An Effective and Efficient Utterance Verification Technology Using Word N-gram Filler Models

Dong Yu, Yun Cheng Ju, Alex Acero

Microsoft Research
One Microsoft Way, Redmond, WA 98052, USA
{dongyu, yuncj, alexac}@microsoft.com

Abstract

In this paper we propose a novel, effective, and efficient utterance verification (UV) technology for access control in the interactive voice response (IVR) systems. The key of our approach is to construct a context-free grammar by using the secret answer to a question and a word N-gram based filler model. The N-gram filler provides rich alternatives to the secret answer and can potentially improve the accuracy of the UV task. It can also absorb carrier words used by callers and thus can improve the robustness. We also propose using a predictor based on the best alternative to calculate the confidence. We show detailed experimental results on a tough UV test set that contains 930 positive and 930 negative cases and discuss types of questions that are suitable for the UV task. We demonstrate that our approach can achieve a 2.14% equal error rate (EER) on average and 0.8% false accept rate if the false reject rate is 2.6% and above. This is a 49% EER reduction compared with the approaches using acoustic fillers, and a 72% EER reduction compared with the posterior probability based confidence measurement.

Index Terms: utterance verification, filler model, word spotting, confidence measure

1. Introduction

Due to the recent progresses in automatic speech recognition (ASR) and dialog management technologies, we see a steady increase in the adoption of the interactive voice response (IVR) systems. In some IVR systems it is required to control the callers' access to sensitive information. For example, in a password reset IVR application, the system needs to verify the caller before it grants the permission to reset the password. Two popular categories of approaches exist to verify a caller: by what he or she owns and by what he or she knows. The speaker identification and verification technology belongs to the first category. The utterance verification (UV) technology, which is the focus of this paper, belongs to the second category.

In the UV task, the system knows the caller's secret answers (in text format) to a set of questions. The system verifies the caller by checking whether the caller has answered these questions (in audio format) correctly. The UV technology differentiates itself from the speaker verification technology in that it does not require callers' spoken samples which require a special enrollment process to get. Even if spoken samples are available, the UV technology can be a complement to the speaker verification technology to increase the accuracy of the caller verification.

Note that two sources of possible errors exist in UV. The first source is the answer itself. The caller may forget some of the answers or the answers may be known by an attacker. The second source lies in the verification technology. The system may interpret a correct spoken answer as wrong (false rejection) or interpret an incorrect answer as correct (false acceptance.) Our focus in this paper is to reduce the errors caused by the second source.

The task of UV has been formulated as a hypothesis test problem and a confidence measure problem in the past. (See Section 2 for more information.) Many approaches have been proposed to solve it [1][2][3][4][5][6][7]. However, these approaches are typically expensive and/or not effective. Some of these approaches do not address phenomena seen in the real UV task and/or are not tested under real condition. For example, in the real UV task, callers may have strong accents and may embed their answers in carrier phrases, e.g., "Em, Seattle," "It's Seattle." or "Seattle, that's right." Their responses may be corrupted by the background noises and may contain words out of the ASR vocabulary, such as those from foreign languages.

In this paper, we propose an effective and efficient UV technology. The novelty of our approach comes from two components. First, we dynamically construct a probabilistic context-free grammar (PCFG) using the secret answer to a question and a word N-gram based filler model. The word N-gram filler provides rich alternatives to the secret answer and can potentially improve the accuracy of the UV task. It can also absorb carrier words used by callers and thus can improve the robustness. Second, we propose using a best-alternative based predictor to calculate the confidence. This confidence algorithm can provide reliable confidence scores without sacrificing the decoding speed. We demonstrate that our approach performs well on a tough UV test set that contains 930 positive and 930 negative cases, and discuss some guidelines on choosing good questions for the UV task.

The remainder of the paper is organized as follows. In Section 2, we introduce the existing UV formulations and algorithms. In Section 3, we describe our UV algorithm, which includes constructing a PCFG by using a generic word N-gram filler model and a best-alternative based confidence measurement. We show detailed experimental results and compare different approaches in Section 4, and conclude the paper in Section 5.

2. Formulations of the UV Task

The UV task can be stated as follows. Given a text transcript w and an utterance o, determine whether o is a presentation of w subject to a cost function f. This problem has been formulated in two different ways.

In its first form, the UV problem is considered as a statistical hypothesis test problem [1][2] with the following null hypothesis and alternative hypothesis:

$H_0 : o$ is a presentation of w

$H_a : o$ is a presentation of something other than w

The task of UV is thus to gather evidence to either accept or reject the null hypothesis. The typical solution to this problem is based on a log likelihood ratio (LLR) testing. In other words, given

$$LLR = \log\left(\Pr\left(H_0 \mid o, w\right)\right) - \log\left(\Pr\left(H_a \mid o, w\right)\right), \qquad (1)$$

the system should accept the null hypothesis if $LLR > \theta$ and reject the null hypothesis otherwise. The threshold θ is determined by the cost function f. The main difficulty with LLR based approach is how to model the alternative hypothesis. In [1][2], anti-models with the same hidden markov model (HMM) structure are adopted for this purpose. However, anti-models need to be trained for each UV task.

In its second form, the UV problem is formulated as a confidence measure problem, and the decision can be made based upon the confidence scores. A widely explored approach is to use the posterior probability of the ASR output as the confidence, i.e.

$$\Pr\left(w \mid o\right) = \frac{\Pr\left(o \mid w\right)\Pr\left(w\right)}{\Pr\left(o\right)} = \frac{\Pr\left(o \mid w\right)\Pr\left(w\right)}{\sum_{w_i} \Pr\left(o \mid w_i\right)\Pr\left(w_i\right)}. \qquad (2)$$

The key to the success of this approach is the accurate estimation of the denominator in (2). Due to the difficulty of estimating the distribution of possible phrases $\Pr\left(w_i\right)$, heuristic methods are used to approximate it. For example, the distribution can be estimated by using a set of general filler models, i.e., all-phone recognition [3], catch-all model [4], etc. Or, it can be estimated from a word lattice based on the forward-backward algorithm [5][6]. The lattice-based approach usually provides the best result if the lattice is rich enough. However, having a rich lattice means keeping more paths in the search space and being less efficient. Besides the posterior-probability-based confidence measures, many people have also proposed using predictors to derive a confidence score from features such as acoustic stability [7], hypothesis density [5], duration, and many others. A good survey on this area can be found in [9].

These two formulations are highly correlated. The main difference between them is whether all the alternatives or all the possible paths are used to compare with the key phrase to be verified. Please note that these approaches usually do not handle well the issues seen in the real UV tasks mentioned in Section 1.

3. An Effective and Efficient Approach

A good UV system should meet the following requirements. First, it can provide enough competitors to the answers (or key phrases) to make estimations of the LLR or confidence measures accurate. Second, it needs the ability to filter out unrelated carrier words and background noises. Third, the confidence measurement should be reliable and consistent under different pruning strategies and operation environments (e.g. different background of users.) In this section, we address these requirements with a word N-gram based filler and a best-alternative based confidence calculation.

3.1. PCFG Using a Word N-gram Filler

To verify an utterance, our system dynamically generates a PCFG using the answer (or key phrase) and a generic word N-gram filler. The PCFG is then used as the language model (LM) for the ASR engine to recognize the utterance. There are two ways to construct the PCFG. Figures 1 and 2 depict grammars constructed using a parallel structure and a sequential structure, respectively. In these grammars, p_1 and p_2 are weights associated with each branch and are usually set to 0.5-0.8 (not sensitive.)

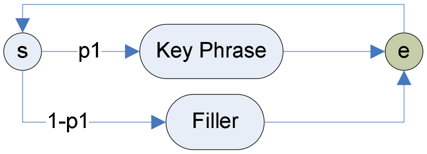

Figure 1. *Parallel Structure*

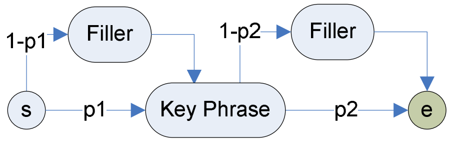

Figure 2. *Sequential Structure*

The key of our approach is to use the generic sharable word N-gram filler (a PCFG) generated by using the approaches described in [8]. The filler used in our system was trained by using the Wall Street Journal data. The word N-gram filler can provide rich competitors to the key phrase. It can also absorb the carriers used by the callers. Compared with the acoustic-based filler models, the word N-gram filler allows the understanding components to determine whether the carrier words are legitimate, in other words, to detect a grouping attack that the caller says things like “Seattle Boston Dallas” as the answer to the question “*What’s your favorite city?”* to trick the system.

One benefit of this approach is its independence to the ASR engine. As long as the ASR engine can support PCFG, this approach can be used. Because no rigorous hypothesis testing is needed during the decoding process, our approach can rely on the pruning of the ASR engine to make the accept/reject decision with high accuracy. The sequential construction tends to reject more [8] and is more suitable to clean conditions where the false accept rate dominates the error, while the parallel construction is more suitable to the conditions where the false reject rate dominates the error.

3.2. The Best-Alternative Based Confidence Measurement

The lattice-based posterior probability has been known to provide the best confidence if the lattice is rich enough. However, keeping a rich lattice usually means low efficiency. If aggressive pruning is applied, as in many commercial ASR systems, the posterior probability estimated over the lattice becomes very unreliable. For this reason, we propose using the best-alternative based confidence measurement that is based on the predictor

$$c = 0.5 - 0.498 \tanh\left(\rho \mathbf{x}^{\mathrm{T}} \mathbf{G} \mathbf{x} + \gamma\right), \qquad (3)$$

where ρ and γ control the balance of the confidence score, and G is a 4x4 matrix. Note that

$$\begin{aligned}\mathbf{x} &= \left[x_0\ \ x_1\ \ x_2\ \ x_3\right] \\ &= \left[1\ \ (bs\text{ - }ac)\ \ (bs\text{ - }bk)\ \ (sc\text{ - }se)\right]\end{aligned} \qquad (4)$$

is a feature vector constructed based on ac (the acoustic log likelihood of the key phrase path), bk (the acoustic log likelihood of the background model, sc (the total utterance log likelihood of the key phrase path, se (the total utterance log likelihood of the best alternative, and bs (the best log likelihood among all paths.) When the feature vector is derived, the segments corresponding to the silences are removed from the calculation. If the key phrase is not in the recognition result the confidence is set to -1. The confidence score using (3) can be estimated on the phrase level or the word level. In our system we calculate the word confidence and derive the phrase confidence in two ways: use the average word confidence and use the minimum word confidence.

The parameters ρ and γ can be predetermined. Let us define the false reject rate (FRR) as

$$FRR = \frac{\#FR}{\#T}, \qquad (5)$$

and the false accept rate (FAR) as

$$FAR = \frac{\#FA}{\#F}, \qquad (6)$$

where $\#T$ and $\#F$ are the number of total positive cases and total negative cases, respectively; $\#FR$ is the number of positive cases that are rejected by the UV system; and $\#FA$ is the number of negative cases that are accepted by the UV system. G can be trained by using a generic training set (not specific to a UV task) to fit the following definition of the confidence.

$$conf = \frac{FRR}{FRR + \rho \times FAR} = \frac{1}{1 + \rho \times FAR/FRR}. \qquad (7)$$

4. Experimental Results

Evaluations have been conducted on a tough UV test set which contains 930 positive cases and 930 negative cases. This test set was collected in Microsoft. The callers range from software developers, program managers, testers, and receptionists. The vocabulary is open. The distribution of the positive (as well as the negative) cases across different questions is shown in Table 1. These utterances cover a great variety of accents (e.g. American, British, Canadian, Indian, and Chinese), styles of speaking (e.g. speed, softness, with and without carrier words), audio channels (e.g. land phone and mobile phone), and noise conditions.

Table 1: *Distribution of Utterances for Different Questions*

Question	# Utterances
What's your mother's maiden name?	119
What's the name of your favorite pet?	116
What's your favorite food?	117
What's your favorite restaurant?	112
What's the title of your favorite movie?	118
What's your favorite city?	121
What's your favorite radio station?	112
Who is your favorite movie star?	115
Total	**930**

We compare the effectiveness of different approaches using the equal error rate (EER)

$$EER = FRR_e = FAR_e. \qquad (8)$$

EER is corresponding to the FRR or FAR where they match.

Table 2 summarizes the performance of eight different approaches in EER, where

- N-gram+Parallel+BestAlt: Use the word N-gram filler in the parallel form, and then use the best-alternative based average word confidence.
- N-gram+Sequencial+BestAlt: Use the word N-gram filler in the sequential form, and then use the best-alternative based average word confidence.
- N-gram+Parallel+BestAlt+Min: The same as N-gram+Parallel+BestAlt except that the minimum word confidence is used.
- N-gram+Sequencial+BestAlt+Min: The same as N-gram+Sequencial+BestAlt except that the minimum word confidence is used.
- Acoustic+Parallel+BestAlt: Use the acoustic filler in the parallel form, and then use the best-alternative based average word confidence.
- Acoustic+Sequencial+BestAlt: Use the acoustic filler in the sequential form, and then use the best-alternative based average word confidence.
- N-gram+Parallel+PosteriorProb: Use the word N-gram filler in the parallel form, and then use the lattice posterior probability based confidence.
- Acoustic+Parallel+PosteriorProb: Use the acoustic filler in the parallel form, and then use the lattice posterior probability based confidence.

Table 2: *Comparison of Different Approaches*

Method	EER
N-gram+Parallel+BestAlt	2.14%
N-gram+Sequencial+ BestAlt	2.22%
N-gram+Parallel+ BestAlt +Min	3.38%
Acoustic+Parallel+BestAlt	4.18%
Acoustic+Sequencial+BestAlt	4.50%
N-gram+Sequencial+BestAlt+Min	5.25%
N-gram+Parallel+PosteriorProb	7.52%
Acoustic+Parallel+PosteriorProb	19.43%

We have four observations from Table 2. First, the word N-gram based filler model outperforms the acoustic filler because the word N-gram filler model provides much richer competitors to the key phrase to be verified. Using the word N-gram filler, we can achieve 2.14% EER, which is 49% less error than the best EER using the acoustic filler. Second, the best-alternative based confidence score is much more reliable than the lattice-derived posterior probability based confidence. This is mainly due to the levity of the lattice quality after the pruning. We have also noticed that for some questions, the EER point is not reachable if the posterior probability is used. Third, using the N-gram filler in the parallel mode gives us a slight gain over that in the sequential mode. This is because the false reject rate dominates the errors in this test set (Figure 3) and the sequential model tends to reject more [8] (i.e. increase the false reject rate even more.) In fact, the sequential model outperforms the parallel model when tested on a clean UV test set (not discussed in this paper.) Fourth, making decisions based on the minimum word confidence score decreases the EER. Using the minimum word confidence as the phrase confidence essentially means that the UV system needs to be confident on each word to accept the phrase. In other words, it tends to reject more than the approaches using the average word confidence as the phrase confidence. Since the false rejection rate dominates this data set, using the minimum word confidence does not help.

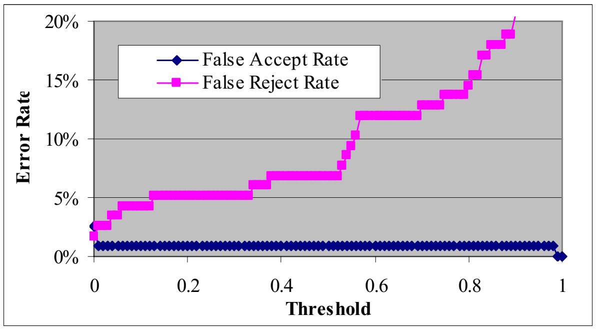

Figure 3. *The distribution of false reject rate and false accept rate against the threshold using the best approach*

Figure 3 shows the distribution of the FRR and the FAR against the threshold using the best approach. It's obvious that the FRR dominates the test set. Note that the users of the UV system care about the FAR more. In our system, we can achieve the FAR of 0.8% if we sacrifice the FRR a little bit to 2.6%, which is corresponding to the threshold of 0.05. Also notice that these results are based on a single question dialog. If two or more questions are used to verify a single caller, we can achieve even better results. We have performed error analysis and noticed that the falsely rejected utterances are mainly from two categories: incorrect pronunciations, especially if the phrase is a foreign name and/or the caller is a foreigner, and bad audio channel. For example, the volume is extremely low sometimes. The falsely accepted utterances are mainly confusable words such as "Jack" and "Jackie."

Table 3: *Comparison of Different Questions*

Question	EER
What's your favorite food?	0.85%
What's your favorite radio station?	1.34%
What's the name of your favorite pet?	1.71%
What's your favorite city?	1.79%
Who is your favorite movie star?	2.17%
What's your mother's maiden name?	2.48%
What's the title of your favorite movie?	3.02%
What's your favorite restaurant?	3.78%

Table 3 compares different questions in EER. According to the table, the question on the favorite food is most suitable for the UV task. The favorite restaurant performs worst. In general, if the answers to the question are likely to be well known English words, the question is good for the UV task. If the answers contain large percent of foreign words, it's not suitable for the UV task. Let's examine a special case – the favorite radio station question. Since radio stations mainly contain abbreviations and numbers, it is listed in Table 3 as the second best question. We want to point out that this result is under the assumption that you have the normalized text answer. If text normalization is not performed right, the EER is as high as 7.59% for this question. This suggests that we prefer a question without normalization over the questions that require normalization.

5. Summary and Conclusions

This work is motivated by the requirement to secure the access to the sensitive information in IVR systems. We described an effective and efficient approach to verify the caller by verifying whether the caller can answer some questions correctly. Our novel approach has two major components: a PCFG constructed using the answer to the question and a generic sharable word N-gram filler, and a best-alternative based confidence calculation algorithm. The purpose of using the word N-gram filler is to provide rich competitors to the answer so that a reliable confidence can be derived. The purpose of using the best-alternative based confidence is to get a robust confidence score even if heavy pruning is conducted by the ASR.

Our experimental results on a tough UV test set demonstrate that the word N-gram filler outperforms the acoustic filler, and the best-alternative based confidence outperforms the posterior probability derived from a heavily pruned lattice. We can achieve a 2.14% EER on average on the test set, and 0.8% of FAR if the FRR is 2.6% and above. Our experiments also indicate that it's preferable to choose questions that do not require normalization and do not contain many foreign words.

6. Acknowledgement

We thank David Ollason, Tal Saraf, and Larry Ockene in Microsoft Speech Server group for helping us collect the data.

7. References

[1] R. Sukkar and C.-H. Lee, "*Vocabulary Independent Discriminative Utterance Verification for Non-keyword Rejection in Subword Based Speech Recognition*," IEEE Trans. Speech Audio Processing, vol. 4, Nov. 1996.

[2] M. G. Rahim, C.-H. Lee, and B.-H. Juang, "*Discriminant Utterance Verification for Connected Digits Recognition*," IEEE Trans. Speech Audio Processing, vol. 5, May 1997.

[3] S. Young, "*Detecting Misrecognitions and Out-of-Vocabulary Words*," in Proc. ICASSP-94, Adelaide, Australia, Apr. 1994, pp. II-21–II-24.

[4] S. Kamppari and T. Hazen, "*Word and Phone Level Acoustic Confidence Scoring*," in Proc. ICASSP-2000, 2000, pp. 1799–1820.

[5] T. Kemp and T. Schaaf, "*Estimating Confidence Using Word Lattices*," in Proc. EuroSpeech-97, 1997, pp. 827–830.

[6] F. Wessel, R. Schluter, K. Macherey, and H. Ney, "*Confidence Measures for Large Vocabulary Continuous Speech Recognition*," IEEE Trans. Speech Audio Processing, vol. 9, Mar. 2001.

[7] M. Finke, T. Zeppenfeld, M. Maier, L. Mayfield, K. Ries, P. Zhan, J. Lafferty, and A. Waibel, "*Switchboard April 1996 Evaluation Report*," DARPA, Apr. 1996.

[8] D. Yu, Y. C. Ju, Y.-Y. Wang, A. Acero, "*N-gram Based Filler Model for Robust Grammar Authoring,*" ICASSP 2006 (to appear).

[9] H. Jiang, "*Confidence Measures for Speech Recognition: A Survey,*" Speech Communication, Vol. 45, No. 4, pp. 455-470, Apr. 2005.