Modelling Non-verbal Sounds for Speech Recognition

Wayne Ward[1]
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

ABSTRACT

When speech understanding systems are used in real applications, they encounter incidental noise generated by the speaker and the environment. Such noises can cause serious problems for speech recognizers not designed to cope with them. We attempt to model these noises by training HMM "noise words" to match classes of noises. The noise words were incorporated into the Sphinx system and performance compared to the system without noise words. Initial results suggest that the technique does increase system performance significantly.

INTRODUCTION

Recent experiments performed by two groups of researchers at CMU have gathered data on subjects using speech recognizers in office-like environments (Rudnicky, et al., 1989, Stern & Acero, 1989). These experiments are presented by the authors in these proceedings. Among other things, they show that non-verbal events (non-stationary noises) do create serious problems for speech recognizers. These sounds are generated both by the speaker and by the environment. Examples of noise generated by the speaker are breath noises, lip smacks, paper rustles, filled pauses, cough, clearing throat, etc. Environmental noise can be phone rings, door slams, other speakers in the background, typing, etc. We attempt to explicitly model classes of noise represented by these events in the context of an HMM based speech recognizer (Sphinx). Subjects were recorded performing the two tasks, spreadsheet and census data (alphanumeric). A significant percentage (approx 10% overall in each task) of the utterances contain phenomena of the type mentioned above. The utterances were transcribed using a set of noise words to represent non-signal events in the recording. Fourteen noise words were used: AH, BEEP, BREATH_NOISE, CLEAR_THROAT, COUGH, DOOR_SLAM, MOUTH_NOISE, MUMBLE, RUSTLE, PHONE_RING, SNIFF, SNEEZE, TAP and THUMP. For each of these noise classes, a phone was added to the phone set and a word consisting of only that phone was added to the lexicon. The standard Sphinx training routines were then used to train context dependent models for all phones except those representing noise. Context independent models were used for the noise phones. The simple word models for noise give no context since they are single tokens, and we did not use between-word models. For recognition, noise words are treated like Silence words. They are allowed to occur after any word, including themselves and other noise words. We use the Sphinx recognizer with only minor modifications to implement transitions to noise words and to allow utterances that are only noise or Silence.

SPREADSHEET TASK

Alex Rudnicky and Michelle Sakamoto gathered a large corpus of examples of users performing a spreadsheet task using voice (Rudnicky et al, 1989). They used an operational speech recognition system, not a PNAMBIC paradigm. The subjects spoke in a spontaneous manner and were recorded using a Sennheiser close talking microphone. The input to their system was continuous, recognition wasn't started by pressing a key just before

---

[1]This research was sponsored by the Defense Advanced Research Projects Agency (DOD), ARPA Order No. 5167, under contract number N00039-85-C-0163. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Defense Advanced Research Projects Agency or the US Government.

speaking. The data used in the present experiment consists of 15 sessions each from 7 speakers (they subsequently recorded more). A session represents approximately 100 utterances. These utterances were divided into a training set and testing set which were then transcribed using noise words. Some "utterances" contained only real words, some contained words and noise, and some were noise alone. Noises loud enough to pass background thresholds were recognized as words even if the subject were not speaking. The recognizer used in their experiment was trained on 4000 read utterances from spreadsheet and calculator tasks. In order to avoid becomming speaker dependent, we used the 4000 speaker-independent read utterances along with 416 noise-word containing spontaneous utterances as our training set. The test set was 7185 spontaneous utterances from the seven speakers (not including any from the training set). The 416 noise utterances in the test set came from only five speakers, so two of the speakers in the test set had not been seen in the training. The original recognizer, Sphinx (SPHX) and the noise word recognizer, Phoenix (PHNX) were both run on the test set. Table 1 shows the word and sentence error rates for each type of utterances.

WORDS and NOISE			
	SPHX	PHNX	% Reduction
sent	47.2	26.9	43.0
word	24.6	11.1	54.9

NOISE Alone			
	SPHX	PHNX	% Reduction
sent	81.4	3.8	95.3

ALL Utterances Containing Noise			
	SPHX	PHNX	% Reduction
sent	70.0	14.2	79.7

WELL-FORMED			
	SPHX	PHNX	% Reduction
sent	22.1	20.1	
word	9.6	8.5	

Table 1: Error Rates for Spread Sheet Task

The WORDS and NOISE results reflect those utterances whose transcripts contained both real words and at least one noise word. In this condition, use of noise words reduced the sentence error rate by 43 percent. The NOISE results are for those utterances whose transcripts consist solely of noise words. The noise word models were very effective at discriminating these events from real speech. Less than four percent produced real word hypotheses. For the original system, 81.4 percent of these noises resulted in hypothesized words from the lexicon. The two previous categories are then combined to give errors for all utterances containing noise. For this test set, the number of sentence errors for utterances containing noise was reduced by a factor of five by using noise words (290 vs 41). The WELL-FORMED condition are those utterances whose transcripts contained only real words from the lexicon.

This was run as a check that using noise words did not degrade performance on clean input. As can be seen, the system with noise words performed at least as well as the original system on clean input.

Census Data Task

Richard Stern and Alejandro Acero gathered data on subjects entering census data (Stern & Acero, 1989). Subjects were asked to spell their name and street address, etc. This is largely an alphanumeric task. The recordings were made in a booth partitioned from the rest of the office, and a Sennheiser close talking mike was used. In this task, subjects were prompted when to speak as opposed to the spread sheet task where recording was continuous. Thus, there were no utterances that contained only noise. As before, the utterances were transcribed using noise words, and the system was trained using these models. The error rate for the utterances containing words and noise is shown in Table 2.

WORDS and NOISE			
	SPHX	PHNX	% Reduction
sent	82.9	63.4	23.5
word	33.9	20.9	38.3

Table 2: Error Rates for Census Data Task

In this task also, the use of noise words significantly reduced error rate for noisy input. As before, there was also no degredation of performance for clean input.

Conclusion

The experiments here were quick studies designed to test the feasibility of using HMM models in the standard framework to model non-stationary noise in the input. The results suggest that the noises that are problematic for close talking mikes in office environments can be modelled with these techniques. We intend to extend and refine these models to give better models of a wider range of events. Much more can also be done with the way that the models are used. Currently, they are allowed to follow all words with no difference in probability. While environmental noise probably is randomly interspersed throughout the signal, this is not true of user generated noise. These noises are more probable at some places than others. Breath noises and rustles are far more common at the beginning and end of utterances, for example. Statistics on occurences of these events can be incorporated into the search as a part of the language model. However, this ability requires that noise words be reliably distinguished from each other. In the data presented, noise words were stripped out for the analysis. Insertions and substitutions of noise words were not counted as incorrect. While noise words were not often confused with real words, they were often confused with other noise words. Better modelling of these events will be required before their "language model" probabilities can be reliably applied.

References

1. Rudnicky, A.I., Sakamoto, M.H., and Polifroni, J.H. Evaluating Spoken Language Interaction. Proceedings of the DARPA Speech and Natural Language Workshop, 1987.

2. Stern, R.M. and Acero, A. Acoustical Pre-Processing for Robust Speech Recognition. Proceedings of the DARPA Speech and Natural Language Workshop, 1987.