Estimating Classifier Performance in Unknown Noise

Ehsan Variani¹, Hynek Hermansky¹²

Johns Hopkins University
¹Center for Language and Speech Processing
²Human Language Technology Center of Excellence
Baltimore, MD, 21218

variani@jhu.edu, hynek@jhu.edu

Abstract
We propose and investigate a non-parametric method for identifying regions of speech that have unexpected distortions not seen in the training data. The method does not require knowledge of correct labels and relies only on divergence between statistics of the test and training data. Our experiments show that the proposed method requires a relatively small amount of test data of the order of several seconds to stabilize, and correlates well with recognition error observed on the test data.

Index Terms: Unexpected distortions, confidence estimation, machine recognition of speech

1. Introduction
It would be beneficial for a machine recognition system to be able to identify potential problems during system operation, or to switch between possible modes of operation in response to unexpected test data. This paper proposes using confidence estimation tools to evaluate the extracted information upstream of the recognition system, in order to predict when performance will be inferior. The existence of such tools would simplify many recognition systems such as speech, speaker, language recognition etc. Current systems require decoding under each hypothesis individually before selecting the most likely hypothesis [2, 9], while having such tools would enable finding the single best hypothesis before decoding.

Another scenario where having such tools is beneficial is in multi-stream speech recognition. Multiple streams of information are extracted from test data and each stream is passed to a phoneme classifier that is trained on known clean data. Streams with higher confidence are then be selected and passed to a fusion block which then adaptively fuses selected streams [6, 7]. The general framework of such systems is shown in Fig. 1.

Figure 1: modifiable classifier based on comparison between training and test statistics.

Having a sufficient amount of training data, the task we are treating in this paper is introducing a measure to calculate divergence between unknown test data and the underlying distribution of known training data in order to estimate reliability of the recognition result. The data we are working on are phoneme posterior probabilities which are estimated by a Multi Layer Perceptron (MLP) trained on known clean training data.

The approach proposed in this paper is extracting appropriate statistics which describe the underlying distribution of the training and test data, and then comparing the statistics using a proper type of divergence function. The benefit of this approach is that we don’t restrict ourselves to any fixed distribution type.

There are some properties that we expect our divergence measure to have. First, we expect it to be well correlated with recognition performance. This is an important property because we are looking for a measure that predicts recognition error without knowing the correct labels. The second property we are aiming for is about test segment length that our measure needs to stabilize. For practical applications, divergence estimation should be done on a relatively short amount of test data. As shown later, the statistical approach proposed here is simple, fast, and highly correlated with recognition error.

The research presented in this paper was partially funded by the DARPA RATS program under D10PC20015 and the JHU Human Language Technology Center of Excellence. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the DARPA or JHU HLTCOE.
calculates the divergence between the test and training data, we introduce a measure which extracts from training data, we introduce a measure which compares these with similar statistics extracted from training data, we introduce a measure which compares these with similar statistics extracted from training data. By comparing these with similar statistics extracted from training data, we introduce a measure which calculates the divergence between the test and training data. The measure uses first and second order statistics of the posteriors generated by a three layer MLP trained on known clean data. The MLP generates a $40 \times 1$ vector $P_i$ for each frame $i$ of data. Each component of the posterior vector is the estimation of a phoneme posterior probability. The first and second order statistics of posteriors will then be calculated as:

$$
\mu = \frac{1}{N} \sum_{i=0}^{N} \phi(P_i)
$$

$$
\Sigma = \frac{1}{N} \sum_{i=0}^{N} \phi(P_i)\phi(P_i^T)
$$

where $N$ is the number of frames in the given window, $\phi(\cdot)$ is some monotonic compressive function (currently the logarithm is used). For a reference statistic we use all training data while for testing we use a short segment of test data.

The extracted statistic will be then compared by a reference statistic. The first order statistic is captured using the Mahalanobis distance metric. The second order statistic is captured using the logarithm of the ratio of the determinants of the autocorrelation matrices.

2.2. The second order variability

Given a window of posteriors (denoted by $P = (P_i, P_{i+1}, ..., P_{i+N})$, where $N$ is the number of frames in the window) from test data, we compare the mean vector of the test data, $\mu_{test}$, to the mean vector of the training data, $\mu$ calculated using all training data. The comparison is done by calculating the Mahalanobis distance between the training mean and the test mean [5]:

$$
d_1 = \sqrt{(\mu_{test} - \mu)^T \Sigma^{-1} (\mu_{test} - \mu)}
$$

where $\Sigma^{-1}$ is the inverse of the autocorrelation matrix derived using all training data and $T$ denotes transpose.

This metric first transforms the difference between sample mean and training mean into an uncorrelated space by multiplying by $\Sigma^{-1}$ and then calculates the Euclidean norm of the transformed data point. This results in a metric that compensates for the effect of the variability of each individual phoneme.

2.2. The second order variability

The average autocorrelation of posteriors over time reveals the pattern by which different phonemes are confused with each other during a given span of time. Fig. 2 shows two autocorrelation matrices for a sample utterance in a clean and a noisy situation. As this figure shows,
Figure 3: Correlation between divergence and recognition error on data sets of utterances of different length $\ell = 1, 2, 4, 8$ and all test utterances of TIMIT database using proposed non-parametric approach.

To evaluate the proposed divergence measure, we used clean TIMIT training data as our reference statistic and 29 classes of test data, including one clean and 28 noisy versions of TIMIT test data. We used four types of noises (babble, buccaneer, white, and pink) in seven SNR levels ranging from 0 dB to 30 dB in 5 dB steps. These noises were added using the FaNT tool [4].

One important parameter of the proposed measure is the minimal length of the test window for calculating a reliable estimate of the divergence value. To calculate this length for the proposed measure, an experiment over a sufficiently large number of utterances of similar length was conducted. The utterances were made by concatenating all TIMIT test utterances into one single long utterance, and then selecting 10000 random segments of fixed length varying from $\ell = 1, ..., 10$ second.

As mentioned earlier, there are two major properties that we expect for the proposed divergence measure: high correlation between divergence and recognition error, and fast stabilization. Different experiments were conducted to evaluate each of these properties for the proposed divergence measure.

The first experiment is evaluating the correlation between the divergence value and the recognition error. Divergence values were calculated for TIMIT utterances and fixed length utterances of $\ell$ second length. Fig. 3 shows the correlation coefficient between divergence and recognition error calculated for all 29 classes of test data. As this figure shows the correlation coefficient increases with the utterance length, which is reasonable.

3. Experimental results

We used a phoneme recognition system based on the Hidden Markov Model Artificial Neural Network (HMM-ANN) paradigm [1]. The hybrid system is trained using PLP features [3] with a 9-frame context extracted from clean TIMIT training data. The initial 61 hand-labels of the TIMIT database are mapped to the standard set of 40 phonemes including silence [8].
since longer utterances lead to better estimation of statistics which result in better estimation of divergence. The correlation stabilizes around 4 seconds. We also plot the correlation between divergence and recognition error for individual noise test classes in Fig. 4. As this figure shows, there is also high correlation between divergence and recognition error for different individual test classes. In addition, Fig. 5 shows scatter plot between recognition error and divergence value for clean and four type of noises (in 0, 10, 20 dB SNR). The divergence and recognition error values are calculated over fixed $\ell = 8$ sec. utterances.

The second experiment investigates the necessary time that divergence value need to stabilize. To calculate this we run experiment over utterances of length $\ell$, where $1 \leq \ell \leq 10$. The divergence values were calculated for each class of test data. Fig. 6 shows how divergence mean and variance changes over time for clean and babble noise in three different SNR (5, 15, 25 dB). Both the mean and the variance of the divergence values decrease as utterance lengths increase. The mean of the divergence value seems to stabilize around 4 seconds which is practical for applications like multi-stream speech recognition [6].

4. Conclusion

We proposed a non-parametric measure to evaluate the divergence between underlying distribution of unknown test data and known training data. The measure is based on comparison between appropriate statistics derived from test data and statistics derived from training data. The statistics used here is derived from phoneme posteriors estimated by a three layer MLP trained on known clean data. Our experiments show that the proposed divergence measure is correlated with recognition error and could be derived from several seconds of test data.

5. References