ISCA Archive
http://www.isca-speech.org/archive

INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

A Novel Confidence Measure Based on Context Consistency for Spoken Term Detection

Haiyang Li, Jiqing Han, Tieran Zheng, Guibin Zheng

School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China
{hyli, jqhan, zhengtieran, zhengguibin}@hit.edu.cn

Abstract

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 word lattice. When calculating the context consistency of a hypothesized word, the proposed confidence measure considers not only the semantic similarity between words but also the uncertainty of the context. To measure the uncertainty of the context, we employ the word occurrence probability, which is obtained by combining the overlapping hypotheses in word posterior lattice. Additionally, we also use two effective measures of semantic similarity to acquire more accurate context consistency for confidence measure. The experiments conducted on the Hub-4NE Mandarin database show that the proposed confidence measure can achieve improvements over the confidence measure which ignores the word occurrence probability of context word.

Index Terms: confidence measure, spoken term detection, context consistency, semantic similarity, word occurrence probability

10.21437/Interspeech.2012-636

1. Introduction

Spoken term detection (STD) is the task which aims to locate all occurrences of terms queried by 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 in 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 detections of the user-defined term in the results of the first step, and further verifies those detections.

In STD, confidence measure is applied to indicate the reliability of detections, and it is crucial to reject false alarms. It is expected that confidence measure can assign high confidence to correct detection and low confidence to 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 [2-6]. 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 [2, 3, 6], and as a high-level confidence measure, this context consistency is effective [2, 3]. 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. And it is assumed that the occurrences of the context words are certain in recognized result. However, the assumption is not true because 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. The context feature vectors are constructed by computing the occurrence counts for each word. As each dimension corresponds to an in-vocabulary word, the vector dimensionality is the word number of the vocabulary. However, this method needs 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.

In this paper, we attempt to overcome the shortcomings of current methods in measuring context consistency. To this end, we propose a novel confidence measure using context consistency based on semantic similarity and lattice. The proposed confidence measure estimates the uncertainty of the context by word occurrence probability, which is obtained by combining the overlapping hypotheses in word posterior lattice. Moreover, the proposed confidence measure does not need any speech data to train additional model for the context. We finally confirm the effectiveness of the proposed confidence measure by further experiments on the Hub-4NE Mandarin database.

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 1-best result for STD, since lattice can provide much more useful information and produce better recall rates [8]. 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 [8]. 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 original lattice produced by speech recognizer, word likelihood is kept

for each hypothesis. A word posterior lattice can be generated by a forward-backward algorithm from original lattice [9], and 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 [10].

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, thus a set of hypotheses H for an utterance O is divided into M groups $G_1, ..., G_M$. Each group $G_i = \left\{g_i^1, ..., g_i^{N_i}\right\}$ is composed of the hypotheses of the same hypothesized word w_i, and these hypotheses overlap each other, where N_i is the element number of G_i. $H = \bigcup_{i=1}^{M} G_i$. And for any integer i and j in $\left\{1, ..., M\right\}$, 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. The word occurrence probability of w_i in τ_i is computed with exclusive accumulated confidence [10] in G_i as:

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

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

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:

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

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

3. Computation of context consistency

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 $\{w_j | j = 1, ..., M \wedge j \neq i\}$. 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 as confidence measure according to the occurrence probability of its context words as:

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

where $SS\,(w_i, v)$ is the measure of semantic similarity between word w_i and word v. Notice that word w_i can be the same as word v in Eq. (4), for a word can appear in the context of the same word. And the semantic similarity will be introduced in section 4 in detail. $H - G_i$ 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. $P\,(v|H - G_i, O)$ denotes the word occurrence probability that v occurs in the context of w_i, and $P\,(v|H - G_i, O)$ is calculated according to Eq. (3) as:

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

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

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

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 occurrence probability of the context word. 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], which ignores the occurrence probability of hypothesized word in context. In that situation, all occurrences of context words are regarded as certain (with $P\,(v|H - G_i, O)$ set as 1) , as a result, Eq. (4) degenerates to the formula for computing the consistency in [2].

4. Measurement of semantic similarity between words

In this section, four methods are employed to measure the semantic similarity: latent semantic analysis, pointwise mutual information, normalized pointwise mutual information, and generalized latent semantic analysis. The first two methods have been applied to compute the context consistency for confidence measure [2, 3, 6]. The last two novel methods are also used to acquire better measurement of semantic similarity for context consistency.

4.1. Latent semantic analysis (LSA)

LSA is used to compute semantic similarity between words [11], and it is based on the assumption that words co-occurring across documents are semantically associated. LSA begins with an inner matrix of co-occurrences, which is constructed between words and documents. Each element of the matrix is word frequency in each document, normalized by document length and word entropy [12]. LSA performs the singular value decomposition (SVD) on the inner matrix to construct a reduced-dimensional space named as semantic space. Then, a vector for each word is formed in the reduced-dimensional space, and similarity between words is measured by the cosine of the angle between their corresponding vectors. And the value of LSA based measure between the same words is defined as 1.

4.2. Pointwise mutual information (PMI)

Mutual information is employed to measure dependencies between random variables. A variant of mutual information is called pointwise mutual information (PMI), which has been applied to measure the semantic similarity between two words [3, 6]. PMI is also a co-occurrence metric, and it normalizes the co-occurrence probability of two words with their individual occurrence probabilities. The definition of PMI is

$$PMI\left(w,v\right)=\log\frac{P_T\left(w,v\right)}{P_T\left(w\right)\cdot P_T\left(v\right)} \tag{6}$$

where $P_T\left(w,v\right)$ is the co-occurrence probability of words w and word v in the same sentence of text corpus T, and $P_T\left(w\right)$ is the occurrence probability of word w in each sentence. To deal with the problem of sparse data, the PMI can be smoothed with the method in [6]. If w and v are the same, then $PMI\left(w,w\right)=-\log P_T\left(w\right)$.

4.3. Normalized pointwise mutual information (NPMI)

PMI is sensitive to low frequency data, and it is hardly interpretable as an association measure [13]. The upper and lower bounds of PMI value are $\min\left[-\log P_T\left(w\right),-\log P_T\left(v\right)\right]$ and $-\infty$ respectively. To improve the measurement of PMI, we use normalized pointwise mutual information (NPMI), which is computed with [13]:

$$NPMI(w,v)=-\frac{PMI\left(w,v\right)}{\log P_T\left(w,v\right)} \tag{7}$$

As a result, the value of PMI can be normalized between -1 and 1. Some special values are set as follows: if only w and v are the same word or occur together as $P_T(w,v)=1$, then $NPMI\left(w,v\right)=1$; if two words occur separately but not together as $P_T(w,v)=0$, then $NPMI\left(w,v\right)=-1$.

4.4. Generalized latent semantic analysis (GLSA)

GLSA extends the idea of LSA by defining a different way to obtain the correlation matrix and has been shown to outperform LSA and PMI on the measure of semantic similarity [14]. GLSA begins with a word-by-word matrix instead of the word-by-document matrix as the inner matrix. Each entry of the inner matrix of GLSA represents the PMI between words in the vocabulary, thus the matrix is symmetric in this case. SVD is also applied to the resulting matrix, and the semantic similarity between the two words is the cosine of angle between corresponding vectors in the reduced-dimensional space as described above.

5. Experiments

5.1. Experimental setup

We evaluate our proposed confidence measure with an STD system on Chinese Mandarin. A speech recognizer is set up to transcribe the utterances to word lattices. The training data set for 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 25ms and 10ms 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 [15], and any word and tonal syllable in Chinese Mandarin 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 the Gaussian mixture components is 8 for each state. The acoustic models are trained using the Baum-Welch update formulas. A vocabulary with 23.1K words is employed and a word trigram model used as language model is trained with 22M text corpus from China Daily (a Chinese newspaper).

To estimate the measure of semantic similarity, another data of text format is also collected from China Daily, and the data is about 16M text including 20,000 documents (295,783 sentences). Chinese word segmentation for text corpus is conducted with Language Technology Platform [16], which is an integrated Chinese processing platform. To construct a stop word list, a method of automatic stop words identification is employed [17], and the number of stop words is 200. The dimension of the reduced space is 150 for both LSA and GLSA.

The test set consists of 4-hour speech (2484 utterances), which is from 1997 Mandarin Broadcast News corpus (Hub-4NE) data [18]. Fifty single-words in vocabulary are selected manually as the query terms for the test, and these words appear 872 times in all test utterances. The performance of the confidence measure is evaluated using the figure-of-merit (FOM) [19], 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.

For comparison, a lattice-based confidence measure is used as a baseline, which works only based on the posterior probability from lattice 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 [10]. The FOM of baseline is 75.7.

5.2. Experimental results

First, we compare two confidence measures, and both of them employ context consistency based on semantic similarity. The first one described in [3] (denoted by "without OP") ignores the occurrence probability of context word, 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 (denoted by "with OP") is our proposed method with the consideration of the occurrence probability of context word. We implement four measures of semantic similarity described in section 4 for the two confidence mea-

sures respectively. 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. 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: *Performance comparison (FOM) of confidence measures based on various measures of semantic similarity.*

Semantic similarity	Without OP	With OP (proposed)
LSA	45.9	54.1
PMI	51.4	56.0
NPMI	51.8	57.8
GLSA	52.4	58.6

Next, we combine the baseline with the two confidence measures respectively. 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 2 presents the performance of the combinations. It can be observed that all the combinations are more effective than the baseline. After combination, the performance of the proposed confidence measure is still better than the one which ignores the occurrence probability of context word. The highest FOM is 82.2 in all combinations, and it is achieved by the proposed confidence measure with GLSA for the semantic similarity.

Table 2: *Performance of linear combination with baseline.*

Combination	Without OP	With OP (proposed)
Baseline	75.7	
Baseline + LSA	77.5	79.9
Baseline + PMI	78.9	80.3
Baseline + NPMI	79.1	81.7
Baseline + GLSA	79.5	82.2

6. Conclusions

In this paper, a novel confidence measure is proposed based on context consistency in lattice for STD. For the computation of context consistency, the proposed confidence measure employs not only the semantic similarity between hypothesized words but also the uncertainty of the context words. The word occurrence probability is calculated to measure the uncertainty by combining the overlapping hypotheses in word posterior lattice. The proposed confidence measure does not need any speech data to train extra model for the context. Additionally, two novel measures of semantic similarity base on NPMI and GLSA are presented for context consistency. The experimental results show that the proposed confidence measure outperforms the method which ignores the uncertainty of context. And the proposed confidence measure also yields a better performance after combined with the lattice-based confidence measure. Moreover, two novel measures of semantic similarity are both shown to outperform the LSA and PMI based measures on experiments for confidence measure.

7. Acknowledgements

This research is supported by the National Natural Science Foundation of China under grants No. 91120303 and No. 61071181.

8. 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, S. Dasmahapatra, "High-level approaches to confidence estimation in speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 7, pp. 406-417, 2002.

[3] G. Guo, C. Huang, H. Jiang, R. H. Wang, "A comparative study on various confidence measures in large vocabulary speech recognition," in *Proc. ICASSP*, 2004, pp. 9-12.

[4] D. Schneider, T. Mertens, M. Larson, J. Kohler, "Contextual verification for open vocabulary spoken Term Detection," in *Proc. Interspeech*, 2010, pp. 697-700.

[5] H. Y. Lee, T. W. Tu, C. P. Chen, C. Y. Huang, L. S. Lee, "Improved spoken term detection using support vector machines based on lattice context consistency," in *Proc. ICASSP*, 2011, pp. 5648-5651.

[6] T. Asami, N. Nomoto, S. Kobashikawa, Y. Yamaguchi, H. Masataki, S. Takahashi, "Spoken document confidence estimation using contextual coherence," in *Proc. Interspeech*, 2011, pp. 1961-1964.

[7] C. Chelba, T. J. Hazen, and M. Saraclar, "Retrieval and Browsing of Spoken Content," *IEEE Signal Processing Magazine*, vol. 25, no. 3, pp. 39-49, 2008.

[8] 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," in *Proc. HLT*, 2006, pp. 415-422.

[9] F. Wessel, R. Schluter, K. Macherey, and H. Ney, "Confidence measures for large vocabulary continuous speech recognition," *IEEE transaction on Speech and Audio Processing*, vol. 9, no. 3, pp. 288-298, 2001.

[10] D. Wang, N. Evans, R. Troncy, S. King, "Handling overlaps in spoken term detection," in *Proc. ICASSP*, 2011, pp. 5656-5659.

[11] T. K. Landauer, P. W. Foltz, and D. Laham, "An introduction to latent semantic analysis," in *Discourse Processes*, vol. 25, pp. 259-284, 1998.

[12] J. R. Bellegarda, "Exploiting latent semantic information in statistical language modeling," *Proceedings of the IEEE*, vo1. 88, no. 8, pp. 1279-1296, 2000.

[13] G. Bouma, "Normalized (pointwise) mutual information in collocation extraction," in *Proc. GSCL*, 2009, pp. 31-40.

[14] I. Matveeva, G. Levow, A. Farahat, and C. Royer, "Generalized latent semantic analysis for term representation," In *Proc. RANLP*, 2005.

[15] C. Huang, Y. Shi, J. L. Zhou, M. Chu, T. Wang, E. Chang, "Segmental tonal modeling for phone set design in mandarin LVCSR," in *Proc. ICASSP*, 2004, vol. 1, pp. 901-904.

[16] W. Che, Z. Li, and T. Liu, "LTP: A Chinese Language Technology Platform," in *Proc. COLING (Demos)*, 2010, pp. 13-16.

[17] F. Zou, F. L. Wang, X. T. Deng, S. Han, "Automatic identification of Chinese stop words," *Research on Computing Science*, vol. 18, pp. 151-163, 2006.

[18] 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

[19] J. R. Rohlicek, W. Russell, S. Roukos, H. Gish, "Continuous hidden Markov modeling for speaker-independent word spotting," in *Proc. ICASSP*, 1989, vol. 1, pp. 627-630.