Automatic Speech Database Verification Method Based on Confidence Measure

Jeomja Kang(ETRI), Hoyoung Jung(ETRI), Sanghun Kim(ETRI)

<Contents>

<Abstract>

Automatic Speech Database Verification Method Based on Confidence Measure

Jeomja Kang, Hoyoung Jung, Sanghun Kim

In this paper, we propose the automatic speech database verification method(or called automatic verification) based on confidence measure for a large speech database. This method verifies the consistency between given transcription and speech using the confidence measure. The automatic verification process consists of two stages : the word-level likelihood computation stage and multi-level likelihood ratio computation stage. In the word-level likelihood computation stage, we calculate the word-level likelihood using the viterbi decoding algorithm and make the segment information. In the multi-level likelihood ratio computation stage, we calculate the word-level and the phone-level likelihood ratio based on confidence measure with anti-phone model. By automatic verification, we have achieved about 61% error reduction. And also we can reduce the verification time from 1 month in manual to 1-2 days in automatic.

* Keywords: Automatic Verification, Speech Database, Confidence Measure, Likelihood Ratio

1. Introduction

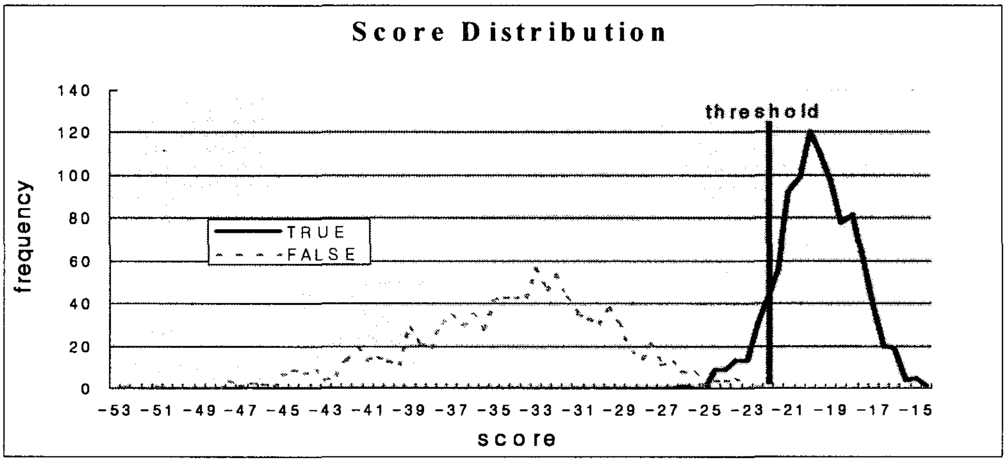
During recent years, the need of end users for speech interface in the various application products such as telephone network, internet, PDA and home appliance have rapidly increased. Therefore, speech recognition technology is now being recognized as the crucial technology for user interface of next generation. For this speech recognition application, a large amount of speech database with high quality is required. So, we constructed common speech database and have distributed this to companies and universities from 2001 up to now, and we are still constructing this. The construction of speech database and its verification is very difficult and also it needs a lot of time and high cost. The process of speech database verification consists of two stages : 1) the check of data format and size and 2) the consistency check of speech data and text transcription data. The second stage is very important, but this is the time consuming job for inconsistency check between speech data and transcription data. This stage is performed through human check(listen to each speech data and then check transcription) or using the speech recognizer. If speech data and transcription data are not consistent each other, we update the transcription data file according to speech data file. In the case of human check, correctness of speech database is very high but it needs a lot of time and high cost. But in the case of using the speech recognizer, the speech database verification is easily and quickly performed but the recognized results can include a lot of error data. Because the speech recognizer returns the recognition result after it chooses the word which has the maximum likelihood probability of lexicon, it may report as the correct data although there is an inconsistency between speech data and transcription data. In this paper, we propose the automatic speech database verification method to solve shortcomings of second stage of process of speech database verification and to get result having a high reliability. For this work, we assume that given transcription is correct and we compare the consistency between speech and transcription. We use the confidence measure to decide whether given transcription is correct or not. If confidence score is lower than predefined threshold, we classify that as error data. Otherwise, confidence score is higher than the threshold, we classify that as the true data. For data classified as the error data, we check the consistency manually between speech and transcription and then update the transcription file if speech and transcription are inconsistent.

The remainder of this paper is organized as follows. In section2, we will describe the overview of automatic speech database verification and four types anti-model. In section 3, we will describe our proposed automatic speech database verification method in detail. In section 4, we will report the experimental results. Finally, we conclude this paper.

2. Overview of Automatic Verification

2.1. Overview

The key idea of automatic speech database verification for efficient time and cost is to use acoustic model, anti-model and confidence measure technology of the utterance verification conceptually based on statistical hypothesis testing[1]-[3]. <Figure 1> shows the concept of our automatic verification using threshold. We regard as the true data if computed likelihood ratio is certainly higher than threshold. Otherwise, if computed likelihood ratio is lower than threshold, then that is regarded as the error data. In this approach, the classified true data absolutely must be less false acceptance rate which depends on quality of released speech database.

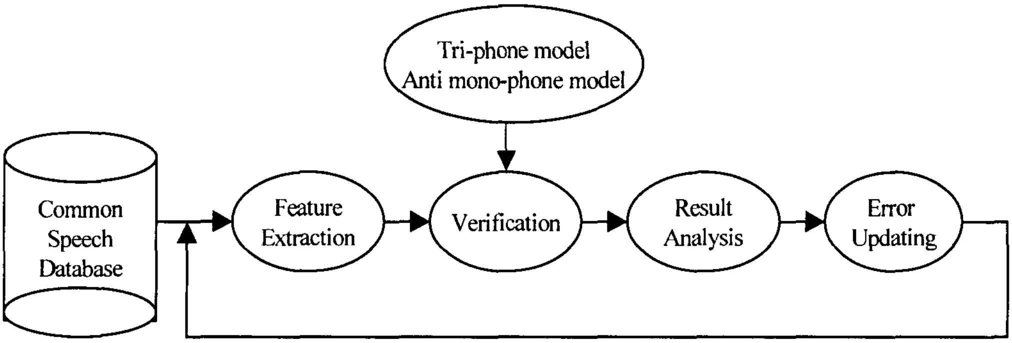


<Figure 1> The concept of our automatic verification

2.2. Procedure

<Figure 2> shows the procedure of automatic speech database verification. After

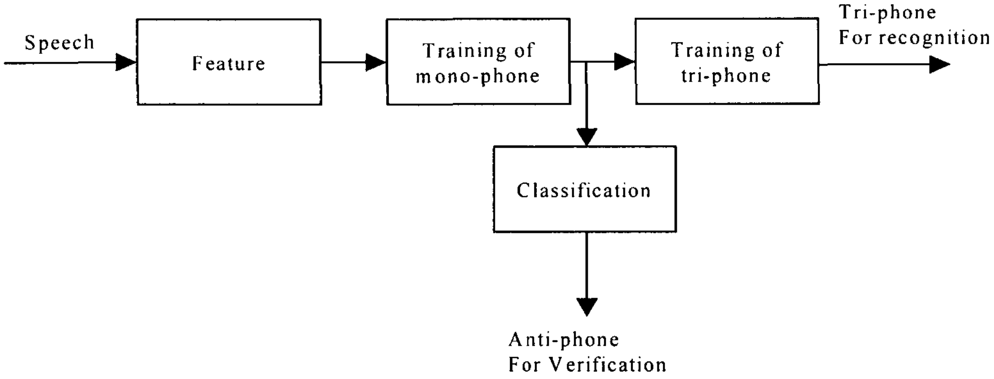
we prepare for the acoustic model and anti-model for verification, we can start automatic verification. This procedure consists of four stages. First, we extract the feature from common speech database. Second, we verify using the tri-phone acoustic model and anti-phone model and determine whether it is true data or not. Third, we analyze the result and finally we update the error data. This procedure is repeated continually.



<Figure 2> The procedure of automatic verification

2.3. Anti-model

The performance of automatic verification depends on how good the models and the anti models are. It is often used with a set of specific model trained for confidence measure. The anti-model can be applied to either the context dependent model or context independent model. Context independent model is used in general. But, there are use example of context dependent model[4]. In padma [4], the performance of context dependent model is better than the performance of context independent model. We use the context independent model because the number of context dependent model is generally more than number of context independent model. In <Figure 3>, the anti-phone model can be generated using the context independent model during train processing.



<Figure 3> Modeling procedure of anti-phone model

For anti-phone, we train the mono-phone model using 3 states, 3 codebooks and 3 gaussian mixture densities. We make four types anti-phone model according to the classification method generated from the trained mono-phone model and the vector quantization(VQ) based on clustering. The VQ based on the clustering trains with all the data that are not associated with the class. The anti-phone classification type through the trained mono-phone model is three types.

- We choose 1 Gaussian mixture which has the maximum weight among 3 Gaussian mixtures.

  : 37 mixture per one state

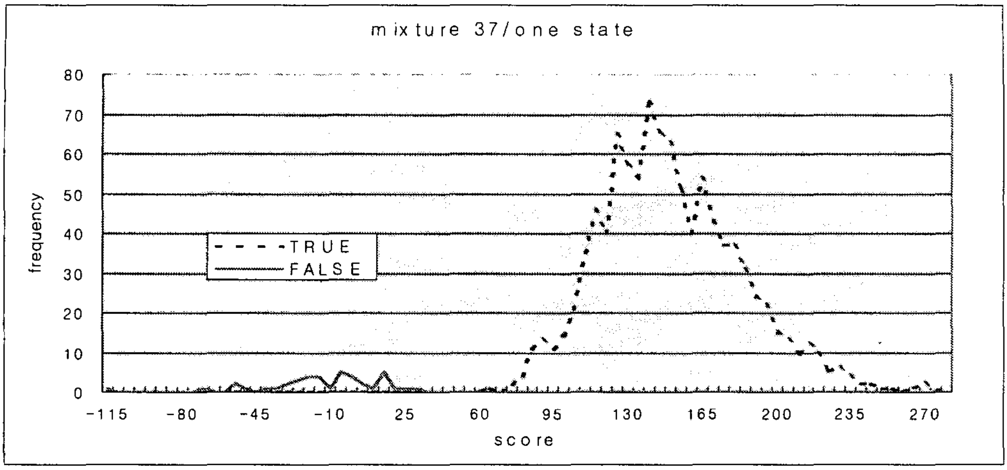
- We choose 2 Gaussian mixture which have the maximum weight value and the second maximum weight value among 3 Gaussian mixtures.

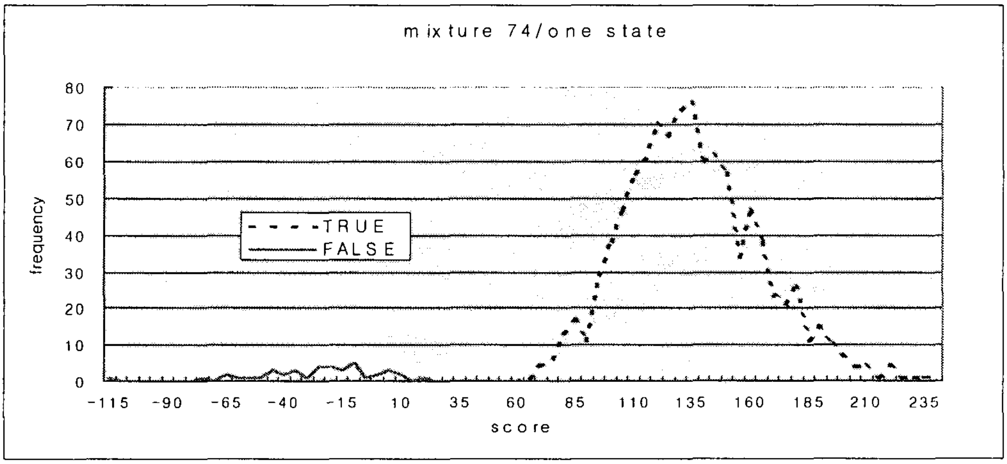
  : 74 mixture per one state

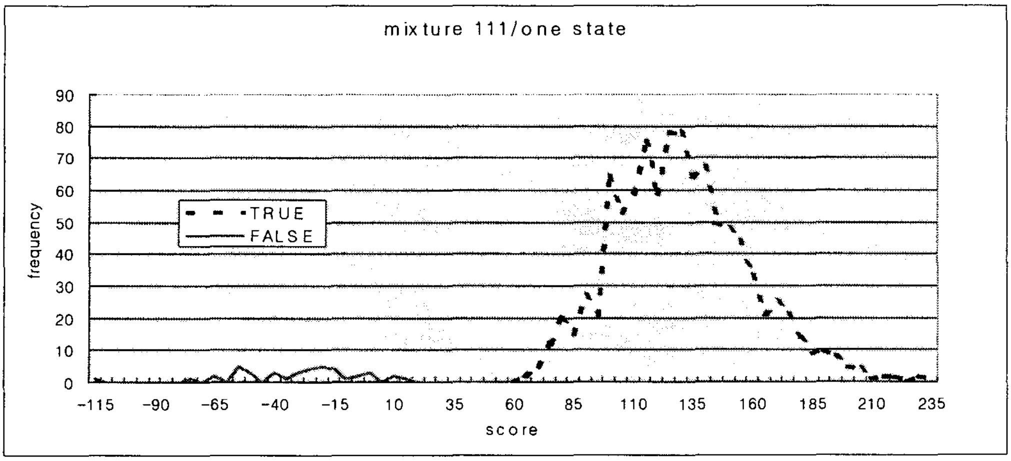
- We choose 3 Gaussian mixture among 3 Gaussian mixtures.

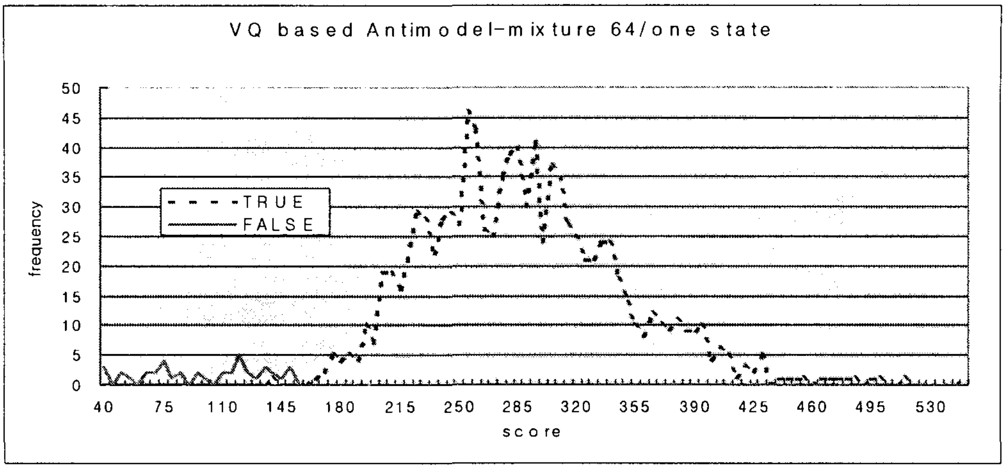
  : 111 mixture per one state

<Figure 4> shows the distance between the probability of distribution of true data and the probability of distribution of false data for each model. We can know that 111 mixture of one state has long distance. Therefore, we choose anti-model of 111 mixture for automatic verification.


mixture 37/one state
80
70
60
50
40
30
20
10
0
frequency
TRUE
FALSE
-115
-80
-45
-10
25
60
95
130
165
200
235
270
score


mixture 74/one state
80
70
60
50
40
30
20
10
0
frequency
TRUE
FALSE
-115
-90
-65
-40
-15
10
35
60
85
110
135
160
185
210
235
score


mixture 111/one state
90
80
70
60
50
40
30
20
10
0
frequency
TRUE
FALSE
-115
-90
-65
-40
-15
10
35
60
85
110
135
160
185
210
235
score

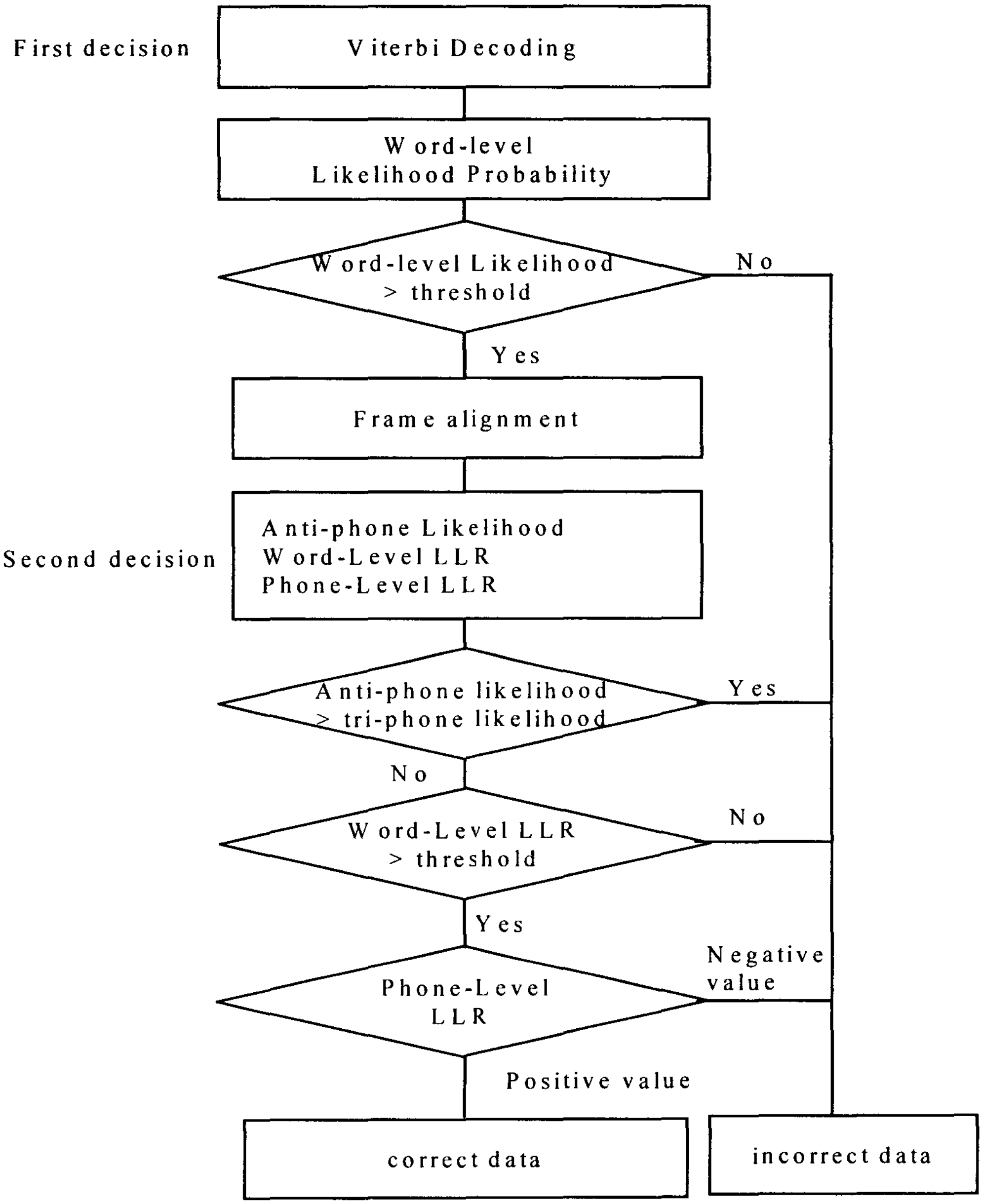


<Figure 4> The distance of each anti-model

3. Automatic Verification Method

3.1. Decision Algorithm

After making the anti-phone model and the tri-phone model, we perform the automatic verification of a large speech database. <Figure 5> shows the automatic verification process for decision which consists of two stages. In the first decision stage, we compute the word-level likelihood probability using viterbi decoding algorithm for estimating incorrect data. If the word-level likelihood probability is higher than threshold, the data are regarded as the correct data. Otherwise, if the word-level likelihood probability is lower than threshold, that is regarded as the incorrect data. We get the frame alignment information through this stage. In the second decision stage, we compute diverse probability for estimating such as the word-level anti-phone likelihood probability, the word-level likelihood ratio probability and the phone-level likelihood ratio probability. If the word-level anti-phone likelihood probability is higher than word-level tri-phone likelihood probability, the data are regarded as the incorrect data. Otherwise, if word-level anti-phone likelihood probability is less than word-level tri-phone likelihood probability, we compute the word-level likelihood ratio and phone-level likelihood ratio. If any one among phones which consists of word has a negative value which is less than a threshold, we regard the data as the incorrect data. Otherwise, the data are regarded as the correct data.



<Figure 5> The algorithm of automatic verification

3.2. multi-level confidence score

In general, the confidence measure can be regarded as a statistical hypothesis testing based on Likelihood Ratio Testing(LRT). In this paper, we use the confidence measure to decide whether input speech data can be accepted as the correct data or not. The confidence is measured by integrating the phone-level likelihood ratio into the word-level likelihood ratio When we compute the confidence based on the statistical hypothesis testing, null hypothesis H_0 represents the input speech containing the given subword and alternative hypothesis. H_1 represents the input speech not containing the given subword. The former is used as the tri-phone acoustic and the latter is used as the anti-phone acoustic model. According to Neyman-Pearson of speech observation vectors, $P(O|H_0)$ is the likelihood of the observation sequence given the subword hypothesis for Lemman, the optimal hypothesis test involves the evaluation of the likelihood ratio. Where O is the sequence the subword S_k, and $P(O|H_1)$ is the likelihood of the observation sequence given the non subword(anti-phone) hypothesis. The hypothesis test is performed by comparing the $LR(O;S_k)$ to a predefined critical threshold τ_k.

$$LR(O;S_k) = \frac{P(O|H_0)}{P(O|H_1)} > \tau_k \qquad (1)$$

We can get Likelihood Ratio(LR) by as follows. If $LR(O;S_k)$ is more than threshold τ_k, we accept the input data. Otherwise, if $LR(O;S_k)$ is less than threshold τ_k, we reject the input data.

First, we compute the phone-level likelihood ratio score, $PLLRi$ of phone i, as follows.

$$PLLR_i = \frac{g_i(O;\lambda_i) - G_i(O;\bar{\lambda}_i)}{T_i} \qquad (2)$$

λ_i : tri-phone model

$\bar{\lambda}_i$: anti-phone model

T_i : number of frame in phone i

$$g_i(O;\lambda_i) = \log[\,P(O \mid \lambda_i)] \qquad (3)$$

$$G_i(O;\bar{\lambda}_i) = \log[P(P \mid \bar{\lambda}_i)] \qquad (4)$$

When we compute the (3) and (4), observation probability of state j about i is as follows. In (5), c_{jk} is mixture weight and j is state of phone and k is number of mixture of state. N is gaussian distribution and u_{jk} is mean vector and U_{jk} is covariance value.

$$b_j(o) = \sum_{i=1}^{k}\{c_{jk}N(o,u_{jk},U_{jk})\} \qquad (5)$$

Therefore, we can get the word-level likelihood ratio as follows. The $WLLRw$ is sum of $PLLRi$ and is normalized the number of phone, N.

$$WLLR_w = \frac{1}{N}\sum_{i=1}^{n} PLLR_i \qquad (6)$$

Computed $PLLRi$ in (2) and $WLLRw$ in (6) is used as algorithm of automatic verification for decision whether input speech data can be accepted as the correct data or not.

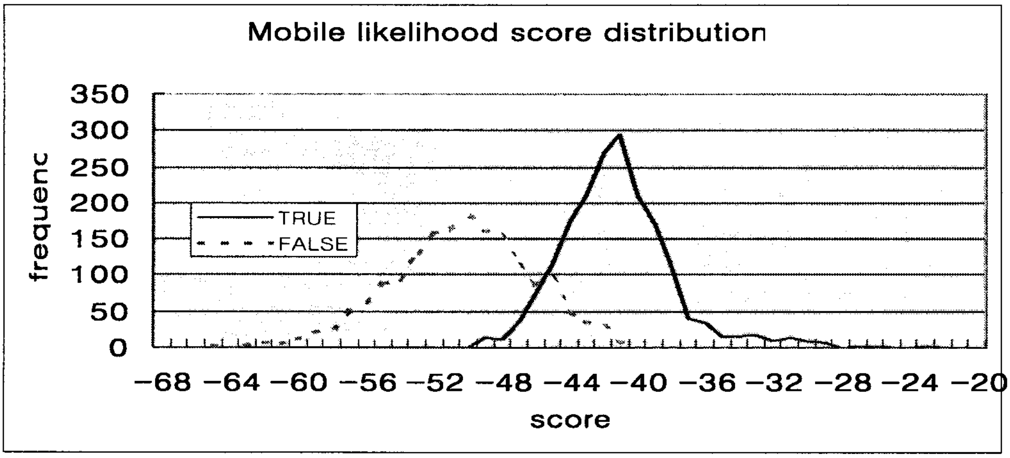
4. Experimental Results

Experiments were carried out on Korean word consisting of wired telephone(117,653 utterance) and wireless telephone(242,489 utterance) which recorded over telephone network in a wide variety of acoustic conditions. The acoustic features basically were 13 dimension Mel-scale Frequency Cepstrum Coefficients(c0c12) and then took the Cepstrum Mean Substracion(CMS). And also, we used 13 delta dimension delta and 13 delta delta dimension. The acoustic model trains the mono-phone which has 1 Gaussian mixture and then it trains the tri-phone using the segment information of mono-phone. The tri-phone trains with 5 mixture Gaussian densities. Automatic verification tool based on windows is very easy for using. <Table 1> shows the experimental result using automatic verification method.

<Table 1> Experimental result of automatic verification

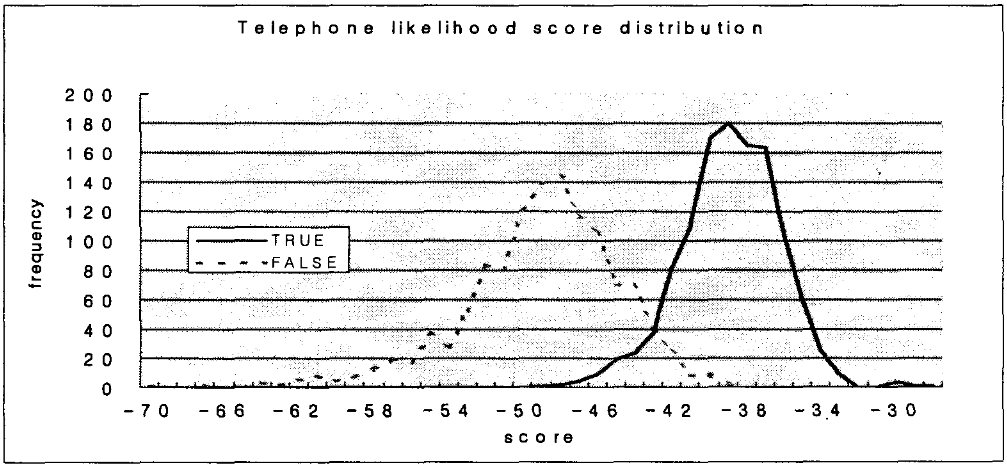
Type	Before Verification	After Verification		Error
	DB Error Rate(%)	Rejection Rate(%)	False Alarm(%)	Reduction(%)
Telephone	0.78	1.33	0.38	40
Mobile	1.25	9.74	0.43	82
Average	**1.02**	**5.54**	**0.41**	**61**

As shown in the <Table 1>, we have achieved the 61% error reduction. The original DB error rate is 1.02% and final DB error rate is 0.41% after automatic verification. The rejection rate means incorrect data which it will be verified manually by people. The difference of rejection rate between telephone and mobile phone depends on the predefined threshold. This means that telephone performance is better than mobile performance. <Figure 6> shows the likelihood score distribution of mobile speech database between the correct data and the incorrect data. In this case, overlap data rate is about 71%.



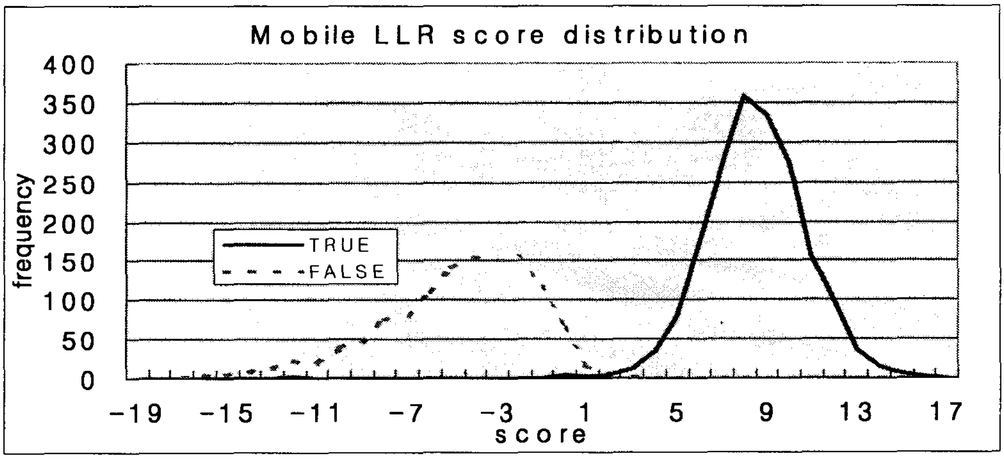
<Figure 6> Mobile word likelihood score distribution

<Figure 7> shows the likelihood score distribution of telephone speech database between the correct data and the incorrect data. In this case, overlap data rate is about 60%.

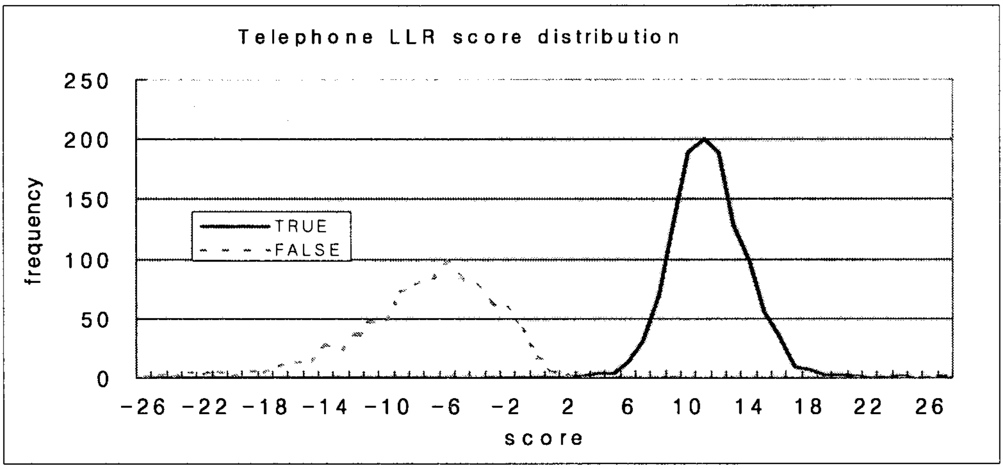


<Figure 7> Telephone word likelihood score distribution

<Figure 8> and <Figure 9> show the word-level likelihood distribution. As shown in Figure, we know that overlap data area is reduced significantly. In <Figure 8>, overlap data rate is about 16.7%. And also we can reduce the verification time from 1 month in manual to 1-2 days in automatic.



<Figure 8> Mobile word likelihood ratio score distribution



<Figure 9> Telephone word likelihood ratio score distribution

5. Conclusions

In this paper, we described the automatic verification method based on the confidence measure for speech database. The automatic verification process consists of the word level likelihood computation stage and the multi-level likelihood ratio computation stage. The experiments have shown that the word level likelihood computation stage has many overlap data but the multi-level likelihood ratio computation stage can significantly reduce the overlap data. And also, the performance of proposed automatic verification method has 61% error reduction and can reduce time and cost required for speech database verification. In the future, we need to extend current algorithm and new algorithm for Korean related to digit and sentence level verification.

Reference

[1] Myoung-Wan Koo, Chin-Hui Lee, Biing-Hwang Juang, "Speech Recognition and Utterance Verification Based on a Generalized Confidence Score", IEEE Trans., *Speech Audio Processing*, Vol. 9, No. 8, pp. 821-832, Nov. 2001.

[2] S. Kamppari, "Word and Phone Level Acoustic Confidence Scoring for Speech Understanding Systems", *Master's thesis*, MIT, 1999.

[3] Eduardo Lleida, Richard C. Rose, "Utterance Verification in Continuous Speech Recognition: Decoding and Training Procedures", IEEE Trans., *Speech Audio Processing*, Vol. 8, No. 2, pp. 126-139, Mar. 2000.

[4] Padma Ramesh, Chin-Hui Lee, Biing-Hwang Juang, "Context Dependent Anti Subword Modeling for Utterance Verification", *ICSLP*, 1998.

접수일자: 2004년 8월 3일
게재결정: 2004년 9월 9일

▶ 강점자(Jeomja Kang)
주소: 305-350 대전광역시 유성구 가정동 161번지
소속: 한국전자통신연구원(ETRI) 음성/언어정보연구부
전화: 042) 860-4880
E-mail: jjkang@etri.re.kr

▶ 정호영(Hoyoung Jung)
주소: 305-350 대전광역시 유성구 가정동 161번지
소속: 한국전자통신연구원(ETRI) 음성/언어정보연구부
전화: 042) 860-1328
E-mail: hjung@etri.re.kr

▶ 김상훈(Sanghun Kim)
주소: 305-350 대전광역시 유성구 가정동 161번지
소속: 한국전자통신연구원(ETRI) 음성/언어정보연구부
전화: 042) 860-5141
E-mail: ksh@etri.re.kr