6th International Symposium on Chinese Spoken Language Processing, pp. 209-212, 16-19 December 2008, Kunming, China

LOCAL MISMATCH PHONE FOR CONFIDENCE MEASURE IN STANDARD AND ACCENTED CHINESE SPEECH RECOGNITION

CAO Wenxiao, LIU Yi, Thomas Fang Zheng

Center for Speech and Language Technologies, Division of Technical Innovation and Development
Tsinghua National Laboratory for Information Science and Technology, Beijing
caowx@cslt.riit.tsinghua.edu.cn, {eeyliu, fzheng}@tsinghua.edu.cn

ABSTRACT

High error rate in speech recognition is largely due to effects of phone local mismatch caused by unclear speaking or noises. In this paper, we propose an approach of using local mismatch phone to improve the reliability of confidence measure. The features of local mismatch phone can be extracted from the recognition phone sequence by computing occurrence frequency of each phone and comparing with a preset threshold. Occurrence frequency is defined as occurrence time of recognition phone in its frame best phone sequence divided by interval. Frame best phone is the symbol of HMM state at the end of maximum likelihood token at certain frame. The effectiveness of this feature is evaluated on standard and accented Mandarin speech databases. It gives significant Equal Error Rate reduction of 19.7% and 8.4%, respectively. In addition to fast computation, this feature is independent of acoustic model, and is convenient for combination with other features.

Index Terms—Speech Recognition, Confidence Measure, N-best, Local Mismatch

1. INTRODUCTION

As a key module of automatic speech recognition (ASR), confidence measure (CM) is able to provide a reliable estimation for recognition result, especially for practical ASR applications [1]. It is well known that the N-best based CM is a common method for unsupervised acceptance /rejection in ASR systems. Using N-best based CM is able to apply online rejection of recognition result so that the unrelated or noise speech input can be filtered.

In recent years, lots of research work has been done to investigate different algorithms based on N-best hypotheses [2, 3, 4, 5, 6]. In [2] and [6] an approach of using the difference in log likelihood between the first and second best hypothesis of state sequences was proposed. Another method shown in [4] estimated posterior probability directly on the N-best list without time information. N-best counting was used as counting the number of each word in the top hypothesis occurring in the same position in the N-best hypotheses [3]. In [5], a confidence score was proposed which is computed by summing the likelihood for all hypotheses that contains the keyword divided by the sum of that in N-best list. The above methods show that using the N-best hypotheses is an efficient way for CM. Most of the above methods try to calculate a likelihood ratio or posterior probability as confidence score. In addition, different CM approaches were studied and compared in order to achieve an efficient usage of posterior probability [3, 7].

In general, the posterior probability can be calculated either based on the confusion network output by decoder (e.g., n-best list, word graph or lattice [3]) or on the acoustic likelihood ratio between recognized hypothesis and an alternative one (e.g., a filler model or network, cohort set based anti-model [8]). However, there are still challenges in CM module for ASR system. Most of previous methods considered a confidence score over sentence. For example, they consider the posterior probabilities over the whole recognition result, and assume that the confidence contribution of different phones is identical. Although geometric combination method tends to give high weight to poorly matched phones, the local mismatch of a certain phone is still not distinguished from others. Hence, in [9], the method of weighting different phone scores according to their discriminative abilities is investigated and proved to be efficient.

We propose an approach of using local mismatch phone to improve the reliability of CM. The features of local mismatch phone can be extracted from recognition phone sequence by computing occurrence frequency (OF) of each phone in its frame best phone sequence over duration and comparing to a preset threshold. We first get the frame best phone for each frame, which is the symbol of HMM state at the end of maximum likelihood token at this frame. These phones construct a frame best phone sequence. Then we calculate OF for each recognition phone, which is defined as occurrence time of this phone in its frame best phone sequence divided by interval. This OF is compared to a preset threshold to determine whether the recognition phone is a local mismatch phone or not. Compared with other methods mentioned above, this approach concentrates on CM for possible local mismatch in recognition result.

This paper is organized as follows. In Section 2 we give the motivation and explanation of using the local mismatch phone in CM. The algorithms for confidence score calculation and feature combination are discussed in Section 3. In Section 4, Experimental results of using local mismatch phone on Mandarin speech recognition are presented. We conclude in Section 5.

2. OCCURRENCE FREQUENCY

2.1. Recognition phone occurrence frequency

Occurrence frequency is defined as the recognition phone occurrence time in its frame best phone sequence divided by interval. The frame best phone is the phone which is the symbol of HMM state at the end of the maximum likelihood token at certain frame.

In the token passing model [10], the log-likelihood of each token in a special frame is an accumulated value and increased by the logarithm of transition probability plus that of observation probability. For a recognition phone, all of its corresponding best tokens over its duration are partial path of the final hypothesized token path, i.e., the recognition sentence. Thus for this recognition sentence, most of its frame token is frame maximum likelihood token. In this approach we assume that frame token of recognition sentence rarely being the maximum likelihood token in a special interval possibly reveals a local mismatch.

2.2. Occurrence frequency and local mismatch phone

Local mismatch in recognition sentence is typically caused by unclear pronunciation or noise. We assume that unclear pronunciation or noise influences at least one phone unit. This results in locally bad ranks over frames of this phone duration. OF describes these locally bad ranks. It is an efficient way for estimating local mismatch probabilities.

Figure 1 provides an evidence for our viewpoint. It is a statistics on recognition result from one of our databases. In this statistics work we categorized phones in any recognition sentence into two categories: locally mismatched phone or not. Locally mismatched phone means the recognition phone is incorrect. The recognition sentence should be rejected. In Figure 1 the horizontal axis denotes different OFs of phone. Vertical axis denotes the number ratio of recognition phone in the category with appointed OF value. Figure 1 shows phones locally mismatched (the real line) and that not locally mismatched (the dotted line) have prodigious different distribution at OF value 0. Locally mismatched phone has much more probability to maintain a zero OF. As explained zero OF is more possibly caused by unclear pronunciation or noise. Thus different distributions at zero OF reveal different confidences of local mismatch.

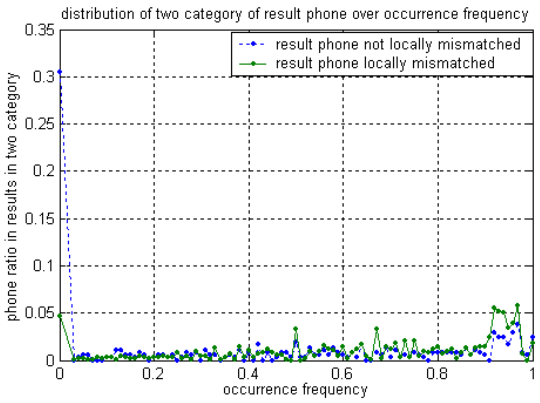

Figure 1: *Statistics over different occurrence frequencies*

3 CONFIDENCE MEASURE CALCULATION OF LCOAL MISMATCH PHONE

3.1. Confidence estimation for local mismatch phone

Recognition result can be represented by phone sequence. Assume $s_i(1 \le i \le n)$ is the best phone candidate at frame i, and n is the total frame number. $\{S_j, 1 \le j \le N\}$ is the recognition phone sequence. $t_s^{S_j}$ and $t_e^{S_j}$ represent the corresponding start and end frame for S_j. We define a symbol function for comparing two phones:

$$\mathrm{sgn}(s_t, S_j) = \begin{cases} 1 & S_j = s_t \\ 0 & otherwise \end{cases} \quad (1)$$

where equation denotes the central phones are the same when the comparing phones are triphones. Then we calculate OF for each recognition phone over its frame interval as:

$$Freq(S_j) = [\sum_{t=t_s^{S_j}}^{t_e^{S_j}} \mathrm{sgn}(s_t, S_j)] / (t_e^{S_j} - t_s^{S_j} + 1) \quad (2)$$

The OF of a recognition phone represents its confidence of being locally mismatched. Threshold is applied on each recognition phone to determine a phone is locally mismatched or not. Thus we define another function:

$$\mathrm{sgn}'(x) = \begin{cases} 0 & x > 0 \\ 1 & otherwise \end{cases} \quad (3)$$

Final confidence score is given by:

$$f(S) = \sum_{j=1}^{N} \mathrm{sgn}'(Freq(S_j) - \theta) \quad (4)$$

where S denotes recognition sentence and θ is a preset threshold. The final confidence score reveals the reliable of recognition result.

3.2. Combination of features

In this section we describe the combination of our proposed feature with other features. In this approach we use normalized likelihood scores (NLS) as baseline of CM. NLS has been investigated as an efficient feature for CM. The computation of NLS bases on Bayes theory.

$$P(W \mid O) = \frac{P(O \mid W)P(W)}{P(O)} \quad (5)$$

This posterior probability is used as confidence score. For each frame, maximum likelihood is given by:

$$f(t) = \max_{W_t}(P(O_t \mid W_t)P(W_t)) \quad (6)$$

Then it is normalized over phone duration:

$$P(p) = [\sum_{t=t_s^p}^{t_e^p} f(t)]/(t_e^p - t_s^p + 1) \quad (7)$$

Likelihoods of phones are summed up and divided by phone number to estimate approximate value of $P(O)$:

$$P(O) \approx \left[\sum_{p \in W} P(p) \right] / n \quad (8)$$

where W represents the recognition result and n is phone number in result. Thus via Eq.(4) we get the confidence score.

The combination uses a punishment strategy. Assuming $S_a(i)$ is confidence score from NLS. The higher the score, the more reliable the result. We use the proposed feature as a punishment exponent as follow:

$$S_f(i) = S_a(i) \times R^b \quad (9)$$

where R is a constant and b is the confidence score from proposed feature. The value of R belongs to [0,1]. The exponent b reveals the most possible amount of local mismatches in recognition result. This combination is different from other combination algorithm like decision tree, neural network and SVM etc. [1] but effective.

4. EXPERIMENTS

4.1. Experiment setup

We evaluated our approach in a Chinese desktop short phrase speech recognition task. There is no word n-gram in these short phrases so that we can isolate the effect of our approach without the influence from high-level information. The databases included two parts: one is standard Chinese speech (Putonghua), and the other is Minnan accented speech. The Putonghua database included 132 speakers (66 males and 66 females) with a total amount of 40 hours, and each speaker had 436 utterances. We used 200 utterances for training. The Minnan database included 35 speakers (18 males and 17 females) with a total amount of 8 hours, and each speaker had 436 utterances too. Training set was from Putonghua speech, and the Dev set was from Minnan and were used to adapt trained model [11]. Finally the adaptation model was tested on two databases. The basic statistics of data sets are listed in Table 1. All speech data were sampled at 16 kHz and 16 bit rate. The HMM topology was three-states, left-to-right without skips. The acoustic features were $13MFCC$, $13\Delta MFCC$ and $13\Delta\Delta MFCC$ plus Cepstral Mean Normalization and energy.

Data Set	Description and Composition
Training Set	Putonghua speech. 120 speakers, 200 utterances per speaker
Dev Set	Minnan accented speech. 20 speakers, 50 utterances per speaker
Testing Set 1	Putonghua speech. 12 speakers, 100 utterances per speakers
Testing Set 2	Minnan accented speech. 16 speakers, 75 utterances per speaker

Table 1: *Basic Statistics for Data Sets*

We used HTK [12] to perform our experiments. The syllable dictionary contained 65 Initials and Finals in total. There were 10629 physical tri-phone HMMs and 3803 physical states after sharing and merging states. State HMM had 14 Gaussian mixture compnents. Lexicon contained 402 toneless Chinese syllables, each syllable related to one pronunciation.

4.2. Experimental results

The final trained model has a sentence correct rate of 90.17% and phone correct rate of 95.90% on standard Mandarin testing set, respectively. These two values on accented testing set are 80.73% and 92.01%, respectively. In the experiments we combine our proposed feature with NLP using the algorithm described in Section 3.2. Algorithm using only NLP is considered as our baseline. Different parameter value in Eq.(9) is evaluated. For estimation of proposed feature, we focus θ at zero in Eq.(4). Equal Error Rate (EER) is used for performance evaluation and the same experiments are carried out on two databases.

Figure 2 shows the EERs of baseline algorithm and combination algorithm in our combination experiment. The horizontal axis denotes different R in Eq.(9), while vertical axis denotes corresponding EER. Figure 2 shows our combination algorithm reduces the EER from 0.239 to 0.192. This makes an absolute reduction of 0.047 and relative reduction of 19.7%. Figure 3 is the same experiment on accented Mandarin testing set. It provides an EER reduction from 0.239 to 0.219 with 8.4% relative reduction.

As the statistics result shown in Section 2.2, local mismatch phone has more probabilities of maintaining zero

OF. In the combination experiments, we use this feature as a punishment. This would lower the confidence of sentence with locally mismatched phones. Thus the combination improves baseline performance. The combination experiments proved that the proposed feature is effective for CM especially in combination with other features.

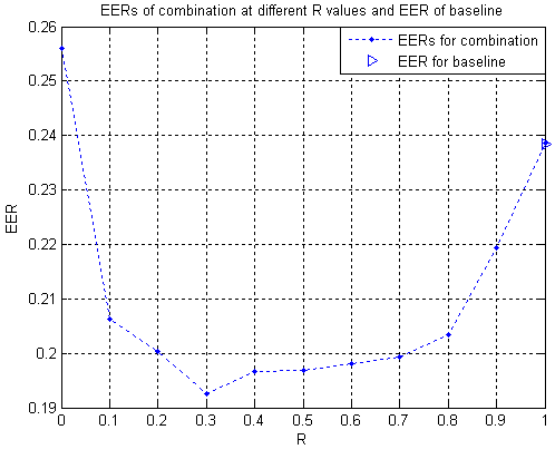

Figure 2. *EERs of combination at different parameter R and EER of baseline on standard Mandarin database*

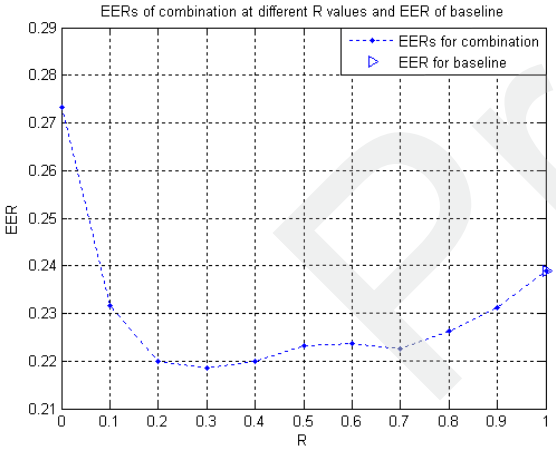

Figure 3. *EERs of combination at different parameter R and EER of baseline on accented Mandarin database*

5. CONCLUSIONS

We described an approach of using local mismatch phone to improve the reliability of confidence measure. The features of local mismatch phone can be extracted from the recognition phone sequence by computing occurrence frequency of each phone in its frames. The occurrence frequency can be obtained from phone recognition results and normalized over its interval. Especially, for each recognition phone sequence, its relevant occurrence frequency is compared to a preset threshold so as to determine the acceptance of the local mismatch phone. The effectiveness of proposed approaches was tested on Standard and Accented Chinese speech corpus. The results on standard Mandarin database have shown an efficient EER reduction of 19.7% and 8.4% on Minnan accented database, respectively. Compared with traditional methods, our local mismatch phone has low computation cost, independent from acoustic model, and is convenient to be integrated with other features of confidence measure features for improving its reliability.

6. ACKNOWLEDGMENTS

We would like to thank He Lei of Toshiba (China) Research and Development Center for his constructive suggestions and discussions of algorithm design as well as his help in running experiments. This work was partly supported by the joint research grant of Nokia and Tsinghua University.

7. REFERENCES

1. Jiang H., " *Confidence Measures for Speech Recognition: A Survey"*. Speech communication: p. 455-570, 2005.
2. Wendemuth A., Rose G., and Dolfing J.G.A. "*Advances in confidence measures for large vocabulary"*. in *ICASSP*. Los Alamitos. pp. 705-708, 1999.
3. Guo G., *et al*. "*A comparative study on various confidence measures in large vocabulary speech recognition"*. in *ISCSLP*. Hong Kong, China. pp.9-12, 2004.
4. Wessel F., *et al*. "*Confidence Measures for Large Vocabulary Continuous Speech Recognition"*. in *IEEE Trans. on Speech and Audio Proc*. pp.288–298, 2001.
5. Weintraub M. "*LVCSR Log-Likelihood Ratio Scoring for Keyword Spotting"*. in *ICASSP*. pp. 297-300, 1995.
6. Bernhard R., " *Obtaining confidence measures from sentence probabilities"*. EUROSPEECH-1997: p. 739-742, 1997.
7. Goronzy, *et al*. "*Prosodically motivated features for confidence measures"*. in *ASR*. pp.207-212, 2000.
8. Lleida E., et al. "*Out-of-vocabulary word modelling and rejection for keyword spotting"*. in *EuroSpeech* pp.1265-1268, 1993
9. Bouwman G., Boves L., and Koolwaaij J. "*Weight -ing phone confidence measures for automatic speech recognition"*. in *COST249*. pp.59-62, 2000
10. Young S. J., N.R.a.J.T., *Token Passing: a Conceptual Model for Connected Speech Recognition Systems*, Cambridge University: CUED, 1989.
11. Liu L.-Q., Zheng T. F., and Wu W.-H.. "*State-Dependent Phoneme-Based Model Merging for Dialectal Chinese Speech Recognition"*. in *Speech Communication*. 50(7):pp.605-615, 2008
12. Young S., *et al*., *The HTK book*: Cambridge University Engineering Department Speech Group, 1995.