $P_1(S|\text{noise})$ as blue and $P_1(S|\text{speech})$ as green histogram. It is clear that using appropriate threshold on average log posterior values, the speech and noise can be easily separated.

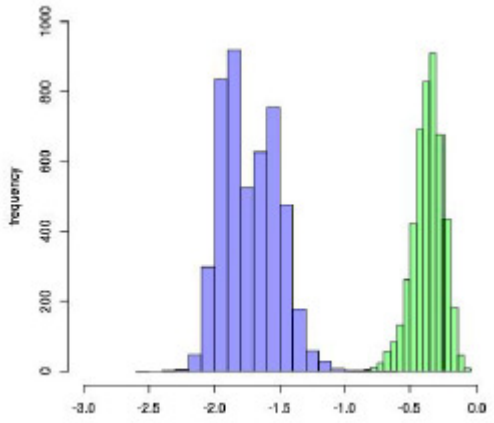

Fig. 7. Speech Posteriors for Noise and Speech Data

Table 1. Results

Threshold τ	True Positives	False Positives
$>$ -0.9	4769	2
$>$ -1	4770	4
$>$ -1.1	4770	10
$>$ -1.2	4770	39

Table 1 shows the true positives and the false positives for different threshold values of average log posteriors.

C. *Softmax Probability Distribution Results*

As in (7), instead of the product of probabilities, we approximate it using the average posterior probability of top phones, for a wave file. ie,

$$P_2(N|\text{wavefile}) \propto \frac{1}{Z}\sum_{i=1}^{Z} ln(P_2(N|\langle p^i, q^i\rangle))$$

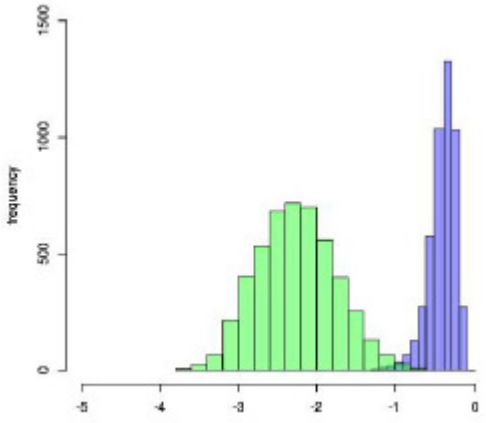

Fig. 8. Speech and Noise Posteriors for Speech Data

Fig 8, plots the results of posterior probabilities, given the speech data. The posteriors are approximated using (7). Blue histogram represents the P_2(S|speech) and green histogram represents the P_2(N|speech).

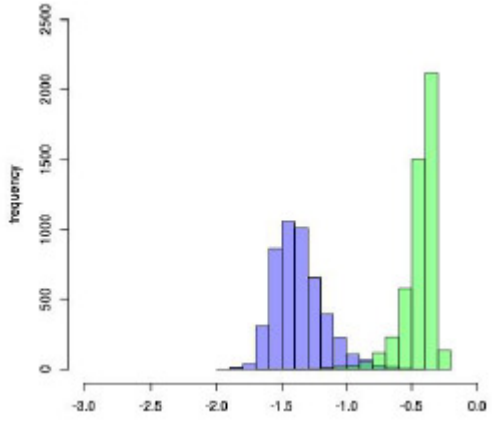

Fig. 9. Speech and Noise Posteriors for Noise Data

Fig 9, plots histogram of speech and noise posteriors, given the noise data as input. Green represents the P_2(*N*|noise), and blue the P_2(*S*|noise).

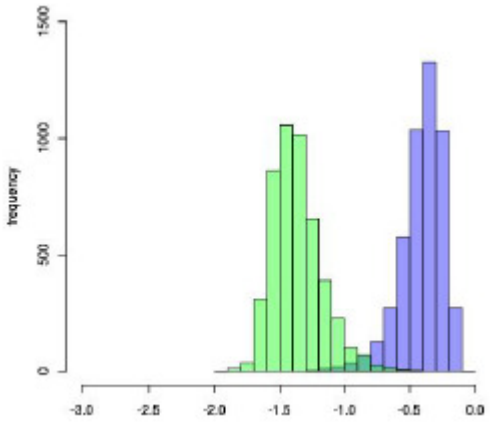

Fig. 10. Speech Posteriors for Speech and Noise Data

Focusing more on true positives and false positives, Fig 10, plots the $P_2(S|\text{speech})$ as blue and $P_2(S|\text{noise})$ as green histogram.

Table 2. Results

Threshold τ	True Positives	False Positives
> -1.0	4741	218
> -0.9	4706	116
> -0.8	4641	47
> -0.7	4516	22

Table 2, shows the true positives and false positives for different value of threshold posteriors.

7. CONCLUSION AND FUTURE WORK

A new approach for detecting whether a wave file consists of speech is presented. A frame classifier is first trained to predict the phone, given a frame. The characteristics of the frame classifier is codified with 2 probability distributions, one on phone chunk size, and another one on softmax probability associated with a phone, given a frame. Posterior distributions are approximated in log domain to reduce the dynamic range of scores. Results are shown separately, to show the effectiveness of both of the approach independently.

In future, we plan to use more spectrum derived features, in conjunction with frame level features to increase the overall accuracy of this approach. Spectrum level features provides vital clues, which could be missed by any parameterized features like mfcc or plp, especially for noise robustness.

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”