Extending Grammars Based on Similar-Word Recognition

Rocio Guillén, Randal Erman

Computer Science Dept., California State University, San Marcos, San Marcos, CA, USA

ABSTRACT

Pronunciation variation is extremely widespread and one of the reasons for recognition errors. In this paper we explore how similar-recognized-words can be used to construct or expand more accurate grammars in a specific domain. The domain that serves as framework for this research is the assessment of depression. Assessment of depression is done via a system that verbally administers a discrete choice questionnaire over the telephone. Several experiments carried out have shown that in spite of the limited number of valid responses, responses uttered by the same speaker may be both correctly or incorrectly recognized. Analysis of the responses incorrectly recognized has provided the elements to formulate new grammar rules. To test the grammars thus expanded a new set of experiments was completed and the results obtained are presented and discussed.

1. INTRODUCTION

One of the problems in automatic speech recognition (ASR) and understanding is the variation in pronunciation. Pronunciation variation is known to be extremely widespread and among the reasons why words are misrecognized. Some authors have modeled pronunciation variation by investigating the characteristics of correctly and incorrectly recognized utterances in interactive voice response systems. Their investigation is based on prosodic analyses of speaker turns in such systems. The combination of prosodic features with other information such as recognized string and acoustic confidence score can be used to predict whether speaker turns will be recognized correctly or not with a high degree of accuracy [1,2]. Other authors consider that multiple pronunciations should be included in the lexicon and phonological rules can be used to generate them. There are different types of linguistic variation that account for pronunciation variation. In general, it is often difficult to define precisely the type of variation modeled in a particular approach, because in most cases it is a combination of different types. Research on pronunciation variation modeling is mainly concerned with testing specific methods to determine whether they improve recognition performance and to what extent [4]. In our research, we explore how similar-recognized-words can be used to construct more accurate grammars in a specific domain. The domain that serves as the framework for our research is the assessment of depression. Participants are interviewed over the telephone by an ASR system VIDAS (Voice-Interactive Depression Assessment System). The interview consists of a set of twenty questions previously recorded both in English and Spanish. Responses range between zero and seven. The recognized responses and the actual responses are recorded automatically for further processing. Once the interview is finished the system scores the recognized responses to assess the level of depression of the client. A high degree of recognition accuracy is crucial in this task. A prototype has been used for interviewing Spanish and English speakers, both male and female. Different grammars have been designed for each language group. Several experiments carried out with the current system have shown that in spite of the limited number of valid responses, responses uttered by the same speaker at different times may be both correctly and incorrectly recognized. Results from analyzing the data have provided the elements to formulate new grammar rules for constraining the selection of candidate responses; that is, the set of hypotheses. The approach for analyzing recognized responses versus actual responses is based on the comparison of contexts in which recognized and actual responses may appear. We are carrying out experiments with the English grammar to determine whether the grammar thus augmented improves the accuracy of the responses, and to what extent. The speech recognizer is a speaker-independent system (IBM) with built-in vocabularies in different languages that has been adapted to work with a telephone.

2. FRAMEWORK

VIDAS is a system that verbally administers a discrete choice questionnaire by presenting pre-recorded prompts in English or Spanish, recognizing spoken responses, scoring the responses, and storing the data. The questionnaire consists of a 20-item self-report screening measure for assessing the frequency of depressive mood and symptoms during the past week. To ease the presentation of the all-audio all-verbal computer-telephone system, the response format was modified from the standard four responses (less than 1 day, 1 to 2 days, 3 to 4 days, and 5 to 7 days) to the actual number of days (0 to 7) in one week. Regardless of the response format, total score variances did not differ for either language group.

Over the telephone, VIDAS instructs the participant on verbally completing the interview (samples of male and female voices are presented) and provides the participant with an opportunity to select the gender of the automated interviewer. When the participant is ready, VIDAS presents the depression items one-by-one, registers the participant's spoken responses, and records voice characteristics. Once the interview is concluded, VIDAS hangs up, scores the responses, analyzes the data, generates a brief interpretive report, and stores the results to be further analyzed [11]. A fragment of an interview is shown in Figure 1.

V: I felt hopeful about the future
P: seven (Not recognized by V)
V: I felt hopeful about the future
P: seven (Recognized by V)
V: I felt that people disliked me
P: what (P did not hear/understand question)
V: I felt that people disliked me
P: three (Recognized by V)
V: My sleep was restless
P: four (Recognized by V)

**Figure 1**: Fragment of an interview, V=VIDAS, P=Participant

Valid responses for English are shown in Figure 2. Additionally, all those words uttered by the participant not in the grammar are recorded. These utterances are not considered when scoring the responses. All recorded information will be used to model a dialog-like interview to allow more spontaneous and conversational speech.

### 3. METHOD

In this section we present first some approaches to modeling pronunciation variation. Next we describe how recognition errors are identified in our system as part of pronunciation variation.

#### 3.1 Pronunciation Variation

Approaches to modeling pronunciation variation can be split into two classes, data-driven and knowledge-based, depending on the source from which information on pronunciation variation is derived. Data-driven methods derive information directly from the signals [5,6], whereas knowledge-based methods derive information from already available sources [7,8,9] such as pronunciation dictionaries and linguistic studies [10]. Our approach is knowledge-based because we are deriving information from pronunciation dictionaries.

Many speech recognizers achieve successful recognition rates by imposing constraints either by specifying the grammar of the sequences of words which are allowed or by providing statistical likelihood’s for word sequences. Some authors [3] have attempted to combine the strengths of hand-built grammars with the strengths of statistical approaches to restrict a speech recognition grammar. We are using a speaker-independent speech recognizer with a pronunciation lexicon that can be constrained grammars defining the vocabulary that is permitted. The vocabulary must be a subset of the pronunciation lexicon. Or, the speech recognizer takes the pronunciation lexicon as the vocabulary that is permitted.

#### 3.2 Recognition Errors in VIDAS

We distinguish between misrecognized responses (utterances) and non-recognized responses (utterances). Misrecognized words are those words that are in the grammar, but do not correspond to the actual response. Non-recognized words are those words that are not in the grammar. When using the lexicon without a grammar there are only recognized and misrecognized words, because the speech recognizer searches for the word or phrase with the maximum likelihood. Our approach for distinguishing between recognized and non-recognized words is based on a comparison between the word that was recognized and the actual response previously recorded.

Most speech recognizers are composed of three elements: the lexicon, the acoustic models, and the language model [4]. Pronunciation variation in the lexicon is modeled by adding pronunciation variants to the lexicon. A drawback is the generation of new errors because these variants can be confused with those of other entries in the lexicon. A solution is to find the set that leads to the best performance. Among the criteria to carry out this task are frequency of occurrence of the variants, maximum likelihood, confidence measures, and the degree of confusability between the variants. We have focused on the lexicon because it is the component that can be modified in the speech recognizer we are using.

### 4. EXPERIMENTAL METHOD

The goal of the experiment was to increase accuracy by having similarly sounding words produce the desired word in the grammar. To accomplish the goal we first decided on two sets of words that we wanted to implement in a grammar. One set was the grammar based on the VIDAS valid responses, and the second set was a random selection of words. The latter was created using the Carnegie Mellon University Pronouncing Dictionary (CMUDICT). CMUDICT is a machine-readable pronunciation dictionary for North American English that contains over 100,000 words and their transcriptions. The dictionary can be used for speech recognition and synthesis. We randomly selected fifty words.

We then asked ten participants to say each word of the desired grammars five times to the speech recognizer and store what it understood. The whole lexicon was used in this task. We then created the constrained grammars A and B shown in Figures 2 and 3. The constrained grammars are context-free and defined in the BNF (Backus-Naur form) the speech recognizer requires.

```<Root>> = zero | one | two | three | four | five | six | seven | yes | no | pardon | repeat | what
```

**Figure 2.** Constrained Grammar A

```<Root>> = abdicite | adrift | automatically | baldness | blast | bolt | chocolate | cinema | coach | crossroads | dare | defensive | death | disputes | eats | epic | fabrics | game | garment | gut | hair | how | infuriate | innocent | keg | lamb | legislate | luggage | march | maybe | musical | nacho | overburden | phantom | pretend | pursuit | redefine | reprint | roses | sands | september | several | strategy | swallow | task | treaty | vault | waif | wet | wipe
```

**Figure 3.** Constrained Grammar B

The next step was to automatically compare the stored words with the words in the grammar to find non-recognized and misrecognized words. Both non-recognized and misrecognized words were mapped to the word that should have been recognized (see example in Figure (4)). The result of the mapping was an extended grammar for constrained grammars A and B (see Figure 5 and Figure 6).
Results from the constrained grammars without any mapping showed a very high accuracy. Both accuracy scores were 90% or above. The 10% that was incorrect were caused do to similar words such as “Bolt” and “Vault,” and also due to the recognition engine just failing to recognize what was said. Charts from the results are shown in Figure 7.

Results from the expanded grammars produced a higher accuracy then the lexicon alone, but were 4% less accurate then just the constrained grammars. Charts from the results are shown in Figure 8, along with a graph (Figure 7) showing a comparison between the three tests.

![Graph of Grammar Accuracy](image)

### 4.1 Experimental Results

There are several interesting things to note about the results from the experiment. The first is the very poor accuracy that the non grammar defined lexicon produced. The average accuracy rate when using the standard lexicon was 66% for grammar A which had 13 words, and 43% for grammar B which had 50 words. These low accuracy rates along with the types of words that were returned, we believe caused the next part of the experiment to suffer.
5. CONCLUSION

Although this method produced little to no positive results in terms on increasing accuracy, we did find clues that will help with the next wave of experiments. We are currently doing more statistically analysis on each individual word in each of the grammars to decipher if there is a correlation between the phonological values of the returned results and improved accuracy of each word.

Also, we have observed the reason, both expanded grammar A and B had a decrease in accuracy was due to the poor initial recognition that returned values completely not similar to the words being spoken for example “Jet Lack” for “chocolate”. This caused the engine to pick up and hence map non similar words when the participants where testing the expanded grammars.

The approach that we have presented may be less or more successful with another speech recognizer. Furthermore, our approach being successful does not necessarily mean that we are modeling pronunciation, but rather correcting for some of the characteristics of the speech recognizer.

6. ACKNOWLEDGEMENTS

This research is supported by a grant from NIH under the MBRS SCORE program.

7. REFERENCES


