A Spoken Dialogue System Based on Keyword Spotting Technology

Pengyuan Zhang, Qingwei Zhao, and Yonghong Yan

ThinkIT Speech Lab, Institute of Acoustics, Chinese Academy of Sciences
Beijing 100080, P.R. China
{pzhang,qzhao,yyan}@hccl.ioa.ac.cn

Abstract. In this paper, a keyword spotting based dialogue system is described. It is critical to understand user's requests accurately in a dialogue system. But the performance of large vocabulary continuous speech recognition (LVCSR) system is far from perfect, especially for spontaneous speech. In this work, an improved keyword spotting scheme is adopted instead. A fuzzy search algorithm is proposed to extract keyword hypotheses from syllable confusion networks (CN). CNs are linear and naturally suitable for indexing. To accelerate search process, CNs are pruned to feasible sizes. Furthermore, we enhance the discriminability of confidence measure by applying entropy information to the posterior probability of word hypotheses. On mandarin conversational telephone speech (CTS), the proposed algorithms obtained a 4.7% relative equal error rate (EER) reduction.

1 Introduction

Research on spoken dialogue systems and their real-world applications have attracted increased attention in recent years [1]. For spoken dialogue applications it is critical to understand the user's requests accurately, since the rest of the system acts based on this recognized result [2]. On the other hand, the performance of current large vocabulary continuous speech recognition (LVCSR) systems is far from perfect, especially for spontaneous speech. In this paper, we focus our interest on the study of recognition strategy obtained during the utterance understanding process. A keyword-based approach is employed to understand spontaneous speech.

The task of keyword spotting is to detect a set of required words in the input continuous speech [3]. It is desirable to achieve the highest possible keyword recognition rate, while minimizing the number of false keyword insertions. A syllable confusion matrix (SCM) is adopted to extract keywords in this paper. The quality of SCM has an obvious influence on the performance of keyword spotting. Based on traditional approaches, we propose an improved SCM. Furthermore, we enhance the discriminability of confidence measure by applying entropy information to the posterior probability of word hypotheses.

The remainder of the paper is structured as follows: Section 2 presents an overview of system framework. In section 3, we introduce the keyword spotting scheme. Experiment results are presented in section 4 followed by the conclusions in Section 5.

J. Jacko (Ed.): Human-Computer Interaction, Part III, HCII 2007, LNCS 4552, pp. 253–261, 2007.

2 Overview

The basic architecture of a spoken dialogue system is illustrated in Fig. 1 [4]. Generally, a spoken dialog system consists of two parts: utterance understanding part and utterance generation part. When receiving a user utterance, the system behaves as follows [5]:

(1) The keyword spotting system receives a user utterance and outputs keyword hypotheses.

(2) The keyword hypotheses are passed to the semantic analyzer. Semantic analyses are performed to convert them into a meaning representation, often called a semantic frame.

(3) The discourse understanding component receives the semantic frame, refers to the current dialogue state, and updates the dialogue state.

(4) The dialogue manager refers to the updated dialogue state, determines the next utterance, and outputs the next content to be delivered as a semantic frame. The dialogue state is updated at the same time so that it contains the content of system utterances.

(5) The surface generator builds the system response, typically as a surface expression (text sentence).

(6) The speech synthesizer generates the system voice using a text-to-speech (TTS) conversion system.

This paper concerns the keyword spotting module of this spoken dialogue system. A novel keyword spotting scheme is proposed to extract keyword hypotheses from syllable confusion networks (CNs).

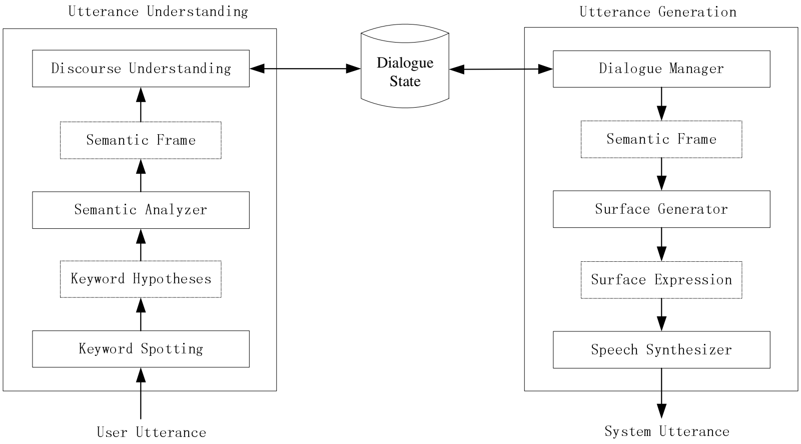

Fig. 1. Module structure of a spoken dialogue system

3 Keyword Spotting Scheme

In our keyword spotting system, search space is built on total Chinese syllables, not specifically for keywords. Syllable recognition is performed without any lexical constraints.

Given a spoken input, a very large lattice is generated firstly. A clustering algorithm is used to translate syllable trigram lattice into a CN [6]. The CN has one node for each equivalence class of original lattice nodes and adjacent nodes are linked by one edge per word hypothesis. We extract keywords from CNs. Generally, confusion matrix is adopted to achieve higher recognition rate in speech recognition system [7-9]. Based on traditional approaches, we generate SCM based on CNs. Entropy information based posterior probability is also applied to reject false accepts.

3.1 Generation of Syllable Confusion Matrix

Confusion matrix is often used for similarity measure. In Mandarin, every character is spoken in a monosyllabic manner. Most of Chinese characters can be expressed by about 1276 syllables which consist of a combination of 409 base syllables and 4 tones. In this work, we build a base syllable confusion matrix which has only 409 entries.

Generally, the syllable confusion matrix is calculated using a syllable recognizer, which recognizes 1-best syllable sequences instead of words [8]. It can be described as the following steps:

(1) Canonical syllable level transcriptions of the accent speech data should be obtained firstly.

(2) A standard Mandarin acoustic recognizer whose output is syllable sequence will be used to transcribe those accent speech data.

(3) With the help of dynamic programming (DP) technique, these recognized syllable sequence are aligned to the canonical syllable level transcriptions. Regardless of insertion and deletion errors, substitution errors are considered. Given a canonical syllable S_m and an aligning hypothesis S_n, we can compute confusion probability:

$$P(S_n \mid S_m) = \frac{count(S_n \mid S_m)}{\sum_{i=1}^{N} count(S_i \mid S_m)} \ . \tag{1}$$

where $count(S_n \mid S_m)$ is the number of S_n which is aligned to S_m. N is the total syllable number in our dictionary.

However, there is a conceptual mismatch between decoding criterion and confusion probability evaluation. Given an input utterance, a Viterbi decoder is used to generate the best sentence. But it does not ensure that each syllable is the optimal one. Instead of 1-best syllable hypotheses, we generate confusion matrix from CNs. N-best hypotheses of each slice are considered. Fig. 2 describes an example of CN. For schematic description, we give top 4 hypotheses in each slice. Corresponding canonical syllable is also presented.

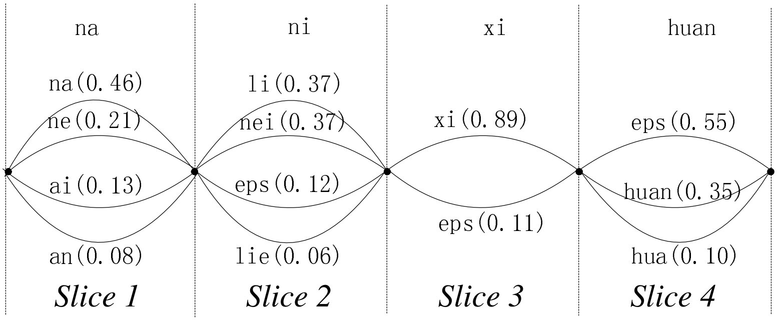

Fig. 2. An example of syllable confusion network

In order to assess whether a CN provides more information than 1-best recognition result, the syllable error rate (SER) on evaluation set is computed. As Table 1 shows, SER of CNs drops significantly compared with 1-best recognition result. That is to say, CNs provide us more effective information.

Table 1. SER of CNs and 1-best recognition result

Methods	SER [%]
1-best recognition result	49.5
CNs	27.1

Recognizer output voting error reduction (ROVER) technology is adopted to align CN and canonical recognition results [10]. We select special slices to generate the confusion matrix. Given a canonical syllable S_m, only slices including S_m are considered. A classification function $\beta(k)$ is defined as:

$$\beta(k) = \begin{cases} 1 & \text{if } S_m \text{ is a most probable syllable in the } k^{\text{th}} \text{ slice} \\ 0 & \quad\quad others \end{cases} . \tag{2}$$

Then, confusion probability can be expressed as:

$$P(S_n \mid S_m) = \frac{\sum_{k=1}^{C} \beta(k) count(S_n \mid S_m)}{\sum_{i=1}^{N} \sum_{k=1}^{C} \beta(k) count(S_i \mid S_m)} . \tag{3}$$

where C is the number of slices in training data, N is the number of syllables in dictionary.

Table 2 presents an example of a confusion matrix. Value in each bracket is the confusion probability between syllables. Obviously, the summation over each row is 1.

Table 2. An example of syllable confusion matrix

ba	ba (0.419)	la (0.074)	ma (0.056)	da (0.043)	...
cai	cai (0.286)	chai (0.107)	tai (0.101)	can (0.088)	...
gen	gen (0.244)	geng (0.106)	ge (0.055)	gou (0.035)	...
lan	lan (0.267)	nan (0.092)	luan (0.062)	lai (0.058)	...
mei	mei (0.416)	nei (0.055)	men (0.051)	lei (0.049)	...
...	...	...	...	...	...

3.2 Fuzzy Keyword Search

Fig.3 describes the block diagram of fuzzy keyword search. For a fast retrieval, arcs of CNs are indexed efficiently. Syllable name and associated posterior probability are also labeled with each arc. In order to exactly locate occurrences of a keyword, we improve the algorithm to record the most probable time information of the syllable in each arc. Moreover, SCM is adopted to improve keyword recognition rate. With the CN and SCM, keyword hypotheses are generated according to the relevance score.

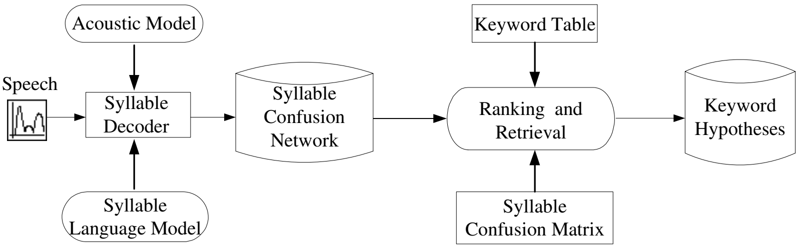

Fig. 3. Block diagram of fuzzy keyword search

In this work, each equivalence class in the CN is defined as a slice. Then CN can be represented as a slice vector $S_N = \{s_1, ..., s_n, ..., s_N\}$. Let the syllable sequence of a keyword Q_M be $\{q_1, ..., q_m, ..., q_M\}$, syllable relevance score $C(m,n)$ is defined as:

$$C(m,n) = \log\{(1-\alpha)p(q_m \mid s_n, O) + \alpha P_{conf}(m,n)\} . \tag{4}$$

$$P_{conf}(m,n) = \sum_{\{\varphi_i \mid \varphi_i \in s_n, \varphi_i \in SimSet(q_m)\}} p(\varphi_i \mid s_n, O) p(q_m \mid \varphi_i) . \tag{5}$$

where $p(q_m \mid s_n, O)$ is the posterior probability, α is a weighting factor, $P_{conf}(m,n)$ is the confusion probability which is simplified by considering q_m's most similar syllables. $SimSet(q_m)$ and $p(q_m \mid \varphi_i)$ are provided by the SCM. The keyword relevance score is calculated by averaging cumulative dynamic programming (DP) score of underlying syllables. The search procedure is to match syllable sequence Q_M with partial slices from the start position of S_N to the end.

3.3 Calculating the Entropy Information

CN is a linear graph transformed from syllable lattice, which aligns links in the original lattice and transforms the lattice into a linear graph in which all paths pass through all nodes. To determine whether a keyword is correct or not, it might be helpful to take all the other arcs in the same slice into account. The posterior probability included in the confusion network is a good confidence measure. Besides posterior probabilities, entropy information obtained in the CN is drawing more attentions in recent years, where not only words in the best path but also words in competing paths are used in computing the probabilities [11, 12].

Entropy measures the difference of posterior probabilities among syllables in the CN, and ambiguity of syllable identity can be better captured by entropy than the posterior probability alone. Entropy for a slice is defined as:

$$E(s_n) = -\sum_{i=1}^{m} p(l_i \mid s_n, O) \log p(l_i \mid s_n, O) \ . \tag{6}$$

where l_i is the syllable in slice s_n, $p(l_i \mid s_n, O)$ is the corresponding posterior probability, m is the syllable number of s_n.

To emphasize the reliability of posterior probability based confidence measure, we propose an entropy-based approach that evaluates the degree of confusion in confidence measure. By incorporating entropy information into traditional posterior probability, the new entropy-based confidence measure of a hypothesized syllable is defined as:

$$C_{entropy}(q_m) = (1-\beta)C(m,n) + \beta E(s_n) \ . \tag{7}$$

Deriving word level scores from syllable scores is a natural extension of the confidence measure. Generally, logarithmic mean is adopted to calculate word confidence. The formula can be written as:

$$CM\left(W\right) = \frac{1}{M}\sum_{m=1}^{M} C_{entropy}(q_m) \ . \tag{8}$$

where M is the number of syllables in w.

4 Experiments

We conducted experiments using our real time keyword spotting system. Acoustic model is trained using train04, which is collected by Hong Kong University of Science and Technology (HKUST) [13]. SCM adopted in this paper are generated using 100 hour mandarin conversational telephone speech (CTS).

4.1 Experimental Data Description

The algorithms proposed in this paper were evaluated on 2005_eval which was provided by HTRDP (National High Technology Research and Development Program). All the data are recorded through landline telephone with local service in real world with environmental noise. All utterances are mandarin conversational speech, but with obvious dialect accent. Speech data are sampled at the rate of 8 kHz with 16 bit quantization. The evaluation set includes 1543 utterances from 14 speakers and the length totals up to 1 hour. 100 keywords were selected randomly as the keyword list. 80 percent is two-syllable Chinese words and others are three-syllable words.

4.2 Experiment Results

A common metric to evaluate the keyword spotting is its equal error rate (EER) which is obtained from the threshold that gives equal false acceptance rate (FA) and false rejection rate (FR). The FA fits the case in which an incorrect word is accepted, and the FR fits the case of rejecting the correct word.

$$FA = \frac{num. of\ incorrect\ words\ labelled\ as\ accepted}{num.\ of\ incorrect\ words} .$$

$$FR = \frac{num. of\ correct\ words\ labelled\ as\ rejected}{num.\ of\ keywords * hours\ of\ testset * C} .$$

where C is a factor which scales the dynamic range of FA and FR on the same level. In this paper, C is set to 10.

Table 3. Effect of two different syllable confusion matrixes

Beam	α	SER [%]	SGD	EER SCM-1 [%]	EER SCM-2 [%]	Relative reduction [%]
0.001	0.01	27.1	18.6	32.7	32.4	0.9
0.01	0.03	30.9	8.9	32.2	31.2	3.1
0.05	0.05	37.2	4.3	35.4	34.9	1.4
0.10	0.10	41.5	3.1	37.4	36.2	3.2

Table 4. EER comparison of different methods

Methods	EER [%]
CN+SCM-1	32.2
CN+SCM-2	31.2
CN+SCM-2+Entropy	30.7

CNs are pruned to contain only those arcs whose posterior probabilities are within a pruning threshold with respect to the best one in each slice. The experiments of two SCMs are summarized in Table 3. The values of syllable graph density (SGD) are also provided. SCM-1 represents the SCM based on 1-best recognition result. SCM-2 is generated based on CN. These results clearly indicate that SCM-2 has more positive impact with different pruning beam. It's interesting to note that when pruning beam increased from 0.001 (without pruning) to 0.1, the relative equal error rate (EER) reduction of using SCM-2 over SCM-1 is ranged from 0.9% to 3.2%. Optimal weighting factor α is increased in consistence.

Table 4 describes the EER performance of various techniques proposed in this paper. As we can see, the improved confusion matrix provides an EER reduction of up to 3.1% in relative. When entropy information is applied, EER has a 4.7% relative reduction.

5 Conclusions

In this paper, we have presented an improved keyword spotting scheme which is applied to a dialogue system. Syllable CNs are applied to extract keyword hypotheses. An improved SCM is also introduced into our keyword spotting scheme. Entropy information is integrated into posterior probability-based confidence measure to reject false accepts. Experiments show that algorithms proposed in this paper achieve 4.7% EER relative reduction.

Acknowledgements. This work is (partly) supported by Chinese 973 program (2004CB318106), National Natural Science Foundation of China (10574140, 60535030), and Beijing Municipal Science & Technology Commission (Z0005189040391).

References

1. Carlson, R., Hirschberg, J., Swerts, M.: Error Handling in Spoken Dialogue Systems. Speech Communication, pp. 207–209 (2005)
2. Akyol, A., Erdogan, H.: Filler Model Based Confidence Measures for Spoken Dialogue Systems: A Case Study for Turkish. ICASSP2004, pp. 781–784 (2004)
3. Heracleous, P., Shimizu, T.: A Novel Approach for Modeling Non-keyword Intervals in a Keyword Spotter Exploiting Acoustic Similarities of Languages. Speech Communication, pp. 373–386 (2005)

4. Higashinaka, R., et al.: Evaluating Discourse Understanding in Spoken Dialogue Systems. ACM Transactions on Speech and Language Processing, 1–18 (2004)
5. Higashinaka, R., Sudoh, K., Nakano, M.: Incorporating Discourse Features into Confidence Scoring of Intention Recognition Results in Spoken Dialogue Systems. Speech Communication, pp. 417–436 (2006)
6. Mangu, L., Brill, E., Stolcke, A.: Finding Consensus Among Words: Lattice-based Word Error Minimization. Eurospeech, pp. 495–498 (1999)
7. Moreau, N., Kim, H-K., Sikora, T.: Phonetic Confusion Based Document Expansion for Spoken Document Retrieval. ICSLP, pp. 542–545 (2004)
8. Liu, M., et al.: Mandarin Accent Adaptation Based on Context-independent/Context-dependent Pronunciation Modeling. In: Proc. ICASSP 2000, pp. 1025–1028 (2000)
9. Yi, L., Fung, P.: Modelling Pronunciation Variations in Spontaneous Mandarin Speech. ICSLP 2000, pp. 630–633 (2000)
10. Fiscus, J.G.: A Post-processing System to Yield Reduced Word Error Rates: Recognizer Output Voting Error Reduction (ROVER). In: Proceedings of IEEE ASRUWorkshop: Santa Barbara, pp. 347–352 (1997)
11. Chen, T-H., Chen, B., Wang, H-M.: On Using Entropy Information to Improve Posterior Probability-based Confidence Measures. In: International Symposium on Chinese Spoken Language Processing, pp. 454–463 (2006)
12. Xue, J., Zhao, Y.: Random Forests-based Confidence Annotation Using Novel Features from Confusion Network. ICASSP 2006, pp. 1149–1152 (2006)
13. http://www.ldc.upenn.edu/