SPEECH RECOGNIZER FOR VOICE CONTROL OF MOBILE TELEPHONE

M. Blomberg, K. Elenius*, B. Lundström[x], L. Neovius**

* *Department of Speech Communication and Music Acoustics, KTH, Box 70014, S-100 44 Stockholm, Sweden*

[x] *Ericsson Radio Systems AB, S-163 80 Stockholm, Sweden*

ABSTRACT

Infovox is marketing a speaker-dependent, pattern-matching word recognition system, developed at KTH. The algorithms in the system have been modified for noise immunity, and performance has been evaluated in moving cars. The main problems were word detection and noise compensation. After simulations we decided to use a close-talking microphone and a "noise addition" method, where we added the measured noise in the moving car to the reference patterns recorded in a parked car. Using this method, the recognition rate was improved from 69% to 97% on a ten-word vocabulary using the best microphone. A more extensive test was performed on the modified recognition system using two cars and twelve speakers, seven male and five female. Most of them were naive speakers. The twenty-word vocabulary contained some confusable words and was trained in a parked car. During 98 sessions, 1,960 words were read under different conditions with an average recognition rate of 86%. With closed windows at 90 km/h the mean was 91%. An open window at the same speed decreased the result to 82%.

INTRODUCTION

In spite of the reported 99+ per cent accuracy of word recognizers together with simplicity of use and attractive pricing, the commercial breakthrough is still pending. One reason could be shortcomings revealed when the systems leave laboratory environments and are exposed to the practical case of, e.g., the voice control of cellular telephones with noisy background, changing microphone distance and different manners of speaking in noise and quiet. Voice control of a vehicular telephone should be a self-evident application for a speech recognition system. Eyes and hands are occupied by driving the car. Simultaneous manipulation of a keypad together with monitoring a dashboard display is a clearly dangerous task for the driver as well as others. There is already legislation in some countries against making telephone calls during driving and other countries will probably follow.

Besides the above-mentioned difficulties, most specification requirements for this application fall well within the performance of a standard isolated-word speaker-dependent recognizer:

- The vocabulary could be restricted to less than 40 words.
- The recognition time is not very critical due to the progress time of the call which is system-dependent and can be tenth of seconds.
- The stringent requirements of simplicity and considerably lower cost than the telephone itself can be met by using standard components like codecs, signal processors and 8-bit CPUs.
- Speaker-dependence and training of the system is acceptable if this can be made in a stationary car.
- If error feedback control is used, recognition accuracy above 90 per cent should be sufficient.

BACKGROUND

Reduction of the noise problem can be done at different stages in the speech recognition process. At the microphone level, high-directivity and close-talk features may be used to improve the signal-to-noise ratio in existing recognition systems. Research with microphone arrays for dynamic focusing to the speaker's position will enable a user to move more freely than with headset microphones or fixed direction microphones (ref 1).

Once the noise signal has been picked up, it will affect the recognition accuracy. To diminish the effects of the noise, one could try to separate it from the speech. This can be done based on statistical knowledge of one of the signals. Methods for doing this has been developed for purposes of speech enhancement (refs 2, 3). Another technique is to adapt the recognition system to the measured environmental noise. The noise can be measured just before and/or after the sampled word. Reference templates trained in a silent environment will have noise added during recognition, to simulate that the training and recognition have occurred under the same environmental conditions. This method has been reported by Klatt and others (refs 4, 5, 6). Still another possibility is to make the recognition algorithm less sensitive to noise. This could be accomplished by giving more weight to high energy regions of the input signal (ref 7).

The word boundary detection problem can be approached by techniques for recognition of continuous speech or by including some samples before and after the detected word endpoints into the sampled words and thus allowing for some uncertainty in the endpoint detection. This method has been used in the Infovox RA-201 and is also later reported by Haltsonen (ref 8). The former method is obviously better, but at moderate noise levels, the latter technique may prove quite adequate and it was also used in this study.

A problem that cannot be solved by the techniques mentioned above is the change of the speaker's voice in high noise conditions. Experiments have shown that this effect can have the same influence on the recognition rate as the noise itself (ref 9).

We decided to test a system using the above techniques of noise adaptation of the reference templates and allowing for some uncertainty in the endpoint detection.

DESCRIPTION OF THE RECOGNITION SYSTEM

The Infovox RA-201 is a speaker-adaptive word recognition system using pattern recognition and dynamic programming. The speech signal is preemphasized by +6dB per octave, sampled by 10 kHz and fed into a NEC 7720 signal processor, which is programmed as a 16 channel filter bank. The filters are spaced according to the critical band scale and the output from the filters is rectified and integrated over 25 ms giving 40 filter sections per second. Seven cepstral coefficients C_1 - C_7 are calculated from the 16 amplitude values. The first coefficient, C_0, that contains the overall energy, is not used, so the cepstral coefficients only describe the shape of the spectrum and are not at all sensitive to the signal amplitude. Every coefficient is represented in one byte. After endpoint detection, the words are linearly normalized to 32 samples giving them a nominal length of 32 * 25 = 800 ms. Each word now occupies 7 * 32 = 224 bytes.

In order to test the proposed noise-adaptive recognition algorithms, we carried out some preliminary experiments on a simulated system using a Data General Eclipse minicomputer. We used recordings of a ten-word vocabulary (a total of 100 test words) by one male and one female
speaker. The recordings were made with three different microphones to compare their performances. We got the best results using a position-adjustable close-talking microphone placed about 10 cm from the mouth of the speaker. In the original Infovox system the endpoint detection was based mainly on the energies in the frequency range from 200 to 500 Hz. The recordings showed that the environmental noise in the car

had a maximum in the same range - especially with all windows closed - making the sampling of the words very hazardous. It turned out that using the summed energy of all filters resulted in a safer endpoint detection.

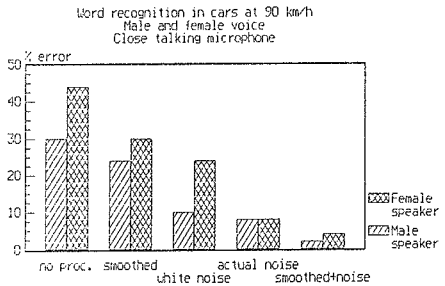

Figure 1. Different noise processing schemes.

The results of some different schemes for noise adaptation are shown in Figure 1 for both speakers. First we see the percentage of errors using the adjustable microphone, the endpoint detection described above and no adaptation to the noise. The amplitude variation between successive filters could be very large. Reducing the difference between filters to 10 dB gave the results shown. Adding white noise instead of the measured noise to the reference templates gave some improvement but not as much as adding the actual noise measured before the sampled word. Finally we see the results using the actual noise plus smoothing. The last results show a substantial improvement of the recognition compared to the first cases.

RESULTS USING TWELVE SPEAKERS

To more extensively test the best method above - noise adaptation plus spectral smoothing - mobile test equipment was developed. The hardware consisted of a standard, transportable PC-compatible computer and an Infovox RA 201/PC speech recognition board. The software was modified to implement the new algorithms. Special testing software was also developed, containing functions for training and recognition with randomized lists, logging and displaying of the results.

The test was run with seven male and five female speakers. It was a rather difficult 20-word vocabulary, with phonetically close pairs like "nio"-"tio", "hem"-"fem" and "Erik"-"Hiby" (nine-ten, home-five, Erik-Hiby), but also some longer words like "Soeharjo" (a name) and "kontoret" (the office). Training the system was made in a parked car repeating each word five times. For safety reasons the recordings were made by the person sitting next to the driver. The speeds varied from 50 to 110 km/h. The conditions also varied with different combinations of window opened/closed, rain or no rain, radio on or off. During 98 sessions, 1,960 words were read under different conditions. The tests were performed without using any sort of rejection of bad matchings and the total number of substitutions were 285, giving a recognition rate of 86%. One-third of the substitutions came from the three phonetically similar word pairs. The mean result was 91% at 90 km/h with closed windows, varying between 78% and 100% per speaker. An open window at the same speed decreased the mean to 82%, the individual results varied between 72% and 97%. We also tested a threshold on the matching distances to reject bad matchings among the 1,960 words. As an example, one of the tested thresholds resulted in 4% substitutions and 22% rejections.

DISCUSSION

We consider the results obtained quite satisfactory, especially when considering the vocabulary. But there are, of course, still some unresolved questions, e.g., weak phonemes at the endpoints can obscure the noise measurements. It is not obvious whether it is better to measure the noise before and/or after a word. Another problem is how to set the energy threshold for endpoint detection - whether it should be fixed or modified by the noise and/or speaking level. The method of rejection could be a fixed threshold rejecting words that have too large a matching distance to the best reference, or a relative threshold rejecting words that have too small a distance to the second best reference. A combination is also possible.

A complication using noise adaptation of the references is that the calculation of the cepstral coefficient of the templates has to be done according to the actual noise of each input word. This means that the reference patterns have to be stored as filter amplitudes, requiring 32 * 16 = 512 bytes per reference compared to 224 bytes when stored as cepstral coefficients. To get a noise-adapted reference, the 16 amplitudes of the noise spectrum have to be compared to each of the 32 spectral sections of the reference template, keeping the maximum of the two amplitudes in each point. To calculate the cepstral coefficients needs a total of 7 * 16 multiplications and additions per template. By using the NEC7720 signal processor for this, the time needed is 12 ms per template which is quite acceptable for a vocabulary of less than fifty words. The time to match references using dynamic programming in the same processor varies between 10 ms and 15 ms.

REFERENCES

1. J. L. Flanagan, "Bandwidth Design for Speech-Seeking Arrays," Proceedings of ICASSP, 1985, Boston.
2. S. F. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction," IEEE Transactions on ASSP, Vol. ASSP-27, No. 2, April 1979.
3. J. E. Porter, S. F. Boll, "Optimal Estimators for Spectral Restoration of Noisy Speech," Proceedings of ICASSP, 1984, San Diego.
4. D. H. Klatt, "A Digital Filter Bank for Spectral Matching," Proceedings of ICASSP, Philadelphia, 1976, pp 573-576.
5. J. N. Holmes, N. C. Sedgwick, "Noise Compensation for Speech Recognition Using Probabilistic Models," Proceedings of ICASSP, 1986, Tokyo.
6. B. P. Landell, R. E. Wohlford, L. G. Bahler, "Improved Speech Recognition in Noise," Proceedings of ICASSP, 1986, Tokyo.
7. H. Matsumoto, H. Imai, "Comparative Study of Various Spectrum Matching Measures on Noise Robustness," Proceedings of ICASSP, 1986, Tokyo.
8. S. Haltsonen, "Improved Dynamic Time Warping Methods for Discrete Utterance Recognition," IEEE Transactions on Acoustics, Speech and Signal Processing, Vol ASSP-33, No. 2, April 1985.
9. P. K. Rajasekaran, G. R. Doddington, J. W. Picone, "Recognition of Speech Under Stress and in Noise," Proceedings of ICASSP, 1986, Tokyo.