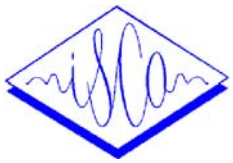

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

AN UTTERANCE VERIFICATION SYSTEM BASED ON SUBWORD MODELING FOR A VOCABULARY INDEPENDENT SPEECH RECOGNITION SYSTEM

Myoung–Wan Koo [1] *and Sun–Jeong Lee*[2]
[1]Multimedia Technology Research Laboratory
Korea Telecom
[2]Junior College of Inchon
Tel. +82–2–526–5090, Fax : +82–2–526–5909, E-mail:mwkoo@smm.kotel.co.kr

10.21437/Eurospeech.1999-75

ABSTRACT

This paper describes a Korean utterance verification system based on subword modeling for a vocabulary independent speech recognition system. We deploy strategy consisting of two modules: recognition and verification, for utterance verification. In the stage of recognition, multiple hypotheses with hypothesized word boundaries obtained through Viterbi segmentation of the utterance are obtained. And likelihood ratio is used as a post-processor for rejecting unlikely hypothesis in the stage of verification. Our study is focused on the verification module. First, we make a comparative study on averaging methods for obtaining the confidence measure for words from the log likelihood ratio based on phone. Three kinds of average techniques were investigated as arithmetic, geometric, and harmonic averages. Second, we study the effect of cohort set, which is the most competitive units to subword units. One cohort set model is trained for each subword. We found out the size of the cohort set for best recognition result. Finally, we present how to model anti-models for each context-dependent units. Three kinds of approaches are studied. The first one is to use cohort set based on context-independent unit to simplify the calculation. The second one is use cohort set based on context-dependent unit, which is obtained by phone recognizer based on context-dependent units. The final one is to use cohort set based on hybrid units. We make a comparative study on each approach.

1. INTRODUCTION

In many telephone services based on automatic speech recognition technology, users unfamiliar with the technology are likely to respond in ways that are less constrained than saying the desired keyword or utterance in isolation. To deal with such responses, the recognizer should be able to accept valid utterance and reject non-valid ones.

Several methods for rejecting non-keyword utterance have been proposed [1] [2]. Sukkar et. al proposed a two pass classifier to reject utterance. And Lleida et. al proposed a likelihood ratio decoder as a one pass classifier [3]. Recently, a hybrid decoder based on generalized confidence score was proposed by Koo et. al [4] [5].

In this paper, we describes a Korean utterance verification system based on subword modeling for a vocabulary independent speech recognition system. Section 2 presents the overview of an utterance verification system. Our study focuses on the verification module. The experimental environment and results are explained in section 3. Speech database and three different of approaches are also discussed here. Finally, we make an conclusion in section 4.

2. UTTERANCE VERIFICATION SYSTEM

2.1. Baseline System

The baseline system which we choose as an utterance verification system is based on two pass classifier consisting of two modules: recognition and verification. In the stage of recognition, likelihood decoder called as Viterbi decoder produces a sequences of hypothesized word labels and hypothesized word boundaries. Hypothesized word boundaries are obtained through Viterbi segmentation of the utterance. Utterance verification is then applied in the stage of verification, applying a hypothesis test to word hypotheses produced by the likelihood decoder. The overview of the base line system is shown in Figure 1. The hypothesis test is based on the computation of the confidence measure. The confidence measures for hypothesized words are computed using the word boundaries and are compared to decision thresholds to accept or reject the hypothesized words. And likelihood ratio between the null and alternate hypothesis for each word hypothesis is used as confidence measures.

2.2. Confidence Measure

The confidence score is based on the log likelihood ratio of correct phone model and corresponding anti-model. To normalize the confidence score, we used sigmoid function which was widely used to resolve dynamic range problem.

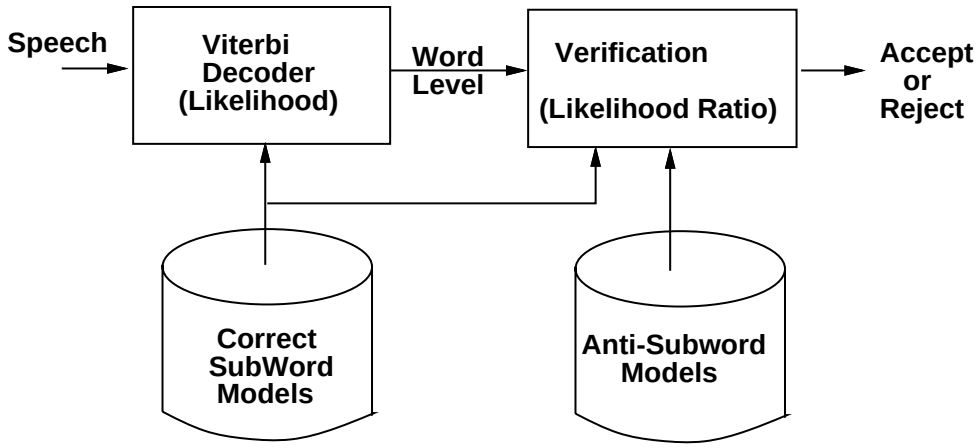

Figure 1: The Overview of an Utterance System

The confidence measure is based on Neyman-Pearson hypothesis testing formulation [6]. Given a sequence of phrase W corresponding to phonetic baseform $U = u_1, ..., u_N$, Verifier may generate the phone-level log likelihood ratio of phone n, LLR_n. The phone-level log likelihood ratio of phone n, LLR_n, is defined as

$$LLR_n = 1/\tau \sum_{t-\tau<l\leq t} LLR_l, \tag{1}$$

where τ is the frame duration for phone n and LLR_l is the log likelihood ratio(LLR) at frame l. The log likelihood ratio for the observed vector, o_l at frame l, is defined as

$$LLR_l = \log \frac{P(o_l/\lambda^c)}{P(o_l/\lambda^a)} = \log \frac{a^c_{ij} b^c_j(o_l)}{a^a_{ij} b^a_j(o_l)}, \tag{2}$$

where λ^c, λ^a are respectively HMM correct models for the unit, anti-models for the corresponding unit, and a^c_{ij}, b^c_j are probabilities for the HMM correct models for the unit, and a^a_{ij}, b^a_j are probabilities for the anti-models for the corresponding unit. The sigmoid limiter is used as a pre-processor of ratio knowledge sources because LLR has the value ranging from $-\infty$ to ∞. The scaled confidence measure at each phone level is defined as

$$CM_p(LLR) = \log \frac{1}{1 + \exp(-\alpha \cdot LLR)} \tag{3}$$

where α are a weighting parameter.

Confidence measure of each word W is obtained from a function of scaled confidence measure for each phone as

$$CMw = f(CMp_1, \ldots, CMp_n) \tag{4}$$

where word consists of n sequences of phones. There are many different ways that confidence measure of each word level can be formed by combining phone-level confidence measures[2][7]. Here, we consider three different averaging methods:geometric mean, arithmetic mean and harmonic mean. The word level confidence measures based on geometric mean, arithmetic mean and harmonic mean from phone level confidence measures are respectively

$$CMw_g = exp(\frac{1}{N} \sum_n \log CMp_n) \tag{5}$$

$$CMw_a = \frac{1}{N} \sum_n CMp_n \tag{6}$$

$$CMw_h = \frac{N}{\sum_n \frac{1}{CMp_n}} \tag{7}$$

where N is the total number of phone sequence for word. For every confidence measures, specific threshold η is set up. If the confidence measure is below the threshold, the keyword is discarded.

$$W = \begin{cases} Accept & \text{if } CMw > \eta \\ Reject & \text{otherwise} \end{cases} \tag{8}$$

2.3. Anti-models

Anti-models are made by using correct models. The procedure for getting each anti-model consists of four steps;

- Step 1: Obtain the segmented information for each phone from training data.
- Step 2: For each phone, run a phone recognizer.
- Step 3: Find top Q phones, and make a phone cohort.
- Step 4: Make an anti-model using the phone cohort.

There are two issues in the upper procedure. The first one is to decide the number of top Q phones for a phone cohort in Step 3. It is important to choose

the proper value for Q because the quality of HMM parameters for anti-models depends on how to make phone cohorts.

The second issue is how to model anti-models when context-dependent correct models are used. There are few study about it[8]. The conventional method is to use context-independent correct models only as basic units for phone recognizer in Step 2. Here the input data for phone recognizer is also labeled as context-independent units. Here, we can consider two different techniques to make each phone cohort using a phone recognizer based on context-dependent correct models. The first one is to use context-dependent models as basic units for the phone recognizer and to use the context-dependently labeled data as input data for phone recognizer. We call the first method as the context-dependent phones. The second one is to use context-independent models as basic units for the phone recognizer and to use the context-dependently labeled data as input data for the phone recognizer. We call the second method as the modified context-dependent techniques. Table 1 shows three different methods for context-dependent anti-models in detail.

3. EXPERIMENTAL RESULTS

3.1. Database

All experiments are done with the database consisting of 1,063 vocabularies including Korean company names as well as some additional keywords needed for retrieving stock information. A total of 62,717 words were used for training and 9,751 words were used for tests. Among test data, 2,089 OOV(Out-Of-Vocabulary) words were included. We designed a 64 context-independent phone set which is appropriate for the recognition of Korean words and expanded it into a 300 context-dependent phone set.

The speech signal is sampled at 8 kHz rate with a μ-law, 8 bit codec and pre-emphasized with a filter whose transfer function is $1 - 0.95z^{-1}$. The pre-emphasized speech is then divided into the frames. Each frame spans 20 msec and is overlapped by 10 msec. Linear Predictive Coding(LPC) analysis is performed and a set of LPC driven cepstral coefficient is computed from the LPC coefficients. The LPC driven cepstral coefficients are weighted by a window $W_c(m)$, of the form

$$W_c(m) = 1 + \frac{Q}{2}\sin(\frac{\pi m}{Q}), \quad 1 \le m \le Q,$$

where Q is the order of LPC. We used four kind of VQ codebooks for the speech recognition: (1) weighted LPC cepstral coefficients, (2) their differences, (3) their second order differences, (4) differenced log power and its second order differences.

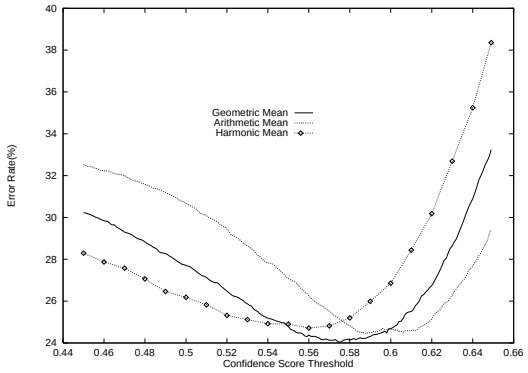

Figure 2: The Error Rates with regarding to Averaging methods

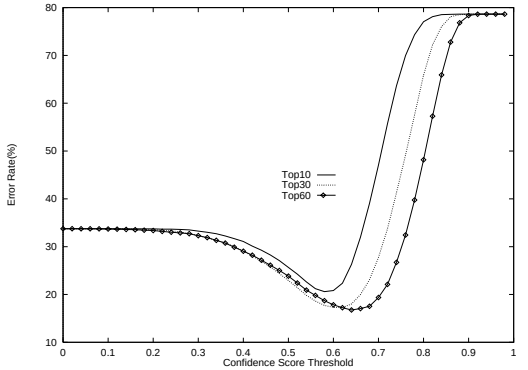

Figure 3: The Error Rates w.r.t. Various Q

3.2. Averaging Methods

We made a comparative study on three different kinds of average methods: arithmetic, geometric, and harmonic means. Figure 2 shows the comparative result with regrads to three averaging methods. We can get the best result with geometric mean.

3.3. Parameter for top Q

We also made a comparative study on various methods. Figure 3 shows the results for various Q's. The results shows that lager Q yields better recognition rate more robust to confidence measure threshold.

3.4. Context-dependent Anti-models

We made an experiment with regards to three different methods for obtaining context-dependent anti-models. Figure 4 shows their results. The results says that the modified context-dependent method gives us the best result.

4. CONCLUSION

In this paper, we described a Korean utterance verification system based on subword modeling for a vocabulary independent speech recognition system.

Table 1: Three Different Methods of Context-dependent Anti-models

Methods	Phone Recognizer	
	Input Data	Reference Patterns
Context-independent	Context-independent	Context-independent
Context-dependent	Context-dependent	Context-dependent
Modified Context-dependent	Context-dependent	Context-independent

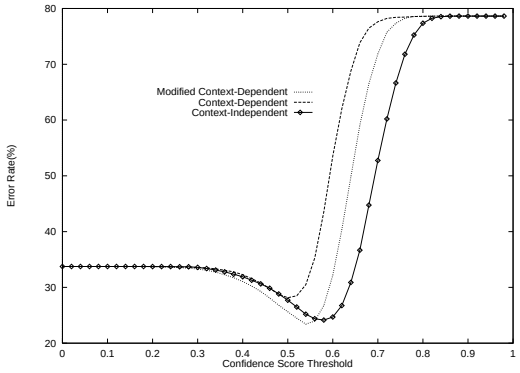

Figure 4: The Error Rates of Context-dependent Anti-models

Our system consist of two modules: recognition and verification, for utterance verification. In the stage of recognition, multiple hypotheses with hypothesized word boundaries are obtained through Viterbi decoder. In the stage of verification, unlikely hypothesis is rejected using confidence measure threshold. We use the likelihood ratio for confidence measure. We focused on the verification module. First, we made a comparative study on some methods for obtaining the confidence measure for words. Three kinds of average techniques were investigated as arithmetic, geometric, and harmonic averages. We can get the best result with geometric average. Second, we studied the effect of Q for phone cohort set. We found out that the largest Q yielded the best recognition result. Finally, we present how to model anti-models for correct context-dependent units. Three kinds of approaches are studied. The first one, context-independent method, is to use only context-independent units as both input data and reference patterns of phone recognizer for obtaining cohort set. The second one, context-dependent method, is to use cohort set based on only context-dependent units. The final one, modified context-dependent method, is to use context-dependent units for input data of phone recognizer and to use context-independent HMM parameters for its reference patterns. We made a comparative study on three approaches. The result shows that the modified context-dependent method gives us the best result.

5. REFERENCES

[1] R. A. Sukkar and J. G. Wilpon, "A two pass classifies for utterance rejection in keyword spotting," *Proc. of Int. Conf. on Acoustics, Speech and Signal Processing*, pp. 451–454, 1993.

[2] E. Lleida and R. C. Rose, "Efficient decoding and training procedures for utterance verification in continuous speech recognition," *Proc. of Int. Conf. on Acoustics, Speech and Signal Processing*, pp. 507–510, 1996.

[3] E. Lleida and R. C. Rose, "Likelihood ratio decoding and confidence measures for continuous speech recognition," in *Proc. Int. Conf. on Spoken Language Processing*, pp. 478–481, October 1996.

[4] M. W. Koo, C. H. Lee and B. H. Juang, "A new hybrid decoding algorithm for speech recognition and utterance verification," *Proc. of IEEE Workshop on Speech Recognition and Understanding* , pp. 507–510, 1996.

[5] M. W. Koo, C. H. Lee and B. H. Juang, "A new decoder based on a generalized confidence score," in *Proc. Int. Conf. on Spoken Language Processing*, pp. 213–216, May 1998.

[6] H.V. Poor, *An introduction to signal detection and estimation* Springer-Verlag, New York, 1988.

[7] S. Cox and R. Rose, "Confidence measures for the switchboard database," *Proc. of Int. Conf. on Acoustics, Speech and Signal Processing*, pp. 511–514, 1996.

[8] P. Ramesh, C. H. Lee and B. H. Juang, "Context dependent anti subword modeling for utterance verification," in *Proc. Int. Conf. on Spoken Language Processing*, October 1998.