ABSTRACT

Many existing search algorithms aim at searching for the best hypothesis from all possible hypotheses with the help of techniques like beam search to reduce the computational cost. These search algorithms are based on the competitive criteria because the best hypothesis is determined after we have the knowledge of all other possible hypotheses. In this paper, we investigate the possible use of a qualifier criterion. A qualifier-based search will make recognition decision without considering all possible hypotheses. Instead of finding the best hypothesis, the system will try to find a hypothesis that can meet with some criteria reflecting the goodness of a hypothesis. We investigate the use of utterance verification measurement as the qualifier measurement used in the search. An utterance verification based speech recognition (UVSR) system is proposed. The search will be terminated as soon as the qualifier criteria are meet without the need to evaluate the likelihoods of all possible hypotheses. The proposed search algorithm with UV performs very closely to the standard speech recognition with UV but with 23% less in the average searching time. If the information of language model is provided, the average searching time can be reduced by 43%.

1. INTRODUCTION

Many efficient search algorithms, such as viterbi search and stack search, have been developed to find the best hypothesis from all possible hypotheses. Some techniques, such as beam search, are applied to reduce the computational cost. These search algorithms are based on the competitive criteria because the best hypothesis is determined after we have the knowledge of all other possible hypotheses.

In this paper, we investigate the possible use of qualifier criteria [7]. In a qualifier-based search, the recognition decision will be made when a hypothesis that meets with some qualifier criteria has been found without having the complete knowledge of all other possible hypotheses. The qualifier-based search will find a good hypothesis instead of the best hypothesis.

We investigate the possible use of the utterance verification measurements as the qualifier criteria. Utterance verification (UV) is a process by which the keyword hypothesis produced by a speech recognizer is verified to determine whether the input speech does indeed contain the recognized strings. The UV techniques have been applied successfully to reject erroneous recognition events and out-of-vocabulary words. In the speaker verification task, very low equal error rate can be achieved by applying the similar techniques.

We present an utterance verification based speech recognition system (UVSR) [5][6]. The UVSR system will try to find a hypothesis with a high confidence level that is determined by utterance verification techniques. If the utterance verification measurements are reliable, the hypothesis with high confidence level may just be the best hypothesis. The system is particularly useful if we can perform a rough preliminary analysis that can generate a list of N-best possible hypotheses according to some likelihood measure using fast pattern matching technique or the information extracted from the language model. With this extra information, the qualifier search can terminate very quickly. A potential application of the UVSR system will be in a multipass system for large vocabulary isolated word recognition where a list of possible hypothesis has been sorted [3][4].

Two utterance verification (UV) measurements are used to determine the confidence level of the hypothesis. The all-phone garbage model based UV measurement is used to guide the search. The N-best likelihood ratio based UV measurement is used to reject the erroneous recognition events. Multiple confidence levels and language model are introduced to improve the system performance. For simplicity, we will present the ideas of the UVSR system in an isolated word recognition task that is evaluated on a field database collected in the telephone environment.

2. UV BASED SPEECH RECOGNITION (UVSR)

Suppose we have a vocabulary of \( M \) words to be recognized and for each word \( m \) in the vocabulary, we build a hidden Markov Model (HMM) \( \lambda_m \). The observation sequence, \( O = \{o_1, o_2, \ldots, o_T\} \), is used to calculate the model likelihood for all models in the vocabulary, \( P(O|\lambda_m) \), for \( 1 \leq m \leq M \). In a standard isolated word recognition, the highest likelihood among the keyword models is selected as the recognized word – that is,

\[
m^* = \arg \max_{m \in M} P(O|\lambda_m) \tag{1}
\]
The probability is evaluated via the Viterbi algorithm and requires on the order of $M \times N^2 \times T$ computations, where $N$ is the number of states of the word model.

### 2.1 UVSR I

The block diagram of the UVSR system is shown in Figure 1. The recognition decision is performed in a model-by-model manner. The list of models, $\{\lambda_1, \lambda_2, ..., \lambda_M\}$, can be sorted in a multipass system or using the prior word probability. To make the problem more challenging, we will assume that a randomly ordered model list is given. For a randomly ordered model list, each model is equal likely to be the correct hypothesis.

The calculation of model likelihood for $m^{th}$ word model is followed by computation of garbage model based UV score $G(O,m)$ using free grammar phone models running parallel with the keyword models,

$$G(O,m) = \frac{\log(P(O|\lambda_m)) - \log(P(O|\lambda_g))}{\text{number of frames}},$$  \hspace{1cm} (2)

where $\lambda_g$ is the all-phone garbage models and $\lambda_m$ is the $m^{th}$ keyword model.

![Figure 1](image)

**Figure 1** Utterance verification based speech recognition system

The UV test is performed as,

$$C(O,m) = \begin{cases} 1, & \text{if } G(O,m) \geq \tau \\ 0, & \text{otherwise} \end{cases}$$  \hspace{1cm} (3)

where $\tau$ is a threshold for GM based UV score. A hypothesis passes the UV test if the $C(O,m)$ is equal to 1. The computation will be terminated and the recognized word will be sent to the user as soon as the system found a hypothesis that passes the UV test. Otherwise, the computation will proceed to $m+1^{st}$ word model. If none of the models pass the UV test, the input utterance will be rejected. The search is based on a qualifier criterion because the decision is made without the knowledge of all possible hypotheses.

### 2.2 UVSR II

It is clear that the order of the models plays an important role in the performance of UVSR II system. For a given observation sequence we may have many keyword models that can pass the UV test depending on the threshold $\tau$ for the UV scores. The recognized word may strongly depend on which model comes first in the model list. The UV test can only tell us whether the current hypothesis is qualified as a good hypothesis but it does not guarantee that this is the best hypothesis, which is usually the correct hypothesis.

To alleviate the above problem, the UVSR II algorithm searches for more than one hypothesis that passes the UV test. The model with highest UV score is then selected as the recognized word – that is

$$m^* = \arg \max_{1 \leq C \leq P \leq M} G(O,m)$$  \hspace{1cm} (4)

where $P \leq M$. The $P$ is determined dynamically so that there are $V_t$ keyword models that pass the UV test from the partial model list $\{\lambda_{k_1}, \lambda_{k_2}, ..., \lambda_{k_{V_t}}\}$, where $V_t \geq 1$.

When the threshold $\tau$ is very high, $V_t$ is set to 1 because the hypothesis that passes the UV test is very likely to be the best hypothesis. When the threshold $\tau$ is low, $V_t$ is set to greater than 1 because more than one keyword models can pass the UV test. This method allows the system to search further when the confidence level is low.

The system performance can be further improved when multiple thresholds are considered. The UV scores can be classified into several confidence levels. Each confidence level is defined by three parameters,

$$\gamma = (\tau, V, \lambda_{\text{top}}),$$  \hspace{1cm} (5)

where $\lambda_{\text{top}}$ is the model of the highest score at confidence level $\gamma$. $V$ is the number of UV-passed hypotheses required to terminate the process and $\tau$ is the threshold for GM based UV score. The system will terminate the computation at confidence level $\gamma$ if there are $V_t$ keyword models whose GM based UV score is greater than the threshold $\tau$. The computation can be terminated at any of these confidence levels.

It was found that in a voice-activated dialing (VAD) system the N-best likelihood ratio (NBLR) based UV score is better than the GM based UV score to reject the mis-recognized event [1]. The NBLR based UV score for hypothesis $k$ is the log of the likelihood ratio of candidate $k$ and $k+1$ in the N-best list and is represented by [2]

$$R(O,k) = \frac{\log(P(O|\lambda_k)) - \log(P(O|\lambda_{k+1}))}{\text{number of frames}},$$  \hspace{1cm} (6)
where $\lambda_k$ and $\lambda_{k+1}$ is respectively the $k^{th}$ and $k+1^{st}$ hypotheses from the $N$-best list produced by sorting the scores of all keyword models.

If the system terminates at confidence level $\gamma$ and $V_\gamma$ is greater than one, we can apply the NBLR based UV score to perform a second UV test. The parameters at each confidence level is redefined as

$$\gamma = (\tau_1, V_\gamma, \lambda_{\text{top}}, \lambda_{\text{2nd}}),$$

where $\lambda_{\text{2nd}}$ is the model of the second highest score at confidence level $\gamma$. The NBLR based UV test aims at rejecting misrecognized events and helps to reduce the false rejection rate. The GM based UV score is used to decide when the correct hypothesis may have been found. The algorithm of UVSRI is summarized as follows:

1. Initialize $m = 1$.
2. Compute HMM likelihood $P(O|\mu_m)$ for word model $m$.
3. Compute $G(O,m)$ and update the parameters of each confidence level.
4. If the computation cannot be terminated at any confidence level, increment $m$ and go to Step (2).
5. If the system terminates at confidence level $\gamma$ and $V_\gamma > 1$, the NBLR based UV test using score $R(O,1)$ is performed.
6. Output the recognition result.

### 3. EXPERIMENTS

The speech corpus used for evaluating the isolated word recognition system is a database of names and other utterances collected between June 1995 and January 1997 from the live Operetta system in Vocalis. This database contains 35060 utterances. We design a vocabulary set consists of 3000 human names.

Continuous density left-to-right HMMs with 3 states were used to model each phone in the recognizer. The observation densities were mixtures of 10 multivariate Gaussian distributions with diagonal covariance matrices. The feature vector consists of 12 cepstral coefficients and 12 delta cepstral coefficients.

#### 3.1 Performance of UV and UVSRI

The baseline performance of the UVSRI system is shown in Figure 2. The UV_GM and UV_NBLR are the standard isolated word recognition systems with UV using garbage model and NBLR method, respectively. The error rate of standard isolated word recognition is 17.07%. With the application of UV techniques using the GM and NBLR methods, the error rate can be respectively reduced by 30% and 60% with about 8% rejection of correctly recognized keyword. The NBLR based UV scores (UV_NBLR) out-performs the GM based UV scores (UV_GM) for rejecting erroneous events.

![Figure 2 Baseline Performance of UVSRI](image-url)

We have purposely created the order of the model list for each test utterance to represent two extreme cases. The UVSR_BEST is the best case assuming that the correct hypothesis for each input utterance is always at the top of an ordered model list. On the other hand, the UVSR_WORST is the worst case assuming that the correct candidate always appears at the end of the ordered model list. These two cases allow us to estimate the limit of the UVSR algorithm.

The UVSRI is evaluated on a randomly ordered model list. This is a more difficult task than a sorted model list. The UVSRI system performs better than the worst case but it is still not a feasible solution. If we rearrange the ordered model list according to the word prior probability which is estimated directly from the database, the performance (UVSRI_WP) becomes closer to the performance of the best case. These results show that the UVSRI algorithm is sensitive to the order of the models and the information extracted from the language model can be very useful. A sorted model list generated from a multipass system [3][4] is also expected to improve the performance.
Figure 3 Average searching time of UVSR.

Figure 3 shows the average searching time of the UVSRI algorithm. When the models are randomly ordered (UVSRI), the average searching time is higher than when the word probability is taken into consideration (UVSRI_WP). The average searching time of standard isolated word recognition with a vocabulary size of 3000 is $3000N^2T$. If the rejection rate of correctly recognized name is about 8%, the average searching time is 1392$N^2T$ for UVSRI compared to 454$N^2T$ for UVSRI_WP. These represent a reduction in searching time of 54% and 85%, respectively. The UVSRI_WP out-performs the UVSRI in both system accuracy and average searching time.

3.2 Performances of the UVSRII Algorithms

We use six confidence levels for the UVSRII algorithm as shown in Table 1. At the $\theta$ confidence level, any threshold less than 120 is allowed. The $\theta$ confidence level controls the rejection rate. Figure 4 shows the recognition results of the UVSRII algorithms. The performance of the UVSRII algorithm is very close to the standard isolated word recognition with NBLR based UV method (UV_NBLR).

If word prior probability is used to rearrange the model list (UVSRII_WP), the recognition accuracy is further improved and achieves a system accuracy of 92.58% and 94.26% by rejecting 7% and 10% of correctly recognized names, respectively. Comparing the performance of UVSRII with or without the language information, it is clear that the recognition accuracy of the UVSRII system does not strongly depend on the order of the model list as in the UVSRI system.

<table>
<thead>
<tr>
<th>Confidence Level, $k$</th>
<th>Threshold, $\xi_k$</th>
<th>Number of UV-Passed Hypotheses, $V_k$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-10</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>-20</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>-40</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>-60</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>-80</td>
<td>11</td>
</tr>
</tbody>
</table>

Table 1 Confidence levels

The average searching time for the UVSRII algorithm is $2300N^2T$ computations for a randomly ordered model list. This represents a 23% reduction in computation time. When the word prior probability is used to rearrange a fixed ordered model list, the average searching time becomes $1700N^2T$ computations, which is a reduction of 43% in average searching time.

Figure 4 Performance of the UVSRII system

4. CONCLUSIONS

In this paper, we investigate the possibility of using a qualifier criterion in searching for the correct hypothesis in a speech recognition task. A qualifier-search can determine the correct hypothesis based on some criteria without the complete knowledge of all possible hypotheses. The utterance verification measurements are used as the qualifier criteria and an utterance verification based speech recognition (UVSR) suitable for large vocabulary speech recognition is described.

Two utterance verification (UV) measurements are used to help making the correct decision. The all-phone garbage model based UV measurement is used to guide the search. The N-best likelihood ratio based UV measurement is used to reject the erroneous recognition events. Multiple confidence levels are introduced to improve the system performance. The UVSR algorithm performs very closely to the standard speech recognition with UV but with 23% less in the average searching time. If the word probability is used, the average searching time can be reduced by 43%. By rejecting 10% of the correctly recognized name, the system accuracy is 94.26%.

The UVSR has been applied successfully to a fixed randomly ordered model list. When the model list is sorted according to the word prior probability, both the accuracy and searching time are improved. In this paper, the order of the hypotheses to be
evaluated is pre-determined. If we can sort the hypotheses dynamically for each input utterance, the accuracy and the searching time are expected to improve further. In a multipass system for large vocabulary isolated word recognition, the sorted list of possible hypotheses is generated from a rough preliminary analysis module [3]. The UVSR system can be invoked to perform an efficient search in this multipass system.

5. REFERENCES


