SMM19, Workshop on Speech, Music and Mind 2019
14 September 2019, Vienna, Austria

New Features for Speech Activity Detection

Punnoose A K

Flare Speech Systems
Bangalore, India
punnoose@flarespeech.com

Abstract

This paper discusses two new features for speech activity detection(SAD), using a multi-layer perceptron(mlp) trained to predict phoneme from acoustic features. The first feature is based on the difference between speech and noise histogram of certain phonemes. A scoring mechanism is formulated to score the softmax probabilities of the frames of a phoneme. The second feature is based on the correlation between softmax probabilities of the edge frames for certain phoneme transitions. A probabilistic approach is formulated to score the phoneme transition. Relevant datasets are used to prove the robustness of the proposed features in terms of speech activity detection.
Index Terms: Speech Activity Detection, Noise Robustness, Feature Engineering

1. Introduction

One of the most crucial requirements of a speech 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 gets 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 results while testing in a 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, then 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 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.

Many recognition engines use mlp for speech decoding in multiple forms. Mlp posteriors can be used in tandem or hybrid architecture [1]. In a very simplistic way, mlp posterior themselves could be used as the recognition feature. In such cases, frame to phoneme mapping is needed. The advantage with such an approach is that frame level decision can be taken depending on the end user application. And the whole hidden Markov modeling(HMM) decoding can be skipped. For isolated word decoding, mlp posteriors are a promising approach.

The issue with the direct frame to phoneme to sentence decoder approach is the presence of many out of order phonemes in the decoded phoneme sequence. A single frame of stop phoneme in the middle of a vowel segment, or vice-versa confuses the sentence decoder. Context dependent phonemes make the matter worse. This paper explores two approaches to assign a confidence score for certain phonemes using mlp decoded parameters.

2. 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 [2, 3]. 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 [4].

Autocorrelation functions and its various derivatives have been used extensively for voice activity detection. Subband decomposition and suppressing certain subbands based on stationarity assumptions on autocorrelation function is used for robust voice activity detection(VAD) [5]. Autocorrelation derived features like harmonicity, clarity and periodicity provide more speech-like characteristics. Pitch continuity in speech has also been exploited for robust speech activity detection [6]. For highly degraded channels, GABOR features along with autocorrelation derived features are also used [7]. Modulation frequency is also used in conjunction with harmonicity for VAD [8].

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 [9].

Another very common method is to use mel frequency cepstral features with classifiers like SVMs to predict speech regions [10]. Derived spectral features like low short-time energy ratio, high zero-crossing rate ratio, line spectral pairs, spectral flux, spectral centroid, spectral rolloff, ratio of magnitude in speech band, top peaks and ratio of magnitude under top peaks are also used to predict speech/non-speech regions [11]. Mlp posteriors have been also used in various noise robustness based tasks in various forms [12,13]. Various neural network architectures like recurrent neural networks(RNN) [14], long short term memory networks(LSTM) [15], deep neural networks(DNN) [16] has been used for speech activity detection.

The rest of the paper is organized as follows. Two new features are introduced and characteristics of these features for

10.21437/SMM.2019-6

noise and speech are explored. Appropriate scoring mechanisms are used to score the features. And finally, the features are benchmarked with standard real-world datasets.

3. Approach & Analysis

An mlp is trained to predict the phonemes, given the plp feature as input. Softmax layer is used as the output layer, and cross entropy is used as the error measure. Given a window of 9 plp frames at the input, mlp outputs a probability vector, where each component corresponds to a phoneme. The phoneme with the highest softmax probability is the classified phoneme for that frame and is labeled the top phoneme for the frame. In this paper, the probability associated with the top phoneme is merely treated as a score which could be used to separate speech and noise. Hence, we use softmax probability and softmax score synonymously.

Voxforge dataset is used to train the baseline mlp. As the Voxforge data is recorded in an unconstrained environment, it is close to real-world conditions. For subsequent model development, other datasets are used. Background noise subset of the CHiME dataset [17] is used as the noise data. Noisy speech dataset NOIZEUS [18] is used as the speech data.

3.0.1. Overall Approach

Run plp frame window feature, from a subset of speech and noise data, through the baseline trained mlp to get a set of training probability vectors. Extract features from these intermediate training probability vectors to train the models for specific top phonemes. To benchmark the approaches, pass the plp window from the testing speech and noise datasets, through the same baseline mlp to get the testing probability vectors. Use the required features from the testing probability vectors to score the models.

As the mlp is trained on one dataset and is used as a tool to make models for speech and noise data, any positive experimental result shows the robustness of the proposed features. The features explored are (i) Score from softmax histogram difference between speech and noise. (ii) Score from softmax probability at certain phoneme transitions.

3.1. Score from Softmax Histogram Difference

Figure 1 shows the softmax probability of the top phoneme /r/ for noise and speech data. The histogram of speech and noise is different but very close. First, we run a hypothesis testing to show that the top probability is indeed different for speech and noise, and can be used for speech vs noise separation.

3.1.1. Hypothesis Testing for Means

As histograms overlap, a student's t-test for independent means is performed to figure out whether the difference between 2 samples is statistically significant. Assuming unknown population variance, the null and alternative hypothesis of the student's t-test are

$$H_0 : \mu_s = \mu_n$$
$$H_A : \mu_s \neq \mu_n$$

where, μ_s and μ_n are the mean probabilities for speech and noise. A $p\text{-}val < 0.02$ and $t = 54.11$, suggests rejection of the null hypothesis and implies that the speech and noise top softmax scores are indeed statistically different. This implies that for the speech, top phonemes are predicted with more

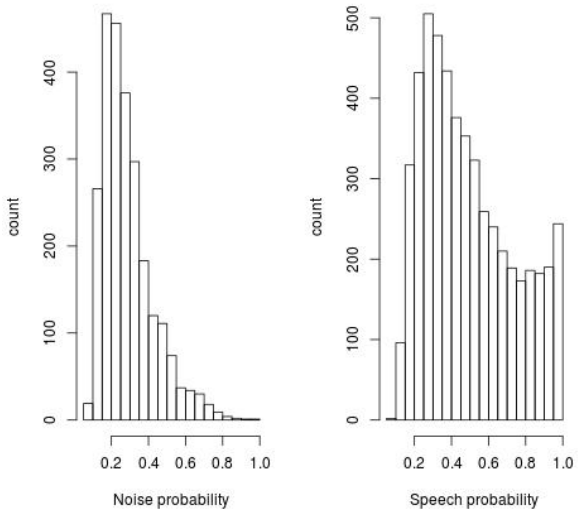

Figure 1: *Histogram for phoneme /r/ for speech and noise*

confidence by the mlp frame classifier as compared to that of noise. In other words, for noise data, a wrongly predicted top phoneme is very close to the second predicted phoneme for the same frame.

3.1.2. Scoring Method

The insight that histogram of the top phoneme is different for speech and noise, can be used to discriminate between speech and noise. A simple function is defined, which scores the softmax probability p of a frame belonging to a phoneme, being generated from speech.

$$d = s_p - n_p \tag{1}$$
$$\hat{d} = d + \tau d \tag{2}$$
$$c = e^{-(1-\hat{d})} \tag{3}$$

where, s_p and n_p is the probability of the softmax score p, calculated from speech and noise histograms respectively. s_p and n_p are defined as,

$$s_p = \frac{C_s[p]}{T_s} \tag{4}$$
$$n_p = \frac{C_n[p]}{T_n} \tag{5}$$

where, $C_s[p]$ is the count of softmax scores in the speech histogram bin where the softmax score p falls, for a particular phoneme. T_s is the total number of softmax score points for the speech data for the same phoneme. $C_n[p]$ and T_n are the respective counts for noise data. Note that the counts $C_s[\ ]$, $C_s[\ ]$, T_s, T_n are computed from the training datasets, where p is an incoming test softmax probability whose score is to be calculated.

In Eq. (1), the probability of a softmax score being generated from the speech is modified by subtracting the probability of the same softmax score being generated from noise. In short,

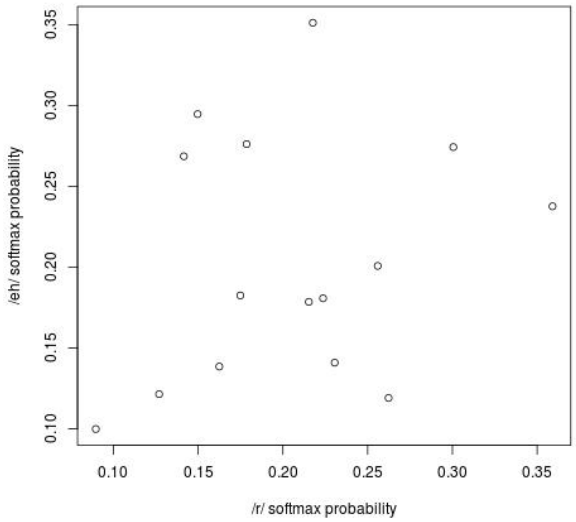

Figure 2: *Softmax prob of /r/ against /eh/ on noise data*

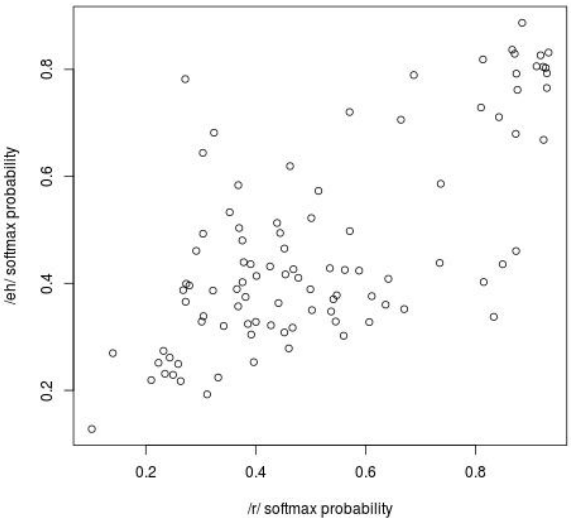

Figure 3: *Softmax prob of /r/ against /eh/ on speech data*

if the probability of a softmax score calculated from speech and noise histogram is the same, then the final score should be minimum. As the difference in probability is minimal between speech and noise histograms, it needs to be amplified in proportion to the difference itself, as in Eq. (2). τ is the amplification factor. Finally, for a softmax probability p, a score is given in Eq. (3). The file level score f_1 is given by

$$f_1 = \frac{1}{N} \sum_i c_i \tag{6}$$

where, N is the number of the occurrence of frames of phoneme p. Note that f_1 is dominated by the frames with a large softmax score difference in Eq. (1).

3.2. Score from Softmax Probability at Phoneme Transition

During a phoneme transition, the top softmax probability of the frames in transition, i.e., the last frame of the first phoneme and the first frame of the phoneme into which the transition is taking place, may exhibit some correlation, for some phoneme pairs. Figure 2 & 3 plots the softmax probability of the edge frames of phoneme transition from top phoneme /r/ to /eh/, for noise and speech respectively.

It is evident from the plots that, for speech, there is a correlation between the softmax probability of the last /r/ frame and the first /eh/ frame in the transition. A measured correlation coefficient $r = 0.61$ for the data in Figure 3 is sufficient enough to exploit this property for speech and noise separation. Note that the range of values of Figure 2 and 3 are different. For noise data, there are very few data points with x-values > 0.35. This warrants us to limit the x values of the speech data, to 0.35. Still there exists a correlation of $r = 0.41$, for curtailed speech data, as opposed to a correlation of $r = 0.21$, for noise data.

3.2.1. Scoring Method

We define an unnormalized joint density function on softmax probability pair (p_a, p_b) for the edge frames in phoneme transition (a, b), from only the speech data.

$$P_S(p_a, p_b) = \frac{1 + C_s\left[p_a, p_b\right]}{v^2 + T_s} \tag{7}$$

where, T_s is the total count of all points in (a, b) space from speech data. $\left[p_a, p_b\right]$ is the probability bin in which the (p_a, p_b) falls and $C_s\left[p_a, p_b\right]$ is the count of the bin $\left[p_a, p_b\right]$. Subscript s denotes speech data. v is the number of bins into which the probability space $[0, 1]$ is divided. Assume x and y space is divided into equal number of bins. Eq. (7) is the Laplace smoothing derived from language modeling. Note that the counts $C_s\left[p_a, p_b\right]$ and T_s are from the speech training dataset.

Given an incoming phoneme transition pair (a, b) and the associated softmax probability pair (p_a, p_b), to calculate the score for the probability pair, use p_a as the independant variable and calculate the conditional expectation value $\hat{p_b}$ using the joint pdf.

$$\begin{aligned} \hat{p_b} &= E[y|x = p_a] \\ &= \sum_y y P_s(y|x = p_a) \\ &= \sum_y y \frac{P_s(p_a, y)}{P_s(x = p_a)} \end{aligned} \tag{8}$$

As the y space is divided into intervals $[y_i, y_{i+1}]$ of duration $\frac{1}{V}$, the expected value $\hat{p_b}$ can be written as,

$$\hat{p_b} = \sum_i m_i \frac{P_s(p_a, m_i)}{P_s(x = p_a)} \tag{9}$$

where,

$$m_i = \frac{y_i + y_{i+1}}{2}$$

Finally, the confidence score of observing a phoneme pair (a, b) with softmax probability (p_a, p_b) is given by

$$g = \begin{cases} e^{-(p_b - \delta_r)} & \delta_r < p_b \\ e^{-(\delta_l - p_b)} & \delta_l > p_b \\ 1 & \delta_l \leq p_b \leq \delta_r \end{cases} \tag{10}$$

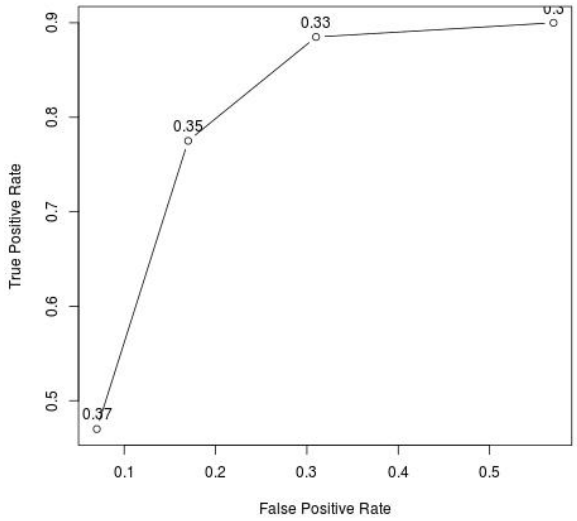

Figure 4: *ROC for* f_1

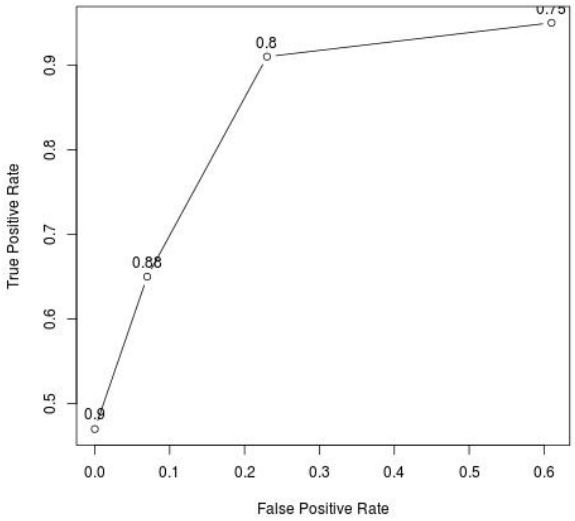

Figure 5: *ROC for* f_2

where,

$$\delta_r = \hat{p_b} + \frac{1}{2V} \tag{11}$$

$$\delta_l = \hat{p_b} - \frac{1}{2V} \tag{12}$$

Eq. (10) is an exponential function which penalizes the absolute difference between the observed p_b and predicted $\hat{p_b}$. There are 2 thresholds, δ_l and δ_r, which acts as an interval to $\hat{p_b}$. If the observed p_b falls in the expected interval $[\delta_l, \delta_r]$, then it is treated as a correct prediction. The file level score f_2 is given by

$$f_2 = \frac{1}{M} \sum_i g_i \tag{13}$$

where, M is the number of (a, b) transitions in the file.

Eq. (7) is not directly used to predict the joint score g, given an incoming testing pair (p_a, p_b), because of the following reason. If all the training data points in (a, b) space are distributed evenly across the joint probability space $[0\ 1]^2$, then for a given p_a, Eq. (7) outputs a not so insignificant density value of p_b. On the other hand, for a widely and evenly distributed p_b, Eq. (10) gives a low score, which is ideal from a precision point of view. Eq. (10) outputs a high value, if the observed p_b is close to the expected value $\hat{p_b}$.

4. Experimental Results

The difference between features f_1 and f_2 is that, for f_1 model construction, noise and speech histogram is required, whereas, for f_2, only speech data is required to build the joint probability distribution. Models(Counts) for f_1 and f_2 are computed from the training subset of both speech and noise data. The features are tested on testing subset, for speech activity detection at the recording level.

f_1 is calculated for the phoneme /r/ with amplification factor $\tau = 1.5$. Figure 4 shows the receiver operating characteristics(ROC) curve for f_1 for various thresholds. Similarly f_2 is computed for the phoneme transition (/r/,/eh/) with $v = 20$, i.e, the probability space is divided into intervals of size 0.05. Figure 5 shows the ROC curve for f_2.

It is clear from the ROC curves that this approach can be used as a reliable add-on to any speech activity detection algorithms. One apparent issue with this approach is the possible scarcity of the frames of a chosen phoneme for f_1 and the frames in phoneme transitions for f_2 calculation. Thus, these approaches are more applicable in continuous speech recognition, as opposed to isolated word recognition.

5. Conclusion and Future Work

Two approaches for speech activity detection, one based on the difference in softmax scores of a phoneme for speech and noise data and another based on the softmax score of edge frames in a phoneme transition, are presented. One concern in using these approaches is probably the scarcity of sufficient frames of a particular phoneme for f_1 and phoneme transitions for f_2. This suggests the usage of these techniques as an add-on to existing speech activity detection approaches. Relevant datasets are used to prove the robustness of the scoring mechanisms.

In the experiments, features derived from a single phoneme and phoneme transition are discussed. The idea is extendible equally to any number of phonemes or phoneme transitions. Context-dependent phoneme transitions may offer more valuable insights to detect speech activity. Presently the experimentation is done at the recording level, but the approach is equally extendible to small chunks in recordings. It is also possible to combine the features to make a composite recording level score for speech activity detection.

Another possible improvement in the calculation of f_2, in particular, is to use frames other than just the edge frames for both the phonemes in transition. The trajectory of the softmax scores of all the frames may provide clues in differentiating speech from noise.

6. References

[1] Herve A. Bourlard and Nelson Morgan. 1993. "Connectionist Speech Recognition: A Hybrid Approach." Kluwer Academic Publishers, Norwell, MA, USA.

[2] Shi-wen Deng & Jiqing Han "Statistical voice activity detection based on sparse representation over learned dictionary," In: Proc. of Digital Signal Processing 2013, vol. 23, pp. 1228–1232

[3] Parvin Ahmadi & Mohsen Joneidi (2014). "A New Method for Voice Activity Detection Based on Sparse Representation," In: Proceedings of 7th International Congress on Image and Signal Processing, CISP 2014, pp. 878–882

[4] Peng Teng and Yunde Jia "Voice Activity Detection via Noise Reducing Using Non-Negative Sparse Coding," IEEE Signal Processing Letters, Vol. 20, Issue. 5, May 2013, pp. 475–478.

[5] 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, pp. 1970–1973

[6] Yiwen Shao & Qiguang Lin "Use of Pitch Continuity for Robust Speech Activity Detection," In: Proc. of ICASSP 2018, pp. 5534–5538

[7] M. Graciarena, A. Alwan, D. Ellis, H. Franco, L. Ferrer, J. H. Hansen, et al., "All for one:Feature combination for highly channel degraded speech activity detection," In: Proc. of INTERSPEECH 2013, pp. 709–713

[8] E. Chuangsuwanich and J. Glass, "Robust voice activity detector for real world applications using harmonicity and modulation frequency," In: Proc. of INTERSPEECH 2011, pp. 2645–2648

[9] Omid Ghahabi, Wei Zhou, Volker Fischer "A robust voice activity detection for real-time automatic speech recognition system" In: Proc. of ESSV 2018, Ulm 2018

[10] Tomi Kinnunen & Evgenia Chernenko & Marko Tuononen & Pasi Frnti & Haizhou Li (2012). "Voice Activity Detection Using MFCC Features and Support Vector Machine."

[11] A. Misra, "Speech/nonspeech segmentation in web videos," in Proceedings of INTERSPEECH, 2012, vol. 3, pp. 1975–1978

[12] Ganapathy S & Rajan P & Hermansky H, "Multi-layer perceptron based speech activity detection for speaker verification," IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2011, pp. 321–324.

[13] Kazemi, A.R. & Sobhanmanesh, Fariborz, "MLP refined posterior features for noise robust phoneme recognition." Scientia Iranica, Trans. D: Computer Science & Engineering and Electrical Engineering, Vol. 18, No. 6, pp. 1443–1449, 2011.

[14] Thad Hughes and Keir Mierle "Recurrent Neural Networks for Voice Activity Detection " ICASSP, IEEE (2013), pp. 7378–7382

[15] Boonkla, Surasak & Sertsi, Phuttapong & Chunwijitra, Vataya & Kurpukdee, Nattapong & Wutiwiwatchai, Chai. (2017). "Robust Voice Activity Detection Based on LSTM Recurrent Neural Networks and Modulation Spectrum." 10.1109/APSIPA.2017.8282048.

[16] Kang, Tae Gyoon et al (2016). "DNN-based voice activity detection with local feature shift technique," 2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, pp. 1–4.

[17] Jon Barker, Ricard Marxer, Emmanuel Vincent, and Shinji Watanabe, "The Third CHIME Speech Separation and Recognition Challenge: Analysis and Outcomes", Computer Speech and Language, 46:605-626, 2017

[18] Hu, Y. and Loizou, P. (2007). "Subjective evaluation and comparison of speech enhancement algorithms," Speech Communication, 49, 588-601.