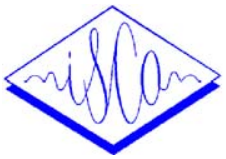


EUROSPEECH '97
5th European Conference on Speech Communication and Technology
Rhodes, Greece, September 22-25, 1997

A COMPARISON OF HUMAN AND MACHINE IN SPEAKER RECOGNITION

Li Liu, Jialong He, and Günther Palm
Abteilung Neuroinformatik
University of ULM, GERMANY
li@neuro.informatik.uni-ulm.de

ABSTRACT

Speaker recognition experiments have been conducted with the publicly available YOHO database to compare the performance of human listeners and computers. Two types of listening experiments have been performed, one is the forced-choice speaker discrimination test which is corresponding to the task of speaker identification. The second experiment of speaker recognition by human listeners is the same-different judgment which is similar to the task of speaker verification. It is shown that the human listeners perform well for the same-different judgment task, but the error rate of speaker discrimination is relatively large. Besides, human listeners are more robust to session variability, while the machine's performance falls off largely when the reference and test utterances are from different recording sessions.

1 INTRODUCTION

Recent years automatic speaker recognition becomes an active research area. The driving force of this increasing interest is due to that there are many promising applications. Rosenberg demonstrated that in some instances speaker recognition by machines can outperform human listeners. This is especially true for short test sentences and unfamiliar sounds [1]. Until now we still have very little knowledge about how a human listener distinguishes speakers and what cues he use. Such kind of knowledge would be surely helpful for designing a more robust and powerful automatic speaker recognition system.

The present study reports several comparison experiments with the publicly available YOHO database. Two human listening experiments have been conduced, one is a forced-choice speaker discrimination test and the other is a same-different judgment test. Corresponding to the two listening tests, two types of speaker recognition experiments: speaker identification and verification, have been performed using the same speech data. Two probabilistic speaker models have been evaluated. In the first model, the distribution density of feature vectors is modeled with a standard multivariate Gaussian function with full covariance matrix, while the second speaker model is the Gaussian mixture model (GMM).

2 SPEECH DATABASE

The speech material were selected from the YOHO database which was collected at ITT to support text-dependent speaker authentication research [2]. There are 106 male and 32 female speakers. Each speaker has four enrollment sessions where in each sessions he/she is prompted to read a series of 24 combination-lock phrases. Each phrase is a sequence of three two-digit numbers (e.g., 35-72-41, pronounced as thirty-five seventy-two forty-one). There are 10 verification trials per speaker consisting of four phrases per trial. The vocabulary consists of 56 two-digit numbers ranging from 21 to 97. The speech was collected in an office environment using a telephone handset connected to a workstation. Thus, the data has slightly wider bandwidth (3.8 kHz) than the telephone line bandwidth, but no telephone transmission degradation. Besides, all sessions took place over the same handset. To save experiment time, only a subset of the database consisting of 20 male and 20 female speakers was used. For the human listening experiments, the signals were played back directly without any further processing. In the automatic recognition, 16 MFCC coefficients were calculated from each voiced segment of speech signals to compose a feature vector. The analysis window size was 32 ms with 16 ms overlapping.

3 SPEAKER RECOGNITION BY HUMAN

3.1 Procedure

Two experiments of speaker recognition by human listeners were performed. The first experiment is a forced-choice speaker discrimination test. This experiment corresponds to the close-set speaker identification by computers. During the test, a sequence of test blocks was played back through earphones. Each test block consists of 5 tokens (a token refers to one digit string). The first token is the reference token, followed by four test tokens corresponding to the four alternative choices labeled as A, B, C and D. The four alternatives were spoken by four different talkers, one of them being spoken by the speaker of the reference token. The subjects were asked to identify which of the four alternative utterances was spoken by the same speaker of the reference token. The

experiment was performed under two different conditions. In the first case, the reference token and the corresponding test token came from the same recording session, while in the second case they were selected from different recording sessions.

The second listening experiment is the same-different judgment test. A pair of tokens was played back each time, the listeners had to judge whether these two tokens were spoken by the same speaker or not. The listeners were also requested to indicate on a three-point scale his confidence in the correctness of his response. Like the first experiment, the two test conditions: the same recording session and different recording sessions, were studied.

3.2 Results

In general, the performance of speaker recognition by human improves with the test going on. In the speaker discrimination experiment, the correct response rate was 62% for the first test session and then gradually increased to 85% after the completion of the 4th session. Even thought the performance is relatively lower when the reference and test utterances are from different recording sessions, the degradation is not so dramatic as in many automatic speaker recognition applications. In other words, human listeners are relatively robust to session variability. No significant difference was found between male and female voices. It is noticed that the human's temporary memory plays an important role in this type of experiments. The subjects must memorize the properties of the reference token in order to compare it to each of the alternatives. The percentage of correct responses by the listener declines as the reference and the correct alternative are separated by an increasing number of incorrect alternatives. The impression of the reference sound becomes more blurred as the time span between the reference and test token increases. The recognition rate is at highest when the token next to the reference is happen to be the correct one.

Recording session	same	different
discrimination rate	77 %	71%
judgment error rate	8%	12%

Table 1 Speaker recognition performance by human listeners.

In the same-different judgment test, the initial error rate was about 20%, but with the test going on, the error rate decreases to less than 8%. Similar to the discrimination experiment, the error rate is higher if the reference and test utterances are from different recording sessions. Table 1 gives the average performances of the listening experiments. The first row gives the correct identification rate obtained from the speaker discrimination experiment, while the second row is the error rate of the same-different test. We summarize the results shown in Table 1 as follows. First, the discrimination rate declines (or the judgment error rate increases) when the reference and test utterances are from different recording sessions, but the performance degradation is not so dramatic as will be shown in the speaker recognition by machine, suggesting that the human listeners are more robust to session variability. Second, the human listeners can distinguish two speakers very well (the same-different test), but it is clearly difficult for human to hold several different voices in head in order to compare them, therefore, the error rate of speaker discrimination is higher than that of the same-different judgment. On the other hand, the automatic system has no such problem, it can store a lot more data and compare them with each other.

4 SPEAKER RECOGNITON BY MACHINE

4.1 Speaker Models

We evaluated two parametric probabilistic speaker models. The first model is the multivariate Gaussian function given by

$$p(\vec{x}) = \frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}} exp\left\{-(\vec{x}-\vec{u})^T \Sigma^{-1}(\vec{x}-\vec{u})/2\right\}$$

The Gaussian model has been studied extensively in the statistical literature and is widely used for speaker recognition [3]. The advantage of using the Gaussian model is that the parameters ($\vec{u}$ and Σ) can be calculated directly from the training data.

For a given test vector sequence, $X = \{\vec{x}_t\}_{t=1}^{N}$, usually the average log-likelihood, $\ell(X) = \frac{1}{N}\sum_{t=1}^{N} log\left(p(\vec{x}_t)\right)$, is used as the measurement of similarity. However, from a series of preliminary experiments, we found that using the Bhattacharyya distance [4] always leads to a better performance. This measurement is defined as

$$B(X) = \frac{1}{8}(\vec{u}_X - \vec{u})^T\left[(\Sigma+\Sigma_X)/2\right]^{-1}(\vec{u}_X - \vec{u}) + \frac{1}{2} ln\left(\frac{\left|(\Sigma+\Sigma_X)/2\right|}{\sqrt{|\Sigma|}\sqrt{|\Sigma_X|}}\right)$$

where $\vec{u}$ and Σ are the mean vector and covariance matrix of the Gaussian model, while $\vec{u}_X$ and Σ_X are the mean vector and covariance matrix of the test vector

sequence. It is easy to show that if the test vector sequence match the model perfectly, then $B(X)=0$. A larger value of $B(X)$ means that the vector sequence is less possible from this model. $B(X)$ was used as the measurement of similarity in our experiments with the Gaussian model.

Recently, the Gaussian mixture model (GMM) becomes very popular and is shown to be able to give a very high speaker recognition performance [5]. The GMM is a weighted sum of several Gaussian functions

$$p(\vec{x}|\lambda)=\sum_{i=1}^{Q} c_i p_i(\vec{x})$$

The parameters of the GMM can not be calculated directly from the training data but can be estimated with the EM algorithm. In practice, due to high computational load and limited available training data, the diagonal covariance matrices are exclusively used in all Gaussian functions. The GMM has stronger modeling capability than the unimodal Gaussian function, but usually needs more training data. Obviously, the feature vectors from one digit string is not enough to reliably estimate the parameters of the GMM, therefore, we conducted normal speaker identification and verification experiments with the GMM.

In the identification experiment, the population size is the same 40 speakers used in the listening experiments, but each model was trained with all data of four enrollment sessions. During the test, an individual string was used as a test utterance, that is, there were 40 test utterances from each speaker and 1600 test utterances in total. In the verification experiment, some additional 20 speakers (10 male and 10 female) were selected as the impostors. The likelihood ratio was used as the score, in each evaluation, there were 1600 scores from the customers and 32000 scores from the impostors. The background speakers were those not claimed but having the same gender as the claimed speaker. In other words, there were 19 background speakers for each model. The background score is the joint probability density of the utterances as described in [5].

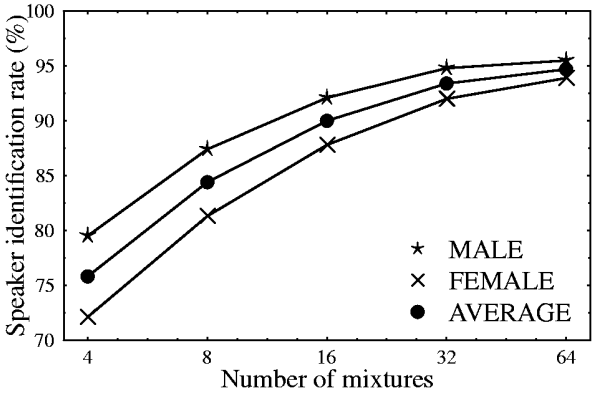
4.2 Identification Experiment

With the Gaussian model, the automatic speaker identification experiment was performed under the same paradigms as the speaker discrimination test. From each test block (5 tokens), five Gaussian models were created, that is, the mean vector and covariance matrix of each Gaussian model were estimated from only one token. Each of the last four models was compared with the first one, the test model that had the smallest Bhattacharyya distance to the reference (the first model) was identified as spoken by the same speaker of the reference. The speaker identification rate with the Gaussian model is shown in Table 2. Comparing with the speaker discrimination test, we see that the machine outperforms the human listeners. This is especially true in the case of the same recording session, but the machine's performance falls off largely when the reference and test utterances are from the different recording sessions. Besides, it is seen that the automatic system has more difficulty to recognize female voices. This conclusion is consistent with that obtained by other researchers [5].

	same recording session	different recording sessions
male	95.6 %	85.6 %
female	94.0 %	78.8 %

Table 2 Speaker identification performance using the Gaussian model under the same evaluation conditions as the forced-choice listening test.

The speaker identification rate vs. the number of mixtures by the GMM is plotted in Figure 1. As expected, the performance improves with the number of mixtures. Again, the identification rate for the male speakers is higher than that for the female speakers. Due to different evaluation conditions, it is not easy to compare these results with the performance of human listeners, but it is quite possible that for the task of identifying a speaker from a large population, the automatic system can do better than humans.



HEI

Figure 1 Speaker identification performance using the GMM.

4.3 Verification Experiment

To simulate the scheme of the same-different judgment test, two Gaussian models were generated from each test pair. The Bhattacharyya distance between these two

models was calculated as the matching score. Figure 2 shows the false rejection rate and false acceptance rate as a function of the decision threshold θ. In this case, the equal error rate (EER) is about 21% at θ=0.54.

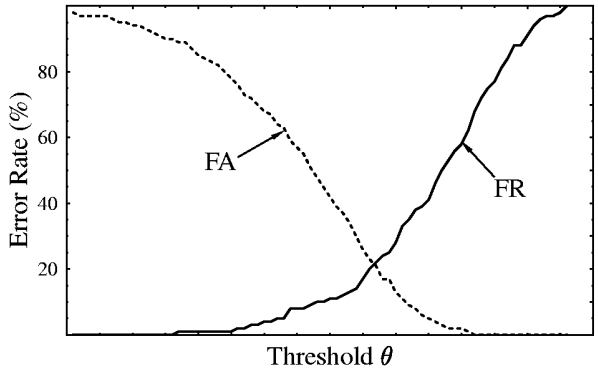


Figure 2 The False rejection rate and the false acceptance rate as a function of the decision threshold θ.

To gain insight into the problem, we reorganized the scores according to the recording sessions and the gender of speakers. Table 3 gives the EER in different situations. From this table, we see that, like speaker identification, the verification accuracy is lower when the reference and test utterances are from different recording sessions. There is still difference between male and female voices, but it is not consistent. Comparing with the same-different test, it is seen that, in general, human listeners perform better than machines for this type of tasks. This is especially true for the different recording sessions.

	same recording session	different recording sessions
male	9.5 %	25.0%
female	12.0%	20.0%

Table 3 The equal error rate (ERR) with the Gaussian model under the same evaluation conditions as the same-different listening test.

The verification performances using the GMM are shown in Table 4. In addition to the EER, this table also gives the false rejection rates at a false acceptance rate of 0.1%. Comparing with the results reported in [5], the error rates shown in Table 4 are much higher. The main reason is that in this experiment an individual string was used as a test utterance, while in Reynolds' paper, he concatenated four digit strings into a long one and used it as a single test utterance. It is known that the error rate decreases with the length of test utterances. Besides, the dimension of feature vectors in this experiment was 16 which is smaller than that used in Reynolds' experiments.

Mixtures	4	8	16	32	64
EER (%)	10.2	7.8	6.4	5.2	4.7
FR (%) @FA=0.1%	62.8	50.3	42.0	33.6	33.9

Table 4 Error rate of speaker verification using the GMM. The equal error rate (EER) and the false rejection (FR) rate at the false acceptance (FA) rate of 0.1%.

5 CONCLUSION

Several experiments have been conducted to compare the performance of speaker recognition by human listeners and by computers. It has been shown that for the task of same-different judgment, which is similar to speaker verification, human listeners can do better than computers. Besides, human listeners are more robust than machines to session variability. For the task of speaker discrimination, which corresponds to speaker identification, due to memory limitation of humans and the interference between different test utterances, computers can outperform the human listeners, especially when the reference is followed by too many alternative choices.

6 REFERENCE

[1] Rosenberg A. E. (1973) "Listener performance in speaker verification tasks," IEEE Trans. on Audio and Electroacoustics, Vol. 21, pp. 221-225.

[2] J. Godfrey, D. Graff and A. Martin (1994), "Public databases for speaker recognition and verification," *ESCA Workshop on Automatic Speaker Recognition, Identification and Verification*, Martigny, Switzerland, pp. 39-42.

[3] Bimbot F., Magrin-Chagnolleau I., and Mathan Luc (1995) "Second-order statistical measures for text-independent speaker identification," Speech Communication, Vol. 17, pp. 177-192.

[4] Fukunaga, K. (1990) *Introduction to statistical pattern recognition*, pp. 97-109, Academic press Inc. San Diego.

[5] Reynolds D. (1995) "Speaker identification and verification using Gaussian mixture speaker models," Speech Communication, Vol. 17, pp. 91-108.