IEICE TRANS. INF. & SYST., VOL.E97–D, NO.3 MARCH 2014

PAPER

Confidence Measure Based on Context Consistency Using Word Occurrence Probability and Topic Adaptation for Spoken Term Detection

Haiyang LI[†], *Student Member*, **Tieran ZHENG**[†], **Guibin ZHENG**[†], *and* **Jiqing HAN**[†a)], *Nonmembers*

SUMMARY In this paper, we propose a novel confidence measure to improve the performance of spoken term detection (STD). The proposed confidence measure is based on the context consistency between a hypothesized word and its context in a word lattice. The main contribution of this paper is to compute the context consistency by considering the uncertainty in the results of speech recognition and the effect of topic. To measure the uncertainty of the context, we employ the word occurrence probability, which is obtained through combining the overlapping hypotheses in a word posterior lattice. To handle the effect of topic, we propose a method of topic adaptation. The adaptation method firstly classifies the spoken document according to the topics and then computes the context consistency of the hypothesized word with the topic-specific measure of semantic similarity. Additionally, we apply the topic-specific measure of semantic similarity by two means, and they are performed respectively with the information of the top-1 topic and the mixture of all topics according to topic classification. The experiments conducted on the Hub-4NE Mandarin database show that both the occurrence probability of context word and the topic adaptation are effective for the confidence measure of STD. The proposed confidence measure performs better compared with the one ignoring the uncertainty of the context or the one using a non-topic method.

key words: *spoken term detection, confidence measure, context consistency, sematic similarity, topic adaptation*

1. Introduction

Spoken term detection (STD) is the task which aims to locate all occurrences of terms queried by a user in large audio archives [1], and it plays an important role in accessing relevant information from spoken documents. A typical STD system can detect a term using two steps. In the first step, a speech recognizer transforms speech signals into transcriptions or lattices. In the second step, a spotter searches all potential hypotheses of the user-defined term in the results of the first step, and further verifies those hypotheses.

In STD, a confidence measure is applied to indicate the reliability of hypotheses, and it is crucial to reject false alarms. It is expected that the confidence measure can assign high confidence to a correct hypothesis and low confidence to a false alarm in a consistent way.

In the last decade, confidence measures based on word context have been widely investigated and proved helpful for speech recognition and STD [2], [3], [5], [6]. The context of a hypothesized word is the set of other hypothesized words in the recognition result of the same utterance [2], [3], [5]. These confidence measures are approached with the idea that a hypothesized word is likely to be a false alarm when it appears to be inconsistent with its context. The context consistency is employed to measure the consistency between a hypothesized word and its context.

The context consistency can be calculated with the measure of semantic similarity between two words, and this context consistency is effective as a high-level confidence measure [2], [3], [6]. The measure of semantic similarity extracts the co-occurrence relationship between words in an utterance over a longer range than the traditional N-gram, which only incorporates semantic coherence in a short range. The measure of semantic similarity can be derived from latent semantic analysis (LSA) [2] or pointwise mutual information (PMI) [3], [6]. For a hypothesized word, each word in the context is called a context word, and the context consistency is formulated as the mean of the semantic similarity measures between the hypothesized word and its context words. In those approaches, it is often assumed that the occurrences of the context words are certain in the recognized result. However, the assumption is not true, since the occurrence of a word in the recognized result is uncertain [7]. Therefore, it is necessary to consider the uncertainty of the context. In [5], context feature vectors of the hypothesized query words are used to calculate the context consistency by support vector machine (SVM) and cosine similarity, with the consideration of the uncertainty of the context. However, this method needs a large amount of detailed speech corpus, including pseudo relevant and irrelevant spoken segments, to select feature vectors and train models for each word in the vocabulary.

To overcome the shortcomings of current methods, we explore two approaches to improve the context consistency based on the measure of semantic similarity. First, we take into account the uncertainty of the context for the context consistency. Second, we also consider the topic information to make the semantic similarity measured more accurate by using text data and speech data with simple labels.

Topic information has been utilized to improve speech recognition as a type of high-level knowledge source. For example, topic consistency is employed as a confidence measure for utterance verification [8], which is a measure of topic match between the input utterance and the application domain from the confidence vector of topic classifica-

Manuscript received July 12, 2013.
Manuscript revised October 15, 2013.

[†]The authors are with the School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China.
a) E-mail: jqhan@hit.edu.cn

DOI: 10.1587/transinf.E97.D.554

tion. Another example is topic adaptation for the language model of speech recognition [9], [10]. A direct but effective method of adaptation is applied by classifying the topic of the target speech and using the language model which is trained adaptively toward that topic [11], [12]. Therefore, topic classification of target speech is needed for both topic consistency and adaptation. Topic classification has been investigated extensively. The approaches as applied to bag-of-words document representations [13] have been successfully ported for topic classification of spoken documents, including naive Bayes [14], [15], SVM [16], [17], and linear classifier [18]. It is obvious that the measure of semantic similarity is also affected by the topic or domain of the document. The measure of semantic similarity between words will change when the topic varies. Therefore, it is necessary to apply the knowledge of the current topic to obtain an appropriate measure of sematic similarity toward that topic.

In this paper, we propose a novel confidence measure based on context consistency using semantic similarity and lattice for STD. The context consistency is computed by considering the uncertainty of the context and the effect of topic. To estimate the uncertainty of context, we calculate the word occurrence probability, which is obtained by combining the overlapping hypotheses in the word posterior lattice. To make use of the topic information, we also propose the topic adaptation for context consistency. The topic adaptation is performed using a direct method, which firstly classifies the spoken document by topic and then computes the context consistency of the hypothesized word with the topic-specific measure of semantic similarity. This paper focuses on the spoken documents as the thematically coherent segments of the speech. These segments may come from thematically segmented multimedia streams [19], [20] or from shorter multimedia documents dealing with one single topic, such as a piece of news. We finally confirm the effectiveness of the proposed confidence measure by further experiments for in-vocabulary term detection on the Hub-4NE Mandarin database.

The proposed confidence measure is derived from the methods based on context consistency [2], [3], [6], and these methods ignore the uncertainty of context. The main contribution of this paper is to compute the confidence measure of context consistency by considering the uncertainty of context and the effect of topic. The work of this paper only needs text data and speech data without transcripts to estimate the measure of semantic similarity, and both types of data are labeled with only topic classes. Compared with the method of context feature vectors [5], the proposed method takes much less effort and time for data labeling. Meanwhile, we regard the other hypothesized words in the utterance as the context of a hypothesized word. Therefore, the proposed method can extract more useful context information from a wider range than the method of contextual verification model [4], which only considers the adjacent words as the context.

2. Word Occurrence Probability Based on Word Posterior Lattice

2.1 Hypotheses from Word Lattice

As a typical representation of speech recognition result, lattice has an advantage over the 1-best result for STD, since lattice can provide much more useful information and produce better recall rates [21]. In this section, we first characterize word lattice, and then describe the representation of word hypothesis based on lattice.

A lattice is a directed acyclic graph used to keep the information about active hypothesis paths during decoding of speech recognition [21]. A word lattice consists of a set of directed arcs and a set of nodes. Arcs represent word hypotheses, while nodes represent relationships among hypotheses. For an original lattice produced by the speech recognizer, the word likelihood is kept for each hypothesis. A word posterior lattice can be generated by a forward-backward algorithm from original lattice [22], and the posterior probability is saved for each word hypothesis. Given the observation of an utterance O, the corresponding word posterior lattice is denoted as L. A set of word hypotheses H can be extracted from L. Each element of the set H is a word hypothesis expressed as $h = (ts[h], te[h], wd[h], pp[h])$, with $ts[h]$ being the start time, $te[h]$ the end time, and $wd[h]$ the identity of the hypothesized word. $pp[h]$ can also be represented as $P(h|L)$, which is the posterior probability for hypothesis h in lattice L.

2.2 Word Occurrence Probability Based on Grouping of Hypotheses

A problem with lattice is the overlap among hypotheses of the same word. In STD, the overlapping hypotheses in the same time span are usually combined into a single detection for the final verification, and this type of combination can improve the performance [23].

For a convenient description, we define the hypotheses by a novel representation with the consideration of the overlaps. A maximum group of overlapping hypotheses for the same hypothesized word is defined as an overlapping group (or a cluster [23]), and each group is obtained by clustering the overlapping hypotheses. Thus, a set of hypotheses H for an utterance O is divided into M groups $G_1, \ldots, G_M$. Each group $G_i = \left\{g_i^1, \ldots, g_i^{N_i}\right\}$ is composed of the overlapping hypotheses of the same hypothesized word w_i, where N_i is the element number of G_i. $H = \bigcup_{i=1}^{M} G_i$. For any integer i and j in $\left\{1, \ldots, M\right\}$, if $i \neq j$, then $G_i \cap G_j = \varnothing$. So any hypothesis h in H can be represented as an element g_i^k ($1 \leq i \leq M$, $1 \leq k \leq N_i$) in a group G_i. The time span $\tau_i = (ms_i, me_i)$ for each group G_i is determined, where ms_i is the minimal start time of the hypotheses in G_i, and me_i is the maximal end time. An example of an overlapping group of hypotheses is shown in Fig. 1 for a lattice. The hypotheses for word w_3 in

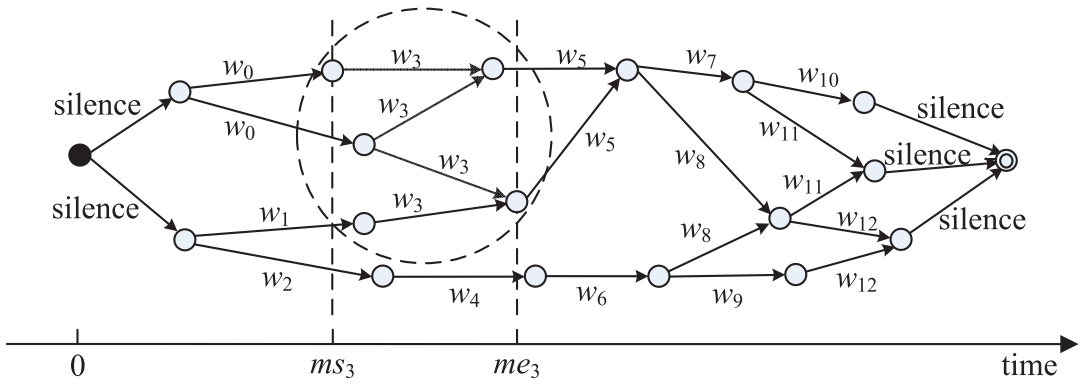

Fig. 1 An example of a hypothesis group in a lattice.

the circle comprise a hypothesis group, since the hypotheses overlap each other in the time duration.

The word occurrence probability of w_i in τ_i is computed with exclusive accumulated confidence [23] in G_i as:

$$P(w_i, \tau_i|O) = 1 - \prod_{\substack{\forall t_1,t_2: \\ ms_i \le t_1 < t_2 \le me_i}} (1 - P(w_i, t_1, t_2|O))$$

$$= 1 - \prod_{\substack{\forall t_1,t_2: \\ ms_i \le t_1 < t_2 \le me_i}} \left(1 - \sum_{\substack{\forall h: h \in G_i \\ \wedge ts[h] = t_1 \\ \wedge te[h] = t_2}} P(h|L) \right) \quad (1)$$

According to Eq. (1), all strict overlaps (with the same starting and ending time) are combined with Bayesian approach, and all non-strict overlaps are combined with evidence approach. $P(w_i, \tau_i|O)$ is also regarded as the final confidence measure for each hypothesis in group G_i.

On the level of overlapping group, the word occurrence probability can also be defined. Firstly, the occurrence probability of word w in an utterance O can be computed with the evidence from a hypothesis group G_i:

$$P(w|G_i, O) = \begin{cases} P(w_i, \tau_i|O) & if \quad w = w_i \\ 0 & if \quad w \neq w_i \end{cases} \quad (2)$$

The word occurrence probability can also be calculated with evidence from multiple groups. A hypothesis set S is an union of several hypothesis groups. The word occurrence probability $P(w|S, O)$ of w with respect to S is calculated as:

$$P(w|S, O) = 1 - \prod_{\substack{\forall i: i = 1,\ldots,M \\ \wedge G_i \subset S}} (1 - P(w|G_i, O)) \quad (3)$$

$P(w|S, O)$ is also computed with the exclusive evidence, since w may appear more than once in several groups of S.

3. Computation of Context Consistency Using Word Occurrence Probability

The context consistency is usually computed as the average semantic similarity measure for confidence measure [2], [3]. However, this context consistency disregards the uncertainty of the context. In this section, we propose a novel framework to calculate the context consistency by using word occurrence probability of context words to consider the uncertainty of the context.

For a hypothesized word w_i, the set of its context words is defined as $B(w_i)$, which is obtained by removing the duplicate words and the common function words from $\{w_j | j = 1, \ldots, M \wedge j \neq i\}$. A stop word list is used to discard the common function words, which usually reoccur in the sentences. The context consistency of w_i can be computed by using the occurrence probability of its context words as:

$$CC(w_i) = \frac{1}{|B(w_i)|} \sum_{u \in B(w_i)} (SS(w_i, u) \cdot P(u|H - G_i, O)) \quad (4)$$

where $SS(w_i, u)$ is the measure of semantic similarity between word w_i and word u. Notice that word w_i may be the same as word u in Eq. (4), for a word can appear in the context of the same word. In this work, four measures of semantic similarity $SS(w_i, u)$ are employed between words w_i and u, including latent semantic analysis (LSA) [2], pointwise mutual information (PMI) [3], [6], normalized pointwise mutual information (NPMI) [24], and generalized latent semantic analysis (GLSA) [25]. Generally, the measure of semantic similarity is estimated with the whole text data, and the effect of topic is ignored. The topic adaptation is studied for the measure of semantic similarity in Sect. 4.

In Eq. (4), $H - G_i$ represents the hypothesis set for the context of w_i, and $H - G_i$ is composed of all hypotheses in H but not in G_i. $P(u|H - G_i, O)$ denotes the word occurrence probability that u occurs in the context of w_i, which is calculated according to Eq. (3) as:

$$P(u|H - G_i, O) = P\left(u \,|\, \bigcup_{\substack{\forall k: \\ k=1,\ldots,M \\ \wedge k \neq i}} G_k, O \right)$$

$$= 1 - \prod_{\substack{\forall k: k=1,\ldots,M \\ \wedge k \neq i}} (1 - P(u|G_k, O))$$

$$= 1 - \prod_{\substack{\forall k: k=1,\ldots,M \\ \wedge k \neq i \\ \wedge w_k = u}} (1 - P(w_k, \tau_k|O)) \quad (5)$$

As shown by Eq. (4), this context consistency not only incorporates the measure of semantic similarity between the considering hypothesized word and its context word, but also takes account of the uncertainty of the context word. The uncertainty of a context word is represented by the word occurrence probability. A context word with higher probability contributes more to the context consistency. The context consistency in Eq. (4) is the expansion of average semantic similarity measure used in [2], [3]. In that situation, all occurrences of context words are regarded as certain (with $P(u|H-G_i,O)$ set as 1), as a result, Eq. (4) degenerates to the formula for computing the consistency in [2].

In this work, we use the context consistency of a word as a type of confidence measure for STD. The context consistency is calculated from the word posterior lattice for each hypothesized word. Moreover, the context consistency can also be combined with the lattice-based posterior probability as a complement to improve the confidence measure [3]. Then, the confidence measure can be employed to verify the potential word detections for STD in the second step.

4. Topic Adaptation of Context Consistency

Obviously, the measure of semantic similarity is affected by the topic or domain of the document. The measure of semantic similarity between words will change when the topic or domain varies. In Sect. 3, we ignore the effect of the topic or domain to the measure of semantic similarity when computing the context consistency. In this section, we propose a topic adaptation method of context consistency to make the measure of semantic similarity toward the topic of the spoken document. This method calculates the context consistency in two steps. It firstly classifies the spoken document according to the topics and then computes the context consistency of the hypothesized word with the topic-adaptive measure of semantic similarity.

4.1 Topic Classification of Spoken Document

For topic classification, we implement a naive Bayes classifier to identify the topic of the spoken documents, which has been shown as an effective probabilistic approach [15], [17]. Assume that $Z=\{z_1,\ldots,z_{N_T}\}$ is the set of N_T different topics. The goal of the classifier is to determine the posterior probability $P(z|d)$ of a topic z ($z\in Z$) given a spoken document d.

4.1.1 Feature of Classification

To classify the documents, words are usually employed as the features with the idea of the bag-of-words, and a text document is represented by the occurrence counts of the individual words present in the document, independent of their ordering [13]. For the spoken documents, the occurrence count for a word is replaced using the expected occurrence count. For a spoken document d, $c_{d,v}$ is the expected occurrence count of word v, and it is estimated by summing the posterior probabilities over all hypotheses of word v in the lattice of d. Thus, the expected count $c_{d,v}$ is allowed to have non-integer values.

Furthermore, the classifiers often preselect a set of topic specific features (i.e., content words) contributing heavily to the determination of the topic, while ignoring the words (i.e., non-content words) contributing nothing to the decision. To select features, we employ a successful method based on the topic posterior probability $P(z|v)$ of topic z given v, and it can be computed using maximum a posterior probability (MAP) estimation as [15]:

$$P(z|v)=\frac{N_{v|z}+N_T P(z)}{N_v+N_T} \tag{6}$$

Here, N_v is the total estimated count that word v appears in all documents, N_T is the number of topics, $N_{v|z}$ is the total estimated count that word v appears in the documents about topic z, and $P(z)$ is the prior probability of topic z as estimated from the training corpus. The words with the top-N posterior probabilities are selected as the features for each topic, and they comprise the vocabulary V for topic classification. Consequently, the feature vector $\mathbf{x}_d$ can be constructed for the spoken document d with each element $c_{d,v}$ for each word $v\in V$.

4.1.2 Naive Bayes Classification

Following Bayes decision theory, we calculate the posterior probability of a topic z given a feature vector $\mathbf{x}_d$ via the Bayes' rule as:

$$P(z|\mathbf{x}_d)=\frac{P(\mathbf{x}_d|z)\,P(z)}{\sum_{i=1}^{N_T}P(\mathbf{x}_d|z_i)\,P(z_i)} \tag{7}$$

Here, $P(\mathbf{x}_d|z)$ represents the likelihood that $\mathbf{x}_d$ is generated as given the topic z. When the statistical independence is assumed between each of the individual words in $\mathbf{x}_d$, $P(\mathbf{x}_d|z)$ can alternatively be approximated by using the expected counts as:

$$P(\mathbf{x}_d|z)\approx\prod_{\forall v\in V}P(v|z)^{c_{d,v}} \tag{8}$$

The probability function $P(v|z)$ is learned from training data using MAP estimation and Laplace smoothing as

$$P(v|z)=\frac{N_{v|z}+N_V P(v)}{N_{AW|z}+N_V} \tag{9}$$

where $N_{AW|z}$ is the total number of all words in the training documents on topic z from vocabulary, N_V is the number of unique words in the V, and $P(v)$ is the prior likelihood of word v occurring independent of the topic. $P(v)$ is estimated from the full collection of training documents using MAP estimation and Laplace smoothing as

$$P(v)=\frac{N_v+1}{N_{AW}+N_V} \tag{10}$$

where N_{AW} is the total count of all words from the N_V word

vocabulary in the training corpus.

4.2 Topic-Adaptive Measure of Semantic Similarity

When the topic class of the spoken document is determined, the context consistency can be computed with the topic-adaptive measure of semantic similarity toward that topic class. In this work, we propose two methods to compute the topic-adaptive measure of semantic similarity, and they are based on the information from the top-1 topic and the mixture of all topics, respectively.

4.2.1 Top-1 Topic

The most probable topic class of d can be given by the decision rule:

$$z^d_{max} = \arg\max_{z_j} P\left(z_j|\mathbf{x_d}\right) \tag{11}$$

This adaptive measure of semantic similarity is computed according to the measure of the most probable topic z^d_{max}:

$$SS^d_{top}(w,u) = SS\left(w,u|z^d_{max}\right) \tag{12}$$

where $SS(w,u|z)$ is the topic-specific measure of semantic similarity within each sentence between word w and word u for topic z. It can be estimated with the text data from topic z by the common semantic similarity, such as LSA, PMI, NPMI, and GLSA. Consequently, the topic-adaptive context consistency of a hypothesized word can be computed for the confidence measure by using Eq. (4) with $SS^d_{top}(w,u)$ as the measure of semantic similarity, when z^d_{max} is determined.

4.2.2 Topic Mixture

The other method is performed using all the topics for d. Thus, the measure is calculated as a mixture weighted with the posterior probability of each topic as:

$$SS^d_{mix}(w,u) = \sum_{j=1}^{N_T} SS\left(w,u|z_j\right) P\left(z_j|\mathbf{x_d}\right) \tag{13}$$

Similarly, the topic-adaptive context consistency of a hypothesized word can be computed for the confidence measure by Eq. (4) and $SS^d_{mix}(w,u)$, when $P\left(z_j|\mathbf{x_d}\right)$ is known.

5. Experiments

5.1 Experimental Setup

We evaluate our proposed confidence measure with an STD system on Mandarin Chinese. Though the proposed confidence measure can be used for out-of-vocabulary (OOV) term detection, it is difficult to collect sufficient text data to estimate the measures of semantic similarity between OOV terms and in-vocabulary (INV) terms. Therefore, we only evaluate the proposed confidence measure for the detection of INV terms.

We use a two-step STD system for evaluation. In the first step of STD, a speech recognizer is set up to transcribe the utterances to the word lattices. From the original lattices, the word posterior lattices are generated by a forward-backward algorithm [22]. For each hypothesized word, the proposed confidence measure is calculated according to the posterior lattice, and the information is stored including the start time, the end time, and the confidence measure. In the second step, the user-defined query word is searched in the results of the first step to get all potential hypotheses, and the confidence measure can be used directly to verify the hypothesized words.

For the speech recognizer, the training data set of acoustic model contains 80-hour news speech and 114-hour reading-style speech. The news speech is recorded from China Central Television, and the reading-style speech is provided by Chinese National Hi-Tech Project 863. The sample rate of all the speech data is 16 kHz. In the front-end, the length and shift of analysis frame are 25 ms and 10 ms, respectively. The used feature is 12th-ordered Mel-frequency cepstral coefficients (MFCCs) and the normalized short-time energy, appending their first- and second-order derivatives (39-dimensional feature). The phone set contains 97 phones [26], and any word and tonal syllable in Mandarin Chinese can be expressed with these phones. The acoustic models are tied-state tri-phone continuous density HMMs. Each HMM has three emitting states with a left-to-right topology, and the number of Gaussian mixture components is 8 for each state. The acoustic models are trained using the Baum-Welch update formulas. A vocabulary with 23.1 K words is employed and a word trigram model used as the language model is trained with 22 M text corpus from People's Daily (a Chinese newspaper).

To estimate the measure of semantic similarity, another text data of news is also collected from People's Daily, and the data is about 16 M text including 20,000 documents (295,783 sentences). Each document is a thematically segmented report, which deals with a single topic. Chinese word segmentation for text corpus is conducted with Language Technology Platform [27], which is an integrated Chinese processing platform. To construct a stop word list, a method of automatic stop words identification is employed [28], and the number of stop words is 200. Therefore, the vocabulary size is 22.9 K for estimating the measure of the semantic similarity. For both the measures of LSA and GLSA, the dimension of the reduced space is 150 to achieve an adequate balance between reconstruction error and noise suppression [29]. The topic class of each document is labeled with the topic of the pages in the newspaper, while the topic label of each sentence is consistent with its document. There are 12 topics used in the experiments (e.g. "Politics", "Economics", "Culture", etc.), and they can cover almost all documents in the newspapers. Then, the topic-specific measure of semantic similarity $SS(w,u|z)$ between words w and u can be estimated with the text data for each topic z.

To train the classifier for the spoken documents, a training set is prepared including 16-hour speech (680 spoken documents) from 1997 Mandarin Broadcast News corpus (Hub-4NE) data [30] and 24-hour speech (922 spoken documents) from China Central Television. The spoken documents are selected from the 12 topics and labeled manually. As described in Sect. 4.1.1, the words with the top-N posterior probabilities are selected as the features for each topic, and N is determined by minimizing the classification error rate (CER) on a development set. The development set consists of 4-hour speech (175 spoken documents) from Hub-4NE. The minimal CER is 10.3% on the development set, when the total number of features are 503 with N as 54 for each topic.

5.2 Evaluation Measure

The test set for STD consists of 4-hour speech (166 documents, 2484 utterances), which is also from Hub-4NE. The performance of speech recognition is evaluated by the lattice error rate (LER), which gives the minimum word error rate of all hypothesized paths through the word lattice. The LER of the recognizer is 9.1% on the test set. The CER is 10.8% for topic classification of spoken documents on that test set. For the test of STD, fifty single-words in vocabulary are selected manually as the query terms, and these words appear 872 times in all test utterances. The performance of the confidence measure is evaluated using the figure-of-merit (FOM) [31], which is defined for keyword spotting task by the average of the word detection rates over a range of 1 to 10 false alarms per keyword per hour of speech.

5.3 Experimental Results

First, we study the effectiveness of the occurrence probability for the confidence measure of context consistency. Here, we compare two confidence measures, and both of them employ context consistency based on semantic similarity. The first one is performed without the word occurrence probability of context word as described in [3], and the context consistency is calculated as the average measure of semantic similarity for the hypothesized word with all context words in its context. The second confidence measure is our proposed method using the occurrence probability of context word to measure the uncertainty of the context. We implement four measures of semantic similarity for the two confidence measures respectively, including LSA, PMI, NPMI, and GLSA. The effect of topic is not considered in these experiments.

In Table 1, the FOMs of the two confidence measures are listed. The proposed confidence measure performs better on all measures of semantic similarity, compared with the confidence measure which ignores the occurrence probability of the context word. This suggests that the occurrence probability of the context word is useful when computing the context consistency. Since considering the uncertainty of context can make the context consistency more accurate, and it provides an improved confidence measure for hypothesized words. It is also shown that both the NPMI and GLSA based measures of semantic similarity outperform the ones using LSA and PMI for confidence measure.

Table 1 The effectiveness of the occurrence probability for the confidence measure.

Semantic similarity	Without occurrence probability	Using occurrence probability
LSA	45.9	54.1
PMI	51.4	56.0
NPMI	51.8	57.8
GLSA	52.4	58.6

Next, we use the topic adaption for context consistency. Here, we compare the proposed context consistency using topic adaption with the context consistency ignoring the topic information, which is regarded as "non-topic" context consistency. Table 2 represents the performance of the proposed context consistency for the confidence measure, and both the measures of semantic similarity based on top-1 topic and topic mixture are shown.

It is observed that the topic adaption can improve the performance of both context consistent considering and ignoring the occurrence probability. This suggests that the topic adaption is helpful for context consistency, and the proposed method can make the measure of semantic similarity computed adaptively. However, the improvements of topic adaptation are not as obvious as the improvements given by occurrence probability. For example, the improvement is 6.2 (from 52.4 to 58.6) by using occurrence probability and GLSA, while the improvement is only 2.2 (from 58.6 to 60.8) with consideration of the topic information. The reason may be that the ability of topic adaption is limited for the measure of semantic similarity in our experiment. Furthermore, we can also see that the confidence measure of context consistency using occurrence probability outperforms the one without occurrence probability, even after employing the topic adaptation. The topic adaptation using the information from the topic mixture performs better than that from the top-1 topic for the measure of semantic similarity.

At last, we combine all proposed confidence measures with a classical confidence measure using lattice-based posterior probability (LBPP) [22], which works without employing the context consistency. For the overlapping hypotheses of the same word, a final confidence measure is computed by combining the posterior probabilities of these hypotheses with Eq. (1), and the time segmentation is estimated by the average time approach [23]. The FOM of LBPP is 75.7. The combinations are conducted by a simple linear interpolation strategy, and the weight of the interpolation is optimized for each combination on a 4-hour development set from Hub-4NE. Table 3 presents the performance of the combinations.

It can be observed that all the combinations are more effective than the LBPP. Since the semantic similarity can extract the word relationship over a longer range without the

Table 2 Performance comparison (FOM) of the topic adaptation.

Semantic similarity	Without occurrence probability			Using occurrence probability		
	Non-topic	Top-1	Mixture	Non-topic	Top-1	Mixture
LSA	45.9	46.7	47.1	54.1	55.1	55.5
PMI	51.4	52.0	52.8	56.0	57.5	57.9
NPMI	51.8	53.0	53.8	57.8	59.2	60.2
GLSA	52.4	53.6	54.4	58.6	59.9	60.8

Table 3 Performance comparison (FOM) of linear combination with the confidence measure using lattice-based posterior probability (LBPP).

Combination	Without occurrence probability			Using occurrence probability		
	Non-topic	Top-1	Mixture	Non-topic	Top-1	Mixture
LBPP	75.7					
LBPP + LSA	77.5	78.1	78.3	79.9	80.6	80.7
LBPP + PMI	78.9	79.4	79.6	80.3	80.9	81.2
LBPP + NPMI	79.1	79.7	79.9	81.7	82.8	83.1
LBPP + GLSA	79.5	80.1	80.5	82.2	83.1	83.4

negative effects of common function words, it can work as an independent linguistic knowledge and complements the *N*-gram language model employed by LBPP. After combination, the performance of the proposed confidence measure considering the occurrence probability of context word is still better than the one which ignores the occurrence probability. We can also see the effectiveness of the occurrence probability for the non-topic context consistency obviously, and the FOM increases from 75.7 to 82.2 with GLSA. The highest FOM is 83.4 in all combinations, and it is achieved by the proposed confidence measure which employs the word occurrence probability and topic adaptation with GLSA as the sematic similarity from the mixture of the topics.

6. Conclusions

In this paper, a novel confidence measure is proposed based on context consistency in lattice for STD. The context consistency is computed with the consideration of the uncertainty of the context and the effect of topic. The uncertainty of context is estimated by word occurrence probability, which is obtained through combining the overlapping hypotheses in word posterior lattice. To handle the effect of topic, the topic adaptation for context consistency is performed using a direct method based on topic classification. The experiments conducted on the Hub-4NE Mandarin database show that the occurrence probability of context word is effective for the confidence measure of STD. The topic adaptation of context consistency is proved helpful, though it needs extra text and speech data with only the topic label. The proposed confidence measure outperforms the one ignoring the uncertainty of the context or the one using non-topic method. Moreover, the proposed confidence measure also yields better performance when combined with the lattice-based posterior probability.

In this paper, we only focus on the proposed confidence measure for INV term detection, and we will investigate the confidence measure for OOV terms in the future. A possible solution to estimate the measure of semantic similarity for OOV terms is to collect the external prior data of text containing the OOV terms automatically through the Internet.

Acknowledgements

Part of this work [32] has been published in the Proceedings of the Annual Conference of International Speech Communication Association (INTERSPEECH), 2012. This research is supported by the National Natural Science Foundations of China (No. 91120303 and No. 91220301) and the Ph.D. Programs Foundation of Ministry of Education of China (No. 20112302110042).

References

[1] National Institute of Standards and Technology (NIST), "The spoken term detection (STD) 2006 evaluation plan," 2006. [Online]. Available: http://www.nist.gov/speech/tests/std
[2] S. Cox and S. Dasmahapatra, "High-level approaches to confidence estimation in speech recognition," IEEE Trans. Speech Audio Process., vol.10, no.7, pp.406–417, 2002.
[3] G. Guo, C. Huang, H. Jiang, and R.H. Wang, "A comparative study on various confidence measures in large vocabulary speech recognition," Proc. ICASSP, 2004, pp.9–12, 2004.
[4] D. Schneider, T. Mertens, M. Larson, and J. Kohler, "Contextual verification for open vocabulary spoken Term Detection," Proc. Interspeech, 2010, pp.697–700, 2010.
[5] H.Y. Lee, T.W. Tu, C.P. Chen, C.Y. Huang, and L.S. Lee, "Improved spoken term detection using support vector machines based on lattice context consistency," Proc. ICASSP, 2011, pp.5648–5651, 2011.
[6] T. Asami, N. Nomoto, S. Kobashikawa, Y. Yamaguchi, H. Masataki, and S. Takahashi, "Spoken document confidence estimation using contextual coherence," Proc. Interspeech, 2011, pp.1961–1964, 2011.
[7] C. Chelba, T.J. Hazen, and M. Saraclar, "Retrieval and browsing of spoken content," IEEE Signal Process. Mag., vol.25, no.3, pp.39–49, 2008.
[8] 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.
[9] R. Rosenfeld, "Two Decades of statistical language modeling: Where do we go from here?" Proc. IEEE, vol.88, no.8, pp.1270–1278, 2000.
[10] J.R. Bellegarda, "Statistical language model adaptation: review and

perspectives," Speech Commun., vol.42, no.1, pp.93–108, 2004.
[11] D. Gildea and T. Hoffman, "Topic-based language modeling using EM," Proc. European Conf. Speech Comm. Technol., 1999, vol.5, pp.2167–2170.
[12] S.H. Bai, C.C. Leung, C.L. Huang, B. Ma, and H.Z. Li, "Building topic mixture language models using the document soft classification notion of topic models," Proc. ISCSLP, 2010, pp.229–232, 2010.
[13] D. Lewis, "Naive (Bayes) at forty: The independence assumption in information retrieval," Proc. ECML, 1998, pp.4–15, 1998.
[14] R. Schwartz, T. Imai, F. Kubala, L. Nguyen, and J. Makhoul, "A maximum likelihood model for topic classification of broadcast news," Proc. Eurospeech, 1997, pp.1455–1458, 1997.
[15] T.J. Hanzen, F. Richardson, and A. Margolis, "Topic identification from audio recordings using word and phone recognition lattices," Proc. IEEE Workshop ASRU, 2007, pp.659–644, 2007.
[16] P. Haffner, G. Tur, and J. Wright, "Optimizing SVMs for complex call classification," Proc. ICASSP, 2003, pp.632–635, 2003.
[17] T.J. Hazen, "MCE training techniques for topic identification of spoken audio documents," IEEE Trans. Audio Speech Language Process., vol.19, pp.2451–2460, 2011.
[18] I. Zitouni, "Constrained minimization and discriminative training for natural language call routing," IEEE Trans. Speech Audio Process., vol.16, no.1, pp.208–215, 2008.
[19] L. Chen, J.L. Gauvain, L. Lamel, and G. Adda, "Unsupervised language model adaptation for broadcast news," Proc. ICASSP, 2003, vol.1, pp.220–223, 2003.
[20] G. Lecorve, G. Gravier, and P. Sebillot, "An unsupervised web-based topic language model adaptation method," Proc. ICASSP, 2008, pp.5081–5084, 2008.
[21] Z.Y. Zhou, P. Yu, C. Chelba, and F. Seide, "Towards spoken-document retrieval for the internet: lattice indexing for large-scale web-search architectures," Proc. HLT, 2006, pp.415–422, 2006.
[22] 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, 2001.
[23] D. Wang, N. Evans, R. Troncy, and S. King, "Handling overlaps in spoken term detection," Proc. ICASSP, 2011, pp.5656–5659, 2011.
[24] G. Bouma, "Normalized (pointwise) mutual information in collocation extraction," Proc. GSCL, 2009, pp.31–40.
[25] I. Matveeva, G. Levow, A. Farahat, and C. Royer, "Generalized latent semantic analysis for term representation," Proc. RANLP, 2005, 2005.
[26] C. Huang, Y. Shi, J.L. Zhou, M. Chu, T. Wang, and E. Chang, "Segmental tonal modeling for phone set design in mandarin LVCSR," Proc. ICASSP, 2004, vol.1, pp.901–904, 2004.
[27] W. Che, Z. Li, and T. Liu, "LTP: A Chinese Language Technology Platform," Proc. COLING (Demos), 2010, pp.13–16, 2010.
[28] F. Zou, F.L. Wang, X.T. Deng, and S. Han, "Automatic identification of Chinese stop words," Research on Computing Science, vol.18, pp.151–163, 2006.
[29] J.R. Bellegarda, "Exploiting latent semantic information in statistical language modeling," Proc. IEEE, vo1.88, no.8, pp.1279–1296, 2000.
[30] Linguistic Data Consortium, 1997 Mandarin Broadcast News Speech (HUB-4NE), ISBN: 1-58563-125-6, 1998. Available: http://www.ldc.upenn.edu/Catalog/LDC98S73.html
[31] J.R. Rohlicek, W. Russell, S. Roukos, and H. Gish, "Continuous hidden Markov modeling for speaker-independent word spotting," Proc. ICASSP, 1989, vol.1, pp.627–630, 1989.
[32] H.Y. Li, J.Q. Han, T.R. Zheng, and G.B. Zheng, "A novel confidence measure based on context consistency for spoken term detection," Proc. Interspeech, 2012, pp.2429–2432, 2012.

Haiyang Li received the B.S. degree in computer science from the Jilin University, Changchun, China, in 2005 and the M.S. degree in computer science from the Harbin Institute of Technology, Harbin, China, in 2007. He was a visiting student in Microsoft Research Asia in 2006. From 2007 to 2009, He was a software engineer in the Hewlett-Packard (HP) Co., Ltd. Dalian Branch. Currently, he is working towards the Ph.D. degree in the School of Computer Science of Harbin Institute of Technology. His research interests include speech signal processing, spoken term detection, and confidence measure.

Tieran Zheng received the B.S degree in electrical engineering from the Harbin University of Science and Technology, Harbin, China in 1993, the Ph.D degree in the School of Computer Science from the Harbin Institute of Technology, Harbin, China in 2008. He was a project manager in the Computer Corporation of Heilongjiang Province and in the Harbin XinZhongXin Co., Ltd in 1993–2001. He was a visiting scientist in the IBM Research Center China from November 2004 to February 2005. Currently, he is an associate professor in the school of Computer Science and Technology at the Harbin Institute of Technology. His research interests include speech signal processing and spoken document retrieval.

Guibin Zheng received the B.S degree in electrical appliances control from Shanghai Jiaotong University, Shanghai, China, in 1996, received the Ph.D degree in the School of Computer Science from the Harbin Institute of Technology, Harbin, China, in 2006. Currently, he is an associate professor in the school of Computer Science and Technology at the Harbin Institute of Technology. His research interests include speech signal processing and audio information retrieval.

Jiqing Han received the B.S., M.S. in electrical engineering, and Ph.D. degrees in computer science from the Harbin Institute of Technology, Harbin, China, in 1987, 1990, and 1998, respectively. From June 1996 to January 1998, he was a Visiting Scientist in System Engineering Research Institute, South Korea. Currently, he is a Professor of the School of Computer Science and Technology, Harbin Institute of Technology. He is a member of IEEE, Vice Chairman of Society of speech processing, Association for Chinese information processing, Vice Chairman of Standing Committee of National Conference on Man-Machine speech Communication, China, member of the editorial board of Journal of Chinese Information Processing, and member of the editorial board of the Journal of Data Acquisition & Processing. Prof. Han is undertaking several projects from the National Natural Science Foundation, 863Hi-tech Program, National Basic Research Program. He has won three Second Prize and two Third Prize awards of Science and Technology of Ministry/Province. He has published more than 180 papers and 3 books. His research fields include speech signal processing and audio information processing.