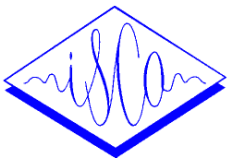

International Symposium on Chinese Spoken Language Processing (ISCSLP 2002)
Taipei, Taiwan
August 23-24, 2002

COMPARISON AND COMBINATION OF COFIDENCE MEASURES IN ISOLATE WORD RECOGNITION

XIONG Zhenyu, XU Mingxing, WU Wenhu

Center of Speech Technology,
State Key Laboratory of Intelligent Technology and Systems,
Department of Computer Science & Technology
Tsinghua University, Beijing, 100084
[xiongzhy, xumx, wuwh]@sp.cs.tsinghua.edu.cn

ABSTRACT

In this paper, we describe our work on the field of confidence measures for isolate word recognition system based on hidden Markov models (HMMs). Three kinds of frame level likelihood ratios are extracted as basic confidence features, and phone level confidence measures are derived from these features. Word level confidence measures are derived from phone level confidence features or from frame features directly. These different kinds of word level confidence measures are experimentally compared on a Chinese name database. The experiment shows that the confidences based on phone level features are better than those derived from frame features directly, and a kind of frame features based on filler model outperforms other two kinds. And then a Fisher linear discriminant projection and a non-linear backpropagation neural network are utilized to combine these different kinds of word level confidence features. An evaluation on the Chinese name database shows that the non-linear network approach exceeds the Fisher linear approach, and improves the performance in comparison to the baseline in which only a single kind of word level confidence feature is used.

1. INTRODUCTION

Speech recognition systems are typically developed for closed set recognition, in which the vocabulary is predetermined fixed and limited, and the models are inadequate. Those systems are not entirely appropriate for real applications where unknown words and noise speech may occur. In the context of command-and-control applications, the recognition system must have the ability to handle the case that a speaker speaks a word which is not within the vocabulary of the system. It must judge the word recognition result of a speaker's input and determine whether we have to 'accept' or 'reject' this result. In other words, it must classify single word utterances into two categories: utterances within the vocabulary which are recognized correctly, and other utterances, namely out-of-vocabulary (OOV) or misrecognized utterances.

To this end, a number of techniques have been developed. In some methods, an explicit OOV word model is added into the model set of recognition system in order to identify potential unknown words during recognition [1, 2]. And for other more methods, a set of confidence features are extracted to estimate recognition reliability for the output of recognition systems. These confidence features may be applied to the acoustic mode [3], or to language model and word graph [4, 5].

In this paper, we address the problem of confidence estimation for HMM-based speaker-independent isolate word recognition. We focus on word level confidence measures derived from purely acoustic features. This means that these features can be extracted from the output of a phonetic classifier, i.e., they can be derived from acoustic observations only. Such features based on language models are not utilized, but they may be able to be combined with acoustic features at a later stage in the processing in the future.

This paper is organized as follows. In section 2, we explain the implementation of the confidence measures. In section 3, we introduce the experimental setup and report the results. Finally we summarize our major findings and outline our future work.

2. IMPLEMENTATION

2.1 Overview

In this paper, the word confidence measures are computed as a post-processing stage after recognition in our recognition system. A hypothesized word composed of a sequence of hypothesized phones and the phonetic boundaries on the observations are derived from the recognition process. Some kinds of frame level likelihood ratios of all observations are calculated for the corresponding components of the word. The word confidence measures are computed via some combination of the frame level features.

2.2 Frame Level Likelihood Ratios

In our system, three kinds of frame level likelihood ratios are utilized: normalized log-likelihood (NLL) scores, modified normalized log-likelihood (MNLL) scores and normalized scores based on filler model. These features are all likelihood for a hypothesized phone normalized by some other likelihood.

As same as described in [6], the NLL score for a boundary model, c_i, given an observation, $\vec{x}$, is expressed as

$$C_{nll}(c_i|\vec{x}) = \log\frac{p(\vec{x}|c_i)}{p(\vec{x})} = \log\frac{p(\vec{x}|c_i)}{\sum_j p(\vec{x}|c_j)P(c_j)} \qquad (1)$$

where $P(c_j)$ is the prior probability for c_j.

In our isolate word recognition system, the prior probabilities are omitted, so $P(c_j)$ is left out in (1), or it can be considered that $P(c_j)$ for each c_j is uniform. Thus equation (1) is simplified as follow

$$C_{nll}(c_i|\vec{x}) = \log \frac{p(\vec{x}|c_i)}{\sum_j p(\vec{x}|c_j)} \quad (2)$$

The MNLL score is derived by replacing $\sum_j p(\vec{x}|c_j)$ with $\max_j p(\vec{x}|c_j)$ in (2):

$$C_{mnll}(c_i|\vec{x}) = \log \frac{p(\vec{x}|c_i)}{\max_j p(\vec{x}|c_j)} \quad (3)$$

The filler model is a fully connected all-phone network [7]. It can be evaluated using a Viterbi beam search. The best path determined by the evaluation of the all-phone network can be considered as alternative hypothesis. Hence, the normalized score based on filler model is expressed as

$$C_{filler}(c_i|\vec{x}) = \log \frac{p(\vec{x}|c_i)}{p(\vec{x}|c*)} \quad (4)$$

where $c*$ is the corresponding model for $\vec{x}$ in the arbitrary phone sequence estimated from the all-phone network.

2.3 Frame and Phone Based Word Confidence

After the frame level likelihood ratios of all observations are calculated, two kinds of word confidence are derived via two different strategies. The one is the frame based word confidence which is calculated as the mean log-likelihood ratio score across all acoustic observations in the word hypothesis. The other is the phone based word confidence. First the features are accumulated for the phones in the word, and then the phone based word confidence is calculated as the average of the scores of all phones in the word. Six different kinds of features are derived and experimentally compared. The features list here

- **NLL-frame:** frame based NLL score
- **MNLL-frame:** frame based MNLL score
- **Filler-frame:** frame based filler model score
- **NLL-phone:** phone based NLL score
- **MNLL-phone:** phone based MNLL score
- **Filler-phone:** phone based filler model score

2.4 Combination

Word level confidence measures can be derived from various features such as NLL scores, MNLL scores or filler model based scores. While it is possible that some single kind of word level features can provide adequate confidence measures, it should also possible to achieve improvements in performance by combining different kinds of features in an appropriate. Significant improvements have been achieved in comparison to some single feature [6].

This paper explores six kinds of phone based word confidence features, and two methods to combine these features to produce a single confidence score for the word: Fisher linear discriminant projection and a backpropagation neural network.

2.4.1 Word Level Features

Six kinds of word level features derived from three kinds of basic frame features are utilized. These features are:

- **Mean NLL Score:** The mean of all phone level scores in the hypothesized word based on NLL frame score.
- **Mean MNLL Score:** The mean of all phone level scores in the hypothesized word based on MNLL frame score.
- **Mean Filler Score:** The mean of all phone level scores in the hypothesized word based on filler model frame score.
- **Minimum NLL Score:** The minimum of all phone level scores in the hypothesized word based on NLL frame score.
- **Minimum MNLL Score:** The minimum of all phone level scores in the hypothesized word based on MNLL frame score.
- **Minimum Filler Score:** The minimum of all phone level scores in the hypothesized word based on filler model frame score.

The three minimum scores represent the lowest scores obtained across all phones. Generally, a low minimum score is an indicator that some portion of the word is not well matched to its hypothesized phonetic unit. As we know, only if all phonetic units are matched well, the hypothesized word can be considered to match well. So these minimum features may provide some different information to those mean features.

2.4.2 Fisher Linear Discriminant Projection

Fisher linear discriminant projection is a means of reducing the multi-dimensional confidence features down to a single confidence score by using a linear projection. The linear projection is determined from training data for a two class discrimination task (correctly and incorrectly hypothesized words). A projection vector $\vec{p}$ is learned from the development data containing correctly and incorrectly recognized word hypotheses. The projection vector is then applied to the word level features vector, $\vec{f}$, of any newly hypothesized word to produce a single word confidence score, C_{fisher}, as follow

$$C_{filler} = \vec{p}^T \vec{f} \quad (5)$$

2.4.3 Neural Networks

A 3-layer neural network classifier with sigmoidal activation function units and 20 nodes in hidden layer was trained, using a mean square error function and standard backpropagation. This backpropagation neural network is

also utilized to combine multi-dimensional confidence features.

3. EXPERIMENTS

3.1 Setup

Experiments are carried out on a speaker-independent Chinese name recognition system. 59 phoneme models and 1 silence model used in the system are estimated with isolate words and continuous utterances by 50 individuals (include male, female). The number of states of each phoneme model is 3, and that of the silence model is 1. The acoustic preprocessing employs 39 cepstral features, including first order and second order derivatives for every features vector.

The employed database contains Chinese name utterances by 20 individuals (10 male, 10 female) who each speaks four passes of 100 given names. There are 70 of these 100 words used as in-vocabulary words, and the rest 30 word used as out-of-vocabulary words. The recognition rate of words in vocabulary is 95.6%.

The database is separated into two parts. 10 speakers' data treated as training data are used to estimate the Fisher linear projection vector and the parameters of the backpropagation neural network, and other 10 speakers' data are used to evaluate confidence features in all experiments. There is no intersection in training data and test data.

To evaluate the performance of confidence features, hypothesized words are classified as correct or incorrect according to the true transcriptions of the utterances. The confidence score for each word is compared against a confidence threshold to judge the hypothesized word is accepted or rejected. The threshold can be varied to control the tradeoff between false acceptances (incorrect words but accepted) and false rejected (correct word but rejected). By varying the threshold, a curve can be plotted to offer a clear interpretation of the performance of confidence measures.

In the initial we compare phone and frame based features derived from different kinds of frame level likelihood ratios. Then we evaluate some features combined via the linear and the non-linear methods.

3.2 Comparison

Performance is measured for the six kinds of features described in section 2.3. As shown in Figure 1, phone based features outperform frame based features remarkably. The results are not surprising. As the results of Viterbi beam search, the duration of observations will be as short as possible for badly matched phonetic units. Phone based features emphasize the low scores for those badly matched phones in comparison to frame based features. And that the reliability of recognition results mainly depends on the worst matched components.

In three phone based features, filler model scores exceed MNLL scores slightly, and outperform NLL scores remarkably. The results indicate that it should be compared against to the best matched unit, not average units, to evaluate the reliability of hypothesized unit.

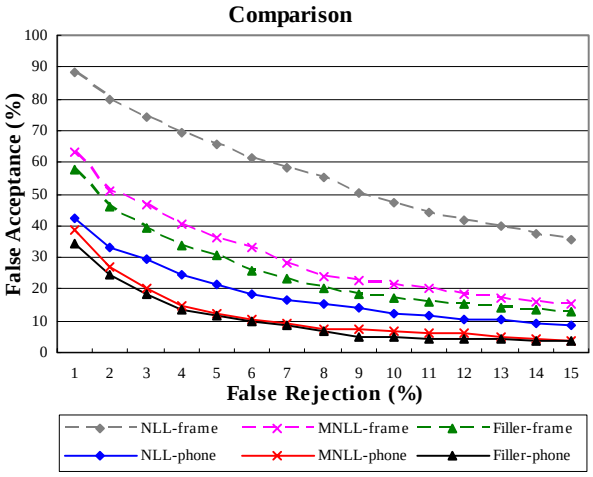

Figure 1. Comparison of features

3.3 Combination

Six features which have been described in section 2.4.1 are combined via Fisher linear approach and neural networks non-linear approach. Figure 2 shows that Fisher approach is not better than that filler model based feature, the best single feature. It is because that the features used in this experiment are similar, and the information provided by these feature are also similar. The Fisher approach is not effective to combine information of these features.

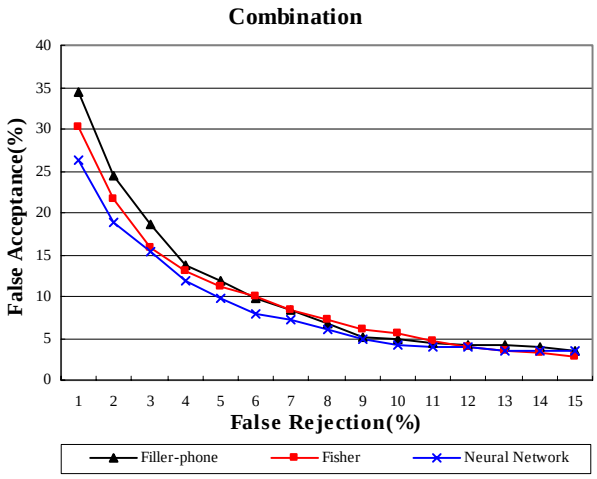

Figure 2. Combination of features

But neural network approach improves the performance in comparison to phone based filler model score which is the best single feature. When the false rejection rate is low, neural network approach reduces the false acceptance rate remarkably. E.g., when false rejection rate is in the range 1~8%, neural network approach reduce false acceptance rate about 15~30%. When false rejection rate is high, e.g. in the range 9~15%, the improvement is not very markedly, the false acceptance rate reduces about 10%. And when false rejection rate is higher than

15%, the performances of neural network approach and filler mode score are uniform. Some utterances are false accepted even some more information is utilized.

4. CONCLUSION AND FUTURE WORK

This paper has compared some kinds of confidence features, and evaluated the performance of the features combined via different methods. A phone based filler model score defeats other single features, and performance improvement can be achieved via a neural network approach.

In this paper, each phonetic unit is weighted equally when the word confidences are made. But a same confidence score maybe corresponds to different reliability for different phones. This will be our future work.

5. REFERENCES

[1] I. Bazzi and J. Glass, "Modeling out-of vocabulary words for robust speech recognition", *Proc. of ICSLP*, 2000

[2] T. J. Hazen and I. Bazzi, "A comparison and combination of Methods for OOV word detection and word confidence scoring", *Proc. of ICASSP*, 2001

[3] T. Schaaf and T. Kemp, "Confidence measures for spontaneous speech recognition", *Proc. of ICASSP*, 1997

[4] T. Kemp and T. Schaaf, "Estimating confidence using word lattices", *Proc. EUROSPEECH*, 1997

[5] F. Wessel, K. Macherey, and R. Schlueter, "Using word Probabilities as confidence measures", *Proc. of ICASSP*, 1998

[6] T. J. Hazen, S. Seneff and J. Polifroni, "Recognition confidence scoring and its use in speech understanding systems", *Computer Speech and Language*, 16, 2000

[7] A. Asadi, R.Schwartz, and J. Makhoul, "Automatic modeling of adding new words to a large-vocabulary continuous speech recognition system", *Proc. ICASSP*, 1991