USING TALKER LOCATION TO DETECT SPURIOUS UTTERANCES IN DESKTOP COMMAND AND CONTROL

Devang Naik

Apple Computer Inc, Cupertino, CA 95014.

ABSTRACT

Hands-free desktop command and control speech recognition suffers from the critical drawback of improperly rejecting spurious conversation. This results in false acceptances of unintended speech commands that can inconvenience the user.

A neural-network approach is proposed to detect spurious conversation by determining talker location. The approach is based on the premise that spoken utterances not directed towards the microphone source tend to be more reverberant and are likely to be spurious.

The method estimates a confidence measure proportional to the amount of reverberation in the endpointed speech signal. The measure is obtained from a neural network that determines if the speech signal was directed to the microphone or was spoken otherwise. The proposed measure can be combined with the acoustic, linguistic and semantic information to improve upon decisions taken by conventional rejection modeling schemes.

1. INTRODUCTION

Current spoken command and control systems typically employ microphones that are desk or computer mounted. Due to cheaper design, these microphones tend to provide directivities that render them sensitive to spurious signals such as background conversations. Consequently, the discrimination between spoken commands and spurious utterances is considerably worse than setups that employ close-talking or noise-cancelling microphones.

Ideally, desktop command and control requires interaction that is uninhibited by the talker location and background conversations. However in practice, poor end-pointing, limited generalization in trained acoustic and language models and practical limitations in the CPU and memory footprint, results in *misfires*. These are spurious utterances falsely accepted as valid spoken commands.

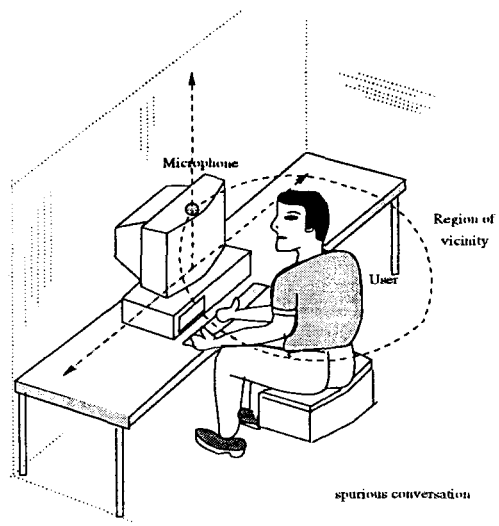

Figure 1: Environment for a desktop speech recognition system.

A high false acceptance rate causes significant inconvenience to the user. Conventional rejection modeling schemes have severe limitations due to poor linguistic generalization of Out-of-vocabulary words and acoustic models.

Hence, there is a need to investigate techniques that complement conventional rejection modeling with little increase in system complexity. By making certain assumptions, it is possible to devise an approach that detects whether or not the speech signal was directed to the microphone. This information can be then augmented with conventional rejection to reduce false acceptances and maintain high in-grammar recognition accuracy.

The subsequent section outlines the assumptions made in the proposed approach, followed by a description and some preliminary experiments.

1.1. Assumptions

Figure 1 shows a typical environment where a desktop speech recognition would be deployed. One can assume

that:

- the environment is typically reverberant, such as an office or a room.
- Spoken interaction always takes place in a region of vicinity to the microphone (illustrated in the Figure 1) and,
- valid spoken commands are always directed to the desktop microphone.

Under the assumptions outlined, desktop speech recognition is *not* independent of talker location. Speech commands that are classified as being acquired outside the region of vicinity are likely to be spurious.

The amount of reverberation in the acquired speech signal increases with distance of the speech source (i.e the talker) from the microphone, in a given environment. A measure proportional to the reverberant content can discriminate if the talker is speaking into the microphone, or further away from it. Such a measure can be then used aposteriori by the recognition search result to reject the end-pointed speech as spurious, or accept it as a valid spoken command.

The semantic nature of the spoken command plays an important role when combined with such a measure. For example, a spoken command such as *Open this file* is likely to be be spoken with the user present at the desktop, whereas, the command, *What time is it?* does not place a locational constraint on the talker. The subsequent section outlines the approach.

2. NEURAL NETWORK BASED SPURIOUS UTTERANCE DETECTION

Feed-forward neural networks trained on extracted features, using the back propagation algorithm, have been shown to perform well in classifying complex nonlinearly separable tasks [5,3].

Generally, features extracted for speech recognition are multidimensional vectors that contain spectral information in the speech signal [2]. These spectral features also implicitly contain information on the amount of reverberation in the acquired speech, the signal-to-noise ratio and the distance from the source. These attributes present in the speech signal could be used to determine the approximate location of the talker.

Typical desktop environments are offices, rooms and auditoriums which have long impulse responses of order greater than hundreds of milliseconds. If the acquired speech signal is highly reverberant then it can be classified as being spoken in the far-field of the desktop microphone. A feed-forward neural network can be trained to approximate a decision surface which allows for the discrimination, as illustrated in Figure 2.

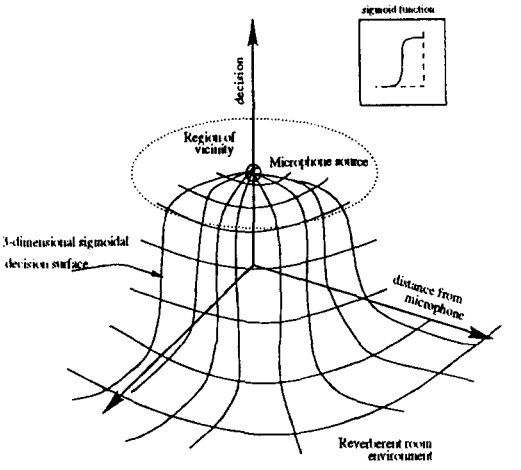

Figure 2: Decision surface approximated by the neural network.

Cepstral features that are used in the speech recognition process are used for classification. The premise of using cepstral features is that they implicitly contain an additive bias component that exists due to the convolutional effect of the impulse response of a reverberant environment [6]. Such a bias term implies a translation in the cepstral feature space, that corresponds to reflective reverberation components in the signal.

Cepstrum based deconvolution has been widely used to compensate for transmission channel degradations and for speech dereverberation [7]. In a typical reverberent environment, the reflective components are less dominant if the speech is directed to the microphone in its vicinity.

Cepstral features extracted from the speech signal are labeled as belonging to two classes,

- undirected speech (in the far field with a label 0.0) and,
- valid directed speech (in the near field as label 1.0).

The neural network is trained to discriminate between these two classes with a high score (1.0) if speech is uttered in the near field and a low score (0.0) otherwise. The output of a neural network trained this way represents the aposteriori likelihood that an observation occurred in the vicinity of the microphone.

The neural network output observed and accumulated over several frames of the acquired speech signal indicates an overall confidence that the source of speech was in the vicinity of the desktop microphone.

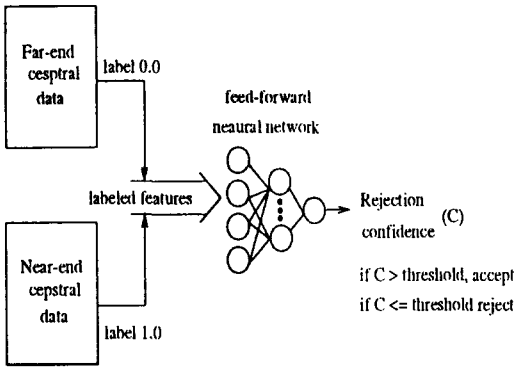

Figure 3: Neural network configuration.

Figure 3 shows the neural network configuration used.

3. EXPERIMENTS

Approximately 5 minutes of speech were collected from each of the 10 speakers in an typical office environment, with an Apple Plaintalk Microphone mounted on top of the computer monitor.

The data collection was carried out independently for the two classes. For one class valid in-grammar speech was collected within a two feet radius of the desktop, directed to the microphone (in near-field).

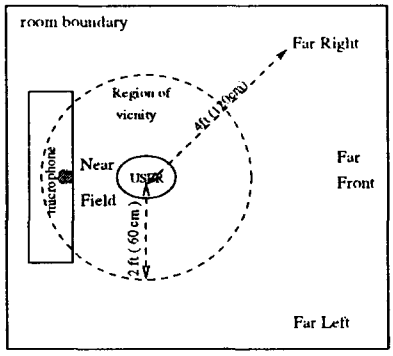

Figure 4: Data collection locations (Top view).

Subsequently, out-of-grammar conversations and spurious conversations of random text were collected away from the desktop (in far-field). Conversations were acquired from three different location points in the room. The locations were approximately four feet to the left, right and front of the desktop and not directed to the microphone, as illustrated in Figure 4, by a top view of the room.

The features extracted in either case were cepstral features weighted on the mel frequency scale (MFCC) [1], derived from a pre-emphasized speech signal, sampled at 16 KHz and analyzed at a frame rate of 10 msec.

A feed-forward neural network with 13 input nodes corresponding to the 13 dimensional MFCC vectors, 10 hidden nodes and 1 output node was trained using the back-propagation algorithm until convergence (less than 1% error).

100 test utterances were collected as near field spoken commands and spurious far-end conversation. Each frame from the test utterance was retrieved through the neural network and the output scores recorded over the entire utterance for various decision thresholds. Each frame was allocated a separate count for scores below or above the arbitrarily fixed threshold. The corresponding likelihood of the utterance being in the near field was computed as,

$$C_{TH} = P(\hat{x}) = \frac{M}{M+N} \quad (1)$$

where M is the number of frames classified as being in the near field, N is the number of frames classified as being in the far field, and $N + M$ is the total number of frames in the test sentence.

The probability in the equation represents the likelihood (C) that the sequence of vectors $\hat{x}$ were generated in the near-field given a threshold TH. Figure 5 plots the neural network outputs of a test sentence spoken from the four different locations. The neural network output for each frame of speech is represented by a point in the figure at the appropriate distance from the microphone.

A ROC curve was plotted to observe the False rejection and False acceptance rate for several different thresholds, as shown in Figure 6. In preliminary experiments the neural network offered an equal-error-rate (EER) of 21.5%. This implies that the rate of falsely accepting a person uttering a far-end conversation, or the rate of falsely rejecting a person, speaking into to the microphone, is 21.5%.

One can adjust the performance curve for the neural network by varying the decision threshold, to falsely accept no utterances that take place further away from the desktop while sacrificing on an increased false rejection rate. The score of the neural network can be used aposteriori to improve upon a false accept of reject decision returned by a conventional recognition system.

4. CONCLUSION AND FUTURE WORK

A neural network approach to measure reverberation content in a speech signal was proposed. Preliminary results corroborate that reverberation information con-

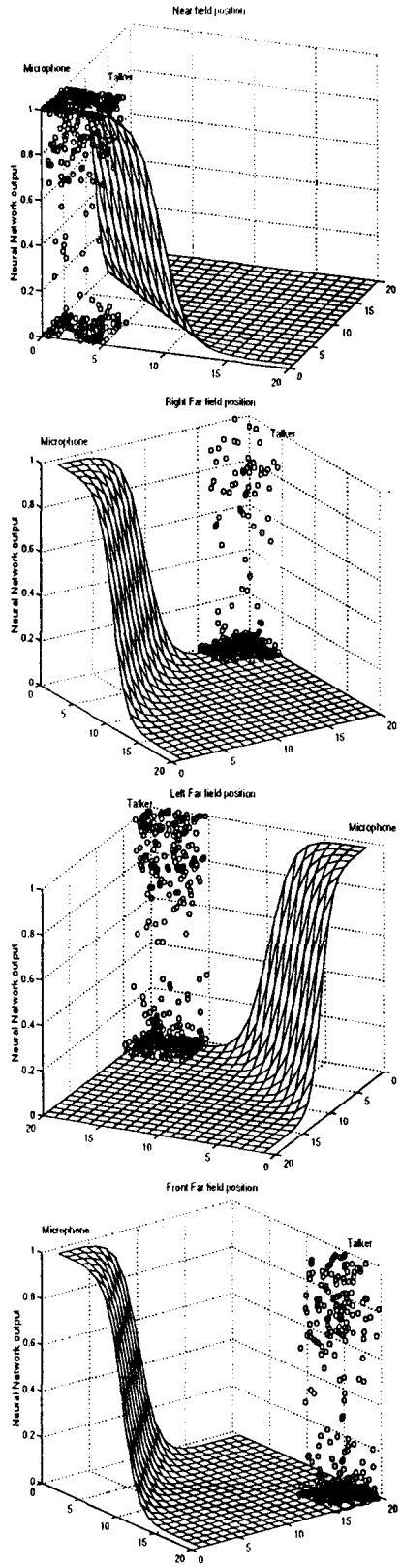

Figure 5: Neural network outputs for speech at different locations. X and Y-axes represent distances from the microphone (x 0.2 feet). Top view shown in Figure 4.

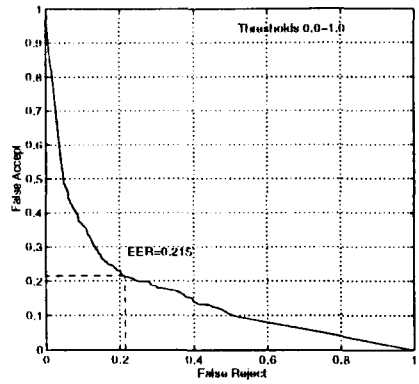

Figure 6: ROC curve outlining False Accept/False reject rate.

tained in the cepstral features can be used to approximate the talker location. This information can be efficiently used to control the misfire rate in a hands-free desktop command and control application.

Future work involves combining this approach with speech recognition for several reverberation environments and provide improved rejection modeling. Applicability of this approach for environment adaptation and speech dereverberation will also be investigated.

5. REFERENCES

1. S. Davis and P. Mermelstein. "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences.", *IEEE ASSP*, 28:357-366, 1980.

2. L. Rabiner and B. Juang. "Fundamentals of Speech Recognition", *Prentice Hall*, Englewood Cliffs, NJ, 1993.

3. R. Duda and P. Hart. "Pattern Classification and Scene analysis", *Wiley*, NY, 1973.

4. C. Bishop. "Neural Networks for Pattern Recognition", *Oxford*, NY, 1995.

5. D. Rumelhart and J. McClelland. "Parallel and Distributed Processing Vol. 1",*MIT Press*, MA, 1986.

6. B. Atal. "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification", *JASA*, 55:1304-1312, 1974.

7. S. Subramanuim, A. Petropulu and C. Wendt. "Cepstrum-Based deconvolution for speech dereveberation", *IEEE Trans. Speech and Audio Proc.*, 4:5:392-396, 1996.