Odyssey 2014:
The Speaker and Language Recognition Workshop
16-19 June 2014, Joensuu, Finland

Comparison of Human Listeners and Speaker Verification Systems Using Voice Mimicry Data

Rosa González Hautamäki[1], *Tomi Kinnunen*[1], *Ville Hautamäki*[1] *and Anne-Maria Laukkanen*[2]

[1]School of Computing, University of Eastern Finland, Joensuu, Finland
[2]Speech and Voice Research Laboratory, School of Education, University of Tampere, Finland
{rgonza, tkinnu, villeh}@cs.uef.fi,
Anne-Maria.Laukkanen@uta.fi

Abstract

In this work, we compare the performance of human listeners and two well known speaker verification systems in presence of voice mimicry. Our focus is to gain insights on how well human listeners recognize speakers when mimicry data is included and compare it to the overall performance of state-of-the-art speaker verification systems, a traditional Gaussian mixture model-universal background model (GMM-UBM) and an i-vector based classifier with cosine scoring. We have found that for the studied material in Finnish language, the mimicry attack was able to slightly increase the error rate in a range acceptable for the general performance of the system (EER from 9 to 11%). Our data reveals that enhancing the audio material by minimizing the differences of data collected in different environments improves the accuracy of the speaker verification systems even in the presence of mimicked speech. The performance of the human listening panel shows that successfully imitated speech is difficult to recognize, even more difficult to recognize a person who is intentionally trying to modify his or her own voice. The average listener made 8 errors from 34 selected trials while the automatic systems had 6 error in the same set.

1. Introduction

The accuracy of speaker verification systems has steadily improved in the recent years due to advances in methods that counteract against undesired channel, environmental noise and session variations. Such systems are gaining demand as a recognition tool in devices and services that require subject's identity verification. Also the accuracy of the systems has a direct relation to the quality of the audio samples and the features used to characterize speakers. Depending on the application, the users of a speaker verification system are considered *cooperative* or *non-cooperative* [1]. A cooperative user wants to have himself or herself correctly recognized to achieve logical or physical access. A non-cooperative user would, voluntarily or not, provide a speech sample with the intention of not getting recognized, by disguising his or her voice. While some voices from cooperative subjects are matched correctly against themselves, others are frequently confused with other speakers' voices [2, 3]. Non-cooperative setting, or attack against an automatic system [4], can be achieved using technical means, such as voice conversion [5], adaptive speech synthesis [6] and replay attacks [7].

In this work, we concentrate on *mimicry attack*, which cannot be easily detected by technical means since the speech is produced by an actual human being and not by a speech synthesis or voice conversion algorithm. Mimicry attack not only raises an interesting non-cooperative scenario that could threaten speaker verification systems, but it is also phonetically relevant to know how speakers can modify their voices to sound like another person.

Previous studies have evaluated mimicked speech and the role of the impersonator, either professional or not, to mimic the targets' speaking characteristics related to prosody, pitch, dialects and speaking style. It has been reported that impersonators are often able to adapt especially the fundamental frequency and occasionally also the formant frequencies towards the target voices [8, 9, 10]. A visual acoustic comparison of the imitator's natural voice and his impersonation, and the target's voice is shown in Fig. 1. In fact, [8, 11] used automatic speaker recognition technology to objectively evaluate the success of voice imitation. The authors in [8] used a prosody-based speaker recognition system and found that fusion of 12 prosodic features increased the impersonator's efficacy. The authors of [11], in turn, evaluated mimicked speech with prosodic features based on intonation, duration and energy. 9-dimensional feature vectors from original and mimicked speech were compared using *dynamic time warping* (DTW) alignment. The best mimic attempt obtained a high speaker similarity score. The authors further carried out a listening test to grade the mimicked speech, and the results indicate agreement between the automatic prosodic system scores and the listeners' opinion. In other studies, focuses have been given on analyzing the vulnerability of speaker verification systems in presence of voice mimicry. For example, in [12, 13, 14], vulnerability of Gaussian mixture model (GMM) based systems was investigated. These studies indicate that if the target of impersonation is known in advance and his/her voice is close to the impersonator's voice, then the chances to spoof an automatic recognizer are increased.

Our recent work on mimicry attack [15] presented a vulnerability study of two well-known speaker verification systems. Though we found that the accuracy of automatic recognizers was not much affected by mimicry, the observation was unreliable due to the technical differences between the target speaker's data harvested from the web and the studio recordings of the impersonator.

This problem can be possibly solved by: a) recording impersonator's and target speakers' voices in the same studio or room, b) using acoustic features that are comparably more robust to additive noise and channel differences such as prosody, or c) include a perceptual test as a human benchmark parallel to the automatic system. Regarding the first option, having an impersonator and target speakers' voices recorded in the same room is not feasible in practice (see [14] for an exception), and the accuracy of prosody-based systems are not yet comparable to spectral-based systems. Human benchmark com-

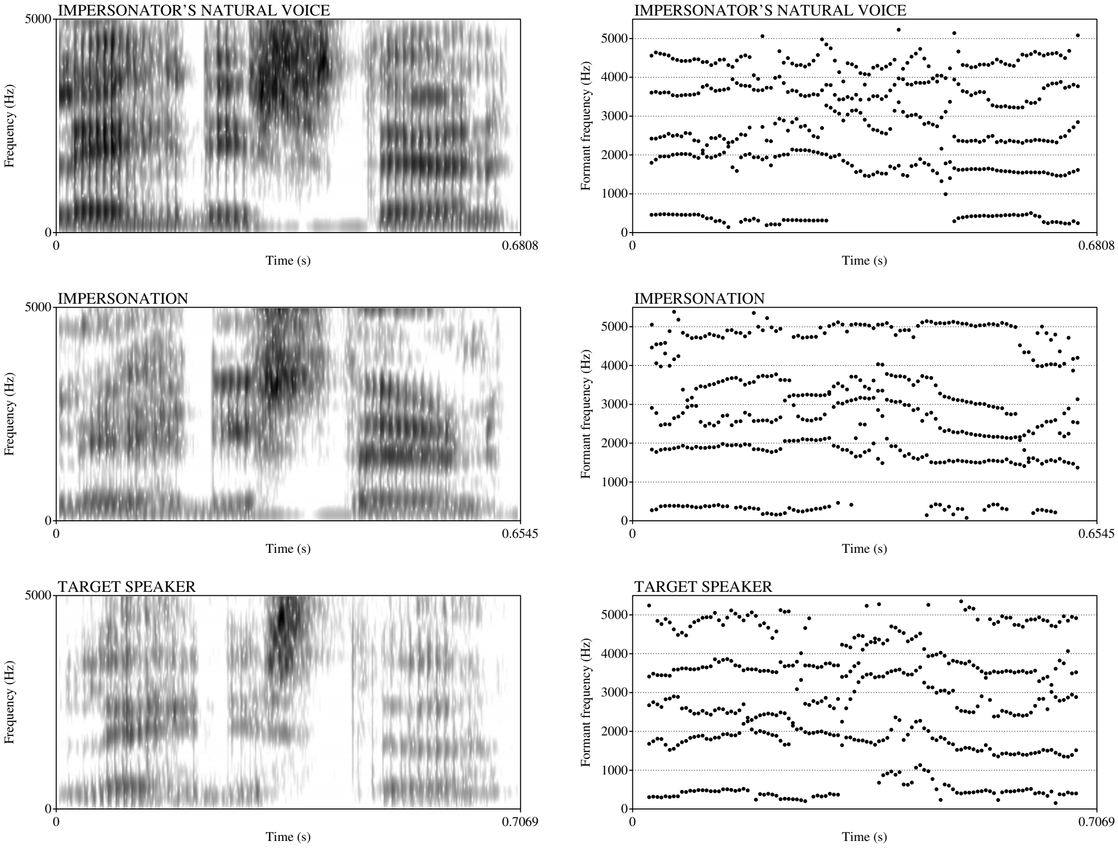


Figure 1: An example of speech impersonation. Spectrogram and formant tracks (F1 through F5) of the impersonator's own voice (top), impersonation (middle) and the target speaker (bottom). Formants computed using Praat. The target speaker is the current president of Finland, Sauli Niinistö. Comparing the top and middle figures, the impersonator can modify his voice away from his natural vocal tract configuration (for instance, F4 is lowered and F5 raised). Even if the formants do not quite match those of the target speaker, the impersonation is perceptually convincing to a native listener.

pared to automatic systems has been used in previous studies [16, 17]. In terms of human assisted speaker verification system [18, 17, 19], such as a forensic system, it is important to know how a non-cooperative subject could either mimic some other speaker or disguise his or her voice. In addition to improving speaker recognition methodology, it is also of great interest to compare human speaker verification performance to that of an automatic system. In this work, we extend it to non-cooperative case involving a professional impersonator. We find out that listeners are able to recognize the speaking style of well-known personalities or identify if a person is "acting" another person's voice. On the other hand, in successful imitation trials, listeners are uncertain about their decisions. The effect is even stronger in disguise trials, where most of the listeners made mistakes. Difficult cases for listeners seem to involve speakers that sound different in separate recording sessions. Also familiarity with the voice in the case of well-known target speakers was considered.

2. Material

The main challenge in studies involving mimicry is the scarcity of the data. The existing data is created for a specific study, which is not publicly available and cannot be considered as a standard evaluation corpus [15]. Not only data collection is expensive, but also finding professional impersonators (voice actors, singers or entertainers), with available time to create the corpus is difficult. In addition, the target speakers are usually well-known public figures and their speech samples are collected from radio interviews and TV programs. As a consequence, there are technical mismatches since the impersonator's voice have been recorded in a studio environment.

2.1. Target speakers

A speech database containing the voice of five well-known Finnish public figure is used for this study, including: the current president of Finland (Sauli Niinistö), a former president (Martti Ahtisaari), a former prime minister (Matti Vanhanen), a theatrical director (Jouko Turkka), and a businessman (Hjallis Harkimo). The actual speakers voice was collected from radio interviews and TV programs where as their voices are mimicked by a professional impersonator whose speech samples were recorded in a quiet studio environment. This voice data was collected by [20] and subsequently, it was used for mimicry attack analysis in [15].

2.2. Technical aspects

To create a scenario in which mimicry attack could threaten a speaker verification system, we focus in telephone quality data. All audio samples were down-sampled to 8 kHz and converted to mono. The audio segments were preprocessed to reduce the differences caused by channel differences and environmental noise. A speech enhancement algorithm based on the log estimation of the complex spectrum of the signal, known as logMMSE [21], was applied to all speech segments. The logMMSE estimator reduces residual noise without greatly affecting the speech signal. For this study, we used the implementation presented in [22]. Other speech enhancing algorithms based on *Wiener* filter were experimented with the mimicry data, and in a subjective comparison it introduced noticeable level of distortion in some of the frequency bands. As noted in [23], the logMMSE speech enhancement method did not degrade significantly the sound quality nor the speech intelligibility, our only requirement for the listening test.

2.3. Corpus design for the experiments

The training material for target speakers includes a maximum of 5 minutes of active speech. The length of the test utterances was set to 20 second chunks from original utterances of varying length. The professional impersonator's natural voice (no mimicry) was recorded reading segments from interviews of the target speakers in addition there are two mimicry samples per target speaker. The trial list of the *baseline case* contains the test segments of the target speakers and the impersonator's natural voice as the impostor. In the *mimicry attack* test, the genuine trials were kept same as of baseline case and impersonator's samples mimicking the target speakers were set as the impostor trials. In this way, the effects for the system performance were compared between the case in which the data included mimicry and the baseline. The trial list contains 27 genuine trials to represent the 5 target speakers and 155 impostor trials. The test segments were set to have the same length and in this study their speech content do not match.

3. Experiments

3.1. Speaker verification systems

In the present study, two speaker verification systems are considered where both utilize a 54-dimensional Mel-frequency cepstral coefficient (MFCC) as feature extractor. The first system is based on a classical **Gaussian mixture model with universal background model** (GMM-UBM) [24]. The other system is based on a recently introduced **i-vector with cosine scoring** [25]. In both systems, the UBM of 512 mixtures is trained with NIST 04, 05, 06 and 08 data.

For the *i-vector with cosine scoring*, given two utterances represented by two vectors in the *i-vector* space [26], the angle between the two vectors, or *cosine similarity*, is considered as a measure of similarity [27]. The i-vector extractor produces 400-dimensional i-vectors and its T-matrix is trained using the same data as UBM plus Fisher and Switchboard corpus.

3.2. Listening test

The listeners in our experiment are native Finnish speakers. 19 female and 15 male with an age range from 20 to 65 years old had participated in a web-based listening test, to compare 34 pairs of speech samples (See form screen in Fig. 5). The listeners are considered naïve since no formal training was required to participate in this experiment. The sample pairs are a subset of the corpus used for the automatic system, with the test segments length of 10 second duration. The length of the speech utterances was set at this size in order to facilitate the listening as the total number of trials is 34. In addition to speech enhancement described in section 2, all the speech samples were further normalized to have the same active speech level. We estimated the active speech level with the implementation provided in the VOICEBOX speech processing toolbox [28] called `activlev` function.

The listening test was conducted in two cities: Joensuu and Tampere where the two collaborating groups in this study are located. Students, acquaintances and coworkers with little or no knowledge of our study were invited to participate in the test. The listening tests were scheduled in a silent office environment of approximately 15 square meters area. In Joensuu, a desktop computer was used with Sennheiser HD 570 headphones, in Tampere a laptop computer with two audio interface devices, Motu Ultralite mk3 and a Roland Quad capture, in addition to AKG and Sony studio headphones was used.

The type of listening trials comprising the test are described in Table 1. The only instruction given to the listeners was to listen to each pair and compare the voices in the samples. The listeners were not told that voice mimicry was included in some of the trials. For each trial, the listeners had to select their decision as one of the following options: *Same speaker, somewhat the same speaker, I cannot tell, somewhat different speaker, different speaker*. After completing the test, the participants were asked to name the speakers that they had identified in the samples and also describe the cues they used to differentiate the speakers.

Table 1: *Distribution of the 34 trials per speaker for the listening test.*

Speaker	Trials		
	Genuine	Impostor	
		Baseline	Impersonator
Martti Ahtisaari	2	2	2
Hjallis Harkimo	2	2	2
Sauli Niinistö	2	2	2
Juoko Turkka	2	2	2
Matti Vanhanen	2	2	2
Impersonator	4	-	-

It is worth mentioning that, during the preliminary test for the listening test, the organizers faced the question whether the task would be too easy for the listeners. Not only the speech samples from the target speakers belonged to different interview segments, but the context of the utterance could also make the comparison more a matter of channel differences or a comparison of the conversation content. However, during the test it became clear that analyzing 34 speech samples for an uninformed listener was not easy as we expected. Most of the participants reported a considerable amount of effort to compare the speakers' voices.

Table 2: Effect of mimicry attack in terms of *equal error rate* (EER %).

Material	Test	GMM-UBM	i-vector Cosine
Original audio	Baseline	11.11	9.03
	Mimicry attack	9.68	11.61
Enhanced audio	Baseline	7.08	0.59
	Mimicry attack	5.52	4.41

4. Results

4.1. Automatic systems results

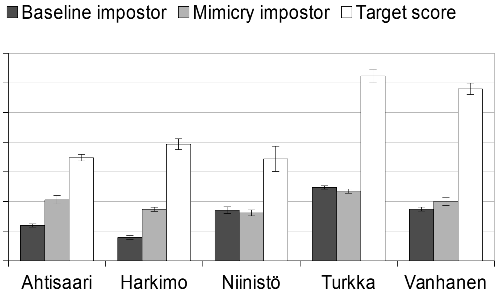
To analyze the effect of imitation spoofing, we present the performance in terms of *equal error rate* (EER) which corresponds to equal miss and false alarm rates. In Table 2, the performance for GMM-UBM system shows contradictory results for the two study cases, this could be the effect of the limited data for the study. Also, we noticed a slight increase in the EER for the i-vector cosine system when mimicry is present. It is noticeable that the systems performed much better when the data was preprocessed by logMMSE enhancement before feature extraction. Even though EER is a standard measure for comparing the accuracies of verification systems for a very large number of trials, in this study it does not provide us enough information due to a limited number of trials. Therefore, an alternative method is used were the scores for each target speaker are analyzed to evaluate the effect of mimicry attacks in speaker verification systems. In Fig. 2, the average recognizer score per target speaker is calculated before the mimicry attack (baseline) and after it (mimicry case). The graphs show the score distribution for the enhanced audio data and include the standard error of the mean (SEM) with confidence range of 95%.

Comparing the heights of the baseline target graphs – a measure of the similarity of our imitator's *natural* voice against a particular target – Ahtisaari and Niinistö appear to be the most similar to the imitator's voice, while Turkka and Vanhanen have lower impostor scores. The same pattern holds for both recognizers. Previous studies [10, 13] have suggested that imitation attacks against "similar" target speakers might be easier than against speakers with very different voice quality. Figure 2 shows that the mimicry scores against the most similar target, Niinistö, is *lower*, while the relative increase is largest with targets like Harkimo and Ahtisaari.

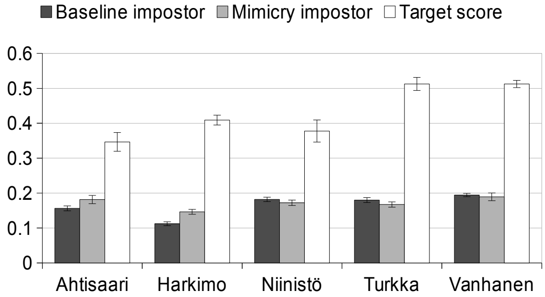
The authors in [12] used a verification system that had high-quality clean input signal and controlled text passages to select the most similar and dissimilar speakers for the impersonator. Our study deals with a scenario with free-text inputs and involve different sound quality target recordings. It is likely that the lexical and channel variation masks the effects of impersonation. To be able to ideally focus on the success of the impersonation, all speech samples should be recorded under the same conditions. However, this is neither practical nor does it represent a threat that could occur in a real world scenario.

4.2. Listening test results

To analyze the difficulty of mimicry attack scenario for human listeners, we carried out a listening test with 34 speech pairs. The small number of trials considered for the listening test allows a thorough trial by trial comparison. The grid in Table 3 shows the listeners' decisions for each of the 34 trials. The errors per trial are shown in terms of false alarms and misses. We identify three main types of trials:



(b) GMM - UBM



(a) ivector-Cosine scoring

Figure 2: Score distribution comparison per target speaker for data duration constraint of enhanced audio samples. The bars also show the standard error of the mean with 95% confidence.

"Easy" trials. The trials with less than or equal to five errors (2, 5, 6, 13, 16, 19, 21, 24, 27, 30) correspond mainly to impostor trials and some genuine trials (11, 15, 18, 23, 32).

Trials with more misses. The group of trials with higher number of errors are mainly genuine trials, for example speech pairs with impersonator's natural voice against his impersonations (trials: 31, 34). These are *disguise* trials because the impersonator attempts to sound different from his natural voice. Other set of trials with more misses correspond to genuine trials as in 7 and 12, these samples are from the target speakers corresponding to different sessions. This perceptual inter-speaker variability could be more related to the context of the conversation: the conversation is more animated, the topic is more attractive, etc.

Trials with more false alarms. One more source of errors are trials with target voice against impersonation as in trials 10, 14, 25.

It is worth noticing that for trials 14 and 25 half of the listeners responded that the samples correspond to the same speaker versus different speaker while these, in fact, were mimicry trials.

We analyze the type of errors the listeners made for target and non-target trials in Fig. 3. Confidence intervals are computed from binomial distribution with 95% confidence region.

Table 3: Listeners total errors trial-by-trial. The errors are shown highlighted. The decision number indicates the confidence level: *1: Same speaker, 2: somewhat the same speaker, 3: I cannot tell, 4: somewhat different speaker, 5: different speaker.*

		LISTENERS																																
Type of trial	Trial #	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	Errors
Genuine	1	3	1	1	5	1	1	1	2	5	1	1	2	1	1	2	1	2	2	5	1	1	2	1	1	1	5	1	2	5	5	1	1	1	4	8
	4	1	1	2	1	2	2	1	1	4	1	2	5	2	1	1	2	4	2	1	4	1	1	1	1	2	5	1	2	2	5	1	1	1	1	5
	7	5	5	4	2	4	5	5	5	5	5	5	5	4	5	5	5	2	5	5	5	5	1	5	5	2	5	2	2	5	5	5	4	1	5	27
	11	1	1	4	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	5	1	1	1	1	2
	12	2	2	4	5	4	5	4	1	5	1	4	4	3	2	2	5	2	2	1	5	5	2	2	1	5	5	5	2	5	5	5	4	2	1	19
	15	1	2	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	0
	18	1	1	1	1	4	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	2	2	1	1	1	1	1	1	1
	22	2	4	2	3	5	2	2	2	5	5	2	1	2	5	2	5	2	4	1	1	4	1	1	4	2	1	2	2	5	5	5	1	1	1	13
	23	1	1	1	2	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0
	26	5	2	2	2	5	1	1	1	5	5	5	3	2	5	5	5	2	1	1	5	1	5	2	1	5	1	2	2	1	5	1	2	2	2	13
	29	1	4	4	2	4	1	1	4	1	2	5	5	1	1	5	1	1	1	1	1	1	1	1	1	1	1	2	2	1	2	1	2	2	1	7
	31	2	2	4	2	5	5	2	5	5	5	5	5	2	5	5	5	2	5	2	2	2	5	4	5	5	5	2	5	5	5	5	4	2	2	22
	32	1	2	1	1	1	2	1	1	1	2	3	1	4	4	1	1	1	1	1	1	1	4	1	2	1	1	2	2	1	2	1	1	2	1	4
	34	5	4	5	5	5	5	2	5	5	5	5	5	4	5	5	5	4	5	5	3	5	5	5	5	5	5	5	5	5	5	5	5	5	2	31
Impostor	2	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	2	5	5	1
	5	5	5	5	5	5	5	5	5	5	5	5	4	5	5	5	5	5	3	5	5	5	4	5	5	5	5	2	5	5	5	5	5	5	5	2
	8	4	5	2	5	5	5	5	5	5	2	5	4	4	5	4	5	5	5	5	5	5	1	5	5	4	2	3	5	5	5	5	5	5	1	6
	13	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	0
	16	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	2	5	5	5	5	5	5	5	5	4	5	2	2
	19	5	5	5	3	5	5	5	5	5	5	5	4	4	2	5	5	2	5	5	5	5	5	1	5	5	5	5	5	5	5	5	4	5	2	5
	24	5	5	5	4	5	5	5	5	5	5	5	5	4	5	4	5	2	5	5	5	5	5	2	5	5	5	5	5	5	5	5	2	5	4	3
	27	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	1	2	5	5	5	5	5	4	5	1	3
	30	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	2	5	5	5	5	5	5	5	5	5	4	5	3	2
	33	5	5	4	4	5	5	5	5	1	5	5	5	5	5	5	5	3	5	5	4	5	2	5	5	5	5	4	5	5	5	5	2	5	1	5
Impostor Impersonator	3	5	5	5	2	5	4	5	5	5	5	5	5	4	5	4	5	5	2	5	2	5	5	5	2	5	5	5	4	1	5	2	2	2	1	9
	6	5	5	5	2	5	5	5	5	5	5	5	5	4	5	5	5	5	4	5	5	5	5	5	5	5	5	5	4	5	5	2	5	2	5	3
	9	5	5	5	5	5	5	2	2	5	5	5	2	5	5	5	5	2	4	5	5	5	5	1	5	5	5	2	4	5	5	4	5	2	5	7
	10	4	5	5	4	5	5	2	3	5	5	4	3	5	5	5	5	5	2	4	5	5	2	5	4	2	5	2	2	1	5	2	2	2	1	13
	14	2	5	4	1	5	2	2	3	5	5	5	3	5	5	4	5	1	2	5	1	5	2	5	2	5	1	2	4	1	5	5	2	2	1	17
	17	2	5	5	3	5	2	5	5	5	5	4	5	5	5	4	5	1	4	4	4	5	2	5	2	5	1	5	2	5	4	5	5	5	2	9
	20	4	4	2	2	5	5	5	1	5	5	2	2	5	5	4	5	5	5	4	5	5	5	4	5	5	1	3	5	5	5	5	4	2	5	8
	21	4	4	5	5	5	4	5	5	5	5	4	5	5	5	5	5	5	1	4	5	5	4	5	5	5	1	4	5	5	5	4	5	2	1	4
	25	5	4	2	4	3	5	5	5	5	5	3	2	4	5	5	5	4	5	1	2	4	3	5	5	2	1	2	1	1	2	2	2	2	1	17
	28	4	2	5	4	5	5	5	5	5	2	5	5	5	5	5	5	1	4	1	5	5	2	5	5	5	1	5	4	1	5	4	5	2	1	9

■ Misses ░ False accepts ◪ Cannot decide

The method used to compute the confidence interval is Clopper-Pearson, the so called "exact" method [29]. The non-target trials are divided in two categories: impersonator's "*natural*" voice (*baseline impersonator*) and mimicry trials. On average, the listeners answered incorrectly for 8 out of 34 trials. The least successful listener made 15 errors and the "*best*" two listeners made only 4 errors. Comparing the impostor and impostor impersonator distributions, it is observed that the intervals for incorrect decision do not overlap. This indicates that the impersonator is able to increase the amount of errors in listeners' decisions in a statistically significant margin when mimicking the target speakers.

The distribution of "*same speaker*" decisions from the 34 listeners is shown in Fig. 4 for each of the target speakers. The graph indicates that the answers corresponding to *same speaker* for genuine trials is higher in most cases except in Turkka's trials. The listeners had difficulty deciding if the trials containing his voice samples correspond to the same or different speaker. The target speaker is a theatrical director and his speech samples are segments from different recordings in which his speaking style changes considerably. This likely caused the listeners to conclude the speaker to be different. Similar confusion was noticed when his voice was compared to the impersonator's natural voice and even more in impersonations of his voice. One reason for confusion could be that just few of the listeners recognized him as one of the target voices indicating unfamiliarity with his voice.

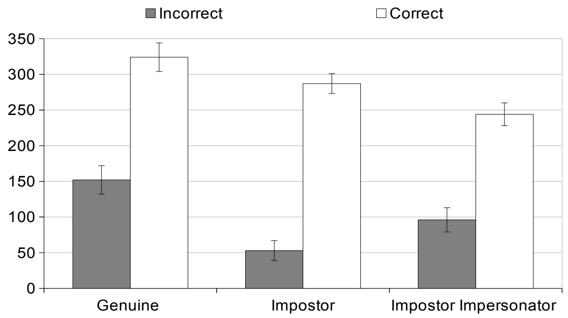


Figure 3: Listening test correct and incorrect decisions by the listeners for target and nontarget (impostor, impostor impersonator) trials.

In the "same speaker" decision for the mimicry trials, we observe an increase in the height of the distributions for all the target speakers except Ahtisaari, indicating that few listeners confused the impersonated samples for the target's voice. After the listening test, most of the listeners were able to name Martti Ahtisaari and Sauli Niinistö as part of the target voices. However, identifying all the speakers in the test did not affect the performance of a listener, for example 2 from the 34 listeners reported correctly the target speakers, nevertheless they made as many errors as the average listener. The two listeners with only 4 errors in the test identified 3 of the 5 target speakers.

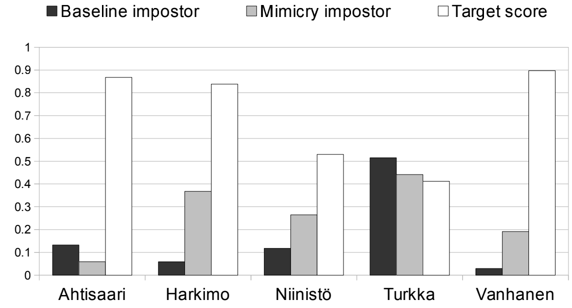


Figure 4: Comparison of "Same speaker" decision distribution per target speakers by the participants of the listening test.

For any automatic system to produce meaningful results, we need to provide a set up with sufficient amount of speech material. On the other hand, for a perceptual test, the amount of test material should be short enough for a listener to perform the voice comparison task without weariness. In this sense, we cannot make a direct and fair comparison between listeners and automatic systems performance. We can, however, evaluate the outcomes and make a qualitative analysis in their performances. Based on these, the 34 speech trials were analyzed by our automatic systems. First, all scores were turned into decisions by finding the optimum bias, with Bayes optimal decision threshold at the origin. We found the bias by logistic regression, with prior probability of seeing a target speaker equals to 0.5 and both, false alarm and miss, costs being to 1. We optimized the bias for the evaluation data directly, so that the results can be seen as the best possible decisions. Both the GMM-UBM and the i-vector systems performed similarly with equal number of errors. Similar to Fig. 3, we divided the decisions to target and nontarget (impostor, impostor impersonator) trials. In contrast to the listening test, we have only one system represented here, so the total number of trials is limited to 34. However, we can observe a similar trend as with the listening test that the impersonation increases the errors on detecting nontarget trials correctly. The total number of errors from the automatic system is 6, comparing this to listening test where on average the listening pool performed 8 errors. However, the best of our human listeners had only 4 errors.

5. Conclusions

We have assessed the accuracy of two speaker verification systems, GMM-UBM and i-vector with cosine scoring, with mimicked data in Finnish language. A perceptual test was also included to analyze if human speaker verification performance of non-expert listeners is affected by the presence of mimicry speech. Our results indicate that, the impersonator was able to increase his automatic speaker verification score only for one of the five target speakers. For the same target speaker, the listeners were however uncertain to decide if the voice of the impersonator was not the same speaker. Even though most of the listeners recognized the speaker correctly with the other target speakers, when the listeners were presented with the impersonator's own voice and a successful imitation of a target speaker (disguise), more than 20 of the 34 listeners judged the speaker to be different. This suggests that human listeners may be more likely to make recognition errors under disguise, rather than mimicry. Comparing human and automatic recognizer performance, we noticed that listeners made 8 errors on average, whereas automatic system made 6 errors on the same data set. In general, listeners performance is affected by the differences in speaking style. In perceptual tests, it is more evident that some speakers are easier to detect than others. On the other hand, automatic system has similar performance across speakers irrespective whether imitation is present or not. For future work, studying mimicry data with more than one impersonator subject would give a wider perspective of how these results extend to other impersonators.

6. Acknowledgements

This work was supported by Academy of Finland (proj. no. 253120 and 253000). The authors would like to thank Tero Ikävalko for his help with the perceptual test. We would like to express our deepest gratitude to all the listeners taking part to the research.

7. References

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

[2] G. Doddington, W. Liggett, A. Martin, M. Przybocki, and D. Reynolds, "Sheep, goats, lambs and wolves: A statistical analysis of speaker performance in the nist 1998 speaker recognition evaluation," in *ICSLP*, 1998, vol. 13.

[3] N. Yager and T. Dunstone, "The biometric menagerie," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 2, pp. 220 – 230, 2010.

[4] N. Evans, T. Kinnunen, and J. Yamagishi, "Spoofing and countermeasures for automatic speaker verification," in *Interspeech 2013*, Lyon, France, August 2013, pp. 925 – 929.

[5] T. Kinnunen, Z.-Z. Wu, K. A. Lee, F. Sedlak, E. S. Chng, and H. Li, "Vulnerability of speaker verification systems against voice conversion spoofing attacks: the case of telephone speech," in *ICASSP 2012*, Kyoto, Japan, March 2012, pp. 4401 – 4404.

[6] P.L. De Leon, M. Pucher, J. Yamagishi, I. Hernaez, and I. Saratxaga, "Evaluation of speaker verification security and detection of HMM-based synthetic speech," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 8, pp. 2280–2290, Oct 2012.

[7] J. Villalba and E. Lleida, "Detecting replay attacks from far-field recordings on speaker verification systems," *Biometrics and ID Management*, pp. 274–285, 2011.

[8] M. Farrús, M. Wagner, D. Erro, and F. J. Hernando, "Automatic speaker recognition as a measurement of voice imitation and conversion," *The Int. Journal of Speech, Language and the Law*, vol. 1, no. 17, pp. 119–142, 2010.

[9] P. Perrot, G. Aversano, and G. Chollet, "Voice disguise and automatic detection: Review and perspectives," in

Progress in Nonlinear Speech Processing, Lecture Notes in Computer Science, pp. 101–117. 2007.

[10] Elisabeth Zetterholm, "Detection of speaker characteristics using voice imitation," in *Speaker Classification II*, Lecture Notes in Computer Science, pp. 192–205. 2007.

[11] Leena Mary, Anish Babu K. K, Aju Joseph, and Gibin M. George, "Evaluation of mimicked speech using prosodic features," in *ICASSP 2013*, Vancouver, May 2013, pp. 7189 – 7193.

[12] Y.W. Lau, M. Wagner, and D. Tran, "Vulnerability of speaker verification to voice mimicking," in *Proc. Int. Symp on Intelligent Multimedia, Video & Speech Processing (ISIMP'2004)*, Hong Kong, October 2004, pp. 145–148.

[13] Y.W. Lau, D. Tran, and M. Wagner, "Testing voice mimicry with the YOHO speaker verification corpus," in *Knowledge-Based Intelligent Information and Engineering Systems (KES 2005)*, Melbourne, Australia, September 2005, pp. 15–21.

[14] J. Mariéthoz and S. Bengio, "Can a professional imitator fool a GMM-based speaker verification system?," Idiap-rr, IDIAP, 2005.

[15] R. González Hautamäki, T. Kinnunen, V. Hautamäki, T. Leino, and A-M. Laukkanen, "I-vector meet imitators: on vulnerability of speaker verification systems against voice mimicry," in *Interspeech 2013*, Lyon, France, August 2013.

[16] A. Schmidt-Nielsen and T. H. Crystal, "Speaker verification by human listeners: Experiments comparing human and machine performance using the NIST 1998 speaker evaluation data," *Digital Signal Processing*, vol. 10, no. 13, pp. 249 – 266, 2000.

[17] V. Hautamäki, T. Kinnunen, M. Nosratighods, Kong Aik Lee, Bin Ma, and Haizhou Li, "Approaching human listener accuracy with modern speaker verification," in *Interspeech 2010*, Makuhari, Japan, September 2010, pp. 1473–1476.

[18] C. Greenberg, A. Martin, G. Doddington, and J. Godfrey, "Including human expertise in speaker recognition systems: report on a pilot evaluation," in *ICASSP 2011*, Prague, Czech Republic, May 2011, p. 5896 5899.

[19] R. González Hautamäki, V. Hautamäki, P. Rajan, and T. Kinnunen, "Merging human and automatic system decisions to improve speaker recognition performance," in *Interspeech 2013*, Lyon, France, August 2013.

[20] Johanna Leskelä, "Changes in $f0$, formant frequencies and spectral slope in imitation," M.S. thesis, University of Tampere, 2011, in Finnish.

[21] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 33, no. 2, pp. 443–445, Apr 1985.

[22] Philipos C. Loizou, *Speech enhancement. Theory and practice.*, Taylor & Francis, USA, 2007.

[23] Yi Hu and P.C. Loizou, "Subjective comparison of speech enhancement algorithms," in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, May 2006, vol. 1, pp. I–I.

[24] D.A. Reynolds, T.F. Quatieri, and R.B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1, pp. 19–41, Jan 2000.

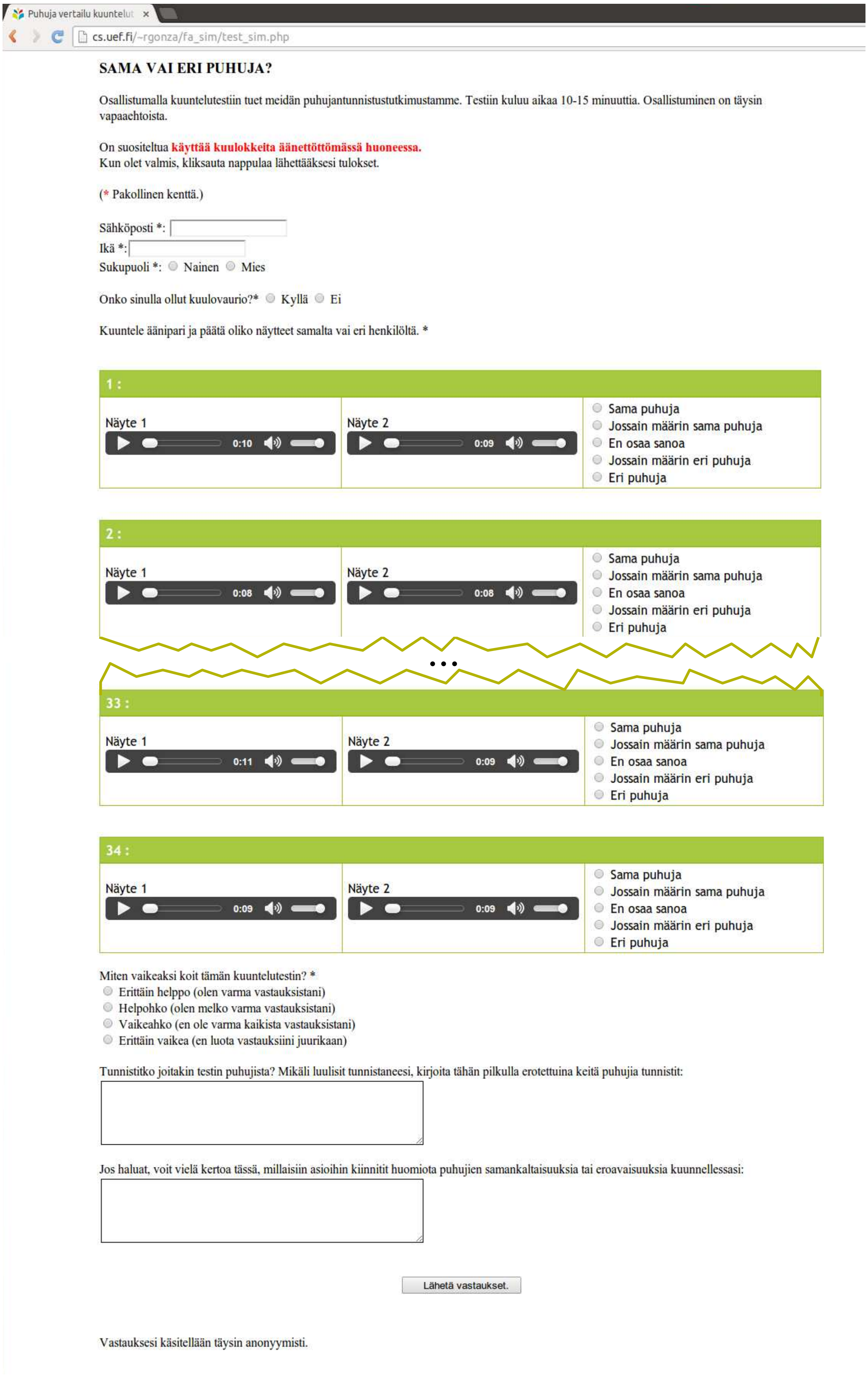
[25] N. Dehak, P.J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *ITASLP*, vol. 19, no. 4, pp. 788–798, May 2011.

[26] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *Interspeech 2011*, Florence, Italy, August 2011, pp. 249–252.

[27] N. Dehak, R. Dehak, J. Glass, D. Reynolds, and P. Kenny, "Cosine similarity scoring without score normalization techniques," in *Proc. Odyssey Speaker and Language Recognition Workshop*, 2010, pp. 71–75.

[28] Mike Brookes et al., "Voicebox: Speech processing toolbox for matlab," *Software, available [January 2014] from www. ee. ic. ac. uk/hp/staff/dmb/voicebox/voicebox. html*, 2006.

[29] C. Clopper and E. S. Pearson, "The use of confidence or fiducial limits illustrated in the case of the binomial," *Biometrika*, vol. 26, pp. 404–413, 1934.


Puhuja vertailu kuuntelut
cs.uef.fi/~rgonza/fa_sim/test_sim.php

SAMA VAI ERI PUHUJA?

Osallistumalla kuuntelutestiin tuet meidän puhujantunnistustutkimustamme. Testiin kuluu aikaa 10-15 minuuttia. Osallistuminen on täysin vapaaehtoista.

On suositeltua **käyttää kuulokkeita äänettömässä huoneessa.**
Kun olet valmis, kliksauta nappulaa lähettääksesi tulokset.

(* Pakollinen kenttä.)

Sähköposti *:
Ikä *:
Sukupuoli *: ○ Nainen ○ Mies

Onko sinulla ollut kuulovaurio?* ○ Kyllä ○ Ei

Kuuntele äänipari ja päätä oliko näytteet samalta vai eri henkilöltä. *

1 :

Näyte 1 0:10	Näyte 2 0:09	○ Sama puhuja ○ Jossain määrin sama puhuja ○ En osaa sanoa ○ Jossain määrin eri puhuja ○ Eri puhuja

2 :

Näyte 1 0:08	Näyte 2 0:08	○ Sama puhuja ○ Jossain määrin sama puhuja ○ En osaa sanoa ○ Jossain määrin eri puhuja ○ Eri puhuja

...

33 :

Näyte 1 0:11	Näyte 2 0:09	○ Sama puhuja ○ Jossain määrin sama puhuja ○ En osaa sanoa ○ Jossain määrin eri puhuja ○ Eri puhuja

34 :

Näyte 1 0:09	Näyte 2 0:09	○ Sama puhuja ○ Jossain määrin sama puhuja ○ En osaa sanoa ○ Jossain määrin eri puhuja ○ Eri puhuja

Miten vaikeaksi koit tämän kuuntelutestin? *

○ Erittäin helppo (olen varma vastauksistani)
○ Helpohko (olen melko varma vastauksistani)
○ Vaikeahko (en ole varma kaikista vastauksistani)
○ Erittäin vaikea (en luota vastauksiini juurikaan)

Tunnistitko joitakin testin puhujista? Mikäli luulisit tunnistaneesi, kirjoita tähän pilkulla erotettuina keitä puhujia tunnistit:

Jos haluat, voit vielä kertoa tässä, millaisiin asioihin kiinnitit huomiota puhujien samankaltaisuuksia tai eroavaisuuksia kuunnellessasi:

Vastauksesi käsitellään täysin anonyymisti.

Figure 5: Web-form for the listening test in Finnish. The listeners were instructed to listen and decide whether the speech samples belong to the same or different speaker. The listener's decision options were: a) Sama puhuja (Same speaker), b) Jossain määrin sama puhuja (Somewhat same speaker), c) En osaa sanoa (I cannot tell), d) Jossain määrin eri puhuja (Somewhat different speaker), e) Eri puhuja (Different speaker).