

Utterance Verification Using Search Confusion Rate and Its N-Best Approach

Kyuhong Kim, Hoirin Kim, and Minsoo Hahn

ABSTRACT—Recently, a variety of confidence measures for utterance verification has been studied to improve speech recognition performance by rejecting out-of-vocabulary inputs. Most of the conventional confidence measures for utterance verification are based primarily on hypothesis testing or an approximated posterior probability, and their performances depend on the robustness of an alternative hypothesis or the prior probability. We introduce a novel confidence measure called a search confusion rate (SCR), which does not require an alternative hypothesis or the approximation of posterior probability. Our confusion-based approach shows better performance in additive noise-corrupted speech as well as in clean speech.

Keywords—Utterance verification, confidence measure, out-of-vocabulary rejection.

I. Introduction

In general, speech recognition systems estimate the acoustic likelihood $\Pr(X|W)$ and language model probability $\Pr(W)$ in order to find the best word sequence W^* given the feature vector sequence X in the utterance.

$$\begin{aligned} W^* &= \arg \max_w \Pr(W|X) \\ &= \arg \max_w \frac{\Pr(X|W) \Pr(W)}{\Pr(X)} \\ &= \arg \max_w \Pr(X|W) \Pr(W) \end{aligned} \quad (1)$$

This maximum likelihood approach has been successfully applied to speech recognition under the assumption that the

prior probability $\Pr(X)$ is constant for the given utterance. As described in (1), the word sequence is independent of the prior probability, and the speech recognizer searches for the most likely word sequence within the predefined vocabulary.

When there is a possibility for the utterance to contain out-of-vocabulary words, the recognizer should be able to decide whether the utterance contains out-of-vocabulary words or not. The posterior probability $\Pr(W|X)$ is one of the best confidence measure candidates. However, ignoring the prior probability $\Pr(X)$ makes it difficult to use the Viterbi search score $\Pr(X|W) \Pr(W)$ for a confidence measure.

Most confidence measures have focused on the practical approximation of the prior probability. Namely, they have attempted to estimate the prior probability using mainly the likelihood ratio or the hypothesis tests from the anti-model [1], the catch-all [2], the cohort model [3], or others. The conventional approaches always require additional computations to obtain an alternative hypothesis, and their performances are dependent on how to model or approximate the alternative models.

In this letter, we introduce a *search confusion rate* (SCR) as a confidence measure for utterance verification. The proposed method has no need to model an alternative hypothesis or to approximate the acoustic prior probability.

II. Search Confusion in Speech Recognition

The accumulated log likelihood obtained by Viterbi search is calculated for every possible state, and low-score paths are pruned to reduce the search space. The momentarily best-scored path could be a part of a recognized result or be pruned later. At the end of the search, only the best-scored path is considered as a recognition result, and all other paths have no

Manuscript received Mar. 28, 2005; revised June 08, 2005.

This work is the result of the URC project sponsored by MIC, Korea.

Kyuhong Kim (phone: +82 42 866 6221, email: kkh@jcu.ac.kr), Hoirin Kim (email: hrkim@jcu.ac.kr), and Minsoo Hahn (email: mshahn@jcu.ac.kr) are with the School of Engineering, Information and Communications University, Daejeon, Korea.

influence on the recognition result. However, these momentarily best-scored paths have meaningful information that can be used for utterance verification.

```
// Initialization
 $\delta_1(i) = \pi_i b_i(x_1), 1 \leq i \leq N_{state}$ 
 $\psi_1(i) = 0, 1 \leq i \leq N_{state}$ 
 $m_1 = \arg \max_{1 \leq i \leq N_{state}} \delta_1(i)$ 
// Recursive step
 $\delta_t(j) = \max_{1 \leq i \leq N_{state}} (\delta_{t-1}(i) a_{ij} b_j(x_t)), 1 \leq j \leq N_{state}$ 
 $\psi_t(j) = \arg \max_{1 \leq i \leq N_{state}} (\delta_{t-1}(i) a_{ij}), 1 \leq j \leq N_{state}$ 
 $m_t = \arg \max_{1 \leq i \leq N_{state}} \delta_t(i)$ 
// Final step
 $p = \max_{1 \leq i \leq N_{state}} \delta_T(i)$ 
 $s_T = \arg \max_{1 \leq i \leq N_{state}} \delta_T(i)$ 
 $m_T = s_T$ 
// Backtracking
 $s_t = \psi_t(s_{t+1}), t = T-1, T-2, \dots, 1$ 
```

Fig. 1. Viterbi search with tracing momentary best states.

1. Search Confusion Rate

In the proposed scheme, the momentarily best states which show the highest Viterbi score are traced frame by frame during a Viterbi search. As depicted in Fig. 1, m_t is the state that momentarily shows the maximum Viterbi score $\delta_t(i)$. After the Viterbi decoding process, a few or many of the momentarily best scored states m_t coincide with the decoded states s_t , whether the recognition result is correct or not. If the recognizer is confused among in-vocabulary words during the Viterbi search, s_t and m_t will be frequently mismatched. This frequent mismatch implies that the spoken word may be an out-of-vocabulary word or seem to be misrecognized as an incorrect word. Thus, the degree of mismatch can be regarded as a confidence measure for the recognition result. To quantify the mismatch, we define the SCR as follows:

$$SCR(S, M) = \frac{1}{T} \sum_{t=1}^T d(\Gamma(s_t), \Gamma(m_t)), \quad (2)$$

$$d(t_a, t_b) = \begin{cases} 0 & \text{if } t_a = t_b, \\ 1 & \text{otherwise.} \end{cases} \quad (3)$$

In (2), $S (= s_1 s_2 \dots s_T)$ and $M (= m_1 m_2 \dots m_T)$ represent the decoded state sequence and momentarily best-scored state

sequence, respectively. Term $\Gamma(s)$ is the triphone that contains the state s , and $d(\cdot)$ is a distance function between two sub-word models which are triphones in this framework. For simplicity, we approximated it to the binary distance metric in (3). The physical meaning of the SCR is how much the recognizer is confused. The SCR score is compared with a threshold to decide whether the recognized result is correct or not. When the SCR falls below the threshold, the recognition result is judged to be correctly recognized.

2. Experimental Setup

To show the effectiveness of our confusion-based measure, we set up a 200 word recognizer and utilize the confidence measure for word verification as a post-processor. We use only Korean PBW (phonetically balanced words) 452 database which contains a large number of phoneme combinations in Korean. It is composed of 452 different words, each uttered twice by 36 males and 36 females. Speech samples in the Korean PBW 452 database are recorded at 16 kHz sampling rate using 16 bit quantization. In training and testing the acoustic hidden Markov models, about 90% and 10% of Korean PBW 452 database are involved, respectively. The number of test words for in-vocabulary and out-of-vocabulary are 3,200 and 4,032, respectively. The baseline system is a speaker-independent speech recognizer in noise-free environments.

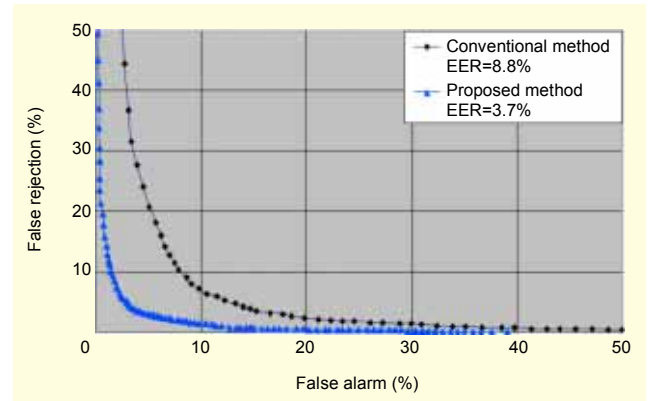


Fig. 2. Receiver operating characteristic curves.

For the performance comparison, we employ the generalized sub-word level confidence measure [1] because it is one of the most popular likelihood ratio test (LRT)-based confidence measures for utterance verification. We estimate a false alarm and false rejection with different thresholds, and the receiver operating characteristic curves in Fig. 2 are then estimated for performance evaluations. From the results, we confirm that our confusion-based confidence measure

outperforms all operational points in a clean condition. Compared with the conventional LRT-based method, our confidence measure shows a 58% reduction in equal error rate (EER). The EER is the error rate when a false alarm is equal to a false rejection.

III. Noise Robustness and N-Best Approach

1. Noise Robustness of SCR

The LRT-based confidence measure makes use of the likelihood ratio which changes drastically with noise corruptions. SCR is expected to be a noise-robust confidence measure compared with the LRT-based method because it utilizes the coincidences of ordered states which are not easily affected by noise corruptions.

2. N-Best SCR

One of the drawbacks of the SCR is that the frame score in (2) is a binary value which sometimes causes an unreliable confidence score. For example, if a recognition result is right but its competing hypotheses are frequently classified into the momentary best state during the Viterbi search, the SCR score will be high and cause a false rejection. The frame SCR with only one momentary best state fails to softly represent how much the recognizer is confused frame by frame; the frame SCR in (2) is a binary representation based on whether it is confused or not. For the purpose of improving the stability of the SCR, our original idea is expanded to an N-best approach by tracing multiple momentary best-scored states during the Viterbi search. The N-best SCR as a confidence measure is defined as follows:

$$SCR_{N\text{-best}}(S, M) = \frac{1}{T} \sum_{t=1}^T \frac{1}{N} \sum_{n=1}^N d(\Gamma(S_t), \Gamma(m_t^n)), \quad (4)$$

$$m_t^n = \arg \max_{\substack{1 \leq i \leq N_{\text{state}} \\ i \neq m_t^{n-1}, \dots, m_t^1}} \delta_i(i), \quad 1 \leq n \leq N, \quad (5)$$

where N_{state} is the number of states. The momentary N-best states m_t^n are traced by (5) during the Viterbi search.

3. Computations for N-Best SCR

In each frame, the computational requirements for N-best SCR are just N comparisons; the LRT-based confidence measure requires two log-likelihood calculations and one sigmoid operation in each frame. Because the processing delay in real-

time speech recognizer applications is a critical constraint, our approach is more attractive, especially in low-power consuming and low-speed hardware-based applications such as speech interface programs in PDAs, cellular phones, and others.

4. Experimental Results under Different Noisy Conditions

For this experiment, we use the Korean PBW database as well as the Korean POW (phonetically-optimized words) database, which was used in [5]. Every utterance is pre-emphasized with a gain of 0.97, and 20 ms Hamming windows are applied with 10 ms overlapping. The feature vector for each frame consists of 13th-order static, delta, and delta-delta mel-frequency cepstral coefficients, resulting in the final 39th-order feature vector. A total of 8,927 tied-state triphones were trained for the baseline speech recognizer. The acoustic models are estimated using the Korean POW database while the Korean PBW database is used for the performance evaluation. From the Korean PBW database, the number of test words for in-vocabulary and out-of-vocabulary are 28,800 and 36,288, respectively.

For the performance comparison, LRT-based [1], SCR, and N-best SCR methods were tested. In this experiment, we set up a speaker-independent, vocabulary-independent speech recognizer under different additional noisy environments. This baseline system shows 97.35% word accuracy for in-vocabulary clean speech. To consider practical noisy environments, we added restaurant, babble, and car noises from the AURORA database [4] into the Korean PBW with 20 dB and 10 dB signal-to-noise ratios, respectively.

Table 1. EERs under different noisy environments (%).

	Restaurants		Babble		Car	
	20 dB	10 dB	20 dB	10 dB	20 dB	10 dB
LRT	39.7	42.9	40.0	42.2	30.8	39.6
SCR	23.1	28.5	23.1	27.6	20.7	31.1
10-Best SCR	20.0	27.5	19.9	26.3	16.9	29.1

Table 2. EERs under clean and multi-condition (%).

	Clean	Multi-condition
LRT	29.4	36.8
SCR	19.1	24.1
10-best SCR	16.1	21.5

As depicted in Tables 1 and 2, our proposed N-best SCR consistently shows better performance in different kinds of noisy environments. In clean and 20 dB signal-to-noise ratio conditions, approximately 50% of the EERs are reduced by N-best SCR compared with the LRT-based method. The multi-condition in Table 2 indicates six different noisy conditions including a clean condition. Even in the multi-condition, the N-best SCR shows about a 41% better EER compared with the LRT-based method.

IV. Conclusions

In this letter, we propose a novel confusion-based confidence measure. Compared with the LRT-based confidence measure, our measure shows better noise robustness owing to ordering characteristics. Also, its computational requirements are negligible. Through several experiments, we confirm that our confusion-based method shows better operational characteristics than an LRT-based measure; this indicates that the proposed search confusion rate is an efficient and effective confidence measure.

The distance function in (3) was approximated to the binary distance metric for simplicity. Our future work will include finding a proper distance measure between states.

References

- [1] M.W. Koo, C. H. Lee, and B. H. Juang, "Speech Recognition and Utterance Verification Based on a Generalized Confidence Score," *IEEE Trans. Speech and Audio Proc.*, vol. 9, no. 8, Nov. 2001, pp. 821-832.
- [2] S. Kamppari and T. Hazen, "Word and Phone Level Acoustic Confidence Scoring," *Proc. ICASSP*, 2000, pp.1799-1820.
- [3] R. Sukkar and C. H. Lee, "Vocabulary Independent Discriminative Utterance Verification for Nonkeyword Rejection in Subword Based Speech Recognition," *IEEE Trans. Speech and Audio Proc.*, vol. 4, no. 6, Nov. 1996, pp. 420-429.
- [4] H. G. Hirsch and D. Pearce, "The AURORA Experimental Framework for the Performance Evaluations of Speech Recognition Evaluations of Speech Recognition Systems under Noisy Conditions," *ISCA ITRW Automatic Speech Recognition: Challenges for the Next Millennium*, Paris, France, Sept. 18-20, 2000.
- [5] H-Y Jeong, "Filtering of Filter-Bank Energies for Robust Speech Recognition," *ETRI J.*, v. 26, no. 3, June 2004, pp. 273-276.