6th International Conference on Spoken
Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

CONFIDENCE MEASURE BASED UNSUPERVISED SPEAKER ADAPTATION

Husheng Li, Jia Liu, Runsheng Liu

Electronic Engineering Department,
Tsinghua University
Beijing, 100084, P. R. China

ABSTRACT

Unsupervised adaptation is the most convenient mode for the user of a speech recognition system. However the performance of unsupervised adaptation is worse than that of the supervised mode because of the recognition errors. This paper introduces a kind of word-lattice based confidence measure to evaluate the reliability of the recognition result and discard the uncertain parts from the adaptation speech. Experiments demonstrate that the confidence can improve the performance of unsupervised adaptation considerably.

1. INTRODUCTION

Speaker adaptation [1] is a powerful means to improve the performance of a speech recognition system with suitable amount of the user's speech. According to whether the correct labels of the speech for adaptation training are known to the system, speaker adaptation can be divided into two modes: supervised and unsupervised. Unsupervised adaptation is the most convenient mode for the user because the user can be adapted during the recognition on-linely, unconscious of the adaptation procedure. However experiments show that the unsupervised adaptation performs worse than the supervised mode due to the wrongly labeled speech unit. Therefore it is desirable to delete these errors from the adaptation speech to remove the negative effect of them.

In recent years, the technique of confidence measure is developed in the field of speech recognition [2][3][4][5]. The confidence measure can be used to evaluate the reliability of the recognition results. Thus we may utilize the confidence measure to find the uncertain part of the recognition result and enhance the capability of the unsupervised adaptation

The paper is organized as follows. The 2nd section introduces the word lattice based confidence measure. The mechanism of unsupervised adaptation is provided in the section 3. The experiment results and conclusion are given in section 4 and section 5.

2. WORD LATTICE BASED CONFIDENCE MEASURE

There are many ways to calculate the confidence measure of the recognition result. We can divide these methods into two groups. One is from the viewpoint of hypothesis test [2][3], in which the most powerful test takes the form of the likelihood ratio of null hypothesis H_0 and alternative hypothesis H_1. The other is based on the pattern recognition with a linear transformation [4] or neural network [5] classifier to distinguish the recognition result into two regions: reliable or unreliable.

In this paper we choose the form of likelyhood ratio as the confidence measure.

$$C(X) = \frac{P(X \mid H_0)}{P(X \mid H_1)} \qquad (1)$$

where X is the speech unit to be tested. The difficulty of calculating (1) lies in the denominator, which can be estimated with the anti-keyword model [2] or on-line garbage model [3]. However the computation of $P(X \mid H_1)$ places heavy burden to the adaptation procedure, especially for the on-line adaptation. In this paper we utilize the information in word lattice to calculate (1), which largely reduces the computation and makes the system more practicable.

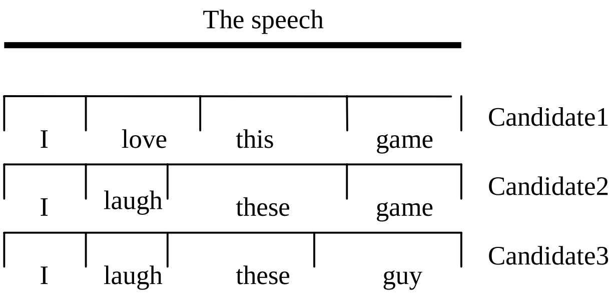


Figure1 An example of Word Lattice

The word lattice is the result of N-Best search. Figure 1 shows an example of word lattice. Let T be the total number of frames, N be the number of candidates, $\{x_t\}_{t=1,2,\ldots,T}$ be the speech features, $\{s_t^n\}_{t=1,2,\ldots,T}$ be the matching score sequence of candidate n, and $\{m_t^n\}_{t=1,2,\ldots,T}$ be the corresponding speech models, where $s_t^n = \log(P(x_t \mid m_t^n))$ is satisfied. The dividend in (1) can be obtained directly from the matching score of each frame, which can be stored during the recognition stage,

$$\log\left(P(X \mid H_0)\right) = \log\left(\prod_{t=t_s}^{t_e} P\left(x_t \mid m_t^1\right)\right) = \sum_{t=t_s}^{t_e} s_t^1 \qquad (2)$$

where t_s and t_e denote the starting and ending frame of speech unit *X*.

In the denominator H_1 denotes all the models other than the model in the 1st candidate. We can confine the range of these alternative models to the models in the corresponding place in the word lattice. Thus the score of alternative hypothesis H_1 can be calculated with the matching scores in the candidates except the best one.

$$\log\left(P(X \mid H_1)\right) = \frac{1}{N-1}\sum_{n=2}^{N}\sum_{\substack{t=t_s \\ m_t^n \neq m_t^1}}^{t_e} s_t^n \qquad (3)$$

Then the confidence measure of speech unit *X* can be obtained by (4),

$$C(X) = \sum_{t=t_s}^{t_e} s_t^1 - \frac{1}{N-1}\sum_{n=2}^{N}\sum_{\substack{t=t_s \\ m_t^n \neq m_t^1}}^{t_e} s_t^n \qquad (4)$$

3. UNSUPERVISED ADAPTATION SCHEMES

The flow chart of the confidence measure based unsupervised adaptation is shown in figure2.

In this paper we adopt MLLR (Maximum Likelyhood Linear Regression) [1] as the adaptation algorithm. The main procedure of adaptation is to calculate the statistics in (5) and (6) for each transformation class which is clustered with likelyhood measure [1] before the adaptation.

$$G^{(i)} = \sum_{m=1}^{M} \frac{1}{\sigma_i^{(m)2}} \xi^{(m)} \xi^{(m)T} \sum_{t}^{T} \gamma_m(t) \qquad (5)$$

$$\vec{k}^{(i)} = \sum_{m=1}^{M}\sum_{t=1}^{T} \gamma_m(t) \frac{1}{\sigma_i^{(m)2}} x_t^i \xi^{(m)T} \qquad (6)$$

where $\xi^{(m)}$ and $\sigma^{(m)}$ denote the mean vector and the diagonal elements of the *m-th* gaussian, x_t means the speech feature of the *t-th* frame and $\gamma_m(t)$ means the probability that x_t is produced by the *m-th* gaussian, and *M* is the number of gaussians in the transformation class. Then the transformation matrix can be solved with the statistics in (5) and (6).

We can calculate the confidence measure on different levels of speech units. In this paper we explore the units of semi-syllable, syllable and sentence. The smaller the unit is, the more efficiently the system utilizes the adaptation speech while the difficulty of rejection also increases. When rejecting an unreliable unit, we simply set $\gamma_m(t) = 0,\ m = t_s, \ldots t_e$.

4. EXPERIMENT RESULTS

4.1 Baseline system

In this paper the speech-controlling private branch exchange (SCPBX) system [5] is used to prove the validity of confidence measure based unsupervised adaptation. The SCPBX system can recognize speaker-independent continuous speech commands, such as “Please dial Mr. Li Husheng in the microwave division”, over telephone channel based on medium vocabulary (<1000 words).

The speech feature used in SCPBX is a 31-dimension vector, which contain 14 MFCC, 14 delta MFCC, normalized energy, delta energy and 2nd order delta energy. The acoustic model is based on the unit of semi-syllable, whose parameters are trained with the 863-Database. The effect of the telephone channel is removed with Cepstral Mean Subtraction (CMS). A set of finite state grammar is used as the language grammar and figure 3 illustrates an example of such grammar

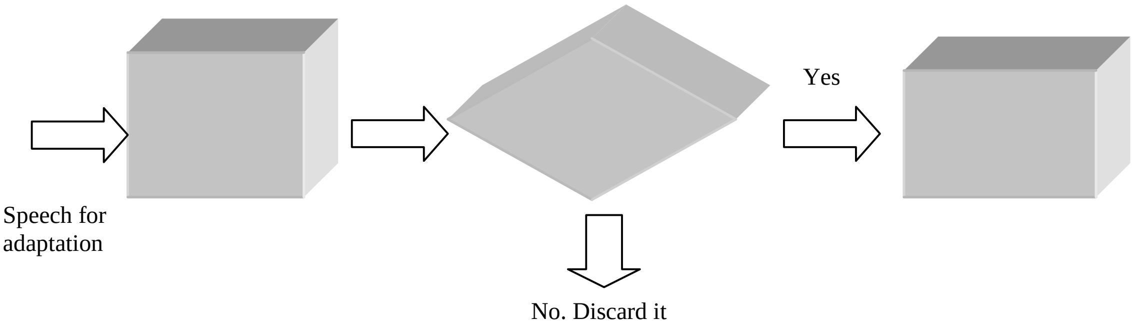


Figure2 The flow chart of confidence measure based unsupervised adaptation

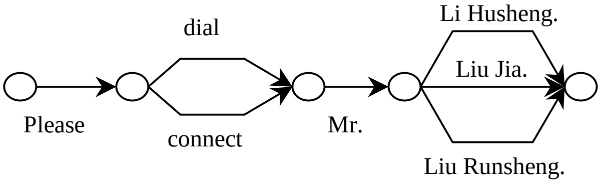


Figure 3 The illustration of the grammar in SCPBX

The system adopts the Multi-Subtree Frame-Synchronous Beam Search Algorithm [6] as the searching algorithm. The average SI error rate without rejection is 24.3%. We choose the pronunciations from 4 speakers to prove the validation of the algorithm proposed in this paper. The performance of the 4 speakers is far below the average due to the heavy accent and the noise in the channel. Each speaker pronounces 205 sentences with 80 sentences to be adapted and 125 sentences to test the performance.

4.2 Performance of confidence measure

Figure 4 shows the ROC (receiver operation curve) of the rejection on different speech units.

First we tested the effect of increasing the number of candidates in 4(a), from which we can draw the conclusion that it is useless to increase the number of candidates to more than 5. Therefore we set the candidate number to 5 in all the rest experiments.

From figure (a) and (b) we can see that the rejection performance of semi-syllable and syllable is about the same. 4(c) proves that the rejection on the sentence performs better than that of the semi-syllable and syllable due to the more information contained in a sentence.

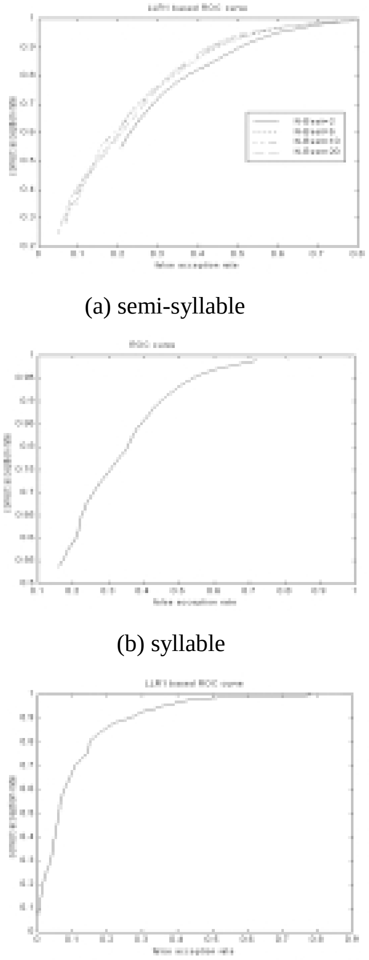
4.3 The performance of adaptation

First we tested the baseline performance of the adaptation system. Table1 lists the baseline error rate before and after the batch-mode adaptation under supervised and unsupervised mode. For each speaker, 80 sentences are used as the adaptation speech and the error rate is evaluated with the rest 125 sentences.

Table1 Baseline performance of adaptation

Speaker	1	2	3	4
Baseline	29.6%	45.6%	29.6%	41.6%
Unsupervised	20.0%	29.6%	23.2%	28.8%
Supervised	11.2%	24.0%	19.2%	26.4%

Table1 demonstrates that the difference between the supervised and unsupervised adaptation is non-negligible if the speaker independent error rate is high. Hence the rejection on the unreliable speech unit is necessary.



(a) semi-syllable

(b) syllable

(c) sentence

Figure4 ROC of rejection on different speech units

Then the confidence measure is introduced to remove the negative effect of the wrongly recognized speech unit. Table2, table3 and table4 show the error rate of the unsupervised adaptation with the rejection unit of semi-syllable, syllable and sentence, where *Th* means the threshold.

Table2 semi-syllable

Speaker	1	2	3	4
Th = 1.7	18.4%	27.2%	20.0%	28.0%
Th = 2.7	17.6%	28.0%	18.4%	27.4%
Th = 3.7	22.4%	26.4%	20.0%	26.4%

Table3 syllable

Speaker	1	2	3	4
Th = 0.7	21.6%	25.6%	21.6%	28.8%
Th = 1.7	22.4%	22.4%	19.2%	28.0%
Th = 2.7	23.2%	27.2%	18.4%	27.2%

Table4 sentence

Speaker	1	2	3	4
Th = -2.2	19.2%	29.6%	21.6%	27.2%
Th = -2.3	18.4%	29.6%	17.6%	26.4%
Th = -2.4	16.0%	31.2%	20.8%	23.2%

The best result is obtained in the case of rejection on semi-syllable when the threshold is 2.7. The differences between the unsupervised and supervised adaptation are relatively reduced by 27%, 35%, 120% and 58% for the four speakers. In the case of rejection on syllable and sentence, the performance can be improved with a speaker-dependent threshold while no common threshold can be found to decrease the error rate for all of the four speakers. Thus the best speech unit to be rejected is semi-syllable.

5. Conclusion

It has been proven that the word lattice based confidence measure can effectively improve the performance of unsupervised adaptation. The best speech unit of rejection is semi-syllable. And the proposed confidence measure is directly from the word lattice, thus reduces the computation and makes the system more practicable.

6. Acknowledgement

This project is supported by National Natural Science Funds (Item: 69975007), National 863 High Technology Projects (Contract No. 863-306ZD13-04-6 and 863-512-9805-10)

7. References

[1] C J Leggetter., Improved Acoustic Modelling for HMMs Using Linear Transformations, Dissertation for PhD, Cambridge, 1995

[2] M G Rahim, C H Lee, B H Juang, “Discriminative Utterance Verification for Connected Digits Recognition”, IEEE Trans. On Speech and Audio Processing, vol. 5, no. 3, pp. 266-277, 1997

[3] J M Boite, H Bourlard, B D’hoore and M Haesen, “A new approach towards keyword spotting”, Proceedings of EUROSPEECH, pp. 1273-1276, 1993

[4] L Gillick, Y Ito, J Young, “A probabilistic approach to confidence estimation and evaluation”, ICASSP1997, pp. 879-883

[5] M Weintraub, F Beaufays, Z Rivlin and et al, “Neural network based measures of confidence for word recognition”, ICASSP1997, pp.887-890

[6] J Liu, K J Hu, S X Pan, J T Jiang, Z Y Wang, “Study of Speech Recognition System based on Private Automatic Branch Exchage”, ACTA ELECTRONICA SINICA, vol. 27, no 1, 1999 (in Chinese)