Speech Input/Output Assessment
and Speech Databases
Noordwijkerhout, the Netherlands
September 20-23, 1989

TESTING SOME ESSENTIAL PARAMETERS OF A WORD RECOGNIZER USED IN CAR NOISE

Mats Blomberg & Kjell Elenius

Department of Speech Communication and Music Acoustics
KTH, Stockholm

ABSTRACT

A speaker-dependent, pattern-matching word recognition system using dynamic programming has been modified to improve the performance in noise. Problems with word detection and noise compensation have been addressed by using a close-talk microphone and a "noise addition" method. The reference templates are recorded in relative silence. The additional environmental noise during the recognition phase is measured and is "added" to the reference templates before using them for template matching. The recognition performance has been tested in moving cars with references recorded in parked cars. Recordings of six male speakers have been evaluated in this report in an effort to test the sensitivity of the recognition system to some essential parameters.

1. INTRODUCTION

The performance of a speech recognition system may be improved by adapting it to the environmental noise. The noise can be measured just before and/or after the sampled word. In many applications, the reference templates are trained in a silent environment. During recognition, they are modified to simulate that the training has occurred under the same environmental conditions (Klatt, 1976; Landell, Wohlford & Bahler, 1986). A noise compensated reference spectral template is created by the following technique. The amplitude value of each channel in each spectral section of the reference is replaced by the amplitude of the corresponding channel in the estimated noise spectral section if the noise amplitude is higher. This simulates that the recorded noise was present during the training session. After noise compensation, each reference spectral frame can be transformed to another form of representation, e. g. cepstrum.

A problem that cannot be solved by this technique is the change of the speaker's voice in loud noise conditions (the "Lombard effect"). Experiments have shown that this effect can have the same influence on the recognition rate as the noise itself (Rajasekaran, Doddington, & Picone, 1986).

The word boundary detection problem can be approached by including some samples before and after the detected word, thus allowing for some uncertainty in the endpoint detection. This method is used in our system and is also reported by Haltsonen (1985).

2. EARLIER RESULTS USING TWELVE SPEAKERS

We modified a standard word recognition system, the Infovox RA-201, to improve its performance in moving cars in order to make it more suitable for voice controlled dialling in mobile telephony. RA-201 is an isolated word recognition system based on mel cepstrum representation from a 16-channel Bark-scale filter bank. We decided to use the above technique of adapting the reference templates to the environmental noise combined with an adaptive threshold for word endpoint detection and also allowed for some uncertainty in the endpoint detection. A rather extensive test was performed using two different cars and twelve speakers, seven male and five female, most of whom were naive speakers (Blomberg, Elenius, Lundström & Neovius, 1987). The experiments were performed during summer in two different cars. The speakers were not driving the car due to safety reasons, but they were sitting in the front seat beside the driver. They were reading the lists in a slow tempo. The reference templates were recorded in a parked car and the performance was tested in a moving car with a typical signal-to-noise ratio of 25 dB. All training and testing sessions were recorded on tape for later processing. During 98 sessions, almost 2000 words were spoken under different conditions. The average recognition rate was 86% on a 20-word vocabulary. One-third of the substitutions came from three phonetically very similar word pairs. With closed windows at 90 km/h the mean was 91%. An open window at the same speed decreased the result to 82%.

3. EXPERIMENTS AND RESULTS

3.1 Experimental conditions

To optimize the parameters of the algorithms, further studies were made on part of the recorded material. Six male speakers of the twelve speakers above were selected, making a total of 38 test lists. The speech level varied a maximum of 14 dB between the speakers. A total of five utterances were not recorded, so there is a total of 755 test words for our experiments. The speech signal was analyzed by the RA-201 and the resulting filter bank frames were stored onto disk for later use as training and test data. The amplitude level of the material was preserved during the digitization. The beginning and endpoints of the utterances were then marked in a semiautomatic way to facilitate an automatic analysis of the recognition results.

3.2 Amplitude normalization and time position of noise measurement

The object of the noise compensation technique is to make the signal-to-noise ratio of the reference template equal to that of the test pattern. This means that the amplitude of the reference pattern must be changed to that of the input utterance before noise is inserted. In this way, varying speech level can be accounted for. The speech level was represented by the average or the maximum amplitude in the utterance. The estimated noise spectrum used for modifying the reference templates was measured over 8 frames (200 ms) before and/or after each word.

Results for different level normalization techniques and noise measurement positions may be seen in fig. 1. The mean energy is integrated over the input word and is the normalizing parameter in this case. Also shown are results using the maximum energy of the words for normalization as well as results for no normalization. It may be seen that the mean energy gives the best performance.

Varying the position of the measurement of the compensating noise spectrum did not have a very significant effect, as can be seen in the same figure. When measuring both before and after the word we used the minimum value of the two for the compensating noise spectrum. For our following experiments we decided to use the mean energy for level normalization and to measure the noise before the word. This means that it may be done at the same time as the adaptive threshold control, which has to be done before each word.

The reference case without any noise compensation gives very good results, because the noise level in these recordings was quite low. These results are very different from some preliminary experiments we had done during winter-time using noisier winter tires with a signal-to-noise ratio of about 15 dB (Blomberg et al, 1987). We therefore made another experiment where we raised the measured noise spectrum by 10 dB and added it to each input frame before making endpoint detection and noise compensation. We consider this to be an acceptable way of simulating increased noise levels when studying methods for the amplitude normalization and the position for measuring the compensating noise. We see that both level normalizing schemes give better performance than the no compensation case.

3.3 Varying the endpoint detection threshold

Preliminary experiments showed that using an overall energy for endpoint detection improved the performance in car noise compared to the 200 - 400 Hz band energy else used in the RA-201. Five samples (125 ms) before and after the detected endpoints were included in the word to reduce the possibility of missing weak initial and final phonemes as well as providing for some uncertainty in the endpoint detection, as discussed above.

We manually decided the optimal energy threshold for each of the 38 recorded lists. We then made some experiments where we changed the threshold in 3 dB steps from -12 dB to +12 dB relative to the optimal value. No adaptation of the threshold was carried out in this test. Results in fig. 2 show that the performance is almost constant over a 10 dB range.

3.4 Using an adaptive threshold

Giving the endpoint detection threshold the same initial values as above but allowing it to adapt to a level 5 dB above the measured noise level, if this level is higher than the initial value, gives the same performance as the above experiment for values -6 dB to +12 dB (fig. 2). If the initial threshold lies below or just above the noise it will adjust itself to a reasonable value and thus improve the performance. However, if it is set too high it cannot recover using this method that was later used in the mobile system.

It should be pointed out that all values for correct recognition percentages in this report only apply to the 755 tested utterances and do not include extra or missing words. The numbers of these are shown separately in fig. 3. From -3 dB to +6 dB there are not very many of these words. The worst performance is for fixed threshold at -12 dB, giving 1184 extra words and 171 missing. The adaptation technique gives very few missed words for low initial thresholds whereas the number of extra words increases at thresholds below -6 dB.

3.5 Some tests using simulated noise

To test the value of threshold adaptation and noise compensation we tried to provoke the system by artificially increasing the noise level by adding a constant to the measured noise spectrum before each word (section 3.2 above). It is, of course, not possible to use this technique to simulate a lower noise level, since we do not know the spectral shape of the tested word below the actual noise.

We simulated a 5 to 20 dB noise increase in 5 dB steps, using three different combinations of adaptation and compensation (fig. 4). One can observe that adapting the threshold to the noise is very critical for the performance. Noise compensation of the templates increases the performance about 10% at a 15 dB noise increase, which is well in accordance with our experience from the preliminary experiment (Blomberg et al, 1987). The total number of both extra and missed words is very low, between 4 and 6, when using threshold adaptation, for all cases between 0 and 15 dB noise increase.

It is not possible, though, to use this method to study how much above the noise level the adaptive threshold should be set, since the actual noise is replaced by a spectral frame at a constant level having no time varying fluctuations. The best results would be achieved for a threshold just above the steady noise. This level would of course cause several word detection errors in real-noise situations.

4. DISCUSSION

The noise level of the two cars used for the experiment was somewhat to low for judging the optimal values of some of the parameters that were to be optimized. However, simulating a louder noise, gave some important indications.

An adaptive threshold related to the noise level has been shown to be of vital importance for the performance. The optimal distance between the threshold and the noise level could not be determined using the technique of simulated increased noise, but the 5 dB we used gave good results for the real noise level. The noise compensation processing is essential when the noise is 10 to 15 dB louder than in our experiments. In this context it should be noted that in a real case, both the microphone distance and the speaking level will change more than under the rather controlled circumstances during these tests and make the signal-to-noise ratio more variable.

Amplitude level normalization is important before making the noise compensation. The mean energy over the utterance gives the best results. It does not seem to be very essential whether the noise used for compensation is measured before and/or after a word. Practical considerations makes it better to do the measurement before each word, at the same time as the adaptive threshold calculations.

ACKNOWLEDGEMENTS

This work has been performed at KTH in cooperation with Infovox AB and Ericsson Radio Systems AB. The recordings were made by Ericsson.

REFERENCES

Klatt, D. H. (1976): "A Digital Filter Bank for Spectral Matching," *Proceedings of ICASSP*, Philadelphia, 1976.

Landell, B. P., Wohlford, R. E. & Bahler, L. G. (1986): "Improved Speech Recognition in Noise," *Proceedings of ICASSP-86*, Tokyo, 1986.

Rajasekaran, P. K., Doddington, G. R. & Picone, J. W. (1986): "Recognition of Speech Under Stress and in Noise," *Proceedings of ICASSP-86*, Tokyo, 1986.

Haltsonen, S. (1985): "Improved Dynamic Time Warping Methods for Discrete Utterance Recognition," *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol ASSP-33, No. 2, April, 1985.

Blomberg, M., Elenius, K., Lundström, B. & Neovius, L. (1987): "Speech Recognizer for Voice Control of Mobile Telephony," *Proceedings of ESCA-87*, Edinburgh, 1987.

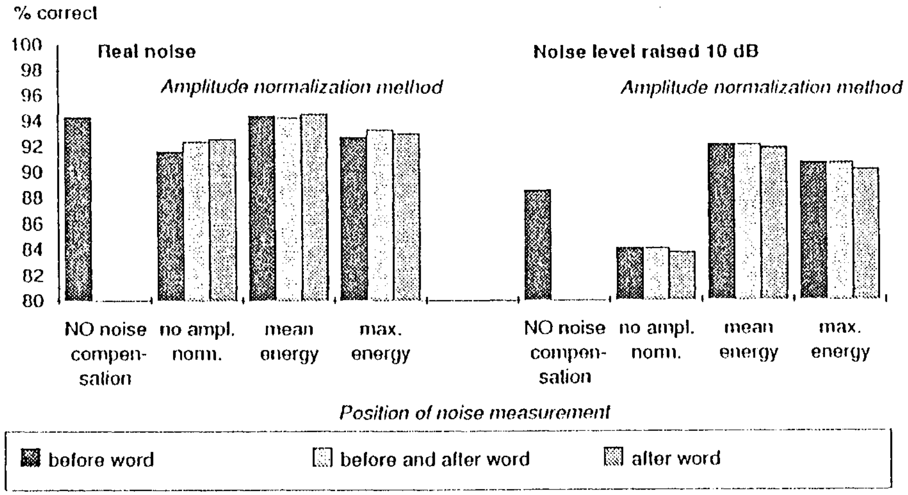

Figure 1. Recognition rates with different amplitude normalization techniques and varying time position of the noise measurement. The noise level is artificially raised on the right hand side.

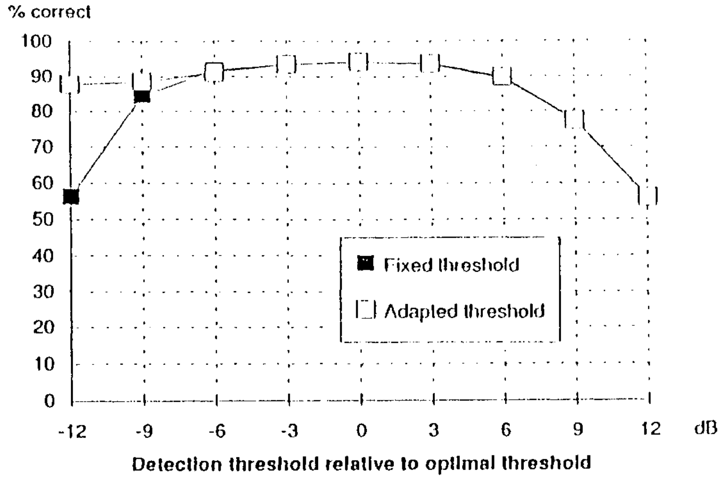

Figure 2. Recognition performance as a function of fixed or adapted endpoint detection threshold.

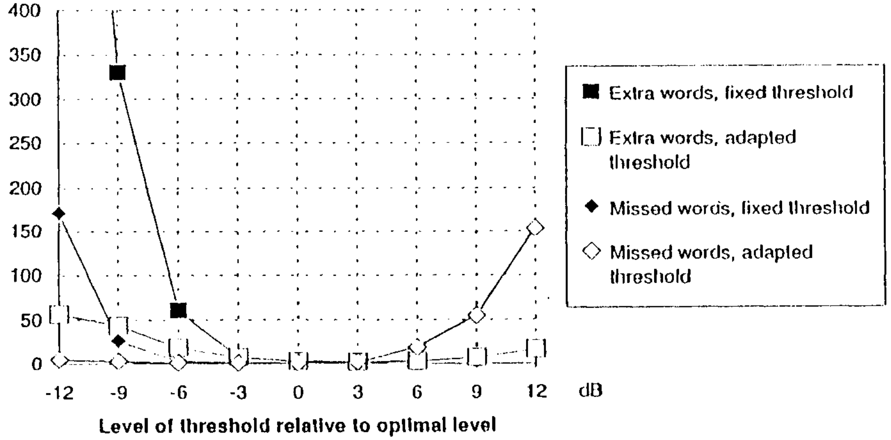

Figure 3. Number of extra and missed word as function of fixed or adapted endpoint detection threshold.

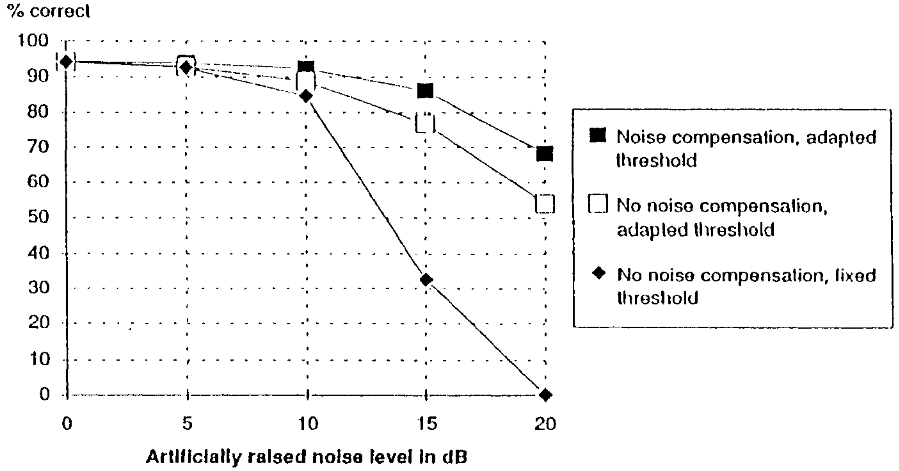

Figure 4. Recognition performance as the noise level is artificially raised in 5 db steps. Three different cases as explained in figure legends.