Acoustical and perceptual study of voice disguise by age modification in speaker verification

Rosa González Hautamäki*, Md Sahidullah, Ville Hautamäki, Tomi Kinnunen

School of Computing, University of Eastern Finland, P.O. Box 111 FI-80101 Joensuu, Finland

Abstract

The task of speaker recognition is feasible when the speakers are *co-operative* or wish to be recognized. While modern automatic speaker verification (ASV) systems and some listeners are good at recognizing speakers from modal, unmodified speech, the task becomes notoriously difficult in situations of deliberate *voice disguise* when the speaker aims at masking his or her identity. We approach voice disguise from the perspective of acoustical and perceptual analysis using a self-collected corpus of 60 native Finnish speakers (31 female, 29 male) producing utterances in *normal*, *intended young* and *intended old* voice modes. The normal voices form a starting point and we are interested in studying how the two disguise modes impact the acoustical parameters and perceptual speaker similarity judgments.

First, we study the effect of disguise as a relative change in fundamental frequency ($F0$) and formant frequencies ($F1$ to $F4$) from modal to disguised utterances. Next, we investigate whether or not speaker comparisons that are deemed easy or difficult by a modern ASV system have a similar difficulty level for the human listeners. Further, we study affecting factors from listener-related self-reported information that may explain a particular listener's success or failure in speaker similarity assessment.

Our acoustic analysis reveals a systematic increase in relative change in mean $F0$ for the intended young voices while for the intended old voices, the relative change is less prominent in most cases. Concerning the formants $F1$ through $F4$, 29% (for male) and 30% (for female) of the utterances did not exhibit a significant change in any formant value, while the remaining $\sim 70\%$ of utterances had significant changes in at least one formant.

Our listening panel consists of 70 listeners, 32 native and 38 non-native, who listened to 24 utterance pairs selected using rankings produced by an ASV system. The results indicate that speaker pairs categorized as easy by our ASV system were also easy for the average listener. Similarly, the listeners made

*Corresponding author

Email addresses: rgonza@cs.uef.fi (Rosa González Hautamäki), sahid@cs.uef.fi (Md Sahidullah), villeh@cs.uef.fi (Ville Hautamäki), tkinnu@cs.uef.fi (Tomi Kinnunen)

more errors in the difficult trials. The listening results indicate that target (same speaker) trials were more difficult for the non-native group, while the performance for the non-target pairs was similar for both native and non-native groups.

Keywords: Voice disguise, voice modification, speaker verification, acoustical analysis, fundamental frequency, formant frequencies, perceptual evaluation

1. Introduction

The human voice carries individual characteristics that can be used to identify the speaker. In *speaker recognition*, the main focus of analysis is on who is speaking rather than what is being said. The human ability to recognize people by their voices is well known, especially in relation to familiar speakers (Schmidt-Nielsen and Stern, 1985). Moreover, the use of technology in the speaker recognition task has increased with the widespread use of personal handheld devices to access information and for daily communications. Nevertheless, whether performed by humans or automatic systems, the speaker recognition task can be challenging as speech is subject to many variations induced by the speaker, the communication scenario and the transmission channel (Campbell, 1997; Hansen and Hasan, 2015; Kinnunen and Li, 2010). State-of-the-art *automatic speaker verification* (ASV) technology (Campbell, 1997; Kinnunen and Li, 2010) has advanced to deal with additive and channel variability, but the *intrinsic*, or speaker-based, variations of the speech remain very challenging. According to Hansen and Hasan (2015), the variations in the speaker's voice characteristics can be affected by *the scenario* or by *the task* performed by the speaker, which may include *vocal effort*, *emotion*, *physical condition* and *voluntary alterations* of the voice.

Voluntary variations of speech can be induced either by *electronic* means, in which speech can be purposefully modified by the use of voice transformation technology (Mohammadi and Kain, 2017; Stylianou, 2009; Clark and Foulkes, 2007); or by *non-electronic* means. Two cases of the latter can be identified. Firstly, the speaker may attempt to be identified as another person by means of *mimicry* or *impersonation* (González Hautamäki et al., 2015; López et al., 2013; Panjwani and Prakash, 2014), such as voice acting or stand-up comedy. Secondly, in a more generic case that does not necessarily involve any specific target voice, the speaker adapts or transforms his or her voice with the aim of concealing his or her *audio identity*. It is this broad form of variation, known as *voice disguise*, that forms the focus of our study. It may involve several variations in speaking style (Perrot et al., 2007; Rodman and Powell, 2000; San Segundo et al., 2013) and is a particularly relevant concern in forensics or audio surveillance. This might include, for example, analysis of an armed robbery or a black-mailing call in which the perpetrator does not wish to be identified later.

Voice disguise may include one or several of the following modifications: *a) forced modifications of the physical vocal cavities*, such as pinched nose, pulled

cheeks, the use of physical obstruction objects (e.g. helmet, face mask (Saeidi et al., 2016), handkerchief over the mouth, pencil or chewing gum (Zhang and Tan, 2008)); *b) changes in the type of phonation*, or modification of the sound source, e.g. imitating a speech defect, or a specific type of phonation such as a creaky, hoarse or falsetto voice (San Segundo et al., 2013); *c) phonemic modification* related to the change in pronunciation, e.g. adopting foreign accent sounds (Leemann and Kolly, 2015) or nasal speech; and *d) prosody-related modifications* in pitch or speech rate (Künzel et al., 2004; Zhang, 2012). A visual example of a speaker's voluntary modification of the voice is shown in Fig. 1, which presents spectrograms and $F0$ contours of the speaker's own voice and two disguised voices.

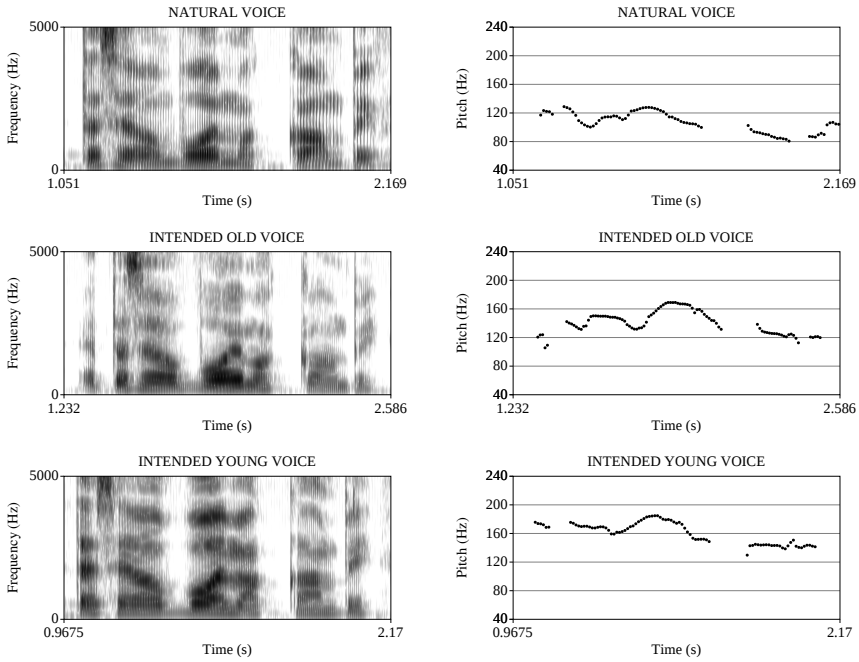


Figure 1: An example of intra-speaker voice variation. Spectrograms (left) and fundamental frequency ($F0$) contour (right) of a male speaker's own voice (top), intended old voice (middle) and intended young voice (bottom) with the same speech content. $F0$ computed using Praat (Boersma and Weenink, 2015). The figure illustrates that the selected speaker raised $F0$ for both, intended old and intended young voice.

Voice disguise is a complex problem that has attracted interest from different research communities. Previous studies on the topic enable one to identify three general perspectives: *vulnerability analysis of ASV systems*, *effects on acoustic parameters* and *perceptual experiments*. Vulnerability analysis mainly addresses voice disguise in terms of target speaker false rejections, and compares ASV system results with and without intentional voice modification. Acoustic analysis focuses on changes in the articulatory and voice source settings, which

are most commonly measured through *fundamental frequency* ($F0$) and formant frequencies. Finally, perceptual evaluations study the performance of human listeners, usually in a controlled environment, in a speaker comparison task that includes disguised voices.

Our preliminary analyses of the effects of voice disguise on modern ASV systems was reported in (González Hautamäki et al., 2016). The experiments indicated the vulnerability of our ASV systems in the presence of disguised voices when the speakers intended old and young voices. In terms of equal error rate (EER), the standard accuracy measure of biometric recognizers, we observed a 7-fold increase for intended old voices for male speakers and 5-fold increase for female speakers. The increase in EER was even higher for the intended young voices: 11-fold for male and 6-fold for female speakers. An analysis of $F0$ histogram distributions for natural, intended old and intended young voices indicated a shift towards higher frequencies for some of the speakers. $F0$ values are expected to be higher for younger speakers and for most of the speech segments the $F0$ increased for intended young voices, while in the case of male speakers it also increased for the intended old voice.

The present study seeks to proceed beyond the population level and the 'average' performance related to the EER metric. Its main objective is to gain a better understanding of the considerable performance loss of our ASV systems against voice disguise by a deeper investigation into the acoustics of disguised speech and an evaluation of the performance of human listeners. It does so by studying the relative change in $F0$ and the difference between formants $F1$ through $F4$, for each speaker caused by disguise. These acoustic features are affected, among many other factors, by biological ageing. Our study addresses a "simulated aging" process using young and old voice stereotypes, rather than biological ageing. In order to quantify the change in formant frequencies, we introduce a novel method to address the joint change in all averaged formant values with respect to their *direction* of change — *none*, *increase* or *decrease* — instead of the raw formant measurements. This sort of discrete descriptive presentation enables us to enumerate all the possible formant change patterns and to study their frequency of occurrence in order to reveal whether any speaker-independent voice disguise strategies can be identified.

In addition to the acoustic analysis, we designed a perceptual experiment to benchmark the performance of human speaker verification accuracy under voice disguise. Our perceptual task includes two novel elements, first, a selection of speech sample pairs, or trials, using the results from the ASV systems implemented in our previous study (González Hautamäki et al., 2016). More specifically, we use the ASV system output to select *easy*, *intermediate* and *difficult* speaker pairs. The test includes trials with and without the presence of voice disguise as well as cases with the same and different speakers. The second element is to compare the performance of native and non-native listeners for its relevance in a forensic setting such as voice-lineups, in which the listeners may be unfamiliar with the speaker's language. Previous studies confirm that the reliability of non-native listeners decreases in speaker recognition tasks (Eriksson et al., 2010; Köster et al., 1997) which is why the results of non-native listeners

in speaker comparison should be considered with caution. Although the accuracy of native vs. non-native listeners under normal voices has been addressed several times (e.g. by Kahn et al. (2011); Hautamäki et al. (2010); Schwartz et al. (2011); Ramos et al. (2011)), the authors are unaware of a previous study that compares the performance of native and non-native listeners with disguised voices for speaker recognition.

The dataset used for this study was collected by the authors and is the same that was used in our preliminary study (González Hautamäki et al., 2016). Our data consists of speech from 60 native Finnish speakers with 31 female and 29 male speakers. We instructed the speakers to not sound like themselves by producing *intended old* and *intended young* voices in addition to their normal modal voices without disguise. The intended vocal age was set to define a disguise strategy that assumes that the speakers have a common knowledge of how stereotypical old and young voices may sound like. In this setting, our experiments dealt with analyzing the effects of disguise in speaker verification accuracy. For our perceptual speaker comparison experiment, we recruited 70 listeners (32 native, 38 non-native), and each listened to the same set of 24 utterance pairs, in which the trial order was randomized for each listener.

The specific research questions that the present study seeks to answer are phrased as follows:

Q1. Is there a significant change in the $F0$ of female and male speakers when attempting voice disguise to sound older or younger? Does it increase or decrease?

Q2. Are there significant differences between the average of the first four formant frequencies of the natural and disguised voices of the female and male speakers?

Q3. Is there any speaker-independent disguise pattern that can be associated with formant frequency variation between natural speech and the studied strategy for disguised speech?

Q4. Is listener performance affected by the presence of voice disguise in a similar way to the performance of our ASV systems?

Q5. Does knowledge of the speakers' native language play a role in making more reliable perceptual speaker comparisons under modal voices and under disguise?

Q6. Is there a particular trial category or listener attribute that affects listener performance in the perceptual speaker recognition task?

2. Previous work on intentional voice modification and vocal ageing

Our study focuses on disguising one's voice identity by means of a specific type of voice modification related to one's perceptual age. Our primary interest is in identity disguise and its detrimental effects on the accuracy of speaker

recognition, while age disguise merely serves as a shared and not too constrained task across our speakers. Given that our speakers are naïve, we do not necessarily expect them to produce particularly convincing old or young voice imitations. Nevertheless, in order to place our findings in the relevant context, and to help us interpret the findings of the acoustic analysis, it is necessary to provide a brief review of both voice disguise and age-related changes on the speaker's voice. These are provided in the following two subsections respectively.

2.1. Voice disguise

Voice disguise have been studied at least for the past four decades, together with its impact on speech perception and speaker recognition. Table 1 presents a summary of our study and selected previous studies. Early studies focused on the acoustical analysis of source characteristics and vocal tract speech parameters (Endres et al., 1971). Subsequently, phonetic and forensic studies focus on the perceptual evaluation of modified voices (Hirson and Duckworth, 1993; Reich and Duke, 1979).

In more recent studies, the vulnerability of automatic systems has been studied, either for speaker verification or forensic applications (Künzel et al., 2004; Kajarekar et al., 2006; Zhang and Tan, 2008). In Künzel et al. (2004), the authors studied the effects of voice disguise on the performance of automatic *forensic speaker recognition* (FSR) system considering only target trials.

The evaluation results of 50 German speakers with three types of disguised voices (high pitch, low pitch and pinched nostrils) only marginally affected the FSR system's performance when the speakers' enrollment speech material contained the same type of disguised voices. By contrast, when the evaluation of disguised voices was performed using natural voice samples for enrollment, the performance was considerably degraded particularly with high- and low-pitch disguised voices. The authors observed that speakers who were not recognized by the system and used disguise by increasing their $F0$, also changed their voice from modal type to *falsetto*, which is one of the most extreme alterations in voice production (San Segundo et al., 2013). This variation affected the spectral features, *mel-frequency cepstral coefficients* (MFCCs), used by the evaluated FSR system that was evaluated. Zhang (2012) evaluated an automatic FSR system performance with raised and lowered $F0$ speech from 11 Chinese speakers. The study indicated that the system performance of raised $F0$ provided 10% recognition rate, while for lowered $F0$ the recognition rate was 55% from a 90% correct recognition for natural voices. The performance of the FSR system was degraded with disguised voices, particularly with raised $F0$ voices.

In the case of ASV systems, Kajarekar et al. (2006) evaluated a state-of-the-art Gaussian mixture modeling (GMM) system in which the speakers that were free to choose the disguise voices and later described their vocal variations with a label. The ASV system indicated a dramatic increase in the false rejection (miss) rate from 7.33% to 39.3% when the system was trained using natural voices. The error was reduced when voice disguise was included in the training phase. In addition, the authors conducted a perceptual speaker verification experiment that included 25 listeners. The human performance was comparable to that of

the automatic system in the case of natural voices. But in the case of disguised voices, the ASV system outperformed the human listeners.

In the same context, our previous study (González Hautamäki et al., 2016) evaluates the performance of six ASV systems. In terms of equal error rate (EER), the ASV systems' configuration performance was degraded with disguised voices. For example, the ivector-PLDA system's performance degraded for male speakers from 2.82% to 19.45% for intended old voice and 30.1% for intended young voice. Similar degradations were observed for female speakers. Such low performance of ASV systems with the disguised data motivated us to explore the possible reasons for this effect in acoustical and perceptual perspectives by considering the early studies of this problem.

From the acoustical perspective of the effects of voice disguise, Endres et al. (1971) investigated voice modifications in terms of the changes in $F0$ and formants by means of speech spectrograms. The authors reported that for disguised voices, the formant positions of vowels or vowel-like sounds shifted to lower or higher frequencies with respect to the natural voice of the same speakers. Only the first formant, $F1$, was found to remain relatively intact. Similarly, the mean $F0$ was affected by deliberate voice modification.

Similarly, Zhang (2012) conducted an acoustical analysis of raised and lowered $F0$ among 11 Chinese speakers. A statistical analysis was conducted for the following acoustic features: $F0$, syllable duration, the intensity and formant frequencies of five selected vowels, and *long term average spectrum* (LTAS) (Kinnunen et al., 2006). The author reported that some speakers were more skillful at adjusting their $F0$ than others and that raising $F0$ was easier than lowering it.

Other relevant studies that focus mainly on the acoustic analysis of disguised voices include those of Amin et al. (2014) and Leemann and Kolly (2015). Amin et al. (2014) studied 27 voices that were produced by three impersonators. The voices did not correspond to any particular target speaker but were defined in relative terms, for example, modified age and speaker's age. The authors studied $F0$, speech rate and formants ($F1$ to $F4$) of six vowel categories. In addition, the *electroglottograph* (EGG) signal for vocal folds activity during voice production was studied. The formant differences across the voices were found to be highly dependent on the vowel category. The authors developed an objective metric based on the vowel-dependent variance of the formants for each disguised voice. In another relevant work, Leemann and Kolly (2015) studied supra-segmental temporal features based on amplitude peaks and voicing features. These features were shown to have considerable between-speaker variation and low within-speaker variation across dialect disguises. The results suggested that imitating another dialect (to sound like a native speaker) is a challenging task. Nevertheless, their findings indicated that those speakers who succeeded in being accepted as native speakers of the imitated dialect may have approximated supra-segmental temporal features of the target dialect. In another recent work, Skoog Waller and Eriksson (2016) investigated how speakers manipulate their voice characteristics to sound either 20 years younger or older than their true age.

Table 1: Selected previous studies in voice disguise and the present study. F: Female, M: Male, FSR: Forensic speaker recognition.

Study	Task	Speakers	Listeners	Speech type	Type of disguise	Evaluation method
Endres et al. (1971)	Speaker identification	1 F, 5 M	n/a	21 samples in German	3 voices freely chosen by the speaker	Acoustic and spectrogram analysis
Reich and Duke (1979)	Speaker identification	40 M	30	Read English sentences	"70-80" years old, hoarse, nasal, slow, 1 freely chosen	Perceptual
Künzel et al. (2004)	Speaker recognition for forensic application	100 M	-	Read call threats in German	Increased pitch, lowered pitch, pinched nose	Automatic FSR system
Kajarekar et al. (2006)	Speaker recognition	32	25	Conversational speech in English	Voices freely chosen, e.g: high and low pitch, dialect and foreign accent imitation	Automatic system and perceptual
Zhang (2012)	Speaker recognition for forensic application	11 M	10 M	Read sentences in Chinese	Raised and lowered pitch	Acoustical, automatic FSR system, perceptual
Amin et al. (2014)	Disguise detection	1 F and 2 M impersonators	18	Read short sentences in English	9 freely chosen, e.g. old and young, cross gender old and young	Acoustical and perceptual
Leemann and Kolly (2015)	Native dialect detection	12 F, 8 M	9 F, 13 M	Read sentences in German	Dialect imitation	Acoustical and perceptual
Skoog Waller and Eriksson (2016)	Speaker's age estimation	18 F, 18 M	47 F, 13 M	Read sentences in Swedish	Intended 20 years younger and older	Acoustical and perceptual
This study	Speaker recognition	31 F, 29 M	26 F, 44 M	Read sentences in Finnish and English	Intended old and young	Acoustical and perceptual

They found that the speakers' $F0$ and speech rate were increased for attempted younger voices and decreased for the attempted older voices.

The effect of voice disguise on human perception has also been studied in different tasks, including speaker identification, disguise detection, and speaker age estimation. With regard to speaker identification, Reich and Duke (1979) studied the speech produced by 40 speakers reading a set sentences in five different speaking modes other than their natural voice: *elderly*, *hoarse*, *nasal*, *slow rate* and *freely disguised* voice. Spectrogram inspections were excluded from the study in order to evaluate more closely the effect of performing the speaker identification only by listening. Two groups of listeners participated in the experiment, namely, *expert* and *näive*. The results indicated that performance of both groups was affected by the presence of disguise. Based on the listeners' performance, speaker identification accuracy for the normal voice was 92% , which was degraded to 59-81% depending on the type of disguise.

Zhang (2012) included a perceptual speaker verification experiment that involved 10 listeners, five of whom knew the speakers (familiar listener group). In the case of voice disguise compared to natural speech, the identification rate in both listener groups (familiar and unfamiliar) was degraded, particularly for raised $F0$. However, the listeners' results were only slightly degraded for lowered $F0$ disguise.

Amin et al. (2014) found that the newly developed objective metric for detecting voice disguise had a large correlation with the results obtained in their perceptual test. The listeners detected disguised voices 56% of the time, which is better than by chance. It is important to note that the speakers in this study were not asked to avoid disguise detection, which gives the listeners' results a lower bound on the speakers' ability to deceive human listeners.

In the task of native dialect detection (Leemann and Kolly, 2015), the perceptual experiment indicated that Bern German listeners detected Bern German speakers 93% of the time for natural speech. However, in the disguised condition, Zurich German speakers were accepted as Bern speakers 40% of the time. The study suggested that imitating a dialect and being accepted as a native speaker by native listeners of that dialect is a challenging task.

The effects of voice disguise in age estimation by listeners was studied earlier by Lass et al. (1982) and was extended by Skoog Waller and Eriksson (2016). Vocal age disguise affected the listeners' performance by a perceived age change of three years, rather than the intended 20 years. The aim of the study contrasts with the present study in which speaker modification is aimed at concealing the speakers' normal voice in order to avoid being identified.

2.2. Age-related voice changes

Several studies investigate the ageing process and its effects on the speaker voice characteristics (Dellwo et al., 2007; Schötz, 2007; Rhodes, 2012). The variations in speech caused by age can be largely attributed to physiological and anatomical changes. These changes are most obvious from childhood to adulthood as the speech production organs grow in size. However, voice changes

continue with increasing age (Harrington et al., 2007). Although the size of the vocal tract remains relatively stable, physical changes occur to the muscles (Dellwo et al., 2007), motor control, and cognitive-linguistic ability (Torre III and Barlow, 2009). The speech of older adults is often characterized by a slow speaking rate, which can be related to reduced cognitive processing and movement of articulators (Torre III and Barlow, 2009; Schötz, 2007; Skoog Waller et al., 2015), such as tongue, jaw, lips, soft palate and larynx. Moreover, the respiratory system changes with increasing age, which is manifested in its effects on breathing and subsequently on the voice. This can also be explained by a decreased lung capacity, the weakening of the muscles involved in breathing, and the stiffness of the thorax (Schötz, 2007), which results from ageing. The changes to the larynx after puberty vary, and affect the fundamental frequency and voice quality (Schötz, 2007; Dellwo et al., 2007). The larynx settings, the degree of adduction and the tension of the vocal folds, combined with sub-glottal pressure, cause speaker variations (Dellwo et al., 2007). In general, *muscle atrophy* is an effect of ageing. Similarly, the vocal folds experience degeneration and atrophy (Schötz, 2007; Torre III and Barlow, 2009). Schötz (2007) explains that the vocal folds become shorter in males. The thin outer layer of tissue thickens in females over age 70, while in males it thickens until the age of 70 and then grows thinner again. Further, the vocal folds become less hydrated due to less secretion of mucous glands, particularly in older males. Finally, muscle atrophy occurs in the facial, mastication and pharyngeal muscles (Schötz, 2007). Age-related changes in the oral cavity, tongue, pharynx and soft palate are described by lose elasticity and decreased sensation (Torre III and Barlow, 2009).

These age-related changes induce changes in the acoustic characteristics of the speech, in which intra-speaker variation is seen as related to neuromotor control, while inter-speaker variations are often related to differences in the ageing process and to other health-related conditions (Torre III and Barlow, 2009), such as those caused by medication, smoking and intoxication. The $F0$, vowel formant frequencies and bandwidths, and speech rate characteristics have been studied to analyze their changes in relation to ageing. The $F0$ of the voice changes throughout adulthood and several studies describe the drop of $F0$ with increasing age (Endres et al., 1971; Harrington et al., 2007; Torre III and Barlow, 2009). With respect to sex differences, the size of the larynx differs between female and male speakers, which means that the $F0$ also differs. Endres et al. (1971) found that the $F0$ distribution becomes narrower with increasing age, indicating that speakers may lose some of their ability to vary their $F0$. Skoog Waller and Eriksson (2016) found the mean $F0$ of modal voices was the same for young females aged 20 to 25 and 40 to 45 but that it was lower for those aged 60 to 65. This was also confirmed in their experiments of age-related disguise. In the case of males, they found that the peak of $F0$ appears at ages 40 to 45. Other age-related studies are mostly longitudinal and report a lowering of the $F0$ for females and males (Harrington et al., 2007). For female speakers, the drop can be significant.

Formants correspond to the resonance frequencies of the vocal tract and differ according to its configuration for the articulation of different voiced sounds,

mostly vowels (Torre III and Barlow, 2009). The first three formants, $F1$, $F2$ and $F3$, are typically evaluated to compare different vowel sounds. An early study (Endres et al., 1971) reported that formants move towards lower frequencies with increasing age. According to a longitudinal study by Harrington et al. (2007), the speakers had lower $F0$ and $F1$, a marginally lower $F2$, and a constant or sometimes higher $F3$ in their later recordings, indicating a shift in the speaker's vowel space. Most studies on age-related changes to formants focus on the production of vowels. A common finding is the lowering of vowel formants which is associated with vowel centralization (Torre III and Barlow, 2009), although the effect is not always seen in all vowels. However, there seems to be no agreement in the formant changes with respect to female and male speakers increasing age (Torre III and Barlow, 2009; Schötz, 2007).

Other acoustic parameters of the voice have been studied in age-related studies, including speaking rate (Skoog Waller and Eriksson, 2016), voice onset time (Torre III and Barlow, 2009), and shimmer (Skoog Waller et al., 2015). However, $F0$ and formant frequencies are the most studied parameters in the studies involving both biological and perceived age. These are considered the primary voice parameters that a listener might focus on to estimate the speakers age, although there is no detailed evidence of how this is accomplished (Skoog Waller et al., 2015; Schötz, 2007). According to Skoog Waller et al. (2015), the age of young speakers is often overestimated, while the age of older speakers is often underestimated.

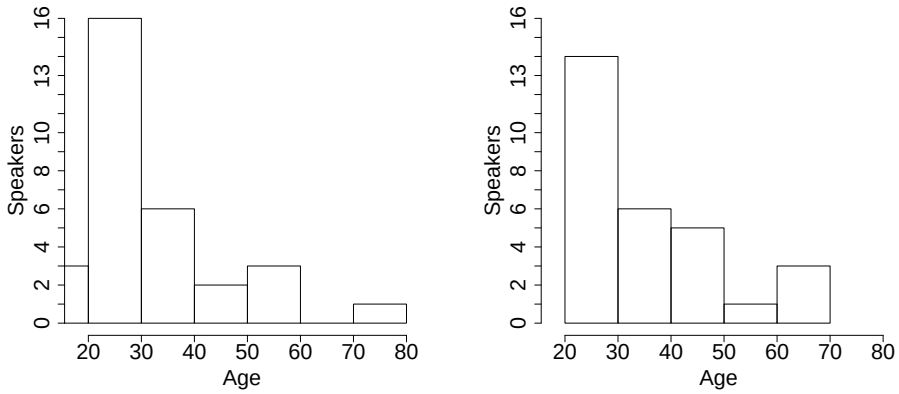
In summary, the impact of age-related voice changes on the various acoustic parameters has been well studied in previous literature. In accordance with the most commonly studied acoustic parameters, we focus on $F0$ and formants in the hope that they may reveal certain aspects of the voice disguise strategies implemented by our speakers.

3. Experimental data

The data collected for our study was first introduced in González Hautamäki et al. (2016). It consists of voice disguise as the *only* intentional modification of the speakers' voices, as opposed to modifications that would involve measures such as physically obstructing one's mouth or nostrils or the use of electronic (software or hardware) voice modifications as discussed by Rodman and Powell (2000). The main instruction given to the participants was *to modify their voices to sound old (imitating an old person) or young (imitating a child's voice)*. The speech data for all the speakers was collected under controlled conditions in the same silent office environment. The participants were all native Finnish speakers and the corpus consisted of reading sentences.

The rationale for asking our speakers to modify their "age" was two-fold. Firstly, rather than giving the speakers a completely free hand (e.g. as in Kajarekar et al. (2006)), we kept the set-up more constrained and comparable across the speakers. Although, the participants were likely to have different interpretations of how and old and young voices sounded, we assumed a certain shared knowledge across the participants, such as younger speakers tending to

have a higher pitch, allowing the possibility of observing speaker-independent disguise strategies. Secondly, rather than specifying that the participants modify their voices in terms of specific physiological parameters, such as pitch or voice harshness, the task was designed to be broader, accessible and intuitive to laymen. Although, the task and the text material was constrained, the speakers had the freedom to interpret how to modify their voices in order to sound older or younger. Overall, we found this recruitment strategy to be successful as our speakers had varied backgrounds with respect to occupation, age, social class, and expertise in voice acting.



(a) Female speakers

(b) Male speakers

Figure 2: Age distribution of speakers in the disguised speech corpus.

A total of 60 speakers participated in the data collection, including 31 females and 29 males, with an age range from 18 to 73 years. Figure 2 shows the age distribution of the speakers. The speakers also self-reported the following information: English proficiency, other known languages, profession, educational level, place of birth, place of residence during elementary education, dialect, experience in voice modification, smoking habits and other freely-worded information that could affect their voice quality and performance of the tasks. All the participants were adults (18+ years old), signed a written consent form to allow the use of their data for research purposes and were rewarded with movie tickets.

Two sessions were recorded per speaker on two different days separated by an average of five days. The recordings had a sampling rate of 44.1 kHz and 32 bits precision. The audio was collected using a portable audio recorder (Zoom H6 Handy Recorder) with an omnidirectional headset microphone (Glottal Enterprises M80), it was also connected to an electroglottograph (EG2-PCX2) in order to record glottal activity in addition to the acoustic microphone data. Moreover, a parallel recording was carried out by voice recording applications

on two smartphones: a Nokia Lumia 635 and a Samsung Galaxy Trend 2. This study focuses on the fundamental question of the extent of within-speaker variation induced by deliberate change in ones voice production for the purpose of disguise, rather than on the technological challenges induced by low-quality smartphone recordings. It therefore only considers the close-talking microphone speech, which has the highest recording quality. Interested readers are pointed to our earlier study (González Hautamäki et al., 2016) in which we analyzed the effect of smart-phone recordings on the accuracy of automatic speaker recognition. The recording set-up is illustrated in Figure 3.

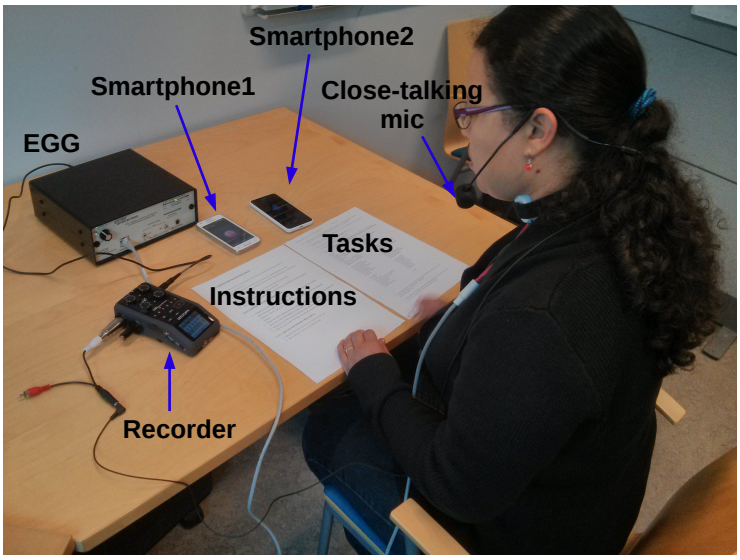


Figure 3: Set-up for the disguised data collection. We simultaneously recorded three acoustic channels (head-mounted close-talking microphone and two smartphones), together with electroglottograph (EGG) recordings of glottal activity. The participants recorded two sessions.

Each participant performed three different tasks per session. The first consisted of reading in the speaker's natural voice without any intentional modification, while the second and third tasks involved modifying one's voice to sound like an old person and a young person (e.g. a child). The read material consisted of two phonetically balanced texts, with a total of 11 sentences in Finnish and two sentences in English, as illustrated in Figure 4. The text material included the Finnish version of the "The Rainbow Passage" and "The North Wind and the Sun" (See Appendix A), plus two TIMIT sentences (Garofolo et al., 1993), SA1 and SA2, in English: "She had your dark suit in greasy wash water all year" and "Don't ask me to carry an oily rag like that".

Each session was recorded in a long audio file without interruptions and manual segmentation was conducted to produce 39 segments per session (13 sentences $\times$ 3 tasks). The segmentation process consisted of manually annotating the beginning and end time stamps of each task and sentence in seconds.

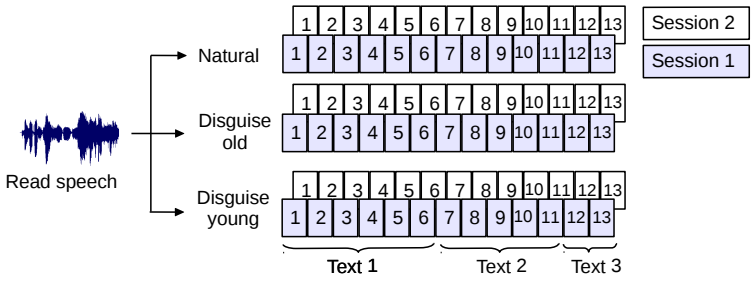


Figure 4: Diagram of the speech collected for this study. The blocks represent segments (sentences) in the text, the Finnish version of "The Rainbow Passage" (Text 1), "The North Wind and the Sun" (Text 2) and two TIMIT sentences in English (Text 3). The details of the sentences are provided in Appendix A.

This annotation was then used to cut the long recordings into sentence long segments. As is common in speaker verification studies, the data was downsampled to 8 kHz to match the sampling rate of our development data. This enabled us to benefit from the use of existing corpora for background modeling and the other necessary steps in setting up our ASV systems.

4. Acoustic analysis of the test material

To analyze the impact of voice disguise, we carried out an acoustical analysis using our test material. We studied the changes implied by voice disguise in $F0$ and formant frequencies $F1$ to $F4$. As mentioned above, these speech characteristics are also affected by biological ageing, which means that the speakers may attempt to produce a certain perceived age by modifying these primary voice parameters.

4.1. Fundamental frequency

We extracted the $F0$ from each utterance in our data using an autocorrelation method (Boersma, 1993) implementation of the Praat software (Boersma and Weenink, 2015). The $F0$ was extracted at 10ms intervals. Given that we had both male and female speakers, the frequency range was set for male speakers between 75 and 400 Hz and for female speakers between 100 and 600 Hz[1].

[1] One important factor in $F0$ estimation is to set the correct range settings. Initially, we experimented with 75 – 200 Hz for men and 100 – 300 Hz for female where the $F0$ values are set to typical values when analyzing modal speech. Such range settings are problematic for the young voice disguise because speakers tend to increase the perceived pitch to higher frequencies above the expected values. In the case of $F0$ range settings 75 – 400 Hz for male and 100 – 600 Hz for female, we found approx. 5 % error in $F0$ estimates. These errors were estimated using randomly selected five female and five male speakers from two sentences per voice type for a total of 60 speech samples.

The mean $F0$ value was taken as a scalar summary of each utterance.

The variation of the mean $F0$ for the modified voices in relation to the speaker's natural voice is defined as follows:

$$\text{Relative change} = \frac{F0_{\text{disguise}} - F0_{\text{natural}}}{F0_{\text{natural}}} \times 100\%, \tag{1}$$

where $F0_{\text{disguise}}$ refers to the average $F0$ of either old or young voice disguise for a specific utterance in Hz. We compute (1) for each utterance (`S1-S13`) for all the 60 speakers and both types of disguise. Figure 5 presents a positive relative change in the $F0$ for young voice disguise for all age groups in both sexes. Considering the old voice disguise, the results are more mixed. The extent of change is generally lower, but it is still neutral or increasing for most of the speakers. For a few female speakers, however, the change is negative for old voice disguise, which indicates that the $F0$ of the disguised voice decreased in comparison to the $F0$ of their modal voices. For 12 female speakers, 11 of whom were under 40 years of age, the change was positive for the intended old voice. In the case of the male speakers, the tendency for the majority of the speakers was to increase the $F0$, while there was no changes for the rest of the speakers. This was observed equally in both the younger and older age groups.

4.2. Formant frequencies

We analyzed the effect of the first four formant frequencies, $F1$ to $F4$, for the case of disguised data. Most of the studies on intra-speaker variation of formants analyze formant changes in isolated, selected vowels (e.g. Amin et al. (2014); Endres et al. (1971); Leemann and Kolly (2015)). In our case, we rather investigated the changes at the utterance level between the speaker's natural voice and the corresponding disguised voices. We extracted the formant frequencies from the voiced frames with Praat that uses Burg algorithm (Childers, 1978) to compute the linear prediction (LP) coefficients used for formant extraction. The formants were extracted at 10ms intervals with a maximum formant frequency set at 5 kHz.

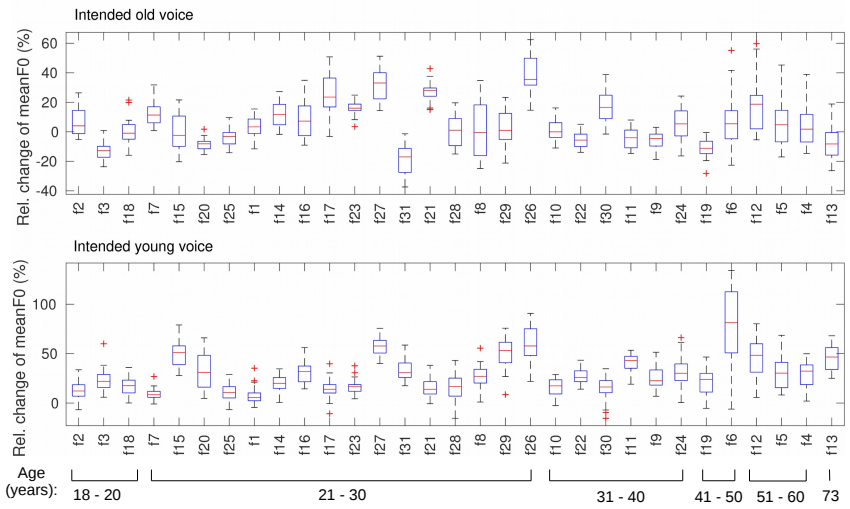
The exact estimation of formant frequencies is known to be challenging, even from recordings in controlled conditions. A number of factors contribute to formant estimation errors. Higher formant frequencies are sensitive to wrong estimates and are susceptible to error-propagation (Xia and Espy-Wilson, 2000) as they depend on the estimate of $F1$ (Singh et al., 2016; Xia and Espy-Wilson, 2000). Some of the known errors in the estimation of $F1$ are related to breathy, nasal or high pitched voices. A common technique for dealing with formant error estimations is to smooth the adjacent frame estimates, or to define the range for which a value of the formant is expected and then eliminate the outlier values. In our analysis, we used all the values extracted for each formant as higher frequencies could also contain important information concerning the way speakers articulated the changes to their voices in the disguise attempts. Therefore, before computing the mean value of $F1$ to $F3$ for each utterance

($F4$, the highest formant, was used as it was), we fitted a bi-Gaussian model to each formant's distribution. This considered the higher and lower frequencies that could otherwise have been considered outside the range of values for the formant value ($F1$ to $F3$). After fitting a bi-Gaussian model to the formant measurements of each utterance, which is detailed in Appendix B, the mean of the lowest component was selected as the representative formant mean of the utterance.

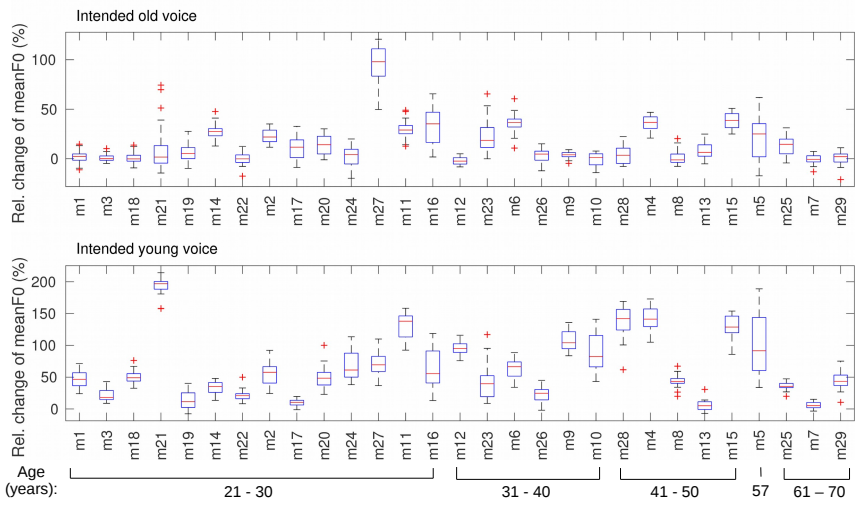
Similarly to the analysis of the $F0$, the mean formant value for each utterance was used to compare the change within the speaker's renditions of the same sentence. The differences were calculated between each naturally produced utterance and its corresponding two disguised cases (old disguise and young disguise). The difference was then reported for each formant frequency across the utterances and their respective rendition in the disguised voice, assigning a value of 1 if the formant increased with respect to the natural voice; -1 if the formant value decreased; or 0 if the difference was not statistically significant. In this way, each utterance was represented by a 4-dimensional average *formant direction change* vector that represented the relative change in the $F1$ to $F4$ estimations. For example, for a given young disguise utterance, the vector $[0\ 1\ 1\ -1]$ indicates no change in $F1$, an increase in $F2$ and $F3$, and a decrease in $F4$, all defined relative to the same but naturally-produced sentence of the same speaker. The difference between the mean formant frequencies was calculated separately for each formant frequency, using the standard deviation of the mean differences of the utterances in the compared condition (See Table C.8 in Appendix C). For a given utterance, if the mean formant difference was above the mentioned values, the formant change was included in the descriptor vector. If not, it was considered that the formant did not show a significant difference.

All the 377 utterances for male speakers and 403 utterances for female speakers were analyzed with respect to their old and young disguise attempts. The occurrences of the formant change patterns were counted in order to identify the most common types of formant variations when the speaker modified his or her voice. Figures 6 and 7 display the 15 most frequently occurring patterns for each speaker sex and disguise condition. The most common variation pattern for both sexes was $[0\ 0\ 0\ 0]$, indicating no statistically significant variation in $F1$ to $F4$. This specific pattern comprises 29% of the male speakers' utterances and 30% of the utterances by female speakers. This indicates that the speakers were able to effect a significant change in at least one of the mean formants studied in the rest of the utterances.

The top patterns of the female speakers exhibited a change in at least one of the formant values. There were more increases in mean formant differences for the young disguise condition, while the old disguise had more decreases in some of the mean formants differences. The increases and decreases in the mean formant differences of the male speakers were more scarce than those of the female speakers, and appeared evenly in the old and young disguise.

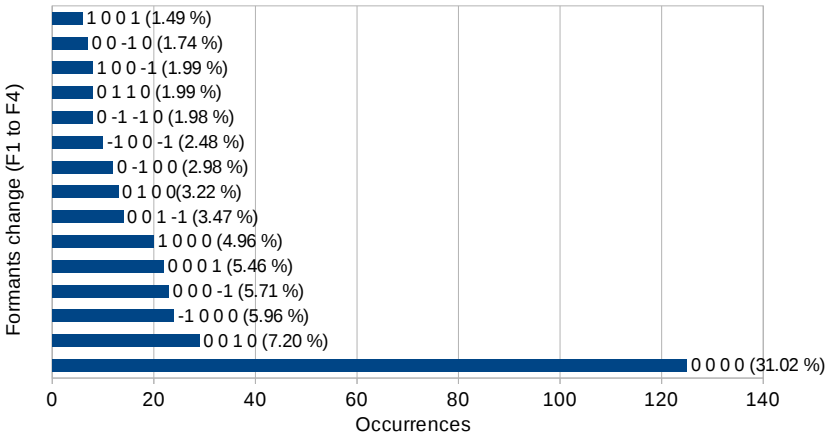


(a) Female speakers

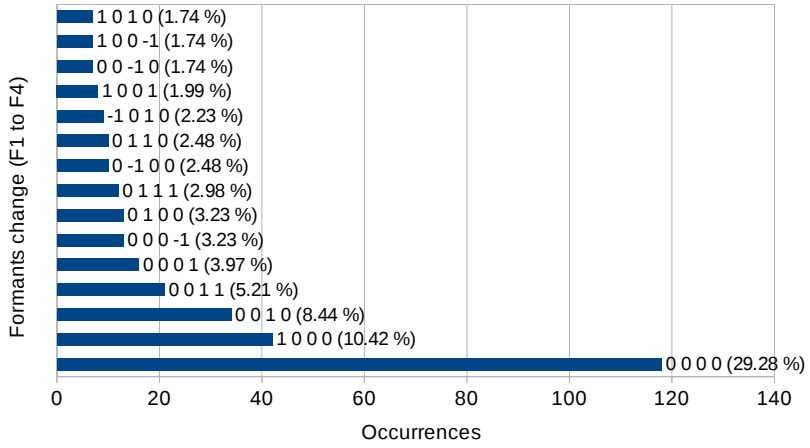


(b) Male speakers

Figure 5: Plot of the relative change in $F0$ between the speakers natural voices and the corresponding utterances with the disguised voices (intended old and young). The speakers are ordered by age in ascending order and the brackets indicate the speakers' age group. The x-axis indicates the speakers label.

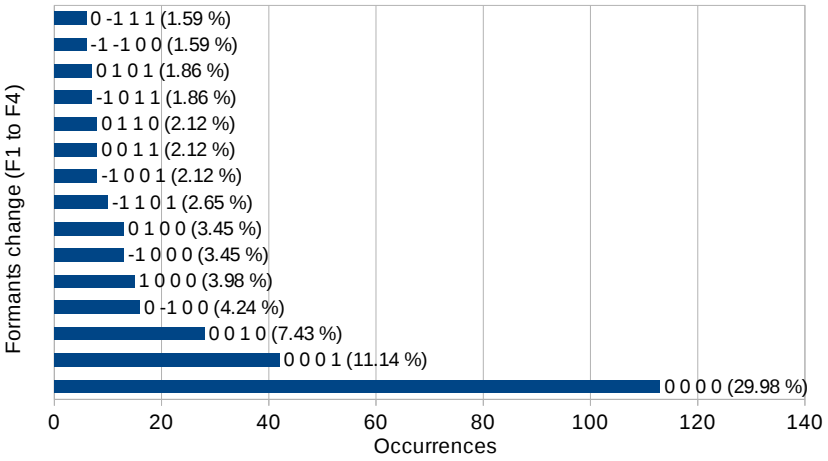


(a) Natural vs. old voice disguise

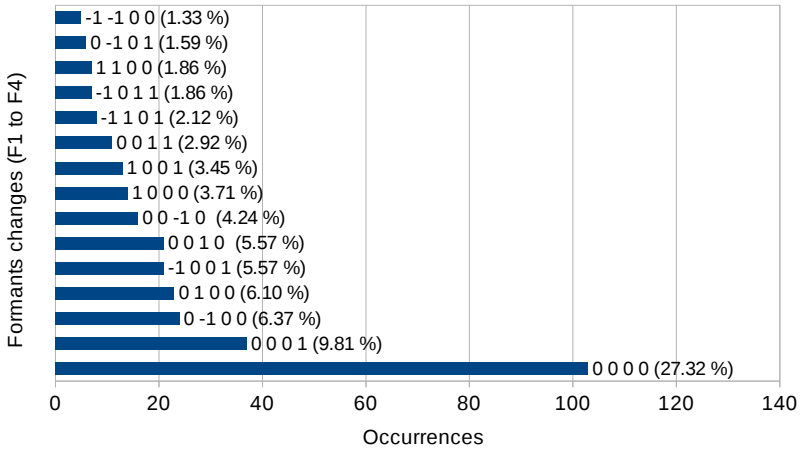


(b) Natural vs. young voice disguise

Figure 6: List of top formant changes between natural and disguised voices for female speakers in this study. The percentage indicates the amount of utterance pairs that exhibit that pattern. Formant pattern $[F1\ F2\ F3\ F4]$ notation: 0 No variation, 1 increase and -1 decrease.



(a) Natural vs. old voice disguise



(b) Natural vs. young voice disguise

Figure 7: Same as Figure 6 for male speakers. List of top formant changes between natural and disguised voices for male speakers in this study. The percentage indicates the amount of utterance pairs that exhibit that pattern. Formant pattern [$F1$ $F2$ $F3$ $F4$] notation: 0 No variation, 1 increase and -1 decrease.

5. Perceptual speaker verification experiment

We have conducted a perceptual experiment in order to evaluate the performance of the listeners. This section details the experimental design and test results.

5.1. Test set-up

Table 2: Performance in terms of equal error rate (EER,%) for *Gaussian mixture model with universal background model* (GMM-UBM) (`Systems 1-2`) and *i-vector* (`Systems 3-6`) systems for female and male speakers with natural voice and two disguised voices: Old and Young. Selected results from González Hautamäki et al. (2016).

		Natural	**Disguise old**	**Disguise young**
Female	System1	10.13	28.45	37.63
	System2	6.88	25.41	35.45
	System3	**5.05**	**24.38**	**31.68**
	System4	7.13	27.71	34.98
	System5	6.92	25.63	33.90
	System6	10.38	29.28	37.65
Male	System1	4.48	21.66	31.40
	System2	4.08	20.55	30.57
	System3	2.82	**19.45**	**30.10**
	System4	3.27	19.84	31.66
	System5	**2.71**	20.79	31.19
	System6	5.14	23.83	35.00

We collected our listeners' responses using a web-based form with 24 pairs of speech samples. The trial selection contained the same number of genuine and impostor trials for both sexes. Given that the listeners cannot evaluate all the possible available trials, we took advantage of the automatic speaker verification (ASV) system performance results reported in González Hautamäki et al. (2016) and included in Table2. The scores produced by the automatic system were used to select a small subset of trials according to their difficulty level: *easy*, *intermediate* and *difficult*. This was achieved by separating the scores from all the ASV systems into same speaker and different speaker distributions, and ranking the trials according to the sum of the scores from the ASV systems. 12 trials were selected , of which six corresponded to different speaker and six to same speaker trials. To maintain the same active speech levels, all the selected speech samples were normalized using the *activlev* function provided in the VOICEBOX speech processing toolbox (Brookes, 2006). Table 3 presents a description of the selected trials. For readability, the trials are grouped here according to the difficulty category, but during the experiment, the trial order was randomized for each listener.

Table 3: Description of the 24 trials selected for the listening test. The trial category (easy, intermediate and difficult) was based on the ASV systems' output scores for target and non-target trials. The trials were further defined by the type of voice samples: both samples had natural voice (N-N), natural vs. old voice (N-O), natural vs. young voice (N-Y). The English language trials are marked with *.

Trial	Trial sex	Category	Trial type (N: natural, O: old, Y: young)	
1	F	Easy	Target	N – N
2	F	Easy	Target	N – N
3	F	Easy	Non-target	Y – N
4	F	Easy	Non-target	N – Y
5	M	Easy	Target	N – N
6	M	Easy	Target	N – N
7	M	Easy	Non-target	O – N
8	M	Easy	Non-target	N – N *
9	F	Intermediate	Target	N – Y
10	F	Intermediate	Target	N - O
11	F	Intermediate	Non-target	Y – N
12	F	Intermediate	Non-target	O – N
13	M	Intermediate	Target	N – Y
14	M	Intermediate	Target	N - O
15	M	Intermediate	Non-target	Y – N
16	M	Intermediate	Non-target	O – N
17	F	Difficult	Target	N – O *
18	F	Difficult	Target	Y – N *
19	F	Difficult	Non-target	N – N
20	F	Difficult	Non-target	N – N
21	M	Difficult	Target	Y – N
22	M	Difficult	Target	N – O *
23	M	Difficult	Non-target	Y – N
24	M	Difficult	Non-target	N – Y

The majority of the participants were *naïve* listeners as no formal training in voice comparison was required. A total of 70 listeners participated in the experiment, including 44 males and 26 females, with an age range from 19 to 63 years old. The experiment took between 15 and 20 minutes on average. The listeners could participate in two different ways. Firstly, the test could be performed in a silent office environment with a set-up prepared by the experimenter, including a desktop computer with an integrated sound card and Sennheiser HD570 headphones. Secondly, the test was also made available online for invited participants. These online listeners needed a computer connected to the Internet, and speakers or headphones, preferably in a silent environment. A majority of 46 of the total 70 listeners performed the experiment online.

Although the majority of the speech material was in Finnish, the experiment was open to all participants regardless of their knowledge of the Finnish language. Of the 70 participants, 32 were native Finnish speakers. The rest of the participants' self-reported proficiency in Finnish varied from none (no knowledge of the language) to intermediate level. The listeners reported their Finnish

and English proficiency using a 5-point scale: *none, beginner, intermediate, advanced* and *native*. The reason for including non-native Finnish listeners was to study whether knowledge of the language plays a role in voice comparison under voice disguise. In addition to their age and sex, the listeners reported their nationality, Finnish skills, English language skills, the presence or absence of hearing problems, their practice of musical instruments, musical training, hobbies related to high-fidelity audio and sound, and work or studies related to language sciences.

5.2. Test results

The listeners compared two speech samples and decided whether they corresponded to same speaker or different speakers. The listeners were *not* informed of the presence of voice disguise in the samples and they could listen to each sample pair as many times as they wanted to. The small number of trials allowed a trial-by-trial analysis of the results: Tables 4 (native listeners) and 5 (non-native listeners) indicate the listeners' decisions for each of the trials, with their errors highlighted.

Considering all the 70 listeners, the average listener made 8.23 errors out of 24. By contrast, the listener *panel*, formed by combining the individual listener's results using the majority vote, made eight errors. The best listeners made only four errors, and correspond to the following listeners: A listener from the non-native group (Listener 1), and Listeners 24, 29 and 32 from the native group. The listeners who made the most errors (13) were both from the non-native group (Listeners 16 and 35).

As expected, most of the errors occurred in the intermediate and difficult trial categories. The easy trials had consistently fewer errors, and no listener made errors in two of the easy trials (Trials 5 and 6). The trials with zero or two errors corresponded to same speaker trials with the speakers' samples in their natural voices (Trials 1, 5, 6). Trial 15, with an intermediate level of difficulty, had only seven errors (out of the 70 listeners). This trial corresponded to a different speaker trial and included one speaker with young voice disguise.

Trials deemed difficult by our ASV systems also had the largest number of listener errors. The trials with the most errors were 12, 14, 19, 20 and 22. They included old voice disguise, which more than half of the listeners classified incorrectly, which also occurred for Trials 13, 18 and 21 that contained young voice disguise. With the exception of Trials 15 and 16 in the intermediate category and Trial 23 in the difficult one, the number of listener errors increased according to the trial's difficulty level and the inclusion of disguised voices.

Three of the trials with English sentences belonged to the difficult trial (Trials 17, 18 and 22). For the automatic system in addition to the disguise task, the speaker variations produced by the effect of the foreign accent reduced the ASV performances. However, it was not conclusive whether this was the case for the listeners as only four trials included English language data.

Table 4: Native Finnish listeners trial-by-trial decisions. 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.*

					Native Finnish Listeners																																	Trial errors
Category	**Sex**	**Type**		**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**	
Easy	**Female**	**Target**	N – N	**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	1	**0**
			N – N	**2**	1	1	1	1	1	1	1	1	1	1	1	1	1	4	1	1	4	1	1	1	1	1	1	1	1	2	1	5	1	1	1	**3**
		Non-target	Y – N	**3**	5	5	5	5	3	5	4	5	5	5	5	5	5	5	5	5	2	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	**2**
			N – Y	**4**	4	5	5	5	5	2	2	5	5	5	2	5	5	5	5	5	5	5	5	5	5	5	5	5	2	5	4	5	5	5	5	5	**4**
	Male	**Target**	N – N	**5**	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	1	1	1	1	**0**
			N – N	**6**	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	1	1	1	1	**0**
		Non-target	O – N	**7**	5	4	5	4	5	5	4	5	5	5	5	4	1	1	5	5	5	4	1	5	5	5	4	5	5	5	5	5	3	5	5		**4**
			N – N *	**8**	4	5	5	4	5	5	5	5	5	5	5	2	5	5	5	5	5	5	1	5	5	5	5	5	5	5	5	5	5	5	4	5	**2**
Intermediate	**Female**	**Target**	N – Y	**9**	1	5	5	4	5	1	4	5	2	5	4	4	5	1	4	2	1	2	1	1	1	5	4	5	1	2	2	2	2	1	2	5	**15**
			N – O	**10**	4	5	5	2	5	5	4	4	4	5	5	5	5	1	5	4	1	4	1	1	1	1	1	1	3	1	2	5	2	2	4	1	**18**
		Non-target	Y – N	**11**	2	5	5	4	2	5	2	2	4	5	5	4	5	1	5	5	1	1	1	1	1	4	5	1	2	5	3	5	5	5	4	2	**14**
			O – N	**12**	4	5	5	4	1	1	2	1	5	1	2	2	5	5	5	4	1	2	1	1	1	1	5	5	1	2	5	1	5	5	5	1	**17**
	Male	**Target**	N – Y	**13**	2	5	5	2	3	1	1	4	5	1	5	3	5	1	5	2	1	1	1	1	5	1	1	1	2	1	3	3	5	4	5	1	**15**
			N – O	**14**	5	5	5	4	1	5	4	5	5	1	5	2	1	5	5	5	5	4	1	1	4	5	5	1	5	1	2	5	2	5	5	1	**21**
		Non-target	Y – N	**15**	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	1	5	5	3	5	5	5	5	5	5	5	5	5	5	**2**
			O – N	**16**	2	4	5	4	1	4	5	5	5	1	5	2	5	5	5	3	5	2	5	5	1	5	1	5	5	2	3	3	4	5	5	5	**11**
Difficult	**Female**	**Target**	N – O *	**17**	1	5	5	1	5	3	2	4	5	1	4	4	3	1	4	2	1	2	5	1	1	3	1	1	2	5	4	2	2	3	4	1	**16**
			Y – N *	**18**	4	5	5	4	1	1	4	2	5	1	2	5	5	1	5	5	1	5	1	1	5	5	1	4	2	1	2	4	2	1	5	1	**16**
		Non-target	N – N	**19**	4	4	5	2	1	1	2	1	1	5	5	1	2	5	5	2	1	4	1	1	5	1	1	5	1	1	3	1	1	1	1	5	**21**
			N – N	**20**	2	1	5	1	1	1	1	5	5	5	1	1	1	1	5	2	1	2	1	1	1	1	4	5	1	4	4	1	2	1	5	5	**21**
	Male	**Target**	Y – N	**21**	1	4	5	4	1	1	5	2	3	1	5	5	5	1	5	4	5	3	5	5	1	1	5	5	1	2	4	5	2	4	5	1	**20**
			N – O *	**22**	4	5	5	4	1	1	4	4	3	5	5	2	5	1	5	2	5	2	1	1	5	5	4	2	4	2	4	1	1	5	5	1	**19**
		Non-target	Y – N	**23**	2	5	5	5	5	5	4	4	5	5	5	5	5	1	5	5	5	5	1	5	5	5	2	5	5	5	4	5	5	5	5	1	**5**
			N – Y	**24**	2	5	5	5	1	5	4	5	5	5	5	5	5	1	5	2	5	5	3	5	5	5	5	5	5	2	4	5	5	5	5	5	**6**
				Errors	**9**	**9**	**8**	**7**	**11**	**7**	**11**	**9**	**8**	**5**	**10**	**11**	**10**	**6**	**9**	**8**	**7**	**10**	**11**	**5**	**8**	**9**	**7**	**4**	**8**	**5**	**7**	**9**	**4**	**8**	**8**	**4**	

Misses
False accepts
Undecided

Table 5: Non-native Finnish listeners trial-by-trial decisions. 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.*

Trial					Non native Finnish Listeners																																												Trial errors
Category	Gender	Type		No.	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	35	36	37	38	
Easy	Female	Target	N – N	1	1	1	1	1	1	1	1	1	1	1	1	3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	4	1	1	1	1	1	1	2
			N – N	2	1	1	1	1	1	1	5	1	1	1	4	1	2	2	4	2	1	1	1	2	4	1	1	2	1	1	1	4	4	2	5	5	1	1	1	1	1	8
		Non-target	Y – N	3	5	4	5	5	5	4	5	5	4	5	5	2	5	4	5	5	5	5	4	4	5	5	5	5	5	5	1	5	5	5	5	5	5	5	5	5	5	2
			N – Y	4	5	4	5	5	5	4	5	5	3	5	4	5	5	5	5	5	5	5	3	2	5	5	5	5	4	5	5	5	5	5	5	5	3	5	5	5	5	4
	Male	Target	N – N	5	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	1	1	1	1	1	1	1	1	1	0
			N – N	6	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	2	2	1	2	1	1	1	1	1	1	1	2	1	1	1	1	0
		Non-target	O – N	7	5	5	5	4	5	5	5	5	5	5	5	3	5	4	5	5	5	4	5	5	5	5	5	5	4	5	5	5	4	4	5	5	1	2	3	5	5	4
			N – N *	8	5	2	1	5	5	5	5	5	5	5	5	2	5	5	5	5	4	5	5	2	5	2	4	5	5	5	5	5	5	5	2	5	5	5	5	1	5	5	7
Intermediate	Female	Target	N – Y	9	4	1	4	1	2	1	5	2	1	1	2	2	4	2	2	4	2	5	2	2	4	5	3	2	1	5	2	5	5	5	1	5	1	2	1	1	5	1	15
			N – O	10	1	2	2	1	3	1	4	2	1	5	1	2	2	2	5	5	4	4	1	2	2	3	4	3	2	2	5	2	5	1	5	1	5	2	2	1	5	1	15
		Non-target	Y – N	11	5	5	5	2	4	5	5	5	5	2	5	2	5	3	5	5	5	5	5	4	5	5	5	5	2	5	5	1	2	4	5	5	4	5	3	5	5	5	8
			O – N	12	4	2	4	2	4	1	5	5	1	2	5	5	5	2	2	2	5	5	3	4	2	4	5	2	2	5	5	1	1	2	4	1	1	5	2	1	5	1	20
	Male	Target	N – Y	13	1	2	5	1	1	5	2	2	1	5	5	3	5	2	4	5	5	5	2	5	2	5	4	5	1	5	5	1	5	1	5	1	1	3	5	1	5	1	21
			N – O	14	5	4	5	2	5	5	5	5	3	5	5	2	5	3	5	5	5	1	3	4	4	5	5	5	2	5	2	1	2	5	5	5	4	5	5	5	5	1	30
		Non-target	Y – N	15	4	5	5	5	5	4	5	4	3	5	5	3	5	5	5	5	5	5	5	5	5	4	5	1	3	5	5	5	4	5	4	5	1	5	5	5	5	5	5
			O – N	16	5	4	5	5	5	5	5	5	3	5	5	4	5	2	5	5	4	5	5	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	3	5	5	5	3
Difficult	Female	Target	N – O *	17	2	5	2	1	5	1	5	2	1	2	2	3	5	2	5	4	5	2	1	5	2	5	5	1	2	5	5	1	4	1	4	1	1	5	2	1	5	1	17
			Y – N *	18	5	4	1	2	4	1	5	4	1	5	5	4	5	2	5	5	2	1	1	5	2	5	5	2	1	2	5	5	5	2	4	1	2	5	3	1	5	1	21
		Non-target	N – N	19	5	2	1	1	2	5	5	2	4	1	2	1	1	2	1	2	2	1	1	1	4	1	1	1	1	5	2	1	5	5	2	1	5	5	1	1	5	1	27
			N – N	20	1	1	2	1	1	1	1	2	1	1	1	2	4	2	2	4	1	2	1	2	2	5	2	4	1	5	1	1	5	5	1	1	1	4	1	1	1	1	30
	Male	Target	Y – N	21	2	5	5	4	5	5	5	2	3	5	5	1	5	5	5	4	5	1	2	4	5	5	5	5	5	5	5	5	4	3	5	5	1	5	3	5	5	5	32
			N – O *	22	1	5	5	1	5	2	5	4	1	5	5	2	5	1	5	5	5	4	5	5	5	5	5	1	2	5	5	5	2	1	5	5	1	1	5	1	5	4	25
		Non-target	Y – N	23	5	5	5	2	2	4	5	2	3	4	5	5	5	5	5	3	5	4	3	5	5	2	5	4	3	5	5	2	5	5	5	5	4	3	3	5	5	4	11
			N – Y	24	4	5	4	5	5	4	5	2	3	5	2	4	5	2	5	3	5	1	1	5	4	5	3	1	1	5	5	1	2	2	3	3	1	5	3	1	5	1	19
			Errors		4	9	8	6	9	5	8	8	9	10	8	11	9	8	10	13	8	7	8	10	6	10	11	8	8	6	8	11	10	6	11	10	8	8	13	7	9	6	

Misses
False accepts
Undecided

It is worth noting that the definition of easy, intermediate and difficult trials was based solely on the ASV systems score distributions. In this sense, the point was to compare whether the listener decisions agree with the categorization of the trial difficulty as judged by the ASV systems. This appears to be the case.

5.3. Factors affecting listener performance

This subsection presents a statistical analysis of the factors affecting the participants' performance in the perceptual experiment. It approaches this by analyzing the listeners' self-reported information and their results. In addition, the trial information is considered in this analysis.

Listener information. The self-reported information was considered as a predictor of listener performance. A *generalized logistic regression model* (Baayen, 2008) was fitted to the listener information in which the correct answers per trial were used as the dependent variable. The listener's information consisted of the following variables: `age`, `sex`, `Finnish` and `English proficiency`, `practices musical instruments`, `musical training`, `high- fidelity related hobbies`, `linguistic education or work`, and `listening device used`. In addition to this information, we collected the listener's opinion concerning the level of difficulty of the test and whether it was performed online or on-site.

We found that *none* of the listeners' details had a statistically significant effect in listener performance. This may indicate that some factors that could have influenced listener performance were not considered or collected and that the model does not fit our data well.

Trial selection effect on listeners performance. In addition to the speaker information, we conducted a similar analysis for trial-specific information. The correct answers from a listener functioned as the dependent variable and the trial information were the factors for the logistic model, which can be seen in Table 6.

Table 6: Trial information defined as the predictors of the model and their corresponding factors.

Predictor	**Factors**
`Category`	Easy, intermediate, difficult
`Sex`	Female, male
`Type`	Target, non-target
`Speech`	Natural, disguise
`Voice`	Natural, old, young
`Language`	Finnish, English

Table 7 presents the statistical analysis for the factors with a significant effect on listener performance. The logistic model indicates a positive effect for the `Category: Easy` factor for both groups, which is particularly significant for the non-native listeners with p-value < 0.001. Both listener groups show a significant negative effect for the trial `Type: Target` factor. It is worth noting that the estimated coefficient for the native speakers is -0.9796 in comparison

to −1.0294 for the non-native listeners. This implies that target (same speaker) trials have a significant effect on the listeners errors, and this is slightly higher for non-native listeners. The condition of the trial, natural or disguise, has a negative effect on listener performance, particularly in the case of the disguise trials with old and young voice disguise. The speaker's sex and language of the trial did not have a significant effect on listener performance.

Table 7: Statistical analysis results for the listener groups using the correct answers as a dependent variable, and the most significant factors based on the trial information. A positive value in the estimate column signifies that a listener with a corresponding factor has a higher probability of giving a correct answer than one with the opposite factor. ***** denotes a statistically significant estimate value.

Listeners	**Factors**	**Estimate**	**Std. Error**	**z value**	**Pr($>\|z\|$)**
Native	(Intercept)	3.883	0.944	4.115	< 0.001 *
	Category: Easy	1.364	0.445	3.062	0.002 *
	Category: Intermediate	-0.006	0.310	-0.018	0.985
	Sex: Male	0.306	0.211	1.447	0.148
	Type: Target	-0.980	0.224	-4.373	< 0.001 *
	Speech: Natural	-0.961	0.486	-1.976	0.048 *
	Voice: Old	-3.392	0.778	-4.361	¡ 0.001 *
	Voice: Young	-2.826	0.842	-3.355	0.001 *
	Lang.: Finnish	-0.176	0.391	-0.450	0.652
Non-native	(Intercept)	3.068	0.814	3.769	< 0.001 *
	Category: Easy	1.932	0.415	4.655	< 0.001 *
	Category: Intermediate	0.659	0.264	2.496	0.012 *
	Sex: Male	-0.201	0.186	-1.081	0.280
	Type: Target	-1.029	0.199	-5.164	< 0.001 *
	Speech: Natural	-1.451	0.443	-3.273	0.001 *
	Voice: Old	-2.462	0.651	-3.784	< 0.001 *
	Voice: Young	-2.250	0.719	-3.13	0.002 *
	Lang.: Finnish	-0.254	0.328	-0.774	0.439

6. Discussion

This work presents a broad study into voice disguise effects with the use of acoustic and perceptual methodologies. Before concluding the study, we present an overview of the results obtained, together with our interpretation according to the research questions formulated at the end of Section 1.

Analysis of acoustic parameters.

Q1. *Is there a significant change in the $F0$ of female and male speakers when attempting voice disguise to sound older or younger? Does it increase or decrease?*

We noticed a systematic increase in the relative change of $F0$ in the case of intended young voice disguise for both sexes and for all age groups. The change was smaller for the intended old voice disguise, but it was still positive or neutral in the case of male speakers. For most of the female speakers who increased their $F0$ for intended old voice were under 40 years of age. There was no change for the rest of the speakers. For eight female speakers, the change was negative for old voice disguise, indicating that these speakers lowered their $F0$ for the disguised voice with respect to their natural voice. Four of these speakers belonged to the young age group. In general, the changes in $F0$ between both intended disguise voices varied between speakers: some implemented extreme variations but most speakers' $F0$ did not vary greatly. The length of the confidence box, per speaker, also indicates the extent of between-utterance variations. A few speakers show a small variance in their performance and maintained their $F0$ stable throughout the disguise task.

Q2. *Are there significant differences between the averages of the first four formant frequencies of natural and disguised voices of the female and male speakers?*

The disguise in vocal tract configurations was measured by means of the averaged $F1$ to $F4$ values, and we introduced a new acoustic analysis method to identify the joint changes in averaged $F1$ to $F4$ formant values. Interestingly, none of the formants changed significantly in 29% of the disguised male utterances. In the case of the female speakers, no significant change was observed in 30% of the utterances. Thus, most of the speakers of both sexes did show a significant change in at least one of the formants. Usually, several formants were jointly changed as a result of disguise. This suggests that, in most cases, the speakers not only modified their larynx settings, but also some of their articulatory configurations. This may be a disguise strategy on the part of the speakers to emulate the changes in vocal tract characteristics that are perceptually related to biological age.

Q3. *Is there any speaker-independent disguise pattern that can be associated with formant frequency variation between natural speech and the studied*

strategy for disguised speech?
With regard to the most common formant direction change patterns, we could not identify any recurring, speaker-independent pattern apart from the "no change" pattern [0 0 0 0]. There may be two possible reasons for this: firstly, the particular participant's interaction with people from different age groups may lead the speaker to have different perceptual impressions of what an imaginary "ideal" old or young voice should sound like. Secondly, even if such an auditory "ideal" would be precise, the speaker may be unable to *reproduce* it consistently. Nonetheless, certain observations were made. For example, many of the top-15 formant patterns in the young voice disguises performed by the female speakers contain 1s, indicates an increase in one or several formant values.

Perceptual speaker verification experiments.

Q4. *Is listener performance affected by the presence of voice disguise in a similar way to the performance of the ASV systems?*
Of the panel of 70 listeners, the average listener made 8.23 errors out of a possible 24, while the entire panel decision based on majority voting made eight errors. The best individual listener, made only four errors. Interestingly, one of the best four listeners was non-native. Our perceptual speaker verification and the ASV systems results were linked in that we selected the listening trials as easy, intermediate and difficult trials based on the ASV systems. The goal was to find out whether or not the listeners followed the same pattern. This was indeed found to be the case: the trials considered easy for the ASV systems were easiest for the listeners, and the trials considered difficult for the ASV systems were also difficult for the listeners. Some trials with an intermediate difficulty level had a similar or slightly lower number of errors than the difficult trials, and they can therefore also be considered as for the listeners. These results were further validated by statistical significance tests, which indicated that trials from the easy category (for ASV) were significantly easier to recognize than the other two categories for both natives and non-natives listeners, with p-values < 0.01.

Q5. *Does knowledge of the speakers' native language play a role in making more reliable perceptual speaker comparisons under modal voices and under disguise?*
To compare the listening ability of native and non-native Finnish speakers, we noted that, their performance was similar for our test data. Both groups made fewer errors in the easy trials and more errors in the difficult trials. Their task was to compare the voices and decide whether the speaker was the same or different, it appears that knowledge of what was said did not provide with an advantage in this task.

Q6. *Is there a particular trial category or listener attribute that affects listener performance in the perceptual speaker recognition task?*

We observed that in the intermediate and difficult categories the target (same speaker) trials were significantly more difficult than the non-target (different speaker) trials for both native and non-native listeners with p-value < 0.00001. Furthermore, the target trials were slightly more difficult for the non-native group than for the native listeners (logistic regression coefficient -1.0294 vs -0.9796). With respect to the listeners' information, we did not find any particular factor that affected the listeners' performance.

7. Conclusions

Verifying the identity of speakers by means of short utterances that include voluntary variations of the voice is a very challenging task for both humans and state-of-the-art automatic speaker verification systems. Therefore, it is important to investigate how speakers manipulate their voices in order to avoid identification. Our case study addressed the impact of voice disguise when the speakers attempt to sound much older or younger than their actual age. To this end, we conducted an acoustical analysis and perceptual speaker verification experiment on a newly collected disguise corpus of 60 native Finnish speakers and a panel of 70 listeners of whom 32 were native Finnish speakers and the rest non-native.

The analysis of the acoustic parameters revealed a considerable increase in mean $F0$ values for both intended young and old voice disguises. The speakers' main strategy for reproducing a stereotypical old or young voice was to increase the $F0$, although some female speakers decreased their $F0$ in attempting an old voice. In the case of male speakers, the $F0$ variations remained neutral or increased for the intended old voice. Given this change in articulation, we analyzed the variations in formant frequencies ($F1$ to $F4$) between natural speech and the disguised voices. We found that, for most of the utterances, the average formant values were changed as a result of disguise. Our results imply that speakers are able to manipulate their vocal characteristics, although the extent of these variations differs between speakers.

With regard to our perceptual speaker verification task, we found a strong correspondence between the decisions made by human and the automatic methods. The selected trials that were difficult for the ASV systems were also difficult for the human listeners, as were the easy trials. With regarding to the performance of native and non-native Finnish listeners, accuracy degraded substantially in both groups in the presence of disguised voices, and particularly in the case of same speaker trials. The non-native listeners had more errors in the different speakers' trials that included disguised voices. In summary, our experiment indicates that knowledge of the speakers' native language was not a substantial help for speaker verification in the context of the disguise set-up.

A step forward towards more robust speaker verification against voice disguise, whether performed by humans or ASV systems, would be to consider the vocal parameters that are more commonly modified by speakers avoiding identification. A system robust to disguise, or extreme vocal modifications, could

consider modeling techniques that include the vocal variation patterns presented in this study. The analysis in this study is based on read speech and the comparisons within and across speakers contained the same text, which facilitated controlled comparisons between utterance pairs. A key point for future work could be to consider spontaneous speech in which other information related to speaker characteristics could be studied. This provides further motivation for the study of vocal parameters, which may be more difficult to modify during disguise.

Our study revealed some of the challenges voice disguise poses to speaker verification by both humans and ASV systems. It also has a few limitations that provide scope for further work. Firstly, the statistical analysis of our data and the effect of the self-reported listener information are naturally limited by how well the model fits our data. The significance of the variables was studied independently, which meant that their interactions were not considered in the model outcome. Secondly, all our experiments were conducted in a clean, controlled and text-constrained set-up in order to systematically analyze the effect of voice disguise and to identify the sources of the differences in the natural and disguised voices. A further study containing disguised voices that are observed "out in the wild", including in noisy environments, telephone channels or voice-over-ip (VoIP) coding artifacts, would therefore be interesting. Given that the relative performance degradation in close-to-perfect conditions is already extremely severe, we would expect further degradations when the voice disguise effects are mixed with noise and channel nuisance factors. Thirdly, our study focuses on disguising one's voice identity by means of a modification related the speaker's perceptual age. The main concern was the collection of data in order to study the detrimental effects on the accuracy of speaker recognition, while age disguise merely served as a relatively non-constrained task across the speakers. Given that our speakers were naïve or had little or no experience with voice modification, they were not expected to produce the most convincing old or young voice imitations. Rather, they simply concealed their voices as best they could. The advantage of this form of data collection was a task that allows a similar disguise strategy but gives certain artistic freedom to the speakers who perform it. However, a more restrictive task that limits the disguise type could allow further perceptual tests. For example, it could indicate whether the age estimation of the perceived voice disguise is in accordance with the intended target age. In this context, future work could involve evaluating the level of success achieved by the speakers in the disguise attempts by means of a perceptual test.

8. Acknowledgments

This study was supported by the Academy of Finland (projects no. 253120, 283256 and 309629), the Finnish Scientific Advisory Board for Defense (MATINE) project no. 2500M-003, and Nokia Foundation. The authors would like to thank Maria Bentz and Prof. Stefan Werner of UEF for their help in contributing to planning and execution of the data collection. Finally, the authors

would like express their sincere thanks to the volunteer speakers and listeners who made this study possible.

References

Akaike, H., 1974. A new look at the statistical model identification. IEEE Transactions on Automatic Control 19, 716–723.

Amin, T., Marziliano, P., German, J., 2014. Glottal and vocal tract characteristics of voice impersonators. IEEE Transactions on Multimedia 16, 668–678.

Baayen, R.H., 2008. Analyzing Linguistic Data: A Practical Introduction to Statistics using R. Cambridge University Press.

Boersma, P., 1993. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound, in: Proc. of the Institute of Phonetic Sciences, pp. 97–110.

Boersma, P., Weenink, D., 2015. Praat: doing phonetics by computer [Computer program]. Version 5.4.09, retrieved 15 June 2015 from http://www.praat.org/.

Brookes, M., 2006. Voicebox: Speech processing toolbox for MATLAB. Software, available [January 2014] from http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html .

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

Childers, D., 1978. Modern Spectrum Analysis. IEEE Press selected reprint series, New York, IEEE Pr.

Clark, J., Foulkes, P., 2007. Identification of voices in disguised speech. The International Journal of Speech, Language and the Law 14, 195–221.

Dellwo, V., Huckvale, M., Ashby, M., 2007. How is individuality expressed in voice? an introduction to speech production and description for speaker classification, in: Müller, C. (Ed.), Speaker Classification I: Fundamentals, Features, and Methods. Springer Berlin Heidelberg, pp. 1–20.

Dempster, A., Laird, N., Rubin, D., 1977. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society. Series B 39, 1–38.

Endres, W., Bambach, W., Flösser, G., 1971. Voice spectrograms as a function of age, voice disguise, and voice imitation. The Journal of the Acoustical Society of America 49, 1842–1848.

Eriksson, A., Llamas, C., Watt, D., 2010. The disguised voice: imitating accents or speech styles and impersonating individuals. Language and identities 8, 86–96.

Garofolo, J., Lamel, L., Fisher, W., Fiscus, J., Pallett, D., Dahlgren, N., Zue, V., 1993. TIMIT acoustic-phonetic continuous speech corpus LDC93S1. Web Download. Linguistic Data Consortium, Philadelphia.

González Hautamäki, R., Kinnunen, T., Hautamäki, V., Laukkanen, A.M., 2015. Automatic versus human speaker verification: the case of voice mimicry. Speech Communication 72, 13–31.

González Hautamäki, R., Sahidullah, M., Kinnunen, T., Hautamäki, V., 2016. Age–related voice disguise and its impact in speaker verification accuracy, in: Proc. Odyssey: the Speaker and Language Recognition Workshop, pp. 277–282.

Hansen, J.H., Hasan, T., 2015. Speaker recognition by machines and humans: a tutorial review. IEEE Signal Processing Magazine 32, 74–99.

Harrington, J., Palethorpe, S., Watson, C.I., et al., 2007. Age-related changes in fundamental frequency and formants: a longitudinal study of four speakers., in: "Proc. Interspeech", pp. 2753–2756.

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

Hirson, A., Duckworth, M., 1993. Glottal fry and voice disguise: a case study in forensic phonetics. Journal of Biomedical Engineering 15, 193 – 200.

Kahn, J., Audibert, N., Rossato, S., Bonastre, J.F., 2011. Speaker verification by inexperienced and experienced listeners vs. speaker verification system, in: Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), Prague, Czech Republic. pp. 5912 – 5915.

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

Kinnunen, T., Hautamäki, V., Fränti, P., 2006. On the use of long-term average spectrum in automatic speaker recognition, in: Proc. 5th Int. Symposium on Chinese Spoken Language Processing (ISCSLP), Singapore. pp. 559–567.

Kinnunen, T., Li, H., 2010. An overview of text-independent speaker recognition: From features to supervectors. Speech Communication 52, 12–40.

Köster, O., Schiller, N.O., et al., 1997. Different influences of the native language of a listener on speaker recognition. Forensic Linguistics. The International Journal of Speech, Language and the Law 4, 11.

Künzel, H., Gonzalez-Rodriguez, J., Ortega-García, J., 2004. Effect of voice disguise on the performance of a forensic automatic speaker recognition system, in: Proc. Odyssey: the Speaker and Language Recognition Workshop.

Lass, N.J., Justice, L.A., George, B.D., Baldwin, L.M., Scherbick, K.A., Wright, D.L., 1982. Effect of vocal disguise on estimations of speakers' ages. Perceptual and Motor Skills 54, 1311–1315.

Leemann, A., Kolly, M.J., 2015. Speaker-invariant suprasegmental temporal features in normal and disguised speech. Speech Communication 75, 97–122.

López, S., Riera, P., Assaneo, M.F., Eguía, M., Sigman, M., Trevisan, M.A., 2013. Vocal caricatures reveal signatures of speaker identity. Scientific reports 3.

Mohammadi, S.H., Kain, A., 2017. An overview of voice conversion systems. Speech Communication 88, 65 – 82.

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

Perrot, P., Aversano, G., Chollet, G., 2007. Voice disguise and automatic detection: Review and perspectives, in: Stylianou, Y., Faundez-Zanuy, M., Esposito, A. (Eds.), Progress in Nonlinear Speech Processing. Springer Berlin Heidelberg, Berlin, pp. 101–117.

Ramos, D., Franco-Pedroso, J., Gonzalez-Rodriguez, J., 2011. Calibration and weight of the evidence by human listeners. the ATVS-UAM submission to NIST human-aided speaker recognition 2010, in: Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), Prague, Czech Republic. pp. 5908 – 5911.

Reich, A., Duke, J., 1979. Effects of selected vocal disguises upon speaker identification by listening. The Journal of the Acoustical Society of America 66, 1023–1028.

Rhodes, R.W., 2012. Assessing the strength of non-contemporaneous forensic speech evidence. PhD thesis. Ph.D. thesis. University of York.

Rodman, R., Powell, M., 2000. Computer recognition of speakers who disguise their voice, in: Proc. of ICSPAT.

Saeidi, R., Huhtakallio, I., Alku, P., 2016. Analysis of face mask effect on speaker recognition, in: "Proc. Interspeech", pp. 1800–1804.

San Segundo, E., Alves, H., Trinidad, M.F., 2013. CIVIL corpus: Voice quality for speaker forensic comparison. Procedia-Social and Behavioral Sciences 95, 587–593.

Schmidt-Nielsen, A., Stern, K.R., 1985. Identification of known voices as a function of familiarity and narrowband coding. The Journal of the Acoustical Society of America 77, 658–663.

Schötz, S., 2007. Acoustic analysis of adult speaker age. Speaker Classification I , 88–107.

Schwartz, R., Campbell, J.P., Shen, W., Sturim, D.E., Campbell, W.M., Richardson, F.S., Dunn, R.B., Granvill, R., 2011. USSS-MITLL 2010 human assisted speaker recognition, in: Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), Prague, Czech Republic. pp. 5904 – 5907.

Singh, R., Gencaga, D., Raj, B., 2016. Formant manipulations in voice disguise by mimicry, in: 2016 4th International Workshop on Biometrics and Forensics (IWBF), pp. 1–6.

Skoog Waller, S., Eriksson, M., 2016. Vocal age disguise: The role of fundamental frequency and speech rate and its perceived effects. Frontiers in Psychology 7.

Skoog Waller, S., Eriksson, M., Sörqvist, P., 2015. Can you hear my age? influences of speech rate and speech spontaneity on estimation of speaker age. Frontiers in psychology 6.

Stylianou, Y., 2009. Voice transformation: a survey, in: Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), Taiwan. pp. 3585–3588.

Torre III, P., Barlow, J.A., 2009. Age-related changes in acoustic characteristics of adult speech. Journal of Communication Disorders 42, 324 – 333.

Xia, K., Espy-Wilson, C.Y., 2000. A new strategy of formant tracking based on dynamic programming., in: "Proc. Interspeech", pp. 55–58.

Zhang, C., 2012. Acoustic analysis of disguised voices with raised and lowered pitch, in: Proc. Int. Symposium on Chinese Spoken Language Processing (ISCSLP), pp. 353–357.

Zhang, C., Tan, T., 2008. Voice disguise and automatic speaker recognition. Forensic Science International 175, 118–122.

Appendix A.

Text fragments read by the speakers

The rainbow passage (*Sateenkaaritarina*)

S1. Kun auringonvalo osuu sadepisaroihin ilmassa, ne käyttäytyvät kuin prismat, ja muodostavat sateenkaaren.

S2. Sateenkaari muodostuu valkoisen valon jakaantuessa useiksi kauniiksi väreiksi.

S3. Nämä muodostavat kauniin pitkän kaaren horisontin yläpuolelle päättyen jonnekin sen taakse.

S4. Legendan mukaan sateenkaaren päässä on padallinen sulaa kultaa.

S5. Ihmiset etsivät sitä kuitenkaan mitään löytämättä.

S6. Kun joku etsii jotain mahdotonta, sanotaan hänen etsivän kultaa sateenkaaren

päästä.

The north wind and the sun (*Pohjantuuli ja aurinko*)
S7. Pohjantuuli ja aurinko väittelivät kummalla olisi enemmän voimää, kun he samalla näkivät kulkijan, jolla oli yllään lämmin takki.
S8. Silloin he sopivat, että se on voimakkaampi, joka nopeammin saa kulkijan riisumaan takkinsa.
S9. Pohjantuuli alkoi puhaltaa niin että viuhui, mutta mitä kovempaa se puhalsi, sitä tarkemmin kääri mies takin ympärilleen, ja viimein tuuli luopui koko hommasta.
S10. Silloin alkoi aurinko loistaa lämpimästi, eikä aikaakaan, niin kulkija riisui manttelinsa.
S11. Niin oli tuulen pakko myöntää, että aurinko oli kuin olikin heistä vahvempi.

Selected TIMIT corpus sentences.
S12. She had your dark suit in greasy wash water all year.
S13. Don't ask me to carry an oily rag like that.

Appendix B.

Long-term formant averages in the presence of outliers

Formant estimation is known to produce errors in which typically higher than expected values are observed (e.g. the $F1$ value for a frame is observed to be in the typical range of F2 formants). Computing long-term averages directly without post-processing could induce bias towards higher frequencies to the mean estimate. A simple technique to cut-off formants measures using fixed thresholds could sometimes remove valid observations or resonances corresponding to high vowels. We would like to use the high frequencies but give them lower weight in the mean estimation. We therefore model formant estimates using the two-component *Gaussian mixture model* (GMM) (Dempster et al., 1977) (bi-Gaussian model), in which the lower Gaussian is assumed to represent the true formant observations while the higher Gaussian represents the spurious or outlier observations. In addition to fitting the bi-Gaussian model, we also fitted a mono-Gaussian model in case a single mode would explain the data better. We select either the bi-Gaussian or mono-Gaussian model using the Akaike information criterion (AIC) (Akaike, 1974).

The probability density function of the bi-Gaussian model is,

$$p(x|\Lambda) = \lambda\mathcal{N}(x|\mu_1, \sigma_1^2) + (1-\lambda)\mathcal{N}(x|\mu_2, \sigma_2^2), \tag{B.1}$$

where F represents the raw formant measurements of a particular formant (F_1, F_2, or F_3) in a particular utterance, and where $\Lambda = \{\lambda, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2\}$, with $\mu_1 \leq \mu_2$, denotes the model parameters; μ_1 and μ_2 are the means, σ_1^2 and σ_2^2 the variances and $0 < \lambda < 1$ the relative proportion of observations in each Gaussian. We estimate Λ separately per each utterance, using the *expectation*

maximization (EM) (Dempster et al., 1977) algorithm with 100 random initializations, of which the model yielding the highest log-likelihood was selected. The 100 random initializations were used to reduce the variance of the estimated model. Finally, the 2nd component was discarded and μ_1 was selected as the formant mean of the particular utterance.

Appendix C.

Standard deviation of the mean differences for the formants

The mean formant differences between the values of naturally produced utterances and their respective two disguise cases was used to measure the level of change. If the difference was above one standard deviation, then the mean formant difference was considered significantly changed. Table C.8 presents the standard deviation per formants separated according to speaker's sex and condition.

Table C.8: Standard deviation (SD) in Hertz of the mean differences for $F1$ to $F4$ between natural voice and both disguised voices.

	Formant	**Old**	**Young**
Female	F1	66.21	54.55
	F2	308.91	312.18
	F3	199.42	250.33
	F4	104.80	97.35
Male	F1	104.9	92.88
	F2	422.21	378.62
	F3	401.99	332.16
	F4	166.89	176.76