PAPER Special Section on Processing Natural Speech Variability for Improved Verbal Human-Computer Interaction

Enhancing the Robustness of the Posterior-Based Confidence Measures Using Entropy Information for Speech Recognition*

Yanqing SUN[†a)], **Yu ZHOU**[†b)], ***Nonmembers***, **Qingwei ZHAO**[†c)], ***Member***, **Pengyuan ZHANG**[†d)], **Fuping PAN**[†e)], ***and*** **Yonghong YAN**[†f)], ***Nonmembers***

SUMMARY In this paper, the robustness of the posterior-based confidence measures is improved by utilizing entropy information, which is calculated for speech-unit-level posteriors using only the best recognition result, without requiring a larger computational load than conventional methods. Using different normalization methods, two posterior-based entropy confidence measures are proposed. Practical details are discussed for two typical levels of hidden Markov model (HMM)-based posterior confidence measures, and both levels are compared in terms of their performances. Experiments show that the entropy information results in significant improvements in the posterior-based confidence measures. The absolute improvements of the out-of-vocabulary (OOV) rejection rate are more than 20% for both the phoneme-level confidence measures and the state-level confidence measures for our embedded test sets, without a significant decline of the in-vocabulary accuracy.

key words: *OOV, speech recognition, confidence measure, entropy information, phoneme-level posterior*

1. Introduction

For many practical speech recognition applications, the rejection of out-of-vocabulary (OOV) words is an important issue. In this research, automatic speech recognition technology is used to recognize the user's speech from words in a word list. A user not familiar with the system may utter OOV words, which are not included in the system's lexicon or in the specific word list. If no measures are taken, the system will always output a recognition result given any input, which may cause incorrect reactions or incur a high cost. To avoid this problem, confidence measures, such as the posterior which has been widely used for speech recognition [1], [2], are computed to make the rejection decision. In case the condition of the test data matches with the condition of the training data, the speech recognition system performs well, as do the posterior-based confidence measures. However, the training and the test data sometimes differ, and the performances of recognition may decline owing to the mismatch, as do the performances of hidden Markov model (HMM)-based confidence measures, such as the posterior. In our research, the training data is well-pronounced English speech, while the test data varies considerably and sometimes contains dialect words or non-native English speech. As a result, the posterior calculated for the training model may not provide a reliable confidence measure for all the test speech, particularly when the training and test data are seriously mismatched. This is why robust confidence measures are needed in this research.

To improve the robustness of the posterior-based confidence measures, many methods utilizing the N best candidate recognition results have been developed, such as the lattice- and confusion-network-based confidence measures in [3] and [4], respectively. By utilizing all the other candidate word arcs with similar time boundaries [5], the generated entropy information can be used to weigh the conventional confidence measures. These improvements have been proven to be complementary and may benefit the posterior-based confidence measures. However, there are some limitations in these methods. For example, when there are few candidate results generated by decoding, sometimes only the best result is available and the performances of the lattice- and the confusion-network-based confidence measures are very poor.

In this study, background information such as the entropy is directly investigated at the speech-unit level. This method shares similar information to the lattice-based confidence measures but may be more comprehensive and is not affected by the recognition. In our research, two levels of speech units are investigated, i.e., senones/states [6] and phonemes [7], although the calculation of entropy does not rely on the levels of the posteriors. To normalize the entropy information to make it comparable with the posterior-based confidence measures, two kinds of entropy-based confidence measures are proposed. Details of the calculations, such as the combination domains, averaging methods used to form higher-level confidence measures, and maximum approximations for practical use, are also discussed. Although this study is focused on experiments on the OOV rejection problems based on a word recognition task, the performances of the confidence measures are improved in a general criterion measured by the equal error rate (EER). Therefore, our proposed algorithm is promising for enhancing the robustness of general confidence measures consist-

Manuscript received December 1, 2009.
Manuscript revised March 1, 2010.

[†]The authors are with the Institute of Acoustics, Chinese Academy of Sciences, China.

*This work was partially supported by the National Science & Technology Pillar Program (2008BAI50B03), and National Natural Science Foundation of China (No. 10925419, 90920302, 10874203, 60875014).

a) E-mail: sunyanqing@hccl.ioa.ac.cn
b) E-mail: zhouyu@hccl.ioa.ac.cn
c) E-mail: zhaoqingwei@hccl.ioa.ac.cn
d) E-mail: zhangpengyuan@hccl.ioa.ac.cn
e) E-mail: panfuping@hccl.ioa.ac.cn
f) E-mail: yanyonghong@hccl.ioa.ac.cn

DOI: 10.1587/transinf.E93.D.2431

ing of the posterior, particularly when the test data does not match the training very well in some practical applications.

This paper is organized as follows. In Sect. 2, the proposed system, as well as a conventional one, is described. The practical details of the posterior calculations are also discussed, including the logarithms, the averaging methods, and the combination domains. In Sect. 3, the two levels of speech units and their confidence measures are introduced. The mining of corresponding entropies is also discussed, which will be used together with the posterior-based confidence measures. The results of experiments carried out separately on the two levels are reported in Sect. 4, with the different combination methods discussed in detail. Conclusions are given in Sect. 5.

2. Entropy Information in the Posterior-Based Confidence Measures

2.1 Proposed Algorithm for Enhancing Posterior-Based Confidence Measures

The algorithm used to enhance the robustness of the posterior-based confidence measures is shown in Fig. 1. For the common posterior-based confidence measures, the likelihoods of all the speech units are calculated and summed to obtain a single posterior for the recognized speech unit. In fact, the posteriors of the other speech units can also be obtained without increasing the complexity compared with that of the calculation of likelihoods, and the posteriors can form background information such as the entropy. As entropy can provide a measure of uncertainty, it is expected by combining it with a single posterior, the robustness of confidence measures can be improved. Details will be given in the following subsections.

2.2 Common Model and Practical Uses in the Calculations of the Posterior-Based Confidence Measures

We assume that there are a total of N_u speech units, which can be phonemes or states. Given an observation O, which is a feature sample used for speech recognition, the posterior probability of being the jth speech unit u_j from the best recognition result is defined as

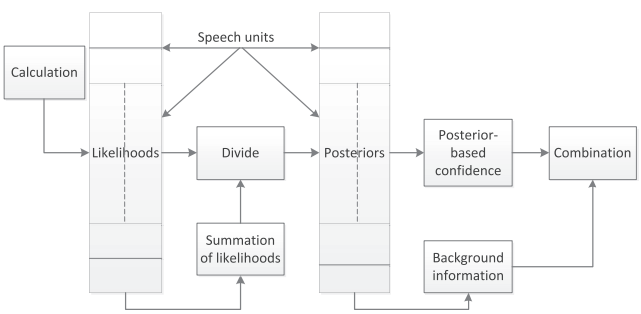


Fig. 1 Algorithm for enhancing the robustness of the posterior-based confidence measures using background information such as the entropy.

$$p(u_j|O) = \frac{p(u_j, O)}{p(O)} = \frac{p(O|u_j)p(u_j)}{\sum_{i=1}^{N_u} p(O|u_i)p(u_i)}, \tag{1}$$

where $p(O|u_j)$ is the likelihood of seeing the observation O given the jth speech unit. Generally, all the speech units are assumed to have equal a priori probability; thus, the terms $p(u_j)$ and $p(u_i)$ can be removed from the numerator and denominator simultaneously. For practical uses, when only the logarithms of the likelihoods are available, we write

$$logl(u_j) = \log(p(O|u_j)) \tag{2}$$

$$logp(u_j) = \log(p(u_j|O)), \tag{3}$$

then the logarithm of the posterior can be calculated as

$$logp(u_j) = logl(u_j) - logadd(u_1, \ldots, u_{N_u}) \tag{4}$$

$$logadd(u_1, \ldots, u_{N_u}) = \log(\sum_{i=1}^{N_u} \exp(logl(u_i))) \tag{5}$$

$$= logl_{max} + \log(\sum_{i=1}^{N_u} \exp(logl(u_i) - logl_{max}))$$

$$\approx logl_{max}.$$

We write $logl_{max} = logl(u_m)$, which is the maximum of the $logl(u_i)$, and if $logl(u_i) - logl_{max} << 0$ when $i \neq m$, the approximation is reasonable.

The posterior can give a direct measure of the probability of a speech unit when the observation is given. However, it does not made good use of the background information, except for in the summation of likelihoods.

2.3 Conventional N-Best-Based Entropy

Instead of simply using the best recognition results, using other candidates in the N best recognition results might provide useful information for enhancing the performance of posterior-based confidence measures in [5], such as

$$C_{entropy} = CM * (1 - E_{avg}), \tag{6}$$

where CM denotes the conventional posterior-based confidence measures and E_{avg} is the entropy information calculated using other candidates. Because it is convenient for implementation, E_{avg} is calculated using confusion networks as [4]

$$E_{avg} = -\frac{1}{\log_2 N} \sum_{i=1}^{N} P(C_i|O) \log_2 P(C_i|O), \tag{7}$$

where N is the number of candidates, C_i is the ith candidate word recognized, and $P(C_i|O)$ is its posterior given the observation O.

2.4 Proposed Speech-Unit-Level Entropy Confidence Measures

When the posteriors of all the speech units are available, the speech-unit-level entropy is defined as

$$H(u) = -\sum_{i=1}^{N_u} p(u_i|O)\log(p(u_i|O)) = -\sum_{i=1}^{N_u} logp(u_i)\exp(logp(u_i)). \quad (8)$$

According to the property of entropy, the lower bound can be reached when only one speech unit takes the posterior of 1 while that of the others is 0, then $H(u) = 0+1*\log(1) = 0$. The upper bound can be reached when all the speech units have an equal posterior, which is $\frac{1}{N_u}$, then $H(u) = -N_u * \frac{1}{N_u} * \log(\frac{1}{N_u}) = \log(N_u)$. Reaching the lower bound indicates that the maximum amount of information is provided and that a certain speech unit can be determined. Reaching the upper bound indicates that the minimum amount of information or no information is provided, and no speech unit can be determined. As $H(u)$ is increased from 0 to $\log(N_u)$, the uncertainty of attributing the observation to a certain speech unit increases.

From the above analysis, $H(u)$ can be used to measure the value of uncertainty. To be comparable to the posterior, with the original range $[0, \log(N_u)]$ of the entropy normalized to $[0, 1]$, two kinds of entropy-based confidence measures are defined as

$$cm_{H_1} = \frac{\frac{N_u}{\exp(H(u))} - 1}{N_u - 1} \quad (9)$$

$$cm_{H_2} = 1 - \frac{H(u)}{\log(N_u)}, \quad (10)$$

where cm_{H_1} and cm_{H_2} are also referred to as the first entropy and second entropy, respectively. As $\log(cm_{H_1} + \frac{1}{N_u-1}) = \log(N_u) - \log(N_u - 1) - H(u)$, for consistency with the logarithm of the posterior, which is often used, the logarithm of cm_{H_1} is simplified to $-H(u)$ to facilitate computation, which means that the negative entropy is equivalent to the logarithm of the posterior. For cm_{H_2}, given in Eq. (10), which is equivalent to the negative entropy, it appears that these two kinds of entropy-based confidence measures have a logarithmic relationship. The performances of these two kinds of entropy-based confidence measures will be compared by performing experiments.

Although entropy information can provide the uncertainty of attributing an observation to a certain speech unit, it is not specifically attributed to a concrete unit but to an abstract unit. Thus, it is impossible to integrate the entropy information with the recognition information at the speech-unit level, and the entropy information can only be referred to as a general confidence measure.

2.5 Normalization and Averaging Methods

Confidence measures are usually calculated for a segment, such as a phoneme segment. As the durations of segments differ from each other, confidence measure is often normalized by the duration to remove its effect. The confidence measure of each frame is considered to be independent, thus, the average of their logarithms over the time range $[t_s, t_e]$ is often used as the final confidence measure as follows to avoid overflow in the computation:

$$cm_g = \exp\left(\frac{\sum_{t=t_s}^{t_e}\log(cm(t))}{t_e - t_s}\right). \quad (11)$$

This is actually the geometric mean of the per-frame confidence measures. For the entropy-based confidence measures, the arithmetic mean is also utilized to convert the per-frame entropies to higher levels of entropy, which are then compared with the geometric mean:

$$cm_a = \frac{\sum_{t=t_s}^{t_e} cm(t)}{t_e - t_s}. \quad (12)$$

2.6 Integration of Entropy Information with the Posterior-Based Confidence Measures

As analyzed in the above sections, both the posterior and entropy information provide useful information for confidence measures. Their combination can provide more information and is expected to achieve a better performance.

Let us reconsider the limitations of the posterior. In most cases, the posterior of a certain speech unit can reflect the confidence of a given observation when the model and speech match. However, in the case of mismatch, the posterior is not sufficiently reliable. The degree of matching can be judged using the entropy information, which can provide a measure of the reliability of the posterior confidence.

A linear combination of posterior and entropy is very simple to use and can be applied in the logarithm domain as well as in the original domain:

$$cm_1(u_j) = \alpha_1 * p(u_j|O) + (1-\alpha_1)cm_H \quad (13)$$

$$cm_2(u_j) = \alpha_2 * logp(u_j) + (1-\alpha_2)\log(cm_H) = \log(p(u_j|O)^{\alpha_2} * cm_H^{1-\alpha_2}), \quad (14)$$

where $\alpha_1, \alpha_2 \in [0, 1]$ are always selected using the development sets and are not changed during the experiments. In particular, the combination in the logarithm domain is a form of nonlinear combination in the original domain, which is similar to the method in [5], where $\alpha_2 = 0.5$. Experiments will be carried out to compare the performances in the two domains by combining the posterior with the proposed entropy confidence measures.

3. Two Levels of Speech Units and Their Posterior-Based Confidence Measures

There are many speech-unit-based posteriors that can be used for confidence measures. Among them, two posterior-based confidence measures can be calculated using the same HMM-based acoustic model, i.e., the state-level posterior and the phoneme-level posterior, which are widely used in speech recognition and pronunciation evaluation. From these posteriors, entropy information can be calculated separately. Details will be discussed in the following subsections.

3.1 State-Level Posterior

When states are the basic speech units in the HMM model, the state-level posterior is very commonly used and can be calculated directly. First, the per-frame log-likelihood of each state is calculated using the probability density functions of each state's multivariate Gaussian mixture model (GMM) as

$$N(x|\mu,\Sigma) = \frac{1}{(2\pi)^{\frac{N}{2}}|\Sigma|^{\frac{1}{2}}} e^{[-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)]} \tag{15}$$

$$p(o_t|u_i) = \sum_{k=1}^{M} p_k * N(o_t|\mu_k,\Sigma_k) \tag{16}$$

$$\begin{aligned} logl_s(u_i) = logadd(& \\ \log\left(\frac{p_1}{(2\pi)^{\frac{N}{2}}|\Sigma_1|^{\frac{1}{2}}}\right) - \frac{1}{2}(o_t-\mu_1)^T\Sigma_1^{-1}(o_t-\mu_1), \ldots, & \\ \log\left(\frac{p_M}{(2\pi)^{\frac{N}{2}}|\Sigma_M|^{\frac{1}{2}}}\right) - \frac{1}{2}(o_t-\mu_M)^T\Sigma_M^{-1}(o_t-\mu_M)), & \end{aligned} \tag{17}$$

where o_t is the observation for the tth frame, p_k, μ_k, and Σ_k are the weight, mean vector, and covariance matrix for the kth ($k \in 1..M$) component of the GMM model for the ith state, and $N(x|\mu,\Sigma)$ is the probability density function of the multivariate Gaussian distribution, respectively.

Second, the logarithms of the per-frame posteriors are calculated for each frame given the state-level time labels using Eq. (4). Here, N_u is the number of states in the HMM model, which is on the order of 10^2 or 10^3, and may be much larger than the number of phonemes. Third, the logarithms of the per-frame posteriors are converted to the phoneme-level confidence measures or higher levels such as that of a whole utterance, using Eq. (11) or Eq. (12).

To calculate the entropy for the state-level posteriors, the logarithm of likelihood summation is calculated using Eq. (5) as the first step. Then the logarithm of all the state posteriors can be simply derived by subtraction as suggested in Eq. (4). Then the state-level posterior-based entropy is finally calculated using Eq. (8), Eq. (9), or Eq. (10).

There are other methods of utilizing the state-level likelihoods for entropy calculation, for example, using the likelihoods of longer segments instead of the likelihoods per frame, such as per state, per phoneme, or per word. Longer segments appear to be more robust. However, they may also lead to a loss of information when combining the likelihoods of shorter segments. The performance of each segment length will be compared by performing experiments.

3.2 Phoneme-Level Posterior

The phoneme-level posterior was proposed in [8], [9]. To calculate the phoneme-level posterior directly for a segment using the same state-based HMM model, a phoneme loop should be constructed first. Then each phoneme should be represented by the corresponding states for the posterior calculation. As the triphone is the basic speech unit for most state-based HMM models, the context phonemes obtained from the recognition of the current phoneme should be utilized for triphone expansions, as demonstrated in Fig. 2. For example, the triphone expansion for ph_k at the current position is $ph(t-1) - ph_k - ph(t+1)$. A special phoneme 'silence' is used as the context phoneme for the first and last phonemes in a sentence.

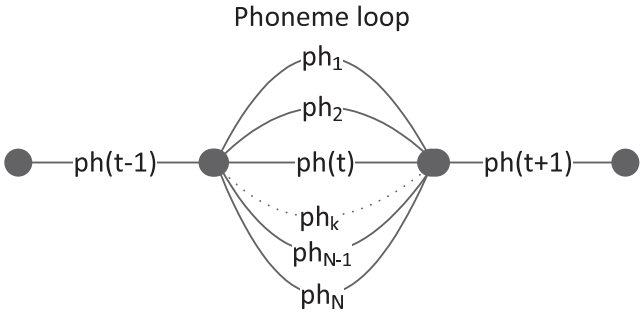


Fig. 2 Construction of triphone loop with context phonemes for the phoneme-level posterior calculation.

To determine the time boundaries of the substates within each triphone expansion, Viterbi decoding was performed for each triphone to map the phoneme segment into its corresponding state sequences and obtain the maximum log-likelihood of each triphone. After the log-likelihood of each triphone is obtained, the posterior of each phoneme can be calculated using Eq. (4). Here, N_u is the number of phonemes in the phoneme set, which is approximately 50.

4. Experiments

From the above analysis, two kinds of entropy confidence measures are proposed for speech-unit-level posteriors, and each of them is combined with the posterior to form the final confidence measure. As they are not specific to a certain speech-unit level, experiments are carried out on two typical levels, i.e., the phoneme-level and state-level, to obtain the performances and their improvements relative to that of the posterior-based confidence measures. The phoneme-level posterior, which is faster in practical embedded systems, will be used principally to obtain our conclusions, which will be verified using the other posterior.

4.1 Experimental Setup

The evaluations involve the recognition of English words in a word list in an embedded device. The THINKIT (speech lab) embedded speech recognition system [10] is utilized, which is based on the HMM model, tree-searching algorithm for a given grammar, and its word list [11]. The acoustic model is triphone based and contains about 3500 states, each modeled by eight Gaussian mixtures. Its features are the 12 dimensions of perceptual linear prediction (PLP) cepstras plus their first-order deltas. The phoneme set is from the CMU pronouncing dictionary, containing 39 phonemes, without counting 'short pause' and 'silence'. Five test sets

Table 1 Construction of the sets used in the experiments and their overall information without any confidence measures, including 5 IV sets and 3 OOV sets, where '*Dev' signifies 'used as development sets', 'list' signifies which word list the test set belongs to, 'size' signifies the number of utterances in the sets, and 'accuracy' signifies the recognition accuracy for the IV set and the correct rejection rate for the OOV set.

	type	id	list	size	speakers	accuracy (%)	date
Test		1	2	1340	9	90.75	10.05
Test		2	3	400	8	94.50	10.31
Test	IV	3	1	291	9	88.32	11.01
*Dev		4	1	272	8	91.54	11.02
Test		5	1	476	14	91.81	11.04
*Dev		1	2	947	-	0	-
Test	OOV	2	3	1907	-	0	-
Test		3	1	1736	-	0	-

are recorded using the board over several days using three different word lists (labeled '1', '2', and '3') at a sampling rate of 8 kHz. When a test set is recognized using a grammar compiled with its corresponding word list, five invocabulary (IV) sets are constructed from the five test sets. Given a word list, an OOV set is constructed by selecting all the utterances whose transcriptions do not exist in the word list. Three OOV sets are constructed for the three word lists. The confidence measure's performance is evaluated in terms of both the recognition accuracy for the IV set and the correct rejection rate for the OOV set.

Among these sets, one IV set and one OOV set are picked as development sets to determine the threshold and combination parameters for each confidence measure system. The performances of the development sets are also included to account for the performance of each confidence measure or combination.

Overall information without any confidence measures for all sets is given in Table 1. The accuracy for the IV set is approximately 90%; however, the recognition system has no ability to reject the OOV set. Using confidence measures, it is hoped that the system can reject as much of the OOV input as possible, while maintaining almost the same accuracy for the IV set.

4.2 Phoneme-Level Confidence Measures

The phoneme-level posterior is calculated in accordance with Sect. 3.2. From the result in Fig. 3, we can see that the EER is about 22.4% for the baseline posterior-based confidence measures, which may cause many false acceptations of utterances from the OOV set. Then the entropy information is utilized to enhance the posterior performance. Two kinds of entropy applied at the speech-unit level (cm_{H_1} in Eq. (9) and cm_{H_2} in Eq. (10)), and different combination domains (Eq. (13) for the original domain and Eq. (14) for the logarithm domain) are compared with the conventional method using the N-best-based entropy (Eq. (6) and Eq. (7)). The detection error trade-off (DET) curves are shown in Fig. 3.

From the above figures, it can be seen that cm_{H_2} performs better than cm_{H_1} (after combination with posterior entropy, the EER decreases from 6.9% to 6.7% in the logarithm domain (shown on the left of Fig. 3), and from 7.3% to 6.4% in the original domain). Both cm_{H_1} and cm_{H_2} perform much better than the traditional N-best-based entropy (19.8%). Only the combination of the posterior and cm_{H_1} in the logarithm domain (Eq. (14)) is chosen in the subsequent experiments for the following reasons:

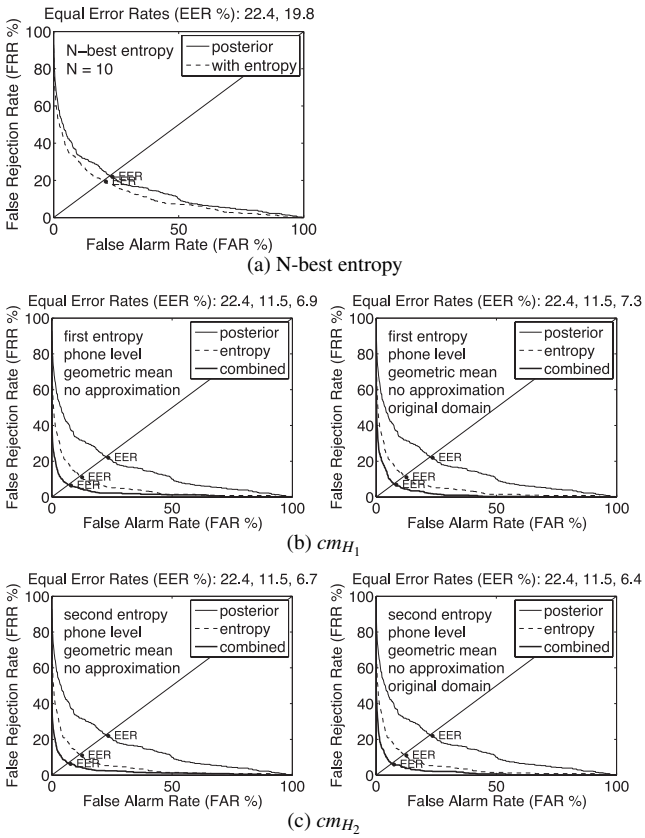
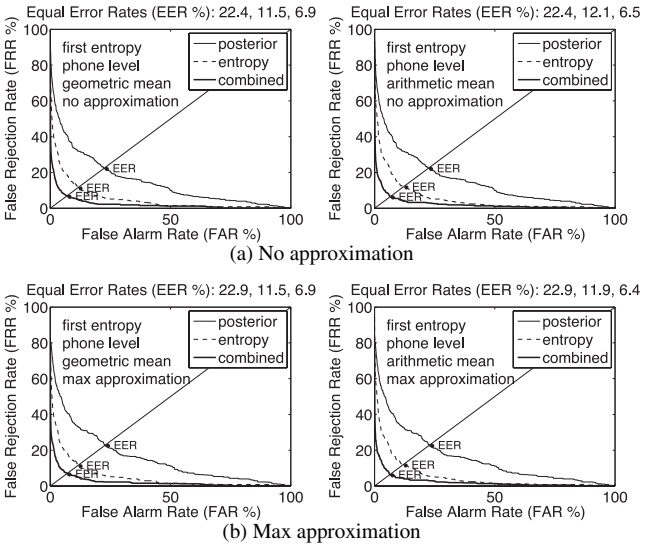


Fig. 3 DET curves of different entropies as well as different combination domains. While the proposed entropies calculated at the speech-unit level show significant improvements in terms of the EER, the two kinds of entropy cm_{H_1} (first entropy) and cm_{H_2} (second entropy) have comparable performance. While the combination of cm_{H_1} and the posterior has a better performance in the logarithm domain (left), the combination of cm_{H_2} and the posterior has a better performance in the original domain (right).

- cm_{H_1} has fewer logarithm operations than cm_{H_2}, and the combination of cm_{H_1} and the posterior in the logarithm domain is easier to calculate when using $-H(u)$ as an approximation of its logarithm.
- Combining cm_{H_1} and the posterior in the logarithm domain gives better performance than their combination in the original domain; the combination weight is 0.524 for each confidence measures, which is almost unbiased.
- This combination allows a comparison with the method proposed in [5] as stated at the end of Sect. 2.6.

For practical uses, the maximum approximation in Eq. (5) is often used to increase the speed of the system but with a certain decrease in performance. The DET curves for the maximum approximation as well as for different averaging methods (Eq. (11) and Eq. (12)) are shown in Fig. 4



(a) No approximation

(b) Max approximation

Fig. 4 Comparison of performances of the entropy calculation for the phoneme-level posterior. No obvious differences are observed between the performances using geometric mean and arithmetic mean, using maximum approximation or not. Thus, for the phoneme-level posterior-based entropy, the effects of both the averaging methods and the maximum approximation are not obvious.

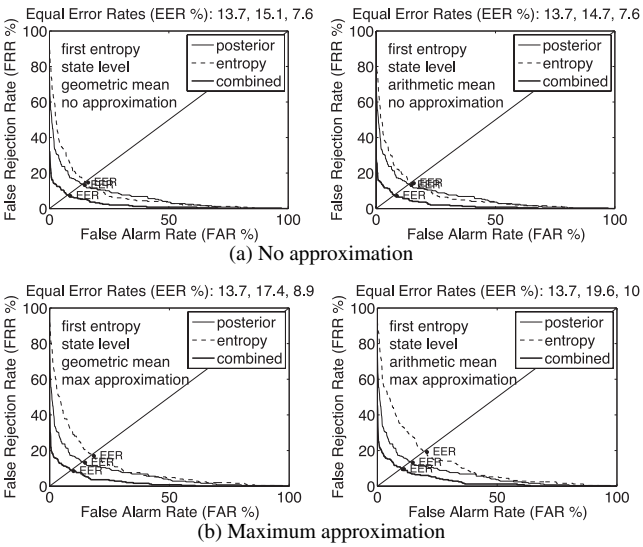
using only the first entropy (cm_{H_1}) to visualize the effects of the maximum approximation and the different averaging methods.

From Fig. 4, it can be seen that the EERs are similar for entropy confidence measures both with and without the maximum approximation (both 11.5% for the geometric mean and approximately 12% for the arithmetic mean); thus, the effects of both the maximum approximation and the different averaging methods can be omitted. Therefore, to facilitate computation, both the combination of the maximum approximation and the arithmetic mean are used for the final phoneme-level posterior-based entropy. For each confidence measure system using different calculation details, the threshold and the combination parameter are tuned for the development sets so that the absolute decline of accuracy for the IV set is less than 1% using the grid search method, and the parameters with the highest correct rejection rate of the OOV set are obtained. The parameters are then applied to the other sets. The detailed performances for all the sets are given in Table 2.

From Table 2 and the above observations, two conclusions can be drawn. First, from the last two columns in Table 2, it can be seen that by combining entropy information with phoneme-level posterior-based confidence measures, the correct rejection rate of the OOV set is increased significantly (more than 25–30% in absolute terms), while the decrease in accuracy for the IV set is negligible (from 90.24% to 90.04% for cm_{H_2}). Second, the two kinds of entropy-based confidence measures are comparable in performance. As cm_{H_2} requires more calculations such as the logarithm which is time-consuming, only cm_{H_1} will be used in the following experiments to increase the speed of calculation.

Table 2 Comparison of phoneme-level confidence measures for development and test sets: posterior and combinations with the two kinds of entropy confidence measures. 'Av' signifies 'average', 'no cm' means that no confidence measure is used, '$+H_1$' means combination with cm_{H_1}, and '$+H_2$' means combination with cm_{H_2}.

	type	id	accuracy or rejection rate(%)			
			no cm	posterior	$+H_1$	$+H_2$
Test	IV	1	90.07	89.55	89.92	89.78
Test		2	94.50	94.00	94.25	94.25
Test		3	88.32	86.94	86.60	85.91
*Dev		4	91.54	90.81	90.81	90.81
Test		5	91.81	89.92	89.92	89.50
		Av	91.25	90.24	90.30	90.05
Test	OOV	1	0	9.00	36.54	47.10
*Dev		2	0	10.98	48.61	57.89
Test		3	0	22.13	40.84	51.27
		Av	0	15.78	42.00	52.09



(a) No approximation

(b) Maximum approximation

Fig. 5 Comparison of performances of the entropy calculations for state-level posterior and its entropy using cm_{H_1}. It is clearly shown that the maximum approximation deteriorates the performance, while the two averaging methods do not have a significant effort on performance, although the geometric mean appears to be more robust. Thus, for the state-level posterior-based entropy, the geometric mean and no approximation are subsequently used.

4.3 State-Level Confidence Measures

Although it is time-consuming, the state-level posterior, which is used widely and is believed to yield better performance, is calculated in accordance with section 3.1. As shown in Fig. 5, the EER is about 13.7% for the state-level posterior-based confidence measures, which is much lower than that of the phoneme-level posterior. To determine whether the consideration of entropy improves performance and whether the maximum approximation affects the performance, the performances of different averaging methods and the maximum approximation are also compared in Fig. 5.

From Fig. 5, it can be seen that the effect of the maximum approximation cannot be disregarded. The perfor-

Table 3 Comparison of state-level confidence measures for the development and test sets: the posterior and the combination of cm_{H_1} with and without the maximum approximation. 'Av' signifies 'average', 'no cm' means that no confidence measure's method is used, '$+H_1$' means combination with cm_{H_1}, and '$H_{1_{maxa}}$' means the combination of cm_{H_1} with the maximum approximation.

	type	id	accuracy or rejection rate(%)			
			no cm	posterior	$+H_1$	$+H_{1_{maxa}}$
Test	IV	1	90.07	88.81	89.55	89.63
Test		2	94.50	93.25	93.50	94.00
Test		3	88.32	86.25	85.57	86.25
*Dev		4	91.54	90.81	90.81	90.81
Test		5	91.81	89.71	90.76	90.97
		Av	91.25	89.77	90.04	90.33
Test	OOV	1	0	33.47	60.72	57.55
*Dev		2	0	56.17	76.66	68.59
Test		3	0	46.37	64.23	59.33
		Av	0	45.34	67.20	61.82

mances of the different averaging methods are different when using the maximum approximation, and the performance is better using the geometric mean. Thus, for the state-level posterior-based entropy, only the geometric mean is subsequently used. The threshold is tuned similarly to that for the phoneme-based confidence measures. As the maximum approximation (Eq. (5)) affects the performance, performances both with and without the maximum approximation are given in Table 3.

In Table 3, from the last two columns it can be seen that the performances of the combination of posterior and entropy information both with and without the maximum approximation are comparable, and both give much better performance than the posterior-based confidence measures as shown in the last three columns. For the state-level posterior-based calculation, since the number of speech units is so large that the maximum approximation can significantly reduce the number of logarithm addition calculations, the maximum approximation may be adopted if the degradation of performance is limited.

The above state-level posterior is calculated for each frame, although it can also be calculated for longer segments. To obtain posteriors for longer segments, the per-frame state-level likelihoods are converted to the likelihoods of three longer segments, i.e., the per-state segment, per-phoneme segment, and per-word segment. As entropy can be calculated for any posterior, DET curves are drawn in Fig. 6 to compare the quantity of information provided by different levels of entropy.

It can be seen in Fig. 6, as the segment length increases, information is lost. The performances of the per-frame entropy and per-state entropy are comparable. The per-phoneme entropy can still provide some information for the per-frame posterior-based confidence measures although its performance is greatly inferior to those of the above two segment lengths. However, the per-word entropy provides no information. To obtain quantitative results, the parameters are tuned for each confidence measure system, and the performances are given in Table 4 except for that of the per-word entropy, which causes a major degradation in performance.

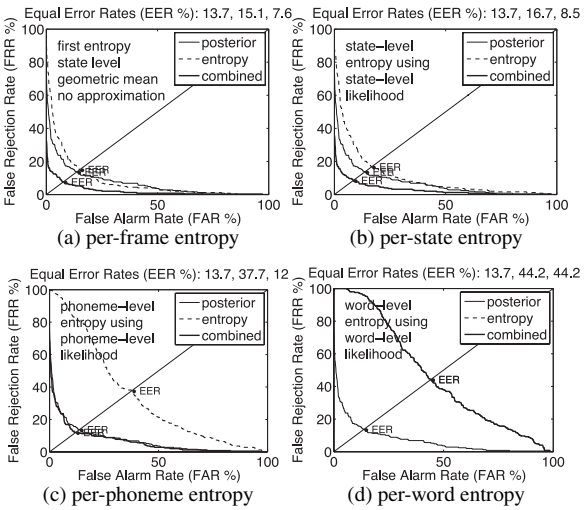


Fig. 6 DET curves showing comparison of the posterior-based entropy derived from different segments of likelihoods: per-frame entropy, per-state entropy, per-phoneme entropy, and per-word entropy. It is clear that the per-frame entropy provides the most information and highest performance, followed by the per-state entropy, with the other two kinds of entropy having almost no information, particularly the per-word entropy.

Table 4 Performance comparison of the posterior-based entropy derived from the likelihoods of different segments: per-frame entropy, per-state entropy, and per-phoneme entropy. '$+H_f$' means using per-frame entropy, '$+H_s$' means using per-state entropy, and '$+H_{ph}$' means using per-phoneme entropy. It is clear that the per-phoneme entropy results in a greater information loss than the per-state entropy.

	type	id	accuracy or rejection rate(%)				
			no cm	posterior	$+H_f$	$+H_s$	$+H_{ph}$
Test	IV	1	90.07	88.81	89.55	90.07	89.40
Test		2	94.50	93.25	93.50	94.25	94.00
Test		3	88.32	86.25	85.57	87.63	86.60
*Dev		4	91.54	90.81	90.81	90.81	90.81
Test		5	91.81	89.71	90.76	91.60	90.34
		Av	91.25	89.77	90.04	90.87	90.23
Test	OOV	1	0	33.47	60.72	38.86	30.20
*Dev		2	0	56.17	76.66	54.17	50.87
Test		3	0	46.37	64.23	46.54	41.30
		Av	0	45.34	67.20	46.52	40.79

From Table 4, it can be seen that the per-frame likelihoods provide the most information on entropy. As the segment length increases, more information is lost. The per-phoneme entropy clearly gives an inferior performance to the per-state entropy.

From the above observations, several conclusions can be drawn. First, combining entropy information with a state-level posterior-based confidence measure, the correct rejection rate of the OOV set is increased significantly (more than 15–20% in absolute terms), with the accuracy of the IV set increased slightly. Second, although the maximum approximation in the state-level posterior entropy calculation affects the performance slightly, it may be adopted to increase the speed of the system. Third, using the entropy information, the phoneme-level confidence measures are comparable to the state-level posterior-based confidence measures

and can be used for fast implementation. However, after using the entropy information, the performance of the state-level confidence measures is still much better than that of the phoneme-level confidence measures (as shown in the last two columns of Table 2 and Table 3, for the same accuracy of the IV set, the difference between the OOV rejection rate is more than 15%).

5. Conclusions

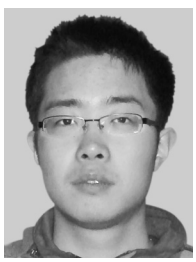
Entropy information, which can provide a measure of the uncertainty of speech, is utilized to improve the robustness of the posterior-based confidence measures. Using two normalization methods, two posterior-based entropy confidence measures are proposed, and their logarithm forms, which are particularly suitable in combination with the logarithm of posterior for practical use, are discussed. On the basis of the HMM framework, two levels of posteriors are analyzed for their calculations of entropy. The experimental results show that the entropy information contributes to typical posterior-based confidence measures, and the improvement is reasonably similar for the development and test sets. The effects of two different averaging methods and the maximum approximation in the entropy calculation are investigated, and no significant effects were observed. However, the system without the maximum approximation is believed to be more robust in combination with posterior and such combinations require more calculations such as logarithm additions. For the state-level confidence measures, the use of the likelihoods of longer segments for the entropy calculation is discussed, and a loss of information is observed as the segment length increases. Using the entropy information, the state-level confidence measures have a better performance than the phoneme-level confidence measures. However, for practical use, the phoneme-level posterior integrated with entropy information is expected to be comparable in performance to the state-level posterior. The proposed confidence measure's framework integrated with entropy is believed to be more robust for practical use, particularly when the training and test sets are mismatched.

Acknowledgements

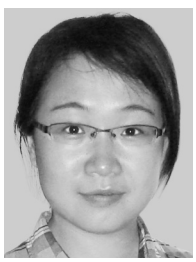
The authors would like to thank all the editors, reviewers and an anonymous professional proofreader, especially professor Akio Kobayashi, for their thorough, extensive, and very helpful comments which helped them greatly in improving the clarity, quality, and presentation of the paper.

References

[1] H. Jiang, "Confidence measures for speech recognition: A survey," Speech Commun., vol.45, no.4, pp.455–470, 2005.
[2] F. Wessel, R. Schluter, K. 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.
[3] A. Allauzen, "Error detection in confusion network," Proc. Interspeech, Aug. 2007.
[4] J. Xue and Y. Zhao, "Random forests-based confidence annotation using novel features from confusion network," Proc. IEEE International Conference on Acoustics, Speech and Signal, pp.I–I, May 2006.
[5] T.H. Chen, B. Chen, and H.M. Wang, "On using entropy information to improve posterior probability-based confidence measures," Chinese Spoken Language Processing (ICSLP), vol.4274, pp.454–463, 2006.
[6] M. Hwang and X. Huang, "Subphonetic modeling with markov states-senone," Proc. IEEE International Conference on Acoustics, Speech, and Signal, vol.1, pp.33–36, March 1992.
[7] Z. Rivlin, M. Cohen, V. Abrash, and T. Chung, "A phone-dependent confidence measure for utterance rejection," Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, pp.515–517, IEEE Computer Society, May 1996.
[8] S.M. Witt and S.J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," Speech Commun., vol.30, no.2-3, pp.95–108, Feb. 2000.
[9] F. Ge, F. Pan, C. Liu, B. Dong, S.D. Chan, X. Zhu, and Y. Yan, "An svm-based mandarin pronunciation quality assessment system," The Sixth International Symposium on Neural Networks (ISNN 2009), pp.255–265, 2009.
[10] J. Shao, J. Han, and Y. Yan, "An efficient approach of constructing search space for embedded speech recognition," NCMMSC 2005, Technical Acoustics, pp.157–160, Beijing, Oct. 2005.
[11] A. Hunt and S. McGlashan,"http://www.w3.org/tr/speech-grammar/," 2004. W3C Recommendation 16 March 2004.



Yanqing Sun received his BE in Electrical Science and Engineering from Nanjing University of Aeronautics and Astronautics (NUAA) in 2005. He is currently a PhD candidate of ThinkIT Speech Laboratory, Institute of Acoustics (IOA), Chinese Academy of Sciences (CAS). His research interests include robust speech recognition and confidence measures.



Yu Zhou received her BE from the School of Electronic Information of Wuhan University in June 2006. She is currently a PhD candidate of ThinkIT Speech Laboratory, Institute of Acoustics (IOA), Chinese Academy of Sciences (CAS). Her research interests include speech signal processing and speech emotion recognition.

Qingwei Zhao received his PhD in Electronic Engineering from Tsinghua University in 1999. He is currently an Associate Professor at ThinkIT Speech Laboratory, Institute of Acoustics (IOA), Chinese Academy of Sciences (CAS). Before joining CAS, he was with Intel as a Senior Researcher. His research interests include spontaneous speech recognition, keyword spotting and automatic pronunciation evaluation.

Pengyuan Zhang received his PhD in Information and Signal Processing from Institute of Acoustics (IOA), Chinese Academy of Sciences (CAS) in 2007. He is currently an Associate Professor at ThinkIT Speech Laboratory, IOA, CAS. His research interests include keyword spotting based on spontaneous speech, and confidence measures in complex environments. He has had over 10 papers published in international and national learning periodicals and conferences.

Fuping Pan received his PhD in Information and Signal Processing from Institute of Acoustics (IOA), Chinese Academy of Sciences (CAS) in 2007. He is currently an Associate Professor at ThinkIT Speech Laboratory, IOA, CAS. His research interests include automatic pronunciation evaluation, speech signal processing and speech recognition.

Yonghong Yan received his BE from Tsinghua University in 1990, and his PhD from Oregon Graduate Institute (OGI). He worked in OGI as an Assistant Professor (1995), Associate Professor (1998), and Associate Director (1997) of the Center for Spoken Language Understanding. He worked in Intel from 1998-2001, chaired the Human Computer Interface Research Council, and worked as Principal Engineer of Microprocessor Research Lab and Director of Intel China Research Center. He is currently a Professor and Director of ThinkIT Speech Laboratory. His research interests include speech processing and recognition, language/speaker recognition, and human-computer interfaces. He has had more than 100 papers published and holds 40 patents.