http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 6 (6): 909-913, 2007
ISSN 1812-5638

Evaluatmg Words of Discrimmative Power in Automatic Speech Recogmition System

[1]V. Meenakshi and [2]V. Shanthi
[1]SRM Arts and Science College, No. 8/4, NH-1 Cheran Street,
Maraimalainagar, Kancheepuram District, 603 209, India
[2]St. Joseph's Engineering College, Chennai, India

Abstract: Confidence measures enable us to assess the output of a speech recognition system. The confidence measure provides us with an estimate of the probability that a word in the recognizer output is either correct or incorrect. A confidence measure is a quantitative estimate of a word's correctness. There are several approaches that incorporates one or two earlier methods to quantify the performance of confidence measures. It can be applied to improve recognition by incorporating extra information into the recognition process or weighting hypothesized words. It can also be used for predicting recognition accuracy and detection of recognizer failure. This study presents an approach that incorporates confidence measures based on both acoustic model and language model for measuring the word-based recognition reliability.

Key words: Confidence measure, out-of-vocabulary, acoustic model, language model

INTRODUCTION

In spite of the multiple efforts done to date on automatic speech recognition technology, its results are not perfect. Every time a recognized word sequence is considered, there is some degree of uncertainty about its correctness. Confidence Measures (CM's) (San-Segundo *et al*., 2000) represent a feasible way to express which of the recognized sequences are likely to be correct and which can be disregarded as incorrect. A rather simple technique, that has shown remarkable results, to generate confidence measures is known as Likelihood Score Ratio (LSR) (Rose and Paul, 1990). It is done by normalizing the likelihood score resulting from the Viterbi decoding process by the likelihood score produced by an alternative recognition network. Word-based confidence measures for speech recognition based on Hidden Markov Models (HMMs) have for some years now been an important research topic. While in the beginning the main interest was to detect Out-Of-Vocabulary (OOV) words and recognition errors in isolated word recognition and to use them as diagnostic tool (Eide *et al*., 1995), several new applications of these measures arose in the more recent years. Confidence measures (Hazen *et al*., 2000) represents a systematic way to express reliability of speech recognition results. The basic idea of most is representing the OOV words with sub-word units. For example, there is some work which uses a separate phone language model for the OOV words which is merged with the in-vocabulary word language model. An extension using automatically learned multi-phone units for use within the separate OOV model is presented (Bazzi and Glass *et al*., 2001). Further detail about this approach for modeling OOV words and experimental results on both JUPITER weather information domain and Broadcast News (Hub4) (Bazzi, 2002). Another approach which uses pairs of graphemes and phonemes as sub-word units (Galescu, 2003). There is also some work done in order to improve readability of text by transcribing OOV words based on phoneme-to-grapheme conversion (Decadt *et al*., 2002). A common approach to confidence measuring is to take profit of the information that several recognition-related features offer and to combine them, through a given compilation mechanism, into a more effective way to distinguish between correct and incorrect recognition results. Here we gave an idea of combining the two common approaches to confidence measures namely acoustic and language model based confidence measures to get the benefit of the both.

CONFIDENCE MEASURE

Let us consider a recognizer generating a sequence of hypothesized word tokens w*i*, i = 1, . . ., n. Associated with these tokens are a sequence of class labels C_i, where ci is defined as

C_i = {1, if w_i is correctly recognized,
0, otherwise.

Corresponding Author: V. Meenakshi, SRM Arts and Science College, No. 8/4, NH-1 Cheran Street, Maraimalainagar, Kancheepuram District, 603 209, India Tel: 9841163537

Associated with the ith token is its true posterior probability pi defined as:

$$p_i = P(c_i = 1/X),$$

i.e., p_i is the probability that the i th token is correctly recognized, given all the information about the recognition process, denoted by X. We also use the notation $p_i(x)$ to indicate that p_i is a function of the feature vector x, $x \in X$. The true posterior probability p_i is unknown a priori and must be estimated. A confidence measure for the recognized token w*i* is an estimate of the true posterior probability p_i

WORD-LEVEL FEATURES

For each hypothesized word in each *N*-best hypothesis, a set of word-level features are extracted from the recognizer to create a confidence feature vector. For this study different features, which have been observed to provide information about the correctness of a word hypothesis, were utilized. These features are:

Mean acoustic score: The mean log-likelihood acoustic score across all acoustic observations in the word hypothesis (where the acoustic score is a zero-centered loglikelihood ratio and not a raw density function score).

Mean acoustic likelihood score: The mean of the acoustic likelihood scores (not the log scores) across all acoustic observations in the word hypothesis.

Minimum acoustic score: The minimum (or worst) log-likelihood score across all acoustic observations in the word hypothesis.

Acoustic score standard deviation: The standard deviation of the log-likelihood acoustic scores across all acoustic observations in the word hypothesis.

Mean difference from maximum score: The average difference between the acoustic score of a hypothesized phonetic unit and the acoustic score of the highest scoring (or best) phonetic unit for the same observation across all acoustic observations in the word hypothesis.

Mean catch-all score: Mean score of the catch-all model across all observations in the word hypothesis.

Number of acoustic observations: The number of phone-level acoustic observations within the word hypothesis.

SCORE DIFFERENCES OF THE TOP WORD CHOICES

The score difference feature is obtained by re-scoring the best word segments with alternative words. These alternative words are extracted from the word lattice whose spectral features overlaps with the time span of the best word. If the scores from the best word and the maximum score of the alternative words are close, we can argue that these two words are confutable and the best word should be assigned a low confidence score. Otherwise, we can assign a high confidence score to it. Score differences can also be calculated between the best word score and average score in a phone lattice. The phone lattice can be generated with looser constraints than the word lattice. This phone lattice could absorb the garbage words and out of vocabulary words. The phoneme scores can better indicate how well the acoustic features fit the phoneme models. When word scores are used, each sub-word unit contributes equally to the total score. In fact, different sub-words should contribute differently to the overall decision.

WORD SCORE PER FRAME

The word score per frame is another attribute which indicates how well the acoustic segment of the word matches the HMM models. A high score indicates that the acoustic features fit the model well and a high confidence score should be given to the word. A low score means the acoustic features do not fit well into the model. This happens when the acoustic features are distorted or models were not trained to cover this situation. For large vocabulary speech recognition, the acoustic units are phoneme based. However, not all the models can be well trained, especially, the stop consonants. The word score per frame is just the average of phoneme scores. We can separate the models into categories each with different weights. The average score can be more indicative for the purpose of rejection.

CONFIDENCE MEASURES BASED ON THE ACOUSTIC MODELS

Conventional HMM-based speech recognizers choose the word or word-sequences W with the highest posterior probability estimate P(W/X) for an observed acoustic observation X. P(W/X) is split up into:

$$P(X|W)P(W) / P(X)$$

where:

P(X/W) = Modeled by sequences of HMMs
P(W) = Stochastical language models or finite grammars
P(X) = A scaling factor that can be omitted because it does not depend on W.

In order to transform the HMM-based likelihoods into posterior probabilities that can be interpreted as confidence measures, Bayes' rule can be applied. For a word w hypothesized for the interval [a,b] this yields:

$$C(w,a,b) = P(w|X_{ab}) = P(X_{ab}|w)P(w)/P(X_{ab})$$

Neglecting the word priors P(W) the confidence measure becomes the observation likelihood for the hypothesized word (the score estimated by the recognizer) weighted by an unconditioned observation likelihood. For a word w within the boundaries a and b (utterance of w is hypothesized to have caused the acoustic observation $X_{ab} = (X_a, \ldots X_b)$ a word-based confidence measure can be defined as function c(w,a,b) with a usual domain of [0, 1]. The higher this function is for a hypothesized word and its hypothesized boundaries, the more confident one can be that it really has been uttered within the interval. Often, the confidence measure of a word w, hypothesized for an acoustic observation sequence $X = (x_1,x_2,x_3,\ldots..x_n)$ is directly interpreted as the word's posterior probability P(w/X). However, especially in continuous speech recognition, the confidence of a word whose position has been hypothesized as the interval [a, b], is often understood as an estimate of the probability that the word starts somewhere around a and ends somewhere around b. In this case, the interpretation as mere posterior probability $P(w/X_{ab})$ is inadequate. Nevertheless, confidence measure and posterior word probability are strongly related.

CONFIDENCE MEASURES FOR CONTINUOUS SPEECH RECOGNITION

In continuous speech recognition measuring word-based confidence mainly faces two additional problems compared to isolated word recognition. On the one hand the hypothesized word boundaries are often incorrect. Ideal substitutions with correct word boundaries (but an incorrect word hypothesis) are rather rare. On the other hand recognition is based not only on the Markov models probabilistic distribution functions, but also on a language model that limits the possible word sequences (word pair grammars, finite grammars) or estimates each word sequence's prior probability (stochastic language models, n-grams).

MEASURES BASED ON THE RESPONSE OF THE RECOGNIZER

A very straightforward approach for measuring confidence in continuous speech recognition is to consider the speech recognizer as a black box, that we cannot or want not look inside, but to let it generate multiple hypothesis and to take the words emission probabilities as their confidence measure. The multiple hypothesis can be set up in various ways. Finke *et al.* (1996) proposed to perform multiple recognition procedures applying different scaling factors for weighing the language model based likelihoods against those based on the acoustic models. Kemp and Schaaf (1997), was compared to the somewhat cheaper alternative of simply taking the N-best or lattice-output of only one recognition procedure. No severe differences to the scaling factor approach have been measured. The multiple recognizer outputs are usually stored in word-lattices N-best lists. This approach provides very useful confidence estimates. (Often, these estimates are even considered as reference for other approaches.) However, this method that is only based on the output of the speech recognizer is extremely expensive with respect to computational time needed for decoding. Thus, for real-time applications, such as dictation or dialogue-systems, methods which do not require additional decoding computations are desirable.

MODEL-BASED MEASURES

Confidence measures that only need little additional computation consider the statistical models themselves. In the following, measures that use the acoustic hidden Markov models, the language model and those that aim to combine them are discussed separately.

MEASURES BASED ON THE ACOUSTIC MODELS

Confidence measures for continuous speech recognition merely based on the acoustic hidden Markov models have been investigated Williams and Renals (1997) for hybrid speech recognizers. It turned out that there is a noticeable degradation of these measures compared to the lattice-based ones. We experienced the same when applying the acoustical confidence measures for hypothesized words of a continuous speech recognizer. These measures neglect the language model

and can hardly cope with the fact that often the acoustic match is fine but the hypothesized word boundaries are wrong. Therefore, in the following we describe our approach to improve the acoustic model-based measures by the incorporation of language model information.

MEASURES BASED ON THE LANGUAGE MODEL

As a mere language model based confidence measure we propose to use an n-gram score weighted by the previous words confidence. In the bigram case, this is formulated by:

$$C_{lm}(w_2)=C_{lm}(w_1)P_{bi}(w_2|w_1)+(1- C_{lm}(w_1))P_{uni}(w_2)$$

for the hypothesized word_ " succeeding the hypothesized word _. Another possible measure that we made use of is the product of the specific word's likelihood and its reverse likelihood (the language model likelihood of the succeeding word). In the bigram case, this yields:

$$C_{lm}(w_2)=p(w_2|w_1)p(w_3|w_2)$$

for the hypothesized word w_2 between the words w_1 and w_3. They show that the measures contain little but at least some information on whether a hypothesized word is correct or not and motivate the combination with an acoustical measure, described in the following paragraph.

A COMBINATION OF ACOUSTIC AND LANGUAGE MODEL-BASED MEASURES

Combining multiple features to result in only one metric can be accomplished in many ways. Neural Networks are a common tool for deriving such a function from training data. Several features were combined this way for measuring the word-based confidence. As we only want to combine two measures, an acoustic and a language model-based one, the product of these two is sufficient. Unfortunately though, just as in continuous speech recognition when combining language model and acoustic model likelihoods, a scaling factor has to be introduced to cope with the different scale and quality of these measures. We use weighted geometric mean to combine the acoustic and duration CMs to obtain a hybrid phone confidence measure.

RESULTS

To test the proposed confidence measures, we conducted experiments using open source speech recognition toolkit.

DATABASE

Experiments were conducted on speaker independent continuous digit recognition task. The vocabulary consisted of 11 words (ten digits and oh). The training set consisted of 12549 digit utterances from 163 speakers and the test set consisted of 12547 utterances from 163 speakers. Mel Frequency Cepstral Coefficients (MFCCs) were used as features for speech recognition. The speech signal sampled at 16 kHz is frame blocked with a window length of 25 m sec and frame shift of 12 m sec. The 13-dimension MFCC vector, delta coefficients and delta-delta coefficients form a 39-dimensional feature vector. Triphone is used as the basic speech modeling unit, modeled by a 5-state left-right HMM. The output density in each state is modeled as mixture of 4 Gaussians. The word error rate of the ASR was 3.1% while the sentence error rate was 8.0%

CONFIDENCE MEASURE RESULTS

To evaluate the performance of the confidence measures, recognized words are compared against the correct transcription of the utterance and each word is classified as correct or wrong. The confidence measures of hypothesized digits are compared against a threshold to either accept or reject the digit. Receiver Operator Characteristics (ROC) plot of the detection rate versus the false acceptance rate - is plotted by varying the confidence threshold between 0 and 1. A confidence measure is good if it has higher detection rates at lower false acceptance rate. The proposed confidence measures are tested for its efficiency in detecting putative errors (erroneous but invocabulary words) as well as Out-Of-Vocabulary (OOV) words. The proposed confidence measures have out performed the baseline confidence measures. Also, the hybrid CM based on acoustic model as well as language model has performed better than the CM based only on acoustic model. To evaluate the performance of the proposed CMs in detecting the OOVs, we simulate recognition errors in the following manner. For every digit (oh, zero, . . . nine), we decode utterances containing the digit using language models not containing the particular digit. The results clearly indicate that the hybrid method has the best performance.

ASSESSMENT OF CONFIDENCE MEASURES

The quality of confidence measures largely depends on the task that they are set up for. A metric proposed by NIST is the relative entropy as described by Kemp and Schaaf (1997). It measures the amount of information that the confidence measure contains on the correctness

of the hypothesized words. The disadvantage of this measure is the need for a transformation of the confidence measure to the interval [0,1] and to an average of the recognizer's correctness in order to be interpreted as posterior word probabilities. This affords the knowledge and incorporation of this correctness figure and allows an additional degree of freedom in scaling that directly effects the metric.

CONCLUSIONS

Here, we described several different ways of evaluating the performance of confidence measures. We introduced the notion of characterizing confidence measures in terms of their discrimination power and bias. Here we presented an approach that incorporates confidence measures based on both acoustic model and language model for measuring the word-based recognition reliability. Even though the earlier approaches in acoustic and language model gives the measure of confidence in speech understanding system, the results clearly indicate that the hybrid method has the best performance.

REFERENCES

Bazzi, I. and J. Glass, 2001. Learning units for domain-independent out-of-vocabulary word modeling. Proceeding of Eurospeech Aalborg.

Bazzi, I., 2002. Modelling out-of-vocabulary words for robust speech recognition. Ph.D Thesis, MIT.

Decadt, B., J. Duchateau, W. Daelemans and P. Wambacq, 2002. Transcription of out-of-vocabulary words in large vocabulary speech recognition based on phoneme-to-grapheme conversion. Proceeding of ICASSP, Orlando, Florida, USA., 1: 861-864.

Eide, E., H. Gish and P.A. Jeanrenaud, 1995. Mielke: Understanding and improving speech recognition performance through the use of diagnostic tools. ICASSP., pp: 221-224.

Finke, M., T. Zeppenfeld, M. Maier, L. Mayfield, K. Ries, P. Zhan, J. Lafferty and A. Waibel, 1996. Switchboard Evaluation Report, DARPA.

Galescu, L., 2003. Recognition of out of vocabulary words with sub-lexical language models. Proceeding of Eurospeech Geneva, Switzerland.

Hazen, T., T. Burianek, J. Polifroni and S. Seneff, 2000. Recognition confidence scoring for speech understanding systems. Proceeding of the ISCA ITRW ASR2000 Workshop on Automatic Speech Recognition: Challenges for the New Millenium, Paris, France, pp: 213-220.

Kemp, T. and T. Schaaf, 1997. Confidence Measures for Spontaneous Speech Recognition. ICASSP: Munich, pp: 875-878.

Rose, R.C. and D.B. Paul, 1990. A Hidden Markov Model based keyword recognition system. Proceedings of 1990 ICASSP, Albuquerque, 1: 129-132.

San-Segundo, R., B. Pellom, W. Ward and J.M. Pardo, 2000. Confidence measures for dialogue management in the CU communicator system. Proceeding ICASSP, Istanbul, Turkey, pp: 1237-1240.

Williams, G. and S. Renals, 1997. Confidence measures for hybrid HMM/ANN speech recognition. EUROSPEECH, Rhodes, pp: 1955-1958.