A Probabilistic Approach for Confidence Scoring in Speech Recognition

Punnoose A K
Flare Speech Systems
Bangalore, India
punnoose@flarespeech.com

Abstract

This paper discusses a method to derive a meaningful confidence score for a speech segment at the phoneme level, using a frame classifier. Multiple functions, which capture various aspects of the frame classifier output, are first introduced. The ability of these functions to discriminate between different phonemes is shown. A probabilistic approach is formulated to combine the functions to get a meaningful confidence score, which reflects the precision of the predicted phoneme chunk. Relevant real-world datasets are used to demonstrate the effectiveness of the proposed confidence scoring mechanism.

1 Introduction

In speech recognition, it is desirable to have a confidence score which has a strong correlation with the correctness of recognition. A low confidence score should imply wrong recognition, and a high score should signal correct recognition. For this, the confidence score should be derived out of features which are not directly used or overlooked, in the speech recognition. Modern automatic speech recognition is done in a multi-level manner. The bottom level corresponds to frame recognition. Next levels are phoneme, word and sentence recognition respectively. In the sentence level, language model plays a key role in the overall word error. The relationship between the frame level accuracy and the word level accuracy follows more of an S-curve. Word accuracy increases gradually as the frame accuracy increases, then it shoots up exponentially, and then gradually slows down. An error in frame level classification can be forgiving than an error in phoneme detection, especially in the case of large vocabulary speech recognition task. Ideally, the confidence scoring should be using low level features which are raw compared to higher level features.

Confidence scoring in speech recognition has a rich literature. A general survey for confidence scoring can be found in (Schaaf and Kemp, 1997; Jiang, 2005; Rose et al., 1995). Confidence scoring was treated as a classification problem with features derived from trained acoustic and language models along with derived word level features (Huang et al., 2013; Weintraub et al., 1997; Wessel et al., 1999; J. Hazen et al., 2002). Another approach is using backward language models (Duchateau et al., 2002). A trained generic confidence scoring mechanism can be recalibrated to output a more meaningful confidence score, by taking into account the end application specific scenarios (Yu et al., 2011). Another approach used for confidence scoring is by using word lattices (Kemp and Schaaf, 1997) and N-best lists (Rueber, 1997).

Multilayer perceptrons(mlp) based posteriors (Lee et al., 2004; Wang et al., 2009; Ketabdar, 2010; Bernardis and Bourlard, 1998) has been extensively used for confidence scoring. mlp posterior based score has the benefit of being at the frame level, rather than at the phoneme level. We propose a confidence scoring mechanism at the phoneme level, using a set of new features derived from an mlp based frame classifier.

The rest of the paper is organized as follows. First, the frame classifier details and datasets used are explained. Certain measures computed from the frame classifier output is explored. Then a set of phoneme level features are derived for confidence scoring. A probabilistic confidence scoring mechanism is formulated using the features derived. And finally, the approach is benchmarked using a test dataset. This is a meta-learning approach, as the confidence scoring stage depends on the output from a trained frame classifier.

Datasets & Definitions: Voxforge data is used for all the experiments. The foremost reason for using Voxforge data is that it is recorded in an uncontrolled environment by people with different accents, mother tongue, etc. This will give the necessary variability in the data and any confidence scoring mechanism derived out of this data will be applicable to a real-world speech based information access.

The whole Voxforge dataset is divided into 3 subsets, d_1, d_2 and d_3. d_1 is used to train the frame classifier. d_2 is fed to the frame classifier to get the output dubbed as d_m, which is eventually used for making distributions and functions needed for confidence scoring. d_3 is used for benchmarking the proposed confidence scoring approach.

1.1 Frame Classifier Details

An mlp is trained to predict phonemes from speech features. Perceptual Linear Prediction Coefficients(plp) along with delta and double delta features are used. Standard English phoneme set is used as the labels. Mini-batch gradient descent is used as the training mechanism. Cross-entropy error is used as the objective for backpropagation training. 3 hidden layers are used and weights of the mlp are initialized randomly between -1 and +1. Given an input, the softmax layer outputs a probability vector, where components of the probability vector correspond to phonemes.

Given a wave file, the frame classifier outputs a sequence of probability vectors, each corresponding to a frame size of 25ms. Each component in the probability vector corresponds to a phoneme and the phoneme with the highest probability is treated as the classified phoneme. The classified phoneme is labeled as the top phoneme for that frame. Define a phoneme chunk as multiple continuous frames, classified as the same phoneme. Chunk duration of a phoneme is the number of frames in that chunk. /p/ denotes the phoneme p.

The subset d_2 is passed through the frame classifier to get a set of classified phonemes and the associated probability vectors. This act as the dataset d_m, which is used to derive a set of features for the confidence scoring. First, we discuss these features by completely disregarding the associated ground truth phoneme label. Next, we use phoneme labels to fit distributions for true positives and false positives for all the features, phoneme wise. Finally, a confidence scoring mechanism with a focus on precision is derived.

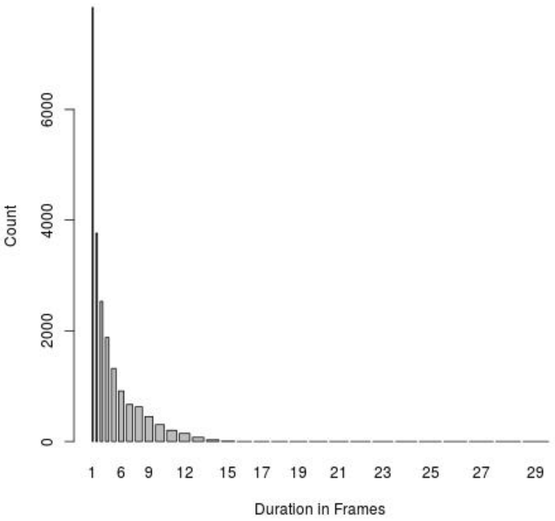

Figure 1: Phoneme /b/: Duration in frames

2 New Features

To derive new features for confidence scoring, three measures are first introduced. These are the duration of phoneme chunks, distribution of softmax probability, and the softmax probability of phoneme chunk. These measures are computed from the frame classifier output d_m, which has a sufficiently large amount of data, that allows us to treat this as a population feature. The measures are analyzed based on the top phoneme detected, framewise, by completely ignoring the ground truth phoneme labels. These measures are finally converted into phoneme chunk level features.

Duration of Phoneme Chunks: Fig 1 and 2 plots the duration of the detected phoneme chunks of /b/ and /ay/ respectively. Note that /b/ tends to have almost zero long phoneme chunks, while /ay/ has relatively plenty long chunks detected by the frame classifier. This difference in the detected duration of different phoneme chunks is significant enough to treat the phoneme chunk duration as a valid variable for confidence score prediction. Converting the counts to a simple discrete distribution,

$$g(k;p) = \frac{C(k)}{\sum_k C(k)} \tag{1}$$

where $c(k)$ is the number of chunks of size k of phoneme p.

Distribution of Softmax Probability: To understand how the highest softmax probability is

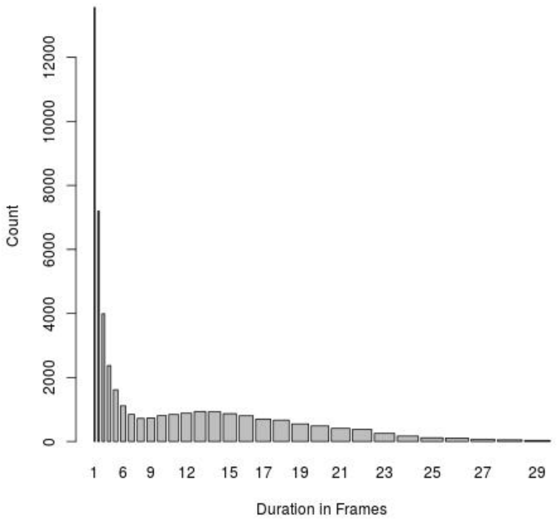

Figure 2: Phoneme /ay/: Duration in frames

distributed generally for different top phonemes, the histogram of probabilities for 2 different top phonemes are plotted for speech data.

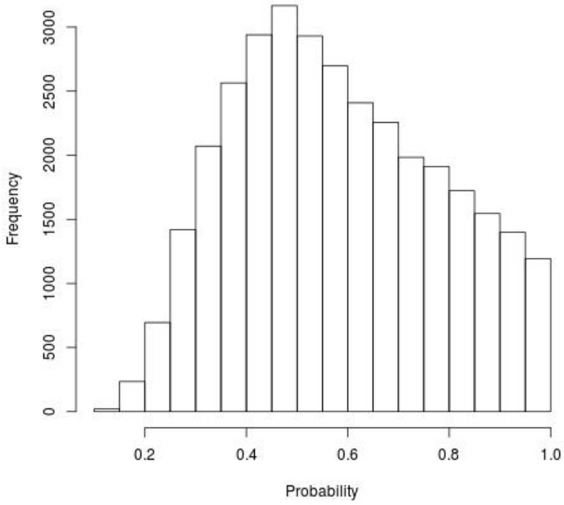

Figure 3: Softmax probability histogram for top phoneme /aa/

Fig 3 plots the histogram of the highest probabilities for top phoneme /aa/, which is a very common phoneme. It is clear from the histogram that the highest count peaks at 0.5 and as it moves up to 1, the count decreases. What it implies is that the number of instances a /aa/ phoneme is predicted with probability [0.9-1] are less than the number instances in which it is predicted with probability [0.4-0.5]. It could be due to the presence of similar sounding phonemes like /ae/, /ah/, etc so that probability gets divided. The issue with probabilities getting divided closely is that it is difficult to assign a meaningful confidence score.

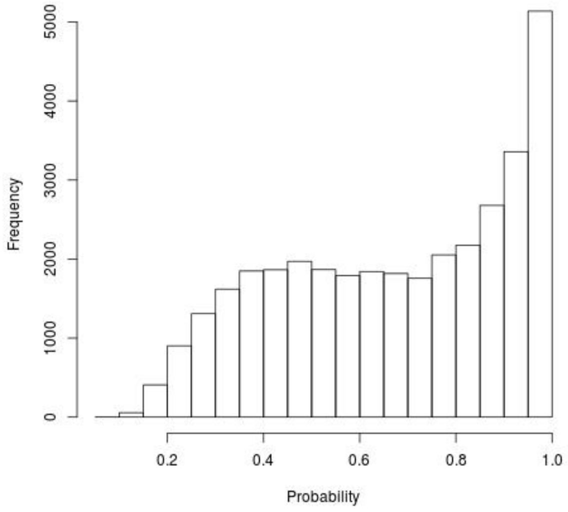

Figure 4: Softmax probability histogram for top phoneme /f/

Fig 4 plots the highest probabilities for the top phoneme /f/. It is apparent that /f/ is predicted with high probability, for most of the instances, rather than getting confused with other phonemes. From the 2 plots, it is clear that the softmax probability of the top phoneme is distributed differently for different phonemes, and could be useful in deriving a robust confidence score.

Softmax Probability of Phoneme Chunks: A related question is whether the softmax probability of a top phoneme is dependant on the neighboring same top phonemes. Fig 5 plots the mean of average softmax probability of phoneme chunks for different phoneme chunks sizes, for /f/. It is clear from the plot that as the phoneme chunk size increases, the average softmax probability also increases.

A strong correlation between the detected phoneme chunk size and mean of average softmax probabilities indicates that any confidence scoring mechanism should take into account the phoneme chunk size. As the above mentioned features are plotted from unlabelled data, it doesn't indicate whether the detected phoneme chunk is indeed correct or not. To assign a confidence score for a phoneme chunk predicted with a softmax proba-

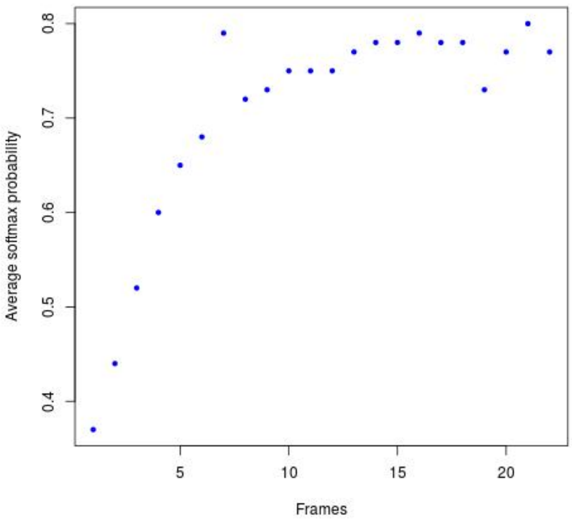

Figure 5: Average softmax probability against chunk size in frames for phoneme /f/

bility, ground truth labels of the phoneme chunks has to be taken into account.

2.1 New Features Using Label Information

Ground truth of the detected phoneme chunk provides the necessary discriminatory information needed to differentiate true positives from false positives. Ground truth label of a detected phoneme chunk is the sequence of true label phonemes. Due to the unpredictability of frame classifier, we consider a predicted phoneme chunk to be correct if at least one of the frames in the predicted phoneme chunk and the ground truth phoneme sequence, have the same phoneme. This is because as the ground truth phoneme sequence is from the output of a forced aligner, there could be the misalignment of phoneme boundaries. For eg, let a phoneme chunk predicted be $[p_1\ \ p_2\ \ p_3\ \ p_4\ \ p_5]$ and let the ground truth be $[q_1\ q_2\ q_3\ q_4\ p_5]$, the phoneme chunk predicted is assumed to be correct, because of the 1 common phoneme at the end of the chunks.

With an objective to maximize the precision of the final scoring system, probabilistic models are fit on the features derived from d_m, to capture multiple aspects of speech. For modeling a specific phoneme, on a particular feature, positive and negative data is first captured. Positive data for a phoneme is the feature values derived from d_m, which eventually is classified as the phoneme correctly. Negative data for a phoneme is the false positive feature values derived from d_m, which incorrectly got classified as the phoneme. In the context of this paper, correctly detected refers to true positives and wrongly detected refers to false positives. Finally, probabilistic models are learned using the positive and negative data, for the particular feature, for the specific phoneme. The features are chunk size, chunk softmax average, chunk average softmax distribution, distinct phoneme count adjacent to the phoneme chunk.

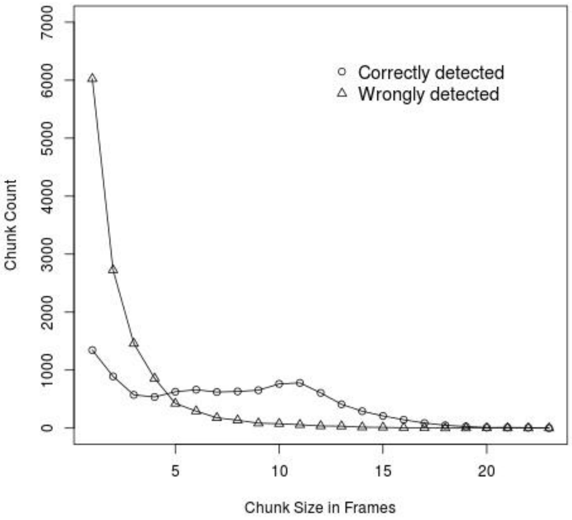

Figure 6: Count of correctly detected and wrongly detected phoneme chunks for /f/

Chunk Size: Phoneme chunk size is a crucial variable that indicates whether the detected phoneme chunk is indeed correct or not. In Fig 6, the true positives and false positives are plotted for phoneme /f/. As the phoneme chunk size increases, the detected phoneme chunk count also increases. Note that the for phoneme chunk size up to 4, the misrecognition rate is very high, which shows that small chunks are more likely to be misrecognized. From these type of plots, distributions on chunk size for true positives and false positives for specific phonemes can be fit.

The fact that even detected chunk size has an effect on the precision is very crucial. This information modeled using a distribution can be used in a full blown large vocabulary recognition engine to rescore the language model probabilities. But the downside is that it assigns a disproportionately low score for very short words.

Chunk Softmax Average: The average of softmax probabilities could be another differentiating variable, between correctly detected and falsely detected phoneme chunks.

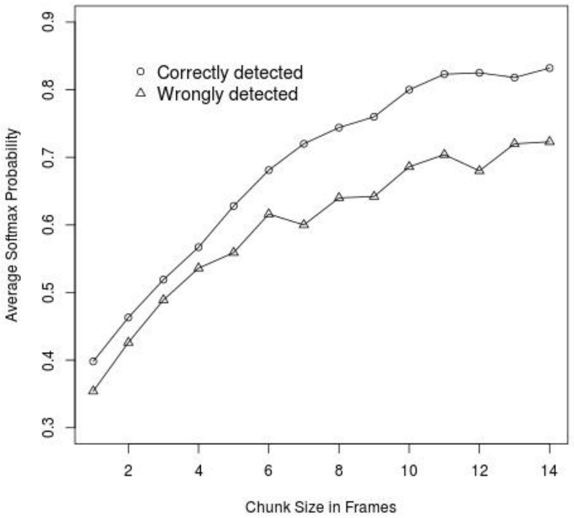

Figure 7: Mean of softmax probability across chunk size

Fig 7 plots the mean of average softmax probabilities of the phoneme chunks for the top phoneme /f/, averaged chunk size wise, for true positives and false positives. As the chunk size increases, the average softmax probability of the chunk increases. The average softmax probability of the phoneme chunk, for the correctly predicted and incorrectly predicted case, is very close to each other and does not provide much discriminatory information, when taken in isolation.

Chunk Average Softmax Distribution: Fig 8 and 9 plots the distribution of average softmax probability of top phoneme /f/, for wrongly detected and correctly detected phoneme chunks respectively. The chunk size used is 9 frames. It is apparent that both histograms are left skewed, but varies in the degree of skewness. The difference in skewness serves as another discriminatory feature.

As the histogram appears to be normally distributed and skewed, a skew normal distribution (Azzalini, 2013) can be used to model the data. A skew normal distribution models data which are normally distributed and skewed either left or right. A random variable Y is said to have a location-scale skew-normal distribution, with location λ, scale δ, and shape parameter α, and denote $Y \sim SN(\lambda, \delta^2, \alpha)$, if its probability density function is given by

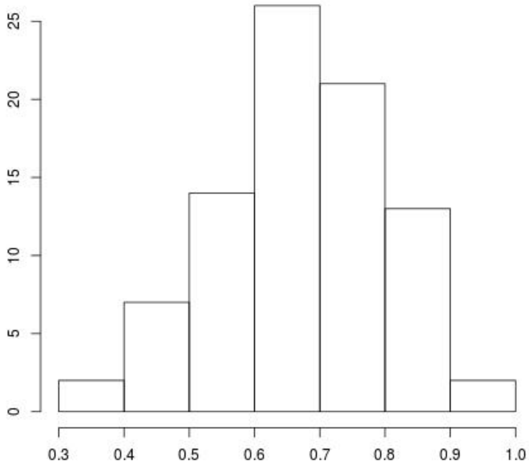

Figure 8: Wrongly detected phoneme chunk softmax probability histogram for /f/

$$f(y; \lambda, \delta^2, \alpha) = \tfrac{2}{\delta}\phi(\tfrac{y-\lambda}{\delta})\Phi(\alpha\tfrac{y-\lambda}{\delta}) \qquad (2)$$

where $y, \alpha, \lambda \in R$ and $\delta \in R^+$. ϕ and Φ denote the probability density function and cumulative distribution function of the standard Normal distribution. If the shape parameter $\alpha = 0$, then the skew normal distribution equals a normal distribution. i.e.,

$$f(y; \lambda, \delta^2, \alpha) = \mathcal{N}(y; \lambda, \delta) \;\text{ when } \alpha = 0$$

Given a dataset, the maximum likelihood estimate(MLE) of parameters λ, δ, α does not have a closed loop solution and are calculated using numerical methods.

Distinct Phoneme Count Adjacent to the Phoneme Chunk: The number of distinct phonemes, in a small window to the phoneme chunk detected, can serve as a source of information on whether the phoneme chunk detected is correct or not.

Fig 10 and 11, plot the number of distinct phonemes in a 5 frame window preceding to the phoneme chunk /p/, where the /p/ is correctly detected and wrongly detected respectively. For the correctly detected case, as seen in Fig 10, the presence of a single phoneme or 2 phonemes are more, in the adjacent left window. This means the detection rate is high if it is a smooth transition between phonemes. Converting this information into a distribution,

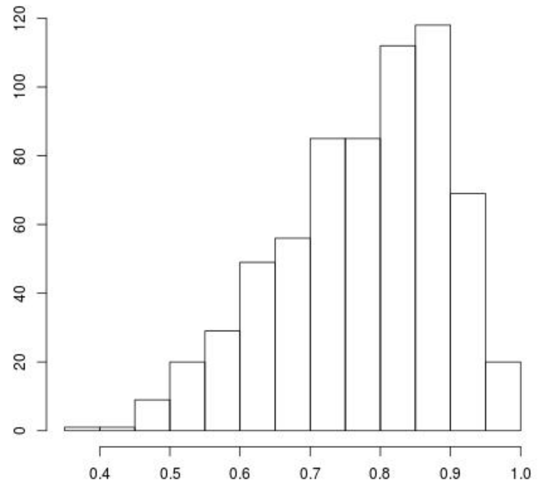

Figure 9: Correctly detected phoneme chunk softmax probability histogram for /f/

$$h(n;p) = \frac{C(n)}{\sum_n C(n)} \tag{3}$$

where $C(n)$ is the count of n distinct phonemes in the left window, of a detected phone p, in the dataset d_m.

3 Confidence Prediction

We seek a confidence score that reflects the precision of the prediction of a phoneme chunk. Assume a phoneme chunk of phoneme p with chunk size k, predicted by the frame classifier, with an average softmax probability s and with n distinct phonemes in the adjoining left window of the phoneme chunk. The confidence score is given by the posterior odds ratio,

$$\begin{aligned}\frac{P(p|s,k,n)}{P(\neg p|s,k,n)} &= \frac{\frac{P(p,s,k,n)}{P(s,k,n)}}{\frac{P(\neg p,s,k,n)}{P(s,k,n)}} \\ &= \frac{P(s|k,n,p)P(k|n,p)P(n|p)P(p)}{P(s|k,n,\neg p)P(k|n,\neg p)P(n|\neg p)P(\neg p)}\end{aligned}$$

With the following conditional independence assumptions,

$$\begin{aligned}P(k|n,p) &= P(k|p) \\ P(s|k,n,p) &= P(s|k,p)\end{aligned}$$

The posterior odds ratio can be written as

$$\frac{P(p|s,k,n)}{P(\neg p|s,k,n)} = \frac{P(s|k,p)P(k|p)P(n|p)P(p)}{P(s|k,\neg p)P(k|\neg p)P(n|\neg p)P(\neg p)}$$

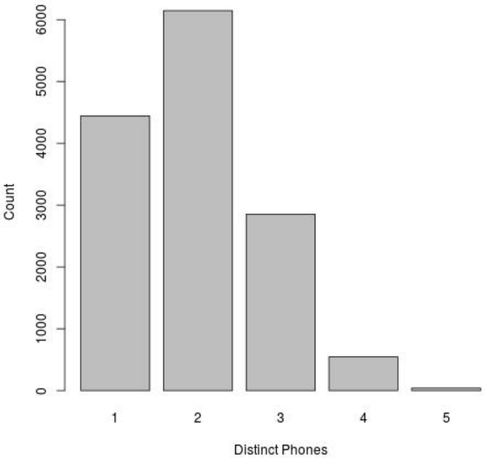

Figure 10: Distinct phoneme count preceding the correctly detected phoneme chunk /p/

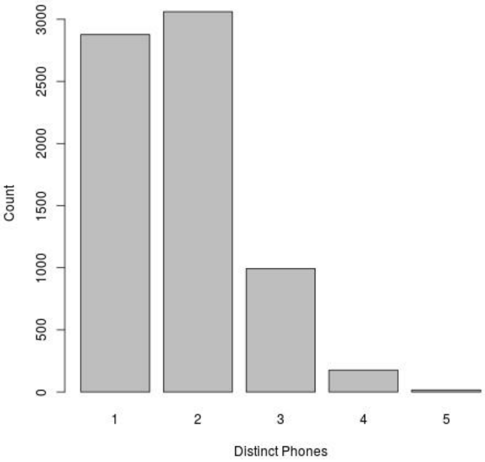

Figure 11: Distinct phoneme count preceding the wrongly detected phoneme chunk /p/

where the distributions are defined as,

$$\begin{aligned}P(k|p) &= g(k;p) \\ P(k|\neg p) &= g(k;\neg p) \\ P(n|p) &= h(n;p) \\ P(n|\neg p) &= h(n;\neg p) \\ P(s|k,p) &= f(s;\lambda_{pk},\delta^2_{pk},\alpha_{pk}) \\ P(s|k,\neg p) &= f(s;\lambda_{\neg pk},\delta^2_{\neg pk},\alpha_{\neg pk})\end{aligned}$$

where $\neg p$ represents the false positives of phoneme p, that is the cases where phoneme chunks are wrongly detected as p. $P(p)$ and $P(\neg p)$ represents the prior probabilities of true positive and false positive cases respectively. A high posterior ratio means high precision as the posterior ratio is directly proportional to the precision of the prediction.

Assuming equal prior probabilities for true positives and false positives, i.e., $P(p) = P(\neg p)$, and in the absence of any other information, the posterior odds ratio reduces to,

$$\frac{P(p|s,k,n)}{P(\neg p|s,k,n)} = \frac{P(s|k,p)P(n|p)P(k|p)}{P(s|k,\neg p)P(n|\neg p)P(k|\neg p)} \quad (4)$$

Equation (4) calculates the posterior probability ratio of the case where the detected phoneme chunk is actually the correct phone, to where detected phoneme chunk is a false positive.

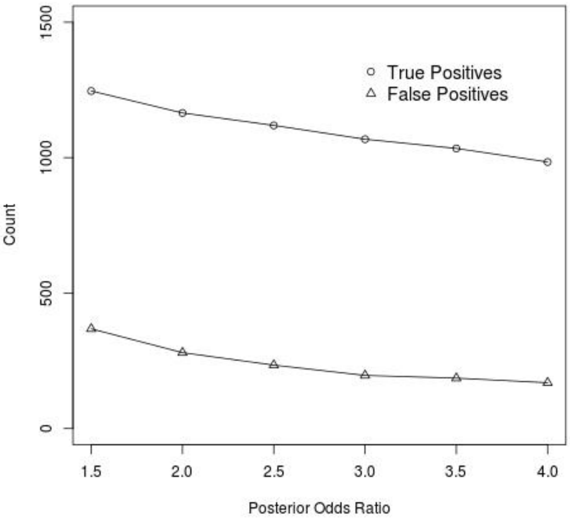

Figure 12: Posterior ratio vs true positives and false positives for /p/

4 Experimental Results

As this approach focuses on confidence scoring for a detected phone, it is the precision that has to be tested. Models $P(k|p)$, $P(k|\neg p)$, $P(s|k,p)$, $P(s|k,\neg p)$, $P(n|p)$, and $P(n|\neg p)$ are built from d_m. For chunk size $k \geq 10$, where the number of instances are less for $\neg p$, data is pooled together and the skew normal distribution is fit. This makes sense as for $k \geq 10$, the average softmax probability of the phoneme chunk varies gradually, as is shown in Fig 7. Testing is done on the subset d_3. Fig 12 plots the posterior ratio vs true positives and false positives for a selected phoneme /p/.

Each point (x, y) in the true positive curve means the following. For posterior odds ratio greater than x, there are y instances of the phoneme chunk /p/ recognized which are true positives. And each point (x, y) in the false positive curve means, for the posterior odds ratio greater than x, there are y instances of phoneme chunk /p/ detected which are false positives. The result is the aggregate of all the chunk sizes together. As false positives decreases, true positives also decreases. Fig 13 plots the precision of phoneme /p/. As the threshold of posterior ratio increases, the precision also increases, but at the expense of true positives. Based on the use case, the best operating point can be selected.

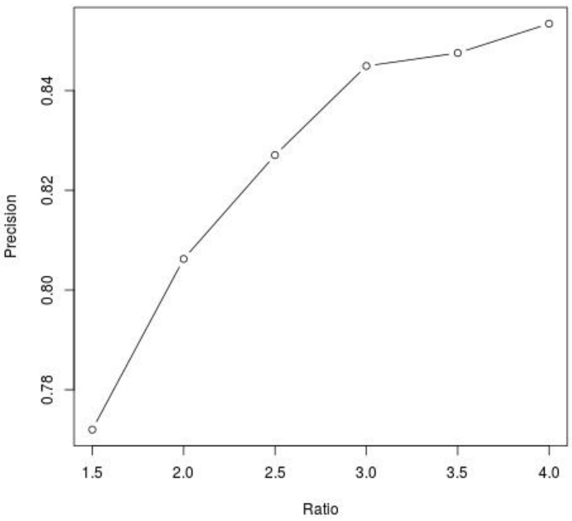

Figure 13: Phoneme /p/ precision

The difference between this approach and a direct posterior based confidence scoring approach (Wang et al., 2009) is in the additional assumptions made on the softmax probability. Characteristics of the posteriors for the true positive and false positives, associated with a phoneme, is incorporated into the probabilistic framework. The focus here is on the precision of the phoneme detection.

5 Conclusion

A new probabilistic approach is presented which provides a confidence score to a phoneme chunk detected by the frame classifier. Predictor variables like the phoneme chunk size, number of distinct phonemes in an adjacent window to the phoneme chunk, the average softmax probability of the phoneme chunk, are explored. A full probabilistic model is specified with conditional independence assumptions to make the distributions simple. The distributions are learned from real-world data. Benchmarking of the approach is done with the sole focus on precision.

This probabilistic model is suitable for adding new variables if the likelihood of new variable value conditioned on various phonemes can be computed. More variables derived independently from acoustic phonetics, time domain, or spectrum, can be easily added to the model. As long as the variables are meaningful and with proper conditional independence assumptions, the confidence score can be calculated without expensive computation.

In this paper, the focus is solely on the precision. This helps in calibrating the confidence scoring mechanism for a certain type of utterances like confirmations in an IVRS system, where a misrecognition is very expensive. In the future, we aim to make a confidence scoring mechanism with an overall goal of improving recall, which suits a host of other applications like recognition from a list of words. Another area of improvement is to use the confidence score of a phoneme chunk to calculate the confidence score of another chunk, possibly in the same word. This requires a language model at the phonetic level to model the short-range dependencies.

References

Adelchi Azzalini. 2013. *The Skew-Normal and Related Families*. Institute of Mathematical Statistics Monographs. Cambridge University Press.

Giulia Bernardis and Hervé Bourlard. 1998. Improving posterior based confidence measures in hybrid hmm/ann speech recognition systems. In *ICSLP*.

J Duchateau, Kris Demuynck, and Patrick Wambacq. 2002. Confidence scoring based on backward language models. volume 1, pages 221–224.

Po-Sen Huang, Kshitiz Kumar, Chaojun Liu, Yifan Gong, and Li Deng. 2013. Predicting speech recognition confidence using deep learning with word identity and score features. In *Acoustics, Speech, and Signal Processing, ICASSP-2013*, pages 7413–7417.

Timothy J. Hazen, Joseph Polifroni, and Stephanie Seneff. 2002. Recognition confidence scoring and its use in speech understanding systems. *Computer, Speech and Language*, 16:49–67.

Hui Jiang. 2005. Confidence measures for speech recognition: A survey. *Speech Communication*, 45:455–470.

Thomas Kemp and Thomas Schaaf. 1997. Estimating confidence using word lattices. *Proceedings of Eurospeech*, pages 827–830.

Hamed Ketabdar. 2010. Improving posterior based confidence measures using enhanced local posteriors. *2010 18th European Signal Processing Conference*, pages 2007–2011.

Akinobu Lee, Kiyohiro Shikano, and Tatsuya Kawahara. 2004. Real-time word confidence scoring using local posterior probabilities on tree trellis search. In *International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 793–796.

Richard Rose, B.H. Juang, and Chin-Hui Lee. 1995. A training procedure for verifying string hypotheses in continuous speech recognition. *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, 1:281–284.

Bernhard Rueber. 1997. Obtaining confidence measures from sentence probabilities. In *EUROSPEECH-1997*, pages 739–742.

T. Schaaf and T. Kemp. 1997. Confidence measures for spontaneous speech recognition. In *Proceedings of ICASSP 97 Volume 2*, pages 875–878.

Dong Wang, Javier Tejedor, Joe Frankel, Simon King, and Jose Colas. 2009. Posterior-based confidence measures for spoken term detection. In *Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, ICASSP '09, pages 4889–4892.

M. Weintraub, F. Beaufays, Z. Rivlin, Y. Konig, and A. Stolcke. 1997. Neural - network based measures of confidence for word recognition. In *Proceedings of the 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '97)-Volume 2*, pages 887–890.

Frank Wessel, Klaus Macherey, and Hermann Ney. 1999. A comparison of word graph and n-best list based confidence measures. In *Proceedings of Eurospeech*, pages 315–318.

Dong Yu, Jinyu Li, and Li Deng. 2011. Calibration of confidence measures in speech recognition. *Trans. Audio, Speech and Lang. Proc.*, 19(8):2461–2473.