# On Effective Speaker Verification Based on Subword Model*

**Sungjoo Ahn** · Sunmee Kang*** · Hanseok Ko****

## ABSTRACT

This paper concerns an effective text-dependent speaker verification method to increase the performance of speaker verification. While various speaker verification methods have already been developed, their effectiveness has not yet been formally proven in terms of achieving acceptable performance levels. This paper proposes a weighted likelihood procedure along with a confidence measure based on subword-based text-dependent speaker verification. Our aim is to remedy the low performance problem in speaker verification by exploring a means to strengthen the verification likelihood via subword-based hypothesis criteria and weighted likelihood method. Experimental results show that the proposed speaker verification method outperforms that of the speaker verification scheme without using the proposed decision by a factor of up to 1.6 times. From these results, the proposed speaker verification method is shown to be very effective and to achieve a reliable performance.

**Keywords: Speaker Verification, Confidence Measure, Subword Model**

## 1. Introduction

Recent advances in the internet have enabled many network related online services to users, providing both convenience and easy accessibility to a variety of services. These services, including electronic commerce, banking applications and information reservation, require highly reliable network securities. Reliable speaker verification provides an added security measure to online services and is rapidly positioning itself as the key to success in the online service industry. As the first step to enable the speaker verification task in these online service systems, a new subscriber is required to register by enrolling his/her voice imprints and creating a user's personal speech model [1]. Speaker verification methods can be categorized into either text-dependent or text-independent.

* This work was supported by grant No. R01-1999-00229 from the Korea Science & Engineering Foundation.

** Dept. of Electronics and Computer Engineering, Korea University
*** Dept. of Computer Sciences, Seokyeong University
**** Dept. of Electronics and Computer Engineering, Korea University
Dept. of Electrical and Computer Engineering, Johns Hopkins University

The former requires the speaker to provide utterances of the password having the same text for both training and verification trails, whereas the latter does not rely on the specific text being spoken.

Most utterance verification systems use subword-based methods to improve performance. In these approaches, various discrimination methods such as neural net and support vector machines are used along with improved confidence measures [2][3][4]. While the subword-based approaches are also used in speaker verification, the challenge is much more demanding than that of utterance verification. Also, in the most recent work in speaker verification, the likelihood of the speaker is calculated with equal weight for each subword model or state [5]. However, the significance of each subword model can vary in the view of input utterances. Thus, it is better to assign appropriate weighting on the likelihood of each subword model than to naively using the likelihood obtained with general Viterbi decoding. As a result, the performance of the on-going speaker verification systems has yet to show any drastic improvement.

To cope with these problems, we propose a two-step procedure, which consists of a weighted likelihood part and a subword based hypothesis decision part. The weighted likelihood step assigns variable weightings to the likelihood of subword model for each speaker. The subword-based hypothesis decision step transforms the likelihood of the subword model for establishing the decision rule that reflects the confidence measure. In this paper, we develop a subword-based speaker verification system by establishing an appropriate confidence measure for each subword model. For decision, the likelihood of each subword model is weighted to the segments assigned to the subword model by the value obtained with the proposed method. We then transform the value to an appropriate confidence measure and use it as a verification score.

This paper is organized as follows. In Section 2, we describe the proposed speaker verification methods. We, then, conduct the representative experiments and discuss the results on the performance of the proposed methods in Section 3. Finally, in Section 4, conclusive remarks are presented.

## 2. Subword Based Speaker Verification

### 2.1 Overview of Speaker Verification

The speaker verification problem can be considered as a form of hypothesis test. That is, to verify the claimed speaker, the likelihood ratio test is applied to an input utterance. If the input utterance is $O = \{o_1, o_2, \cdots, o_T\}$ and the speaker claim to the model $\lambda_c$, the probability of likelihood $P(\lambda_c | O)$ is calculated by using the Viterbi decoding algorithm [1][6].

$$likelihood\ \ ratio = \frac{\Pr(O \text{ is from the claimed speaker})}{\Pr(O \text{ is not from claimed speaker})} = \frac{\Pr(\lambda_c | O)}{\Pr(\lambda_r | O)} \quad (1)$$

where $\lambda_r$ is the anti-speaker model of the claimed speaker.

Based on this likelihood ratio and threshold, we verify the claimed speaker.

2.2 Subword based Speaker Verification

Each piece of the password information is represented by a word, $W$, which in turn is equivalently characterized by a concatenation of a sequence of subwords, $\{s_n\}_{n=1}^{N}$, where $s_n$ is the $n$th subword and $N$ is the total number of subwords in the input utterance.

We then apply the subword models, $\lambda_1, \ldots, \lambda_N$, in the same order of the subword sequence $S$ to decode the input utterance. This process is known as "forced decoding" or "forced alignment", in which the Viterbi algorithm is employed to determine the maximum likelihood segmentations of the subwords. That is, if the input utterance is $O = \{o_1, o_2, \ldots, o_T\}$ and the number of subwords is $N$, the input utterance is segmented into $N$ parts of subword models [3].

$$X = \{O_1, O_2, \ldots, O_N\} = \{O_1^{t_1}, O_{t_1+1}^{t_2}, \ldots, O_{t_{N-1}+1}^{t_N}\} \quad (2)$$

where $O_N$ is the segmented sequence of observations corresponding to subword $s_n$.

Given a decoded subword $s_n$ in an observed speech segment $O_N$ we need a decision rule by which we assign the subword to either hypotheses $H_0$ or $H_1$. Following the definition, $H_1$ means that observed speech $O_N$ consists of the actual sound of subword $s_n$, and $H_0$, is the null hypothesis. For the binary-testing problem, one of the most useful tests for decision is the Neyman-Pearson lemma. For a given number of observations $K$, the most powerful test, which minimizes the error for one class while maintaining the error for the other class constant, is a likelihood ratio test.

$$r(O_n) = \frac{P(O_n | H_0)}{P(O_n | H_1)} = \frac{P(O_n | \lambda_n)}{P(O_n | \bar{\lambda}_n)} \quad (3)$$

This equation is the likelihood ratio of the subword $s_n$ where $\lambda_n$ is the speaker's subword model $s_n$, and $\bar{\lambda}_n$ is the anti-speaker's subword model $s_n$ respectively.

For normalization, an average frame log-likelihood ratio (LLR), $R_n$, is defined as

$$R_n = \frac{1}{l_n}[\log P(O_n \mid \lambda_n) - \log P(O_n \mid \bar{\lambda}_n)] \tag{4}$$

where $l_n$ is the number of frame for subword $n$.

2.3 Confidence Measure

For an effective speaker verification, we need to define a function to combine the results of subword tests. A confidence measure $M$ for an input utterance $O$ can be represented as

$$M(O) = F(R_1, R_2, \ldots, R_N) \tag{5}$$

where $F$ is the function to combine the LLR's of all subwords in the input utterance[3]. That is, a subword unit based confidence measure($M$) can be computed from the likelihood ratio between the correct and alternative hypothesis models. Several confidence measures have been proposed for utterance verification but hardly any for speaker verification. We denote two of them as $M_1$ and $M_2$ as follows:

$$M_1 = \frac{1}{L}\sum_{n=1}^{N} l_n R_n \tag{6}$$

where $N$ is the total number of subwords in the utterance, and $L$ is the total number of frames of the utterance, $L = \sum_{n=1}^{N} l_n$. Furthermore,

$$M_2 = \frac{1}{N}\sum_{n=1}^{N} R_n \tag{7}$$

Here $M_1$ is an average score over all frames and all subwords. Each of the subword score $R_n$ is weighted by its duration. $M_2$ is an average LLR of all subwords and independent of individual duration.

For effective speaker verification, we define a different confidence measure, $M_3$. The above two confidence measures have a large dynamic range, which is undesirable. A preferable statistic should have a stable and limited numerical range so that a common threshold can be established for all subwords for simplicity.

One way to limit the dynamic range of the subword confidence measure is to use a sigmoid function of the form.

$$u(n) = \frac{1}{1 + \exp(-\alpha * (R(n) - \tau))} \tag{8}$$

where $\alpha$ and $\tau$ are constants which control the slope of the function and the shift of the smoothing function, respectively. Figure 1 shows the relationship of the sigmoid function $u(n)$ with various $\alpha$ values. The sigmoid function is generally used in discriminative training method such as the minimum classification error (MCE) and generalized probabilistic descent (GPD) training to approximate the misclassification error count.

Using this sigmoid function, we define yet a different confidence measure $M_3$ as follows,

$$M_3 = \frac{1}{N} \sum_{n=1}^{N} u(n) \tag{9}$$

Thus the range of the subword confidence measure is compressed to fall within the interval [0,1].

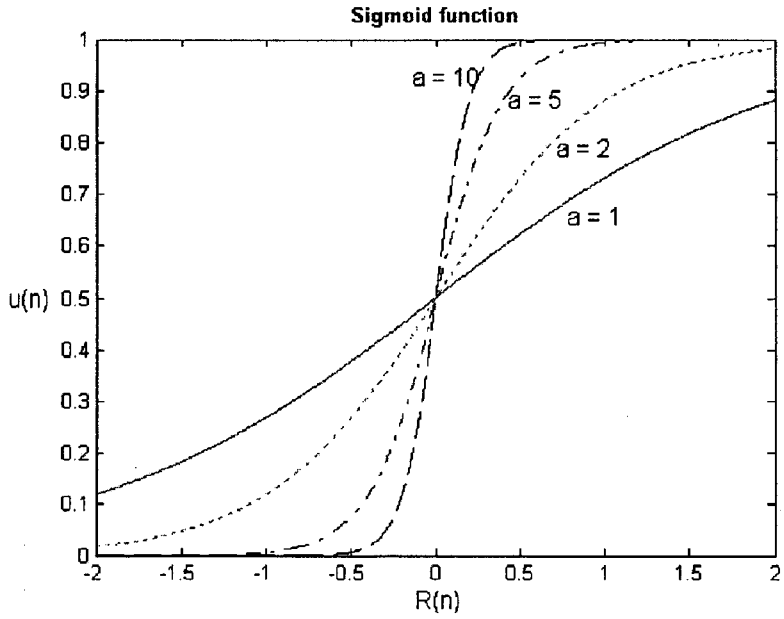


Figure 1. Sigmoid function u(n)

2.4 Weighting the Subword Model

We propose an effective weighted likelihood method for speaker verification by assigning appropriate weighting on the likelihood of each subword unit.

In the most recent work of subword based speaker verification, the likelihood of the speaker is calculated with equal weight for each subword unit. However, the significance of each subword unit is different in the view of input utterances. Thus, it is better to assign appropriate weighting on the likelihood of each subword unit than blindly using the likelihood obtained with general Viterbi decoding. We suggest a procedure to quantify the contribution of each subword segmentation, in the decision scheme, according to the following procedure.

The distance $D$ between average LLR scores on the client's training data and the LLR scores on a development set of impostor data for a given subword unit $n$ is defined as follows.

$$D_n = \begin{cases} (\mu_C^n - \mu_I^n)/\sigma_I^n & \text{if } \mu_C^n > \mu_I^n \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

where $\mu_C^n$ and $\mu_I^n$ are the mean LLR scores of a subword model $n(R_n)$ on the client's training utterances and on the impostor utterances, respectively, and $\sigma_I^n$ is the standard deviation of the LLR scores($R_n$) on the impostor utterances. Thus, the weight of each subword unit is obtained using the following equation.

$$w(n) = \frac{D_n^\gamma}{\sum_{j=1}^{N} D_n^\gamma}, \quad n = 1,\ldots, N \tag{11}$$

where $w(n)$ is a weight of a subword unit $n$ and $N$ is the total number of subwords in the input utterance. Each element $w(n)$ shows the accuracy of a given subword model, producing an observation vector $O_n$. Clearly, if $w(n)$ is small for other values of $n$, it does not contribute less to the decision process. It is reasonable to conclude that this particular weight contributes to the verification process.

In decision, the LLR contribution of each subword model $n$ is weighted by $w(n)$ to the segments assigned to the subword model $n$. Also, in Equation (11), the exponent $\gamma$ controls the balance of the weighting scheme: $\gamma$ =0 gives equal weights to all subword models, if $\gamma$ increases the more discriminative models get a larger weight. However, too high $\gamma$ value is given to a small part of the test utterance, and as a result, only a few subwords determine the final LLR on the verification. So the preference should go to more stable quality measures on each subword level. The following two confidence measures are obtained reflecting the desirable qualities.

$$M_4 = \frac{1}{N}\sum_{n=1}^{N} w(n)R_n \tag{12}$$

$$M_5 = \frac{1}{N}\sum_{n=1}^{N} w(n)u(n) \tag{13}$$

## 3. Experiments

In this section, we perform various representative experiments to make the performance comparisons of those proposed candidate methods as discussed in Section 2.

### 3.1 Experimental Condition

To show the effectiveness of proposed speaker verification methods, text-dependent speaker verification experiments were performed using the methods discussed in Section 2.

The speech database used in the experiments contains isolated Korean words uttered by 35 speakers (17 male and 18 female). Each of these speakers uttered 15 times his/her password and uttered 3 times other speakers' passwords. The speech samples were recorded using microphones in an office environment and sampled at 8 kHz. In the experiments, each speech signal was parameterized using 12 MFCC (Mel-Frequency Cepstral Coefficient) plus log-energy, and their first and second derivatives. The analysis window size was 25 ms with 10 ms overlap.

Acoustic phoneme models for the speaker produced 2-state left-right continuous density HMMs with two Gaussian mixtures per state. 9 utterances of each client were used to test the false rejection and 2 utterances of imposters were used to test false acceptance. We used the number of training to 6 and the number of development set for imposter data to 34 for each speaker in this experiment.

The performance of speaker verification is evaluated with the Equal Error Rate (EER), meaning that the False Rejection Rate is equal to the False Acceptance Rate. While EER measures the error rate committed by the verification system the score can be interpreted as how accurately the model is constructed. High EER reflects an inadequately constructed model while low EER indicates more faithful model construction.

### 3.2 Experimental Results

We experimented with various confidence measures using the proposed method. As the baseline experiments, a general speaker verification experiment was conducted first to compare the performance of the proposed methods. In the experiment, the subword

phoneme models for each speaker produced 2-state left-right continuous density HMMs with two Gaussian mixtures per state. The results obtained are shown in Table 1. As can be seen, Table 1 shows the performance in EER over the two confidence measures ($M_1$, $M_2$). From this result, it is evident that the method using the measure $M_1$ is better than that of $M_2$. This is because of the independent assumption of the individual subword model durations.

Table 1. Comparison of EER with baseline confidence measure

| method | EER(%) |
|---|---|
| $M_1$ | 2.15 |
| $M_2$ | 3.1 |

In the following experiment, we applied the sigmoid function based confidence measure ($M_3$) to the speaker verification. Table 2 shows a comparison of EER with two parameter values $(\alpha, \tau)$. As shown in Table 2, the best performance was obtained when the tau($\tau$) is 0 and alpha($\alpha$) is 10. However, the value of $\tau$ is more important than that of $\alpha$. That is, the performance dramatically varies with the change of the value $\tau$ but not so with $\alpha$. Comparing this result with Table 1, it shows that the likelihood transformation method based on sigmoid function ($M_3$) attains the performance improvement.

Table 2. Comparison of EER with various parameter values

(confidence measure = $M_3$)

| tau($\tau$) \ alpha($\alpha$) | 1 | 2 | 5 | 10 |
|---|---|---|---|---|
| -2 | 1.87 | 2.1 | 2.22 | 2.3 |
| 0 | **1.67** | **1.71** | **1.66** | **1.64** |
| 2 | 1.73 | 1.83 | 1.73 | 1.89 |
| 4 | 1.81 | 1.87 | 2.11 | 2.1 |
| 6 | 2.72 | 3.3 | 3 | 3.3 |

In the following experiment, we applied the weighted likelihood method as described above. Table 3 shows the comparison of EER against $\gamma$ in calculating the weigh value. As shown, the performance is significantly better than the results of Table 1 but lower than those of Table 2. This is because of the LLR's large dynamic range of each subword model. However, the best performance is obtained when the value of $\gamma$ is 2. This value is used in next experiment.

Table 3. Comparison of EER against the value of gamma($\gamma$) in the weighting value
(confidence meaure = $M_4$)

| gamma($\gamma$) | 0.3 | 0.5 | 0.7 | 1 | 2 | 4 | 8 |
|---|---|---|---|---|---|---|---|
| EER(%) | 2.15 | 2.15 | 2.12 | 2.10 | **2.03** | 2.04 | 3.50 |

The final the proposed experiment was performed with a confidence measure based on the subword model using the sigmoid function and weighted likelihood method. The results obtained are shown in Table 4. From the results, it was shown that when the best performance is obtained when the value of $\tau$ is 0 and the value of $\alpha$ is 10 as like in Table 2. Although Table 4 does not show the results of other cases of $\gamma$, the similar results were obtained as in Table 4. In comparison with previous results, the proposed method outperforms the other methods. Also, the value of $\tau$ is shown to be very important. This is because the performance dropped sharply when the value of $\tau$ was varied. Thus from these results, selecting the optimal parameter value was shown to be an important consideration.

Table 4. Comparison of EER with various parameter values when $\gamma$=2
(confidence measure = M5)

| tau($\tau$) \ alpha($\alpha$) | 1 | 2 | 5 | 10 |
|---|---|---|---|---|
| -3 | 3.3 | 6 | 7.8 | 9 |
| -2 | 3 | 3.76 | 5.2 | 5.5 |
| -1 | 1.74 | 1.93 | 2.8 | 3.55 |
| 0 | **1.48** | **1.41** | **1.32** | **1.31** |
| 1 | 2 | 2.2 | 3.6 | 4.1 |

The above experimental results show that the proposed speaker verification method using the weighted likelihood procedure along with a confidence measure using sigmoid transform outperforms that of the speaker verification scheme without using the proposed method by a factor of up to 1.6 times. That is, from the experimental results, we demonstrated that the proposed speaker verification methods indeed significantly improve speaker verification performance.

## 4. Conclusions

In this paper, we proposed a weighted likelihood procedure along with a confidence

measure based decision scheme focused on subword-based text-dependent speaker verification. Representative simulations showed that the proposed speaker verification method indeed significantly improves speaker verification performance. By applying the proposed methods, the speaker verification method using the weighted likelihood method and the confidence measure based decision criteria outperforms that of the speaker verification scheme not using the proposed method by a factor of up to 1.6 times. From these results, we showed that the proposed text-dependent speaker verification method is very effective and achieves a reliable performance. Future study should continue to further increase the verification performance and to find the optimal parameters along with optimal confidence measures.

## References

[1] Lee, C. H., F. K. Soong & K. K. Pliwal. 1997. "Automatic Speech and Speaker Recognition-Advanced Topics." *Kluwer Academic Publishers.* Second Printing, 31-56.

[2] Rahim, M. G., C. H. Lee & B. H. Juang. 1997. "Discriminative Utterance Verification for Connected Digits Recognition." *IEEE Trans. On Speech & Audio Processing*, Vol. 5, No. 3, 266-277.

[3] Li, Q., B. H. Juang, Q. Zhou & C. H. Lee. 2000. "Automatic Verbal Information Verification for User Authentication." *IEEE Trans. On Speech & Audio Processing*, Vol. 8, No. 5, 585-596.

[4] Charlet, D., G. Mercier & D. Jouvet. 2001. "On Combining Confidence Measure for Improved Rejection of Incorrect Data." *Eurospeech 2001*, 2113-2116.

[5] Lleida, E. & R. C. Rose. 2000. "Utterance Verification in Continuous Speech Recognition: Decoding and Training Procedures." *IEEE Trans. On Speech & Audio Processing*, Vol. 8, No. 2, 126-139.

[6] Furui, S. 1994. "An Overview of Speaker Recognition Technology." *ESCA Workshop on Automatic Speaker Recognition, Identification and Verification*, 1-9.

Received: Jan. 27, 2002.
Accepted: Mar. 5, 2002.

▲ Sungjoo Ahn
Dept. of Electronics and Computer Engineering, Korea University
5Ka-1, Anam-dong, Sungbuk-ku, Seoul, 136-701, Korea.
Tel: +82-2-927-6115 (O) Fax: +82-2-3291-2450
H/P: 011-9099-0258
E-mail: sjahn@ispl.korea.ac.kr

▲ Sunmee Kang
Dept. of Computer Science, Seokyeong University
16Ka-1, Chongnung-Dong, Sungbuk-ku, Seoul, 136-704, Korea
Tel: +82-2-940-7291 (O) Fax: +82-2-919-0345
H/P: 011-9760-7144
E-mail: smkang@skuniv.ac.kr

▲ Hanseok Ko
Dept. of Electronics and Computer Engineering, Korea University
5Ka-1, Anam-dong, Sungbuk-ku, Seoul, 136-701, Korea.
Tel: +82-2-3290-3239 (O) Fax: +82-2-3291-2450
H/P: 011-9001-3239
E-mail: hsko@korea.ac.kr

Currently on Sabbatical Leave at:
Dept. of Electrical and Computer Engineering, Johns Hopkins University
Baltimore, MD, USA.
Tel: 410-516-7199 (O) 410-922-7907 (H)
E-mail: hsko@clsp.jhu.edu