Rejection in Speech Recognition Systems with Limited Training

Aruna Bayya
Rockwell Semiconductor Systems
4311 Jamboree Road, Newport Beach, CA.

ABSTRACT

In this paper, we propose a new rejection criterion applicable specifically to limited-training speech recognition systems such as Speaker-Dependent (SD) recognition systems. The new criterion uses confidence measures as well as speaker-specific out-of-vocabulary (OOV) models. The OOV models are created from the same training data that is available to create the in-vocabulary (IV) word models. We describe the method for creating these speaker-specific out-of-vocabulary models from limited training data. We also define a fairly robust confidence measure to reject the OOV words. The results presented in this paper demonstrate the effectiveness of the new criterion in a SD recognition task under various conditions.

1. BACKGROUND

Wide acceptance of Automatic Speech Recognition (ASR) systems depends not only on the recognition rates, but also on the user-friendliness of the system. Poor user interface and failures in completing a task successfully could be two most frustrating factors for the users of ASR systems. To avoid this frustration, even the speech recognition systems used in simple tasks such as a small vocabulary isolated-word recognition, should be designed to provide the users with graceful recoveries from system failures as well as user's mistakes. The recovery from these failures can be accomplished by rejecting the utterance and prompting the users for the same spoken input again.

Several rejection criteria have been proposed in the past for both speaker-independent (SI) and speaker-dependent recognition tasks. A typical approach to rejection of out-of-vocabulary words has been to include an explicit model which represents all the out-of-vocabulary words. This model is usually referred to as out-of-vocabulary model or filler model [1,2]. The out-of-vocabulary model is often derived from many samples of out-of-vocabulary words, non-speech sounds such as clicks & pops, and also samples from background noise/silence signals. Shown in Figure 1 is a simplified version of a network consisting of N in-vocabulary word models and K out-of-vocabulary models each one representing one of the above categories of sounds. Rejection or acceptance of a spoken utterance is determined by measuring the closeness of the utterance to the OOV models.

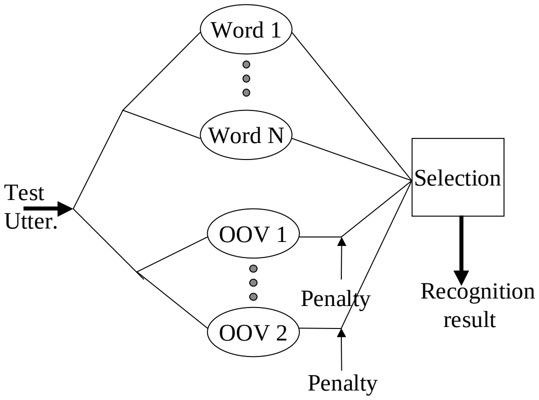

Figure 1: Rejection with Garbage Models

Another approach to rejection is to use an absolute threshold for the scores. In this case, the system relies on the a priori knowledge of score distributions for both in-vocabulary words and the out-of-vocabulary words/sounds. For example, the smoothed histograms of in-vocabulary word scores and out-of-vocabulary sound scores may be as shown in Figure 2.

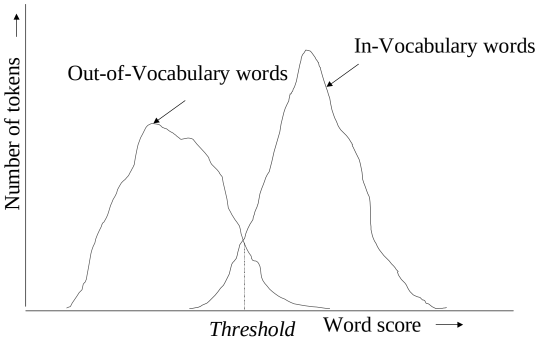

Figure 2: Distribution of scores

Based on the distributions, a threshold is determined for the score. If the best score for an utterance is higher than the pre-determined threshold, the spoken utterance is declared as the word associated with the model scoring the highest. If not, the spoken utterance is rejected. As it is clear from Figure 2, for this approach to be effective, the distributions should be non-overlapping or minimally overlapping.

A third approach to rejection is an extension of the threshold-based rejection. In this method, a confidence measure is obtained from the set of the scores. To measure the confidence level, the scores for all the word models are arranged in a descending order with the best (highest) score being at the top of the list. A simple-minded confidence score is defined as:

$$cm = \frac{1}{S_1}\left(S_1 - \frac{1}{(K-1)}\sum_{i=2}^{K} s_i \right)$$

where S_1 is the highest or best score and $S_2 \ldots S_K$ are the next K best scores. Then, all the spoken utterances with $cm < x\%$ are rejected where x is user-defined number.

While these approaches have been shown to be successful in speaker-independent recognition systems, applied independently, they are not very effective in speaker-dependent recognition systems for the following reasons:

1. Knowledge of score distributions is crucial to threshold-based rejection. However, in SD recognition systems, the score distributions are not available when the user begins to use the system. Even when they are available, they vary significantly from speaker to speaker. Hence, it is not possible to design a single criterion that is optimum for all users. Also, due to limited training data in the SD recognition system, the models for the vocabulary words are usually not as robust as they are in the case of SI recognition. Hence the distribution of scores for in-vocabulary words and the distribution of scores for out-of-vocabulary words/sounds are not as well separated as it is shown in Figure 2. In such a situation, simple application of score thresholds will not yield low false rejection rate at low false alarm rate. If a tight threshold is applied, the false rejection rates will be unacceptably high. If a low threshold is applied, the number of resulting false alarms would be too high.

2. While using the confidence measures is slightly superior to using score thresholds, that criterion is also somewhat weak when the word models are created from only one repetition of the word. This is particularly true when the vocabulary contains similar sounding word sets. In such cases, the confidence level may always be low when scoring the confusing word sets, thus resulting in high rejection rates for those words.

3. The designer of a SD speech recognition system is faced with another challenge when using out-of-vocabulary models. Since the in-vocabulary word models derived from severely limited training data are not robust, creating a reliable out-of-vocabulary model is crucial to the success of rejection in SD recognition [3]. However, creating a reliable out-of-vocabulary model for SD system is a difficult task due to lack of additional speech/non-speech samples. Even though, some speaker-independent training data may be obtained from other sources, in most of the applications, almost no speaker-specific data is available. The out-of-vocabulary models created from SI training data are not as effective as speaker-specific out-of-vocabulary models.

Since any of the above rejection criteria, applied individually, will not yield satisfactory results, a combination approach is more suitable for SD applications. In the following section, we describe one such method.

3. NEW REJECTION CRITERION

We propose to accomplish the rejection by using a combination of confidence measures and an out-of-vocabulary model. The confidence measure is a function of the distance of the highest score to the next K-best scores where K is 2 or 3. The current implementation includes the case of K = 1. In other words,

```
If (best_score – second_best_score) < 75% of best_score
then
        Reject the utterance
else
        Declare the recognized word
```

To increase the rejection rate for out-of-vocabulary words without affecting the false rejection rate, we also use a speaker-specific out-of-vocabulary model. For each speaker, a separate speaker-dependent out-of-vocabulary model is derived from the speech collected during the training phase. Since the speech collected during the training phase corresponds to the vocabulary words, a collage of sounds is formed by scrambling the order of frames. This new sequence of frames, while retaining the speaker-specific characteristics, does not carry any of the acoustic characteristics of the vocabulary words, thereby making a useful training material to obtain a reliable out-of-vocabulary model. Many sets of new sequences can be prepared by ordering the frames in different ways. Repeating this process of scrambling, as much training material as needed can be created. Since the acoustic properties of the new training material and the original training token are different, the out-of-vocabulary model can be created even when one training token for only one word is available. It can be updated as more training tokens are available from the speaker.

The scrambling or rearranging can be performed at frame level or at state level segmentation. These two different techniques are presented in the form of a flow diagram in Figures 3 and 4. In the frame level segmentation, the speech is segmented into frames of arbitrary length (40-50 msecs). Then, these frames are ordered in several different ways, to form a set of new tokens. These tokens are then used in the training process to build a model for out-of-vocabulary speech. Any of the scrambled speech samples available from previously trained words are also included in the training process.

In the current implementation a 50 msec. segments are used in deriving two training samples from one token. In the first sample, each frame is switched with its neighboring frame. The second sample is prepared by switching every 'i^{th}' frame with '$(T\text{-}i)^{th}$' frame where '*T*' is the total number of frames in the utterance.

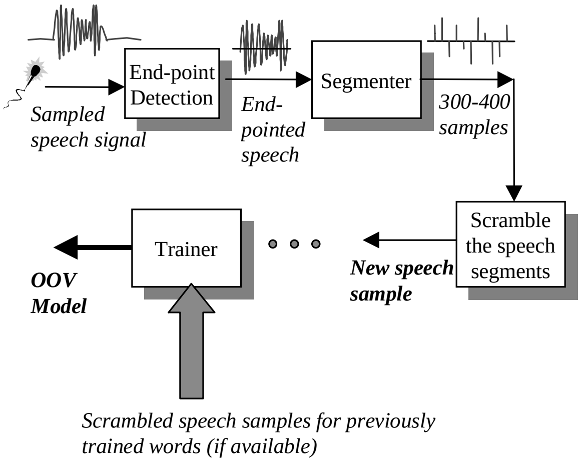

Figure 3: Creating training speech for garbage model

One advantage of this approach is that even the training tokens in compressed form can be used to build an out-of-vocabulary model. Hence, the creation of an out-of-vocabulary model can be performed off-line. Updating the out-of-vocabulary model to accommodate new training material is easier and usually results in a more robust out-of-vocabulary model. Our experiments have shown that the performance degradation by using compressed speech in almost unnoticeable. Because of the limitations on available resources, this is a requirement for some applications such as 'Voice-Dialing' in a cellular phone.

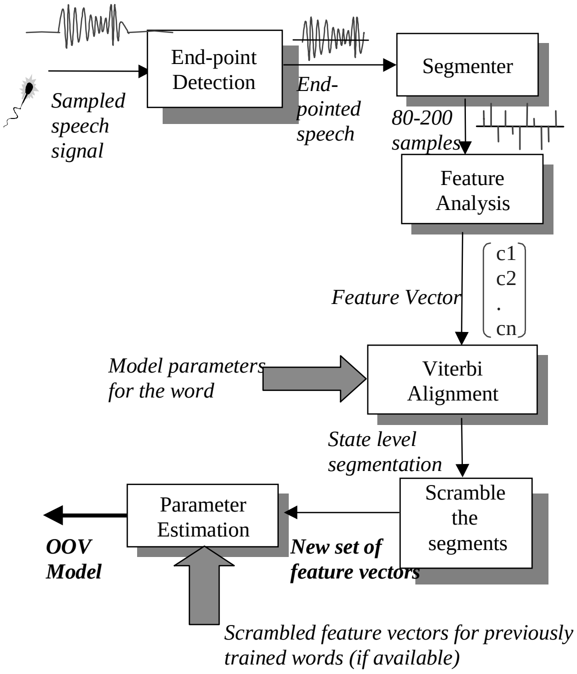

Figure 4: Creating training vectors for garbage model

Another way of creating training material is to use state-level segmentation. In this approach, the feature vectors are computed for each frame (of length 10-20 msecs). Then the utterance is matched with itself by using previously computed word model. The matching process (Viterbi scoring) results in a state segmentation. In other words, each frame of the word is assigned to a HMM state in an optimal way. Then, the scrambling is done at a state level. The order of the states is modified while keeping all the frames in a state together. This rearrangement is more intuitive and more accurate because each state in a HMM is supposed to represent an acoustic event or acoustic unit.

As before, each way of ordering the states will result in a training sample. Since the feature extraction and state segmentation is already done, the new state sequences are used directly in the parameter estimation as shown in Figure 4. Any of the scrambled state sequences available from previously trained words are also used in the out-of-vocabulary model parameter estimation.

In the current implementation, only one training sample (state sequence) is derived from each utterance. This sequence is obtained by switching every 'i^{th}' state with '$(N\text{-}i)^{th}$' state where '*N*' is the total number of states in the token. The out-of-vocabulary model will become more robust as more words are introduced into the vocabulary.

In both approaches, the amount of training material created and the selection of a small set of scrambled sequences from all the possible sequences is determined based on the computing/memory resources available for a particular application.

4. RESULTS

To measure the effectiveness of the new criterion, it is implemented in a HMM-based speaker-dependent recognition system. In this system, the user is required to say each vocabulary word only once during the training phase. Hence only one token for each word is available for creating in-vocabulary word models as well as the out-of-vocabulary model. In all the experiments, the vocabulary size is fixed to be 20. All the results presented in this report assume the availability of all the utterances at the time of building out-of-vocabulary model. However, the new criterion is also proved to be extremely robust even in the case of incremental training (where only partial data is available to create the out-of-vocabulary model initially, and the model is updated as more words are added to the vocabulary).

In the following, we present results of applying the new criterion to various databases. The databases selected for this purpose are representative of various applications. The first database (43 speakers, open microphone, 8 repetitions of each word) represents a situation where both the training and the recognition are performed in the same environment. However, the recognition is carried out over a long period of time. The second database (10 speakers, open microphone and telephone, 16 repetitions of each word) represents mismatched equipment where the training is done using one type of microphone and the recognition tests are performed using a different microphone. The third database (7 speakers, open microphone and the handset, 20 repetitions of each word) is representative of mismatched conditions. In this case, the training data consists of speech samples recorded in a quiet environment while the recognition tests are performed on data recorded in quiet as well as noisy environments.

The results reported here include the average recognition rate on 20 vocabulary words, rejection rate for 10 out-of-vocabulary words and false rejection rate for in-vocabulary words at different confidence levels. To prove the superiority of performance of the new criterion, it is compared with threshold-based criterion and also with the approach using SI out-of-vocabulary model.

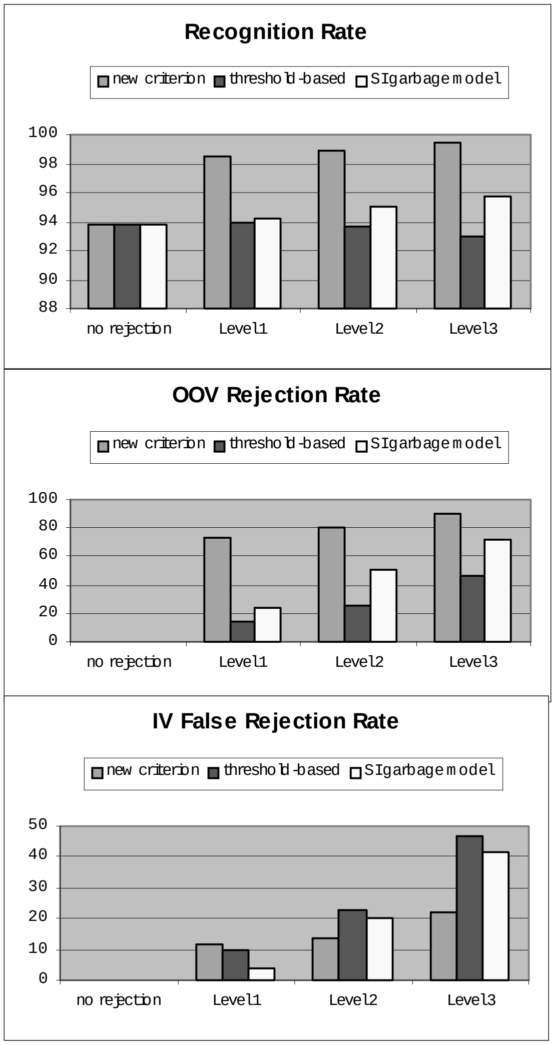

Figure 5: Results of rejection on 43 speaker database

As it can be observed from Figure 5, the new criterion is consistently better than the other two approaches in rejecting the out-of-vocabulary words and improving the recognition rate while maintaining a low false rejection rate.

	rejection criterion	% recog with no rejection	% recog with rejection	% false rejection rate
10 spks	New	88.36	97.2	23.1
	Thrsh	88.36	90.21	21.1
	SI model	88.36	90.2	20.73
7 spks	New	86	95.6	18.7
	Thrsh	86	86.7	38.2
	SI model	86	89	24.2

Table 1: Results of rejection on tasks with mismatched recording equipment and mismatched background conditions

It is noted from the table that the new criterion outperforms the other two criteria even in the case of mismatched conditions. At similar rejection levels, the effective recognition rates are higher when the new criterion is used.

Finally, to show that new criterion is invariant to data storage options available in a specific application, we show the results of using the new criterion when the training samples are stored in compressed form and can be retrieved after decoding the compressed speech. The following table shows the recognition rate, false rejection rate and the rejection rate for out-of-vocabulary words resulting from tests on Database 1 using a G729A Codec (standard in the GSM wireless applications).

	IV % recog. rate	IV % false rejection rate	OOV % rejection rate
uncompressed	99.25	23.51	85.75
compressed	99.16	22.1	84.34

Table 2: Results of rejection using compressed speech

From Table 2, it is clear that compressing the data there by losing some quality of the speech has insignificant influence on the performance of the new criterion.

5. SUMMARY

A simple, but powerful rejection criterion is proposed for limited-training speech recognition tasks such as SD recognition. The results of SD tests on various databases indicate the robustness of the criterion to various recording methods, background conditions and storage options.

6. REFERENCES

1. J. Wilpon et. al., “Automatic Recognition of Keywords in Unconstrained Speech Using Hidden Markov Models”, *IEEE Trans. ASSP*, 1990, pp. 1870-1878.
2. R. C. Rose and D. B. Paul, “ A Hidden Markov Model Based Keyword Recognition System”, *Proc. ICASSP-90*, pp. 129-132.
3. Vijay Raman and Vidhya Ramanujan, “Robustness Issues and Solutions in Speech Recognition Based Telephony Services”, *Proc. ICASSP-97*, pp. 1523-1526.