Using Word Confidence Measure for OOV Words Detection in a Spontaneous Spoken Dialog System

Hui SUN, Guoliang ZHANG, Fang ZHENG, and Mingxing XU*

Center of Speech Technology, State Key Laboratory of Intelligent Technology and System
Tsinghua University, Beijing, 100084, China
[sunh, liang, fzheng, xumx]@sp.cs.tsinghua.edu.cn
* Beijing d-Ear Technologies Co., Ltd., fzheng@d-Ear.com, http://www.d-Ear.com

Abstract

Developing a real-life spoken dialogue system must face with many practical issues, where the out-of-vocabulary (OOV) words problem is one of the key difficulties. This paper presents the OOV detection mechanism based on the word confidence scoring developed for the d-Ear Attendant system, a spontaneous spoken dialogue system. In the d-Ear Attendant system, an explicit filler model is originally used to detect the presence of OOV words [1]. Although this approach has a satisfactory OOV detection rate, it badly degrades the accuracy of in-vocabulary (IV) detection by 4.4% absolutely (from 97% to 92.6%). Such the degradation will not be acceptable in a practical system. By using a few commonly used acoustic confidence features and some new context confidence features, our confidence measure method not only is able to detect the word level speech recognition errors, but also has a good ability for OOV words detection with an acceptable false alarm rate. For example, with a false rejection rate of 2.5%, the false acceptance rate of 26% is achieved.

1. Introduction

For most domain-specific dialogue systems, the presence of the out-of-vocabulary (OOV) words is inevitable since the vocabulary size is always finite. And these OOV words often lead to recognition errors. In our d-Ear Attendant dialogue system, the great majority of recognition and understanding errors is caused by such OOV words.

The OOV words detection method presented here is developed for the d-Ear Attendant system, a spontaneous spoken dialogue system. The goal of the d-Ear Attendant system is to route calls to the desired person or department. Considering the characteristic of this domain, our system only concentrates on the keywords such as person names, department names etc. Accordingly the recognition accuracy in this paper is represented by the keyword detection rate. This characteristic makes the dialogue system simple in the language understanding and dialogue management; however, it is not so easy as it seems to be, because we must face with many practical difficulties, including spontaneous speech, OOV words, and the impact of the environment, etc. What's more, this task requires higher accuracy than other spoken dialogue systems, since users are unwilling to spend more extra time "talking" with the machine, we cannot expect to improve the understanding for the user's intention by interaction. In order to effectively solve this problem, the extended grammar network [1] directed speech recognition algorithm is brought to the d-Ear Attendant system. But the result of OOV words detection is not quite satisfying yet.

A popular method for OOV words detection is to incorporate some forms of filler or garbage models to absorb such OOV words [2, 3]. In our previous work, we used a type of online filler model to detect the OOV words, achieving an OOV words detection rate of 76.5%. But this filler model degrades the accuracy of in-vocabulary (IV) data greatly. As we have mentioned above, the accuracy is so important in this task that such a degradation caused by the OOV filler model is not acceptable.

The confidence measure is useful in most applications of speech recognition, which is mainly for recognition errors rejection. Since most OOV words in the sentence are reflected by recognition errors, the confidence measure can be used to detect them. In the d-Ear Attendant system, two levels of confidence features, acoustic features and context features, are computed and combined to decide whether a word should be rejected or not. Especially, the context features proposed in this paper are shown important for confidence scoring. For the in-grammar test set, we achieve an OOV words rejection rate of 76.5% at a false rejection rate of 2.8%, while reducing the in-vocabulary detection accuracy only from 97% to 94.3%.

In the remaining part of this paper we first briefly introduce the d-Ear Attendant system and its filler model for OOV words detection. Then in section 3, the features used for confidence measure, including the acoustic features and context features, as well as the method for combining all the confidence features, are presented. A series of experiments and results are given in section 4. Finally, we conclude the paper and present several directions for the future work.

2. The d-Ear Attendant system description

The spontaneous speech phenomena and the OOV words are two kinds of key issues that the d-Ear Attendant system has to solve for practical applications. Users' spontaneous speech contains many spoken phenomena, such as lengthening, repetition, speech repair, coughing and so on. In this paper, an OOV word stands for a word class (of person name or department name) outside the definition of keywords. To deal with the spontaneous phenomena in users' speech, we use a robust speech recognition algorithm directed by the extended grammar network. It incorporates the filler model into the grammar network. Such a combination of the filler model and the grammar network improves the system in two aspects: the recognition accuracy and the robustness.

This robust speech recognition algorithm employs the online filler models to improve the robustness of the system. The filler model is integrated into the grammar network to absorb some garbage in users' sentences. In other words, the algorithm regards the average of probability scores of the

top-*n* hypothesis as the score of the filler. With such a filler model, a high accuracy can be achieved even the users speak to the machine more spontaneously. The result indicates that the filler model can be used to efficiently absorb the garbage caused by the spontaneous speech, so naturally we think such a kind of filler model can also be used to detect the OOV words. Figure 1 shows the grammar network with the filler model for OOV words.

Figure 1: Filler model for OOV words detection

We achieve an OOV words detection rate of 76% with the above filler model, and it seems satisfying for OOV words detection. Unfortunately, this method leads to a keyword detection rate reduction from 97% to 92.6% for in-vocabulary data. Considering the high requirement for the accuracy in this domain, such degradation cannot be tolerated. Thereby, another method based on the confidence measure for OOV words detection is proposed.

3. Features for confidence measure

The imperfect performance of the current speech recognition technology makes the confidence measure research important for the speech recognition applications [4]. The research on confidence measure mainly aims at reducing the negative effect to the dialogue system's performance caused by the mis-recognition and the OOV words. Two levels of confidence features, acoustic features, and context features, are used in our OOV words detection method. In the following part, we will describe these features in detail.

3.1. Acoustic features

Five acoustic confidence features are investigated, most of which have been shown to be promising [4, 5].

- Word-level normalized log-likelihood (NLL) score: the average of the phone-level normalized log-likelihood scores. The NLL score for a hypothesized phone w given acoustic observation $\vec{x}$ is:

$$c(w \mid \vec{x}) = \log \frac{p(\vec{x} \mid w)}{p(\vec{x})} \tag{1}$$

- Minimum phone-level normalized log-likelihood score in the word: this feature ensures that confidence scoring can be used to reject the keyword whose word-level normalized log-likelihood score is acceptable, but actually with a very unreliable part in it.
- Number of active paths: the number of active paths in the search space at the end of current word.
- Number of similar hypotheses: the number of the other hypotheses with the similar location in the word-graph.
- Number of similar paths: the number of the paths containing the current hypothesis in the word-graph.

3.2. Context features

It is shown that confidence features coming from language model (LM) perform better than decoder based acoustic features in spoken dialogue system [6]. In other words, the features representing the context characteristic in language level are more useful for the confidence scoring. However, in our system, the extended grammar network instead of LM is used to direct the recognition algorithm, which makes it impossible to extract the commonly used LM confidence features such as the LM back-off behavior, which have been proven helpful. But we still think the context features are important for confidence scoring, and therefore try to extract some context features different from the LM features.

- Behavior of the word ahead of the current keyword hypothesis: if the word ahead of the current keyword hypothesis is a filler or NULL.
- Acoustic confidence features for the word ahead of the current keyword hypothesis: the five acoustic confidence features for the word ahead of the current keyword.
- Behavior of the word after the current keyword hypothesis: if the word after the current keyword hypothesis is a filler or NULL.
- Acoustic confidence features for the word after the current keyword hypothesis: the five acoustic confidence features for the word after the current keyword.
- The number of fillers in the whole sentence: the total filler number in the sentence.
- The ratio of frames between all fillers and the whole sentence: the frame number of all fillers divided by the total frame number of the sentence. This feature reflects the reliability of the sentences.

3.3. Classifier

The Fisher Linear Discriminate Analysis (FLDA) is used to combine all the above confidence features. For two-class discrimination, a linear discrimination project vector is learned from the training data to produce a single word confidence score. This vector reduces the multi-dimensional confidence features down to a single one. This can be mathematically expressed as

$$c = \vec{p}^{\,T} \vec{f} \tag{2}$$

where $\vec{f}$ is the vector formed by the confidence features, $\vec{p}$ is the linear discrimination project vector for dimension reduction, and c is the resulted single confidence score. All the confidence measure experiments in the remainder of this paper adopt this FLDA approach.

4. Experiments and results

4.1. Overview

The data used in the experiments are all telephone speech. We collected over 100 calls, each of which contains several sentences that simulate the real case when people call somebody. We divide the data into four sets.

- In-grammar and in-vocabulary data set (IG_IV): the data in this set strictly accord with the grammar definition,

and each sentence must contain a person name in the keyword list.

- In-grammar and out-of-vocabulary data set (IG_OOV): this data set shares a similar characteristic to the IG_IV except that the person name contained in each sentence is not in the keyword list.
- Out-of-grammar and in-vocabulary data set (OOG_IV): the sentences in this set do not strictly comply with the grammar network that directs the decoding. There may be some filler words and spontaneous phenomena, such as lengthening, repetition, etc. But in each sentence there is a person name in the keyword list.
- Out-of-grammar and out-of-vocabulary data set (OOG_OOV): like OOG_OOV set, the sentences here do not obey the grammar either. In addition, although in each sentence there is a person name, the name is out of the keyword list.

For the confidence measure training, we use 1,000 sentences in the IG_IV set and 1,000 sentences in the IG_OOV set. And 2,000 sentences from the above four sets with each set consisting of 500 sentences form the test set. The training data and the test data are completely different.

4.2. Evaluation of confidence measure

There are two criteria to evaluate the confidence measure, the False Reject rate (FR) and the False Accept rate (FA).

$$\text{False Reject Rate} = \frac{\text{Number of rejected words that are correctly recognized}}{\text{Total number of correctly recognized words}}$$

$$\text{False Accept Rate} = \frac{\text{Number of accepted words that are misrecognized}}{\text{Total number of misrecognized words}}$$

FR and FA should be both considered for evaluating whether the confidence measure is effective. The relationship between FR and FA can be clearly shown in the Receiver Operating Characteristic (ROC) curve that can be used to reflect the performance of the confidence measure. The Equal Error Rate (EER) is another evaluation metric. An ERR point is one in the ROC curve where the False Reject rate equals the False Acceptance rate. It is generally considered that the minimum error rate is achieved at this point.

The confidence score for each keyword is compared against a threshold to decide the rejection or acceptation of the keyword. The threshold can be varied to balance the FR and FA.

5. Experiments and results

To avoid the negative effect of the spontaneous phenomena, we first do our experiment in the in-grammar data test set, including IG_IV set and IG_OOV set. In this experiment, there are two classes of recognition errors; one is of the mis-recognition errors that are small in amount occupying only 2.9%, while another is caused by the OOV words each of which is falsely recognized as a certain keyword.

In Figure 2, the results of the confidence measure for in-grammar test set are given with two ROC curves. The dotted curve shows the result of the word confidence scoring when only using five acoustic confidence features, and the solid one gives the result when applying all the confidence features. By using all the confidence measure features, 71% of recognition errors can be correctly detected while only 2.5% of correctly recognized keywords incorrectly rejected.

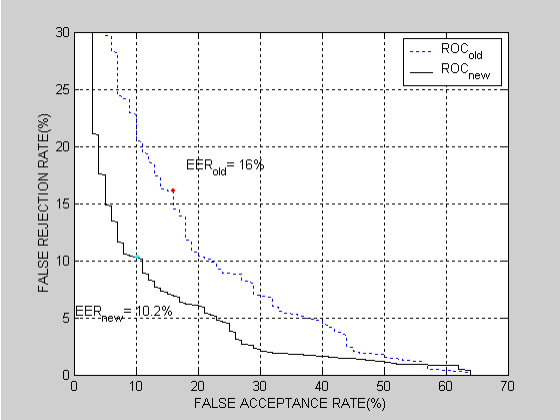

Figure 2: ROC curve on the in-grammar test set

This result indicates that the context features are helpful for confidence measure. It greatly reduces the false acceptance rate at a certain false rejection rate. For example, the false acceptance rate reduction is 41% when the false rejection rate remains at 5% (See table 1).

Table 1: Comparison of confidence measure results using different confidence features

Confidence features	**False Acceptance rate (FA)**	
	FR =2.5%	**FR = 5%**
Acoustic features	45%	39%
Acoustic features + context features	29%	23%
FA Reduction	**35.6%**	**41%**

Considering only the errors caused by OOV words, the confidence scoring method outperforms the filler model described in section 2. For in-grammar test set, our confidence measure method degrades the performance of in-vocabulary data less than the filler model, from 97% to 94.3% (see table 2).

Table 2: Accuracy of IG_IV test when using different OOV words detection mechanism (the original accuracy is 97%)

	Accuracy of IG_IV set (OOV words detection rate =76.5%)
Filler Model	92.6%
Confidence Measure	94.3%

5.1. Experimental result on all test sets

The d-Ear Attendant System would inevitably encounter some spoken phenomena, such as coughing, repetition, speech repair, and so on. So we have to evaluate our word confidence scoring across the out-of-grammar test set.

Figure 3 shows the result of confidence measure applied to the test set, including the in-grammar set and the out-of-grammar set. Compared with the above experiments, this result is not so good; context features seem little useful for out-of-grammar data. After analyzing carefully, we can see there are two main reasons leading to the result. Firstly, the test data do not accord with the training data because the out-of-grammar data in the test set are quite different from the in-grammar data. Secondly, the characteristics of the out-of-grammar data determine that context features are not so helpful for the confidence scoring because the fillers in the sentence disconnect the keyword and other words.

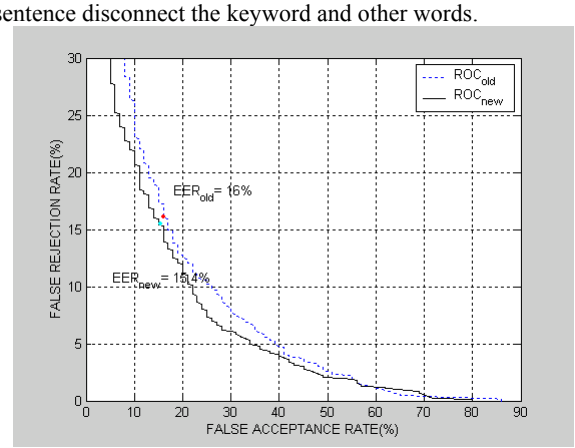

Figure 3: ROC curve on all the test data

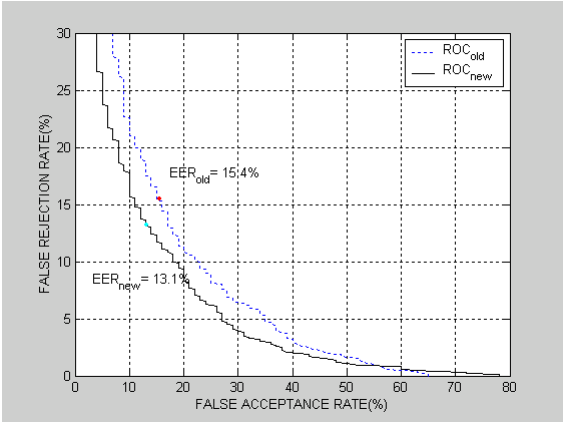

Figure 4: ROC curve after adding some out-grammar data to the training set

For the first problem, we collect another 100 OOG_IV -like sentences and another 100 OOG_OOV-like sentences, and add them to the training set. Figure 4 shows the ROC curves after enriching the training data. It can be clearly found that the performance increases to a certain extent. In fact, the confidence measure performance across the in-grammar data is also slightly improved by enriching the training set. At the false rejection rate of 2.5% the false acceptance rate is reduced from 29% to 26%, while the false acceptance rate from 23% to 17.5% at a false rejection rate of 5% (see table 3).

Table 3: False Acceptance Rate of test data after enriching the training data

	IG Data		**IG+OOG Data**	
	FR =2.5%	**FR = 5%**	**FR =2.5%**	**FR =5%**
Acoustic features	45%	39%	43%	35%
Acoustic features + context features	26%	17.5%	38%	27%

6. Conclusion and future work

A mechanism of word confidence scoring is studied and tested for OOV words detection. Some new context confidence features suitable for the d-Ear Attendant system have been proposed. Compared with the OOV filler model, the proposed word-level confidence scoring mechanism achieves higher accuracy for the in-vocabulary data with the same OOV words detection rate.

But there are still some issues having not been successfully solved. For example, the confidence measure performance is not so good for out-of-grammar data. In the future, we will try to find some new confidence features for the out-of-grammar data. Additionally, methods on how to build a more effective filler model for OOV words detection and how to combine the OOV filler model with word confidence scoring mechanism will also be focused on.

7. Acknowledgements

The authors would like to thank LI Jing for his work on data collection and his kindly help.

8. References

[1] Zhang, G.-L., Ling, J., Zheng, F., and Wu, W.-H., "Robust speech recognition algorithm directed by extended grammar network in dialogue system," submitted to *the Journal of Tsinghua University* (*Science and Technology*).

[2] Bazzi, I. and Glass, J., "Modeling Out-of-vocabulary Words for Robust Speech Recognition," *Proc. ICSLP,* Beijing, 2000.

[3] Manos, A. and Zue, V., "A Segment-based Spotter Using Phonetic Filler Models," *Proc. ICASSP,* Munich, 899-902,1997.

[4] Hazen, T. J., Seneff, S. and Polifroni, J., "Recognition confidence scoring and its use in speech understanding systems," *Computer Speech and Language (2002) 16, 49-67.*

[5] Kamppari, S. and Hazen, T.J., "Word and Phone Level Acoustic Confidence Scoring," *Proc ICASSP2000*, Istanbul, Turkey, June 2000

[6] San-Segundo, R., Pellom, B., Hacioglu, K., Ward, W., and Pardo, J.M., "Confidence measures for Spoken Dialogue Systems," *Proc ICASSP2001,* Salt Lake City, USA, 2001.