HUMAN *VS.* MACHINE SPEAKER IDENTIFICATION WITH TELEPHONE SPEECH

Astrid Schmidt-Nielsen† and Thomas H. Crystal‡

†U.S. Naval Research Laboratory; ‡IDA Center for Communications Research

ABSTRACT

An experiment compared the speaker recognition performance of human listeners to that of computer algorithms/systems. Listening protocols were developed analogous to procedures used in the algorithm evaluation run by the U.S. National Institute of Standards and Technology (NIST), and the same telephone conversation data were used. For "same number" testing, with three-second samples, listener panels and the best algorithm had the same equal-error rate (EER) of 8%. Listeners were better than typical algorithms. For "different number" testing, EER's increased but humans had a 40% lower equal-error rate. Other observations on human listening performance and robustness to "degradations" were made.

1. INTRODUCTION

The development of automatic speaker verification technology has been greatly facilitated by the recent development of uniform training and test materials and of procedures and metrics for evaluating verification performance [1,2]. For many, the performance goal for automatic verification is to obtain the verification accuracy of human listeners. The 1998 NIST Speaker Recognition Evaluation [3] presented an opportunity to examine where the automatic algorithms stand *viz. a viz.* this goal and to examine how human listeners perform given the data and the guidelines for the evaluation. We could also compare human performance on this collection of spontaneous, conversational telephone speech with human performance for other data sets and testing paradigms [5,6].

The challenge of designing a paradigm analogous to the one for evaluating computer algorithms that can be implemented within time and cost constraints makes very evident the differences between human and machine. The differences and the budget influenced many of the design choices and reflect that individuals have limited memory, easily become fatigued and won't work efficiently 24 hours a day, even with additional compensation.

What we have tried to uncover and report are the similarities and differences in recognition performance by humans and machines and the conditions under which human recognition is more robust. And, we can report that, given three-second test samples, listeners -- acting as a panel – do outperform automatic algorithms.

2. TEST METHOD

2.1 The Basic Task and The Test Data for Speaker Identification

The basic speaker recognition task is the same for both machine and human. Performance is averaged over a set of target speakers. For each target speaker there are training examples, for modeling the target, and test samples. The test sample set contains speech from the target, for measuring detection, and speech of non-targets or foils, for measuring false alarms

The speech data for the speaker recognition evaluation was prepared by NIST and drawn from Switchboard II [1]. It consists of five-minute recordings of separate sides of telephone conversations between pairs of strangers, mediated by a robot operator and recording system. Participants both initiate and receive calls. Participants initiate calls from many different phone numbers, so their voices are recorded through a variety of microphones and channels. All the calls received by a given participant come to the same number. Each side of each conversation is evaluated as to its distortion, noisiness and microphone type, e.g., electret or carbon button. One-minute training segments and 3-second, 10-second and 30-second test segments are extracted from the conversation sides. The test segments were selected automatically so some, especially the shorter ones, may contain mostly non-speech such as laughs, snorts, loud noise or even silence.

In the evaluation of algorithms, a number of rules and restrictions were imposed, many of which cannot be applied equally to people and machines. Some examples: The use of information about other test segments or about other talkers in the test set was not allowed in evaluating a given test segment. Knowledge of whether training and test were from the same or different phone numbers and handset type was allowed. Knowledge of talker sex was given; test samples were always from a talker of the same sex as the target.

2.2 Listening Task and Test Design

Within the constraints of time and cost, the human listener tests paralleled the algorithm evaluation as closely as possible. The human tests used 3-second test samples and a smaller number of talkers and test samples than for evaluating algorithms. "Two-session" training was used: each of the two one-minute samples was from a different call from the same phone number. The listening task was designed to compensate for the fact that humans do not have perfect memory of the training materials and that it was not possible to erase their memory of previous tests without causing undesirable damage to the system. Listeners were trained and tested on one target talker at a time. After 2 minutes of training on a given target, listeners were tested on 21 3-second test samples taken from different conversations, approximately half being from the target and half from foils. On each trial, listeners heard the test sample, a 3-second reminder taken from the training speech and a repeat of the test sample. Each listener then used a 10-button response box to record a same/different judgement with five levels of confidence. Short tones were used to alert the listeners to the start of each trial and to help them keep the samples and reminders separate. After the test samples for one target were completed, the training and tests for the next target were presented. Every effort was made to prevent memory for previously presented targets from influencing later performance. Once a given target talker had been evaluated by a listener panel, it was assumed that the particular talker was known to the listeners, and that talker was never presented to the same listeners as a foil in later test.

Up to eight listeners at a time were tested in a sound attenuated room. The stimuli were presented simultaneously to all the listeners through headphones, and the next sample was not presented until every listener had responded. In all, there were 65 listeners (34 females and 31 males), with each panel hearing a total of 36 talkers (18 males and 18 females) over a day of testing with interspersed rest periods. Male and female talkers were tested in alternate sessions, separated by rest periods. Each target talker was heard by approximately 16 listeners. In all, 144 target talkers, 72 males and 72 females, were evaluated, for a total of 3024 test samples.

3. RESULTS

3.1 Individual and Combined Responses

The machine recognition tests required both a "hard" decision as to whether a given sample was or was not from the target talker and a similarity rating, with higher values indicating greater similarity. The 10-button response box likewise allowed for a hard decision (whether the selection was SAME or DIFFERENT), and numerical values for similarity were assigned to each of the 10 confidence levels, from 0 for most confident DIFFERENT to 9 for most confident SAME. As is to be expected with biological detection systems, there were large individual differences in listener performance, both in the ability to discriminate talkers and in bias (how strict or lenient a listener was in rejecting or accepting foils). The overall error rate for individual listeners ranged from 15% to 48%. This variability in performance was the reason for using multiple listeners for each test sample, and our primary interest was in determining how well listeners as a group performed, rather than in individual decisions.

Both the need for combining individual responses and the results of doing it are shown in the Detection Error Tradeoff (DET) curves [4] of Fig. 1. Miss rate *vs.* false-alarm rate is plotted for individual scores and for combining methods. The DET curves were generated by NIST with the same routines used in evaluating algorithms. Equal error rates, (EER's) the intersection of the DET curves with the diagonal, are given in Table 1. The data in Fig. 1 and Table 1 are for the "All-number" condition, i.e., not restricted to the "same-number" or the "different-number" conditions described below.

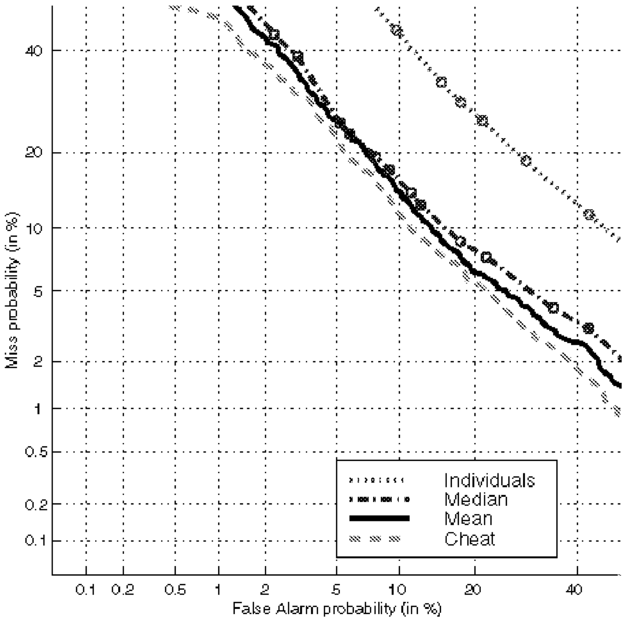

Figure 1. Detection Error Tradeoff (DET) plots for different methods of combining listener responses, for "all-numbers."

Combining Method	Error	Z-score
Individuals (none)	0.23	0.739
Median	0.125	1.15
Mean	0.120	1.175
Minimum Variance	0.1084	1.235
Log-Likelihood ("Cheat")	0.1058	1.249

Table 1. Equal error rates for different combining methods, for the "all-number" condition.

The curve in the upper right shows the uncombined, individual responses. The circles represent the cumulative errors for each of the 10 possible responses, in order of similarity and cumulated over listeners and test samples. Only six circles (connected by straight lines to improve readability) show up because the other four are outside the range (50%) of the axes. The intersection of this curve with the diagonal shows that the listeners, treated individually, have a collective equal error rate of about 23%. The other curves show the results of different methods of combining all the responses to a given test to get a single score or category judgement for that test. This usually meant combining 16 responses (the experimental design called for two panels of eight to make each judgement). The next curve, labeled Median, was obtained by finding the median response value. Twelve of 19 possible points (10 response values and 9 in between) are within range. The resulting equal error rate of approximately 12.5% indicates that combining divides the error rate in half. The curve labeled Mean was obtained by averaging the response values. This gave a slight improvement over computing the median. Averaging 16 responses gives many numerical values, and the individual points are not circled. Means will be used for the comparisons with computer algorithms and the analysis of robustness. The curve for "cheating" and the remaining table entries, which show that crime doesn't pay very well, were obtained by "optimally" calibrating each of the 10 responses for each individual listener, as explained next.

Different people use the response scales differently, and it makes sense to calibrate each listener before combining the scores. However, as the listeners' responses are also influenced by what is in the stimulus set, this calibration should be done using an independent stimulus set. Not having enough data to do this, we resorted to a cheating experiment in which the same data set was used for both calibration and evaluation. Two "optimal" calibrations were attempted and the results are shown in Table 1. In one experiment, each response value was a listener-specific log-likelihood. This was computed as the logarithm of the quotient of the frequency of a given response for target samples divided by the frequency for foil samples. This technique gave the "Cheat" DET curve in Fig. 1. In the other experiment, each response value was the one that would minimize the mean square error given that a target should receive a score of 1; a foil, -1. Table 1 shows that a cheating optimization of response values gives only 1.6% absolute (13% relative) lowering of the equal error rates. This is an upper bound of what might be achieved by an honest calibration. It vindicates the use of the integers 0 to 9 as an accurate characterization of listeners' perceptual space.

3.2. Human *vs.* Machine

The first comparison of listeners to algorithms is made for the "same number" condition, which was the principal condition used in the 1998 evaluation of algorithms. In this condition, the error rates are computed from a subset of the test samples: the only target test samples used are from conversations using the same number as the training samples and the foil test samples used are those from the same type of handset. A comparison of listeners with the best performing computer algorithm and two other "typical" algorithms is shown in Fig. 2. It can be seen that human recognition was at least as good as the best system and noticeably better than the typical systems that participated in the evaluation.

The relative robustness of listener responses under changes in signal characteristics is illustrated by the comparison of "same-number" to "different-number" performance in Fig. 3. In the "different-number" analysis condition," the target test samples are from different phone numbers than the training and there was no restriction on the handset type for the foil samples. Both humans and algorithms did more poorly when the test and training data were from different numbers than from the same number, but the loss was greater for algorithms than for humans.

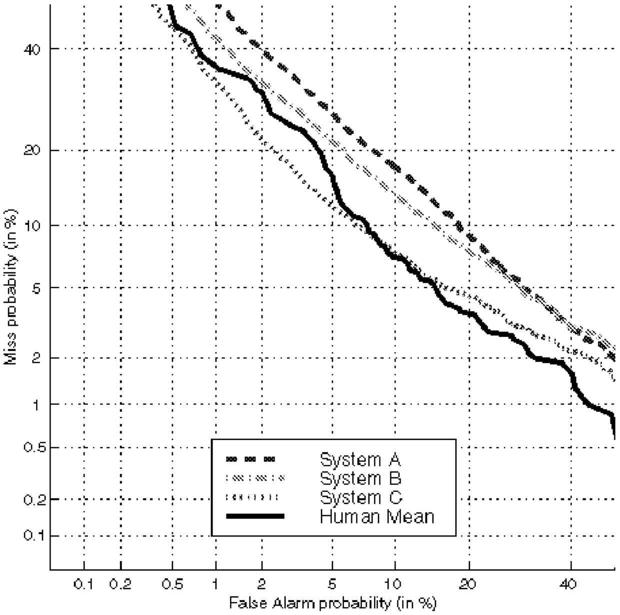

Figure 2. DET plots comparing human performance to the best and two typical algorithms (same-numbers data).

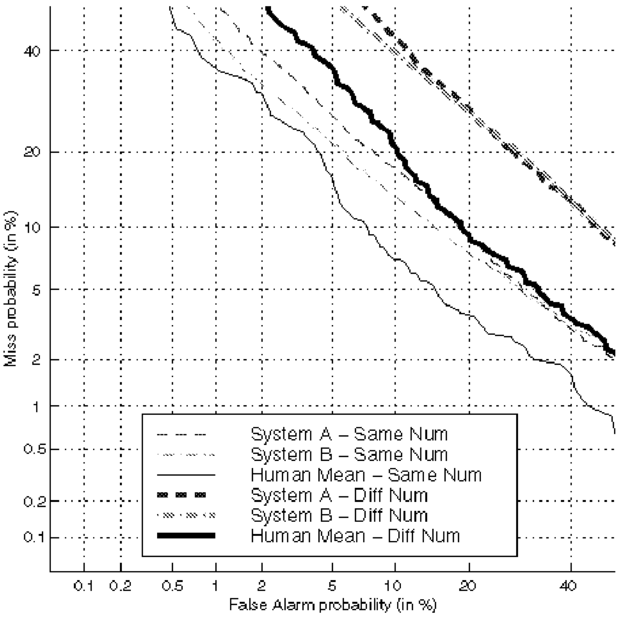

Figure 3. DET plots comparing human listeners to two typical algorithms, for "same-number" *vs.* "different number".

For this data set, unlike most algorithm results and several previous human recognition studies [5,6], humans recognized the female talkers better than they did the male talkers. An analysis by listener sex showed essentially the same pattern, in that both male and female listeners recognized the female talkers better. Interesting as it would be to find same-sex/opposite sex effects, this suggests that the 72 female talkers were in fact more easily recognized than the 72 male talkers were. In many earlier studies, the talker sets were so small (5-10 male and female talkers) that differences in recognition could as easily be due to talker selection as to talker sex. Most studies of human speaker recognition have been conducted with read speech (sentences, words, single vowels) and not with spontaneous speech. Table 2 shows the results of one study [5] of 10 male and 10 female talkers reading sentences compared with individual performance in the present study. Recognition with controlled text is considerably higher than with spontaneous speech.

	Read Sentences		**Telephone Speech**	
	Ave	**S.E.**	**Ave.**	**S.E.**
Male talkers	94.7%	0.63%	76.2%	0.70%
Female talkers	85.0%	0.80%	77.5%	0.80%

Table 2. Human speaker-recognition average percentage correct and standard error (S.E.) as a function of the type of material and the sex of the talker.

Previous work [7,8] using multi-dimensional scaling (MDS) to relate perceived talker distances to objective measures of the speech signal suggests that, with a three dimensional scaling solution, two of the dimensions seem to be related to measurable acoustic parameters of the speech signal, but the third dimension is not. The combination of target talkers and test samples in the experimental design was selected in a way that allowed us to collect pairwise similarity data for a subset of 24 male and 24 female talkers. An exploratory MDS analysis (3 dimensions) for these talkers, comparing human distance data with one of the machine algorithms, found significant correlation between human and machine for two of the dimensions, and not for the for the third. People often recognize talkers by their speech habits, accent, choice of vocabulary, etc. Although modern speech recognition algorithms are getting better at detecting variations in pronunciation, these cues are still harder for machines to detect than for people.

4. CONCLUSIONS

Human listeners show tremendous individual variability, and we explored ways of combining listener data to arrive at a group decision. We found that the group mean worked well, and the human results were very competitive with the best computer algorithms in the same handset condition. When different handsets/phone numbers were used, human performance degraded somewhat, but not as badly as algorithm performance. Both human and algorithm performance also went down when the signal was degraded by background noise, crosstalk, poor channel conditions, etc., but again the humans were more robust for the worst conditions. Unlike the computer algorithms, humans performed better with female than with male voices.

The greater human robustness to certain degradations can be explained in part by the fact that humans depend heavily on speech habits (pronunciation, word choice, characteristic laughs, etc.) when recognizing talkers. Some machine recognition systems use word recognition to detect variations in pronunciation. More exploitation of the cues used by humans may be the next step to additional progress.

5. ACKNOWLEDGEMENTS

Crucial financial support of NRL's participation was provided by NAVSPAWARSYSCOM. This research benefited greatly from the support and contributions of George Doddington, Doug Reynolds and Dave Pallett and his Spoken Language Technology Group at NIST. Among them, Alvin Martin provided invaluable technical advice and Mark Przybocki not only provided excellent advice but also devoted considerable time and talent to producing our figures. John Tardelli, Beth Kreamer, and Paul Gatewood of Arcon, Inc. made invaluable contributions in developing the programs for data presentation and collection and conducted the listening tests. Thanks also to Janet Stroup for her help in accurately assembling the properly randomized test lists.

6. REFERENCES

1. Przybocki, and Martin, A. “NIST Speaker Recognition Evaluations”, *Proc. First International Conference on Language Resources and Evaluation*, Granada, 28-30 May 1998, 331-335.

2. Przybocki, M. and Martin, A. “NIST Speaker Recognition Evaluation 1997”, *Proc. RLA2C*. Avignon, 20-23 April 1998, pp. 120-123.

3. Reynolds, D. Doddington, G., Przybocki, M. and Martin, A. “Statistical Analysis of Speaker Verification Results from the NIST 1998 Evaluation,” *Proc. ICSLP98*, Paper Number: 0608

4. Martin, A., Doddington, G., Kamm, T., Ordowski, M. and Przybocki, M. “The DET Curve in Assessment of Detection Task Performance,” *Proc. EuroSpeech 1997*, 4, 1895-1898.

5. Schmidt-Nielsen, A. and Brock, D. “Speaker Recognizability Testing,” *Proc. ICASSP-96*, vol. II 1149-1152, Atlanta, GA, May 1996.

6. Schmidt-Nielsen, A. and Stern K. R. “ Recognition of Previously Unfamiliar Speakers as a Function of Narrowband Processing and Speaker Selection,” J. Acous. Soc. Amer., 79, 1174-1177.

7. Schmidt-Nielsen, A. “Perceiving Talker Differences,” Abs. of the Psychonomic Soc., 1, 55, 1996 (Abstract).

8. Necioglu, B., Clements, M. A., Barnwell, T. P., and Schmidt-Nielsen, A. “Perceptual Relevance of Objectively Measured Descriptors for Speaker Characterization,” Proc. ICASSP-98, 1998.