Proceedings of the 5th World Congress on Intelligent Control and Automation, June 15-19, 2004, Hangzhou, P.R. China
第五届全球智能控制与自动化大会 2004 年 6 月 15 – 19 日，中国杭州

Robust Speech Recognition Directed by Extended Template Matching in Dialogue System

Guoliang Zhang, Hui Sun, Fang Zheng* and Wenhu Wu
Department of Computer Science & Technology
University of Tsinghua
Beijing City, 100084, China
{liang & sunh & fzheng}@cst.cs.tsinghua.edu.cn

***Abstract* - In recent years, the research on dialogue systems becomes more and more important with the ever-increasing demands. However, the automatic speech recognizer of dialogue system is not satisfying because of its bad performance for spontaneous/casual utterances. This paper presents a novel extended template matching (ETM) strategy, which imports the filler models of a keyword spotting (KWS) strategy into the template matching (TM) strategy. Because this recognition strategy not only makes use of the context information and the background knowledge by grammar template, but also adopts filler models to match extraneous speech and non-speech signals, it achieves high recognition accuracy and good robustness. The experiments show the ETM outperforms the TM and the KWS in both the reading style and the spontaneous style testing sets.**

Index Terms – Keyword spotting, ETM, Search algorithm, Speech recognition.

I. INTRODUCTION

The human-machine dialogue interface is an integration of speech recognition and language understanding technologies. Generally speaking, a human-machine spoken dialogue system consists of four modules: speech recognition, syntactic and semantic analysis, dialogue management, and response generation. Different from the in-laboratory speech system, the main goal of a dialogue system is to achieve a real-world pragmatic task, e.g. to find out a best route to a location, or to book an air ticket. Thus the user's utterances are usually in a spontaneous/casual style, often with some spontaneous phenomena, such as noise, murmur/unclear, disfluency, coughing, laughing, lip smack etc., and out-of-vocabulary (OOV) words. Though many efforts have been made to deal with the recognition of spontaneous utterance, the performance of automatic speech recognizer is still not satisfying.

There exist three kinds of speech recognition strategies that can be used in dialogue system. The first and the simplest way is the TM, in which the input utterances are explicitly restricted in the grammar template network. Because in the strategy what will be said can be predicted in advance, it will achieve good performance when the speaker's utterance matches the templates exactly, whereas, and quite bad performance when unpredicted speeches are encountered. The second is the KWS [1], which just focuses on the keyword segments so as to be unaffected by spontaneous utterance. But it cannot achieve high performance due to the lack of using any other knowledge. The third is the continuous speech recognition with a domain oriented language model built from a great deal of databases by a statistic method [2]. It has best performance among the three strategies, but its robustness is still not good because of the shortage of sufficient training corpus. In brief, the performance of the automatic speech recognizer of the dialogue system is still not satisfying.

The ETM strategy is presented in this paper that is used to deal with spontaneous utterance. It is the combination of the TM strategy and the KWS strategy, where the filler models of KWS are imported into the grammar template of TM. As a consequence, the ETM strategy has high recognition accuracy and good robustness.

II. FRAMEWORK

It is well known that the robustness of speech recognition is one of the most important factors in dialogue systems. Because no approach can restrict the user's utterances in which a lot of spontaneous phenomena, such as cough, repetition, pet phrase, and OOV often occur, robustness is the key evaluation criterion of speech recognition in dialogue systems.

The TM strategy is based on the idea that the dialogue system can probably guess what the user intend to say depending on the context information and the background knowledge of the dialogue system each time the user is about to utterance. In this strategy, a grammar template including all the sentences that the user might speak should be summarized offline by analyzing a great deal of data. And during the recognition process, the speech recognition of this strategy will be processed under the direction of the grammar template. Because the strategy supposes the user's utterance will match the grammar template exactly, whether the recognition performance of the template matching strategy is good or not depends on the speech quality. If the user's utterance matches the grammar template exactly, the strategy will achieve high recognition accuracy; on the contrary, it will be low. The KWS strategy is also widely used in dialogue systems. It just pays attention to the keywords and skips other segments because of the application of the filler model. Therefore,

* Also Beijing d-Ear Technologies Co., Ltd., *fzheng@d-Ear.com*, *http://www.d-Ear.com/*.

when there are some spontaneous phenomena in the user's utterances, it still can spot the keywords without being affected by the utterance's low quality. Though the KWS has better robustness, it has an obvious disadvantage that the recognition accuracy is lower than that of the TM while the user's utterance matches the grammar template, because KWS does not use the very important context information and the background knowledge. On the whole, we can see the advantage of the one strategy is the disadvantage of another strategy between the above two speech recognition strategies.

This paper presents an ETM strategy, in which the filler models of KWS are imported into the traditional TM to match the unexpected segment of the user's utterance. Because the filler models may be added into the grammar template more than one location with different weights, the extended grammar templates will become more complicated than the traditional grammar template. Where the filler models may be added into is up to the experts by analyzing the database similarly. In fact, the ETM is a combination of the TM and the KWS; so obviously, the speech recognition directed by the extended grammar template will lead to high recognition accuracy and good robustness, which makes the strategy attractively in dialogue systems.

Figure 1 gives some examples to compare the search network and function of the three strategies. On the left side it shows that the search network of ETM is almost the combination of the traditional grammar template and the KWS. By inserting some filler arcs, the grammar network will be turned into the extended grammar network. In addition, the examples on the right side of the figure prove that ETM has good accuracy and robustness.

The import of the filler models will not make the accuracy of TM get worse. Just as in KWS, the filler models will have a smaller weight than the other segments in the extended grammar, which ensures that the probability score of the filler models is high for the unexpected segment in user's utterance while lower in other situation.

When the extended grammar network is designed, another point that should be paid attention to is that a route which can go from the start node to the end node just along the arcs that representing the filler models must exist in the grammar network, which may ensures that the grammar includes all kinds of user's utterance, in other words, unexpected segment will never appears in the utterance for the grammar. In the worst situation that the utterance is out of the grammar totally, the ETM will degenerate as the KWS.

III. IMPLEMENTATION

A. Filler models in KWS

Now, there are two methods that can be used to implement the filler model in KWS: to use the Hidden Markov Model (HMM) to model extraneous speech exactly, or on the contrary, to use the online filler models [3] method which does not attempt to explicitly model extraneous speech or non-speech noise.

In the HMM approach, the performance to a great extent depends on the ability of the filler models to match arbitrary speech and non-speech signals without swallowing the keywords. We can explicitly train some filler HMMs to model the entire background environment and extraneous speech or make use of the models that are used to describe the lexicon keywords to define the filler models [4]. During the search process, in both the filler HMM and the keyword HMM the probability score will be calculated for each frame speech. In contrast, the online approach does not attempt to explicitly define filler models. Instead, in this approach the local filler probability scores for each time frame will be calculated directly as the average of the N best local probability score of the context independent or context dependent models used to describe the keyword models. In this way, the local score for the filler models will never be the best one but will always be in top candidates.

In contrast with the online approach, the former approach can be understood easily, but it is difficult to construct the filler HMM with perfect performance, because of complicated

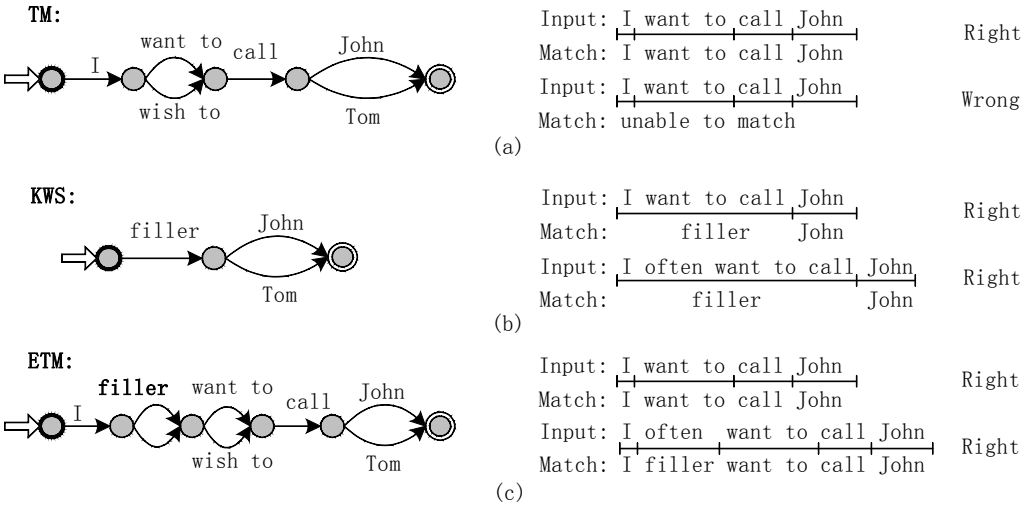

Fig. 1 Illustration for three recognition strategies' topology network and some examples.

variety of the background environment and extraneous speech. Furthermore, a bigger difference between the training environment and testing environment will lead to worse performance in the approach. Though the online approach is not easy to be understood, experiments show the performance is better than the HMM approach and the recognition speed is fast.

B. *Filler models in ETM*

Figure 2 shows the construction of the search network of the ETM; what differs from the traditional TM is the consideration of the sub network that specially describes the filler models, which can be obtained from search network of KWS directly. Similar to the KWS, there are two approaches to define the filler models: the HMM approach and the online approach. We only adopt the online filler models method in our experiments.

IV. CONFIDENCE MEASURE

The imperfect performance of the current speech recognition technology makes the research of confidence measure seem more important for speech recognition applications. The research on confidence measure (CM) mainly aims at reducing the bad effect to the dialogue system's performance caused by the misrecognition and the OOV words. Most CM techniques can be integrated into the ETM as easily as into the TM or the KWS. Here, only the acoustic level confidence measure technique is used in our experiments, because the experimental dialogue system is rather simple.

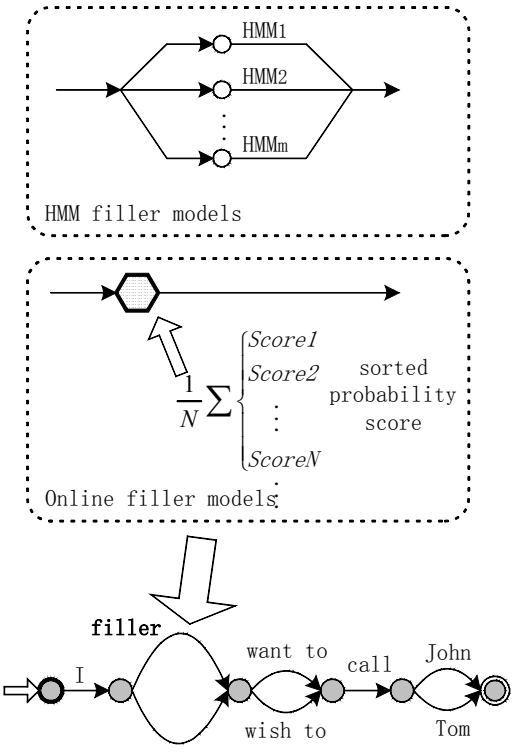

Fig. 2 Approaches to integrate the filler model into ETM.

Five acoustic confidence features are investigated in this part as follows:

- Word-level normalized probability score: frame-level normalized probability score can be calculated using:

$$C_{frame}(c_i \mid \vec{x}) = \log \frac{p(\vec{x} \mid c_i)}{p(\vec{x} \mid c^*)}. \quad (1)$$

  where $p(\vec{x} \mid c^*)$ is the score generated by the all-phone network [5]. The phone-level normalized probability score is the average of the frame-level scores of all frames in the phone while the word-level score is the average of the phone-level scores of all phones in the word.
- Minimum phone-level posterior score in the word;
- Number of active paths: the number of active dynamic path in the search space at the end of current word hypothesis;
- Number of similar hypothesis: the number of the other hypothesis with the similar location in the word graph;

In addition, the Fisher Linear Discriminate Analysis is used to train a linear discrimination project vector that is learned from training data and to produce a single word confidence score.

V. EXPERIMENTS

A. *Background*

All experiments are based on the *d-Ear Attendant*: a dialogue system like the Call Centre, in which the speaker will tell whom he wants to call on telephone, then the system recognizes the person's name and connect the speaker with the person. The system supposes the user will cooperate with the system, namely each user's utterance must contain a person's name. So only the recognition of name in utterances is considered in our experiments. Though the dialogue system is very simple, the purpose of the experiments is to evaluate the performance of speech recognition in dialogue systems, instead of that of the semantic understanding; therefore, the dialogue system is suitable to be an experiment platform. By analyzing a great deal of data, we summarize a grammar template that includes all kinds of user's utterance, which will be adopted as the template that directs the decoding in the TM strategy and the ETM strategy in the following experiments.

B. *Database*

Both the training set and the testing set are collected by telephone at an 8k sample rate. The context dependent extended initial/final set [6] are taken as the speech recognition units, with each modelled by a three-states HMM using HTK. The training set of acoustic models includes about 100k utterances belonging to about 400 speakers, and the content of training utterances are independent of the dialogue system. Four testing sets are used in our experiments; each of them includes 500 utterances and every utterance has one and only one person's name within a list of 110 person's names

totally. Whereas, the lexical style of three testing sets are completely different each other.

- Testing set I is a reading style set, complying with the grammar network summarized for *d-Ear Attendant* strictly;
- Testing set II is a spontaneous style set, the sentences in this set do not strictly comply with the grammar network, one or two spontaneous phenomena, such as noise, murmur/unclear, coughing, laughing, and OOV, are introduced into each utterance;
- Testing set III has a similar characteristic as that in the testing set II except the person name in each utterance is not in the name list.

C. Experimental results

1) Comparison without CM: Here we compare three recognition strategies: TM, KWS, ETM, without CM, however it is known that each testing utterance contains one and only one person's name. The recognition accuracy of person's name in below tables is adopted as the criterion to evaluate the performance.

TABLE I
COMPARISON OF RECOGNITION PERFORMANCE FOR DOMAIN SPECIFIC DATA WITHOUT CM

Strategies	Testing set I (%)	Testing set II (%)
TM	96.6	27.2
KWS	88.0	81.6
ETM	97.2	90.2

Table I shows that the recognition performance of TM is better than that of KWS in the situation testing utterances complying with the dialogue grammar, otherwise the TM is much worse than KWS. But ETM achieves the best performance in both cases, which proves that ETM has the advantage of high recognition accuracy and good robustness, which guarantees ETM is able to deal with the spontaneous speech in dialogue system.

Another experiment is designed to evaluate the recognition performance of ETM for other domain utterances. Testing set I and testing set II are selected as the testing databases as in the previous experiment, but the grammar template used in the ETM is replaced by a new grammar template designed for stock enquiry.

TABLE II
COMPARISON OF RECOGNITION PERFORMANCE FOR DOMAIN UNSPECIFIC DATA WITHOUT CM

Strategies	Testing set I (%)	Testing set II (%)
KWS	88.0	81.6
ETM	83.6	78.2

The experimental results in Table II show that the recognition performance of ETM is almost equivalent to KWS when the testing data are taken from another domain, which proves that the ETM will degenerate as the KWS in the worst situation that the speakers do not cooperate with the dialogue system.

2) Comparison with CM: In this experiment, we suppose each utterance includes one and only one person's name inside or outside of the name list. Now, the CM technique is integrated into the recognition strategies to confirm the recognition results. Here, the recognition accuracy and rejection accuracy are adopted as the criterion to evaluate the performance. The testing set II and the testing set III are used in the experiments altogether.

TABLE III
COMPARISON WITH CM

Strategies	Recognition accuracy (%)	Rejection accuracy (%)
KWS	77.5	65.0
ETM	85.5	61.8

From the comparison results, the rejection accuracy of the ETM strategy with CM is almost equal to KWS, but the recognition accuracy is much higher than that of the KWS. So we can draw the conclusion that the ETM with CM can do a good job in dialogue systems.

Though a relative dialogue system is adopted as the experimental platform, the spontaneous phenomena involved in the experiments are independent of the knowledge domain of the dialogue system and they will affect almost all dialogue systems. Therefore, the experimental conclusions in this paper are applicable to other dialogue systems.

VI. CONCLUSIONS

Presented in this paper is a new method named ETM, which imports the filler models of the KWS into the TM. Because in this strategy filler models are adopted to match between the extraneous speech and the non-speech signals, which ensure that the utterance out of extended grammar will never occurs, it has good robustness. Additional, filler models do not have bad effect on the recognition performance of TM, so the ETM strategy has the advantage of high recognition accuracy and good robustness, which are crucial characteristics to build a good dialogue system. Plentiful experimental results prove the advantages of the ETM.

REFERENCES

[1] F. Zheng, "Studies on Approaches of Keyword Spotting in Unconstrained Continuous Speech," *Thesis for doctor degree,* China, Tsinghua University, May 1997

[2] Y.-Y. Wang, M. Mahajan, and X. Huang, "A Unified Context-Free Grammar and N-Gram Model for Spoken Language Processing," *Proceedings of ICASSP2000,* pp. 1639-1642, Istanbul, Turkey. 2000

[3] H. Bourland, B. D'hoore, and J.-M. Boite, "Optimizing recognition and rejection performance in wordspotting systems," *Proceedings of ICASSP1994.* v.1, pp. 373-376. Adelaide, Australia, 1994

[4] R.-C. Rose, D.-B. Paul, "A hidden Markov model based keyword recognition system," *Proceedings of ICASSP1990.* v.1, pp. 129-132, Albuquerque, NM. 1990

[5] S.-R. Young, "Detecting misrecognitions and out-of-vocabulary words," *Proceedings of ICASSP1994,* v.1, pp. 21-24. Adelaide, Australia, 1994

[6] J.-Y. Zhang, F. Zheng, J. Li, "Improved Context-Dependent Acoustic Modeling for Continuous Chinese Speech Recognition," *Proceedings of EuroSpeech2001,* pp. 1617-1620. Aalborg, Danmark. 2001