UTTERANCE VERIFICATION USING IMPROVED CONFIDENCE MEASURES BASED ON ALIGNMENT CONFUSION RATE IN CHINESE DIGITS RECOGNITION

Shilei Zhang, Danning Jiang, Yong Qin

IBM China Research Lab

{slzhang,jiangdn,qinyong}@cn.ibm.com

ABSTRACT

In this paper, we explore an approach to improved confidence measures based on a novel alignment confusion rate (ACR) which integrates alignment information from two different modeling unit sets in Chinese digits recognition system. Both Initial-Final (IF) phone set and Head-Body-Tail (HBT) models have proven to obtain good recognition performance for connected digit strings. These two different modeling can produce similar results but with different time-marked word boundaries. The objective of our proposed method is combining posterior probability with alignment confusion rate score provided by word alignment of IF-based results to HBT-based reference results that minimizes word error rate to get an effective confidence measure for utterance verification. The efficiency of the proposed algorithm is demonstrated with various experiments on data collected from car-kit microphone.

Index Terms—Chinese digits recognition, utterance verification, confidence measure

1. INTRODUCTION

Much of the application work has been developed for connected digit recognition, such as recognition of telephone, credit card, or personal identification numbers. Current spoken digits recognition systems are capable of achieving very high recognition accuracy in matched conditions. In practice, the limited amount of training data and the mismatch between training and testing environments nevertheless slightly reduce the recognition rate. Although the system generally obtains high recognition accuracy, there are always occurring some incorrect recognition results that can make it difficult to use in the application. Sometimes, it is very difficult to obtain significant recognition accuracy improvements using the current state-of-the-art speech recognition technology. Therefore, utterance verification (UV) based on confidence measures (CM) are used to evaluate reliability of recognition results which does not directly try to improve the recognition accuracy, but detects incorrect recognitions to increase the usability of practical user-friendly digits recognition systems. A good confidence measure can largely benefit automatic Chinese digits recognition systems in many practical applications. The application could use corrective action to the utterances which are likely to be erroneous.

Previous work on confidence measure has been reported. In [1], Jiang summarizes most research works related to confidence measures which have been done during the past 10 – 12 years and present all these approaches as three major categories. Because the causes of recognition errors can be many and varied, it would be difficult to use a single indicator of the reliability of an output to sufficiently flag all errors. Therefore, some investigators have attempted to combine multiple confidence measures [2, 3, 4, 5, 6]. Such combinations can take many different forms, but only focus on single modeling set and system. In this paper, we will compare and analyze the performance between phone-based IF units and word-based HBT units in Chinese digits recognition system. Then, we propose a hybrid approach combining conventional method and a novel CM score, called alignment confusion rate (ACR), which sufficiently exploit word alignment information provided by two different levels of modeling sets systems.

The rest of the paper is organized as follows: In section 2 we review the spoken digits recognition system and compare the recognition results of different modeling sets. Section 3 describes the conventional and proposed confidence measure schemes. In section 4 experiments on real data are presented and the results will be discussed. We will draw some experimental conclusions in section 5.

2. CHINSES DIGITS RECOGNITION

2.1. Two modeling unit sets

Modeling units play a very important role in state-of-art speech recognition systems [7]. The design and selection of them will directly impact the performance of final speech recognition engine. Correspondingly, there are several layers of units to be considered: phones, syllables and words [8, 9, 10]. Phone-based units, such as Initial-Final (IF) tonal models, are reasonable since they are more trainable and generalizable. Word-based units should be a good choice for Chinese digits. One such model set, referred to as Head-Body-Tail (HBT) models, has been used effectively in connected digit recognition. HBT models are a case of sub-word modeling where the sub-word units are not phonetic units, but rather, represent the beginning, middle, and end of a word. The center of each word, represented by the body model, is a context independent unit. Context dependency information is incorporated in the head and tail models. The above two unit sets for digits recognition are shown in Table 1.

Table 1. *Two modeling unit sets.*

Digit	IF units		HBT units		
0	LI	IG2	H0	B0	T0
1	Y	AO1	H1	B1	T1
	Y	I1	H1_1	B1_1	T1_1
2	GS	ER4	H2	B2	T2
3	S	AN1	H3	B3	T3
4	S	IH4	H4	B4	T4
5	W	U3	H5	B5	T5
6	LI	OU4	H6	B6	T6
7	QI	I1	H7	B7	T7
8	B	A1	H8	B8	T8
9	JI	OU3	H9	B9	T9

These two modeling system all uses content-dependent models to model inter-word dependencies, each unit being modeled by a 3-state left-to-right HMM with self-loops and forward transitions which is trained on features obtained from a feature space minimum phone error (fMPE) transformation [11]. Inter-word silence and silence at the beginning and end of the utterances are modeled by two separate phones. Both the context independent and context dependent models use speaker independent, continuous density HMMs, with varying number of states and mixtures.

2.2. Database

In this experiment, we used the spoken connected digit database collected over car-kit microphone in car under different speed conditions such as parked car, media speed, high speed which includes 61 hours' pure digital data from about 1189 speakers. All above data were used for training acoustic models for Chinese digits recognition. To evaluate the efficiency and robustness of the two modeling system, we conducted experiments on two testing corpus recorded from the same conditions as training data. The first corpus consisted of 3 data sets each containing about 580 utterances ranging in length from 2 to 5 digits, corresponding to three speed environments: parked car (T0_var1), media speed (T1_var1) and high speed (T2_var1) respectively; Speech data with variant length 6-15 digits used in this second group consisted of 3 data sets each containing about 580 utterances, corresponding to three speed environments: parked car (T0_var2), media speed (T1_var2) and high speed (T2_var2) respectively.

2.3. Experiment results and analysis

2.3.1. Results

Perceptual linear prediction (PLP) features were used as features for Chinese digits recognition. The speech signal sampled 16 kHz is frame blocked with a window length of 20 ms and frame shift of 10 ms.

Table 2. *The performance of WER and SER (%).*

Corpus	IF models		HBT models	
	WER	SER	WER	SER
T0_var1	1.9	7.64	2.1	8.50
T1_var1	1.8	7.48	1.7	7.30
T2_var1	2.3	9.26	1.8	7.37
T0_var2	1.3	13.64	1.4	14.20
T1_var2	1.9	18.96	2.0	18.27
T2_var2	2.3	22.84	2.3	22.10

The trained IF model set consists of 18 phonetic units which are represented by 300 state Gaussian mixture model containing 5291 Gaussians totally. The trained HBT model set consists of 33 phonetic units which are represented by 407 state Gaussian mixture model containing 12099 Gaussians totally.

Table 2 shows the performance comparison of HBT models and IF models using word error rate (WER) and sentence error rate (SER) as the evaluation metrics. As can be seen in the Table, both HBT models and IF models get low WER and SER and reach the similar performance, taking into account that the model size of HBT is slightly larger than that of IF models. Table 3 shows the three most common errors and percent of total errors using HBT models and IF models respectively. Generally speaking, the short digits strings dominate the errors. The common errors of the two modeling units in this corpus are remarkably similar to each other.

Table 3. Comparison of common errors.

Errors	IF models		HBT models	
	Digit	Percent (%)	Digit	Percent (%)
Insertions	5	49.7	5	49.2
	2	22.1	0	15.2
	0	15.4	2	12.9
Deletions	5	54.0	5	54.3
	2	30.5	2	27.4
	0	8.1	0	8.8
Substitutions	0->6	22.6	0->6	27.1
	2->8	19.6	2->8	21.1
	1->6	9.0	9->6	6.3

2.3.2. Analysis

The above detail analysis of insertion, deletion and substitution errors from the recognition results can find that two modeling unit sets generate the similar error distributions. Only about 10 percent of error utterances of IF-based recognition results are different from that of HBT models; otherwise, the important finding is that these two different level modeling tend to produce relatively different word boundaries even when they have the same recognition results. In other words, all the pairs of utterances can be divided into two categories depending on the alignment mapping relation with time-marked word boundary information provided by the two modeling sets: relatively similar alignment and relatively different alignment.

- If the two modeling sets all can sufficiently descript the acoustic characteristic of recognized utterances, two systems can produce correct results and similar time-aligned digits sequences. In other words, for these two sentences to be correctly aligned, the corresponding words are identical and almost fully overlapping.
- A relatively different alignment can result from two different reasons. Firstly, different modeling methods lead to diverse acoustic probability distributions; secondly, the match between the speech input and recognized utterance is poor, so the different modeling units will decode with less-trained probability distributions. When seeming the HBT-based results as reference and aligning it with IF-based results, we can analysis time alignment information based on corresponding word boundaries. Under the situation, even the two systems produce the same results, they will produce relatively bad time alignment; On the other hand, these two systems can also produce a small part of the recognition utterance results including diverse corresponding results which can be labeled as three types of token: substitution, insertion, deletion, respectively. The detail of the alignment process is given in the next section.

Based on the above analysis, we will exploit the alignment information provided by two levels of different unit sets to propose a novel CM score called alignment confusion rate.

3. PROPOSED CONFIDENCE MEASURES FOR UTTERANCE VERIFICATION

In this section, we will review the conventional CM methods based on posterior probability and propose new approach using alignment confusion rate. We also describe a method to combine the two CMs to obtain a hybrid improved CM.

3.1. Confidence score using posterior probability

As stated in [1, 12, 16], it is well known that the posterior probability in the standard maximum a posterior (MAP) decision rule is a good candidate for CM in speech recognition since it is an absolute measure of how well the decision is. However, it is very hard to estimate the posterior probability in a precise manner due to its normalization term in the denominator. In this study, word lattice-based method is adopted to approximate it. Usually, one word lattice is generated by the ASR decoder for every utterance. Then the posterior probability of each recognized word or the entire hypothesized sentence can be calculated based on the word-lattice from an additional post-processing stage.

Given a digits transcript U and an utterance X, the posterior probability based confidence measures denoted CM_{pp} is defined as:

$$CM_{pp}(U,X) = P(U \mid X) = \frac{P(X \mid U)P(U)}{P(X)} = \frac{P(X \mid U)P(U)}{\sum_{U_i} P(X \mid U_i)P(U_i)} \quad (1)$$

The posterior probability calculated from a word lattice can approximate the true $P(U \mid X)$ pretty well. Therefore, the resultant confidence measures generally achieve better performance than other common CMs.

3.2. Proposed score using alignment confusion rate

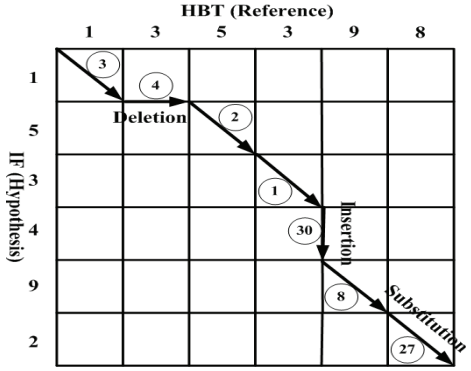

Figure 1: *Illustration of alignment task for digits results.*

When HBT-based results and IF-based results are regarded as reference transcriptions and decoding hypothesis respectively, an alignment process similar to recognition performance evaluation [13] can be carried out. Dynamic-programming alignment algorithm is applied to map the IF-based hypothesized words to HBT-based reference words that minimizes the word error rate. An example of optimal alignment is shown in Figure 1. The square matrix represents the entire possible set of alignments, and the heavy line path is the chosen alignment. The heavy line path in Figure 1 can be interpreted as follows: A diagonal move represents a mapping of a HBT-based reference word and an IF-based system word to each other. Words that are identical are correct, while words that are different are substitution tokens. A vertical move represents an un-mapped system word which is an insertion token in the IF-based system output. Similarly, a horizontal move in the path represents a deletion of a reference word.

Taking start and end times into account, pairs of mapped words can be time-aligned based on corresponding boundaries, which contains useful time alignment information. The numbers labeled inside the circles along the heavy line path are the dissimilarities, called alignment confusion, between pairs of corresponding word boundaries in number of frame length based on the above time alignment. As mentioned in section 2.3.2, to compute alignment confusion rate, it is reasonable to note that different alignment confusion computation should be used for different types of tokens based on alignment information. Recognized IF-based words are compared against the HBT-based reference transcription of the utterance and each word can be labeled as correct or incorrect involving time-alignments. For the "correctness" of an IF-based word, we define the alignment confusion as un-overlap time length; for incorrectness (substitution and insertion) of a word, we define the alignment confusion as the duration of the word; for deletion of a word, we define the alignment confusion as a small penalty value. Then we can get the confidence measure provided by alignment confusion rate as:

$$CM_{ACR} = 1 - ACR = 1 - \frac{all\ alignment\ confusion}{sentence\ duration} \quad (2)$$

3.3. Combination of two confidence scores

Since the two kinds of confidence scores are of different information, better performance will be achieved if combining them together. For computational reasons, we use linear combination to combine these two scores:

$$CM(U,X) = \alpha CM_{pp}(U,X) + \beta CM_{ACR}(U,X) \quad (3)$$

where α, β $(\alpha + \beta = 1)$ are the acoustic CM weight factors, which can be determined according to discriminative training on training data [14]. For simplicity, they are set to 0.5 in our experiments.

Utterance verification can be seen as a post-processing block to examine the reliability of the hypothesized recognition result. The task of UV has been formulated as a hypothesis test problem based on the above confidence scores. The confidence measures of hypothesized digits are compared against a threshold to either accept or reject the digit string. Under the framework of UV, we first propose two complementary hypotheses, namely the null hypothesis H_0 and the alternative hypothesis H_1 as following:

- H_0 : X is correctly recognized
- H_1 : X is wrongly classified

Then we test H_0 against H_1 with the following rules to determine whether we should accept the recognition result or reject it.

$$CM(U,X) \begin{cases} > \tau & : H0 \\ < \tau & : H1 \end{cases} \quad (4)$$

where τ is the critical decision threshold.

4. EXPERIMENTS RESULTS

In this section, we carry out experiments to evaluate the performance of proposed method and discuss the relevant results. Receiver Operator Characteristics (ROC) [15], plot of the detection rate versus the false alarm 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 alarm rate. A higher ROC curve indicates a better confidence measure method. Figure 2 is the ROC curves plot

of the baseline posterior probability, ACR and the proposed confidence measures on the three testing sets and digits recognition systems mentioned in section 2. The dashed thin and thick lines represent the accuracy of baseline posterior probability CM and ACR CM, respectively, while the solid line represents the accuracy of the proposed hybrid CM. As shown in the Figure, any single CM is difficult to reach a high ROC curve; on the other hand, we can consider posterior probability and ACR as different sources of information, so a better performance is achieved when combining the two confidence scores. On average, the proposed method gave more than 10% relative acceptance rate improvement at 20% false alarm rate over single CM method on all testing corpora.

In practical application, the confidence scores computed as above must be used to take the final decision of accepting or rejecting a hypothesis. For each recognition result, a confidence value describing the goodness of the match between the uttered and recognized word is computed. If the confidence of the recognized digits utterance falls below a pre-determined threshold value, the recognition result is rejected and the user is asked to repeat the last utterance. As the recognition accuracy requirements change depending on the application, the optimal threshold value must be adjusted individually for each system. Usually, a confidence score of 0.9 can ensure a large number of incorrect utterances can be rejected without deleting too many correct recognition results.

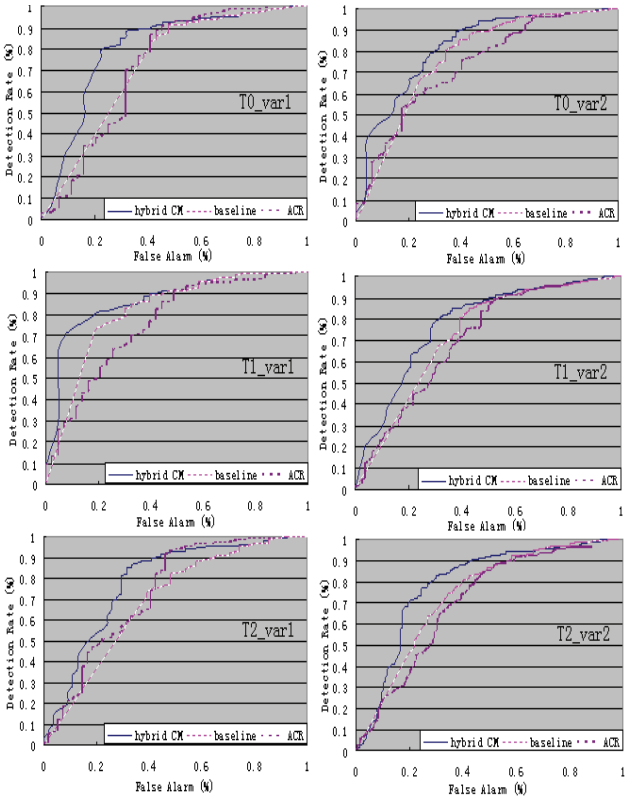

Figure 2: *curves for baseline CMs and proposed hybrid CM.*

5. DISCUSSION AND CONCLUSIONS

It is inevitable that a speech recognition system will make some errors. Therefore, it is desirable to improve the system performance through utterance verification. In this paper, we have investigated a new design CM called alignment confusion rate which integrate alignment information of two different modeling units depending on the optimal word-based alignment in Chinese digits recognition. The most important attribute of the technique is that it sufficiently exploits the timing information included in recognition results provided by the two modeling units and takes into account the effect of different alignment token types. Conventional CM score and alignment confusion score are combined using linear combination to obtain a hybrid improved CM. The hybrid CM shows significant improvement over the CM based on single indicator. Generally speaking, Taking into account time-alignment information obtained from two different levels of modeling systems can improve the efficiency of confidence measure. With respect to the computational complexity, the decoding speed of digits recognition is fast enough to realize the proposed CM method for practical application. Hence, the suggested utterance verification technique can helps us to develop user-friendly Chinese digits recognition systems.

6. REFERENCES

[1] H. Jiang, “confidence measures for speech recognition: A survey”, Speech communication 2005.

[2] D. Charlet, G. Mercier, G. Jouvet, “On Combining Confidence Measures for Improved Rejection of Incorrect Data”, Proc. of Eurospeech '01. Aalborg, pp. 2113-2116, 2001.

[3] Kyuhong Kim, Hoirin Kim and Minsoo Hahn, “Utterance Verification Using Search Confusion Rate and Its N-Best Approach”, ETRI Journal, vol.27, no.4, pp.461-464, Aug. 2005.

[4] G. Greenland, W. Wong and H. Kunov, “Improving utterance verification using additional confidence measures in isolated speech recognition interfaces”, IEEE International Conference on Acoustics, Speech, and Signal Processing 2005 on Volume 1, pp. 81 – 84, March 18-23, 2005.

[5] O. Viikki, K. Laurila, P. Haavisto, “A Confidence Measure for Detecting Recognition Errors in Isolated Word Recognition”, Proc. International Conference on Speech Science and Technology, pp. 67-72, Adelaide, Australia, 1996.

[6] B. Tan, Y. Gu, T. Thomas, “Word Level Confidence Measures Using N-best Sub-hypotheses Likelihood Ratio”, Proc. of Eurospeech '01. Aalborg, pp. 2565-2568, 2001.

[7] Chao Huang, Yu Shi, Jianglai Zhou, Min Chu, Terry Wang, Eric Chang, “Segmental Tonal Modeling for Phone Set Design in Mandarin LVCSR”, vol.1, pp.901–904 ICASSP 2004.

[8] J. Sturm and E. Sanders, “Modelling Phonetic Context using Head-Body-Tail Models for Connected Digit Recognition”, Proceedings ICSLP2000, Beijing, China, 2000.

[9] C. J. Chen, H. P. Li , L. Q. Shen and G. K Fu, “Recognize Tone Languages Using Pitch Information on the Main Vowel of Each Syllable”, Proc. ICASSP’2001, Salt Lake City, USA, 2001.

[10] W. Chou, C.-H. Lee, B.-H. Juang, “Minimum error rate training of inter-word context dependent acoustic model units in speech recognition”, Proceedings International Conference on Spoken Language Processing, pp. 439-442, 1994.

[11] D. Povey, B. Kingsbury, L. Mangu, G. Saon, H. Soltau, and G. Zweig, “fMPE: Discriminatively trained features for speech recognition”, in Proc. IEEE ICASSP, Philadelphia, USA, 2005.

[12] Yu Dong, Ju Yun-Cheng and Acero Alex, “An effective and efficient utterance verification technology using word n-gram filler models”, in ICSLP 2006, pp. 1408-1412, 2006.

[13] National Institute of Standards and Technology, “2004 rich transcription (RT-04F) evaluation plan”, Tech. Rep, 2004.

[14] K. Vertanen, “An overview of discriminative training for speech recognition”, Technical Report, University of Cambridge, 2004.

[15] G. Williams, “A Study of the Use and Evaluation of Confidence Measures in Automatic Speech Recognition”, Technical Report, CS-98-02, Dept. Comp. Sci., Sheffield University, 1998.

[16] Wessel F, Schluter R, Macherey K, Ney H., “Confidence measures for large vocabulary continuous speech recognition”, IEEE Trans. on Speech and Audio Processing, 9(3):288-298, 2001.