Approaching Human Listener Accuracy with Modern Speaker Verification

Ville Hautamäki, Tomi Kinnunen, Mohaddeseh Nosratighods, Kong-Aik Lee, Bin Ma, and Haizhou Li

Institute for Infocomm Research (I^2R), A*STAR, Singapore
School of Computing, University of Eastern Finland, Joensuu, Finland
School of Electrical Engineering and Telecommunication, University of New South Wales, Australia
{*vishv,kalee,mabin,hli*}*@i2r.a-star.edu.sg, tkinnu@cs.joensuu.fi, hadis@unsw.edu.au*

Abstract

Being able to recognize people from their voice is a natural ability that we take for granted. Recent advances have shown significant improvement in automatic speaker recognition performance. Besides being able to process large amount of data in a fraction of time required by human, automatic systems are now able to deal with diverse channel effects. The goal of this paper is to examine how state-of-the-art automatic system performs in comparison with human listeners, and to investigate the strategy for human-assisted form of automatic speaker recognition, which is useful in forensic investigation. We set up an experimental protocol using data from the NIST SRE 2008 core set. A total of 36 listeners have participated in the listening experiments from three sites, namely Australia, Finland and Singapore. State-of-the-art automatic system achieved 20% error rate, whereas fusion of human listeners achieved 22%.

1. Introduction

It is a long-believed fact that while computers are faster in processing large amounts of data, they cannot outperform human accuracy in real-world pattern recognition tasks. Human beings are outstanding in recognizing spoken words (speech content) under varying conditions including background noises, transmission channels, reverberation and presence of other interfering speakers. The reason is that humans rely on several different levels of information in the speech signal to recognize others from voice alone. These cues might be a certain usage of words, speaking habits or a unique style in a person's laughter. It is complicated to extract the speaking habit or style of a person automatically. Therefore automatic systems mostly rely on low-level spectral features to discriminate speakers. However, spectral features are susceptible to any environmental and intra-speaker variation and, compared to human-based detection systems, they are usually less robust under severe mismatched conditions.While the best-performing automatic speech recognition (ASR) systems can already handle some of these conditions quite well, it remains a great engineering challenge to make the systems robust under all those conditions [1].

What about the speaker and language recognition accuracy of human beings? It can be argued that the speech content (words) and the affective cues (emotions and attitudes) are the most important information for social communication between human beings. But what would be the advantage of being able to recognize different speakers and languages? It can be hypothesized that, at the best, the speaker and language cues are of secondary importance. It is of great scientific interest, then, to know whether the automatic methods could outperform human being(s) in speaker and language recognition tasks. When developing new speaker and language recognition methods, should we take the human being as our benchmark? Such questions are also of great importance for forensic audio analysis where a mixture of automatic and semi-automatic methods and aural recognition are commonly used [1, 2].

A couple of studies have compared human and machine performance in the speaker [3, 1, 4] and language [5] recognition tasks. In this paper we focus on the speaker recognition task (Table 1). One of the most extensive comparisons between aural and automatic systems has been conducted a decade ago [3]. In that study, human speaker recognition performance was compared to three automatic systems on the NIST 1998 speaker recognition evaluation (SRE) data. The average human equal error rate (EER) of all trials was 23 %. The accuracy was improved to 12 % after combining all the listeners' verification scores by averaging. In matched channel conditions, human mean and best automatic systems gave both an EER of 8 %. However, in channel-mismatched condition, human mean was 14 % EER whereas machine accuracy was degraded to 24 % EER, supporting the assumption that humans are more robust under signal distortions. It should be noted, however, that while human *average* was good, there were large variances between the individual listeners. It is also noteworthy that the listening experiment in [3] was done in a controlled laboratory environment where the listeners needed to make decisions within short intervals.

More recently, in [1] human speaker recognition performance was compared against automatic system in a forensic setting. Unlike in [3] where the listening was strictly controlled, the subjects could listen to the material as long as they wanted. The material included forensic material (French polyphone-IPSC02 corpus) from 10 speakers under three different conditions. There were as many as 90 listeners, each listening to 25 verification trials. The accuracy was compared to *Gaussian mixture model* (GMM) recognizer using *perceptual linear prediction* (PLP) features. The conclusions were similar to [3]: under channel mismatch, human listening pool outperformed the automatic system. It was found, interestingly, that the automatic system outperformed humans in the matched channel conditions.

In another study [4], focusing mainly on speech disguise but also comparing average human accuracy to a more modern Gaussian mixture model - universal background model (GMM-UBM) [6] system, the authors had a self-collected corpus with 32 speakers recorded in four different sessions. In two or more of the sessions the speakers were asked to disguise their voice to *not* sound like themselves. The listener pool included 25 listeners and, similar to [3], the listening was done under controlled set-up where the listeners could not play with the sam-

Table 1: Previous work comparing human and automatic systems in speaker recognition task. Missing values marked by the n/a were not available in the original papers.

Study	Corpus	Data type	Channel	Machine system	No. trials	No. listeners	Machine (EER)	Human (EER)
Schmidt-Nielsen & Crystal [3]	NIST 1998 subset	Telephone	Mixed	n/a	3172	65	n/a	12%
			Matched	n/a	n/a	65	8%	8%
			Mismatched	n/a	n/a	65	24%	14%
Alexander *et al* [1]	IPSC02 subset	Telephone	Mixed	GMM	apprx.500	90	n/a	n/a
			Matched	GMM	n/a	90	4%	16%
			Mismatched	GMM	n/a	90	30%	30%
Kajerekar *et al* [4]	Self-collected	Lab	Clean	GMM-UBM	2484	25	0.05%	0%
Hautamäki *et al* **[this study]**	NIST 2008 subset	Telephone & interview	Mixed	GMM-UBM with JFA	40	36	20%	22%

ples. The conclusion from the study was clear but different from both [3] and [1], in the favor of automatic method: the GMM-UBM system systematically outperformed human average, and for the disguised test segments the difference was even greater. It should be noted that the data was recorded in laboratory condition without any channel variability.

Even though the methodologies, experimental set-ups and automatic system configurations in the cited studies are diverse, an interesting pattern is that the recent recognition systems are getting closer to or even performing better than humans. During the past 4 to 5 years, the automatic systems have seen dramatic accuracy improvements (from range $\sim [5-10]\%$ EER to $\sim [1-3]\%$ EER on typical NIST SRE core tasks in SREs 2004–2008 corpora), thanks to advanced statistical channel compensation techniques in the so-called *supervector* space [7, 8, 9]. In particular, the *joint factor analysis* (JFA) [8] which represents the state-of-the-art of the field. We are, however, unaware of any comparisons between JFA-compensated automatic system and human listeners. Have we finally reached the point where human is systematically worse than the automatic system? This is the main question we address in this paper.

In addition, differently from the previous studies [1, 3, 4], we utilize the most recent NIST 2008 SRE corpus in our comparisons. A special feature of this corpus is inclusion of microphone data from an interview scenario. Subjects who already participated in telephone conversations were also invited to participate in a separate recording session in a room built for the purpose. Subject's voice was simultaneously recorded using a number of different far-talking microphones. NIST 2008 SRE core task contains subtasks where one of the recordings for the trial pair is from interview microphone speech and the other from telephone conversation. We hypothesize that the automatic methods outperform human listeners even under this severe channel degradation.

2. Experimental Setup

2.1. Task Definition

The task considered in this paper is *speaker verification*, that is, deciding whether two given utterances are spoken by the same speaker. A single utterance pair which the decision must be declared for (either by human or automatic method) is called *trial*. For the automatic method, one of the utterances is considered as the enrollment utterance and the other one the test utterance. For human listeners, however, we did not impose such artificial training/test division but the listeners could listen to the samples in any order and as much as they wanted to. This is different from previous studies [3, 4] where the listening was controlled.

2.2. Selection of Trials

A total number of 40 verification trials were selected from NIST 2008 speaker recognition evaluation corpus. Half of the trials were male-male trials and half female-female trials (no cross-gender were selected as these are considered too easy). In our experimental setup, we set the detection costs to unity and prior probability of the target speaker to $p_{\text{target}} = 0.5$. These costs and probability of target/non-target were also reported to the listeners. With this in mind, *equal error rate* (EER) approximately equals the classification error for the whole set.

All trials in the core test were first scored using the automatic fusion system described in [10]. For practical reasons we chose English-only trials. The threshold at the equal error rate (θ_{EER}) was first determined from the set of scores. A set of 20 trials, denoted herein as *hard*, were selected by their closeness to the EER threshold. In this set, the target trials (i.e. the utterances originating from the same speaker) were usually from severely mismatched channel conditions, whereas the non-target trials were those with quite distinctive higher-level features such as accent, speaking style and intonation. To get diversity into our trial pool, we selected another set of 20 trials which the I4U fusion system [10] classified correctly. We refer to this set as *easy*. In the easy set, channel mismatch is not very apparent. On the other hand, as seen in the spectrograms of Fig. 1, there are severe channel mismatches in the *hard* set.

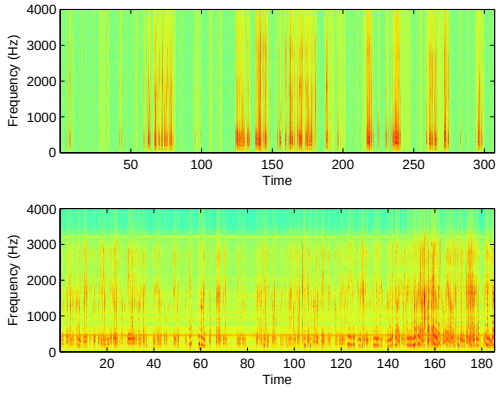

Figure 1: Female-female trial from the *hard* set.

2.3. Listening Setup

Previous studies [3, 4, 5] comparing human performance to automatic system have suggested a controlled environment for the listening exercise, that is, subjects had to listen in a specific room and make decisions under supervision. However, it is difficult to put a time constraint on listeners. Typically, when a difficult task is given to the listener to make the final decision, (s)he prefers to browse both utterances simultaneously to look for the similar speech patterns or the proof that speakers are definitely different. This fact motivated us to design our experiments in a way that listeners have the freedom to do the test in their own workstations and pace regardless of the length of listening time.

The verification trials were delivered as two-channel audio files over FTP-transfer from the listeners in three different sites. A total number of 36 listeners were recruited from Singapore, Australia and Finland. The participants were graduate students, post-docs or senior researchers in computer science, electrical engineering or speech technology. None of the listeners possessed any formal training in forensic speaker recognition, hence our listeners are characterized as "naive". There is also large linguistic variability between the listeners: none of the listeners was native English speaker and spoken English proficiency was not uniform over all listeners. First languages of the listeners included, for example, Mandarin, Finnish, Farsi, Vietnamese and Malay.

2.4. Automatic System

In the speaker recognition task, the duration of enrollment utterances generally varies from a few seconds to a few minutes. Under this constrained condition, a speaker model is always obtained based on *the universal background model* (UBM) via an adaptation process referred to as *maximum a posteriori* (MAP) estimation [6]. Given a test sample X, the confidence score $s(X)$ is defined as the log-likelihood ratio

$$s(X) = \log p(X|\lambda) - \log p(X|\theta), \quad (1)$$

where λ and θ represent the speaker-dependent GMM and the UBM, respectively. For the same speaker, the GMM models estimated from different training utterances and channels are generally different. Channel compensation is therefore necessary to make sure that the test data obtained from a different channel (than the training data) can properly be scored against the speaker models.

Joint factor analysis (JFA) [8] is a modeling technique, built on top of the GMM-UBM framework, to address the issue of speaker and session variability. More specifically, JFA assumes that the speaker and session variabilities can be modeled separately with two low dimensional subspaces. The crux of JFA is to determine these subspaces and characterize the speaker and channel variability jointly in terms of the so-called *latent factors* under these subspaces. The EER threshold was determined from all NIST SRE 08 core trials and that threshold was then used for making the binary decisions. In other words, we assume that complete system is free of calibration errors.

In the experiment, the gender-dependent UBMs, with 1024 Gaussian mixtures, were trained using NIST SRE04 data. The speaker subspace (with 300 speaker factors) was trained using the data drawn from the Switchboard Cellular and Switchboard Landline. The channel subspace (with 200 channel factors) was trained using the data drawn from NIST SRE04, SRE05 and SRE06. Score normalization was applied with t-norm followed by z-norm.

2.5. Fusion of Human Decisions and Cross-Validation

We used the simple *majority voting* fusion strategy. It needs no tuning and was found out to be robust method, only losing slightly over trained systems [3]. We have noticed that fusing all listener decisions is not an optimal fusion strategy. We have thus decided to experiment with varying the listener fusion pool compostion, by dividing the 40 trial set randomly into two subsets, thus mixing hard and easy trials and finding the best listener pool for each. Then found listener pool was applied to the other set. For each randomly selected subset we find the individual classifier performances. We then start to build successive set of classifier pools by starting with the all 36 in the same pool and then leaving out one by one worse performers, until only the best performer is remaining. From these pools we pick the one that gives best performance for the training set.

We also found best individual human listener for each set in that way and applied his/her decisions on the testing set. Similarly, JFA decisions were also applied. In this cross-validation setting, we will gain confidence on our results

3. Results

Performance of human listeners and JFA system on the 40 trials are summarized in Table 2. For the hard set, some listeners (12 in total) performed worse than chance level. Of course, we could take negation of each listener decision, whose error rate was above 50%. But we have no confidence that the negation operation generalizes to unseen data, so we decided to leave decisions as is. Average performance of hard set was 40.42%, which was improved to 25% using majority vote. Similarly, in easy set human average is improved from 26.25% to 20% using majority vote. Single human outperforms the majority vote of all 36 listeners. However, in Fig. 4 we notice that majority vote is more stable (with average testing error being 29.9%, when best individual is only 33.5% and JFA achieving on average 20.3%). Using best individual as a fusion rule, there is higher chance than with voting fusion to get error of more than 50%.

Table 2: Comparing the basic statistics of the two sets of trials using the pool of listeners and JFA system. Statistics include, minimum, maximum and average human error rates.

Data set	Min	Max	Avg	Fusion	JFA
Hard	20%	60%	40.42%	25%	40%
Easy	10%	45%	26.25%	20%	0%

Figure 2 displays DET plot comparing JFA and majority voting of all 36 listeners. Voting results are turned to score by a ratio of listeners that agree with the majority decision, by $s = N_{\text{true}}/N$, where N_{true} is the number of votes for true and N is the total number of votes. Human and machine perform comparably around the EER region but JFA outperforms humans at the low false acceptance region. A likely reason is that naive listeners tend to decide "different speaker" under channel mismatch which causes increase in false rejections. The JFA system, in turn, is designed for tackling the speaker and channel variations. Distribution of optimum human fusion pool size (Fig. 3) indicate that seldom all the listeners fuse; a subset of only 3 (good) listeners provides best results.

Figure 4: Distributions of classification testing errors from the randomized cross-validation trials (average error rates are: individual 33.5%, fusion 29.9% and JFA 20.3%).

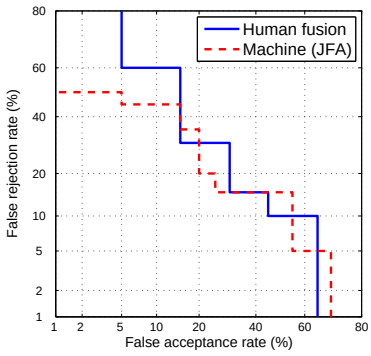

Figure 2: Human vs. machine (JFA) over all 40 trials.

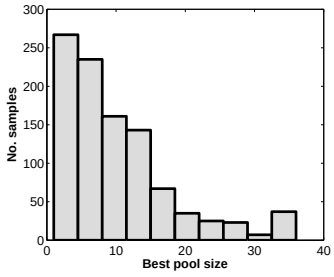

Figure 3: Distribution of the empirically best classifier pool size. Estimated from the training data.

4. Conclusions

We hypothesized that advances in speaker verification technology have resulted in systems surpassing human accuracy. We have partly being able to show that JFA system achieved 20% error rate, whereas fusion of human decisions resulted in 22% on the whole set. Using cross-validation we also found that JFA is the best performer with average error of 20.3%, voting fusion with 29.9% and using only one single best human 33.5%.

5. Acknowledgment

Work of V. Hautamäki and T. Kinnunen was supported by the Academy of Finland (projects 131298, 132129). The authors thank Dr. Julien Epps and the listeners at UNSW, I2R and UEF for their support and participation in the listening experiments.

6. References

[1] Anil Alexander, F. Botti, D. Dessimoz, and A. Drygajlo. The effect of mismatched recording conditions on human and automatic speaker recognition in forensic application. *Forensic Science International*, pages S95–S99, 2005.

[2] P. Rose. *Forensic Speaker Identification*. Taylor & Francis, London, 2002.

[3] A. Schmidt-Nielsen and T.H. Crystal. Speaker verification by human listeners: Experiments comparing human and machine performance usin the nist 1998 speaker evaluation data. *Digit. Sign. Proc.*, 10:249–266, 2000.

[4] S.S. Kajarekar, H Bratt, E Shriberg, and R de Leon. A study of intentional voice modifications for evading automatic speaker recognition. In *Speaker Odyssey 2006*.

[5] David van Leeuwen, Michaël de Boer, and Rosemary Orr. A human benchmark for the NIST language recognition evaluation 2005. In *Speaker Odyssey 2008*.

[6] D.A. Reynolds, T.F. Quatieri, and R.B. Dunn. Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing*, 10(1):19–41, Jan 2000.

[7] A. Solomonoff, W.M. Campbell, and I. Boardman. Advances in channel compensation for SVM speaker recognition. In *ICASSP*, pages 629–632, March 2005.

[8] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel. A study of inter-speaker variability in speaker verification. *IEEE Trans. Audio, Speech and Language Processing*, 16(5):980–988, July 2008.

[9] T. Kinnunen and H. Li. An overview of text-independent speaker recognition: from features to supervectors. *Speech Communication*, 52(1):12–40, January 2010.

[10] Haizhou Li, Bin Ma, Kong-Aik Lee, Hanwu Sun, Donglai Zhu, Khe Chai Sim, Changhuai You, Rong Tong, Ismo Karkkainen, Chien-Lin Huang, Vladimir Pervouchine, Wu Guo, Yijie Li, Lirong Dai, Mohaddeseh Nosratighods, Thiruvaran Tharmarajah, Julien Epps, Eliathamby Ambikairajah, Eng-Siong Chng, Tanja Schultz, and Qin Jin. The I4U system in NIST 2008 speaker recognition evaluation. *ICASSP*, 0:4201–4204, 2009.