OOV Utterance Detection based on the Recognizer Response Function

Erica Bernstein and Ward R. Evans
The MITRE Corporation
1820 Dolly Madison Blvd.
McLean, VA 22102
email: egb@mitre.org or wrevans@mitre.org

ABSTRACT

This paper addresses the problem of out of vocabulary (OOV) utterance detection for spoken language systems in an open microphone environment. This problem is becoming crucial as use of spoken language systems grows beyond the research laboratory. In the past this problem has been addressed in the context of keyword spotting, e.g., for connected digits in a telephone environment and more recently in OOV word detection in a large vocabulary continuous speech recognition system.

We develop a novel technique for designing a lexical garbage model that takes advantage of application specific knowledge and any potential bias in the recognizer. We do this through the formulation of a recognizer response function.

1. BACKGROUND

The problems of garbage modeling, utterance verification, and OOV detection are important and difficult problems to be solved by automatic speech recognition systems. These problems arise in many applications; from large vocabulary continuous speech recognition to small vocabulary keyword spotters. We address this problem in the context of a small vocabulary spoken language system called VOICE MOUSE.

VOICE MOUSE is a spoken language interface for X-windows. It allows the user to access windows by name, voice macros, launch applications, and rename windows. It was designed to be used by researchers at MITRE as they performed daily tasks on their workstations. In addition VOICE MOUSE records all spoken interaction, thus allowing later data analysis. The system is speaker independent with a vocabulary of 40 to 100 words. One novel feature is the use of a dynamic grammar, which allows the introduction of window names "on the fly." For maximal convenience the systems works in an open microphone mode. Since the system is used in a normal office environment, it is necessary to deal with environmental noise, spontaneous speech artifacts, and with extraneous speech. The fact that our system is a fairly simple real world system makes it ideal to investigate various approaches to these problems. Furthermore, the extraneous speech from he open-mic environment makes this problem different form the usual small vocabulary word spotting tasks in that the OOV utterances are less predictable.

In real world systems integration applications, it is not possible to modify the commercial recognizers directly. Thus, we were interested in developing an solution that could be implemented through the language model. Naturally, this led us to consider how we could construct word models for OOV utterances. To this end, we develop an explicit junk model that is not trained directly on acoustic data.

Previous approaches to OOV utterance detection, or garbage modeling have used either acoustically trained garbage models, [3], [4] or score normalization procedures [5]. The approach of [1] in using discriminant analysis for online garbage modeling motivated us to consider application dependent garbage modeling. The approach of [2] for designing lexical fillers led us to the idea for the recognizer response function.

2. RECOGNIZER RESPONSE FUNCTION

Experience has shown that recognizer errors are not random, but seem to be a function of the language model, the training of the acoustic models, and environmental factors, such as the user and room acoustics.

The recognizer response function (RRF) is constructed from transcriptions of all utterances recognized as well as the target speech. This allows us to determine the bias of the recognizer towards particular phones given the particular language model. This approach should be contrasted with methods that normalize recognizer scores with the output of a phoneme recognizer [5] since we obtain the phonemic transcriptions from the recognized utterance. In this way we take into

account the predisposition of the recognizer to select the phonemes given the language model. In other words the RRF is the conditional probability of a phone given the language model and the recognizer.

RRF(phone)=P(phone | lang. model, recognizer)

OOV utterances are represented a string of junk words, each word consisting of a single phoneme that belongs to the lexical model which we derive from the RRF. The phones are rank ordered according to the RRF and are included into the junk model by decreasing order of occurrence. If all phones are included then all OOV utterances are detected but also an unacceptable number of legitimate utterances are classified as junk. In classical detection theory terms we would say that the probability of detection is one, but that the false alarm probability is too high. Thus, false alarms require the user to repeat a legitimate utterance or take other action while a missed detection could result in inappropriate system behavior.

The RRF as defined above is not convenient to compute and does not lend itself to computation "on the fly." Thus, we examined several alternative formulations.

The first formulation is given by:

RRF1(phone) = P(phone | SRO) - P(phone | transcription)

where SRO means speech recognizer output. In this formulation, the RRF is difference between what the recognizer identified and what was actually said. Computation of this formulation requires task specific data.

The second formulation:

RRF2(phone) = P(phone | OOV, lang. model),

that is, the probability that the phone was identified from an OOV utterance. If the recognizer transcribes all in-vocabulary utterances correctly, this formulation is equivalent to formulation 1. Computation of this formulation requires only a representative set of OOV utterances.

Thus, the error rate of the recognizer for in-grammar utterances will help determine the appropriate formulation.

The third formulation:

RRF3(phone) = P(phone | lang. model)

where the probability is computed directly from the language model, in which all legal utterances are regarded as equally likely. Note this formulation does not require SRO output and hence, no collection of data.

3. JUNK MODELS

Once a RRF is selected the phonemes are rank ordered by frequency of occurrence. Then the most likely phonemes are selected to make the OOV word models. Each word model consists of a single phone and each OOV utterance is modeled as a string of OOV words. Figure 3.1 illustrates the incorporation of the junk model into the recognition grammar.

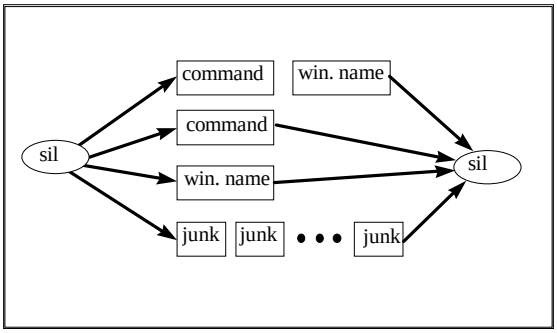

Figure 3.1

For each recognizer response function we evaluated several junk models incorporating different numbers of phones. A junk model was tested using randomly selected phones and found not to perform well.

4. EXPERIMENTAL RESULTS

Using these approaches we designed several junk models. For each model we computed a probability of detection (of junk) and a probability of false alarm (legitimate utterance classified as junk). This allowed us to construct a ROC curve as used in classical detection theory.

The data were collected during normal daily use of the VOICE MOUSE.

In figure 4.1 we compare the phoneme frequency distributions for formulation 1 and formulation 2. Since these distributions are virtually identical, we based our junk models only on formulation 2 and 3.

Figure 4.2 is a ROC curve comparing performance of junk models derived from formulations 2 and 3. Each point on the ROC curve is computed from a different junk model. Clearly, formulation 2 yields better OOV detectors than formulation 3, with the best yielding a probability of detection of .98 and a false alarm probability of .09, whereas for formulation 3 the best junk model yielded a .93 probability of detection and .32 false alarm probability. This is due to the fact that formulation 3 does not include information about the bias of the recognizer. We illustrate this fact in figure 4.3, where we compare the frequency distribution of phones for OOV utterances vs. the frequency distribution based solely on the language model.

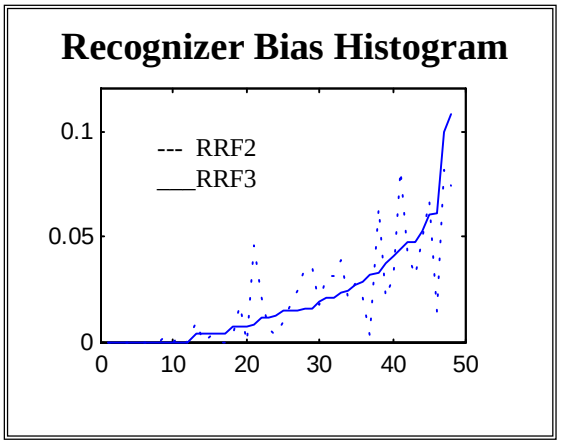

Figure 4.3

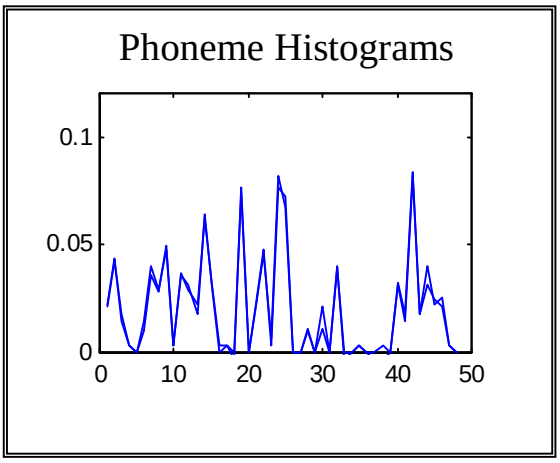

Figure 4.1

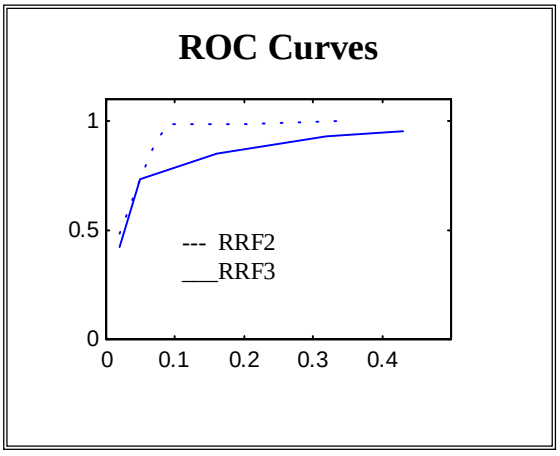

Figure 4.2

5. CONCLUSIONS

We believe that the formulation of a recognizer response can lead to the development of effective junk models which do not require modification to the recognizer. Estimating recognizer bias leads to significant improvement over junk models based solely on language model predisposition. The third formulation may be useful for recognizers which enjoy a low level of bias or in applications which are not sensitive to false alarms.

6. REFERENCES

[1] J. Caminero., et.al, *"On-line Garbage Modeling with Discriminant Analysis for Utterance Verifications,"* ICSLP 1996, Philadelphia, PA.

[2] R. Meliani, and D. O'Shaughnessy., *"New Efficient fillers for Unlimited Word Recognition and Keyword Spotting,"* ICSLP 1996 Philadelphia, PA.

[3] M.G. Rahim, C.H. Lee and B.H. Juang, *"Robust Utterance Verification for Connected Digits Recognition"*, ICASSP-05, Detroit, pp. 285-288.

[4] R.C. Rose, B.H. Juang and C. H. Lee, *"A Training Procedure for Verifying String Hypothesis in Continuous Speech Recognition"*, ICASSP-95, Detroit, npp. 281-284.

[5] S. Young and W. Ward, *"Recognition Confidence Measures for Spontaneous spoken Dialog,"* EUROSPEECH 1993, Berlin, Germany