Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 72 (2015) 13–31

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Automatic versus human speaker verification: The case of voice mimicry

Rosa González Hautamäki [a,*], Tomi Kinnunen [a], Ville Hautamäki [a], Anne-Maria Laukkanen [b]

[a] Speech and Image Processing Unit, School of Computing, University of Eastern Finland, P.O. Box 111, FI-80101 Joensuu, Finland
[b] Speech and Voice Research Laboratory, School of Education, University of Tampere, FI-33014 Tampere, Finland

Received 22 July 2014; received in revised form 1 May 2015; accepted 3 May 2015
Available online 8 May 2015

Abstract

In this work, we compare the performance of three modern speaker verification systems and non-expert human listeners in the presence of voice mimicry. Our goal is to gain insights on how vulnerable speaker verification systems are to mimicry attack and compare it to the performance of human listeners. We study both traditional Gaussian mixture model-universal background model (GMM-UBM) and an i-vector based classifier with cosine scoring and probabilistic linear discriminant analysis (PLDA) scoring. For the studied material in Finnish language, the mimicry attack decreased lightly the equal error rate (EER) for GMM-UBM from 10.83 to 10.31, while for i-vector systems the EER increased from 6.80 to 13.76 and from 4.36 to 7.38. The performance of the human listening panel shows that imitated speech increases the difficulty of the speaker verification task. It is even more difficult to recognize a person who is intentionally concealing his or her identity. For Impersonator A, the average listener made 8 errors from 34 trials while the automatic systems had 6 errors in the same set. The average listener for Impersonator B made 7 errors from the 28 trials, while the automatic systems made 7 to 9 errors. A statistical analysis of the listener performance was also conducted. We found out a statistically significant association, with $p = 0.00019$ and $R^2 = 0.59$, between listener accuracy and self reported factors only when familiar voices were present in the test.

Keywords: Voice imitation; Speaker recognition; Mimicry attack; Listening test

1. Introduction

Speaker verification (Campbell, 1997; Reynolds, 2002) is the task of recognizing persons from their voices. The accuracy of speaker verification systems has steadily improved in the recent years due to advances in channel, noise and inter-session compensation techniques, making the technology available for tailored applications. Automatic speaker verification (ASV) technology is generally used under three scenarios. Firstly, *authentication* applications involve verifying the identity of a cooperative user who demands physical or logical access. Secondly, a *forensic* scenario involves comparing two speech samples to determine whether they originate from the same or different subject. Finally, *screening* and *indexing* applications involve searching a particular target speaker from large amounts of unlabeled data.

One of the increasing concerns in practical uses of ASV technology is vulnerability of the recognizers to intentional circumvention (Wu et al., 2015). In the first case, authentication, this refers to dedicated effort to manipulate one's speech so that an ASV system would misclassify the attacker's sample to originate from the target (client). There are

* Corresponding author.
E-mail addresses: rgonza@cs.uef.fi (R. González Hautamäki), tkinnu@cs.uef.fi (T. Kinnunen), villeh@cs.uef.fi (V. Hautamäki), anne-maria.laukkanen@uta.fi (A.-M. Laukkanen).

http://dx.doi.org/10.1016/j.specom.2015.05.002
0167-6393/

four main types of such *spoofing attacks* (Evans et al., 2013; Wu et al., 2015): mimicry, replay (Villalba and Lleida, 2011), speaker-adapted speech synthesis (De Leon et al., 2012) and voice conversion (Kinnunen et al., 2012). A common feature of all spoofing attacks is that the attacker uses non-zero effort to circumvent an ASV system, for instance, with financial motivation. This is different from the latter two use cases, forensics and screening, where the person in question may desire *not* to be detected as him/herself, and is therefore being considered to be non-cooperative. This type of circumvention, with an intention to provoke false rejections (misses), is known as *evasion* or *obfuscation* (Alegre et al., 2014). Similar to spoofing, evasion could be achieved by both technical means (for instance, by adding reverberation) and by disguising one's speech by, for instance by raising F0 or imitating a foreign accent (Zhang and Tan, 2008; Kajarekar et al., 2006). We should also point out that some speakers, without any voluntary effort to spoof or evade recognizers, tend to be confused with other users (Doddington et al., 1998; Yager and Dunstone, 2010). In this study, we focus on scenarios with intentional speech modification, namely, mimicry.

Speech mimicry is an interesting research phenomenon for several reasons. Firstly, most readers are likely to be familiar with talented impersonators (often stand-up comedians) in their mother tongue who are able to create funny, yet convincing-sounding impersonations of politicians or other public figures. We, as ASV researchers, are frequently asked whether such impersonators would be able to spoof ASV systems; a general belief is that human listeners can be fooled but ASV system accuracy is not affected by mimicry attacks. Table 1 summarizes some of the previous speaker recognition studies for mimicry data. Secondly, studying mimicry as a potential spoofing technique is also relevant. Detection of technical spoofing attacks, such as speech synthesis and voice conversion, can already to a certain extent be achieved by designing discriminative features known to differentiate synthetic and natural utterances (De Leon et al., 2012; Alegre et al., 2013; Wu et al., 2012). Clearly, such countermeasures are inapplicable for detection of impersonation produced by a real human being, making mimicry a challenging test case for spoofing countermeasure development, and particularly interesting for forensic and speech security applications. Thirdly, looking from the perspective of the impersonator, ASV technology could be used as an objective feedback tool to evaluate the similarity of one's impersonations against the intended target speaker. Such technology might help, for instance, actors to help practicing idiosyncratic speech of their characters.

The general challenges related to studies that involve mimicry include lack of a standard corpus for evaluation and technical mismatches. While there are standard and public corpora to benchmark speaker verification systems under zero-effort imposture, this is not the case regarding mimicry attacks; professional impersonators are not easily available to provide speech samples, and target speakers are often public figures whose samples are collected from public sources. Naturally, mismatches of audio recordings arise when professional impersonators' speech is collected in a studio environment and the target speakers' recordings from TV and radio interviews. An alternative way to analyze the mimicry attack is to include a perceptual test as a benchmark parallel to automatic system analysis. A human benchmark, compared to automatic systems in a zero-effort imposture setting, has been used in previous studies (Schmidt-Nielsen and Crystal, 2000; Hautamäki et al., 2010). In terms of human assisted speaker verification system (Greenberg et al., 2011; Hautamäki et al., 2010; González Hautamäki et al., 2013a), 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 the present study, we analyze voice mimicry attacks with audio material from the speakers described in Section 3, extending our preliminary analyses reported in González Hautamäki et al. (2013b, 2014). The current study extends these preliminary studies both regarding data and analyses. Firstly, we have collected fresh data from a new impersonator who mimics four additional target speakers presented in neither González Hautamäki et al. (2013b) nor in González Hautamäki et al. (2014). Secondly, a new human benchmark involving a large listening panel was also added.

Overall, our major contribution is an up-to-date analysis of mimicry attacks against state-of-the-art automatic speaker verification systems accompanied by a relatively large-scale human benchmark. Earlier studies on mimicry attacks (Table 1) have included classical spectral

Table 1
Some of the previous studies on mimicry data and the present study. Previous studies concentrate on acoustical analysis and Gaussian Mixture Model (GMM) with and without universal background model (UBM).

Study	Target language	Target speakers	Impersonators	Speaker verification
Lau et al. (2004)	English	6	2 naïve	GMM
Lau et al. (2005)	English	6	2 professional linguists, 4 naïve	GMM
Mariéthoz and Bengio (2005)	French	3	1 professional, 1 intermediate and 1 naïve	GMM-UBM
Zetterholm (2007)	Swedish	9	2 professional, 1 amateur	Auditory analysis by a panel
Farrús et al. (2010)	Spanish-Catalan	5	2 professional	Prosodic system
Panjwani and Prakash (2014)	English	53	3 professional and 13 naïve	GMM-UBM
This study	Finnish	8	2 professional	GMM-UBM, i-vector cosine and i-vector-PLDA, perceptual test

GMM-based speaker recognition systems as well as auditory analyses but, as far as we know, none within a single study and to the same extent as in the present study. In comparison to classical speaker verification systems, i-vector systems (and other recent methods) are comparatively more robust against intersession, noise and channel variations but it is unknown how vulnerable they are to mimicry. Lastly, our recognizer pool involves both a traditional Gaussian mixture model - universal background model (GMM-UBM) system (Reynolds et al., 2000) and a state-of-the-art i-vector system (Dehak et al., 2011) with two back-end scoring techniques: cosine scoring (Dehak et al., 2011) and probabilistic linear discriminant analysis (PLDA) (Prince and Elder, 2007). We make a detailed comparison of the performance of these systems in the recognition task with that of our human panel. The purpose is to study whether automatic systems are vulnerable to mimicry spoofing and also how a pool of human listeners performs the speaker verification task when impersonated speech is present. Different from González Hautamäki et al. (2013b, 2014), we further analyze the performance of automatic systems and human listeners to identify the factors that affect their accuracy in the verification task and whether there is a performance difference between mimicry and disguise conditions. We also explore the possible differences of performance between listeners and the two impersonators.

2. Imitation and speaker verification

Previous studies have evaluated mimicked speech and the success of the impersonator, either professional or not, in mimicking the targets' speaking characteristics related to spectral characteristics, prosody, dialects and speaking style. It has been reported that impersonators are often able to adapt especially the fundamental frequency ($F0$) and occasionally also the formant frequencies towards the target speakers (Farrús et al., 2010; Perrot et al., 2007; Zetterholm, 2007). An example of visual acoustic comparison of an imitator's natural voice and his impersonation, and the target speaker's voice is shown in Fig. 1. Farrús et al. (2010) and Mary et al. (2013) used automatic speaker recognition technology to objectively evaluate the success of voice imitation. Farrús et al. (2010) used a prosody-based speaker recognition system and found that fusion of 12 prosodic features increased the impersonator's efficacy. Mary et al. (2013), in turn, evaluated mimicked speech with prosodic features based on intonation, duration and energy. 9-dimensional feature vectors from the original target and mimicked speech were compared with the help of

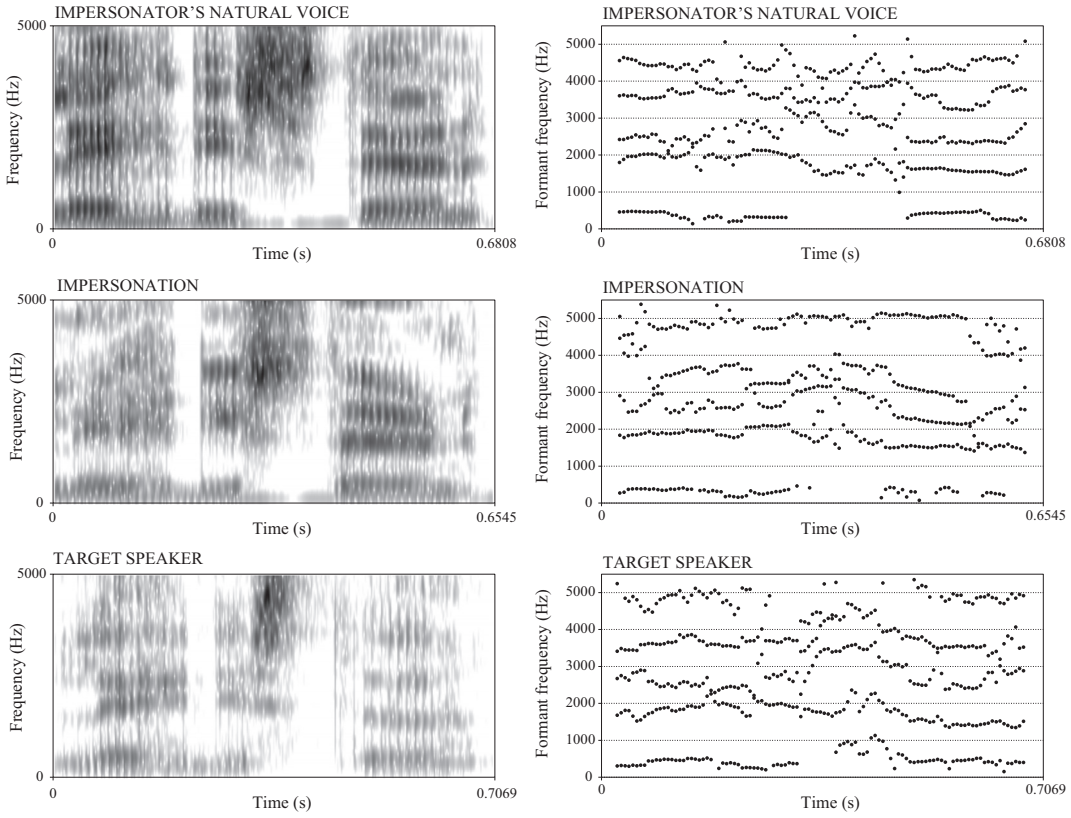


Fig. 1. An example of speech impersonation. Spectrograms (to the left) and formant tracks (F1–F5) of the impersonator's own voice (top), impersonation (middle) and the target speaker (bottom). Formants computed using Praat. The target speaker is the current president of Finland, Sauli Niinistö. Comparing the top and middle figures, the impersonator can modify his 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.

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, the results indicating an agreement between the automatic prosody system scores and the listeners' opinion. In other studies, foci have been on analyzing the vulnerability of speaker verification systems in the presence of voice mimicry. For example, in Lau et al. (2004, 2005) and Mariéthoz and Bengio (2005), the vulnerability of spectral 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 closer to the impersonator's voice, then the chances to spoof an automatic recognizer are increased.

Professional impersonators, especially in entertainment, mimic most striking characteristics related to prosody, voice quality, dialect and speaking style of a target speaker. Different techniques used by professional imitators were studied in Farrús et al. (2010) and Zetterholm (2007). The studies found that the impersonators are able to adapt their fundamental frequency and the formant frequencies towards the target voices. This reveals a potential vulnerability of the automatic speaker recognition systems that mainly utilize spectral features. Fig. 2 shows an example of the *fundamental frequency* $F0$ contour. We observe a difference between the $F0$ of the impersonator's natural voice and his impersonation of the target speaker.

Farrús et al. (2010) attempted to quantify how much a speaker is able to approximate others' voices, by focusing on a selection of prosodic and acoustic features from two professional impersonators that imitated Spanish politicians. The authors used an automatic speaker recognition system based on both prosodic and acoustic features. The prosodic parameters included duration of words, means and ranges of $F0$, as well as jitter and shimmer measurements. In their imitation experiment, the identification error rate increased when score level fusion of the prosodic features was performed.

Lau et al. (2004) used the YOHO corpus in their experiments in an interesting way. The authors used recordings from two naïve impersonators (native Chinese, living in Australia more than 7 years) with no experience in mimicry. Having recorded the natural voices of the impersonators, the authors used a spectral GMM system to select 3 different speakers from YOHO: the *closest, intermediate* and *furthest* speaker. Then the impersonators read all of the 40 training utterances from the three speakers, listened to the target speaker's samples and tried to imitate them. There were four recording sessions for both impersonators because the authors wanted to find out whether the imitators become better with more training. It was concluded that, indeed, the verification errors increased as a function of training times. An interesting observation was that both of these "naïve" impersonators were falsely accepted by the system as the target speakers they were imitating. However, this was true *only* for the closest speaker. Neither imitator was able to be accepted as the intermediate or the furthest

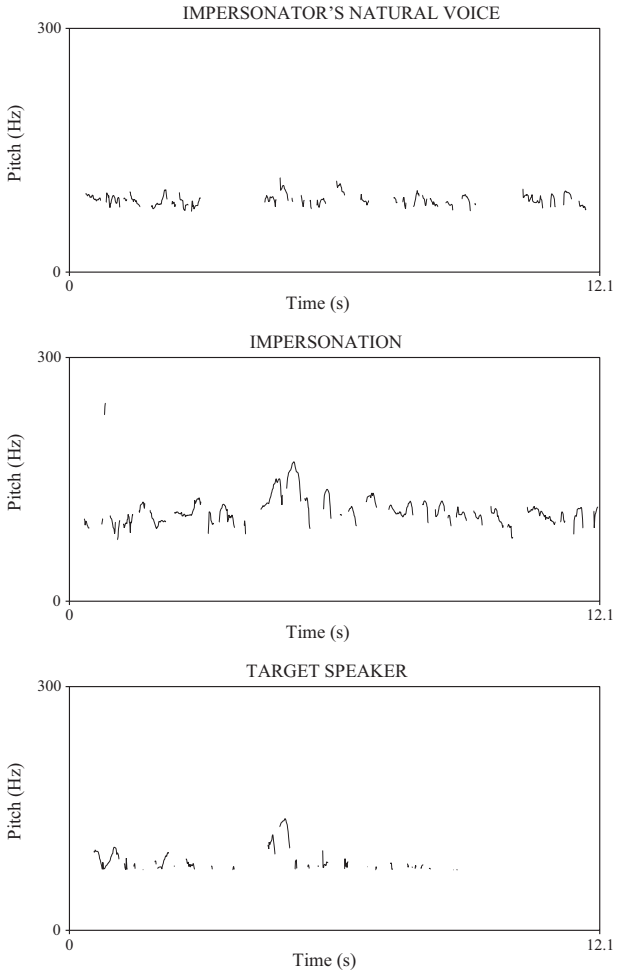


Fig. 2. An example of speech impersonation showing the fundamental frequency ($F0$) contour of the impersonator's own voice (top), impersonation (middle) and the target speaker (bottom) with the same speech content. $F0$ computed using Praat.

speaker. This seems to suggest that the speakers whose vowel space is close to that of the imitator tend to be easily imitated, likely due to similar articulatory constraints. If the articulators are very different, it will be difficult or impossible to modify the speech sufficiently towards the target.

In another study by Lau et al. (2005), the authors tested two groups of imitators: professional and non-professional. The professional group consisted of two linguists (one female and one male). Four other naïve imitators (two Chinese male and female, and two Australian, male and female) formed the non-professional group. Similar to Lau et al. (2004), some target speakers were selected from YOHO, this time only the most similar speakers in the sense of GMM likelihood score. Three recording sessions for each imitator were taken. For the first professional (female linguist) the *false acceptance rate* (FAR) increased from practically 0% to 60%; for the male linguist, (only) to 10% FAR, using the same threshold setting. For the amateur female imitators (1 Chinese, 1 Australian), the numbers were around 20% and 30%

FAR. The Australian male showed similar results to the male linguist. However, the Chinese male achieved as high as 60% FAR. The study suggests that, independently of whether the imitators were professional or not, the error rates were increased by imitation, and that knowledge in Linguistics may not help in imitation, at least in the case when the voices of the target speakers are similar to the impersonator's natural voice.

In a recent study (Panjwani and Prakash, 2014), the authors also used both professional and non-professional imitators. The novelty in their approach was to use crowdsourcing, via Amazon Mechanical Turk[1], to obtain the non-professional imitators. Out of 176 non-professional imitators reached, only 6 were deemed successful, and 3 professional imitators were selected from the pool of 25 potential candidates. Each impersonator was presented with a one closest matching target speaker who he was asked to impersonate. This procedure is in contrast to our study, where 2 professional impersonators were asked to impersonate multiple targets, thus we were able to obtain knowledge of the impersonators' skill variance in terms of target speakers. In addition, Panjwani and Prakash (2014) used GMM-UBM for speaker comparison and their study did not include perceptual tests. However, they were able to find non-professional imitators that succeeded in increasing their average score significantly, but not exceeding the target self score. This suggests that imitation ability is target specific, which is also in line with the findings of the present work.

Zetterholm et al. (2004) presents a comparison between human perception and a speaker verification score to measure how close to the target the impersonator's voice could get. A professional impersonator mimicking two target speakers is the basis for the study. HMM-based speaker verification system's score was used to measure the success of the impersonator to get closer to the target speaker's voice. The authors found that there were perceptual differences between the impersonator's natural voice and the target speakers' and between the impersonations. The perceptual experiment included 22 listeners who were asked to compare two samples and decide which one was closest to a reference sample. The authors concluded that there was no correlation between the perceptual test results and the score provided by the automatic system. The authors also reported that the perceptual test was demanding for the listeners who could do better in a standard verification test which did not include mimicry.

3. Material

The main challenge in studies involving mimicry is the scarcity of data. The existing data is created for a specific study, which is not publicly available and cannot be considered as a standard evaluation corpus. Not only is data collection 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 need to be collected from radio interviews and TV programs. As a consequence, there are necessarily technical mismatches since the impersonator's voice samples have been recorded in a studio environment.

3.1. Target speakers

A speech database containing the voice of eight well-known Finnish public figures is used for this study (see Table 2). The speech data for the first five target speakers and Impersonator A was collected by Leskelä (2011), and it was used for a preliminary mimicry attack analysis in González Hautamäki et al. (2013b, 2014). For the present study, speech samples from another impersonator (Imp B) and three additional target speakers were collected.

3.2. Technical aspects

To study a scenario in which mimicry attack could threaten a speaker verification system, we focus on data with 8 kHz sampling rate. Since the majority of ASV systems studies use mainly data with sampling rate of 8 kHz or 16 kHz, all the recorded audio samples for this study were down-sampled to 8 kHz. The audio segments were preprocessed to compensate for technical mismatches induced by channel differences and environmental noise. A speech enhancement algorithm based on the logarithmic estimation of the complex spectrum of the signal, also known as logMMSE (Ephraim and Malah, 1985), was applied to all speech segments. The logMMSE estimator reduces the residual noise without greatly affecting the speech signal. For this study, we used the Matlab-based implementation presented in Loizou (2007) and provided on a DVD that accompanies the book. Other experiments were conducted using algorithms that utilize *Wiener* filter but according to informal subjective comparison, they introduced noticeable level of distortion in some of the frequency bands. Therefore, these approaches were not included in the present study. As noted in Hu and Loizou (2006), the logMMSE speech enhancement method does not significantly degrade either the sound quality or intelligibility, which were the only requirements for our listening test.

3.3. Material for automatic speaker verification tests

In the present study, the training material used for the verification systems consisted of a maximum of 5 min of active speech from each of the target speakers. The test segments, for the ASV system experiments, were set to 20 s in duration, chunked from the original long recordings. The professional impersonators' natural voices (no mimicry) were recorded when they read segments from interviews

[1] http://www.mturk.com/mturk/.

Table 2
Impersonator and target speakers. YLE = Yleisradio, Finnish national public broadcasting company. The braces indicate the selected target speakers by the impersonators.

	TARGET SPEAKERS					
	Label	**Name**	**Position**	**Source material**	**Duration (mins)** Train	Test
Imp. A	TS2	Hjallis Harkimo	Politician, businessman	YLE Radio	5:00	2:02
	TS3	Sauli Niinistö	Current president of Finland	YLE Radio	5:01	3:00
	TS4	Jouko Turkka	Theatrical director	YLE Arch.	5:00	2:05
	TS5	Matti Vanhanen	Former prime minister	YLE Radio	5:00	1:50
Imp. B	TS1	Martti Ahtisaari	Former president, UN mediator	YLE Radio	5:01	2:20
	TS6	Andy McCoy	Rock musician, singer	YLE Radio, Arch	5:20	4:20
	TS7	Sakari Kuosmanen	Singer, actor	YLE Radio, TV	5:04	4:16
	TS8	Pertti "Spede" Pasanen	Actor, TV presenter, film producer	YLE Radio, TV	5:10	4:21
	IMPERSONATORS					
	Imp A	Impersonator A	Impersonator, singer	Studio recording	10:60 (own) 5:52 (imp.)	
	Imp B	Impersonator B	Impersonator, musician and stand-up comedian	Studio recording	11:00 (own) 11:16 (imp.)	

of the target speakers. Additionally, both impersonators recorded two mimicry samples per each of their target speakers. The sample pairs or trials contain test segments of the target speakers, to be called *genuine* trials, and the impersonator's samples as the *impostor* trials. For the *baseline case*, the impostor trials consisted of the impersonator's natural voice. In the *mimicry attack*, the impersonator's samples mimicking the target speakers were used as the impostor trials. In this way, the effects for the system performance are compared between the cases when the data includes or does not include mimicry.

One of the relevant variables to study is text-dependency. In its usual definition (Hébert, 2008; Furui, 1997), text-dependent recognition implies that the training and test utterances are matched in their lexical content, in other words, the test sample is a subset of the training utterances. In our study, this cannot be fully studied as the target speaker material originates from free-worded interviews without prompted text. However, the impostor test samples, either naturally spoken or impersonated ones, are matched with the target test utterances. To this end, we defined the following scenarios:

Text independent: The speech content for the training and test utterances do not match in their lexical content. Also the test segments lexical content do not match between impersonator's samples and target ones.

Same text: The training and test utterances do not match in content, but the test segments for each target speaker, genuine and impostor trials, match in their lexical content.

The trial lists for the mimicry case contain segments as defined in Table 3.

It is worth mentioning that we intended to have both impersonators uttered the same text for TS1. Impersonator B requested a different recording for this speaker and for that reason the lexical content for TS1 does not match across the impersonators' samples in this test.

3.4. Material for listening tests

For the listening tests, 68 speech samples for Impersonator A and 56 samples for Impersonator B were selected out of the material that consisted of 20 s duration samples. The samples were further cut into 10 s samples and in addition to speech enhancement described in

Table 3
Test trials for mimicry attacks with 20 s samples.

Test	Trials	Impersonator A	Impersonator B
Text independent	Genuine	50	48
	Mimicry Impostor	31	48
Same text	Genuine	23	24
	Mimicry Impostor	38	48

Table 4
Distribution of the 34 trials per speaker for Listening test 1. Test samples of 10 s in duration.

Speaker	Number of trials		
	Genuine	Baseline impostor	Mimicry impostor
TS1	2	2	2
TS2	2	2	2
TS3	2	2	2
TS4	2	2	2
TS5	2	2	2
Imp A	4	–	–

Table 5
Distribution of the 28 trials per speaker for Listening test 2. Test samples of 10 s in duration.

Speaker	Number of trials		
	Genuine	Baseline impostor	Mimicry impostor
TS1	2	2	2
TS6	2	2	2
TS7	2	2	2
TS8	2	2	2
Imp B	4	–	–

Section 3, all the speech samples were further normalized to have the same active speech level. We estimated the active speech level using the `activlev` function provided in the VOICEBOX speech processing toolbox (Brookes et al., 2006). This function implements ITU-T P.56 recommendation of objective measurement of active speech level in a given audio file.[2]

The length of the samples for the listening test was considered. Trials including a pair of 20 s segments would have taken at least 40 s to answer for a listener. For this reason, we also considered the amount of material to use during the test, since listeners are prone to fatigue or boredom if the listening task is long or tedious. It is mentioned in Bech and Zacharov (2007) that perceptual evaluations requirements consider sessions of 20 min a good idea, and 30–40 min sessions acceptable. We decided to include 34 trials for the Listening test 1 (Imp A) and 28 trials for the Listening test 2 (Imp B) with a test duration estimated at 30 min for each listener.

Tables 4 and 5 show the distribution of the number of trials per speaker, 2 genuine trials per impersonator can be classified as disguise cases.

These trial lists were also evaluated by our speaker verification systems for comparison, even though it is expected for the systems to perform poorly with short utterances of 10 s in contrast to the 20 s data.

4. Automatic speaker verification systems and a human panel

4.1. Automatic speaker verification systems

In the present study, two widely applied automatic speaker verification approaches are considered. Both utilize a 54-dimensional Mel-frequency cepstral coefficient (MFCC) as feature extractor (see Sys11 in (Saeidi et al., 2013)). The first system is based on a classical *Gaussian mixture model with universal background model* (GMM-UBM) (Reynolds et al., 2000). The other one is a state-of-the-art *i-vector* system (Dehak et al., 2011). The scoring of the extracted i-vectors is performed with *cosine scoring* (Dehak et al., 2011) or *probabilistic linear discriminant analysis* (PLDA) (Prince and Elder, 2007). For completeness, we will briefly describe each system in the following.

4.1.1. GMM-UBM system

The basis of these ASV systems is the so-called *universal background model* (UBM), which is a GMM estimated from a large speech corpus (Reynolds et al., 2000). The intention of the UBM is to model the general feature space distribution of speech. In the GMM-UBM system, the target speaker models are obtained via *maximum a posteriori* (MAP) adaptation from the UBM. The background model also acts as an impostor hypothesis model, i.e. "not the target speaker". The verification score is then the log-likelihood ratio of the test utterance generated by the target model and that generated by the UBM. All the speaker verification systems described here use the same UBM of 512 Gaussians, estimated from Fisher, Switchboard and the 2004, 2005, and 2006 NIST *Speaker Recognition Evaluation* (SRE) corpora.

4.1.2. i-vector with cosine scoring

The i-vector approach, in contrast to GMM-UBM, models each utterance as a low-dimensional i-vector. It stems from the idea that the MAP adaptation, when performed to the mean vectors only, results in a supervector of concatenated means. One can compare supervectors using a multitude of pattern recognition methods. Specifically, utterance-dependent GMM mean supervector ($\mathbf{s}$) is defined in Dehak et al. (2011),

$$\mathbf{s} = \mathbf{m} + \mathbf{T}\mathbf{w}, \tag{1}$$

where $\mathbf{m}$ is the UBM supervector, $\mathbf{T}$ is a low-rank rectangular matrix and $\mathbf{w}$ is a latent variable with a prior distribution of $\mathcal{N}(\mathbf{0}, \mathbf{I})$. Residual in (1) is assumed to be distributed as $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_{\text{ivec}})$, where $\mathbf{\Sigma}_{\text{ivec}}$ is typically copied directly from the UBM. The posterior distribution of $\mathbf{w}$ is also a Gaussian, resulting in a closed-form expression for the posterior mean, which is the extracted i-vector. The $\mathbf{T}$-matrix is estimated using an expectation–maximization (EM) algorithm from an external data set, in these experiments from NIST SRE 2004, 2005, and 2006, Fisher and Switchboard corpora. The i-vector dimensionality (the rank of $\mathbf{T}$-matrix) was set to 400.

Prior to scoring, the i-vectors are first post-processed by *radial Gaussianization* (Garcia-Romero and Espy-Wilson,

[2] http://www.itu.int/rec/T-REC-P.56/e.

2011), with the intention that the i-vectors better obey the Gaussian distribution assumptions (Kenny, 2010). For the *i-vector with cosine scoring*, given two utterances represented by their corresponding i-vectors, the angle between the two vectors, or *cosine similarity*, is used as a measure of similarity (Dehak et al., 2010).

4.1.3. Probabilistic linear discriminant analysis

Probabilistic linear discriminant analysis (PLDA) (Prince and Elder, 2007) is a generative model of i-vectors. It allows scoring and session compensation in the i-vector space. Radial Gaussianization is also used as pre-processing for the PLDA scoring. Formally, the PLDA model for speaker j and recording i is:

$$\mathbf{w}_{i,j} = \boldsymbol{\mu} + \mathbf{F}\mathbf{h}_j + \mathbf{G}\mathbf{c}_i + \boldsymbol{\epsilon}_{i,j}, \quad (2)$$

where $\boldsymbol{\epsilon}_{i,j}$ is distributed as $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ and $\boldsymbol{\Sigma}$ is a diagonal matrix. The latent variable vector $\mathbf{h}$ is called *speaker factor* in the speaker recognition terminology. Correspondingly, the latent variable $\mathbf{c}_i$ is called *channel factor*, which describes the position of the recording i in the noise subspace. The hyper-parameters $\boldsymbol{\mu}, \mathbf{F}, \mathbf{G}$ and $\boldsymbol{\Sigma}$ are estimated using an EM algorithm from the development set which, in our case, is the same set as in the **T**-matrix estimation. It is noteworthy that PLDA hyper-parameter estimation needs speaker labels for each utterance, whereas the UBM and **T**-matrix do not. In our experiments, the rank of **F** was set to 200 and the rank of **G** to 0. The $\boldsymbol{\Sigma}$ was then allowed to be of full rank. This model is known as the *simplified PLDA* (Kenny, 2010).

The scoring in PLDA is the log-likelihood of the two hypotheses that either the i-vectors from the test and train share the same $\mathbf{h}$ or that they are generated by two different latent variable vectors ($\mathbf{h}_1$ and $\mathbf{h}_2$).

4.2. Listening test

The participants in the listening panels were native speakers of Finnish (for description, see Table 6). They participated in a web-based listening test, to compare 34 pairs of speech samples for Impersonator A and 28 pairs for Impersonator B (See an example of the web-form screen in Appendix A). The listeners are considered naïve since no formal training was required to participate in this experiment. In fact, much effort was paid in recruiting listeners who were neither familiar with speech science or technology, nor with our research topics in general.

The listening tests were conducted in two cities, Joensuu and Tampere, where the two collaborating groups in this study are located. The listening tests were scheduled in a silent office environment of approximately 15 square meters area. In Joensuu, a desktop computer with integrated sound card was used with Sennheiser HD 570 headphones, in Tampere, two laptop computers with audio interface devices, Motu Ultralite mk3 and Roland Quad capture, and AKG and Sony studio headphones were used. The data for the listening tests was collected in different times, the time difference between the two tests was almost 4 months.

Table 6
Listening panel participants.

Listening test	Female	Male	Total	Age range
Impersonator A	19	15	34	20–65 years
Impersonator B	16	16	32	18–46 years

The type of listening trials comprising the listening tests were described in Tables 4 and 5. The only instruction given to the listeners was to listen to each pair and compare the speakers 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 make their decision out of five given options: *Same speaker, somewhat same speaker, I cannot tell, somewhat different speaker, different speaker*. After completing the test, the participants were asked to voluntarily report the speakers that they had recognized in the samples and also to describe the cues they had used to differentiate the speakers in the sample pairs.

It is worth mentioning that, during the preliminary evaluation 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, after analyzing the test results, it became clear that for uninformed listeners the task was not as easy as was expected. Most of the participants reported a considerable amount of effort to compare the speakers' voices and, as we will see below, made several errors.

5. Results

5.1. Automatic systems

To analyze the effect of imitation spoofing, we present the performance of the verification systems in terms of *equal error rate* (EER) which corresponds to the operating point with equal miss and false alarm rates. We calculate the EER using the implementation for ROC convex hull (ROCCH-EER) method from *Bosaris Toolkit*[3].

5.1.1. Text independent test

In Table 7, it can be seen that, unlike the other two systems, the performance of the GMM-UBM system shows inconsistent results in the baseline and the mimicry attack cases for the data set of Impersonator A. Regarding the i-vector Cosine and PLDA systems, however, there is a notable increase in EER when mimicry is present. For the data set of Impersonator B, the systems demonstrated

[3] http://sites.google.com/site/bosaristoolkit/home.

Table 7
Effect of mimicry attack in terms of *equal error rate* (EER %) for text independent test samples of 20 s duration.

Material	Test	GMM-UBM	i-vector Cosine	i-vector PLDA
Impersonator A	Baseline	8.17	5.96	1.86
	Mimicry attack	6.83	8.96	2.47
Impersonator B	Baseline	6.58	5.33	6.15
	Mimicry attack	12.21	17.44	7.69
Pooled	Baseline	10.83	6.80	4.36
	Mimicry attack	10.31	13.76	7.38

similar performance, showing an increase in EER in the presence of mimicry. Pooling scores from both test sets reveals a consistent increase in EER for the i-vector systems in mimicry attack compared to the baseline case.

The *detection error trade-off* (DET) curve (Martin et al., 1997) is a standard tool for assessing the accuracy of speaker verification systems beyond a single operating point. The usefulness of DET curves is limited for this study due to sparse data. A more insightful analysis can be obtained by studying the response of the recognition systems for individual target speakers; these are shown in Figs. 3 and 4. The graphs display the recognizer's average scores per target speakers before (baseline) and after the attack (mimicry). The *standard errors of the mean* (SEMs) are also shown, with the confidence range limits set to 95%.

In Fig. 3, comparing the heights of the baseline impostor bars relative to the genuine bars – a measure of the similarity of our imitator's *natural* voice against a particular target – TS1 and TS3 appear to be the most similar to Impersonator A's voice, while TS2 and TS5 have lower scores. The same observation can be made for the graphs of all three systems. Also, comparing the height of the bars for the case of mimicry impostor, we observe that there is an increase in the scores towards the target voice for Impersonator A, in particular TS1 and TS2, for the i-vector systems scores. These increments, however, are not significant when we take into account the overlap in the confidence intervals for both speakers. Regarding Impersonator B in Fig. 4, his natural voice seems to be more similar to that of TS6. TS7 and TS8 mimicry impostor scores increased towards the target speakers scores for i-vector Cosine system.

Previous studies (Zetterholm, 2007; Lau et al., 2005) have suggested that imitation attacks against "similar" target speakers might be easier than against speakers with very different voice quality. Fig. 3, however, indicates that the mimicry scores against the most similar target, TS3, was *lower*, while the relative increase was largest for target speaker TS2 in the i-vector systems. Similarly in Fig. 4, the mimicry scores for TS6 are systematically decreased over the baseline impostor case.

The authors in Lau et al. (2004) used a speaker verification system which had high-quality clean input signal and controlled text passages to select the most similar and

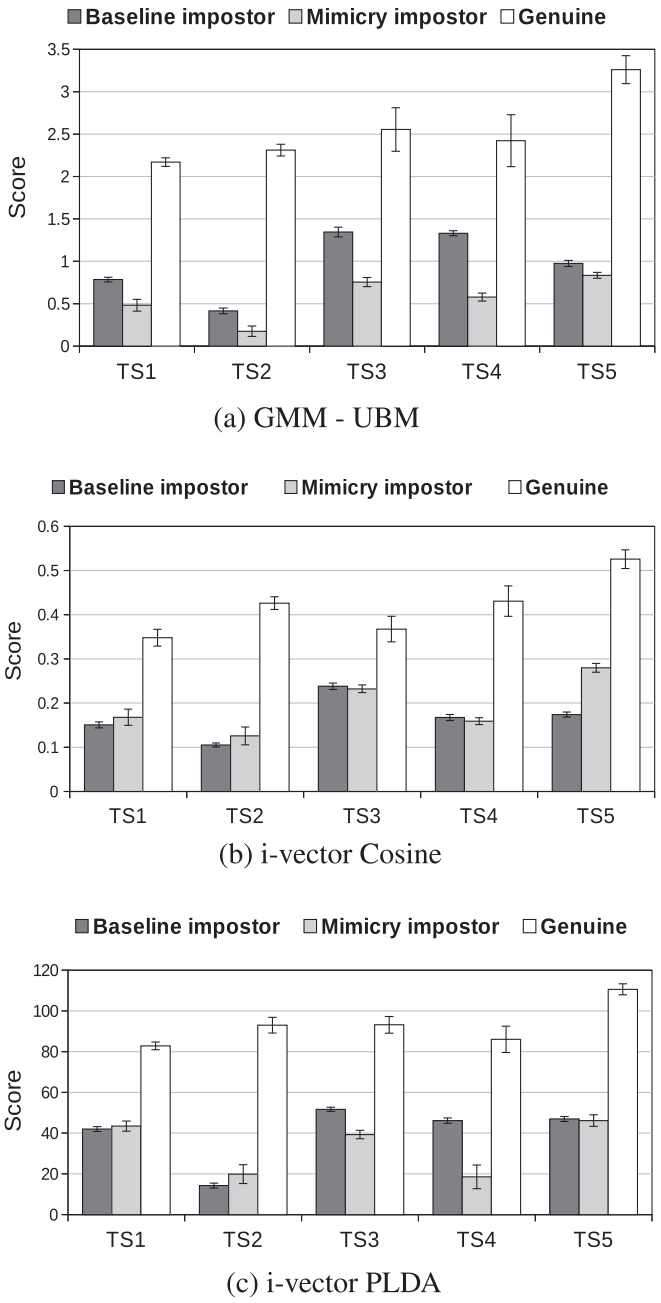
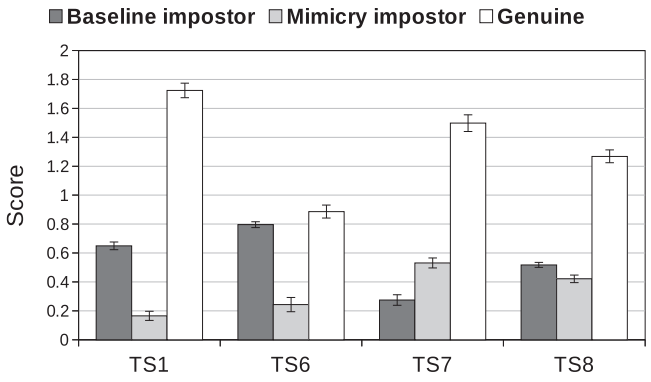


Fig. 3. Score distribution comparison per target speaker for Impersonator A. Test samples were 20 s in duration and non-matching speech content. The bars also show the standard error of the mean with 95% confidence. Baseline impostor refers to impersonator's own voice score, and mimicry impostor refers to impersonation score.

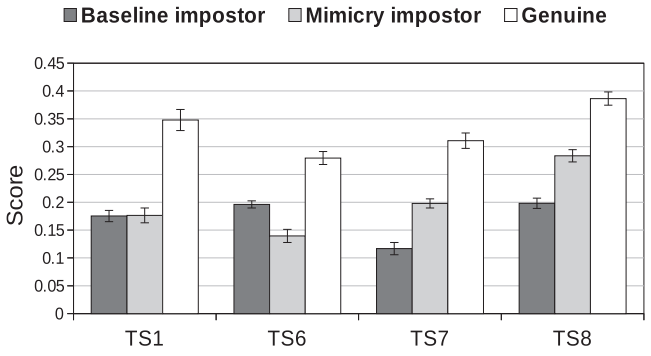
dissimilar speakers for the impersonator. Our study deals with a scenario with free-text inputs and includes recordings of varying quality. To be able to ideally focus on the success of the impersonation, all speech samples should be recorded under same conditions. However, as described above, this was not practical in our case.

5.1.2. Same text test

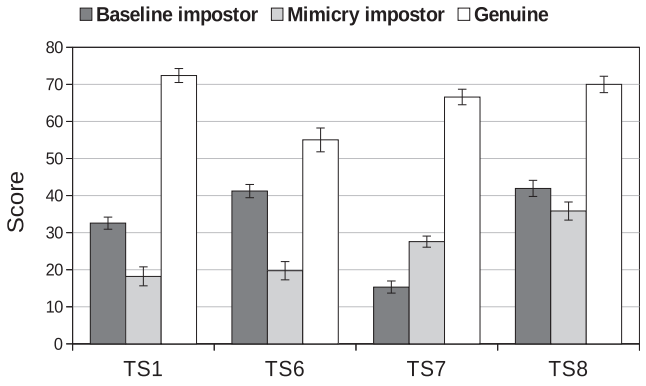
For the case in which the speech content of test samples match for genuine and impostor trials, we observe in



(a) GMM - UBM



(b) i-vector Cosine



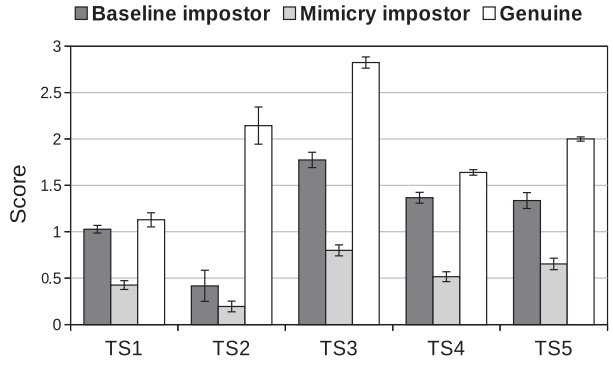
(c) i-vector PLDA

Fig. 4. Score distribution comparison per target speaker for Impersonator B. Test samples were 20 s in duration and non-matching speech content. The bars show the standard error of the mean with 95% confidence.

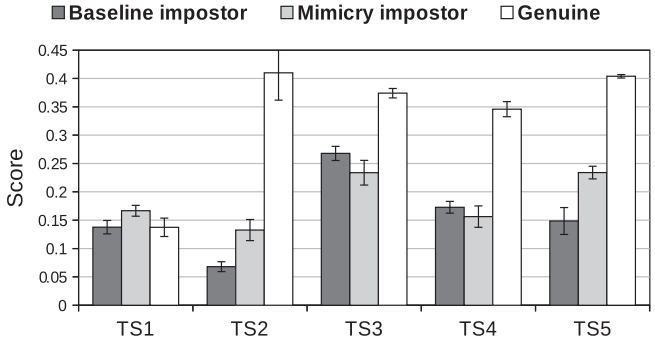
Table 8
Effect of mimicry attack in terms of *equal error rate* (EER %) for Same text test of 20 s samples.

Material	Test	GMM-UBM	i-vector Cosine	i-vector PLDA
Impersonator A	Baseline	22.22	16.48	13.04
	Mimicry attack	16.08	14.93	14.25
Impersonator B	Baseline	2.08	9.03	1.29
	Mimicry attack	7.58	17.36	2.11
Pooled	Baseline	10.38	13.73	7.16
	Mimicry attack	11.21	15.76	7.74

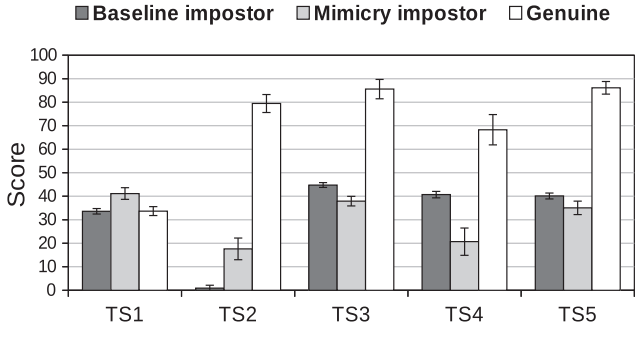
Table 8 an increase in EER for all the systems when the data from the impersonators is pooled. However, for Impersonator A, GMM-UBM and i-vector cosine systems show inconsistent results between the baseline and the mimicry attack test. Only the i-vector PLDA shows a slight increase whereas the EERs of the two other systems decreased. For Impersonator B, all the systems experienced systematic degradation. In Fig. 5, the score distributions for baseline impostor and genuine show target speakers TS1, TS3 and TS4 as the most similar to Impersonator A, specifically TS1, which we observe for all the three



(a) GMM - UBM



(b) i-vector Cosine



(c) i-vector PLDA

Fig. 5. Score distribution comparison per target speaker for Impersonator A. Test sample duration 20 s, and matching speech content. The bars also show the standard error of the mean with 95% confidence. Baseline impostor refers to impersonator's own voice score, whereas mimicry impostor refers to impersonation score.

systems. For the mimicry impostor, the score distribution comparison shows an increase towards the target voices of TS1 and TS2 for the i-vector systems. It can be seen in Fig. 6 for Impersonator B and i-vector Cosine and GMM-UBM systems, that TS6 still is the most similar voice to the impersonator's own voice but the mimicry impostor score bar shows a decrease in the scores away from the target score. In the mimicry impostor distributions, we observe an increase of scores towards the targets TS7 and TS8 in the i-vector systems, however, this increase is not significant for TS8.

Comparing the results between the text-independent (Table 7 and Figs. 3 and 4) and Same text (Table 8 and Figs. 5 and 6) tests, we observe similar average trends in the systems' performance for the pooled data, while the score distribution differences for individual speakers are less conclusive. For the score distribution of both tests, in the case of Impersonator A, TS1 does not show a significant difference between the average scores for the 3 systems, and there is an increase in the mimicry impostor score for the i-vector systems when the speech content matches. For Impersonator B, the score distributions for both data sets do not show differences.

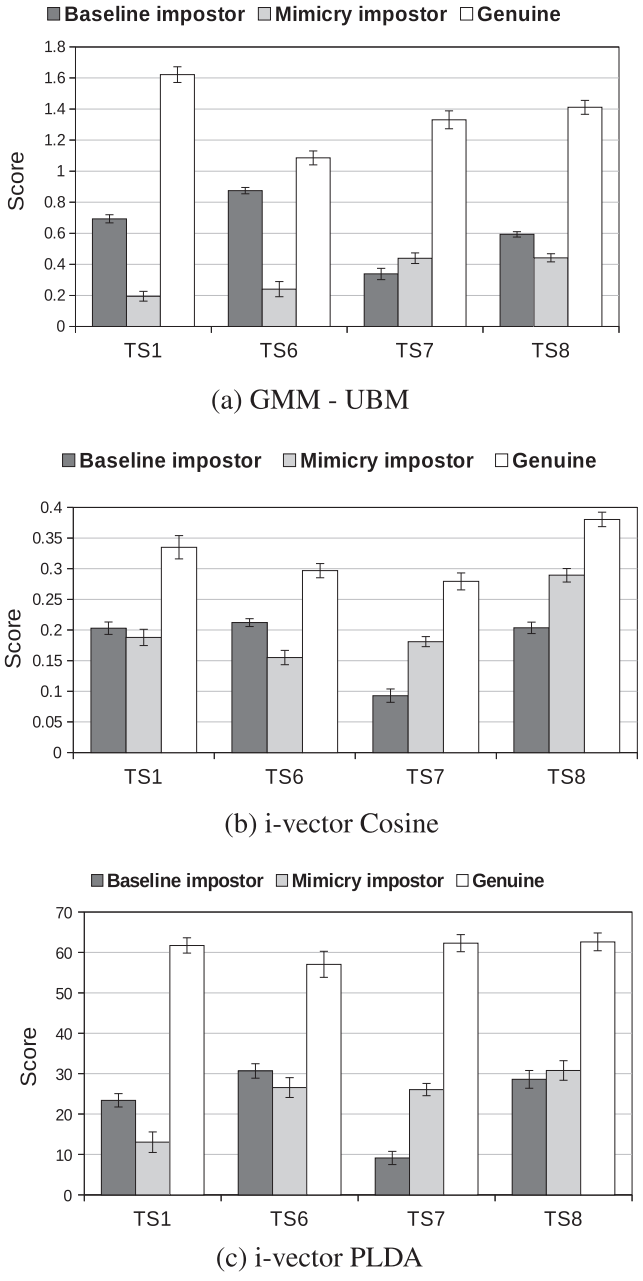


Fig. 6. Score distribution comparison per target speaker for Impersonator B. 20 s test samples duration and matching speech content. The bars also show the standard error of the mean with 95% confidence.

5.2. Listening test

To analyze the difficulty of the mimicry attack scenario for human listeners, we carried out two listening tests. In Listening test 1, we had 34 speech sample pairs including five target speakers (TS1 to TS5) and Impersonator A. In Listening test 2, we selected 28 speech sample pairs representing 4 target speakers (TS1, TS6, TS7 and TS8), and Impersonator B. Table 9 shows the EERs for the listening test trials, both separately and pooled. For Listening test 1 (Impersonator A trials), i-vector PLDA system had the lowest EER (14.29%), followed by the GMM-UBM system (15.00%).

For Listening test 2 (Impersonator B trials), i-vector Cosine showed the lowest and i-vector PLDA the higest EERs. For the listening panel performance, we considered the listener decisions corresponding to the following definition: 1: same speaker, 2: somewhat same speaker, 3: I cannot tell, 4: somewhat different speaker, 5: different speaker. These listener votes, or integers from 1 to 5, were interpreted as a score and summed up to give a verification result for the whole panel, in contrast to a single listener. We used the same methodology as in Hautamäki et al. (2010). The listening panel outperformed the automatic systems by a wide margin for both listening tests.

We observe that the automatic system EERs for the listening tests samples are considerable higher that those in Tables 7 and 8. The low performance of all automatic systems is likely caused by the very short duration speech segments used in the listening tests in contrast to the 20 s automatic system experiments. Our result is in line with that of recent experiments with 10 s train and test condition in NIST SRE 2008, where EERs between 19% and 21% were reported (Kanagasundaram et al., 2013). We also notice that the test with Impersonator A is easier than that

Table 9
Performance of the automatic systems and listener panel in terms of equal error rate EER (%) for the listening tests with 10 s in duration. Impersonator A with sample pairs of Listening test 1, and Impersonator B with sample pairs of Listening test 2.

	GMM-UBM	i-vector Cosine	i-vector PLDA	Listening panel
Impersonator A (34 trials)	15.00	21.43	14.29	9.09
Impersonator B (28 trials)	27.27	25.00	31.25	16.67
Pooled	20.20	22.22	23.08	12.50

with Impersonator B for both the human panel and the automatic systems.

The small number of trials considered for the listening tests allows a thorough trial-by-trial comparison. The grid in Table 10 shows the listeners' decisions for each of the trials in the Listening test 1. The errors per trial are shown in terms of false alarms and misses. We identify three main types of trials in Listening test 1, as follows:

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

Trials with more misses. This group corresponds to "difficult" genuine trials, for example speech pairs with the impersonator's natural voice against his impersonations (trials 31 and 34). These are *disguise* trials because the impersonator attempts to sound like someone else (one of the targets). Other trials with more misses are trials 7 and 12 recorded in different sessions. Even if human listeners tend to be more forgiving to channel or session differences than automatic systems, the perceptual inter-speaker variation may be more affected by the content or speaking style in the speech sample pairs: the conversation topic may be more interesting or the speaking style more lively in one sample compared to the other one.

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 noting that for the mimicry trials 14 and 25, half of the listeners responded that the samples corresponded to the same speaker, while the other half responded the opposite.

The grid in Table 11 shows the listeners' decisions for each of the 28 trials for the second listening test corresponding to Impersonator B. Again, we identify three groups of trials:

Easy trials. Consist of all baseline impostor trials, a few genuine trials (11, 12, 15, 19) and some mimicry trials (10, 14, 23). Universally easy trials are 2, 15 and 25 on

Table 10
Listeners total errors trial-by-trial for Listening test 1. The errors are shown highlighted. The decision number value indicates the confidence level of the judgement: *1: Same speaker, 2: somewhat 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	Panel 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
Baseline 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	4	5	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	5	1	2	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
Mimicry impostor	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
	Errors	6	5	9	10	9	6	5	8	9	7	9	12	4	7	5	6	9	9	5	8	4	12	7	7	6	14	10	5	11	10	9	11	10	15	

Misses False accepts Cannot decide

Table 11
Listeners total errors trial-by-trial for Listening test 2. The errors are shown highlighted. The decision number value indicates the confidence level of the judgement: *1: Same speaker, 2: somewhat same speaker, 3: I cannot tell, 4: somewhat different speaker, 5: different speaker.*

		LISTENERS																															
	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	Panel errors
Genuine	1	2	5	5	1	5	1	1	5	1	5	5	1	1	2	1	1	4	1	5	1	1	1	4	1	1	2	1	1	5	2	1	4	11
	4	5	5	2	5	1	5	1	1	2	2	2	1	5	1	5	1	2	5	5	2	4	5	5	2	1	5	5	1	5	1	5	2	15
	7	4	5	2	5	5	2	5	5	5	5	5	5	4	3	5	5	5	1	5	5	4	5	4	5	1	4	1	1	4	1	1	5	24
	9	2	1	4	1	5	1	1	1	1	1	5	4	1	2	2	1	2	2	5	4	5	1	2	5	1	2	1	1	1	2	5	4	10
	11	2	1	1	1	1	1	1	1	1	1	1	1	1	1	4	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	12	4	1	4	1	1	1	5	1	1	5	3	1	1	1	1	1	1	1	1	2	1	1	1	1	1	2	1	1	1	1	1	1	5
	15	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0
	17	3	5	1	5	2	1	5	1	5	5	3	1	2	1	5	1	5	1	5	1	4	4	4	2	1	4	1	5	1	1	1	2	15
	19	1	5	1	1	1	1	1	1	1	1	1	1	1	1	4	1	1	1	1	1	1	1	1	1	1	1	1	1	5	1	1	1	3
	22	2	5	4	1	1	1	5	5	5	2	5	1	2	4	4	5	2	5	1	1	2	1	1	1	1	2	1	1	1	2	1	4	11
	24	5	5	5	1	3	1	5	1	5	2	5	5	5	2	5	4	5	5	5	4	5	5	5	5	1	4	1	5	5	4	1	5	24
	27	1	1	2	5	1	1	5	1	5	4	1	1	2	1	1	5	4	1	5	1	1	1	1	1	1	1	1	1	1	1	1	1	7
Baseline 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	4	5	5	5	5	5	5	0
	5	5	5	5	5	5	5	5	1	5	5	5	5	2	5	5	5	5	5	5	5	5	5	5	5	5	4	5	5	5	1	5	1	4
	8	5	5	5	1	3	5	5	5	5	2	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	4	5	5	3
	13	5	5	4	5	5	5	5	5	5	5	5	5	4	5	5	5	5	5	5	5	5	5	3	5	5	5	5	5	5	5	5	5	1
	16	5	5	5	5	5	5	5	5	5	5	3	5	5	5	5	5	5	5	5	5	5	5	5	5	5	4	5	5	5	5	5	4	1
	20	5	5	5	5	5	5	5	1	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
	25	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
	28	5	5	5	5	3	5	5	1	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
Mimicry impostor	3	2	5	5	5	5	5	5	1	1	5	5	1	5	1	2	1	1	5	5	4	5	1	5	1	5	2	5	1	5	5	2	5	13
	6	2	5	1	5	5	5	1	1	5	1	1	1	5	5	5	5	5	1	5	1	1	5	1	5	5	1	5	5	1	1	5	1	15
	10	2	5	4	5	5	5	5	1	5	1	5	5	5	5	5	5	5	5	1	5	5	5	5	4	5	5	5	5	5	4	5	5	4
	14	5	5	3	5	5	5	5	1	5	5	5	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	2	5	5	3
	18	4	5	4	5	5	5	5	5	5	4	5	2	5	5	5	1	5	5	1	1	5	5	5	4	5	2	5	5	5	5	5	2	6
	21	4	5	4	1	5	5	1	1	5	1	1	5	2	5	5	1	5	1	1	5	5	5	1	5	5	1	5	5	5	4	5	1	12
	23	4	5	4	5	5	5	5	1	5	5	1	5	5	5	5	5	5	5	1	5	5	5	5	5	5	4	5	5	5	5	5	5	3
	26	5	5	5	5	5	5	5	1	5	5	3	5	5	5	5	4	5	5	1	2	5	2	5	2	5	2	5	1	1	4	5	5	9
	Errors	8	7	7	6	6	1	8	13	6	9	12	6	5	3	8	7	6	5	12	6	6	6	8	5	0	9	1	4	7	4	3	9	

Misses False accepts Cannot decide

which *all* the 32 listeners responded correctly. Note that no such trials were observed in the case of Listening test 1.

Trials with more misses. Trials considered the most difficult for the listeners are 7 and 24, which again correspond to the impersonator's natural voice against his impersonations (disguise). Other notably difficult trials are 4 and 17. Trial 4 is a genuine trial for TS7, and trial 17 is a genuine trial for Impersonator B's natural voice. Both trials contain samples from different sessions.

Trials with more false alarms. This case includes, again, mimicry samples (trials 3, 6 and 21).

The distribution of the "*same speaker*" decisions from the 34 listeners of Listening test 1 is shown in Fig. 7 for each of the target speakers. The graph indicates that the answers corresponding to the *same speaker* for genuine trials are higher in most cases except for the trials of TS4. The listeners had difficulty in deciding whether the trials containing the speech samples corresponded to the same speaker or different speakers. Here, the target speaker is a theatrical director, and his speech samples are segments from different recordings in which his speaking style changes considerably. This made the listeners conclude that the speaker was different. A similar confusion was noticed when the samples of TS4 were compared to the impersonator's natural speech and even more in impersonations of TS4. One reason for the confusion could be unfamiliarity

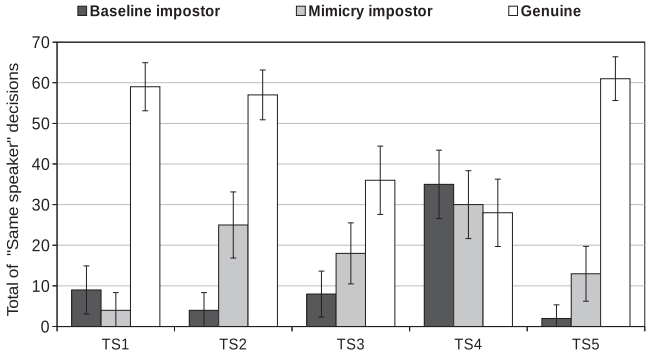


Fig. 7. Distribution of "Same speaker" decision per target speakers. Results from Listening test 1. Baseline impostor refers impersonatos own voice and mimicry impostor refers to impersonations.

with his voice, only 5 from 34 listeners recognized his voice to be one of the target speakers.

In the "same speaker" decisions for the mimicry impostor trials (Fig. 7), we observe an increase towards the target speakers TS2, TS3 and TS5 relative to baseline (no mimicry). The increase is higher for target speakers TS2 and TS5. The case of TS1 indicates that very few listeners were confused with the impersonated samples from the target speaker's genuine samples. For TS4, as explained above, the height of the distributions do not show clear differences. After the listening test, most of the listeners were able to name Martti Ahtisaari (TS1) and Sauli Niinistö (TS3) as being included in the target speakers. However, identifying all the speakers in the test did not affect the performance of a listener. For instance, 2 out of the 34 listeners correctly reported the target speakers, but made as many errors as the other listeners on average. The two listeners with only 4 errors in the test identified 3 out of the 5 target speakers.

Regarding the "same speaker" decisions for Impersonator B (Fig. 8) in the baseline impostor case, listeners did not make many errors. We can observe an increase of errors in the mimicry impostor trials for the four target speakers, where the increase is significant for two target speakers, TS6 and TS8.

For any automatic system to produce meaningful results, we need to provide a set up with a sufficient amount of speech material. On the other hand, the duration of a perceptual test should be short enough for a listener to perform the voice comparison task without weariness. In this sense, making a direct and fair comparison between the performance of listeners and that of the automatic verification systems is challenging for the type of material and classifiers used in this study. We can, however, evaluate the outcomes and make a qualitative analysis of the performances. To this end, the 10 s samples of Listening test 1 and 2 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 observing a genuine trial equaling to 0.5 and both false alarm and miss costs being set to 1 (Brümmer and du Preez, 2006). We optimized the bias for the evaluation data directly so that the results can be seen as the best possible (oracle) decisions. The errors for the listeners and the automatic systems are shaded in Table 12.

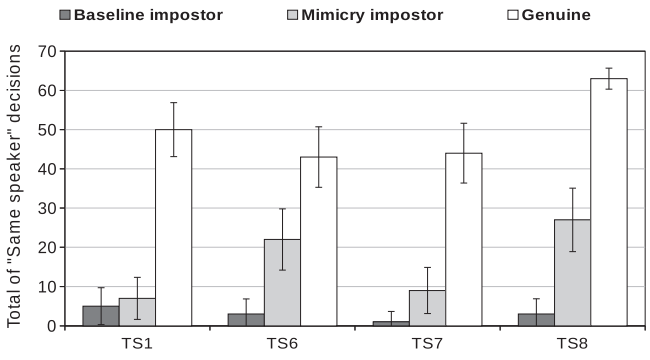


Fig. 8. Same as Fig. 7 but for Listening test 2.

Both the GMM-UBM and the i-vector Cosine systems performed similarly with equal number of errors. We can observe a similar trend as with the listening panel, that impersonation increases the errors in comparison with the correct detection of the baseline impostor trials. The total number of errors from the automatic systems are 5 (i-vector PLDA) and 6 (GMM-UBM and i-vector Cosine), while the listening pool made 8.15 errors on average. However, the best of our human listeners made only 4 errors, listeners number 13 and 21 in Table 10. All the automatic systems had an error in the genuine trial 29 for TS2. The GMM-UBM and the i-vector cosine systems made errors in trial 1 (a genuine trial for TS1). The i-vector Cosine and PLDA systems had an error for the impostor trial 5 for TS5.

In Listening test 2, Table 13 indicates that the total number of errors for the automatic systems is 7, 8 and 9 for the GMM-UBM and the i-vector systems respectively, while the average listening pool performed 7.25 errors. The best listener for this test made no errors, as can be seen for listener 25 in Table 11. We noticed that for trials 10, 14 and 19, all the automatic systems failed to answer correctly. Trial 19 is a genuine trial for TS1, the samples are extracted from two different interviews. Trials 10 and 14 included mimicry samples for TS1 and TS7, respectively. Trial 9 corresponds to a genuine trial for TS6 in which both i-vector systems reported errors.

5.3. Factors affecting listener panel performance

After completing the listening tests, we asked the listeners to report the names of the speakers they had recognized in the speech samples (second last free-text box in Appendix A). Based on the name lists, we computed the numbers of correctly and incorrectly identified target speakers. In Listening test 2, we collected more background information of the participants than in the previous test. These new binary *factors* were: play musical instrument, have formal musical training, have hi-fi as a hobby, and have experience in Linguistics. In Listening test 1, no listener could recognize the impersonator's identity, but in Listening test 2, 16 out of 32 listeners could name Impersonator B correctly. For that reason we used the binary factor whether the listener recognized the impersonator only in the data from Listening test 2.

To assess the importance of the above mentioned factors, we used the linear regression model where the listener's recognition accuracy was set as the dependent variable and the explanatory variables are the factors. In the case of Impersonator A, the obtained coefficient of determination was $R^2 = 0.23$ and the corresponding p-value 0.17, leading to non-significant statistical association between self-reported listener factors and listener

Table 12
Trial-by-trial comparison of total errors by listeners and automatic verification systems for Listening test 1 (10 s sample duration).

Type of trial		Trial #	Errors: 34 Listeners	Panel decision	GMM-UBM	i-vector Cosine	i-vector PLDA
Genuine	TS1 – TS1	1	8	0	1	1	0
	TS5 – TS5	4	5	0	0	0	0
	TS4 – TS4	7	27	1	1	0	0
	TS5 – TS5	11	2	0	0	0	0
	TS3 – TS3	12	19	1	0	0	0
	ImpA – ImpA	15	0	0	0	0	0
	TS1 – TS1	18	1	0	0	1	0
	TS4 – TS4	22	13	0	0	0	0
	ImpA – ImpA	23	0	0	0	0	0
	TS3 – TS3	26	13	0	0	0	1
	TS2 – TS2	29	7	0	1	1	1
	TS3 mimicry – ImpA	31	22	1	0	0	0
	TS2 – TS2	32	4	0	0	0	0
	TS3 mimicry – ImpA	34	31	1	0	0	0
Baseline impostor	TS2 – ImpA	2	1	0	0	0	0
	ImpA – TS5	5	2	0	0	1	1
	ImpA – TS4	8	6	0	0	0	0
	ImpA – TS5	13	0	0	0	0	0
	TS1 – ImpA	16	2	0	0	0	0
	TS3 – ImpA	19	5	0	0	0	0
	TS3 – ImpA	24	3	0	1	0	0
	TS2 – ImpA	27	3	0	0	0	0
	TS1 – ImpA	30	2	0	0	0	0
	ImpA – TS4	33	5	0	0	0	1
Mimicry impostor	TS3 – TS3 mimicry	3	9	0	0	0	0
	TS1 – TS1mimicry	6	3	0	0	1	0
	TS1 – TS1 mimicry	9	7	0	0	0	0
	TS4 mimicry – TS4	10	13	0	0	0	0
	TS2 – TS2 mimicry	14	17	1	0	0	1
	TS3 mimicry - TS3	17	9	0	0	0	0
	TS2 – TS2 mimicry	20	8	0	0	0	0
	TS5 mimicry – TS5	21	4	0	1	0	0
	TS4 mimicry – TS4	25	17	1	1	0	0
	TS5 mimicry – TS5	28	9	0	0	1	0
		TOTAL	-	6	6	6	5
		AVERAGE	8.15				

Misses False accepts

recognition accuracy. Perceived difficulty of the listening test factor had a single highest p-value of 0.15. Interestingly, when the same factors were used in Listening test 2, we obtained $R^2 = 0.59$ with $p = 0.00019$, resulting in a significant statistical association between accuracy and the self-reported listener data. The factor that was found to be statistically significant, for Listening test 2, was the number of correctly identified speakers, with $p = 0.0003$. The p-value of the perceived difficulty of the test is 0.13, which is in line with the results of Listening test 1. It is noteworthy that R^2 shows that the model explained 59% of the variation in the data. When all the factors were added to linear regression model of the data from Listening test 2, we observed a decrease of the adjusted R^2, in the complete model adjusted R^2 was 0.47, while in the reduced model it was 0.51. This indicates that the newly collected listener data did not bring extra explanatory information. The correct identification of the target speakers still remained the only significant explanatory variable with $p = 0.019$.

We can explain the difference between these two data sets by noting that both the target speakers and the impersonator himself in Listening test 2 are currently more known media personalities in Finland than those in Listening test 1. From the dataset 1, only the current president of Finland, Sauli Niinistö (TS3), is in the national media weekly. The artists Andy McCoy (TS6) and Pertti Pasanen (TS8), from the dataset 2, have very peculiar and recognizable speaking styles, making the identification easier for the listeners. Impersonator A is no longer active

Table 13
Trial-by-trial comparison of total errors by listeners and automatic verification systems for Listening test 2 (10 s samples).

Type of trial		Trial #	Errors				
			32 Listeners	Panel decision	GMM-UBM	i-vector Cosine	i-vector PLDA
Genuine	TS1 – TS1	1	11	0	0	0	0
	TS7 – TS7	4	15	0	0	1	0
	TS7 mimicry – ImpB	7	24	1	0	0	0
	TS6 – TS6	9	10	0	0	1	1
	TS8 – TS8	11	1	0	1	0	0
	TS7 – TS7	12	5	0	0	0	1
	TS8 – TS8	15	0	0	0	0	0
	ImpB – ImpB	17	15	0	0	0	0
	TS1 – TS1	19	3	0	1	1	1
	Imp B – TS6	22	11	0	0	0	0
	ImpB – TS7 mimicry	24	24	1	1	0	1
	TS6 – ImpB	27	7	0	0	0	0
Baseline impostor	TS7 – ImpB	2	0	0	0	0	0
	TS1 – ImpB	5	4	0	0	0	0
	ImpB – TS8	8	3	0	0	0	0
	ImpB – TS1	13	1	0	0	0	1
	TS7 – ImpB	16	1	0	0	0	0
	TS6 – ImpB	20	1	0	0	0	0
	ImpB – TS8	25	0	0	0	0	0
	TS6 – ImpB	28	2	0	0	0	0
Mimicry impostor	TS6 mimicry – TS6	3	13	0	0	0	1
	TS8 mimicry – TS8	6	15	0	1	0	1
	TS1 mimicry – TS1	10	4	0	1	1	1
	TS7 – TS7 mimicry	14	3	0	1	1	1
	TS7 mimicry – TS7	18	6	0	0	1	0
	TS8 – TS8 mimicry	21	12	0	0	1	0
	TS1 mimicry – TS1	23	3	0	1	0	0
	TS6 – TS6 mimicry	26	9	0	0	1	0
		TOTAL	.	2	7	8	9
		AVERAGE	7.25				

Misses False accepts

in TV and radio comedy circle, whereas Impersonator B has had a TV show on a national channel continuously for years and also released a new music CD in 2014.

6. Conclusions

In this work, we assessed the accuracy of three automatic speaker verification systems: GMM-UBM, i-vector with cosine scoring and i-vector with PLDA scoring, for mimicked data in Finnish language. This study includes, for the first time, a wide analysis of the performance of state-of-the-art automatic speaker verification systems in the presence of voice mimicry. A perceptual test was also included for two reasons, firstly, to find out whether speaker verification performance of non-expert (naïve) listeners is affected by the presence of mimicry, and secondly, to set a comparative benchmark parallel to automatic evaluations.

In this study, we observed that for the pooled trials of 20 s duration in the case of data where the speech content did not match (Table 7), the automatic systems EERs was reasonably low ranging from 4.36% to 13.76%. For the case of matched speech content, the obtained EERs ranged from 7.16% to 15.76% (Table 8). Comparing the performance of the most accurate system in these two cases, we observed that the accuracy of the i-vector with PLDA scoring system was affected specially in the text independent case.

In another experiment that includes test samples of 10 s in duration (Table 9) —samples used in the listening tests— the automatic systems performed poorly as expected, due to the short duration of the utterances. Here, the listening panel outperformed the three automatic systems by a wide

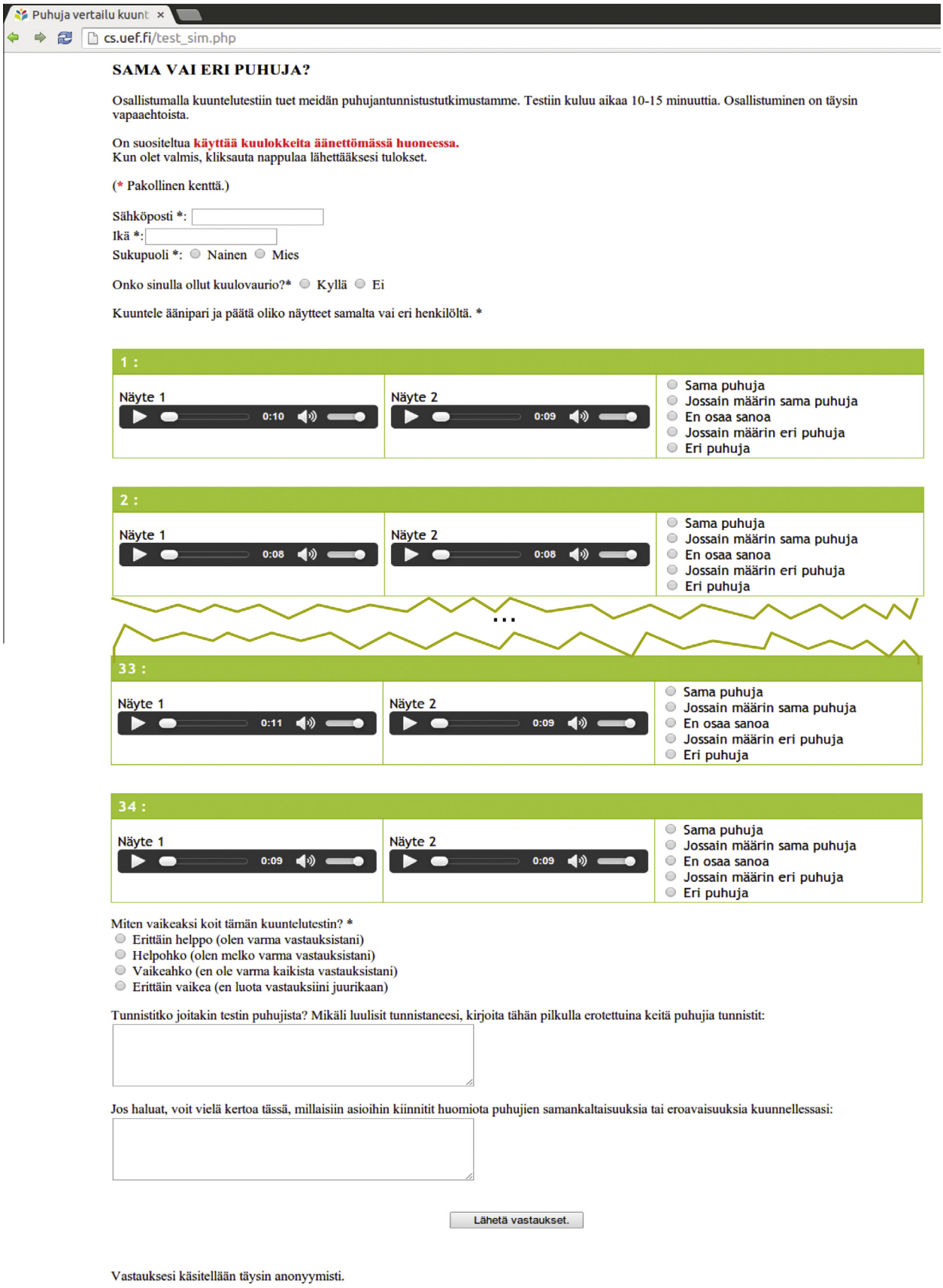


Fig. A.9. Web-form for the listening test 1 in Finnish. The listeners were instructed to listen and decide whether the speech samples belong to the same or different speaker. The listener's decision options were: (a) Sama puhuja (Same speaker), (b) Jossain määrin sama puhuja (Somewhat same speaker), (c) En osaa sanoa (I cannot tell), (d) Jossain määrin eri puhuja (Somewhat different speaker), and (e) Eri puhuja (Different speaker).

margin. When comparing humans and automatic systems, it is important to keep in mind that the participants in the listening test are familiar with the target speaker's voice since they are well-known figures and many listeners have an advantage of substantially more training material obtained by following Finnish media. That is, familiarity with the target speakers gave an advantage to listeners over automatic systems, as was seen in Section 5.3.

Interestingly, the EER of two automatic systems was increased under mimicry as Tables 7 and 8 indicate. For instance, the best system (i-vector PLDA) degraded from EER of 4.36% to EER of 7.38% under impersonation "attack". Degradation of performance was also observed with i-vector Cosine scoring, too. For GMM-UBM system, the degradation of performance was observed only in the Same text results where EER increases from 10.38% to 11.21%. However, the per-target score distributions (shown in Figs. 3–6) clearly indicate that for most targets, our two impersonators were *not* able to increase their ASV system scores significantly. This apparent discrepancy across the "classic" EER measure and target-specific score distributions can be attributed to the fact that one global threshold is placed in case of EER, whereas in target-specific score distributions, each target has its own threshold. The EER metric was chosen mainly for convenience as it is widely used in evaluating speaker verification (and other biometric systems) accuracy. Our set-up, however, differs from traditional speaker verification evaluation in terms of both the trial count (extremely small) and the fact that our main goal is to learn the effect of "before" (zero-effort) versus "after' (dedicated) effect of impersonation on a target-by-target and impersonator-by-impersonator basis, rather than in a average terms where all trials are pooled in a single error measure. Future work, therefore, should carefully address the choice of evaluation methodology and objective metrics, keeping these constraints and differences in mind.

A few earlier studies have reported that speakers whose voices are closer to the impersonator's voice might be easier to imitate. Referring back to the observations of target-specific score distributions in Figs. 3–6, we did not observe such effect systematically. This could be because, despite the multitude of normalization techniques applied, MFCC features are sensitive to not only to changes in the voice quality (due to mimicry) but also to the changes in channel (due to different recording conditions). This would suggest studying speaker similarity in terms of prosody or other features less susceptible to mismatch in data quality.

Most of the listeners did quite well in the perceptual test under normal conditions (no mimicry nor disguise). However, when the listeners were presented with the impersonators own voice and with a "good" impersonation of a target speaker (disguise trial), according to Tables 10 and 11 at least 75% (in Listening test 1, 22 out of 34 and in Listening test 2, 24 out of 32) of the listeners judged the trials incorrectly. The two impersonator were able to disguise their own voice, which lead many listeners to conclude "different speaker". However, the impersonated voices were not confused with the given target speakers in most cases. This suggests that human listeners may be more likely to make recognition errors under disguise rather than mimicry. In addition, listeners' performance is enhanced if they can identify the target speakers. In fact, this was the statistically significant factor that affected listener performance. Familiarity with the target voices has likely a major effect that should be investigated further in future work. We also observe from Figs. 7 and 8 that listener performance had a high variance across the target speakers.

A major thrust of the present work goes into analyzing the performance of a single random listener versus automatic system. Our results in Table 9 indicate that considering the whole panel instead of just the single listener can boost the human performance considerably. This result gives rise to a few interesting questions. How to select a minimal sized listening panel that does not have high variance of performance? Can we derive an auxiliary measure where listening panel and automatic system can be fairly compared? These are possible topics of our future work. In addition, we plan to incorporate more mimicry data to the study and include expert listeners in the perceptual test for future studies.

Acknowledgments

This work was supported by the Academy of Finland projects number 253120, 253000 and 283256. 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 as well as to the two impersonators for their willingness to produce the needed audio data for this study. We further thank Dr. Md Sahidullah for his help on language proofing the earlier version of this work.

Appendix A. Listening test web form

See Fig. A.9.

References

Alegre, F., Vipperla, R., Amehraye, A., Evans, N. 2013. A new speaker verification spoofing countermeasure based on local binary patterns, in: Proc. Interspeech.

Alegre, F., Soldi, G., Evans, N. 2014. Evasion and obfuscation in automatic speaker verification. In: Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2014), Florence, Italy.

Bech, S., Zacharov, N., 2007. Perceptual Audio Evaluation–Theory, Method and Application. John Wiley & Sons.

Brookes, M. et al. 2006. Voicebox: Speech Processing Toolbox for Matlab. <www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html [January 2014].

Brümmer, N., du Preez, J., 2006. Application-independent evaluation of speaker detection. Comput. Speech Lang. 20, 230–275.

Campbell Jr., J.P., 1997. Speaker recognition: a tutorial. Proc. IEEE 85, 1437–1462.

Dehak, N., Dehak, R., Glass, J., Reynolds, D., Kenny, P., 2010. Cosine similarity scoring without score normalization techniques. In: Proc. Odyssey Speaker and Language Recognition Workshop, pp. 71–75.

Dehak, N., Kenny, P., Dehak, R., Dumouchel, P., Ouellet, P., 2011. Front-end factor analysis for speaker verification. ITASLP 19, 788–798.

De Leon, P., Pucher, M., Yamagishi, J., Hernaez, I., Saratxaga, I., 2012. Evaluation of speaker verification security and detection of HMM-based synthetic speech. IEEE Trans. Audio Speech Lang. Process. 20, 2280–2290.

Doddington, G., Liggett, W., Martin, A., Przybocki, M., Reynolds, D., 1998. Sheep, goats, lambs and wolves: a statistical analysis of speaker performance in the nist 1998 speaker recognition evaluation. In: ICSLP.

Ephraim, Y., Malah, D., 1985. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. IEEE Trans. Acoust. Speech Signal Process. 33, 443–445.

Evans, N., Kinnunen, T., Yamagishi, J., 2013. Spoofing and countermeasures for automatic speaker verification. In: Interspeech 2013, Lyon, France. pp. 925–929.

Farrús, M., Wagner, M., Erro, D., Hernando, F.J., 2010. Automatic speaker recognition as a measurement of voice imitation and conversion. Int. J. Speech Lang. Law 1, 119–142.

Furui, S., 1997. Recent advances in speaker recognition. In: Audio-and Video-based Biometric Person Authentication, pp. 235–252.

Garcia-Romero, D., Espy-Wilson, C.Y., 2011. Analysis of i-vector length normalization in speaker recognition systems. In: Interspeech 2011, Florence, Italy, pp. 249–252.

González Hautamäki, R., Hautamäki, V., Rajan, P., Kinnunen, T., 2013a. Merging human and automatic system decisions to improve speaker recognition performance. In: Interspeech 2013, Lyon, France. pp. 2519–2523.

González Hautamäki, R., Kinnunen, T., Hautamäki, V., Leino, T., Laukkanen, A.M., 2013b. I-vector meet imitators: on vulnerability of speaker verification systems against voice mimicry. In: Interspeech 2013, Lyon, France. pp. 930–934.

González Hautamäki, R., Kinnunen, T., Hautamäki, V., Laukkanen, A.M., 2014. Comparison of human listeners and speaker verification systems using voice mimicry data. In: Speaker Odyssey 2014, Joensuu, Finland. pp. 137–144.

Greenberg, C., Martin, A., Doddington, G., Godfrey, J., 2011. Including human expertise in speaker recognition systems: report on a pilot evaluation. In: ICASSP 2011, Prague, Czech Republic. pp. 5896–5899.

Hautamäki, V., Kinnunen, T., Nosratighods, M., Lee, K.A., Ma, B., Li, H., 2010. Approaching human listener accuracy with modern speaker verification. In: Interspeech 2010, Makuhari, Japan. pp. 1473–1476.

Hébert, M., 2008. Text-dependent speaker recognition. In: Benesty, J., Sondhi, M., Huang, Y. (Eds.), Springer Handbook of Speech Processing. Springer, Berlin Heidelberg, pp. 743–762.

Hu, Y., Loizou, P., 2006. Subjective comparison of speech enhancement algorithms, in: 2006 IEEE International Conference on Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings, pp. I–I.

Kajarekar, S.S., Bratt, H., Shriberg, E., de Leon, R., 2006. A study of intentional voice modifications for evading automatic speaker recognition. In: The Speaker and Language Recognition Workshop, 2006. IEEE Odyssey 2006, pp. 1–6.

Kanagasundaram, A., Dean, D., Gonzalez-Dominguez, J., Sridharan, S., Ramos, D., Gonzalez-Rodriguez, J., 2013. Improving short utterance based i-vector speaker recognition using source and utterance-duration normalization techniques. In: Interspeech 2013, Lyon, France. pp. 2465–2469.

Kenny, P., 2010. Bayesian speaker verification with heavy-tailed priors. In: Speaker Odyssey (2010).

Kinnunen, T., Wu, Z.Z., Lee, K.A., Sedlak, F., Chng, E.S., Li, H., 2012. Vulnerability of speaker verification systems against voice conversion spoofing attacks: the case of telephone speech. In: ICASSP 2012, Kyoto, Japan. pp. 4401–4404.

Lau, Y., Wagner, M., Tran, D. 2004. Vulnerability of speaker verification to voice mimicking. In: Proc. Int. Symp on Intelligent Multimedia, Video & Speech Processing (ISIMP'2004), Hong Kong. pp. 145–148.

Lau, Y., Tran, D., Wagner, M. 2005. Testing voice mimicry with the YOHO speaker verification corpus. In: Knowledge-Based Intelligent Information and Engineering Systems (KES 2005), Melbourne, Australia. pp. 15–21.

Leskelä, J., 2011. Changes in *F*0, Formant Frequencies and Spectral Slope in Imitation. Master's Thesis. University of Tampere. (in Finnish).

Loizou, P.C., 2007. Speech enhancement. In: Theory and practice. Taylor & Francis, USA.

Mariéthoz, J., Bengio, S., 2005. Can a Professional Imitator Fool a GMM-Based Speaker Verification System? Idiap-RR. IDIAP.

Martin, A., Doddington, G., Kamm, T., Ordowski, M., Przybocki, M., 1997. The DET curve in assessment of detection task performance. In: EUROSPEECH, Rhodos, Greece, pp. 1895–1898.

Mary, L., Anish Babu, K.K., Joseph, A., George, G.M., 2013. Evaluation of mimicked speech using prosodic features. In: ICASSP 2013, Vancouver, pp. 7189–7193.

Panjwani, S., Prakash, A., 2014. Crowdsourcing attacks on biometric systems. In: Symposium on Usable Privacy and Security (SOUPS), Menlo Park, USA, pp. 257–269.

Perrot, P., Aversano, G., Chollet, G., 2007. Voice disguise and automatic detection: review and perspectives. In: Progress in Nonlinear Speech Processing. Lecture Notes in Computer Science, pp. 101–117.

Prince, S.J., Elder, J.H., 2007. Probabilistic linear discriminant analysis for inferences about identity. In: IEEE 11th International Conference on Computer Vision, 2007. ICCV 2007. IEEE, pp. 1–8.

Reynolds, D., 2002. An overview of automatic speaker recognition technology. In: 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. IV-4072–IV-4075.

Reynolds, D., Quatieri, T., Dunn, R., 2000. Speaker verification using adapted Gaussian mixture models. Digital Signal Process. 10, 19–41.

Saeidi, R., Lee, K.A., Kinnunen, T., Hasan, T., Fauve, B., Bousquet, P.M., Khoury, E., Martinez, P.L.S., Kua, J.M.K., You, C.H., Sun, H., Larcher, A., Rajan, P., Hautamäki, V., Hanilci, C., Braithwaite, B., Hautamäki, R.G., Sadjadi, S.O., Liu, G., Boril, H., Shokouhi, N., Matrouf, D., Shafey, L.E., Mowlaee, P., Epps, J., Thiruvaran, T., van Leeuwen, D.A., Ma, B., Li, H., Hansen, J.H.L., Bonastre, J.F., Marcel, S., Mason, J., Ambikairajah, E., 2013. I4U submission to NIST SRE 2012: a large-scale collaborative effort for noise-robust speaker verification. In: Interspeech 2013, Lyon, France. pp. 1986–1990.

Schmidt-Nielsen, A., Crystal, T.H., 2000. Speaker verification by human listeners: experiments comparing human and machine performance using the NIST 1998 speaker evaluation data. Digital Signal Process. 10, 249–266.

Villalba, J., Lleida, E., 2011. Detecting replay attacks from far-field recordings on speaker verification systems. Biometrics ID Manage., 274–285

Wu, Z., Kinnunen, T., Chng, E.S., Li, H., Ambikairajah, E., 2012. A study on spoofing attack in state-of-the-art speaker verification: the telephone speech case. In: Proc. Asia-Pacific Signal Information Processing Association Annual Summit and Conference (APSIPA ASC).

Wu, Z., Evans, N., Kinnunen, T., Yamagishi, J., Alegre, F., Li, H., 2015. Spoofing and countermeasures for speaker verification: a survey. Speech Commun. 66, 130–153.

Yager, N., Dunstone, T., 2010. The biometric menagerie. IEEE Trans. Pattern Anal. Mach. Intell. 32, 220–230.

Zetterholm, E., 2007. Detection of speaker characteristics using voice imitation. In: Speaker Classification II. Lecture Notes in Computer Science, pp. 192–205.

Zetterholm, E., Elenius, D., Blomberg, M., 2004. A comparison between human perception and a speaker verification system score of a voice imitation. In: 10th Australian International Speech Science and Technology Conference SST2004, pp. 393–397.

Zhang, C., Tan, T., 2008. Voice disguise and automatic speaker recognition. Forensic Sci. Int., 118–122