WORD AND PHONE LEVEL ACOUSTIC CONFIDENCE SCORING

Simo O. Kamppari and Timothy J. Hazen

Spoken Language Systems Group
Laboratory for Computer Science
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139 USA

ABSTRACT

This paper presents a word level confidence scoring technique based on a combination of multiple features extracted from the output of a phonetic classifier. The goal of this research was to develop a robust confidence measure based strictly on acoustic information. This research focused on methods for augmenting standard log likelihood ratio techniques with additional information to improve the robustness of the acoustic confidence scores for word recognition tasks. The most successful approach utilized a Fisher linear discriminant projection to reduce a set of acoustic features, extracted from phone level classification results, to a single dimension confidence score. The experiments in this paper were implemented within the JUPITER weather information system. The paper presents results indicating that the technique achieved significant improvements over standard log likelihood ratio techniques for confidence scoring.

1. INTRODUCTION

Because the speech recognition systems of today remain far from perfect, the process of discovering errors in recognition remains an important task. In earlier work we examined this problem at the utterance level [6]. By examining various features extracted from the results of the recognition and understanding components of the system, a decision on whether or not to accept or reject the system's hypothesized understanding of an utterance was made. This approach was successful at rejecting a large number of utterances which contained out-of-vocabulary words, severe noise or non-speech events, poorly articulated speech, misrecognized words, etc. However, the system was limited in that it could only accept or reject an entire utterance but was unable to accept or reject individual words or phrases contained within an utterance.

In this work we look to extend our confidence scoring approach to the level of words, thus allowing a finer grained analysis of the output from the recognition process. The goal is to develop word level confidence scores which can serve as robust indicators of the correctness of word hypotheses. In a spoken language system, these scores could help determine what portions of a user's query the system has recognized correctly and what portions of the utterance the system had difficulty recognizing. Ideally, these scores would enable the system to target potential misunderstandings and take measures to correct, clarify, or confirm them before performing any misguided actions based on an incorrect recognition string.

This paper focuses on word level confidence scores derived from purely acoustic features. Specifically, the research focuses on various features that can be extracted from the output of a phonetic classifier, i.e., features that can be derived from acoustic observations only. This means that features based on language model outputs are not utilized, even though their use has proven to be effective in past work [1, 7]. However, our goal is to develop an accurate acoustic confidence measure which could be combined with features from a language understanding component at a later stage in the processing.

2. IMPLEMENTATION

2.1. Overview

In this paper the derivation of a word level acoustic confidence metric is a two step process incorporated into the SUMMIT speech recognition system [3]. First acoustic confidence scores are calculated for the underlying components of each word. In this case, the recognizer scores observations extracted from *landmarks*, which are potential phonetic boundaries proposed by a segmentation algorithm. These landmarks are scored using context-dependent *diphone* boundary models. This is similar to a standard Hidden Markov Model (HMM) approach with the exception that the proposed landmark observations included measurements which span multiple frames and do not occur at a fixed rate. A hypothesized word is thus composed of a sequence of hypothesized diphones. After the landmarks have been scored, a word confidence score is computed via some combination of the underlying diphone scores.

2.2. Phone Level Scoring

The acoustic features are primarily based on two common phonetic classification scoring approaches: normalized log-likelihood (NLL) scoring and maximum *a posteriori* probability (MAP) scoring. This work builds on previous work which has dealt with these techniques [7]. The MAP score for a boundary model, c_i, given a landmark observation, $\vec{x}$, is expressed as:

$$C_{map}(c_i|\vec{x}) = \mathrm{P}(c_i|\vec{x}) = \frac{\mathrm{p}(\vec{x}|c_i)\mathrm{P}(c_i)}{\mathrm{p}(\vec{x})} \qquad (1)$$

Similarly the equivalent NLL score is expressed as:

$$C_{nll}(c_i|\vec{x}) = \log\left(\frac{\mathrm{p}(\vec{x}|C_i)}{\mathrm{p}(\vec{x})}\right) \qquad (2)$$

This research was supported by DARPA under contract DAAN02-98-K-003, monitored through U.S. Army Natick Research, Development and Engineering Center.

In both of these scoring techniques, the likelihood of the hypothesized model is normalized by a generic *catch-all* model $\mathrm{p}(\vec{x})$, which can be expressed as:

$$\mathrm{p}(\vec{x}) = \sum_{j=1}^{N_c} \mathrm{p}(\vec{x}|c_j)\mathrm{P}(c_j) \quad (3)$$

The MAP score also utilizes the prior probability of the diphone model to produce a true probability measure which varies between 0 and 1. The NLL score is expressed in the log domain and can be viewed as a zero-centered score where positive scores are good and negative scores are bad.

2.3. Word Level Scoring

Word level confidence scores can be derived by extracting various measurements or *features* from the underlying phone level NLL and MAP scores (or other recognition output features) and then combining them together in some fashion to produce a single word level confidence score. In this work twelve word level features are derived from the results produced by the recognizer. A summary of these twelve features is presented in Table 1.

The first four features are simply averages of C_{map} and C_{nll} over all observations in a word. Both arithmetic and geometric means are utilized. The two means have distinct behaviors depending on the underlying scores. The geometric mean can be heavily biased by poorly scoring observations, whereas the arithmetic mean is less sensitive to small outliers, and thus more indicative of the average ability of a model's ability to account for the observations.

In addition to the above means several other features were utilized. The standard deviations for the C_{map} and C_{nll} scores, σ_{map} and σ_{nll}, are used as indicators of the consistency of the scores across the word. A high arithmetic mean along with a low standard deviation indicates consistently high phonetic scores across the whole word. A high standard deviation means that the phonetic scores are widely dispersed, and hence not consistent across all the phones.

The three minimum scores, $C_{min-map}$, $C_{min-map-internal}$, and $C_{min-nll}$, represent the lowest scores obtained across all observations. Generally, a low minimum score is an indicator that some portion of the word is not well matched to its hypothesized phonetic unit.

The arithmetic mean p^A is the average ability of the catch-all model to account for the acoustic observations in a word. This score is independent of the hypothesized string, but is an indicator of how well matched the observed acoustics are to the typical acoustics observed in the training data.

The last two features are N_{nbest} and N_{land}. While these two are only indirectly a function of the acoustic evidence, they can be correlated with correctness. N_{nbest} is the number of competing hypotheses in the *n*-best list. The fewer hypotheses there are, the better the models are doing at discriminating between competing hypotheses. N_{land} is the number of landmarks within each word. Generally, longer words are more acoustically distinct than shorter ones, thus the chance of confusion is much smaller for longer words.

Feature	**Description**
C^A_{map}	Arithmetic mean of C_{map} scores
C^A_{nll}	Arithmetic mean of C_{nll} scores
C^G_{map}	Geometric mean of C_{map} scores
C^G_{nll}	Geometric mean of C_{nll} scores
σ_{map}	Standard deviation of C_{map}
σ_{nll}	Standard deviation of C_{nll}
$C_{min-map}$	Minimum C_{map} score in word
$C_{min-map-internal}$	Min. internal C_{map} score in word
$C_{min-nll}$	Minimum C_{nll} score in word
p^A	Arith. mean of *catch-all* model score
N_{nbest}	Number of utts. in *n-best* list
N_{land}	Number of landmarks in word

Table 1: A complete list of word level features used for confidence scoring.

2.4. Combining Word Level Features

2.4.1. Overview

While is is possible that some of the word level features can provide adequate confidence scores on their own, improvements in performance over the single best features should be possible by combining the features in an appropriate fashion. Significant improvements may be possible if the features provide complementary information. This paper explored two methods for analyzing and combining the full set of features: probabilistic hypothesis testing and Fisher Linear Discriminant Analysis (FLDA) [2].

2.4.2. Hypothesis Testing

The probabilistic hypothesis testing approach utilizes two probabilistic models which are applied to the vectors of word level features, $\vec{f}$. The model M_C models the features of words that were correctly recognized, while the model M_I is for words which were incorrectly recognized. During word level confidence scoring, a simple hypothesis testing ratio between the two models can be computed to generate a word level confidence score, C_{ht}, as follows:

$$C_{ht} = \frac{\mathrm{p}(\vec{f}|M_C)}{\mathrm{p}(\vec{f}|M_I)} \quad (4)$$

This research explored the use of mixture Gaussian models (both full covariance and diagonal) for representing M_C and M_I.

2.4.3. Fisher Linear Discriminant Analysis

Fisher Linear Discriminant Analysis (FLDA) is a means of reducing a set of measurements to a single measurement using a linear projection. The linear projection is determined from training data for a two class discrimination task (correctly and incorrectly hypothesized words in this case). An FLDA projection vector, $\vec{w}$, is learned from the development data containing correctly and incorrectly recognized word hypotheses. The projection vector is then applied to the word level feature vector, $\vec{f}$, of any newly hypothesized word to produce a word confidence score, C_{flda}, as follows:

$$C_{flda} = \vec{w}^t\vec{f} \quad (5)$$

2.5. Catch-all Model Estimation

In a real-time recognition system, the computation of the *catch-all* model $\mathrm{p}(\vec{x})$ becomes an issue. A large number of context-dependent diphone models are typically required for adequate performance. However, because pruning is typically performed during the search to reduce computation, only a fraction of the diphone models may actually be computed for any given landmark. In order to maintain real-time performance it is not feasible to compute the value of $\mathrm{p}(\vec{x})$ directly because it requires the computation of all diphone models. In order to reduce the computational burden, a method for estimating $\mathrm{p}(\vec{x})$ is proposed.

This method is based on a binary bottom-up clustering of all of the Gaussian components in the catch-all model. At each iteration of the bottom-up clustering, the two most similar Gaussians are found using a weighted *Bhattacharyya* distance metric. These two Gaussians are then combined together to form a new single Gaussian, which is an ML estimate of the sum of the separate models. The new Gaussian then replaces its two constituent Gaussians in the next iteration. Each iteration reduces the number of Gaussian components by one. The process is continued until the estimated model is reduced enough for it to be computed efficiently during recognition. Details of the clustering algorithm and distance metric can be found in [5].

3. EXPERIMENTS

3.1. System Description

To evaluate the word confidence scoring techniques, the utterances used for the evaluation process were actual spontaneous queries collected over the telephone by the JUPITER weather information system [8]. The word confidence scoring techniques are applied to the recognition results for the recognizer used by the JUPITER system [4]. The version of the recognizer used for these experiments had a vocabulary of 1893 words and was trained on 20064 utterances. A development set of 3437 utterances was used to train the hypothesis testing models and the FLDA projection vector. A test set of 2405 utterances was used to evaluate the confidence scoring techniques. The word error rate of the recognizer was 19.4%.

3.2. Evaluation Metrics

To evaluate the performance of confidence metrics, hypothesized words are compared against the true transcription of the utterance with each hypothesized word being classified as *correct* or *incorrect*. The confidence scores for each word are then compared against a confidence threshold and the hypothesized words are either *accepted* or *rejected*. The threshold can be varied to control the tradeoff between false alarms (incorrect words that are accepted) and detections (correct words that are accepted). By varying the confidence score threshold, a receiver operating characteristic (ROC) curve can be plotted.

Performance can also be measured in terms of a *figure of merit* (FOM), which measures the performance of a system at or around a particular operating point on the curve. In our system it is desirable to maintain a high detection rate at the expense of increased false alarms. To capture this condition our figure of merit measures the area under the ROC curve in the range of .8 to 1.0 for correct acceptances. This area is then normalized by the total area in this range to produce an FOM whose optimal value is 1. A *chance* FOM of 0.1 is achieved by random guessing.

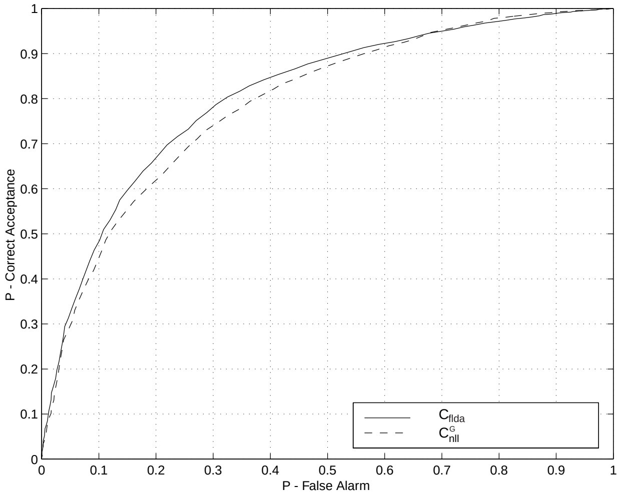

Figure 1: The ROC curves indicating relative word level confidence performance for the single best feature C_{nll}^{G} vs. the FLDA combined feature set C_{flda}.

Feature	Figure of Merit
C_{nll}^{G}	0.4114
C_{nll}^{A}	0.3782
$C_{min-map}$	0.3617
C_{map}^{G}	0.3546
$C_{min-nll}$	0.3018
C_{map}^{A}	0.2591
C_{flda}	0.4502
$chance$	0.1

Table 2: The figure of merit performance of six individual word features and the FLDA combined feature confidence scores.

3.3. Word Level Feature Performance

Of the 12 proposed word level features the geometric mean of the C_{nll} scores, C_{nll}^{G}, was the single best performing feature. In general the NLL based scores outperformed the MAP scores, leading to the conclusion that the priors do not improve confidence scoring performance. Excluding the priors, as is the case with the NLL based scores, allows the acoustic evidence to speak for itself. It is also interesting that the geometric means consistently outperform their arithmetic mean counterparts, for both the NLL and MAP based scores. This result can be accounted for by the characteristic behaviors of each of the means. The geometric mean allows a single low score to pull down the score for the whole word, where as an arithmetic mean can be immune to a single low score, especially if many values are averaged. Table 2 shows the FOM performance for the 6 best performing individual word level features.

3.4. Combining Word Level Features Performance

In a set of preliminary experiments, the FLDA approach for combining features performed significantly better than the probabilistic hypothesis testing approach. Because hypothesis testing performed so poorly in these initial experiments it was abandoned

Reduction	Figure of Merit
None	0.4502
75%	0.4451
95%	0.4316
99%	0.4161
99.5%	0.4092

Table 3: Effects of *catch-all* model reduction on figure of merit performance.

early, and the FLDA approach was adopted for the remainder of our experiments. Figure 1 illustrates the relative performance of the FLDA combination method vs. the single best feature C_{nll}^{G}. Table 2 shows the FOM performance for the FLDA combined score, C_{flda}, as compared to the six best individual word feature scores. This table illustrates a significant increase in performance from using all of the features instead of using just the best single feature.

3.5. Performance of Estimated Catch-all Model

It was hoped that the size of the *catch-all* model could be significantly reduced without harming performance. Table 3 shows the FOM performance for the FLDA derived confidence score when reducing the *catch-all* model size using the estimation procedure discussed in Section 2. The initial *catch-all* model was defined by 11433 mixture Gaussian components. The percentages on the left hand column of the table indicate the reduction in the number of Gaussian components. A 99.5% reduction corresponds to a catch-all model which is defined by only 57 mixture Gaussian components. Thus a 99.5% reduction in the size of the catch model resulted in only a 9% relative reduction in the FOM.

3.6. Effects of Word Content Classes

When computing the word error rate for a recognizer, all words contribute equally to the performance measure. However, as speech recognition is often used in conjunction with some understanding component it is clear that some words are more important than others. From the perspective of understanding, function words like *a*, *an* and *the* have little value while content words, which depend highly on the domain, are very important. As this paper revolved around a weather information domain, words describing locations of interest, dates, and weather conditions were the most important types of words for understanding the user's request. For our experiments, the entire vocabulary of JUPITER was hand-classified into two categories: high and low content words. Words in the high content category are crucial to understanding while words in the low content category contain little or no information relevant to the final understanding of the utterance.

The results of this analysis were encouraging. The confidence scores extracted for high-content words were significantly more accurate than the confidence scores for low-content words. This result can most likely be attributed to the observations that the high-content words tend to be longer in length, more acoustically distinct, and more carefully articulated than the low-content words. Table 4 shows the performance for both the combined score, C_{flda}, and the best single word level feature, C_{nll}^{G}, for high- and low-content words.

One should note that the performance of the combined score, C_{flda}, is significantly better for the high-content words then the low-content words. On the other hand, the difference in performance between low-content and high-content words using the single feature C_{nll}^{G} is significantly smaller. This indicates that the added value of using the full set of features is most pronounced when examining the words which are most important to the correct understanding of the utterance.

Feature	Content Type	Figure of Merit
C_{flda}	High	0.5249
C_{flda}	Low	0.4311
C_{flda}	All	0.4502
C_{nll}^{G}	High	0.4297
C_{nll}^{G}	Low	0.4102
C_{nll}^{G}	All	0.4114
$chance$		0.1

Table 4: Figure of merit performance values for C_{flda} and C_{nll}^{G} on content classes *high*, *low*, and all words.

4. CONCLUSIONS & FUTURE WORK

This paper has presented a method for word level acoustic confidence scoring which combines multiple features using a Fisher linear discriminant analysis technique. This approach performs significantly better than a standard normalized log-likelihood approach. This performance improvement is even larger when examining only the high-content words which are most important to the understanding of a query. The next step of our work is to begin incorporating confidence scores into the dialogue component of our system. It is our hope that these scores can be useful for providing informed feedback to the user about potential misrecognitions that the system may have incurred.

5. REFERENCES

[1] L. Chase, "Word and acoustic confidence annotation for large vocabulary speech recognition," in *Proc. Eurospeech*, Rhodes, 1997.

[2] R. Duda and P. Hart, *Pattern Classification and Scene Analysis*. New York: John Wiley & Sons, 1973.

[3] J. Glass, J. Chang, and M. McCandless, "A probabilistic framework for feature-based speech recognition," in *Proc. ICSLP*, Philadelphia, 1996.

[4] J. Glass, T. Hazen, and L. Hetherington, "Real-time telephone-based speech recognition in the jupiter domain," in *Proc. ICASSP*, Phoenix, 1999.

[5] S. Kamppari, *Word and Phone Level Acoustic Confidence Scoring for Speech Understanding Systems*. Master's thesis, MIT, 1999.

[6] C. Pao, P. Schmid, and J. Glass, "Confidence scoring for speech understanding," in *Proc. ICSLP*, Sydney, 1998.

[7] M. Weintraub, F. Beaufays, Z. Rivlin, Y. Konig, and A. Stolcke, "Neural-network based measures of confidence for word recognition," in *Proc. ICASSP*, Munich, 1997.

[8] V. Zue, *et al*, "JUPITER: A telephone-based conversational interface for weather information," *IEEE Trans. on Speech and Audio Processing*, vol. 8, no. 1, January 2000.