Approximated Posterior Probability for Scoring Speech Recognition Confidence

Kyuhong Kim(ICU), Hoirin Kim(ICU)

<Contents>

<Abstract>

Approximated Posterior Probability for Scoring Speech Recognition Confidence

Kyuhong Kim and Hoirin Kim

This paper proposes a new confidence measure for utterance verification with posterior probability approximation. The proposed method approximates probabilistic likelihoods by using Viterbi search characteristics and a clustered phoneme confusion matrix. Our measure consists of the weighted linear combination of acoustic and phonetic confidence scores. The proposed algorithm shows better performance even with the reduced computational complexity than those utilizing conventional confidence measures.

* Keywords: Speech recognition, Utterance verification, Confidence measure

1. Introduction

The general, speech recognition systems estimate the acoustic likelihood $\Pr(A|W)$ and the language model probability $\Pr(W)$ to find the best word sequence W^* from the feature vector A of the utterance. This maximum likelihood approach has been successfully applied to speech recognizers under the assumption that the acoustic prior probability $\Pr(A)$ is a constant for the given utterance. As described in equation (1), the word sequence is independent of the prior probability. And the speech recognizer searches the most likely word sequence within the predefined vocabulary.

$$W^* = \arg\max_W \Pr(W \mid A) = \arg\max_W \frac{\Pr(A \mid W)\Pr(W)}{\Pr(A)}$$
$$= \arg\max_W \Pr(A \mid W)\Pr(W) \quad (1)$$

When there is a possibility for the utterance to contain an OOV(out-of-vocabulary), the recognizer should be able to decide whether the utterance contains OOVs or not. The posterior probability $\Pr(W|A)$ is one of the best confidence measure candidates. However, because of ignoring the acoustic prior probability $\Pr(A)$, it is difficult to use the recognition score $\Pr(A|W)\Pr(W)$ as a confidence measure. Many confidence measures have focused on the practical approximation of the prior probability. Namely, they have attempted to estimate the prior probability, mainly based on the likelihood ratio and the hypothesis tests, from the catch-all, the anti-model, the cohort model, etc. Recently, the graph, the lattice, and the N-best approaches have reported the best performances with rather increased complexity for the first two than the third [1-4]. In this paper, we propose a new algorithm to calculate recognition confidence for the utterance verification by approximating the posterior probability. The posterior probability is approximated and divided into acoustic and phonetic confidence scores. The acoustic confidence score is based on the likelihood ratio of the decoded sub-word model to the constrained anti-model, while the phonetic confidence score is calculated by utilizing a clustered phoneme confusion matrix. In this paper, we firstly describe the approximation procedures and assumptions. Secondly, the estimation of the acoustic and the phonetic confidence score is introduced. Finally, we describe experimental setups with the algorithmic performance analysis

2. Approximated posterior probability for confidence measure

The posterior probability $\Pr(W|A)$ may be one of the best candidates for the confidence measure. The posterior probability can be represented as product-sum of the acoustic likelihood $\Pr(p|A)$ and the phonetic likelihood $\Pr(W|p)$ over all possible sub-word sequences p as described in equation (2). When the product-sum deviation from N-best phone sequences is very small, we can approximate the posterior probability by limiting the search space from all the possible phone sequences to N-best ones as follows.

$$\Pr(W \mid A) = \sum_{p} \Pr(W \mid p) \Pr(p \mid A) \approx \sum_{N-best\ p} \Pr(W \mid p) \Pr(p \mid A) \tag{2}$$

Moreover, if multiple pronunciation lexicons are not used, only the 1-best case may be considered in equation (2). So the posterior probability can be further approximated as follows [5].

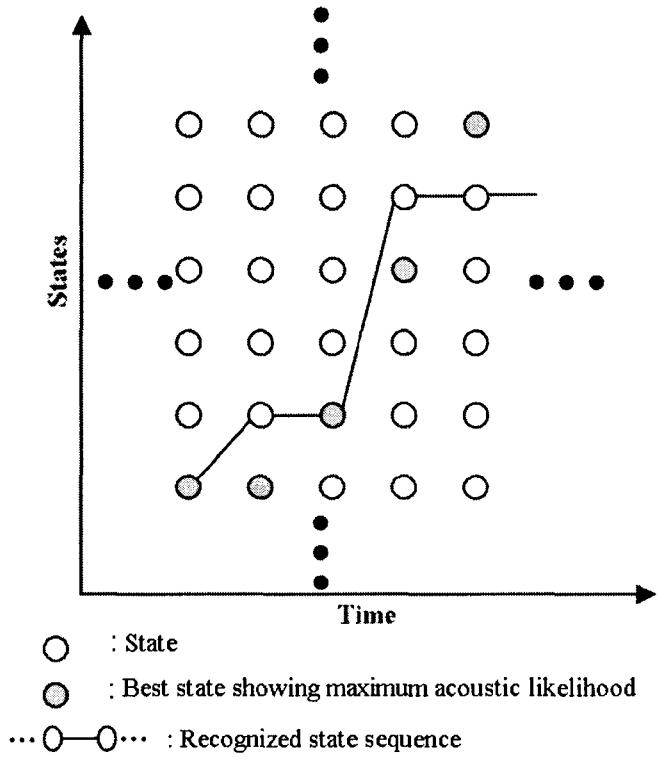
$$\Pr(W \mid A) \approx \sum_{N-best\ p} \Pr(W \mid p) \Pr(p \mid A) \approx \Pr(W \mid p^*) \Pr(p^* \mid A) = \tilde{\Pr}(W \mid A) \tag{3}$$

In equation (3), the approximated posterior probability $\tilde{\Pr}(W \mid A)$ consists of the phonetic and the acoustic probabilities. In scoring recognition confidence, it is natural to assume that the hypothesized word sequence W is already determined by the speech recognizer. We try to estimate two probabilities by using a clustered confusion matrix and the likelihood ratio of hypothesized sub-word models to their constrained anti-models. Cox, et al. had used only the phonetic probability $\Pr(W|p)$ in their meta-model, and then the coincidence ratio was adopted as the confidence measure [5]. On the contrary, our proposed method utilizes the acoustic as well as the phonetic confidence measureuence of frames classified as speech or non-speech frames.

2.1. Best phoneme sequence for acoustic and phonetic confidence measures

Before estimating the acoustic and the phonetic likelihoods in equation (3), the unknown best phoneme sequence p^* having the maximum likelihood at a given

instance is traced by using the Viterbi search. We assume that candidate models within the Viterbi beam usually include the cohort model, the neighborhood information, and the best phoneme sequence [4]. The best phoneme sequence $p^* = p_1^* p_2^* \cdots p_T^*$ within the Viterbi beam is traced by selecting the state showing the maximum acoustic likelihood for every frame.



<Figure 1> Best state sequence and recognized state sequence in Viterbi search space

The best phoneme sequence can be different from the recognized one because the latter is determined by the different maximum criterion that utilizes both the accumulated acoustic and the language model likelihoods. <Figure 1> shows the detailed description of the best and the recognized state sequence search. The proposed algorithm finds one best state for each frame and the phoneme sequence can be easily constructed by simply concatenating the central phones because our network is a triphone-based one. The best phoneme is represented in equation (4).

$$\Pr(A_t \mid p_t^*) = \max_{p_t \in beam} \Pr(A_t \mid p_t)$$

$$p_t^* = \arg\max_{p_t \in beam} \Pr(A_t \mid p_t) \tag{4}$$

2.2 Acoustic confidence measure

The acoustic probability $\Pr(p^*|A)$ is represented by the Bayesian rule, and then we assume that the phonetic probability $\Pr(p)$ is uniform and the summation of likelihoods along all possible phone sequences is separated into two terms of the best-phone model and the anti-phone model. Finally, the acoustic confidence score $CM_{acoustic}$ can be approximated to the likelihood ratio between the phone and the anti-phone likelihood as follows.

$$\Pr(p^* \mid A) = \frac{\Pr(A \mid p^*)\Pr(p^*)}{\sum_{all\ p} \Pr(A \mid p)\Pr(p)} = \frac{\Pr(A \mid p^*)\Pr(p^*)}{\Pr(A \mid p^*)\Pr(p^*) + \Pr(A \mid p^{*anti})(1 - \Pr(p^*))}$$

$$= \frac{1}{1 + \frac{\Pr(A \mid p^{*anti})(1 - \Pr(p^*))}{\Pr(A \mid p^*)\Pr(p^*)}} \approx \frac{\Pr(A \mid p^*)}{\Pr(A \mid p^{*anti})} = CM_{acoustic}(A, p^*) \tag{5}$$

Here, $\Pr(A|p^{*anti})$ can be approximated with the phone $p^{*anti} \approx \arg\max_{all\ phone\ within\ beam,\ phone \neq p^*} \Pr(A \mid p)$, because other phone likelihoods may be assumed to be relatively negligible. The anti-phone model is defined as the model showing maximum likelihood except the given phone model. Hence the approximated anti-phone model becomes the second-best model within the beam space. In fact, the approximated likelihood ratio is conceptually identical to the cohort model based hypothesis test. We propose a frame-level acoustic confidence measure as follows.

$$LCM_{acoustic}(A, p^*, t) = \log \Pr(A_t \mid p_t^*) - \log \Pr(A_t|p_t^{*anti}) \tag{6}$$

Because the best phoneme model and its anti-model can be simultaneously traced during the Viterbi search, additional calculations for confidence scoring are negligible.

2.3 Clustered phoneme confusion probability

The recognized word W can be represented as a phoneme sequence $W = q_1 q_2 \cdots q_T$. For simplification, we have assumed that the phonetic descriptions are statistically independent and the prior probability $\Pr(\cdot)$ is uniform. The phonetic probability is proportional to the product of frame-based phonetic confusion probabilities.

$$\Pr(W = q_1 q_1 \cdots q_T \mid p = p_1 p_1 \cdots p_T) = \prod_{t=1}^{T} \Pr(q_t \mid p_t) = \prod_{t=1}^{T} \frac{\Pr(q_t \mid p)\Pr(q_t)}{\Pr(p_t)} = \prod_{t=1}^{T} \Pr(p_t \mid q_t) \tag{7}$$

Because the relation between q_t and p_t cannot be known in advance, we try to model it with a confusion matrix. We firstly separate the training set into the correct and the incorrect word sets after a closed recognition experiment, and then estimate the histograms for all phonemes. A phoneme confusion probability is estimated by using this histogram. We propose a frame-level phonetic confidence measure as follows.

$$LCM_{phonetic}(W, p^*, t) = \log \Pr(p_t^* \mid q_t) \tag{8}$$

2.4 Proposed confidence measure

We propose a frame-level confidence measure as the weighted sum of the acoustic and the phonetic confidence measures.

$$LCM(W,t) = \alpha \cdot LCM_{acoustic}(A, p^*, t) + \beta \cdot LCM_{phonetic}(W, p^*, t) \tag{9}$$

Here, both α and β can have the value between 0 and 1. This confidence measure is used to verify the utterance to be acceptable or not by comparing the confidence sc ore with a prefixed threshold γ as shown in equation (10).

$$LCM_{Proposed}(W) = \frac{1}{t_e - t_s}\sum_{t=t_s}^{t_e} LCM(W,t) \begin{matrix} accept \\ \geq \\ < \\ reject \end{matrix} \gamma \quad (10)$$

Here, t_s and t_e are the start and the end frame of the utterance W.

3. Experiments and results

Typical conventional and our proposed confidence measures are evaluated for OOV detection experiments. A CDHMM-based vocabulary independent speech recognizer is employed as our baseline. As described in <Table 1>, both the Korean phonetically balanced word (PBW) and the Korean phonetically optimized word (POW) databases are used for performance evaluation[6].

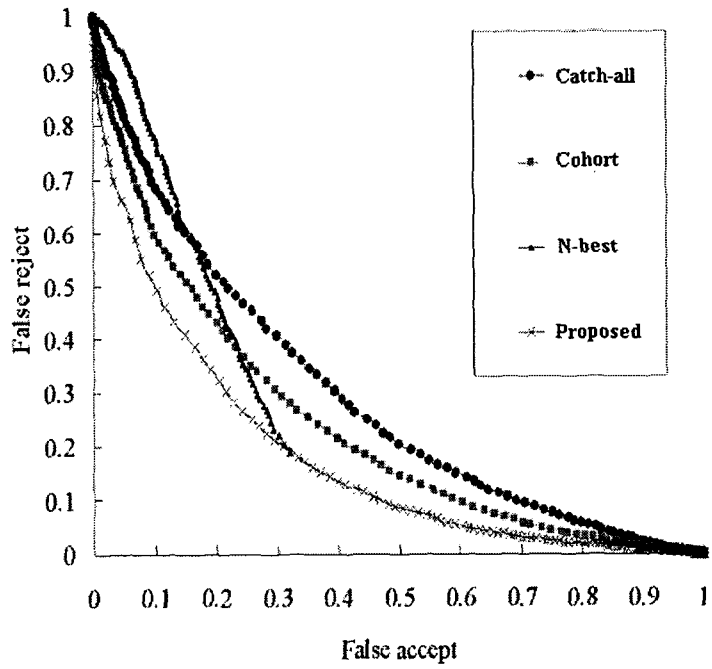
<Table 1> Korean PBW and POW speech database

	PBW	POW
Vocabulary size	452	3,848
Number of speakers	72	40
Sampling rate	16 kHz	16 kHz
Utterances	65,088	37,521

Every utterance is pre-emphasized with 0.97 and 20 msec Hamming windows are applied with 10 msec overlapping. Our feature vector for each frame consists of 13th-order original, the delta, and the delta-delta Mel-Frequency Cepstral Coefficients (MFCC) resulting in the final 39th-order feature vector. 8,927 tied-state triphones were used for our baseline speech recognizer. The acoustic models and the confusion probability matrix are estimated with the POW database while the PBW database is used for the performance evaluation. The number of in-vocabulary and OOV words are 3,851 and 252, respectively.

For the comparative performance analysis, we evaluate the performances of the most popular confidence measures, i.e., the catch-all with 46 clusters, the cohort, and the N-best methods [7-8]. To make use of N-best scoring in OOV detection, 100-best candidates are extracted for every utterance, then the coincidence ratio is evaluated. For the phonetic probability evaluation, the sub-words, which are 8,927 tied-state

triphones, are clustered into 46 models for memory efficiency. The weighting coefficients, α and β, are all set to 0.5. All our comparative performance evaluation results are depicted in <Figure 2>. In this figure, we can confirm that our proposed method always produces the best performance. <Table 2> shows the equal error rate of the conventional and proposed methods for OOV rejection. When compared to acoustic confidence measures based on the catch-all and the cohort model, the proposed confidence measure reduces the equal error rate about 28.7% and 16.2%, respectively. Also, even when compared to the N-best score, our proposed method reduces it about 8.3%. In addition, our confidence measure does not require any additional acoustic likelihood calculation, which is necessary for the confidence measure based on the conventional log likelihood ratio tests, because the acoustic likelihoods are already calculated during the Viterbi search.



<Figure 2> Receiver operating characteristics

<Table 2> Performance evaluation results

	Catch-all	Cohort	N-best	Proposed
EER	35.5%	30.2%	27.6%	25.3%

4. Conclusion and further works

In this letter, we have proposed a new, simple confidence measure as the weighted linear combination of the acoustic and the phonetic confidence scores. In order to validate the usefulness of our new confidence measure, we applied it to the OOV rejection task with a vocabulary independent speech recognition system. The results confirm that our proposed confidence measure always outperforms the conventional algorithms, i.e., 8.3% error reduction rate improvement even when compared with that of the N-best method, while guaranteeing the reduced computational complexities.

Our future works may include the noise robustness test of our proposed measure, verification power for confusable words, application to the key word spotting, and finally, usefulness verification in continuous large vocabulary speech recognition.

References

[1] F. Wessel, R. Schluter, et al., "Confidence measures for large vocabulary continuous speech recognition", *in Proc. of IEEE Trans. on Speech and Audio Processing*, vol. 9, no. 3, Mar. 2001.

[2] T. Kemp and T. Schaaf, "Estimating confidence using word lattices", 5th Euro. Conf. on Speech Communication and Technology, Greece, pp.827-830, Sept. 1997.

[3] L. Gillick, Y. Itou, and J. Young, "A probabilistic approach to confidence estimation and evaluation", *in Proc. of ICASSP*, pp. 879-882, 1997.

[4] H. Jiang and C. H. Lee, "A new approach to utterance verification based on neighborhood information in model space", *in Proc. of IEEE Trans. on Speech and Audio Processing*, vol. 11, no. 5, Sept., 2003.

[5] S. Dasmahapatra and S. Cox, "Meta-models for confidence estimation in speech recognition", *in Proc. of ICASSP*, pp.1815-1818, 2000.

[6] Y. Limand Y. Lee, "Implementation of the POW (phonetically optimized words) algorithm for speech database", *in Proc. of ICASSP*, pp.89-92, 1995.

[7] R. Sukkar and C. H. Lee, "Vocabulary independent discriminative utterance verification for nonkeyword rejection in subword based speech recognition", *in Proc. of IEEE Trans. on Speech and Audio Processing*, vol. 4, no. 6, Nov. 1996.

[8] S. Kamppari and T. Hazen, "Word and phone level acoustic confidence scoring", *in Proc. of ICASSP*, pp.1799-1820, 2000.

접수일자: 2004년 11월 15일
게재결정: 2004년 12월 10일

▶ 김규홍(Kyuhong Kim)
주소: 305-714 대전광역시 유성구 문지로 119번지 한국정보통신대학교
소속: 한국정보통신대학교(ICU) 공학부
전화: 042) 866-6221
E-mail: kkh@icu.ac.kr

▶ 김회린(Hoirin Kim)
주소: 305-714 대전광역시 유성구 문지로 119번지 한국정보통신대학교
소속: 한국정보통신대학교(ICU) 공학부
전화: 042) 866-6139
E-mail: hrkim@icu.ac.kr