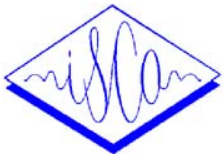

EUROSPEECH '97
5th European Conference on Speech Communication and Technology
Rhodes, Greece, September 22-25, 1997

SPEAKER RECOGNITION BY HUMANS AND MACHINES

Herman J.M. Steeneken and *David A. van Leeuwen*
TNO Human Factors Research Institute,
Soesterberg, The Netherlands
Tel. +31 346 356269, FAX +31 346 353977, E-mail: steeneken@tm.tno.nl

ABSTRACT

Speaker recognition with human listeners and with an automatic system were compared. Eight male and eight female speakers were involved. Also the effect of the speech quality was investigated: wide band, telephone band and two signal-to-noise conditions of 6dB and 0dB.

It was found that for both methods the male speakers are slightly better recognized. One to two words are sufficient, in the wide band condition, for correct subjective recognition. The automatic recognition requires a slightly longer utterance.

10.21437/Eurospeech.1997-610

1. INTRODUCTION

Recognition of a speaker becomes more and more important at surveillance of communication channels (forensic, military) and with command and control operations.

For our study three aspects were considered: (1) how do humans perform (human bench mark), (2) which parameters are defining the human recognition performance, (3) what is the relation with automatic speaker recognition systems.

In order to study these aspects a set of speech samples is required of individuals who are known in a certain community. Also the subjects who have to identify a certain speaker should be part of that community. We selected 16 well known individuals from our institute and 8 individuals from elsewhere as speaker. Selection of these individuals was balanced on gender and age.

Speaker recognition also depends on the quality of the speech signals, therefore various conditions were used including wide-band speech, telephone-band speech, and two conditions with noise (SNR +6 dB, 0 dB). For this purpose noise samples were used with a spectrum shaped according to the long-term speech spectrum.

The automatic speaker recognition was based on an algorithm which uses a description of the signal by the co-variance in the spectral domain.

For a representative evaluation on automatic speaker recognition the data base used for the experiments was disseminated in a wider community. Various laboratories are presently evaluating these data and the results will be presented at the conference.

2. EXPERIMENTAL SET-UP

2.1 Experimental design

The recognition experiments were performed both with humans and with an in house automatic speaker recognition system. A human recognition performance experiment was performed earlier at our Institute by Boxelaar and Pols (1986) for speech coder assessment.

For the present experiment we used 16 known speakers and 8 unknown speakers (not known at our Institute). The speech samples were all from the same sentence in order to exclude confounding with the text material.

One Dutch test sentence was used for the experiments, the total length was approximately 7s. We used various short samples of this sentence successively.
The utterances were respectively: (1) "man", (2) "afkomst", (3) "De zesenvijftig-jarige man", (4) "De zesenvijftig-jarige man is van Turkse afkomst", (5) "De zesenvijftig-jarige man is van Turkse afkomst en woont al tientallen jaren in de gemeente".

For the training of the automatic system a different sentence was used in order to get text independent recognition.
Four conditions with different speech qualities were used: wide-band, telephone band, and wide-band speech mixed with speech noise (spectrum equal to long-term speech spectrum) at two signal-to-noise ratios (SNR). Hence a fair equal masking of all frequency bands was obtained. The SNR's were + 6 dB and 0 dB respectively.

2.2 Subjective recognition
For the subjective experiments 16 subjects from our Institute were used. All subjects have worked there for many years so they all know the speakers used in the experiment quite well. The 16 speakers (8 male and 8 female) were selected from the Institute's population with the condition that all were well known individuals in the Institute community. Additional to the 16 "known" speakers, 8 "unknown" individuals were selected from outside the Institute. (4 males, 4 females).

During the experiments each listener was situated individually in a sound booth. The speech samples were presented binaurally by headphone. The sequence was a short item (see 2.1) of one word after which the subject was asked to reveal the identity of the speaker or ask for the next, slightly longer, utterance. Responses were made by using a note-book on which the photographs of all 16 possible speakers were displayed. Also a box "unknown male", "unknown female", and "unknown" was displayed. The subject could respond by using a mouse controlled pointer on the screen. If a space-bar response was given the next, longer, speech utterance was presented. If an identification response was given the next, randomly selected, speaker stimulus was presented. The 24 speakers at the four conditions were selected randomly. Hence, 96 trials were judged. As the presentation was ended by giving a response (which may point to the wrong speaker identity) a proportional scoring measure was not available.

The scores for a correct response ranged from "1" to "5", representing stimuli of increasing duration. The penalty for incorrect responses was a score higher than 5. We did an analysis for various values of this penalty (6 to 10).

2.3 Automatic recognition
A similar experiment was performed with an automatic speaker recognition system (based on the second order statistics of speech spectra, Bimbot et al. 1995). This system was also trained for the speech samples for the "unknown" speakers in order to allow for confusions which may be related to the listener responses.

Two experiments were performed: one with a training of the system based on wide-band speech and one with a training with speech signals corresponding to the four test conditions. The recognition system did not provide a robust measure of confidence for the match between the presented item and the best fitting training template. Scores tended to vary with condition, stimulus length, and other uncontrolled phenomena. We therefore also used a similar scoring method as used with the subjective tests, hence, as a function of the utterance duration a score was given similar to the utterance sequence number at which correct recognition was given (i.e., 1-5), false responses were labelled "6".

3. RESULTS

3.1 Subjective results
In Fig. 1 the mean recognition performance is given for the three groups of speakers (male, female, and unknown), and the four conditions. Incorrect responses were scored as "6". In all conditions the male speakers have the best recognition score (lowest value), followed by females. The "unknown" speakers give the highest value for the recognition score.

There is a difference in recognition performance between the conditions. For all speaker groups the wide-band condition leads to the highest recognition score, followed by SNR 6 dB, telephone band width, and SNR 0 dB. These results do not depend on the score chosen for a false recognition (6-10).

We performed a statistical analysis on the subjective data by using ANOVA. It was found that the differences between speaker group

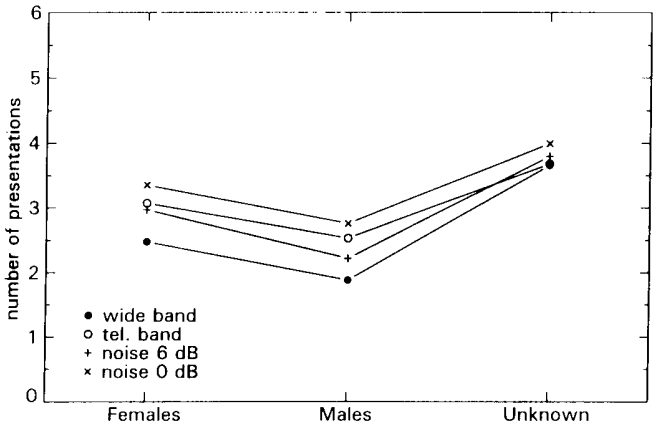

Fig. 1 Mean subjective recognition for the three groups of speakers and the four conditions (Score related to the number of stimuli required for correct response, with "6" as penalty for incorrect response).

(male, female, and unknown) as well as between conditions were significant ($p<0.01$). In Fig. 2 for each speaker and condition the mean number of stimuli is given for correct recognition. For the males M1 - M5 on average less than 2 stimuli were required for correct recognition (one or two short words) in the wide band condition. Even for the 0 dB noise condition less than three stimuli were required. This shows the perfect human speaker recognition performance for short samples even under adverse conditions.

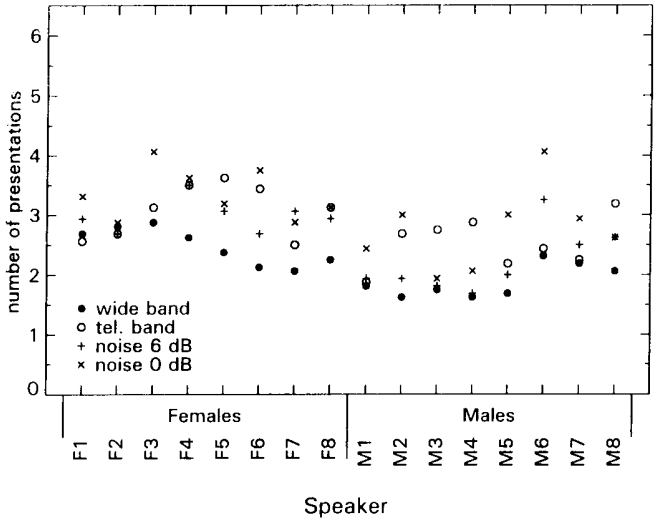

Fig. 2 Mean number of stimuli required for correct subjective response for each individual speaker and the four conditions (Score related to the number of stimuli required for correct response, with "6" as penalty for incorrect response).

3.2 Automatic recognition results

For the automatic recognition we had to train the system with speech utterances different from the test sentence. The sentence was the same for across speakers. We used short samples (approx. 5s) of wide-band speech and also in a separate trial degraded speech samples similar to the test conditions.

The same score as used for the subjective tests was used for evaluation of the results. In Fig. 3 these scores for the four conditions in the case of specific training are given.

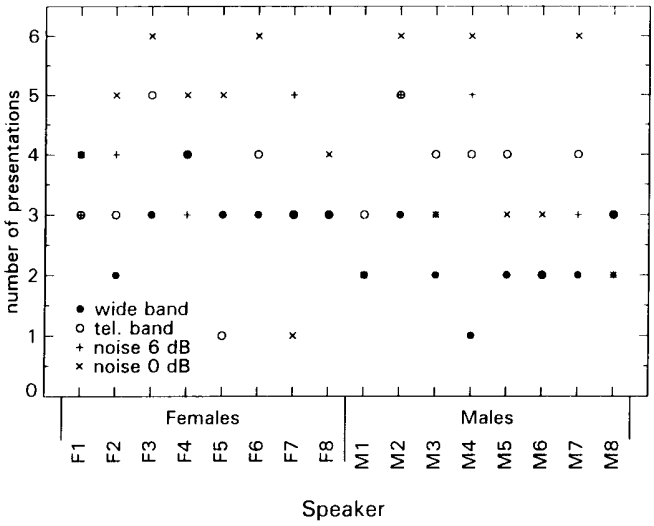

Fig. 3 Mean number of stimuli required for correct automatic recognition for each individual speaker and the four conditions (Score related to the number of stimuli required for correct response, with "6" as penalty for incorrect response).

The quantities shown in this figure are identical to Fig. 2 which was obtained for subjective recognition. The difference is however, that Fig. 2 is based on the mean score for 16 observers and Fig. 3 on one response of our automatic system. For the automatic recognition stimulus 2 was different from the stimulus used for the subjective tests. Here the stimulus consisted of the combined "man afkomst" because this is a more fair comparison with the subjective test for an automatic system which does not "remember" the first stimulus.

4. DISCUSSION AND CONCLUSION

4.1 Subjective recognition

Listeners need only a short utterance for a correct recognition of male speakers, only one word of moderate length was required. The recognition performance decreases for telephone-band speech and speech with noise. For our set of females a slightly lower recognition rate was observed. For the eight female speakers a same recognition performance for wide-band speech was obtained as for male speakers at a SNR of 6 dB. The difference between wide-band and telephone band is smaller for female speakers which may be related to the smaller frequency range of female speech.

4.2 Automatic recognition

Similar results with the automatic spectral analysis based approach were obtained. Also the ranking between male and female responses and for the four conditions were similar. We did not investigate specific analysis parameters of the system. The spectral analysis was based on MFCC (Mel Frequency Cepstral Coefficients). The algorithm for automatic recognition is presently under further development.

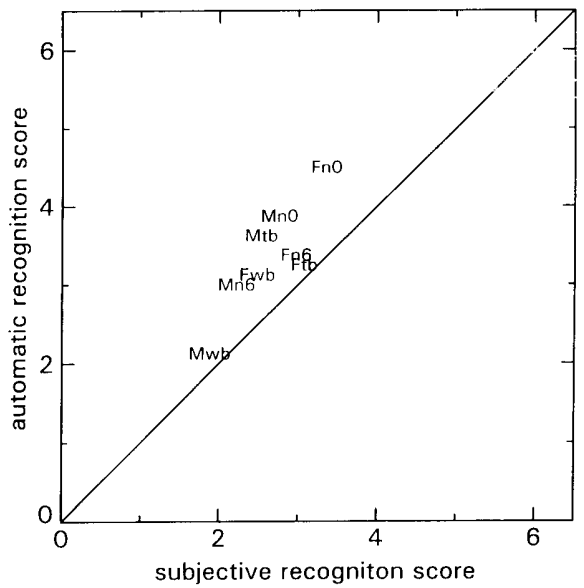

Fig. 4. Relation between subjective and objective speaker recognition, for female and male speakers and four speech quality conditions.

4.3 Comparison of the two methods

The relation between the two recognition methods is given in a scatter diagram (Fig. 4). In this diagram the mean recognition score for female and male speakers and for the four conditions is used.

The correlation coefficient obtained with a linear regression between the two recognition methods is r = 0.83 (based on only 8 data points), for the gender specific scores (4 data points) r = 0.78 for females, and r = 0.99 for males.

In Fig. 4 a different relation between subjective and machine recognition for gender is observed. As this is only an example for the specific automatic system that we used, a more general view will be obtained by additional experiments with other systems. For this purpose the data base was disseminated to another research group, results are expected before the Eurospeech 97 conference.

5. ACKNOWLEDGEMENT

The authors wish to thank Simon Alberts, who performed the subjective experiments during his practical study at our Institute.

6. REFERENCES

Boxelaar, G.W., and Pols, L.C.W. (1986) "PB-word intelligibility and speaker identification of 5 medium bandwidth coders: a pilot study.

Bimbot, F., Magrin-Chagnolleau, I., and Mathan, L. (1995) "Second-order statistical measures for text-independent speaker identification". Speech Comm. **17**, 177-192.