LETTER

Utterance Verification Using State-Level Log-Likelihood Ratio with Frame and State Selection

Suk-Bong KWON[†a)], *Nonmember and* Hoirin KIM[†], *Member*

SUMMARY This paper suggests utterance verification system using state-level log-likelihood ratio with frame and state selection. We use hidden Markov models for speech recognition and utterance verification as acoustic models and anti-phone models. The hidden Markov models have three states and each state represents different characteristics of a phone. Thus we propose an algorithm to compute state-level log-likelihood ratio and give weights on states for obtaining more reliable confidence measure of recognized phones. Additionally, we propose a frame selection algorithm to compute confidence measure on frames including proper speech in the input speech. In general, phone segmentation information obtained from speaker-independent speech recognition system is not accurate because triphone-based acoustic models are difficult to effectively train for covering diverse pronunciation and coarticulation effect. So, it is more difficult to find the right matched states when obtaining state segmentation information. A state selection algorithm is suggested for finding valid states. The proposed method using state-level log-likelihood ratio with frame and state selection shows that the relative reduction in equal error rate is 18.1 % compared to the baseline system using simple phone-level log-likelihood ratios.

key words: *utterance verification, confidence measure, likelihood ratio testing, state-level log-likelihood ratio, frame selection, state selection*

1. Introduction

Many automatic speech recognition (ASR) systems were developed by many researchers, and some of them showed good recognition performance in specific conditions such as sufficient trained data, small vocabulary size, excellent devices, clean environments, and so on. Nowadays, the verification of recognition results is very important issue for making more intelligent ASR system because many applications want speech recognition system to be used in real field. The technique verifying recognition results with acceptance or rejection is called utterance verification (UV). It requires an algorithm computing the reliability or probability of correctness for recognition results. We call the value of the reliability or probability of correctness as confidence measure (CM). Up to now, various types of CMs have been proposed and implemented. Especially the likelihood ratio testing (LRT)-based [1]–[3] and *a posterior* probability-based [1], [4] have shown good performance in utterance verification systems. Also the algorithms combining some CMs to cope with diverse causes of mis-recognition were proposed. SVM (Support Vector Machines), FLDA (Fisher's Linear Discriminant Analysis), neural network, decision tree [1], and Bayesian [5] are representative fusion methods. Moreover, the neighborhood context-dependent acoustic models were used as anti-phone models in some research [2].

Usually phone-level log-likelihood ratio (PLLR) shows good performance in utterance verification, but it has some limitations such as obscure phone segmentation information, existence of bad frames in the input speech, and insufficient acoustic models and anti-phone models. Since the goal of speech recognition process is to find the best state sequence with maximum accumulated log-likelihood through Viterbi search, phone segmentation information obtained from speech recognition process is obscure. Short pause intervals and rapid transition intervals in the input speech are recognized certain phone in a word. Also, acoustic models and anti-phone models are insufficient to cover diverse pronunciation and coarticulation effect. Thus PLLR is unreliable in some cases because it is computed on these conditions. We suggest an utterance verification system using state-level log-likelihood ratio with frame and state selection to overcome some limitations of PLLR. Section 2 represents the overview of utterance verification system used in this paper and baseline LRT-based confidence measure. In Sect. 3, state-level log-likelihood ratio (SLLR) is explained. Also frame and state selection algorithms are described. Experiments and results of the proposed methods are described in Sect. 4. In the last chapter we give our conclusions and discuss future works in SLLR-based systems.

2. Utterance Verification

2.1 System Overview

The utterance verification is executed as the post process after speech recognition usually. Here, since we will use state-level confidence measure for computing word-level confidence measure, we need to obtain state segmentation information from the ASR system. We can obtain state segmentation information from our speech recognition system (EchoS-1.0) [6] easily because speech recognition process is to find a state sequence with maximum log-likelihood through Viterbi search algorithm. Figure 1 shows the utterance verification process using SLLR with state and frame selection algorithm. First, we obtain state segmentation information from speech recognition system and a log-likelihood sequence on the recognized state sequence. Additionally, a silence-based feature vector angle sequence is calculated for frame selection algorithm. In frame selec-

Manuscript received June 30, 2009.

Manuscript revised October 31, 2009.

[†]The authors are with Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea.

a) E-mail: sbkwon12@kaist.ac.kr

DOI: 10.1587/transinf.E93.D.647

Copyright © 2010 The Institute of Electronics, Information and Communication Engineers

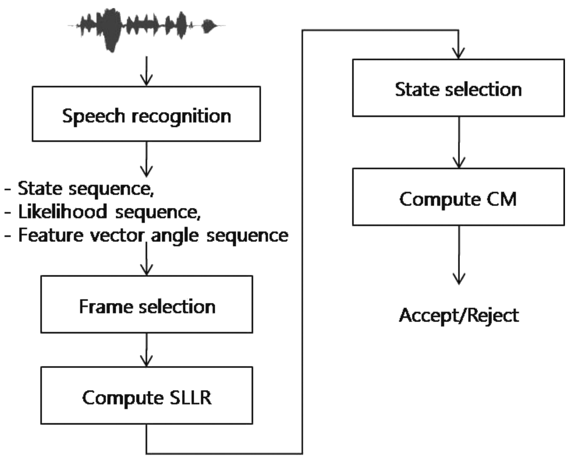

Fig. 1 Utterance verification system using state-level log-likelihood ratio with frame and state selection.

tion process, we find proper frames to compute more reliable confidence measures. Next, we compute SLLR on these selected frames. The state selection process does a role to remove ambiguous states that happened by coarticulation effect and pronunciation transition mainly. Finally we compute word-level confidence measure of a recognized word for decision of acceptance or rejection.

2.2 LRT-Based Confidence Measures

In this paper, we use the LRT-based confidence measure as a baseline [3]. The probabilistic hypothesis testing algorithm is background theory of the LRT-based confidence measure. The role of anti-phone models is competing with the recognized phone in order to obtain the reliability. The type of anti-phone models affects the performance of utterance verification, but the difference of performance is subtle. We use monophone-like-sets except a self-monophone model as anti-phone models. They can be easily obtained from training process of acoustic models for speech recognition. The PLLR is computed as

$$PLLR(ph) = \frac{\log P(X_{ph}|\lambda_{ph}) - \log P(X_{ph}|\bar{\lambda}_{ph})}{\tau(ph)} \quad (1)$$

where $\tau(ph)$ is the number of frames and X_{ph} is the feature vector sequence recognized as phone ph. λ_{ph} and $\bar{\lambda}_{ph}$ are an acoustic model and its anti-phone model of recognized phone ph, respectively. The word-level log-likelihood ratio (WLLR) is calculated by averaging sigmoid values of PLLRs.

$$WLLR(w) = \frac{1}{n_p(w)} \sum_{j=1}^{n_p(w)} sigmoid(PLLR(ph_j)) \quad (2)$$

where $n_p(w)$ is the number of phones composing a recognized word w. The sigmoid function is used to normalize PLLR into a value between 0 and 1.

3. State-Level Log-Likelihood Ratio with Frame and State Selection

3.1 State-Level Log-Likelihood Ratio

We use PLLR as the basic unit in our baseline LRT-based utterance verification system. It is calculated on a phone segment obtained from speech recognition process. Acoustic models used for speech recognition are hidden Markov models (HMMs) containing three states and basic phone unit is tied-state triphone. The states in HMM represent different characteristics of pronunciation and their distributions of log-likelihood are also different. In the middle state, mean of log-likelihoods is large and variance of them is low relatively. The middle state reflects more stable characteristics of pronunciation, so SLLR of the middle state is more important than others. This tendency happens on monophone-based anti-phone models. When we compute PLLR in baseline system, log-likelihood of anti-phone model is calculated without state segmentation information of a recognized phone. Hence we suggest SLLR to obtain more reliable confidence measure by reflecting these characteristics. The SLLR is computed as

$$SLLR(ph, s) = \frac{\log P(X_{(ph,s)}|\lambda_{(ph,s)}) - \log P(X_{(ph,s)}|\bar{\lambda}_{(ph,s)})}{\tau(ph, s)} \quad (3)$$

where $X_{(ph,s)}$ is feature vector sequence and $\lambda_{(ph,s)}$ is the acoustic models of s-state in recognized phone ph. $\bar{\lambda}_{(ph,s)}$ is the s-state in its anti-phone model of recognized phone ph. $\tau(ph, s)$ is duration of s-state in phone ph. The PLLR is calculated with sigmoid values of SLLRs as following

$$PLLR(ph) = \sum_{s=1}^{S} \frac{\omega_s \tau(ph, s)}{\tau(ph)} sigmoid(SLLR(ph, s)) \quad (4)$$

where S is number of states in an acoustic model. S is three in this paper. ω_s is the weight of s-state in a phone ph. ω_s is obtained empirically. The first and third states in HMM reflect transition characteristics of pronunciation and coarticulation effect. The middle state reflects relatively stable characteristics of pronunciation. Since the probability that the middle state have bad frames is lower than others, confidence measure of the middle state has a greater proportion of PLLR. We obtained same ω_s for s-state regardless of phones in this paper. $\tau(ph)$ is duration of a recognized phone ph. The WLLR is calculated with Eq. (2).

3.2 Frame and State Selection

In speech recognition process, unnecessary frames could be included in the recognition results. Since the goal of speech recognition is to find the best state sequence with maximum accumulated log-likelihood on given search network, unnecessary frames do not influence in recognition process severely because the number of them is relatively small compared to all frames of a recognized word. When we compute SLLR of a recognized state, these frames cause less reliable confidence measure because the size of these unnecessary frames in a state is relatively large and anti-phone models do not compete with a recognized state properly. Some silence intervals in front of the first phone and in rear of the last phone could be included state segmentation information in a recognized word. Short pause intervals within the input speech could be contained in certain recognized state. Also there are feature vectors that change between two consecutive phones rapidly. All of them disturb computing reliable confidence measure in spite of correct speech recognition results. Figure 2 shows that the input speech has some short pause intervals and rapid transition intervals.

In this paper, we suggest an algorithm to find useful frames for computing reliable confidence measure by deleting silence intervals and short pause intervals. For deletion of silence and short pause intervals, we use silence-based feature vector angle and log-likelihood sequence obtained from speech recognition process. The silence-based feature vector angle is computed as

$$\theta(x_t) = \cos^{-1}\left(\frac{x_t \circ x_s}{\sqrt{x_t \circ x_t}\sqrt{x_s \circ x_s}}\right) \tag{5}$$

where $\circ$ is the inner product operator between two vectors and x_s is the feature vector calculated with some frames that represent silence in front of the input speech. x_t is the feature vector at time t. The frame selection function is

$$FS(x_t) = \begin{cases} 0, & l(x_t) > thr_{sil} \text{ and } \theta(x_t) < \pi/4 \\ 1, & \text{otherwise} \end{cases} \tag{6}$$

where $l(x_t)$ is log-likelihood of feature vector x_t, and the thr_{sil} is the threshold for decision of silence with log-likelihood. New SLLR is re-estimated on the selected frames with Eq. (3).

The first and third states in HMM are trained for modeling transition characteristics of phones. Also, it is difficult to decide an accurate state between a recognized phone and its neighborhood phone in some cases and some states can be recognized forcedly on the search network nevertheless there are not matched feature vectors. Hence when we compute PLLR with SLLRs with Eq. (4), we suggest an algorithm to decide the useful states. The state decision function is

$$\begin{aligned} &SS(ph, s) \\ &= \begin{cases} 0, & \tau(ph, s) < thr_s \text{ and } \bar{l}(X_{(ph,s)}) < thr_l \\ 1, & \text{otherwise} \end{cases} \end{aligned} \tag{7}$$

where $\bar{l}(X_{(ph,s)})$ is mean of log-likelihood of feature vector sequence $X_{(ph,s)}$, thr_s is the threshold of the size of frames for deletion of a state, and thr_l is the log-likelihood thresholds for decision of invalid state. Rapid transition intervals are deleted by Eq. (7).

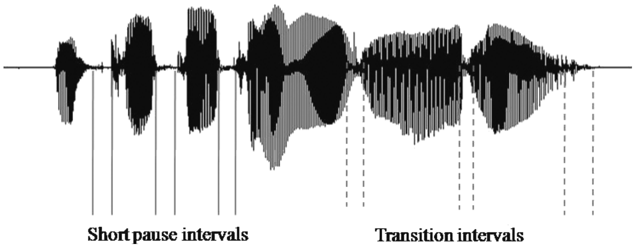

Fig. 2 Waveform of the input speech with short pause intervals and transition intervals. (16 phones exist in this input speech)

4. Experiments and Results

4.1 Experimental Environments

We used POW (Phonetically Optimized Words) 2002 database collected in ETRI (Electronics and Telecommunications Research Institute), South Korea to train acoustic models for speech recognition. Also, anti-phone models were trained with same database. The tied-state triphone CDHMM (Continuous Density Hidden Markov Models) type is used as the acoustic models. Each HMM includes 3 states and 7 Gaussian mixture components are contained in each state. The type of feature vector is a 39-dimension feature vector, consisting of 12 MFCC (Mel-Frequency Cepstrum Coefficients), 12 delta MFCC, 12 delta-delta MFCC, energy, delta energy, and delta-delta energy. CMN (Cepstral Mean Normalization) is adopted. The anti-phone models are phone-like-sets except a self-monophone model. That is, for 46 trained monophone models, each corresponding anti-phone model is obtained by a set of other monophone models except itself. We used a development data set to find thresholds to compute confidence measures in this paper such as thr_{sil}, thr_s, and thr_l. The development data set consists of 5,250 in-vocabulary utterances. The evaluation data set contains 3,675 utterances for in-vocabulary and 1,575 utterances for out-of-vocabulary. Here, the development data set and the evaluation data set are POI (Point of interest) database. Here, POI is a specific point location that someone may find useful or interesting used in navigation system such as address, office name, park name, building name, and so on.

4.2 Experimental Results

It was known that the LRT-based confidence measure works well in utterance verification systems. So the LRT-based confidence measure was used as baseline confidence measure in our utterance verification system. Table 1 shows that the performance of utterance verification system using SLLR is better than that using PLLR. Furthermore, when we ues SLLR with frame and state selection algorithms, this proposed method shows that the relative reduction in equal error rate was 18.1 % compared to the baseline system. In an experiment using only SLLR, the improvement is subtle because confidence measures are more reliable on some

Table 1 Performance comparison of utterance verification methods using baseline word-level log-likelihood ratio (WLLR) and state-level log-likelihood ratio with frame and state selection. (EER: Equal Error Rate, ERR: Error Reduction Rate)

CM type	*EER(%)*	*ERR(%)*
Baseline	17.11	-
SLLR	16.25	5.0
SLLR with frame selection	15.12	11.6
SLLR with frame and state selection	14.01	18.1

states, but they are not on some states. The main reason of this phenomenon is that the unnecessary frames could be used to compute SLLR and the size of these frames in segmented state interval is relatively larger than that in segmented phone interval. So we added frame selection algorithm on the SLLR-based system in a next experiment. We obtained 11.6 % error reduction rate. From this result, we know that unnecessary frames prevent from computing more reliable confidence measure. In the last experiment, we used all frame and state selection algorithms on the SLLR-based system. It shows the best performance among our experiments by deleting some useless states additionally. Since the speech is pronounced continuously with coarticulation effect and diverse types, some useless states in the state sequence can be recognized forcedly. Therefore it is efficient to delete these states.

5. Conclusions

In this paper, SLLR using state segmentation information was proposed in replace of PLLR using phone segmentation information for obtaining more reliable confidence measure. Since states in HMM represent different characteristics, it is more reliable to compute SLLR on matched states between a recognized phone and its competing anti-phone model and to give proper weights on states when computing PLLR with SLLRs. Generally, state segmentation information is less precise than phone segmentation information and the input speech has bad frames such as silence, short pause and rapid transition intervals. As the number of frames for computing SLLR become small, bad frames effect SLLR severely. So it is important to find useful frames or intervals for more reliable SLLR. Hence we proposed frame selection algorithm to select proper frames. The state selection algorithm finding valid states also improves the utterance verification performance. For overcoming the limitations of baseline system such as obscure phone segmentation information, existence of bad frames in the input speech, and insufficient acoustic models and anti-phone models, we used SLLR with frame and state selection algorithm. When a mis-recognized word is very similar with its correct word, it is still difficult to verify a recognized word with the proposed method. Hence we will research efficient algorithms to solve this problem by focusing on ambiguous intervals when computing confidence measure of a word as further works.

References

[1] H. Jiang, “Confidence measure for speech recognition: A survey,” Speech Commun., vol.45, no.4. pp.455–470, April 2005.

[2] H. Jiang and C.-H. Lee, “A new approach to utterance verification based on neighborhood information in model space,” IEEE Trans. Speech Audio Process., vol.11, no.5, pp.425–434, Sept. 2003.

[3] K.-S. Moon, Y.-J. Kim, H.-R. Kim, and J.-H. Chung, “Out-of-vocabulary word rejection algorithm in Korean variable vocabulary word recognition,” IEEE International Symposium on Circuits and Systems, vol.5, pp.53–56, May 2000.

[4] F. Wessel, R. Schluter, M. Macherey, and H. Ney, “Confidence measures for large vocabulary continuous speech recognition,” IEEE Trans. Speech Audio Process., vol.9, no.3, pp.288–298, March 2001.

[5] T.-Y. Kim and H. Ko, “Bayesian fusion of confidence measures for speech recognition,” IEEE Signal Process. Lett., vol.12, no.12, pp.871–874, Dec. 2005.

[6] O.-W. Kwon, H.-R. Kim, S.-B. Kwon, S.-R. Kim, G.-C. Jang, Y.-R. Kim, B.-W. Kim, C.-D. Yoo, and Y.-J. Lee, “Development of a Korean large vocabulary continuous speech recognition platform (ECHOS),” Proc. O-COCOSDA 2007, pp.108–111, Dec. 2007.