III. SPEECH RECOGNITION

A. NOISE CANCELLING MICROPHONES FOR AUTOMATIC SPEECH RECOGNITION *

H. Sohlström

Abstract

Automatic speech recognition as well as man-to-man communications in noisy environments require noise cancelling microphones. A number of such microphones are studied. Special attention is given to a contact microphone. The test procedure is described and the results are discussed. The contact microphone is found to give better sound quality than expected.

Introduction

Automatic Speech Recognition systems are leaving the laboratory stage. Several systems are today commercially available[6]. If a system is to be useful its operation must be unaffected by background noise. This can be achieved in three ways. The first - and best - way is to reduce the noise level. The second is to use a noise cancelling microphone. The third is to extract the phonetic information from the waveform in a way that makes the system immune to noise[2].

Noise cancelling microphones have been used for a long time in man-to-man communications. The situation is, however, somewhat different in speech recognition systems depending on the special set of parameters adopted[7].

In the present study several noise cancelling microphones were tried both for man-to-man communication and speech recognition. One of the microphones was of the contact type, i.e. it was to be fixed upon the speaker in order to pick up vibrations rather than sound. This type of microphone is used in very noisy environments, for example in aircrafts. Because of the different principle used in this microphone special interest was paid to it.

The speech recognition system was a phonetically oriented system developed by Mats Blomberg and Kjell Elenius. A short description of the system is given in Blomberg and Elenius, 1978[1].

* Thesis work 1977 under supervision of Mats Blomberg and Kjell Elenius.

The Microphones

The different types that were examined in the study were chosen among those available at the Dept. of Speech Communication. The microphone normally used with the recognition system was a Sennheiser MD 421 which is a commonly used dynamic cardiod microphone. This was of course included in the study together with the more special types of noise cancelling microphones. They were represented by a headset microphone from Sennheiser and a microphone capsule, Hosiden KUC 7001. The microphone from Sennheiser was of the dynamic type while the Hosiden capsule was of the electret type.

The contact microphone was developed by Peter Branderud and Jan Gauffin at the Dept. of Speech Communication and is produced by Special Instrument AB, Stockholm. It is a simple piezoelectric accelerometer mounted in a round capsule with a diameter of 15 mm and a thickness of 6 mm.

Measurements

What measurements could give relevant information about the microphones? The measurements that were made could be divided into three groups.

The first one was the "classical" frequency response measured in an anechoic room at different directions to the sound source in a horizontal plane. The measurements were also repeated at several distances from the sound source. At a distance of 2 m the sound wavefront could be considered to be plane. The shortest distance at which measurements were done was 1 cm. At this distance the microphone definitely "sees" a spherical sound wavefront. Some of the microphones were of the differential type. This means that they suppress plane sound waves compared with spherical. To give a measure of this effect the microphones were subjected to the same sound pressure from a distant and an adjacent source.

Since the contact microphone responses to vibrations, its frequency response was recorded using an accelerometer calibrator.

The noise cancelling microphones are designed to be used very close to the mouth of the speaker. Ideally, it should also be tested in this position. The ideal test signal is speech. In the second group of measurements the microphones were tested in the way they are designed to work - at the right distance and with speech.

Speech is by no means a regular source of sound. To allow correct comparisons we made simultaneously recordings from two microphones of a person reading a short test. The recordings were made in an anechoic room. One of the microphones was a pressure sensitive dynamic microphone from Sennheiser, MD 211. This microphone was the "reference" high quality microphone with which all the other were compared. The other recording was made with the microphone under test.

The average amplitude distribution as a function of frequency for the two recordings was then computed. This was done with our CD 1700 computer and a 51-channel spectrum analyzer. The differences between the distributions could then be interpreted as deviations from ideal responses.

To permit analysis of the separate speech sounds, as they were transduced by the microphones, recordings of VCV and CVC words were made. Also in this case each microphone was compared to the reference microphone.

For the contact microphone the measurements in the second group proved much more relevant. Its response was very much dependent upon its position on the speaker. Several positions were tried. Two positions were found to be representative, each in its own way. The two positions were on the forehead and on the neck just under the chin and halfway towards the ear, Fig. III-A-1. If the microphone had been positioned closer to the larynx it would only have picked up a signal dominated by the fundamental and much of the formant pattern would have been lost.

The third group of the measurements were performance tests using the speech recognition system.

The recognition system works with isolated words. A standard vocabulary of 41 words was chosen. The words in this vocabulary are the words used in Swedish, when spelling out words over the telephone (Adam, Bertil, Cesar, etc.), the numbers 0-9 in Swedish and the words "miss" and "mellanslag" (space). This vocabulary was spoken five times with each microphone. A reference recording was made simultaneously with the MD 211 microphone mentioned earlier.

Recordings were made both in an anechoic room and in a normal room where tapes with different kinds of noise were being played back. The noise level was up to 90 dB (lin)[5].

It should be mentioned that the words were read by the author, unfortunately in a rather hoarse voice. This accounts for the overall low recognition rates.

Before the actual recognition test with each microphone, the system had to "listen" to a number of repetitions of the vocabulary in order to extract statistical information about formant freuuencies, sound duration etc. This information is used in the recognition process. This procedure will be referred to as "learning". When the system was to operate in noise, the learning could be done either in silence or with the noise used. Both cases were studied.

Results and discussion

The measurement on the Sennheiser MD 421 did not give any surprising results. The frequency response for different directions to the sound source is shown in Fig. III-A-2. The rejection of sounds from the rear of the microphone is a bit uneven over the frequency range, but as this microphone is not designed for use in noisy environments this is of little importance. As can be seen from Fig. III-A-2 the response rises some 10 dB from the low end of the spectrum to the high. This changes as the microphone is moved closer to the sound source. At a distance of 1 dm the response is well balanced. The tests with speech do not give any more information about the microphone. The microphone is a good example of a dynamic cardiod microphone.

Sennheiser headset microphone is a differential microphone, designed to be used very near the speaker's mouth. If it is used far from the sound source it has a very uneven frequency response, rising steeply with frequency. Sennheiser has published diagrams showing a rather flat response at a distance of 1 cm.

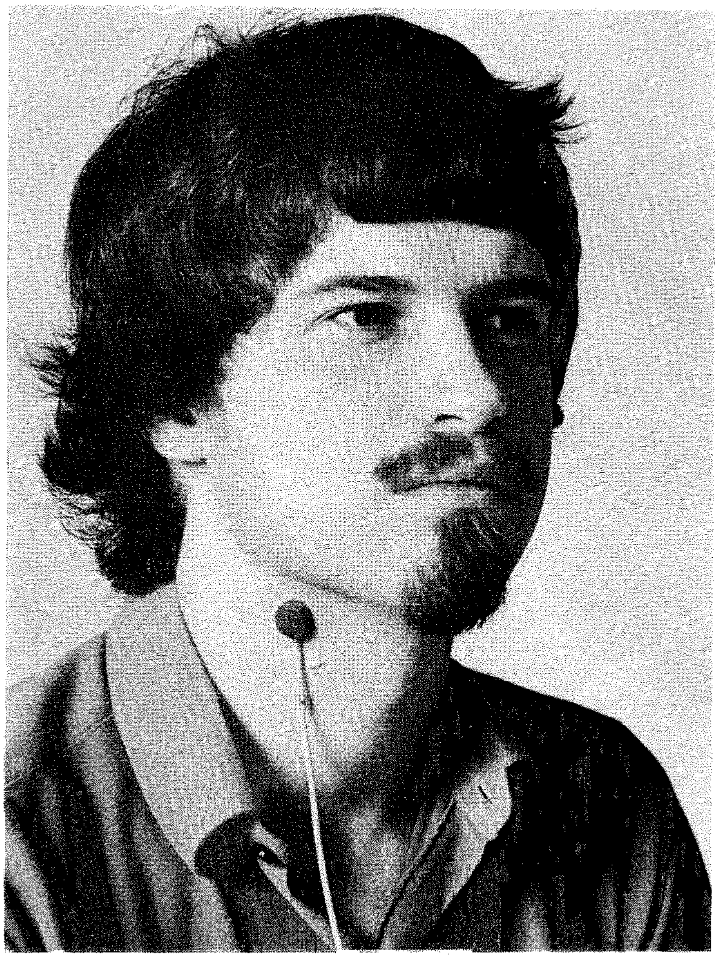

Fig. III-A-1. Microphone position on the neck.

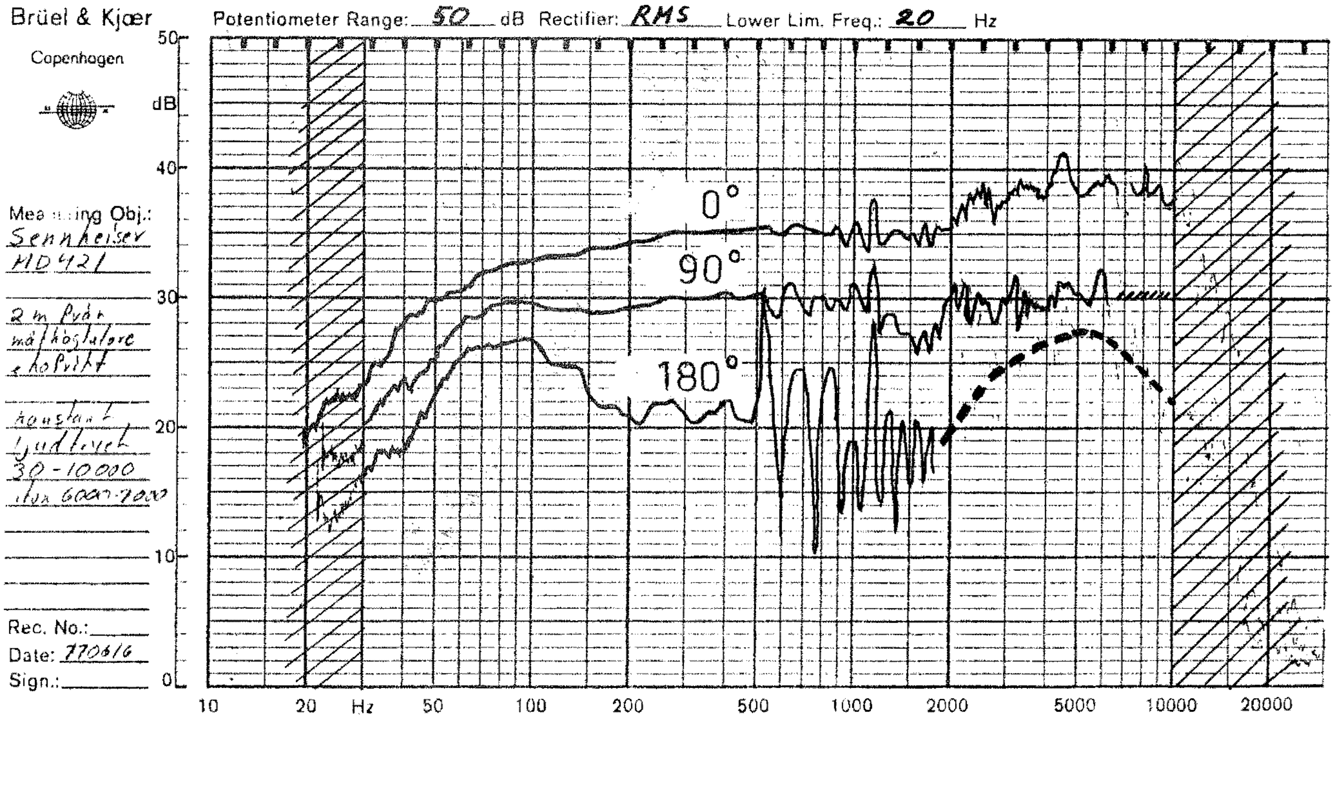

Fig. III-A-2. Frequency response, long distance, MD 421.

For our microphone this could not be duplicated with the test setup used. The sound source was a Brüel & Kjær Artificial Voice 4215 modified to make the "mouth opening" smaller, more like the recent 4219. The sound pressure was held constant with the aid of a measuring microphone, controlling the output from the generator. The result obtained can be seen in Fig. III-A-3. The response is far from flat. There is apparently a strong resonance in the microphone at about 8 kHz. The rejection of sounds from distant sources is good. It varies from 10 to 30 dB through the audible range.

The average spectrum for the short text confirms the impression from Fig. III-A-3, see Fig. III-A-4.

A closer examination of the different speech sounds revealed some breath noises but this is almost unavoidable with close talking microphones. The breath noises showed up as an increased low frequency level.

The KUC 7001 microphone has a frequency response that looks far better than that of the Sennheiser microphone. Fig. III-A-5 shows the response 1.0 cm from the sound source. The rejection of distant sound sources is about the same as that of the Sennheiser headset microphone. More breath noises could be heard with this microphone than with the preceding one. There are two possible explanations for this. This microphone has a better low frequency response and the noises can therefore more easily be heard. The second possible reason is that it is in fact only a "naked" microphone capsule without any protective screening against the air stream from the mouth. The average spectrum confirms that this microphone gives a good reproduction of speech.

As mentioned above, the position of the contact microphone has a great influence on the results. The response for speech transmitted through the tissues of the neck or face is selectively frequency dependent. The damping seems to be greatest in the soft tissues, especially for high frequencies. The bone structure of the face seems to have a much lower damping. This agrees well with what has been reported by others.

In Fig. III-A-7 the average speech spectrum with the microphone on the forehead is compared with the reference condition. The

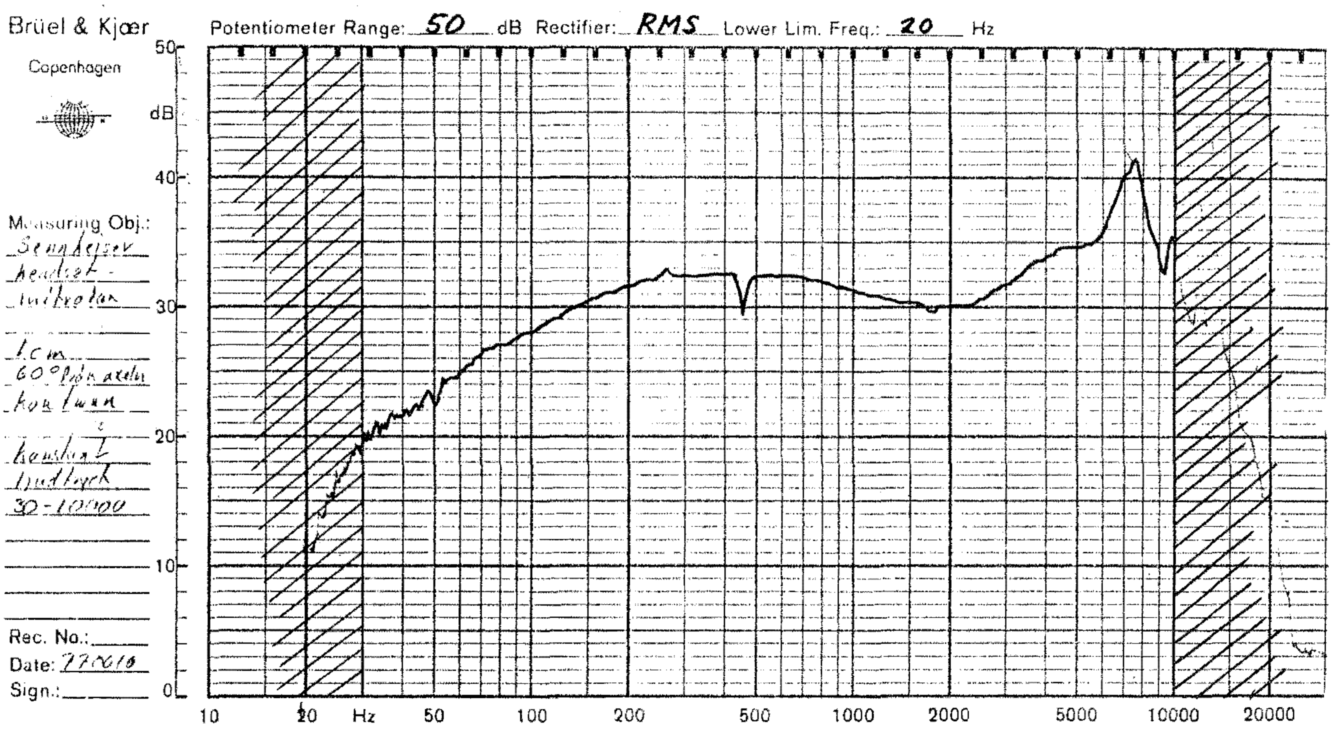

Fig. III-A-3. Frequency response, 1.0 cm from the sound source. Sennheiser headset microphone.

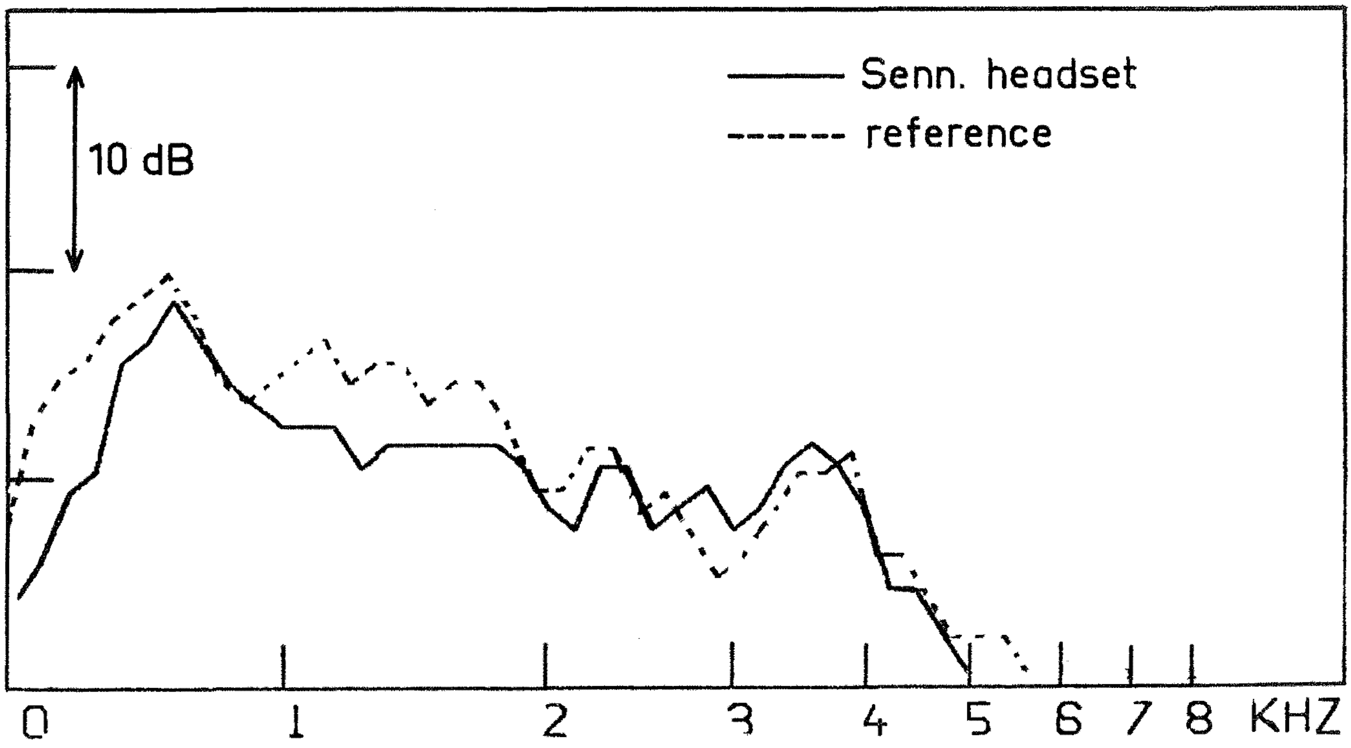

Fig. III-A-4. Average spectrum, Sennheiser headset microphone.

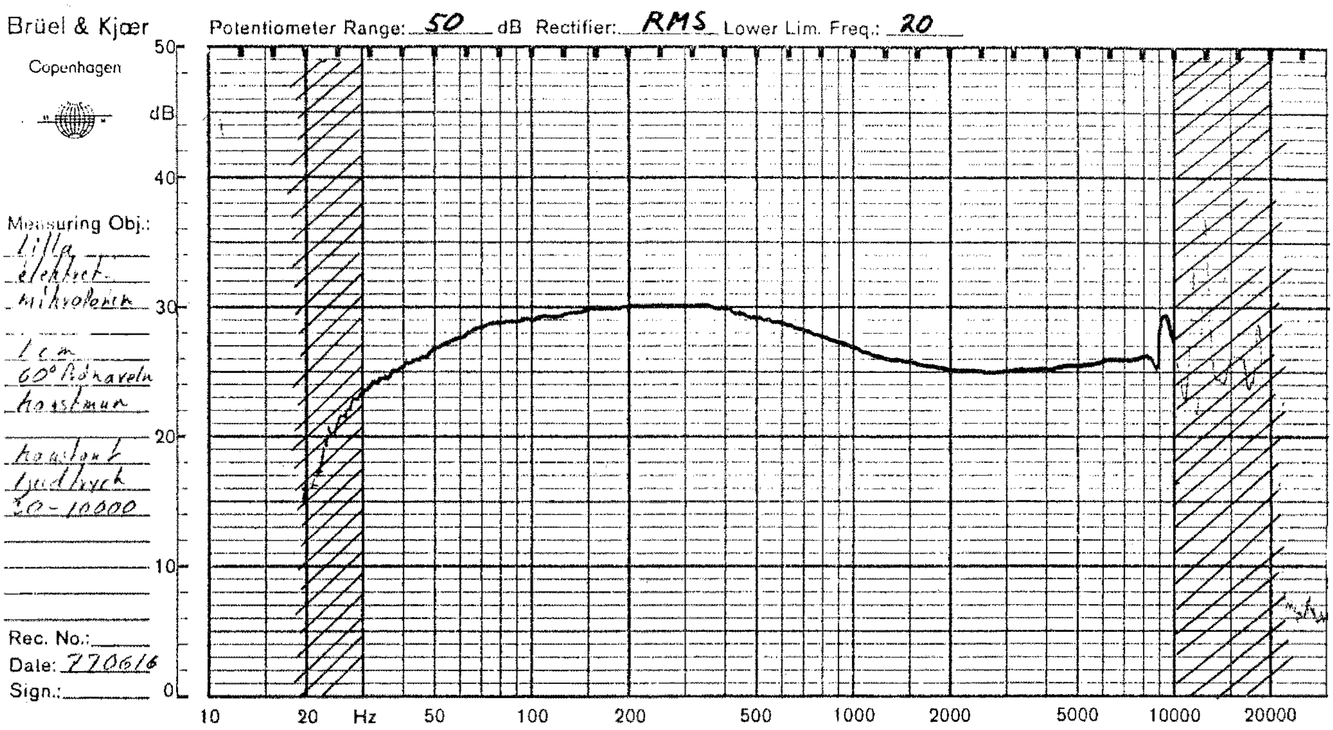

Fig. III-A-5. Frequency response, 1.0 cm from sound source. KUC 7001.

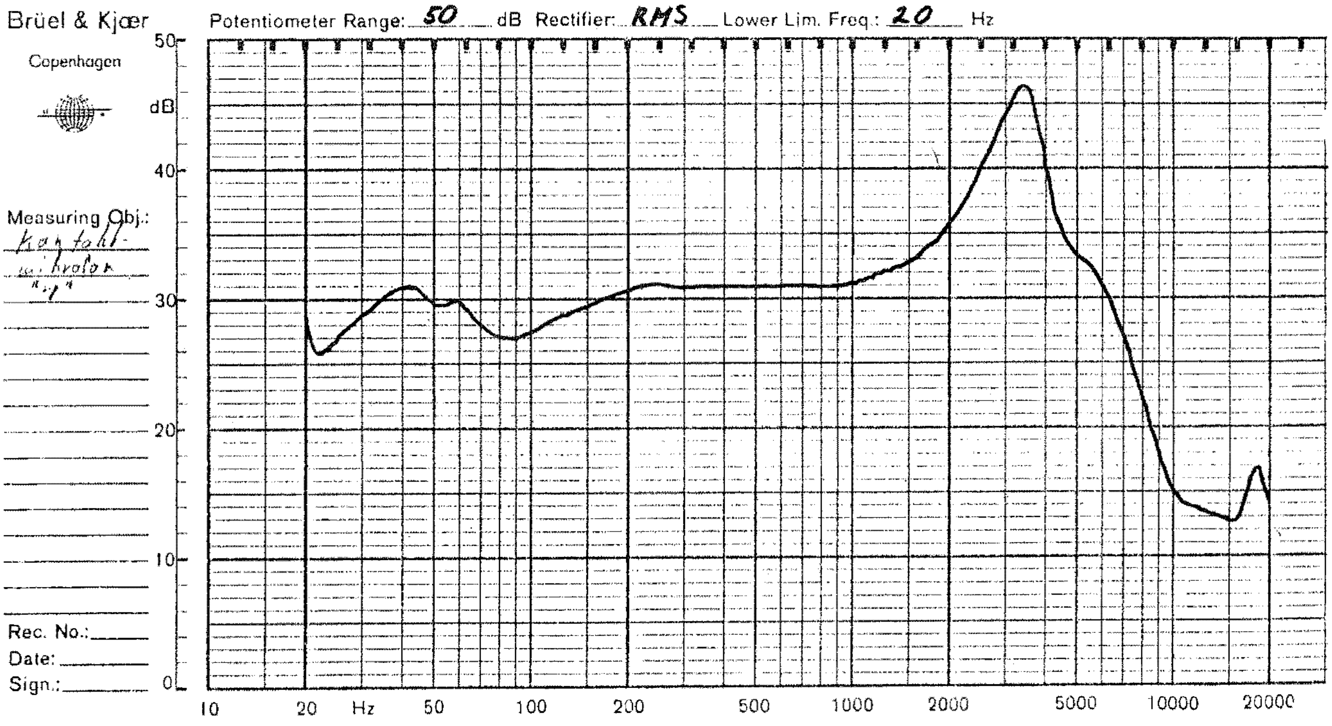

Fig. III-A-6. Frequency response, contact microphone.

difference clearly reflects the 3.5 kHz peak of the contact microphone in Fig. III-A-6. This would imply that the vibrations on the forehead have a spectrum not too different from that of the sound transmitted through the mouth. While this is true on the average there are gross deviations for separate speech sounds.

With a contact microphone the "output" from the speech apparatus is not taken from the normal place. Variation in formant amplitude pattern may thus be expected.

The greatest difference was found in the nasals. Fig. III-A-8 shows the spectrum of [n] in the word [ɑ:na]. The high level of the upper formants is expected, considering the position of the microphone close to the nasal cavities.

A similar change takes place with [s]. As can be seen from Fig. III-A-9 the low frequency limit is approximately 2 kHz with the contact microphone, while in the reference recording this bound is in the vicinity of 3.5 kHz. A probable explanation would be that the normal 3.5 kHz limit is set by a high pass filtering at the passage out of the mouth. The contact microphone, when placed on the forehead receives its signal, probably to a large extent from pressure variations in the mouth. In this way the high pass filtering of [s] is bypassed.

Vowels are affected in a similar manner. Formants that are associated with high energy in the mouth cavity would be dominant at the forehead also. Examples of spectra for a much affected [o:] and an only slightly affected [y:] vowel is shown in Fig. III-A-10 and III-A-11. The interpretation is, however, not quite evident.

All vowels recorded with the contact microphone seem to have an extra resonance, "formant", at 2.2 kHz. Fig. III-A-12 shows an example of this. The resonance at 2.2 kHz is not the normal third formant. F3 was very weak in the reference recording.

The cause of this "extra formant" could be a resonance in the bone structure of the forehead. A study of these resonances and how they work in speech and singing would be interesting[3,4].

When the contact microphone is put on the neck, the response is different. The average spectrum is shown in Fig. III-A-13. This

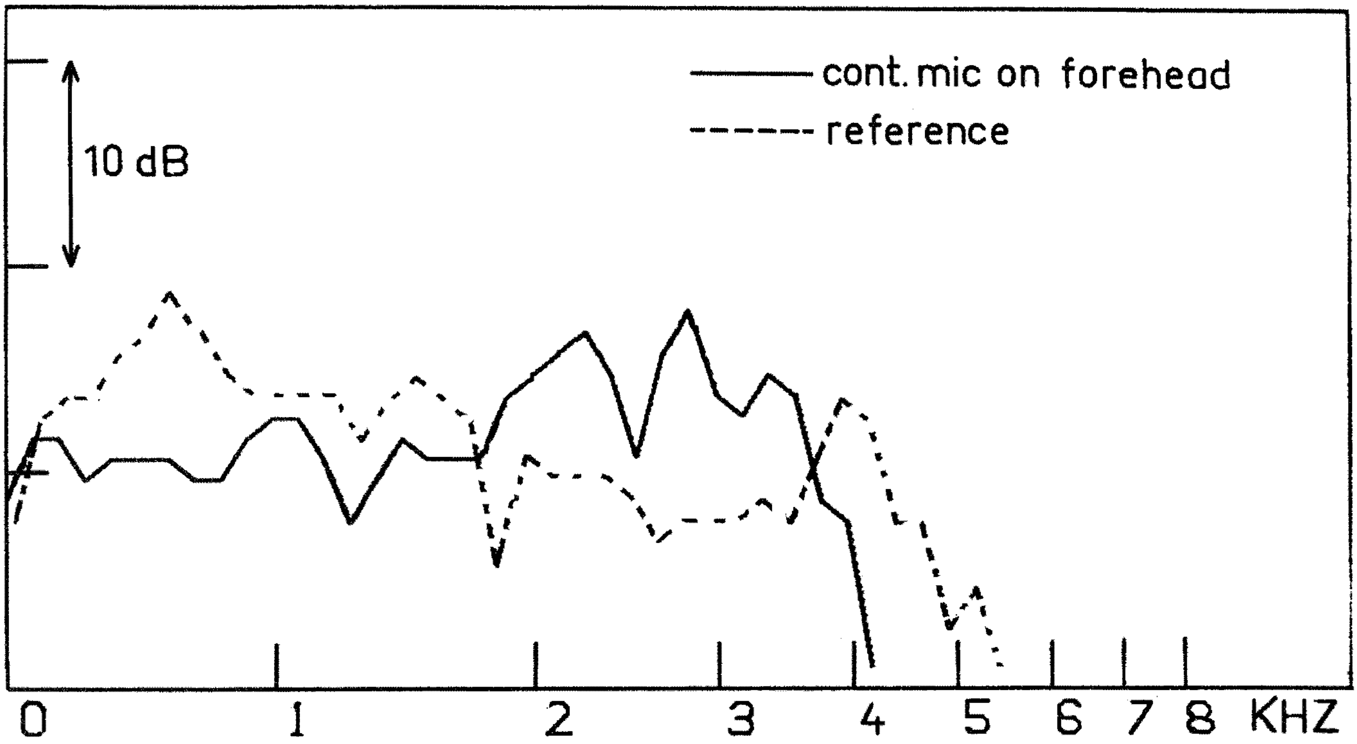

Fig. III-A-7. Average spectrum, contact microphone on forehead.

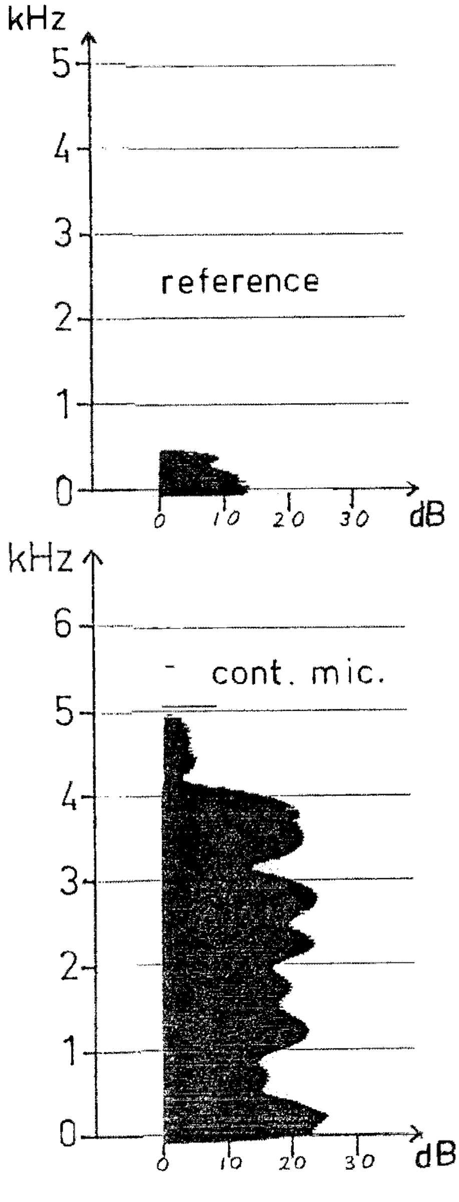

Fig. III-A-8. Spectrum for [n], contact microphone on forehead.

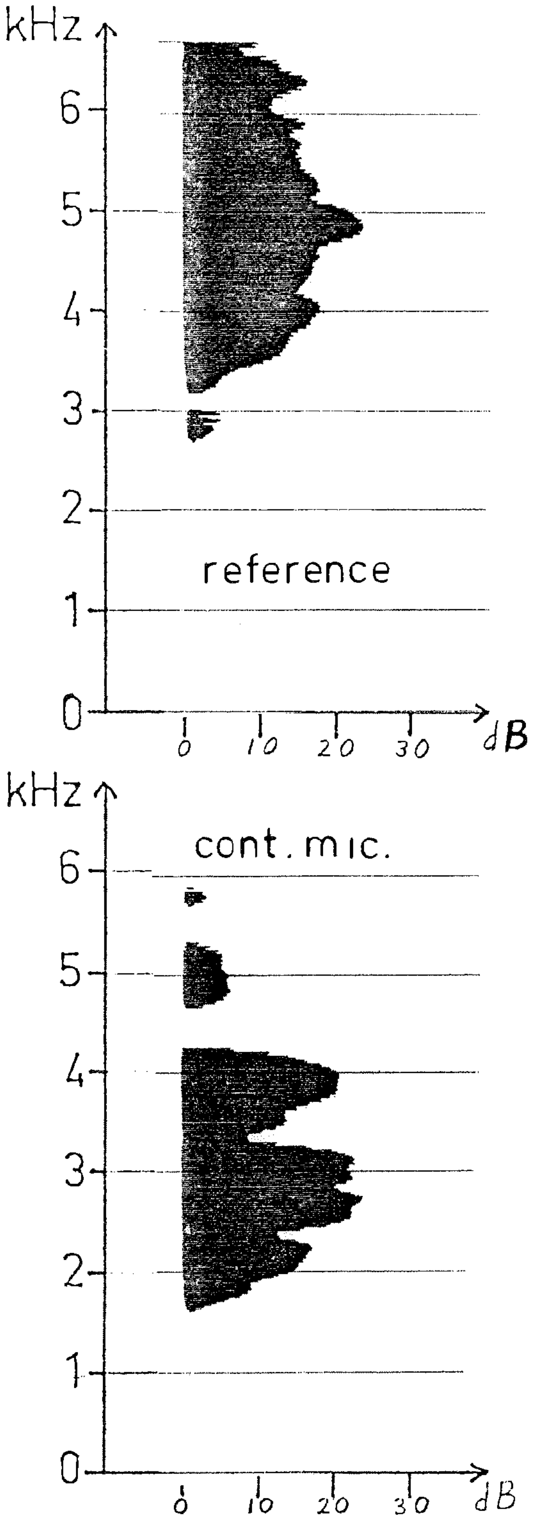

Fig. III-A-9. Spectrum for [s], contact microphone on forehead.

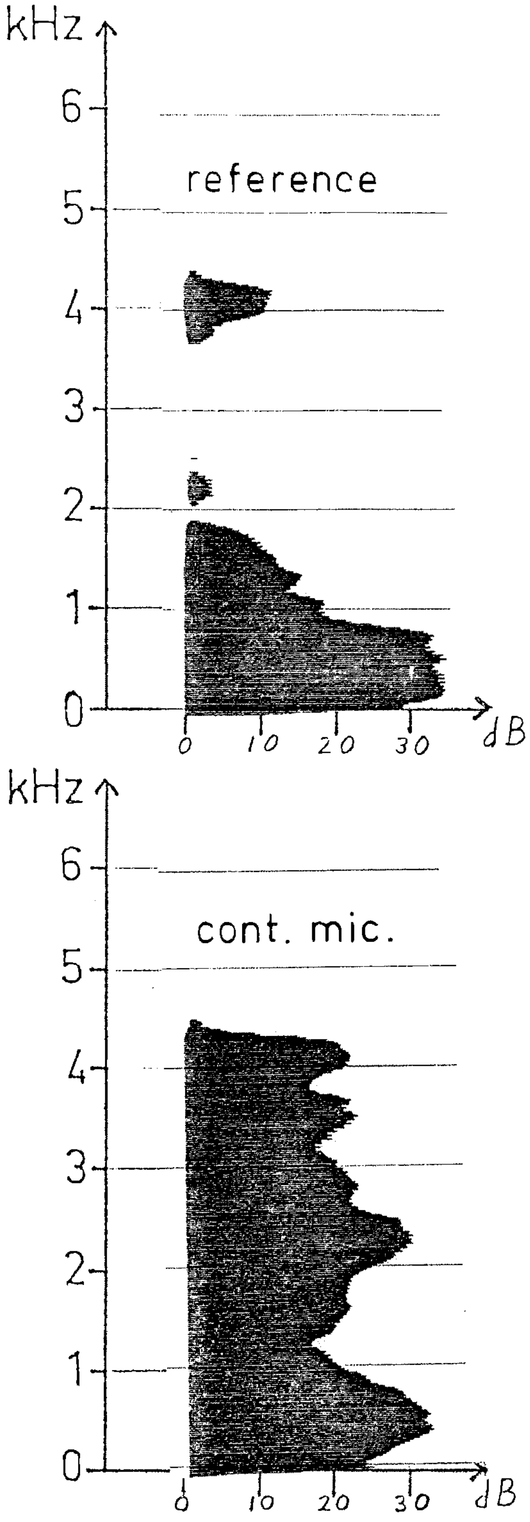

Fig. III-A-10. Spectrum for [o:], contact microphone on forehead.

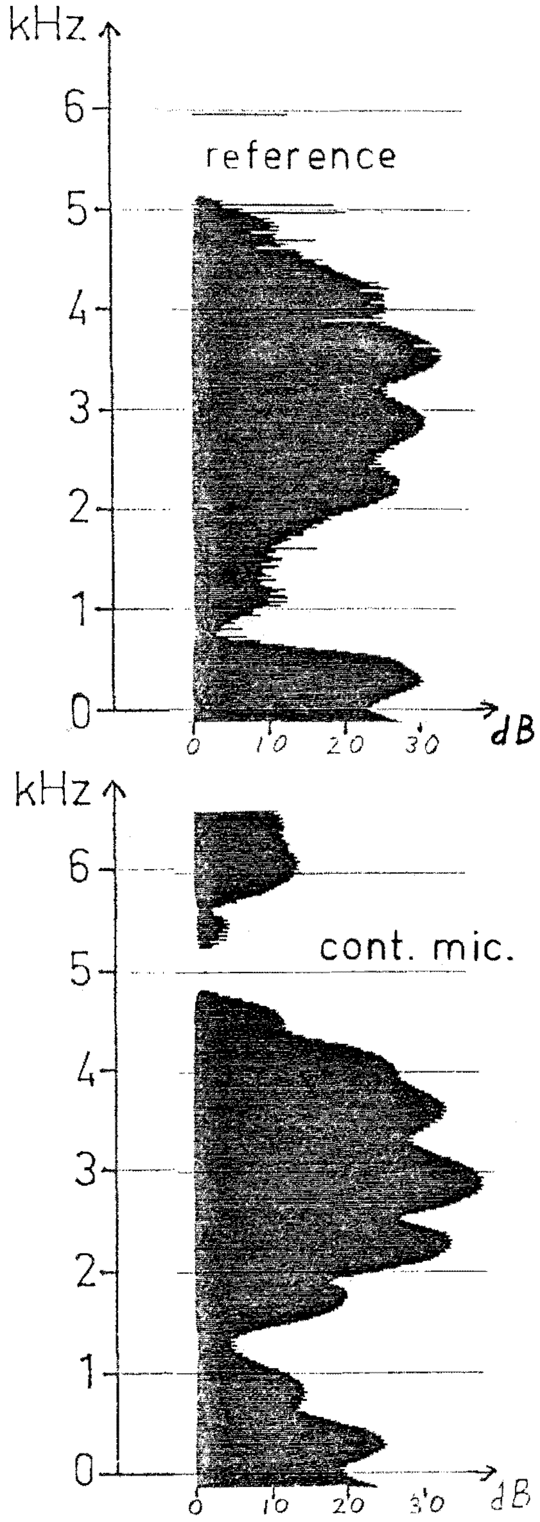

Fig. III-A-11. Spectrum for y: , contact microphone on forehead.

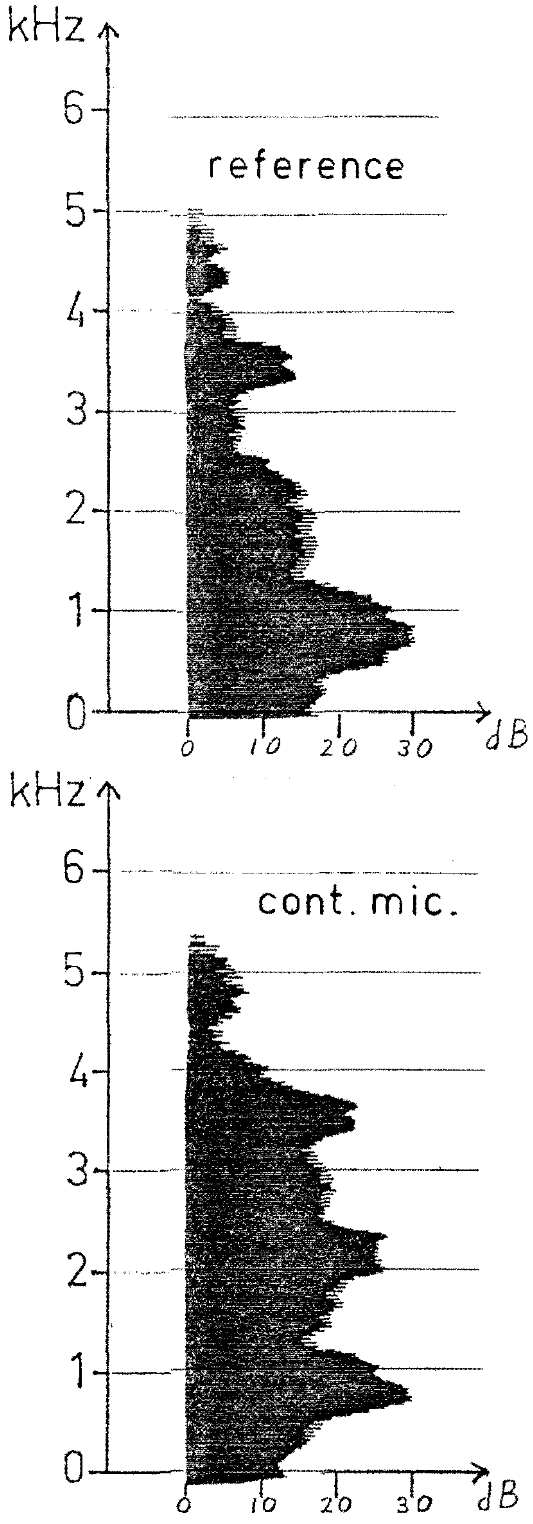

Fig. III-A-12. Spectrum for [ɑ:], contact microphone on forehead.

differs from that of the reference microphone. A standard pre-emphasis filter with a slope of +6 dB per octave from 200 Hz to 5 kHz was tried to compensate for the fall-off of the spectrum. With this filter the spectrum looked like Fig. III-A-14. Without it the speech sounded strange and was hard to understand. With the filter the sound quality improved. The result was better than with the microphone placed on the forehead. The formant pattern showed less distortion. Unfortunately, the fricatives and plosive bursts were weak. The rejection of sounds transmitted through the air was also better with the microphone in this position due to the higher signal level. From most points of view this position was more desirable. For man-to-man communications the sound quality was good enough, but of course not as good as with an ordinary microphone. At high noise levels, however, the rejection of noise is more important for over all intelligibility than the precise reproduction of speech sounds[5].

Recognition results

The percentage of words that were correctly recognized can be found in the following table.

	Learning and recog-nition in silence	Learning in silence and rec. in "traffic noise" 70 dB (lin)average	Learning in silence and rec. in "computer room noise" 90 dB(lin)	Learning and recognition in "computer room noise"
Senn. MD 421	82.4			
reference	93.2			
Senn. headset	83.9	≈25	<25	not possible
reference	89.3			
Hosiden KUC 7001	92.2	<25	<25	not possible
reference	87.7			
Cont. mic. on the forehead	80.5			
reference	93.7			
Cont. mic. on the neck	85.4	48.6	53.6	77.0
reference	82.2			

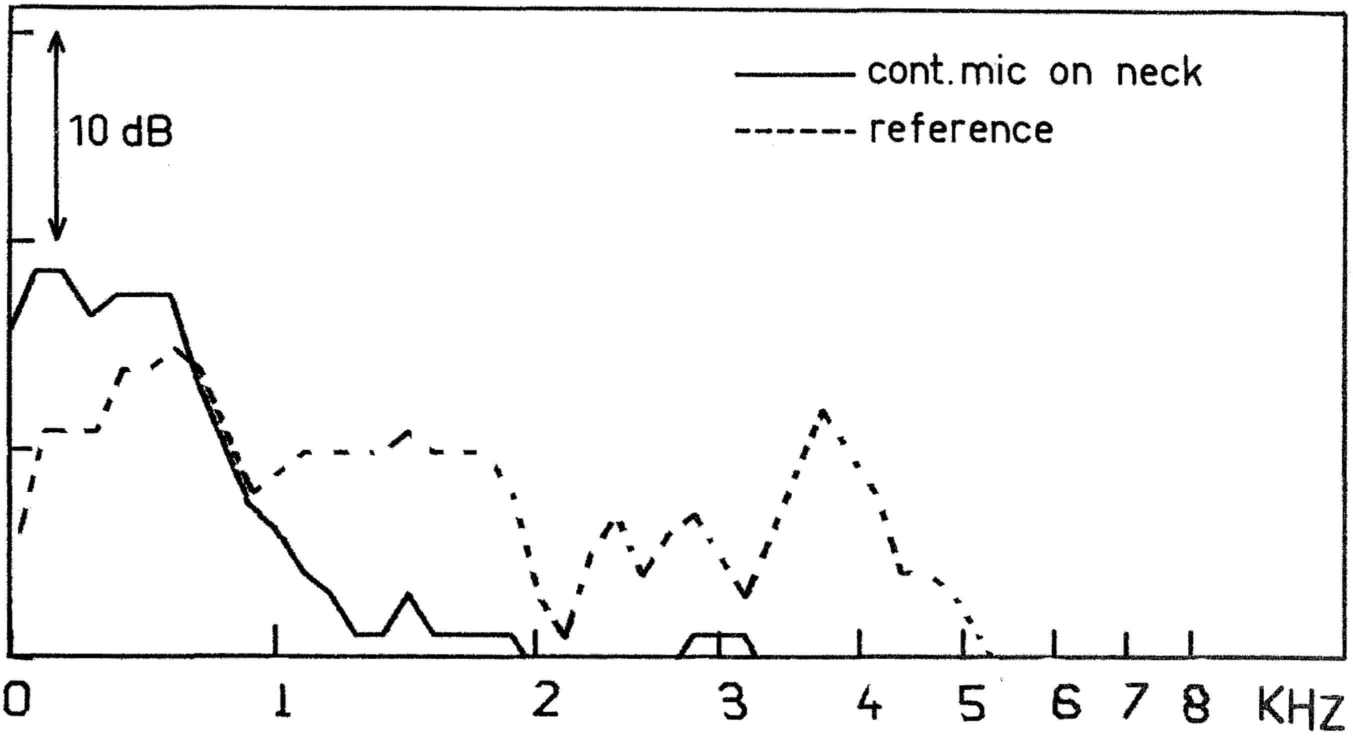

Fig. III-A-13. Average spectrum, contact microphone on neck.

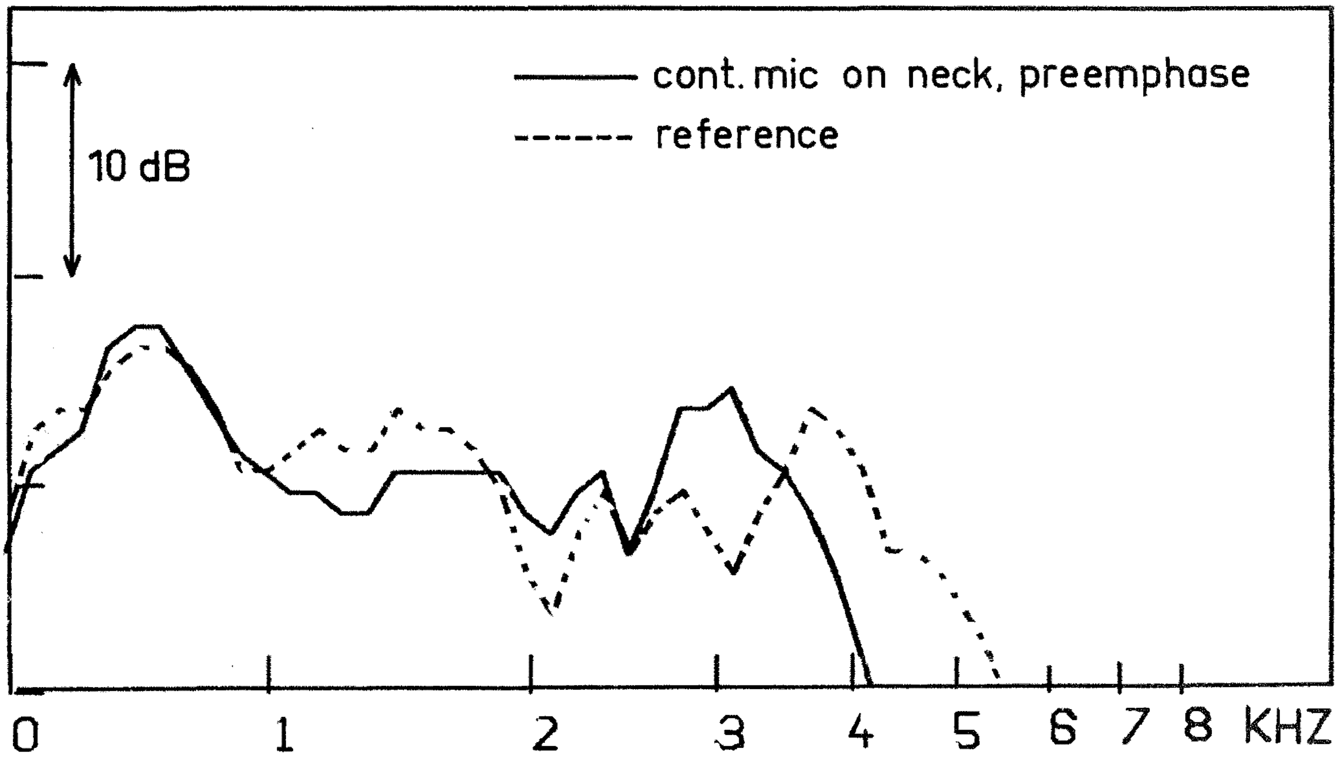

Fig. III-A-14. Average spectrum, contact microphone on neck, extra pre-emphasis.

The percentage correctly recognized words with the MD 421 seems a bit low, compared to the reference recording. A closer examination showed a weakness in the learning process that accounted for a specific confusion. There is little reason to blame the microphone for this. Apart from these errors, the results are comparable to that of the reference recording. This agrees well with the measurement result.

The headset microphone has also given a rather poor result. The errors are of a rather random character, i.e. they are evenly distributed in the confusion matrix. These errors are probably due to the poorer sound quality of this microphone. With noise up to 90 dB(lin) the recognition was not very successful. It should be mentioned, however, that the noise rejection of this and the following microphone is very good, when compared to the MD 421.

The other noise cancelling microphone, the Hosiden capsule, gave a better result, which is consistent with a good sound reproduction quality. With "traffic noise" the result with this microphone is not as good as with the Sennheiser microphone. The reason is the slightly better background damping of the Sennheiser microphone.

Contact microphone on the forehead did not work as well as contact microphone on the neck.

Considering the fact that the recognition system was not in any way optimized for the contact microphone, the results with a high quality contact microphone on the neck are promising.

Conclusions

The study did not reveal anything spectacular about the normal microphones. The contact microphone proved to give a very good noise rejection and an acceptable sound quality.

The speech recognition system worked well also with the contact microphones. This was not expected, considering the distortion that this type of microphone introduces. Apparently the contact microphone is a good alternative in situations with high noise levels. The sound transmission to the contact microphone under operation conditions is only very roughly described. A better understanding of this requires more research.

refs. on p. 38.

References

(1) BLOMBERG, M.E. & ELENIUS, K.O.E.: "A phonetically based isolated word recognition system", paper presented at the 96th meeting of the Acoustical Society of America, Hawaii.

(2) DENES, P.B.: "Automatic speech recognition: Old and new ideas", 1974 IEEE Symp. on Speech Recognition, Invited Papers, Academic Press 1975.

(3) FANT, G.: Acoustic Theory of Speech Production, Mouton 1960 (2nd edition 1970).

(4) FANT, G. & PAULI, S.: "Spatial characteristics of vocal tract resonance modes", paper presented at SCS74, Stockholm; in Speech Communication, Vol. 2, Almqvist & Wiksell 1975.

(5) FLETCHER, H.: Speech and Hearing in Communications, D. van Nostrand Co., Inc. 1953.

(6) MARTIN, T.B.: "Applications of limited vocabulary recognition systems", 1974 IEEE Symp. on Speech Recognition, Invited Papers, Academic Press 1975.

(7) MOORE, R.K.: "Evaluating speech recognizers", IEEE Trans. on Acoustics, Speech and Signal Processing, Vol. ASSP-25.