PAPER

Novel Confidence Feature Extraction Algorithm Based on Latent Topic Similarity

Wei CHEN[†a)], **Gang LIU**[†], **Jun GUO**[†], ***Nonmembers***, **Shinichiro OMACHI**[††], **Masako OMACHI**[†††], ***Members***, ***and*** **Yujing GUO**[†], ***Nonmember***

SUMMARY In speech recognition, confidence annotation adopts a single confidence feature or a combination of different features for classification. These confidence features are always extracted from decoding information. However, it is proved that about 30% of knowledge of human speech understanding is mainly derived from high-level information. Thus, how to extract a high-level confidence feature statistically independent of decoding information is worth researching in speech recognition. In this paper, a novel confidence feature extraction algorithm based on latent topic similarity is proposed. Each word topic distribution and context topic distribution in one recognition result is firstly obtained using the latent Dirichlet allocation (LDA) topic model, and then, the proposed word confidence feature is extracted by determining the similarities between these two topic distributions. The experiments show that the proposed feature increases the number of information sources of confidence features with a good information complementary effect and can effectively improve the performance of confidence annotation combined with confidence features from decoding information.

key words: *speech recognition, confidence annotation, confidence feature, latent topic similarity*

1. Introduction

Along with the rapid development of automatic speech recognition (ASR) technology, an ASR system is continuously widely applied. However, the performance of speech recognition is still far from perfect, and a large number of errors still exist in the recognition results. To evaluate the reliability of recognition results accurately, confidence annotation is applied to postprocessing after the initial speech recognition results are obtained. The confidence annotation is used to determine the reliability of the hypothesis produced by a speech recognizer and could be regarded as a pattern classification issue [1]. It classifies the units of recognition results into the following two classes: the correctly recognized class and the incorrectly recognized class based on a single confidence feature or a combination of different features. The unit of confidence annotation is usually a word and could also be a frame, a phone, a sentence, and so on [1]–[4].

Manuscript received November 4, 2009.
Manuscript revised March 11, 2010.

[†]The authors are with the Pattern Recognition and Intelligent System Laboratory, Beijing University of Posts and Telecommunications, Beijing 100876, China.
[††]The author is with the Graduate School of Engineering, Tohoku University, Sendai-shi 980–8579 Japan.
[†††]The author is with the Faculty of Science and Technology, Tohoku Bunka Gakuen University, Sendai-shi 981–8551 Japan.
a) E-mail: cwsunshine@gmail.com

DOI: 10.1587/transinf.E93.D.2243

Recently, how to extract an effective word confidence feature has always been the key issue of confidence annotation based on a word, and there have been many associated research studies [2]–[7]. Currently, word confidence features are always extracted from decoding information and can be mainly classified into two types as below.

The first type of word confidence feature is extracted from an acoustic or language model level, and mainly includes word normalized acoustic likelihood score, language model score, word duration, the number of phones per word, acoustic stability, log-likelihood ratio, and so on [1], [7]–[9]. This type of confidence feature is regarded as the most conveniently extracted feature.

The second type is based on lattice, word graph, N-best list, confusion network, and so on, which record a decoding process explicitly including each candidate path, start time, end time, probability information of each candidate word, and so on. On the basis of the obtained information, several types of feature, such as posterior probability, lattice density, N-best homogeneity score, and so on, could be extracted for each target word [7], [9]–[11]. Now, the posterior probability based on a word graph is shown to be one of the most useful single confidence features [11].

A human hearing experiment has shown that people can only clearly hear 70% of all known syllables in continuous speech and can be guided mainly by high-level information sources, such as semantics and syntax, in the case of unclear speech [12]. Since the performance of ASR has been determined by the ability of ambiguity resolution and error correction in postprocessing, the confidence feature extracted from high-level information sources that are statistically independent of decoding information becomes very important. However, it is still difficult for machines to extract this type of confidence feature effectively at the ASR postprocessing stage. In recent years, although there have been several related research studies such as those on parse quality [13], the number of semantic slots filled [14], and in-domain confidence [15], the methods used in these studies are limited to either a simple, low vocabulary spoken dialogue task or a specific domain application of a back-end system.

A latent Dirichlet allocation (LDA) [16] model has been found useful in unsupervised language model (LM) adaptation in ASR [17]–[20] and spoken language translation [21] systems. In this paper, a novel confidence feature extraction algorithm based on latent topic similarity, which

is called the latent topic similarity-based confidence feature extraction (LTS) algorithm, is proposed. LDA is used to calculate each word topic distribution and context topic distribution in one recognition result, and the proposed word confidence feature is extracted by determining the similarities between these two topic distributions. The experiments show that the proposed feature increases the number of sources of confidence features with a good information complementary effect and can effectively improve the performance of confidence annotation combined with confidence features from decoding information.

This paper is divided into five parts: firstly, background information and organization are shown; secondly, LDA is introduced; thirdly, the proposed confidence feature extraction algorithm based on latent topic similarity is emphatically shown; and then, the experimental results are given; lastly, we present our conclusion.

2. LDA Model

Recently, statistical topic models have been widely applied to many fields, such as text classification and information retrieval, and have achieved good performance [22], [23]. Given a corpus, a topic model can extract latent topics by analyzing numerous statistics so as to obtain understandable and relatively stable latent topic knowledge, which can also be regarded as short descriptions for documents in a large-scale corpus.

LDA is an unsupervised topic model that has been proposed recently [16]. It is a generative probabilistic hierarchical model with three levels including word, topic, and document. In LDA, each document is represented as random mixtures over latent topics, where each topic is characterized by a multinomial distribution over words. The topic proportions for a document are treated as a draw from the Dirichlet distribution. The words are obtained in the document by repeatedly choosing a topic assignment from those proportions and then drawing a word from the corresponding topic. In Fig. 1, the graphical model of LDA is shown.

In Fig. 1, α and β are two hyperparameters of LDA. Given a corpus D including M documents containing K topics expressed over V unique words, for each document d in the corpus, the word sequence $\overrightarrow{w}$ with the length N_d could be described as $\overrightarrow{w} = (w_1, w_2, \ldots, w_{N_d})$, where w_i is the i-th word in the sequence and assigned to the topic z_i, which is an element of the topic sequence $\overrightarrow{z} = (z_1, z_2, \ldots, z_{N_d})$ and z_i=j represents the topic assignment of w_i to the topic $j(j = 1 \ldots K)$.

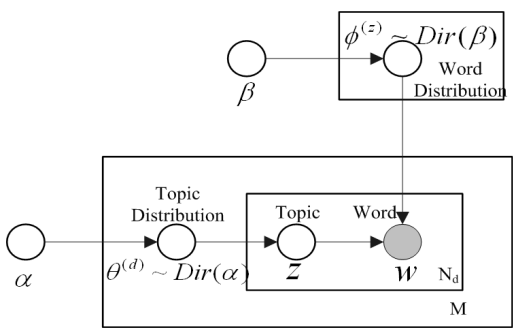


Fig. 1 LDA graphical model.

According to the notation in the previous paper [23], the two variables ϕ and θ are defined as

$$\phi_j^{(w)} = P(w|z = j), \tag{1}$$

$$\theta_j^{(d)} = P(z = j|d). \tag{2}$$

In LDA, some distributions are defined as

$$\theta \sim Dirichlet(\alpha), \tag{3}$$

$$\phi \sim Dirichlet(\beta), \tag{4}$$

$$z_i|\theta_{z_i}^{(d)} \sim Multinomial(\theta_{z_i}^{(d)}), \tag{5}$$

$$w_i|z_i, \phi \sim Multinomial(\phi_{z_i}^{(w_i)}). \tag{6}$$

Only the posterior distribution $P(\overrightarrow{z}|\overrightarrow{w})$ requires sampling; however, concerning the calculation complexity brought by computing a probability distribution over a large discrete state space [22], it is impractical to make accurate inference and parameter estimation. Hence, Gibbs sampling [24], [25] is adopted for approximation. As one of the simple classes of Markov chain Monte Carlo (MCMC) sampling methods, Gibbs sampling enables the construction of a Markov chain converging to a certain target probabilistic distribution and then draws the samples approaching that distribution from the Markov chain [22]–[24]. Moreover, the full conditional posterior distribution $P(z_i|\overrightarrow{z_{-i}}, \overrightarrow{w})$, where $\overrightarrow{z_{-i}}$ denotes all the word assignments of z_n, where $n \neq i(i,n = 1 \ldots N_d)$, is used for approximation. The sampling equation is shown as

$$P(z_i = j|\overrightarrow{z_{-i}}, \overrightarrow{w}) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(.)} + V\beta} \cdot \frac{n_{-i,j}^{(d)} + \alpha}{n_{-i,.}^{(d)} + K\alpha}, \tag{7}$$

where $n_{-i,j}^{(w_i)}$ is the number of times that the word w_i in a corpus is assigned to the topic j; $n_{-i,j}^{(.)}$ is the total number of times that words in a corpus are assigned to the topic j; $n_{-i,j}^{(d)}$ is the number of times that the topic j appears in the document d; $n_{-i,.}^{(d)}$ is the total number of times that topics appear in the document d; and $n_{-i,j}^{(w_i)}$, $n_{-i,j}^{(.)}$, $n_{-i,j}^{(d)}$, and $n_{-i,.}^{(d)}$ all exclude the assignment at the time z_i=j.

After several iterations, the topic assignments for words could be obtained, and the parameters ϕ and θ could be estimated using Eqs. (8) and (9).

$$\phi_j^{(w)} = P(w|z = j) = \frac{n_j^{(w)} + \beta}{n_j^{(.)} + V\beta} \tag{8}$$

$$\theta_j^{(d)} = P(z = j|d) = \frac{n_j^{(d)} + \alpha}{n_{\cdot}^{(d)} + K\alpha} \tag{9}$$

Here, $n_j^{(w)}$ represents the number of times that the word w is assigned to the topic j; $n_j^{(\cdot)}$ represents the total number of times that words are assigned to the topic j; $n_j^{(d)}$ represents the total number of times that words are assigned to the topic j in the document d; and $n_\cdot^{(d)}$ represents the word number in the document d.

3. Confidence Feature Extraction Algorithm Based on Latent Topic Similarity

The proposed LTS algorithm is based on the following assumptions: each document has unified topics for which each word in the document serves, and the topic distributions of the words in one document are similar.

From these assumptions, the speech recognition result of one utterance is called the document, and the unified topic distribution of the document is called the context topic distribution. In the document, namely, the recognition result, if the topic distribution of one word is closer to the context topic distribution, this word will probably be recognized more correctly. In contrast, if the topic distribution of one word is different from the context topic distribution, this word will probably be less correct. In this paper, both context and word topic distributions could be estimated by LDA, and the latent topic similarity between these two distributions is proposed as our word confidence feature.

Suppose that the corpus D includes M documents and V unique words, and K is the number of topics. Both $\phi_j^{(w)}$ and $\theta_j^{(d)}$ could be calculated using Eqs. (8) and (9), respectively. The proposed confidence feature of w_i in the document d is defined as the similarity between the topic distribution of w_i and the context topic distribution of the document d, which is shown as

$$LTS(w_i) = Sim(Topic\_dis(w_i), Topic\_dis(d)), \quad (10)$$

where
$LTS(w_i)$ is the proposed confidence feature of w_i;
$Topic\_dis(w_i)$ is the topic distribution of w_i;
$Topic\_dis(d)$ is the context topic distribution of d; and
$Sim()$ is the latent topic similarity function.
The following are three problems that should be solved:

a) how to calculate the word topic distribution by LDA;
b) how to accurately estimate the context topic distribution of one document, which reveals the unified topics of the document; and
c) how to measure the similarity between the two distributions in a) and b).

The solutions of these problems will be introduced respectively as below.

3.1 Word Topic Distribution

In Eq. (10), the K-dimensional vector $Topic\_dis(w_i)$ is the topic distribution of w_i in the document d. $\phi_j^{(w_i)} = P(w_i|z_i = j)$ and $\theta_j^{(d)} = P(z = j|d)$ could be firstly calculated by LDA inference, and the topic distribution of the word w_i could be calculated using $\phi_j^{(w_i)}$ and $\theta_j^{(d)}$ shown in Eq. (11).

$$\begin{aligned} &Topic\_dis(w_i) \\ &\quad = (H(w_i, z_i = 1), H(w_i, z_i = 2), \ldots, H(w_i, z_i = K)), \end{aligned} \quad (11)$$

where

$$\begin{aligned} H(w_i, z_i = j) &= P(z_i = j|w_i) \\ &= \frac{\phi_j^{(w_i)} * P(z_i = j)}{P(w_i)}. \end{aligned} \quad (12)$$

In Eq. (12),

$$\begin{aligned} P(w_i) &= \sum_{j=1}^{K} P(w_i, z_i = j) \\ &= \sum_{j=1}^{K} P(w_i|z_i = j) * P(z_i = j) \\ &= \sum_{j=1}^{K} \phi_j^{(w_i)} * P(z_i = j), \end{aligned} \quad (13)$$

$$\begin{aligned} P(z_i = j) &= \sum_{m=1}^{M} P(z_i = j, d_m) \\ &= \sum_{m=1}^{M} P(z_i = j|d_m) * P(d_m) \\ &= P(d) * \sum_{m=1}^{M} \theta_j^{(d_m)}. \end{aligned} \quad (14)$$

In Eq. (14), the prior probability of the document $d_m (m = 1 \ldots M)$ is assumed to obey a uniform distribution, indicating that

$$P(d_m) = P(d), m = 1 \ldots M. \quad (15)$$

3.2 Context Topic Distribution

In Eq. (10), the K-dimensional vector $Topic\_dis(d)$ is the context topic distribution of the document d, which could be calculated as

$$\begin{aligned} &Topic\_dis(d) \\ &\quad = (L(d, z = 1), L(d, z = 2), \ldots, L(d, z = K)). \end{aligned} \quad (16)$$

In the document d, topics reflected by every word should be consistent with the unified topics of the document d. Concerning the fact that the unified topics of the document d are always determined by certain words with strong topic orientations, which are called anchor words, the context topic distribution of d could be calculated using the arithmetic mean of word topic distributions of these

anchor words. Therefore, how to find anchor words in the document becomes the key issue of context topic distribution calculation. Considering that for confidence annotation, misrecognized words may occur in the recognition results, there are two requirements for anchor words. Firstly, the anchor words should have to be recognized much more correctly with high posterior probabilities. Secondly, these words should have strong topic orientations.

Suppose that there are L anchor words corresponding to the word sequence $\overrightarrow{A} = (A_1, A_2, \ldots, A_L)$ in the document d. The j-th dimension of $Topic\_dis(d)$, namely, $L(d, z = j)$, could be calculated as

$$L(d, z = j) = \frac{1}{L} * \sum_{i=1}^{L} H(A_i, z = j). \tag{17}$$

The explicit algorithm of finding anchor words in the document d is shown in algorithm 1.
algorithm 1:

a) Analyze the document d, and record the posterior probability of each word in d.
b) Set the threshold of posterior probability $PPThresh$. If the posterior probability of a word in d is larger than $PPThresh$, this word will be added to the authentic class $CClass$; In contrast, if the posterior probability of a word in d is smaller than $PPThresh$, this word will be added to the unauthentic class $MClass$.
c) Count the number of words named by $Cnum$ in the authentic class $CClass$. If $Cnum$ is equal to 0, the words in $MClass$ will all be added to $CClass$.
d) For each word w in the document d, calculate the $Topic\_dis(w)$ of the word w using Eq. (11) and record the maximum $H(w, z = j)$ of $Topic\_dis(w)$ represented by $max\_prob(w)$, where

$$max\_prob(w) = \max_{j=1\ldots K} H(w, z = j). \tag{18}$$

Here, $max\_prob(w)$ corresponds to the topic orientation intensity of the word w.
e) Set the ratio named $Aratio$ for selecting anchor words, and the number of anchor words can be written as

$$L = INT(Cnum * Aratio) + 1, \tag{19}$$

where the function INT() is the top integral function.
f) Sort the words in $CClass$ by $max\_prob(w)$ in step d) by descending order and select L words as anchor words of document d.

After finding the anchor words, $L(d, z = j)$ could be calculated using Eq. (17), and $Topic\_dis(d)$ could then be obtained.

3.3 Latent Topic Similarity

Kullback-Leibler (K-L) divergence, which is a meaningful statistical measure, is widely used to measure the difference between different probabilistic distributions [26]. In this paper, K-L divergence is adopted to measure the latent topic similarity.

The topic distribution $Topic\_dis(w)$ of the word w in d and the context topic distribution $Topic\_dis(d)$ of the document d are respectively represented by $M1$ and $M2$ as,

$M1$:$Topic\_dis(w)$; $M2$:$Topic\_dis(d)$.

The K-L divergence between $M1$ and $M2$ is calculated as,

$$D_{KL}(M1||M2) = \sum_{j=1}^{K} H(w, z = j) * \log(\frac{H(w, z = j)}{L(d, z = j)}). \tag{20}$$

To eliminate the use of the reference model, the symmetric K-L divergence is defined as

$$Sym\_KL = \frac{1}{2}\{D_{KL}(M1||M2) + D_{KL}(M2||M1)\}. \tag{21}$$

The latent topic similarity between the word and context topic distributions is calculated using Eq. (21), and is proposed as our confidence feature for confidence annotation.

4. Experiments

4.1 Experimental Data

Our continuous mandarin speech recognizer is trained on the National High-Tech Research and Development Plan of China (863) speech recognition database, which is the read-speech database of newspaper under the silent condition [27]. The data is coded into 25 ms frames with a frame shift of 10 ms and parameterized to 39-D vectors consisting of 12 MFCCs and normalized log energy + delta + acceleration coefficients. Context-dependent initial and final models are used as basic units for acoustic hidden Markov models (HMMs). The specifications of the recognizer are as follows:

1) The speaker-independent female training set of the 863 speech recognition database is used for acoustic modeling (30 hours, 41 female speakers, 520 utterances per speaker);
2) 24718 physical cross-word triphones are created, each triphone is composed of 5 emitting states, each of which is modeled by a single Gaussian mixture, and 83740 states are generated by tree clustering;
3) The vocabulary size is 20196;
4) A bigram language model is trained on 6M words from People Daily Corpus, and its perplexity is 102.

The HMM training/decoding software used for all experiments is HTK v3.4 [28].

The LDA training set includes 311029 sentences from People Daily Corpus. After removing certain meaningless words, such as empty words, adverbs, and so on, and

Table 1 Summary of training, test, and development sets.

	Training Set	Test Set	Development Set
#Speakers	22	11	9
#sentences	11957	6140	4683
#words	67935	35235	27805
#WER	10.35%	8.89%	9.29%

deleting the stopped and low-frequency words whose occurrences are lower than 3, the training set has 94641 word terms and 6168185 words. Phan's GibbsLDA++ [29] is used for LDA parameter estimation and inference. Keeping the empirical values $\alpha = 0.5$ and $\beta = 0.1$, LDA parameter estimation iterates 1000 times, and inference iterates 30 times. By comparison, the topic number K=50 is used.

4.2 Experimental System

Considering that a support vector machine (SVM) is commonly used in many classification problems and considered to be an accurate and robust confidence classifier in confidence annotation [9], [13], [30], we adopt a SVM as our confidence classifier. Utterances of the female test set of the 863 speech recognition database, which are 30 hours in length and consist of 42 female speakers, each of which speaks about 520 utterances and is independent of the speakers of the female training set, are recognized, and the recognition results are processed by removing stopped words and used in the confidence annotation database including 22780 sentences and 130975 words. Three subsets of the database are separately used as training, test, and development sets. Their speakers and sentences are mutually exclusive, and word error rates (WERs) are recorded. We summarize the information in Table 1.

The process of confidence annotation in our experiments includes the following four parts: initialization, SVM training, SVM recognizing, and annotation performance evaluation, which are described below.

1. Initialization:
   By comparing with the references, each word in the recognition output from each sentence could be annotated into the following two classes: the correctly recognized word class and the incorrectly recognized word class.
2. SVM training:
   a) Different confidence features or a combination of different features of each word in the training set are extracted.
   b) A two-class SVM is trained to classify the correctly recognized word class represented by C^+ and the incorrectly recognized word class represented by C^-.
3. SVM recognizing:
   a) Different confidence features or a combination of different features of each word in the test set are extracted.
   b) Each word in the test set is classified by the trained two-class SVM in step 2 to determine whether the word is correctly recognized (C^+) or not (C^-).
4. Annotation performance evaluation:
   Comparing the classification result of each word in the test set in step 3 with its corresponding label annotated in step 1, annotation error rate (AER) [3], false acceptance rate (FAR) [4], [9], and false rejection rate (FRR) [4], [9] are calculated using Eqs. (22), (23), and (24), respectively.

$$AER = \frac{Number\_WR}{Total\ number\ of\ words} \tag{22}$$

$$FAR = \frac{Number\_FA}{Total\ number\ of\ words} \tag{23}$$

$$FRR = \frac{Number\_FR}{Total\ number\ of\ words} \tag{24}$$

Here,
$Number\_WR$ is the number of incorrectly annotated words;
$Number\_FA$ is the number of incorrectly recognized words, which are classified as C^+;
$Number\_FR$ is the number of correctly recognized words, which are classified as C^-.

In our experiments, the radial basis function (rbf)

$$k(x, x_i) = \exp(-\gamma|x - x_i|^2) \tag{25}$$

is adopted as the kernel function of the SVM, where γ is set to 0.5 in our experiment.

Different types of confidence feature are listed in Table 2.

Table 2 Different types of confidence feature.

Feature	Meaning
PP	Word posterior probability [7], [11] $Mean_{pp} = 0.95, \sigma_{pp} = 0.15$
LLD	Word-frame-based acoustic likelihood score [1], [8], [9]
Lg	Word language model score [1], [8], [9]
LTS	Proposed word confidence feature based on latent topic similarity

In Table 2, the PP is calculated using the confusion network [10]. By analyzing the PPs of all the words in the confidence annotation database, the mean PP represented by $Mean_{pp}$ is equal to 0.95, and the standard deviation of the PP represented by σ_{pp} is equal to 0.15. LLD and Lg could be obtained during decoding. For the word w, LLD is the acoustic score per frame, and Lg is the score $P(w|v)$ of the bigram language model for w and its context v.

4.3 Parameter Selection of Algorithm 1

In this paper, the context topic distribution obtained by anchor words is necessary for the proposed word confidence feature LTS. In algorithm 1, how to find anchor words in one

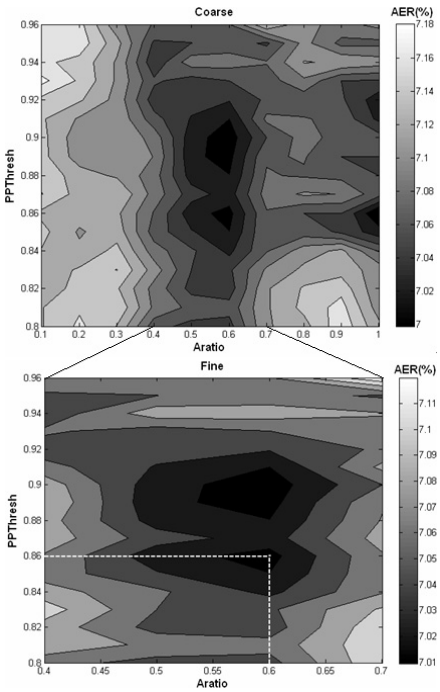


Fig. 2 Parameter selection using development set.

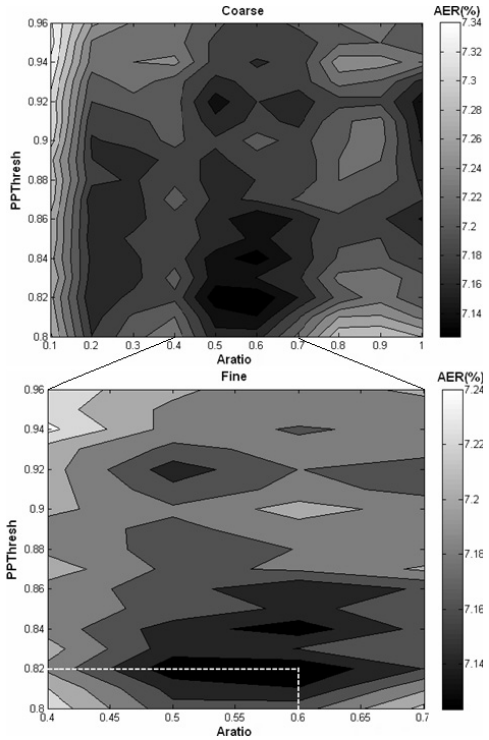


Fig. 3 Parameter selection using test set.

document is shown. Those words whose posterior probabilities are larger than $PPThresh$ are firstly added to the authentic class to guarantee the correctness of anchor word recognition, and then, the first $Aratio$ of the words with stronger topic orientations in the authentic class are selected as anchor words to calculate the context topic distribution. Since anchor words markedly affect the context topic distribution, the $PPThresh$ and $Aratio$ of algorithm 1 should be selected in advance.

Because LTS has the advantage of a good information complementary effect, and confidence annotation using LTS combined with confidence features extracted from decoding information could achieve higher performance characteristics than that using only LTS, the combination of LTS and PP is adopted as a confidence feature in these experiments. Annotation error rates (AERs) of different confidence annotation systems using different $PPThresh$ and $Aratio$ values are computed separately, and the corresponding optimal $PPThresh$ and $Aratio$ values are determined from the development set by the full grid search of the AER surface, where $Aratio$ ranges from 0.1 to 1 in increments of 0.1 and $PPThresh$ varies from 0.80 corresponding to $Mean_{pp} - \sigma_{pp}$ to 0.96 around $Mean_{pp}$ in increments of 0.01. These optimized parameters selected in the development set are then used in the test set.

The AER contours at various $PPThresh$ and $Aratio$ values are shown in Fig. 2 for the development set. The coarse-scale plot shows the AER contours in the full parameter range. Fine-scale contours of lower AER regions are shown in a small range.

In Fig. 2, the lateral axis represents $Aratio$, and the vertical axis represents $PPThresh$. The darkness represents the AER of specific $PPThresh$ and $Aratio$. In the coarse scale plot, it can be seen clearly that the low-AER region mainly lies in the area where $Aratio$ ranges from 0.4 to 0.7. Hence, a fine-scale contour of a lower AER region is shown in a smaller range of $Aratio$, and 0.86 and 0.6 are determined as the values of $PPThresh$ and $Aratio$ for the lowest AER, respectively. In Fig. 3, parameter selection using the test set is shown.

In Fig. 3, using a similar parameter selection method, $PPThresh$ and $Aratio$ could be determined from the test set by determining the optimal performance, which is the upper bound obtained by closed test set tuning. Here, 0.82 and 0.6 are respectively determined as the values of $PPThresh$ and $Aratio$. These parameters could be regarded as optimal references for the parameters determined from the development set.

In Fig. 4, to verify whether $PPThresh$ is necessary in algorithm 1, the performance characteristics of two confidence annotation systems, where $PPThresh$ is set to 0.86 and $PPThresh$ is set to 0, namely, $PPThresh$ is ignored, are compared using the test set. $Aratio$ varies from 0.1 to 1 in increments of 0.1.

In Fig. 4, the lateral axis represents $Aratio$ ranging from 0.1 to 1, and the vertical axis represents AER. It can be observed that the performance characteristics of the confidence annotation system using $PPThresh$ are much higher than those of the system not using $PPThresh$.

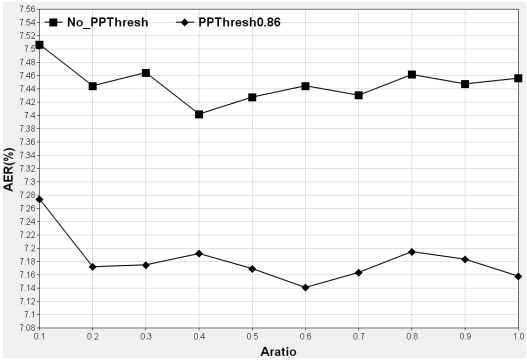


Fig. 4 Necessity verification of *PPThresh*.

Table 3 Annotation accuracy comparison.

Feature	Optimal (%)	AER (%)	FAR (%)	FRR (%)
Accept All	8.89	8.89	8.89	0.00
LTS	8.89	8.89	8.89	0.00
PP	7.53	7.53	4.60	2.93
PP+LTS	7.12	7.14	4.72	2.42
PP+LLD	7.58	7.58	5.16	2.42
PP+LLD+LTS	7.13	7.16	4.78	2.38
PP+Lg	6.85	6.85	4.63	2.23
PP+Lg+LTS	6.51	6.43	4.57	1.86
PP+LLD+Lg	6.85	6.85	4.76	2.09
PP+LLD+Lg+LTS	6.54	6.48	4.62	1.86

4.4 Comparison Using Different Confidence Features

In our experiments, a baseline system for the test set is used for performance comparison. It is obtained by accepting all recognition words, achieving an AER of 8.89%.

In Table 3, the performance characteristics of several confidence annotation systems using different confidence features or a combination of different features in the test set are compared. The column Optimal shows the performance characteristics of confidence annotation systems using the *PPThresh* 0.82 and *Aratio* 0.6 determined from the test set, which are the upper bounds obtained by closed test set tuning. The columns AER, FAR, and FRR show the performances characteristics of confidence annotation systems using the *PPThresh* 0.86 and *Aratio* 0.6 determined from the development set.

In Table 3, the performance of the confidence annotation system using LTS is not improved compared with the baseline system. The LTS feature has a disadvantage in that it cannot be used alone for confidence annotation. However, confidence annotation systems using LTS combined with confidence features extracted from decoding information could achieve much higher performance characteristics with lower AER, FAR, and FRR than those not using LTS, which proves that LTS has the advantage of a good information complementary effect. In addition, the performance characteristics obtained using LTS in the column AER are similar to those in the column Optimal.

By using PP as a confidence feature, the relative AER reduction is determined to be 15.3% compared with that of the baseline system, and when a combination of PP and LTS (PP+LTS) is used, AER relatively decreases by 5.2% compared with that of the system using PP and by 19.9% compared with that of the baseline system. Because a large information redundancy exists between PP and LLD, the performance of the system using PP+LLD is worse than that of the system using PP, and the relative AER reduction of the system using LTS combined with PP and LLD is 5.9% compared with that of the system using PP+LLD.

In Table 3, the AER of the system using PP+Lg is higher than those of the systems using PP+LLD and PP+LTS, and the AER of the system using PP+Lg relatively decreases by 9.04% compared with that of the system using PP. There are two reasons that explain why the performance characteristics of the confidence annotation system using PP+Lg are higher than that of the system using PP+LTS:

a) LDA has a stronger capability of processing long documents than short documents. However, the documents, which are speech recognition results for confidence annotation in our experiments, are always short documents.
b) The language model used in our experiments is bigram which has a higher precision of text modeling than LDA.

Because the information sources of LDA and the language model are relatively independent and show a small information redundancy, the confidence annotation system using PP+Lg+LTS shows the best performance among all related experimental systems. The associated AER relatively decreases by 6.17% compared with that of the system using PP+Lg, 14.66% compared with that of the system using PP, and 27.68% compared with that of the baseline system. The AER of the confidence annotation system using PP+LLD+Lg relatively decreases by 22.89% compared with that of the baseline system. If our proposed feature LTS is combined with PP+LLD+Lg, the AER of the confidence annotation system using PP+LLD+Lg+LTS relatively decreases by 5.42% compared with that of the system using PP+LLD+Lg and 27.08% compared with that of the baseline system.

Our proposed confidence feature LTS is proved to be effective by these experiments. LTS increases the number of information sources of confidence features with a good information complementary effect and can effectively improve the performance of confidence annotation combined with the confidence features from decoding information.

5. Conclusions

To describe and analyze speech recognition results more accurately and improve the performance of confidence annotation more effectively, numerous confidence features were proposed; however, how to extract high-performance confi-

dence features and make full use of high-level information sources such as semantics and syntax remains a problem. Our proposed feature in this paper has the advantage of a good information complementary effect and can effectively improve the performance of confidence annotation combined with the confidence features from decoding information. Compared with that of the baseline system, the maximum relative reduction in AER 27.68% could be achieved when LTS is combined with PP and Lg, and AER relatively decreases by 6.17% compared with that of the confidence annotation system using PP+Lg and 14.66% compared with that of the system using PP. Concerning the fact that the LTS feature has a disadvantage in that it cannot be used alone for confidence annotation, how to improve the algorithm to overcome this shortcoming will be our future work.

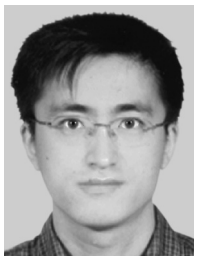
Acknowledgments

This study is supported by the National Natural Science Foundation of China (60705019), the National High-Tech Research and Development Plan of China (2006AA010102 and 2007AA01Z417), the NOKIA project, and the 111 Project of China under Grant No. B08004.

References

[1] L. Chase, Error-Responsive Feedback Mechanisms for Speech Recognition, Ph.D. Thesis, Carnegie Mellon University, April 1997.
[2] D. Bansal and M.K. Ravishankar, "New features for confidence annotation," Proc. ICSLP-98, vol.6, pp.2391–2394, 1998.
[3] R. Zhang and A. Rudnicky, "Word level confidence annotation using combinations of features," Proc. 7th European Conference on Speech Communication and Technology, pp.2105–2108, 2001.
[4] B. Mison and R. Gopinath, "Robust confidence annotation and rejection for continuous speech recognition," Proc. IEEE ICASSP, vol.1, pp.389–392, 2001.
[5] R. San-Segundo, B. Pellom, and W. Ward, "Confidence measure for dialogue management in the CU communication system," Proc. IEEE ICASSP, vol.4, pp.697–700, 2000.
[6] 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.
[7] H. Jiang, "Confidence measures for speech recognition: A survey," Speech Commun., vol.45, pp.455–470, 2005.
[8] T. Schaff and T. Kemp, "Confidence measure for spontaneous speech recognition," Proc. IEEE ICASSP, vol.2, pp.887–890, 1997.
[9] A. Kobayashi, K. Onoe, S. Homma, S. Sato, and T. Imai, "Word error rate minimization using an integrated confidence measure," IEICE Trans. Inf. & Syst., vol.E90-D, no.5, pp.835–843, May 2007.
[10] L. Mangu, E. Brill, and A. Stolcke, "Finding consensus in speech recognition: Word error minimization and other applications of confusion networks," Comput., Speech Lang., vol.14, no.4, pp.373–400, 2000.
[11] R.C. Rose, B.H. Juang, and C.H. Lee, "A training procedure for verifying string hypotheses in continuous speech recognition," Proc. IEEE ICASSP, pp.281–284, 1995.
[12] Z.Y. Huang, HNC (Hierarchical Network Concept) Theory, Tsinghua University Press, Beijing, 1998.
[13] C. Pao, P. Schmid, and J. Glass, "Confidence scoring for speech understanding systems," Proc. ICSLP, pp.815–818, 1998.
[14] D. Guillevic, S. Gandrabur, and Y. Normandin, "Robust semantic confidence scoring," Proc. ICSLP, pp.853–856, 2002.
[15] I. Lane and T. Kawahara, "Verification of speech recognition results incorporating in-domain confidence and discourse coherence measures," IEICE Trans. Inf. & Syst., vol.E89-D, no.3, pp.931–938, March 2006.
[16] D. Blei, A. Ng, and M. Jordan, "Latent dirichlet allocation," J. Machine Learning Research, vol.3, pp.993–1022, Jan. 2003.
[17] Y.C. Tam and T. Schultz, "Language model adaptation using variational bayes inference," Proc. Interspeech, pp.5–8, 2005.
[18] A. Heidel, H.A. Chang, and L.S. Lee, "Language model adaptation using latent dirichlet allocation for topic inference," Proc. Interspeech, pp.2361–2364, 2007.
[19] Y.C. Tam and T. Shultz, "Correlated latent semantic model for unsupervised LM adaptation," Proc. IEEE ICASSP, pp.41–44, 2007.
[20] Y.C. Tam and T. Schultz. "Correlated bigram LSA for unsupervised language model adaptation," Proc. Neural Information Processing Systems (NIPS), pp.1633–1640, 2008.
[21] Y.-C. Tam and T. Schultz. "Bilingual LSA-based translation lexicon adaptation for spoken language translation," Proc. Interspeech, pp.2461–2464, 2007.
[22] T. Griffiths and M. Steyvers, "Finding scientific topics," Proc. Nat. Acad. Sci., vol.101, pp.5228–5235, 2004.
[23] M. Steyvers and T. Griffiths, "Probabilistic topic models," in Latent Semantic Analysis: A Road to Meaning, ed., T. Landauer, D. Mcnamara, S. Dennis, and W. Kintsch, Laurence Erlbaum, 2005.
[24] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the bayesian restoration of images," IEEE Trans. Pattern Anal. Mach. Intell., vol.PAMI-6, no.6, pp.721–741, 1984.
[25] G. Heinrich, "Parameter estimation for text analysis," Technical Report, University of Leipzig, 2004.
[26] T.M. Cover and J.A. Thomas, Elements of information theory, Wiley, New York, 1991.
[27] Z. Yiqing, "Text design for continuous speech database of standard chinese," Chin. J. of Acoustics, vol.18, no.1, 1999.
[28] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X.A. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, The HTK Book (for HTK Version 3.4), Cambridge University Engineering Department, 2006.
http://htk.eng.cam.ac.uk/docs/docs.shtml
[29] X.-H. Phan, http://gibbslda.sourceforge.net/
[30] I.R. Lane, T. Kawahara, T. Matsui, and S. Nakamura, "Out-of domain detection based on confidence measures from multiple topic classification," Proc. IEEE ICASSP, vol.1, pp.757–760, 2004.

Wei Chen received his B.E. degree in Automation from Beijing University of Posts and Telecommunications (BUPT) in 2006, and he is currently a 2nd year Ph.D. student of the Pattern Recognition and Intelligent System (PRIS) Laboratory in BUPT, where he is engaged in speech recognition and machine learning research.



Gang Liu received his B.E. degree in 1997 and M.E. and Ph.D. degrees in Signal and Information Processing from BUPT in 1999 and 2002, respectively. He is currently an associate professor of the School of Information and Communication Engineering (SICE), BUPT. His research interests include pattern recognition, speech signal processing, and audio information retrieval.

Jun Guo received his B.E. and M.E. degrees from BUPT in 1982 and 1985 respectively, and his Ph.D. degree from Tohuku-Gakuin University, Japan in 1993. He is currently a professor and the dean of SICE in BUPT. His research interests include pattern recognition theory and application, information retrieval, content-based information security, and network management. He has published over 200 papers, some of which were published in popular international journals or conferences including Science, IEEE Trans. PAMI, IEICE, ICPR, ICCV, SIGIR, and so on.

Shinichiro Omachi received his B.E., M.E., and Doctor of Engineering degrees in Information Engineering from Tohoku University, Japan, in 1988, 1990, and 1993, respectively. He worked as a research associate at the Education Center for Information Processing at Tohoku University from 1993 to 1996. Since 1996, he has been with the Graduate School of Engineering at Tohoku University, where he is currently a professor. He was a visiting associate professor at Brown University, Providence, RI, from 2000 to 2001. His research interests include pattern recognition, computer vision, image processing, image retrieval, and data mining. He received the MIRU Nagao Award and IAPR/ICDAR Best Paper Award in 2007. Dr. Omachi is a member of the IEEE, the Information Processing Society of Japan, and the Japanese Society of Artificial Intelligence.

Masako Omachi received her B.E., Master of Information Sciences, and Doctor of Engineering degrees from Tohoku University, Japan, in 1994, 1996, and 1999, respectively. She worked as a research associate at Tohoku Bunka Gakuen University from 1999 to 2004 and as a lecturer from 2004 to 2008. She is now an associate professor of the Faculty of Science and Technology, Tohoku Bunka Gakuen University. Her research interests include image processing, image retrieval, and image recognition. She received the MIRU Nagao Award in 2007.

Yujing Guo received her B.E. degree in Communication Engineering from Beijing Information Technology Institute (BITI) in 2007, and she is currently pursuing her M.E. degree in PRIS from BUPT, where she is engaged in speech recognition research.