Confidence Measure (CM) Estimation for Large Vocabulary Speaker-Independent Continuous Speech Recognition System

Yaxin ZHANG*, Raymond LEE**, and Anton MADIEVSKI**
Motorola Labs,
* China Research Center
** Australian Research Center
yaxin.zhang@motorola.com

Abstract

In this paper we report a study for confidence measure estimation in a large vocabulary speaker-independent continuous speech recognition system. A hybrid confidence measure estimation algorithm was developed. The final confidence measure consists of a number of confidence parameters which are generated from the different processing levels of the recognition system. A Parameter Reliability Analysis (PRA) algorithm was proposed to combine the confidence parameters to form the final confidence measure. The approach was applied to a large vocabulary speaker-independent continuous speech recognition system and obtained superior performance.

1. Introduction

As the speech recognition is deployed in an increasing number of applications, the systems need to be flexible enough to deal with a wide range of user responses and behaviors, such as heavy accent, hesitations, lip smacks and heavy breath, pause within a word, false starts and sounds like um's and ah's. Users may also respond with speech in which some or all of the words are out of the recognizer's vocabulary definition. Another common problem is that users often do not follow the required task or grammatical constraints, even though the system gives a very clear instruction. Recent applications on speech recognition technology demand for reliable systems tending to achieve correct results in a large number of tasks and environments. However, despite the large efforts done to date, current state of the art speech recognition systems do not achieve absolutely correct results. Every time a recognized word sequence is considered that there is, inherently to it, some degree of "uncertainty'" about its correctness. Therefore, it is necessary to build up a confidence measure of how corresponding to the input utterance is the resulting word sequence in order to affirm that the recognition output as correct or incorrect.

In this paper, we propose a new and advanced approach, which is a hybrid in the sense that it combines a set of features derived from different levels of recognition process, to address the above mentioned questions. This approach has been tested on the real-world problem and it outperforms any algorithm known to us.

2. The system

Figure 1 shows a block diagram of this CM estimation approach in a large vocabulary speaker-independent continuous speech recognizer. The technique consists of a number of sub-techniques in which each of them generates one or more parameters to deal with the various aspects of the CM for a large vocabulary task.

There are 10 parameters, denoted a_i (i=1~10) in the block diagram, generated from the different techniques. We will give the descriptions for these parameters in the following sections.

3. The speech recognizer

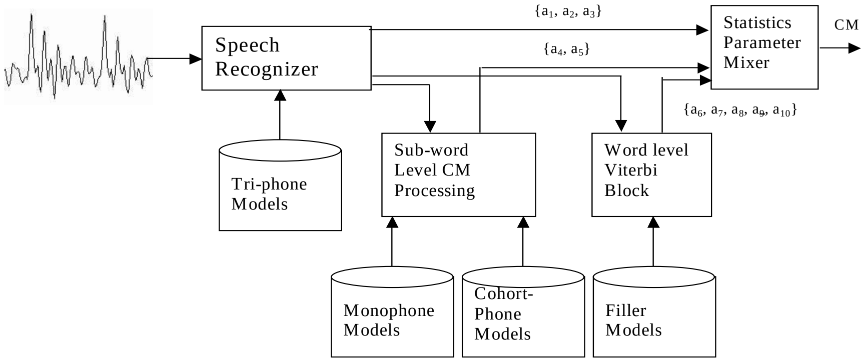
Our speech recognizer is based on the Motorola Lexicus continuous speech recognition system [6]. Tied state, multiple mixture Gaussians continuous density hidden Markov models (HMMs) are used for the tri-phones. The recognizer implements a one-pass time synchronous search for the decoder. Stochastic N-gram is used as the language model for the recognition.

The output for each utterance from the time synchronous search is a word-graph. N-best decoding [4] is then performed on the word-graph to generate the list of N best word sequence. The N=1 rank word sequence is the recognition result in which confidence parameters are extracted for each word in the sequence. Phonetic and word alignments for the first rank word sequence are also generated and are passed to other parts of the CM estimation system for phone based and word based confidence parameters estimation respectively.

First, a_1 is the acoustic score for the word from the recognizer's tri-phone model set. This provides an absolute measure of the likelihood that the speech segment of the word matches the tri-phone sequences of the word's pronunciation.

Second, a_2 is the voting score from the N-best list. Each word sequence in the N-best list is

Figure 1, Block diagram of the CM estimation



dynamically aligned to the best word sequence. The confidence parameter a_2 for each word in the best word sequence is then obtained by counting the number of times it is aligned to the same word in the N-best list. The count is then normalized by N giving a measure of how probable that a word is hypothesised in the N best list.

Third, the confidence parameter a_3 is obtained for each word from the stochastic N-gram. This provides a measure of how probable that the word is correctly recognized according to the linguistic knowledge provided by the N-gram models.

4. Sub-word level CM processor

A sub-word level CM processor is basically a two-pass phoneme based speech recognizer. The mono-phone model set has 41 key mono-phoneme models trained from all the correctly recognized phoneme segmentations of the training database. And the cohort phone set has 41 corresponding cohort-phoneme models trained from the incorrectly recognized phonemes in the training database. Therefore for each phoneme there are two corresponding models, one key mono-phone model and one corresponding cohort phoneme model.

For each phoneme-level segment of the input speech utterance, the CM processor fetches the corresponding mono-phone model from the mono-phone model set and cohort phoneme model from the cohort model set, and then calculates its likelihood with these two models. The two likelihood scores generated from these two models are denoted as S_k and S_c respectively. The parameter a_4 is defined as the accumulation of the log-likelihood ratio (LLR) of these two scores (N is the number of phonemes in a word).

$$a_4 = \frac{1}{N}\sum_{i=1}^{N} \frac{\log(S^i{}_k)}{\log(S^i{}_c)} = \frac{1}{N}\sum_{i=1}^{N} \log(S^i{}_k - S^i{}_c)$$

5. Word level Viterbi block

In this sub-processor we generate 6 word level parameters. The parameter a_5 is the length of the word segment represented by the numbers of frames or milliseconds. The other 5 parameters are the recognition scores for the 5 filler models.

Traditionally one comprehensive filler model is used to deal with the all garbage input. It is observed that one filler often fail to cover the wide range of the garbage inputs. In our system we design 5 fillers, each of them represents a class of garbage words with a certain word length in terms of number of phonemes of the word. For example, filler 1 is trained by all the training words which have one or two phonemes. Filler 2 is trained by all the three-phoneme training words, and so on. Finally, filler 5 is trained by all the words which have six or more phonemes. The filler models trained in such a way representing more explicitly the garbage inputs and therefore extract more out-of-vocabulary (OOV) words.

6. Statistical parameter mixer

To optimally extract the reliability information laid inside the 10 parameters, a Parameter Reliability Analysis (PRA) algorithm was proposed. The PRA algorithm assesses each the parameter based on the two classes of training words, correctly and incorrectly recognized words. The assessment includes a statistical analysis, and reliability estimation for the parameters.

Suppose that we have the probability distributions of a parameter a_i for the correctly recognized training data and incorrectly recognized training data. The means, denoted as μ_i and $\tilde{\mu}_i$, and the variances, denoted as δ_i and $\tilde{\delta}_i$, are estimated from the distributions respectively.

Since each the parameter shows a different individual performance represented by the degree of the separation of the two distributions, we need to weight each the parameter to get maximum usage of the advantage of the each parameter. In general the more the two distributions of a parameter separate, the better the parameter is. We would of course weight such parameter heavier. Suppose that we have a parameter a_i. Its mean and variance in the distribution for the correctly recognized training data are μ_i and δ_i. Its mean and variance for the incorrectly recognized training data are $\tilde{\mu}_i$ and $\tilde{\delta}_i$. We need to align the parameters to a unified reference point where we set the threshold to make acceptance/rejection decision. The reference point Δ_i for each parameter a_i would be

$$\Delta_i = \mu_i - W_i \times \delta_i = \tilde{\mu}_i - \tilde{W}_i \times \tilde{\delta}_i$$

Here W_i and $\tilde{W}_i$ are weights for the parameter a_i in the case of correctly recognized training data and incorrectly recognized training data respectively. The weights can be solved by the equations,

$$\begin{cases} W_i^{\;2} - \tilde{W}_i^{\;2} = 2 \ln \dfrac{\tilde{\delta}_i}{\delta_i} \\ W_i \delta_i + \tilde{W}_i \tilde{\delta}_i = abs\left(\mu_i - \tilde{\mu}_i\right) \end{cases}$$

The following equation aligns the all parameters and adds them to form a final confidence measure.

$$CM = \sum_{i=1}^{10} W_i \times (a_i - \Delta_i)$$

7. Experiments and results

We tested the CM estimation method on the Air Travel Information System (ATIS) task. We decided to use the ATIS sentences of the Macrophone database. The Macrophone database consists of telephone speech data and exhibits a lot of non-speech artifacts such as line noise, breath noise, lip smacks etc. and hence provides a suitable testing environment for our CM method. We used 6600 sentences from the database for training and 940 ATIS sentences for testing. The front-end feature vector consists of 12 MFCCs, frame energy and their delta and delta-delta parameters giving a total of 39 dimensions. The speech recognizer's model set consists of 3000 PDFs where each PDF consists of 10 Gaussian mixtures. Tri-gram language model trained from the ATIS text data was used as the grammar for recognition. A baseline recognition word accuracy of 91.9 % was obtained for the test sentences.

Three confidence parameters are generated from the speech recognizer as described in Section 3. N is chosen to be 10 for the N-best based voting score. We have done some experiment with different values of N and found that value of N greater than 10 did not improve the performance of the parameter. We also experimented with other possible confidence parameters that can be generated from the word-graph such as word and phone duration but found that they have very small correlation to the correct confidence scores.

To verify and reject the out-of-vocabulary, we applied the CM estimation method to a small telephone command/control word database, which has 47 speakers (22 males and 25 females). Each speaker utters to a telephone handset 28 command/control words which are out of the vocabulary definition in ATIS.

Figure 2 shows the results of word rejection performance in Receiver Operator Characteristic (ROC) plots, by using the above calculated confidence measure. False rejection means rejection of correctly recognized words, and correct rejection implies rejection of misrecognized words or OOV words respectively. Table 1 shows the results at some fixed false rejection points.

Figure 2. Word-Level Rejection Results in ROC

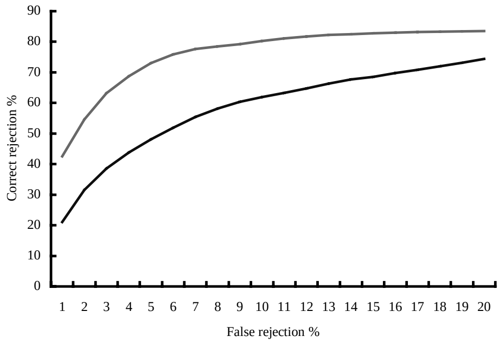


Table1. Word-Level Rejection Results in Table

False Rejection	Correct Rejection	
	Misreco-gnitions	OOV
1.0%	21.1%	42.5%
5.0%	48.1%	73.0%
10.0%	61.9%	80.3%
20.0%	74.4%	83.5%

8. Conclusion and discussion

In a large vocabulary speaker-independent continuous speech recognition system, confidence measure estimation is always a difficult task. When the system is deployed in the noisy or distortion environments, such as telephone environment, the recognition performance degradation will emphases the importance of confidence measure, which leads to the rejection of most misrecognized words. On the other hand, CM estimation is also seriously affected by the noise and speech distortions. In such cases, we often find that any single CM parameter does not contribute a steady and reliable verification performance for the recognition results. We observed that some CM parameter performs well in one condition but fails in other conditions. Based on these observations we proposed a robust hybrid confidence measure estimation algorithm, which makes use of a combination of a number of CM parameters estimated from different processing levels of the speech recognition system.

There are many different ways to obtain CM parameters. The selection of those parameters is based on two principles: (1) good single parameter performance and, (2) its orthographicability with other CM parameters. We can assess the performance of a single CM parameter by its PDF graph. The second point is also important. For example, a parameter generated from phoneme level acoustic score would not be orthographic with a parameter of word level acoustic score if the word score is an accumulation of phoneme scores. Except the general analysis, the PDF graph of the combined CM parameters also can be used to assess their orthographicability. Another important issue is the combination strategy for those selected CM parameters. Looking at the PDF of each individual CM parameter, we can see that the shapes and positions of these PDFs are different. To make the maximum utility of these CM parameters, we need to adjust the shape and position of the CM parameters to a unified shape and position. A Parameter Reliability Analysis (PRA) algorithm was created to handle the CM parameter combination.

We use this strategy to generate a confidence measure and make the word level recognition result verification. At 5% false rejection we achieved 48.1% correct rejection for the misrecognized words and 73.0% for the OOV words respectively. We believe this is a superior result for the comparable tasks.

The future work may focus on two aspects, looking at better CM parameter and investigating other parameter combination strategy. An artificial neural network may be used as the CM parameter mixer since we can use its memory capability to discover the hidden information in the CM parameters.

9. References

1.Young SJ and Woodland PC, “State Clustering in HMM-based Continuous Speech Recognition, Computer Speech and Language”, Vol. 8, No. 4, 1994, pp 369-384

2.Juang B-H, “Maximum-Likelihood Estimation for Mixture Multivariate Stochastic Observations of Markov Chains”, AT&T Technical Journal, Vol. 64, 1985, pp. 1235-1249

3.Davis and Mermelstein, “Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences”, IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. 28, pp. 357 – 366

4.Xavier Aubert and Hermann Ney. N, “Large Vocabulary Continuous Speech Recognition Using Word Graphs”, Proceedings International Conference on Speech and Signal Processing, 1995, pp. 49-55

5 Odell J.J., Valtchev V., Woodland P.C. and Young S.J., A One pass Decoder Design for Large Vocabulary Recognition, Proceedings ARPA Workshop on Human Language Technology, 1994, pp. 405-410

6. Sreeram V. Balakrishnan, “Effect of Task Complexity on Search Strategies for the Motorola Laxicus Continuous Speech Recognition System”, Proceedings of International Conference on Spoken Language Processing, 1998.

7. Sukkar, R.A, Chin-Hui Lee, “Vocabulary independent discriminative utterance verification for nonkeyword rejection in subword based speech recognition”, IEEE Transactions on Speech and Audio Processing, Volume: 4 Issue: 6, Page(s): 420 –429, Nov. 1996.