Confidence Measure Based Incremental Adaptation for Online Language Identification

Shan Zhong, Yingna Chen, Chunyi Zhu, and Jia Liu

Department of Electronic Engineering,
Tsinghua University,
Beijing 100084, China
zhongshan00@mails.tsinghua.edu.cn

Abstract. This paper proposes an novel two-pass adaptation method for online language identification by using confidence measure based incremental language model adaptation. In this system, we firstly used semi-supervised language model adaptation to solve the problem of channel mismatch, and then used unsupervised incremental adaptation to adjust new language model during online language identification. For robust adaptation, we compare three confidence measures and then present a new fusion method with Bayesian classifier. Tested on the RMTS(Real-world Multi-channel Telephone Speech) database, experiments show that using semi-supervised language model adaptation, the target language detection rate rises from 73.26% to 80.02% and after unsupervised incremental language model adaptation, an extra rise over 3.91% (from 80.02% to 83.93%) is obtained.

Keywords: Language Identification, Language Model Adaptation, Confidence Measure, Bayesian Fusion.

1 Introduction

Nowadays, with the development of speech recognition system, many material uses have been proposed. Language identification (LID) is a technique to identify the language being spoken from a sample of speech by an unknow speaker.

As the global economic community expands, there is an increasing need for automatic spoken language identification services.Many techniques have been develeped rapidly for this task,such as Parallel Phone Recognizer followed by Language Modeling (PPRLM) [1], Gaussian Mixture Model (GMM) classifier [2], and so on [3]. But there is a common problem with all of these techniques. How our system runs robustly with the mismatch of different microphones, telephone channels, accent, background noises, and so on. Sometimes the mismatch will be pestilent in the real-world tasks. To solve this problem, we propose a novel two-pass adaptation system for online robust language identification by using confidence measure based incremental language model adaptation.

The paper is organized as follow. In section 2 we describe three confidence measures (CMs) in detail, and then fusing them with Bayesian classifier. Section 3

J. Jacko (Ed.): Human-Computer Interaction, Part III, HCII 2007, LNCS 4552, pp. 535–543, 2007.

discusses our language model adaptation methods. Experiment results are shown in Section 4 followed by the conclusion given in section 5.

2 Bayesian Fusion of Confidence Measures

In speech recognition, CMs are used to evaluate the reliability of recognition results. Comparing to Gaussian model classifier or max-likelihood classifier, the CM based method is more robust and with better performance in practical LID tasks.

In our LID system, with the difference of online garbage models, three kinds of CMs are employed. We described and compared these three CMs, and then presented a new CM fusion method by using Bayesian classifier. The experiment results will be given later.

2.1 Best_Lan Confidence Measure （CM_{BS}）

CM_{BS} is the difference between the log-likelihood of the first and second candidates in a N-best decoding approach, normalized by the length of the utterance.

$$C(L_i, X) = \frac{1}{n}[\log(P(X|L_i)) - \log(P(X|L_j))] \ . \tag{1}$$

where X represents the observed vector sequence, n stands for the frames of the utterance, L_i is the first candidate language and L_j represents the second.

This measure is a simple but classical confidence measure when the N-best decoding is available. Because the garbage model of the second candidate is the most competitive one, the confidence score can well distinguish the languages.

2.2 Avg_Lan Confidence Measure （CM_{AVG_N}）

The idea of CM_{AVG_N} is similar to CM_{BS}, but it calculates the distance between the first candidate language and the average of the residual N-best candidates.

$$C(L_i, X) = \frac{1}{n_i}\log(P(X|L_i)) - \frac{1}{N-1}\sum_{j=1, j\neq i}^{N}\frac{1}{n_j}\log(P(X|L_i)) \ . \tag{2}$$

To use CM_{AVG_N}, we can get better use of the information of the decoding result by arithmetical average of the N-best candidates. But from the physical model, it is obvious that the candidates with higher matching scores should be more competitive to the identification result. To deal with this problem, the third algorithm is shown below.

2.3 Post_Lan Confidence Measure （CM_{POST}）

The posterior probability $P(L|X)$ is a splendid confidence score when the observed speech vector sequence is X. By the Bayesian rule, $P(L|X)$ can be split up as follows:

$$P(L|X) = \frac{P(X|L)P(L)}{P(X)} = \frac{P(X|L)P(L)}{\sum_i P(X|L_i)P(L_i)} \quad . \tag{3}$$

If L is viewed as equal, $P(L|X)$ can be expressed:

$$P(L|X) = \frac{P(X|L)}{\sum_i P(X|L_i)} \quad . \tag{4}$$

Here, the posterior probability confidence measure is constructed by the N-best candidates $\sum_i P(X|L_i)$. If $\sum_i P(X|L_i)$ is considered as the online garbage model, the third confidence measure is proposed as CM_{POST}:

$$C(L_i, X) = \frac{1}{n_i}\log P(X|L_i) - \log\sum_{j=1}^{N}\exp(\frac{1}{n_j}\log P(X|L_j)) \quad . \tag{5}$$

2.4 Bayesian Fusion of Confidence Measures （CM_{fusion}）

Since the three CMs are of different information, better performance will be achieved after merging them together. These works are CM combining, shown in Fig.1. Recent efforts on CM combining include linear discriminant analysis (LDA) based CM combining [4], support vector machine (SVM) classifier [5], boosting [6], and others.

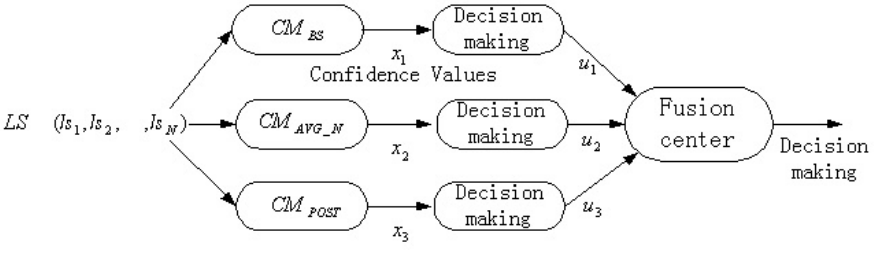

Fig. 1. Bayesian fusion of confidence measures

We apply the Bayesian classifier in which individual CMs are used as features for making a decision whether the recognition result is correct or incorrect as described in [7]. This approach is concerned about the estimation of the two classes and finds Bayes optimal decision boundary. From the Bayesian classification rule in binary cases, the following decision rule is expressed as:

$$\begin{cases} \omega_1 & if \quad \dfrac{P(x_1, x_2, \cdots, x_N \,|\omega_1)}{P(x_1, x_2, \cdots, x_N \,|\omega_0)} > Th \\ \omega_0 & if \quad \dfrac{P(x_1, x_2, \cdots, x_N \,|\omega_1)}{P(x_1, x_2, \cdots, x_N \,|\omega_0)} < Th \end{cases} . \tag{6}$$

where $x_i = j, j \in \{0,1\}$ means that the *i*th individual decision chooses the class ω_j. If we assume the independence of local decisions, then the left-hand side of the (6) can be factored as:

$$\prod_{S_1} \frac{P(x_i = 1|\omega_1)}{P(x_i = 1|\omega_0)} \prod_{S_0} \frac{P(x_i = 0|\omega_1)}{P(x_i = 0|\omega_0)} = \prod_{S_1} \frac{1 - P_{M_i}}{P_{F_i}} \prod_{S_0} \frac{P_{M_i}}{1 - P_{F_i}} . \tag{7}$$

where $s_k = \{i | u_i = k\}$ is the set of local decisions for ω_k. P_{M_i} and P_{F_i} represent the probabilities of miss and of false alarm of the *i*th local decisions, respectively. Substituting (7) into (6) and taking the logarithms leads to:

$$\begin{cases} \omega_1 & if \quad \sum_{i=1}^{N} [x_i \log \dfrac{1 - P_{M_i}}{P_{F_i}} + (1 - x_i) \log \dfrac{P_{M_i}}{1 - P_{F_i}}] > Th \\ \omega_0 & if \quad \sum_{i=1}^{N} [x_i \log \dfrac{1 - P_{M_i}}{P_{F_i}} + (1 - x_i) \log \dfrac{P_{M_i}}{1 - P_{F_i}}] < Th \end{cases} . \tag{8}$$

which is a weighted voting of local decision reflecting the reliability of each local decision maker.

3 Two-Pass Language Model Adaptation

The PPRLM is our basic system for the LID task [8]. The front-end HMM based phone recognizers tokenize the incoming speech utterance into a sequence of phones, and the probability that this sequence of phones generated by each language model is calculated. Finally, we can decide which language it is by the scores.

Thus the decoding sequences with high confidence scores can be used for adaptation. In our LID system, two parts of adaptation are contained. One is for the front-end phone recognizer, and the other is for the language model. Nevertheless,

experiments show that language model adaptation is much more effective than the adaptation of phone recognizer, with the need of less adaptation data and clearly lower computation cost at the meanwhile [9]. Because it is very appropriate for our online adaptation system, this paper is focused on the language model adaptation.

During language model adaptation, the language of each adaptation data has to be recognized first. Then, each speech utterance in the adaptation data set is decoded to several phone sequences automatically through each phone recognizer. As a result, the transcriptions of each speech utterance for its corresponding language models that follow each phone recognizer are gained. Finally, we use these new transcriptions to build an adapted language model with the linear merging method. For a word w_i in n-gram history h, with parameter λ,

$$P^{s+a}(w_i \mid h) = \lambda P^s(w_i \mid h) + (1-\lambda) P^a(w_i \mid h) \quad . \tag{9}$$

where $0<\lambda<1$ is the weight of the source model s and $1-\lambda$ is the weight of new adaptation model a. In fact, it can just be viewed as a maximum a posterior (MAP) adaptation strategy, given observation sample x, the MAP estimate is obtained as the model of the posterior distribution of θ denoted as $g(\cdot \mid x)$

$$\theta_{MAP} = \arg\max_{\theta} g(\theta \mid x) \quad . \tag{10}$$

3.1 Semi-supervised Language Model Adaptation

Different from supervised LM adaptation, semi-supervised LM adaptation means that only the languages of the adaptation data are available (see Fig.2). The transcriptions of each speech utterance for its corresponding language models which follow each phone recognizer are gained from front-end phone recognizers.

Giving the exact transcription of each speech utterance by manual work needs great patience and also great deal of time. But an experienced listener can estimate the language of the speech very easily and quickly. So the semi-supervised LM adaptation is effective, and its workload is reasonable.

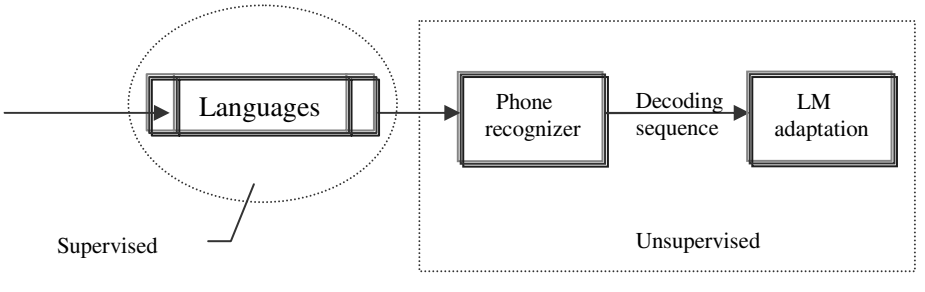

Fig. 2. Block diagram of semi-supervised LM adaptation

3.2 Confidence Measure Based Online Unsupervised Language Model Adaptation

After the semi-supervised LM adaptation, the performance has been greatly improved. But we can get further improvement by online unsupervised LM adaptation [10]. During the adaptation shown in Fig.3, we first send the incoming unknown speech utterance into our PPRLM system, and then use the language scores with high confidence to guide the model adaptation. A threshold for confidence score is set in order to ensure that almost all the utterance used for adaptation are correctly recognized.

Because of the online unsupervised adaptation, the LM is matching the testing domain step by step. Compared with the initial input utterance, the data tested later provide better accuracy. So after the whole testing process, the LM is optimized. In our LID system, the optimal LM is used to re-estimate the input utterance to get an additional accuracy improvement.

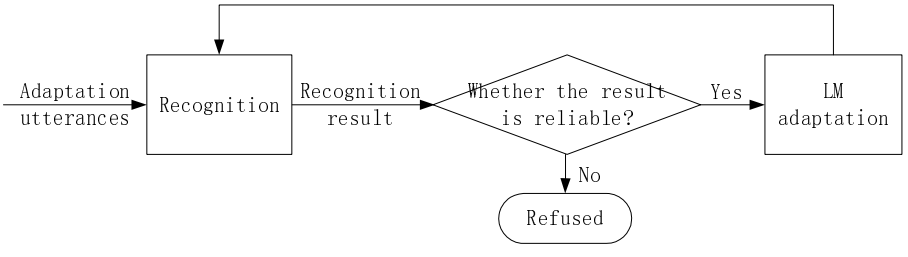

Fig. 3. Block diagram of CM based online unsupervised LM adaptation

4 Experimental Results

4.1 Speech Corpus

Real-world Multi-channel Telephone Speech (RMTS) is a speech corpus collected from different telephone channels in real-life phone-call situation. All the data come from one side of conversation and are presented as standard 8-bit 8 kHZ mu-law digital telephone data. There are almost 30 languages with three target languages in this corpus: Chinese, English and Russian. Each segment was prepared to use an automatic speech activity detection algorithm to identify intervals of speech, which were then concatenated and cut into short segments with duration of 35 seconds each to form the test segments. Thus, we can use RMTS to evaluate the goodness of our proposed system in different telephone channels.

4.2 Comparison of Confidence Measures

Fig.4 shows that CM_{BS} outperforms CM_{AVG_N} and CM_{POST}. It is because that CM_{BS} stands for the distinction of the most competitive two candidates. CM_{POST} is

approximate to the posterior probability by sufficient models. But actually in our experiment, only three garbage models are offered. That is why CM_{POST} does not perform well here.

The experiment indicates us how to adjust the weighting factors when fusing the above three CM. CM with better detection rate should have larger weight. With the well adjusted parameters, CM_{fusion} greatly improves the detection performance and gets the best result. So the following unsupervised LM adaptation experiment is based on CM_{fusion}.

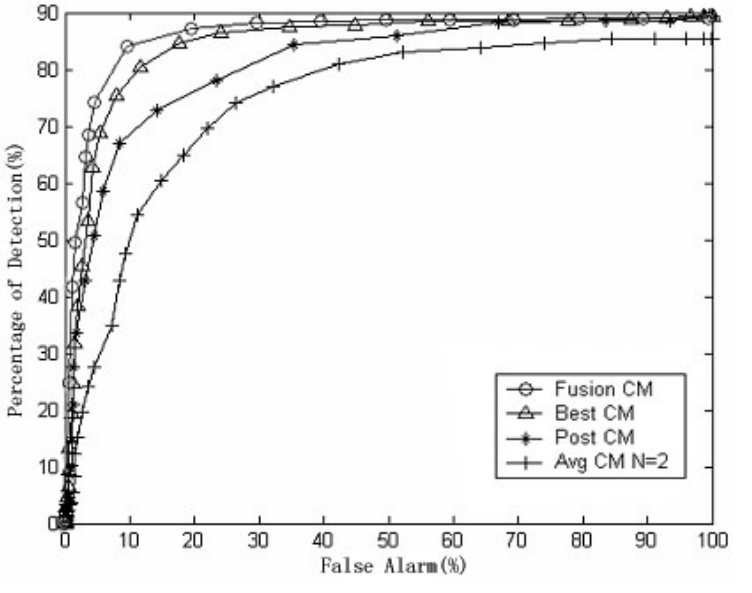

Fig. 4. Language detection rates with different confidence measures

4.3 Language Model Adaptation

In our LM adaptation process, semi-supervised LM adaptation is first used in three different telephone channels. Fig.5 shows that after the process, the detection rate of the target language rises from 73.26% to 80.02%, and the average rises from 70.85% to 77.45%.

In the course of testing, the online unsupervised LM adaptation works. Illustrated in Fig.6, at fist, the adaptation is inconspicuous for the sparseness of the data accumulated. But as the test data accumulated,3 hours in the experiment, the LM is matching the testing domain and the detection rate rises gradually. An extra rise of the target language detection rate over 3.91% (from 80.02% to 83.93%) is obtained, and the average over1.91% (from 77.45% to 79.36%).

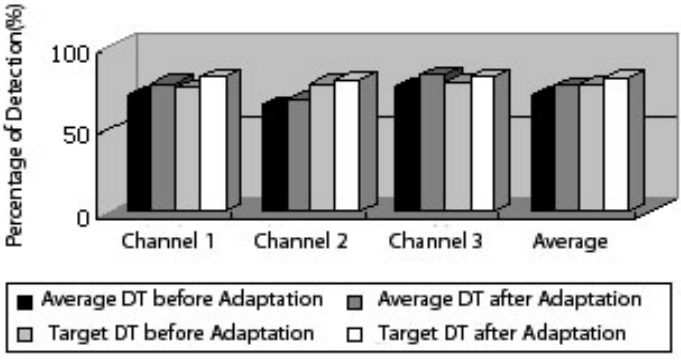

Fig. 5. Performance of semi-supervised LM adaptation in different telephone channels

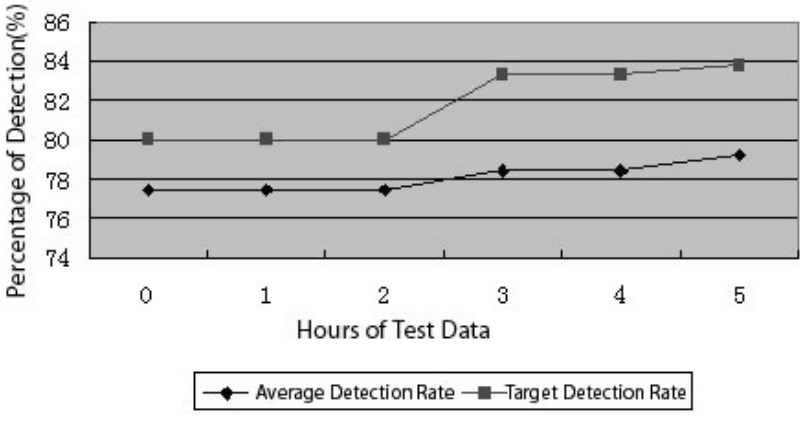

Fig. 6. Detection rates during online unsupervised LM adaptation

5 Conclusions

This paper presented an improved two-pass adaptation method for online language identification by using confidence measure based incremental language model adaptation. The experiments results show that this method can clearly improve the system performance, and make it more robustly in different channels. However, we should be careful in choosing good CM features for combining so as not to raise the estimation problem. For further improvement, our future work will apply this method not only to the language model, but also to the acoustic model.

Acknowledgements. This project is supported by the National Natural Science Foundation of China (NSFC) (60572083).

References

1. Muthusamy, Y.K., Barnard, E., Cole, R.A.: Reviewing automatic language identification. IEEE Trans. Signal Proc. Magn 11(4), 33–41 (1994)
2. Torres-Carrasquillo, P.A., Reynolds, D.A., Jr Deller, J.R.: Language identification using Gaussian mixture model tokenization. In: Proc. ICASSP '02, vol. 1, pp. 757–760 (2002)
3. Zissman, M.A., Berkling, K.M.: Automatic language identification. Speech Communication 35(1-2), 115–124 (2001)
4. Kamppari, S., Hazen, T.: Word and phone level acoustic confidence scoring. In: Proc. ICASSP '00, Istanbul, Turkey, pp. 5–9 (2000)
5. Zhang, R., Rudnicky, A.: Word level confidence annotation using combinations of features. In: Proc. EUROSPEECH '01, Aalborg, Denmark, pp. 2105–2108 (2001)
6. Moreno, P.J., Logan, B., Raj, B.: A boosting approach for confidence scoring. In: Proc. EUROSPEECH '01, Aalborg, Denmark, pp. 2109–2112 (2001)
7. Kim, T-Y., Ko, H.: Bayesian fusion of confidence measures for speech recognition, Signal Processing Letters, IEEE, vol. 12(12), pp. 871–874 (December 2005) Digital Object Identifier 10.1109/LSP.2005.859494
8. Shizhen, W., Jia, L., Runsheng, L.: Language Identification Using PPRLM with Confidence Msasures. In: Proceeding of ICSP2004, pp. 683–686 (2004)
9. Chen, Y., Liu, J.: Language Model Adaptation and Confidence Measure for Robust Language Identification. In: Proceeding of ISCIT 2005, vol. 1, pp. 270–273 (2005)
10. Bacchiani, M., Roark, B.: Unsupervised Language Model Adaptation. In: IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, vol. 1, pp. 236–239 (2003)