# Voice Disguise in Automatic Speaker Recognition

MIREIA FARRÚS, Universitat Pompeu Fabra, Republic of Catalonia

We, humans, are able to identify other people even in voice disguise conditions. However, we are not immune to all voice changes when trying to identifying people from voice. Likewise, automatic speaker recognition systems can also be deceived by voice imitation and other types of disguise. Taking into account the voice disguise classification into the combination of two different categories (deliberate/non-deliberate and electronic/non-electronic), this survey provides a literature review on the influence of voice disguise in the automatic speaker recognition task and the robustness of these systems to such voice changes. Additionally, the survey addresses existing applications dealing with voice disguise and analyses some issues for future research.

CCS Concepts:•**Computing methodologies** → **Speaker recognition;**

Additional Key Words and Phrases: speaker recognition, voice disguise, voice imitation, voice conversion, channel degradation, robustness

**ACM Reference Format:**
Mireia Farrús, 2017. Voice Disguise in Automatic Speaker Recognition. *ACM Trans. Embedd. Comput. Syst.* 9, 4, Article 39 (March 2010), 24 pages.
DOI: 0000001.0000001

## 1. INTRODUCTION

Speaker recognition is the ability to identify others from a spoken sentence, taking into account those individual and characteristic speech features. In forensic applications, speaker recognition started as a human process, and early in the 1960s, speech forensic scientists made the first attempts to use speech spectrograms to recognise speakers [Kersta 1962; Stevens et al. 1968; Bolt et al. 1969; Tosi et al. 1972]. At that time, computer technology was still in early stages, and was not sufficient to become a complementary tool to the phonetic interpretation of the spectrograms. But as computer technology improved over time, the use of machines became more and more popular, leading to what nowadays is known as automatic speaker recognition.

Speech is a naturally produced signal, with a very low intrusiveness for humans, which can be easily stored and transmitted. Therefore, human voice has become a strong characteristic for speaker recognition [Reynolds et al. 2002], and the amount of related applications has considerably grown over the recent decades, such as secure access control to physical and electronic sites, transaction authentication, law enforcement, forensics, speech data management for mail browsing applications or intelligent answering machines, and customisation of devices and smart systems.

In order to make all these applications work, it is essential to find those idiosyncratic characteristics of the speech signal that make able to identify individuals. People take into account many diverse levels of information contained in the human voice [Schmidt-Nielsen and Crystal 2000], and these levels are related to several aspects

This work is supported by the Spanish Ministry of Economy through the Ramón y Cajal programme.
Author's address: M. Farrús, Department of Information and Communication Technologies, C/ Roc Boronat 138, 08018 Barcelona, Catalonia.

DOI: 0000001.0000001

of the voice: a specific word usage, a characteristic timbre, a rough voice, etc. Traditionally, these levels have been hierarchised from low level —associated to the vocal tract and to the learned articulatory configurations [Rabiner and Juang 1993; Gish and Schmidt 1994; Campbell 1997]— to high level —mostly related to the learned speaking habits.

Human speaker recognition performs rather reliable within a small population and in a high degree of familiarity between the speaker and the listener. What is more difficult —although feasible—, is to recognise others in voice disguise conditions [Reich 1981]. Nevertheless, both humans and automatic systems are exposed to several external threats: human listeners are not immune to voice changes, and they can also be deceived by voice imitation and other types of disguise. Likewise, automatic speaker recognition systems are also vulnerable to voice disguise. In some contexts, they are even less robust to voice disguise due to the lack of knowledge usually attributed to humans. The vulnerability of speaker recognition systems to impersonation has been largely tested and reported in literature; see e.g. [Lau et al. 2004; 2005], in which the experiments performed showed that an impostor could successfully attack the system if he knew of the clients of the database and both had similar voices. In control client access related applications, such vulnerability turns into a huge security problem. Therefore, an accurate analysis of voice disguise is needed in research in order to determine the optimal features to be used for automatic speaker recognition. Voice disguise analysis —and voice imitation in particular— can be used to find out which features are the most difficult ones to imitate or modify, leading to a more specific knowledge that will improve speaker recognition against spoofing attacks.

Taking into account some idiosyncratic characteristics of both human and automatic speaker recognition, and using Rodman's classification [Rodman 1998; Rodman and Powell 2000] as a take-off point —in which disguise is classified into the combination of two different categories: deliberate/non-deliberate, electronic/non-electronic— the aim of this survey is to provide a literature review on the influence of voice disguise in the speaker recognition task. To this end, this paper is structured as follows. Section 2 presents an overview of the existing types of voice disguise, describing their main characteristics. Section 3 outlines the influence of the well-known source-filter model in speaker recognition. Section 4 reviews the robustness of existing state-of-the-art speaker recognition systems against the mentioned voice disguises in terms of several speech features (mainly spectral, prosodic and voice quality features), in order to analyse how automatic speaker recognition reacts in front of such voice alterations. Section 5 analyses some issues for future research, specifically what still needs to be done to ensure a robust speaker recognition task and to avoid the effects of voice disguise. Finally, conclusions are sketched in Section 6.

## 2. VOICE DISGUISE

DNA, iris and fingerprints are some of the most known biometric identifiers that share the characteristic of being highly permanent over time. Voice, instead, is highly variable over time due to ageing, illness, emotional stress, and other non-deliberate factors. Moreover, voice is also modified by deliberate reasons, such as impersonating, speaking a foreign accent, etc. These voice changes, whatever the cause or the objective, are manifested as *voice disguise*. In this light, Rodman defined it as "any alteration, distortion or deviation from the normal voices, irrespective of the cause" [Rodman 1998].

Although voice disguise includes any voice modification, it can be caused by several factors. On the one hand, voice can be deliberately or non-deliberately changed. People may want to produce a deliberate disguise for different aims, making an extra effort to achieve it. But in many cases, our voice suffers uncontrollable changes during our lifetime, or even over the day. On the other hand, voice can be modified naturally or by

means of electronic devices. Based on these two dimensions, [Rodman 1998] classifies voice disguise based on four categories: (1) non-deliberate and non-electronic modifications, (2) non-deliberate and electronic modifications, (3) deliberate and non-electronic modifications, and (4) deliberate and electronic modifications.

In the current section, the influence of voice alteration on speech and the easiness of imitating speech features are outlined. The main characteristics and examples of each type of disguise following the two-dimensional classification described above are explained next and summarised in Table I.

### 2.1. Non-Deliberate Voice Disguise

People are subject to several changes in their voices due to uncontrolled causes. Most of them are naturally caused by modifications that affect the normal development of our body: ageing, illness, etc. Other modifications can be found when using electronic devices in the communication process. We will refer to them as non-deliberate non-electronic and non-deliberate electronic disguises, respectively.

*2.1.1. Non-Electronic Disguise.* The most common examples of non-deliberate and non-electronic disguises are a hoarse or breathy voice, a voice alteration due to emotional changes, intoxication —essentially by alcohol—, or the voice variation over time, i.e. ageing. The main acoustic characteristics of each of them are briefly described next.

*Hoarse and breathy voices.* Hoarseness is the colloquial expression for *disphonia*, an alteration of the voice quality usually manifested by breathy, rough or strained voices, and normally caused by abnormal situations or illnesses such as laryngitis or vocal cord nodules [Sulica 2011]. In 1967, [Yanagihara 1967] suggested that hoarseness was mainly characterised by the interactions of three factors: "(1) noise components in the main formant of each vowel, (2) high frequency noise components above 3000 Hz, and (3) loss of high frequency harmonic components". [Yumoto et al. 1982], [Kojima et al. 1982] and [Yumoto 1988] demonstrated that the hoarseness degree can be quantified by measuring the harmonics-to-noise ratio (HNR). On the other hand, breathy voices are typically characterised by an increase in spectral noise [Wayland et al. 1995]. Other acoustic parameters like jitter —the $F0$ cycle-to-cycle variations— and shimmer —the amplitude cycle-to-cycle variations— have been used over many years to detect voice pathologies; therefore, they are also known as *voice quality* parameters. Such pathological voices are usually related to higher values of both jitter and shimmer [Michaelis et al. 1998; Kreiman and Gerratt 2005]. In some recent works [Zhang and Lin 2017], the characteristics of disguised whispery voices in terms of intensity, syllable duration, formants and long-term average spectrum (LTAS) have also been analysed.

*Emotional changes.* The effects of emotional state on speech have been widely studied, especially when dealing with synthesised speech. One of the first works on the relationship between emotions and acoustic parameters was carried out by [Williams and Stevens 1972], who concluded that "anger, fear and sorrow situations tended to produce characteristic differences in fundamental frequency contour, average speech spectrum, temporal characteristics, precision of articulation, and waveform regularity of successive glottal pulses". The study, however, also showed that the same emotional situation was not consistent between speakers in terms of acoustic attributes. [Johnstone and Scherer 1999] and [Johnstone 2001] also demonstrated the existence of a relationship between acoustic parameters such as the fundamental frequency floor and range, the distribution of spectral energy, the jitter, and the speaker attitude and emotions. [Gobl and Chasaide 2003] explored the influence of voice quality on the communication of attitudes and emotions, pointing

out the non-existence of an unambiguous correspondence between affect and voice quality, but a cluster of affective attributes associated with an acoustic quality. Most of the studies focus on the relationship with pitch variables for being easier to measure [Scherer 1986]. However, other characteristics should be taken into account, such as speech rate and intensity [Scherer 1986][Williams and Stevens 1972; Carlson et al. 1992] or pausing structure [Cahn 1990], being voice quality parameters the key to differentiate emotions [Scherer 1986; Scherer et al. 1984; Murray and Arnott 1993]. [Narayana and Kopparapu 2009b] also show how the accuracy of a speaker recognition system decreases when the speaker is under stress or emotion, by quantifying the inherent stress contained in the speech of speaker using pitch, amplitude and duration. The detection of stressed speech levels has been shown to be highly tied to the Lombard effect [Zollinger and Brumm 2011] and includes "increased vocal effort, greater duration of words due to increased vowel length, shifts in formant locations for vowels, increased formant amplitudes, and deletion of some word-final consonants" [Markowitz 1996; 2007].

*Intoxication.* A good deal effort has also been put on determining how speech is altered under intoxication effects. In [Klingholz et al. 1988], several features were analysed under both sober and alcohol intoxicated conditions: frequency distributions of $F0$, signal-to-noise ratio (SNR), ratio of first to second formant frequencies ($F1/F2$), variation speed of $F0$, $F1$ and $F2$, and the long-term average spectrum (LTAS). While frequency variation speeds were not found to be altered by intoxication, LTAS, $F0$ and SNR —and especially the combination of the two latest ones— were able to discriminate quite well between sober and intoxicated conditions. On the other hand, $F1/F2$ was only modified with high levels of alcohol in blood. Other studies concluded that certain changes such as a $F0$ raising and a slowing of speaking rate occurred in intoxication —although they were not universal— while no significant changes were found in terms of vocal intensity [Hollien et al. 2001b; Hollien et al. 2001a], neither in voice onset times of the occlusives $/d/$ and $/t/$ [Swartz 1992].

*Ageing.* Speech production experiences both anatomical and physiological changes throughout life [Schoetz 2007]. In order to know how these changes develop, several studies have also been carried out that analyse the long-term voice changes over time. One of the first related studies indicated five characteristics as predictors of perceived age: "voice tremor, laryngeal tension, air loss, imprecise consonants, and slow articulation rate" [Ryan and Burk 1974]. On the other hand, [Ferrand 2002] found a significant lowering of HNR and apparent $F0$ differences in elderly speakers, while no significant differences were encountered in jitter. Contrarily, works such as [Linville 2001] reveal that jitter, shimmer, overall $F0$ statistics and long-term average spectra are highly correlated with ageing. Other studies such the one carried out by [Hartman 1979] revealed that the most prominent features when judging the age of the speakers could be classified according to $F0$, voice quality, articulation and speech rate. An increase of amplitude perturbation and a decrease of the reading rate were also detected as indicators of increasing age [Bruckl and Sendlmeier 2003]. To date, it is widely assumed that the most important correlates of speaker age are "features related to speech rate, sound pressure level, and $F0$" [Schoetz 2007].

*2.1.2. Electronic Disguise.* Non-deliberate electronic disguise refers mainly to speech degradation or distortion due to channel effects, as for instance, telephone transmission, use of microphones, etc. In fact, apart from the size of the population used in the automatic speaker identification task, the degradation introduced by noisy communi-

Table I. Different kinds of voice disguise. Classification and references.

| Disguise | Types | References |
|---|---|---|
| Non-deliberate & non-electronic | Hoarseness | [Yanagihara 1967][Kojima et al. 1982][Yumoto et al. 1982] [Yumoto 1988][Wayland et al. 1995][Yanagihara 1967] [Michaelis et al. 1998][Kreiman and Gerratt 2005] [Sulica 2011] |
| | Emotional stress | [Williams and Stevens 1972][Scherer et al. 1984] [Scherer 1986][Cahn 1990][Carlson et al. 1992] [Murray and Arnott 1993][Johnstone and Scherer 1999] [Johnstone 2001][Gobl and Chasaide 2003] [Markowitz 1996; 2007] |
| | Intoxication | [Klingholz et al. 1988][Swartz 1992] [Hollien et al. 2001b][Hollien et al. 2001a] |
| | Ageing | [Ryan and Burk 1974][Hartman 1979][Linville 2001] [Ferrand 2002][Bruckl and Sendlmeier 2003][Schoetz 2007] |
| Non-deliberate & electronic | | [Reynolds et al. 1995][Benesty et al. 2005] |
| Deliberate & non-electronic | | [Markham 1997][Zetterholm 2003] |
| Deliberate & electronic | | [Mashimo et al. 2001][Mashimo et al. 2002] [Hosom et al. 2003][Duxans 2006] |

cation channels is considered one of the largest factors affecting the system performance. This is stated in [Reynolds et al. 1995], in which some experiments are shown in order to analyse the performance loss with respect to various telephone channel degradations. Therefore, a lot of effort has been put on investigating noise reduction techniques in order to increase the speech communication and intelligibility in channel degradation situations. In this light, Schroeder pioneered such research in 1960 at Bell Labs [Benesty et al. 2005].

### 2.2. Deliberate Voice Disguise

Deliberate voice disguise depends on the speaker. [Kunzel 2000], for instance, reported differences in the strategies used between men and women. Nevertheless, whatever the way and the purpose are, deliberate modifications are an intentional alteration of the voice. As the non-deliberate voice disguise, the deliberate one is also classified into electronic and non-electronic. This section focuses on both kinds of voice alterations, analysing their main acoustic characteristics.

*2.2.1. Non-Electronic Disguise.* Voice imitation is innate in humans and can be found in human communication by means of language acquisition, impersonation, and voice transformation [Zetterholm 2003], as shown in Table II. In some way, impersonation and voice transformation are the most deliberate ones, since in both cases the speakers pretend to be someone else. Imitation encountered in language acquisition is essential to learn a mother tongue, as well as to learn foreign languages and for community integration by adapting new speaking manners. The language acquisition process is manifested by the word repetition, reproduction of syntactic structures, and phonetic reproduction, among others [Markham 1997].

Impersonation is a sort of imitation whose aim is to reproduce the voice of someone else [Markham 1997]. Entertainment is the main goal for impersonators, who have the ability to pretend to be a different person, being capable to target and imitate the most prominent speech features of the selected speaker. For stage entertaining, an impersonator tries to copy also the body language, as well as other non-verbal cues. Contrarily, when the audience is not able to see the impersonator, more focus is needed on vocal features. Wherever the entertainment takes place, the selection and exaggeration of the most prominent features is an essential characteristic to reach a successful imitation [Zetterholm 2003].

Table II. Voice imitation scope, aims and characteristics [Farrús 2008].

| **Scope** | **Aim** | **Characteristics** |
|---|---|---|
| Language Acquisition | First and second languages<br>Speaking manner adaptation | Imitation of several real speakers<br>Repetition of words<br>Reproduction of syntactic structures<br>Phonological/phonetic acquisition |
| Impersonation | Entertainment | Imitation of a specific person<br>Vocal and non-vocal imitation<br>Exaggerations |
| Voice Transformation | Hide identity | Imitation of a fictitious person<br>No exaggeration |

Another aim of voice imitation is to hide a specific identity. In this case, changes in the vocal tract are performed, so that some voice characteristics such as fundamental frequency, accent, prosody, voice quality, etc. are modified. In contrast with impersonation, the modified features are not exaggerated.

The variety of features to imitate is usually large: some of them are more related to a specific geographical or social environment —which would be the case of dialects and sociolects—, whereas other characteristics are more related to the individuals themselves —in which case we would talk about a special timbre and voice quality, among others—. Therefore, a successful imitator must focus on both groups of features [Zetterholm 2003]. However, as [Laver 1994] reports, a good imitation can turn into a difficult task due to large organic differences, being male and female voices an extreme example [Pittam 1994]. More recent works [Delvaux et al. 2017] use also LTAS in order to assess the ability to impersonate by comparing perceptual and acoustic features in controlled speech.

The importance of being a professional imitator is also studied in [Zetterholm 2003], by comparing both amateur and professional imitators imitating the same speakers. Although all imitations were different in terms of acoustic characteristics, the target speaker was clearly identified, and the same prominent features were targeted: ”pitch, dialect, rhythm, pausing, phonetic pronunciation of specific segments, and individual habits such as hesitations and loud breathing”. The only significant difference was found in intonation: whereas the professional impersonators were able to successfully copy the intonation patterns of the target voices, the amateur one did fail in the correct $F0$ range.

Dialect and accent modification is another sort of disguise, reported in works such as [Shuy 1990], among others. Some other studies are focused on whether human perception is affected by theatrical accent and dialect modification [Machlin 1975; Halloran 2003]. This disguise is not initially aimed at fooling people; however, since it could be used in a criminal setting, it is relevant to investigate how convincing dialect imitation is in general.

*2.2.2. Electronic Disguise.* Electronic disguise is the use of a device in order to alter or modify the natural speech. When performed in a deliberate way, it is very often found in the form of voice conversion, which is the transformation of the voice of a *source speaker* into the voice of a *target speaker*, in order to resemble a chosen target voice. So, technically speaking, it is a sort of electronic imitation of someone else's voice. Voice conversion is performed by means of a transformation function. During this transformation, the physical characteristics of the voice are modified, but the content of the message is kept as it is [Duxans 2006; Mashimo et al. 2001]. Nowadays, many recent works have gained a lot of insight into the voice conversion field by means of deep learning techniques (see, e.g. [Nakashika et al. 2013; Chen et al. 2014; Mohammadi and Kain 2014]).

Table III. Voice conversion applications [Farrús 2008].

| **Applications** | **Examples** |
|---|---|
| TTS Customisation | Creation of any target voice<br>(including intra- and cross-gender conversion)<br>Customisation of speech-to-speech translation output |
| Foreign Language Learning | Assistance for students to get a proper pronunciation |
| Medical Aids | Intelligibility improvement of an abnormal speech<br>Appropriate hearing aids design for specific hearing problems |
| Entertainment | Creation of voices of famous actors or people who are not alive<br>Assistance to singers in a karaoke |

Voice conversion technology is included in several applications such as speech synthesis, foreign language learning or speech-to-speech translation [Duxans 2006]. Voice conversion, for instance, can be added at the output of a synthesizer in order to customise the system [Mashimo et al. 2001], or be used in foreign language learning, by generating a proper phonetic pronunciation and intonation in their own voice [Mashimo et al. 2001; Mashimo et al. 2002; Duxans 2006]. For speech impaired people with abnormal speech, voice conversion systems can be used to improve their intelligibility [Hosom et al. 2003], or to design hearing aids by transforming those frequency ranges that can not be heard by some people [Duxans 2006]. Moreover, voice conversion can be found in entertainment scenarios such as karaokes and film dubbing. Table III summarises some examples of voice conversion applications.

Voice conversion has been shown to be effective when trying to fail automatic speaker recognition systems, as it will be seen in section 4. However, trying to deceive human listeners becomes a difficult task. In [Huckvale and Kristiansen 2012], for instance, a series of experiments with several degrees of electronic disguise using pitch scaling and vocal tract length scaling reported that the speaker identification accuracy was still high under these voice distortions. Only when an extreme disguise was applied —a pitch increase of 12 semitones and a vocal tract length reduction of 20%— the human recognition was significantly lower. In line with these experiments, [Clark and Foulkes 2007] performed a listening test over artificially disguised voices in which the $F0$ was modified. Disguises above +8 semitones and -8 semitones yielded the lowest scores, while the identification over the range between -8 and +8 semitones was still reasonably good enough.

## 3. SOURCE AND FILTER PARAMETERS IN AUTOMATIC SPEAKER RECOGNITION

The speech production system is normally described as two different processes: (a) the sound generation, and (b) the acoustic filtering of the speech sounds. The former —the *source*— takes place in the larynx, whereas the latter —the *filter*— is placed in the vocal tract. In this *source-filter model* [Fant 1960] (Figure 1), unvoiced sounds are represented as a random white noise. The linear model of speech production assumes that, given a voiced or an unvoiced source $U(z)$ that produces a voiced or an unvoiced speech $S(z)$, respectively, the filter consists of three cascade-based filters: glottal pulse $G(z)$, vocal tract $V(z)$, and lip radiation $L(z)$ [Fant 1960; Flanagan 1972] (Figure 2).

Low-level information has been traditionally associated to the speech signal features derived from the filter processes, which are related —in a complex way— to the vocal tract and to the learned articulatory configurations [Rabiner and Juang 1993; Gish and Schmidt 1994; Campbell 1997], and are referred to as the speech spectrum. Some of the most used spectral features in speech applications include LPCC (Linear Prediction Cepstral Coefficients, [Makhoul 1975]), MFCC (Mel Frequency Cepstral Coefficients, [Davis and Mermelstein 1980]), and PLP (Perceptual Linear Prediction Coefficients, [Hermansky 1990]).

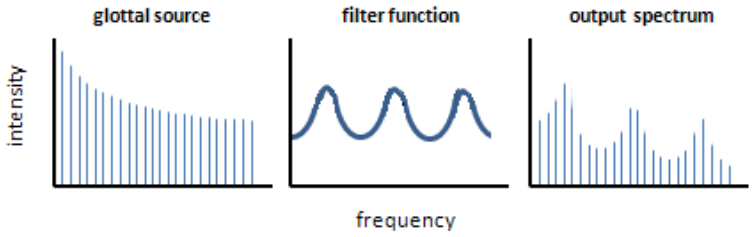


Fig. 1. Representation of the source-filter model output.

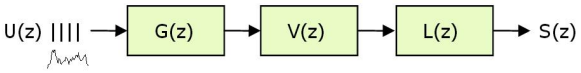


Fig. 2. Linear speech production model.

In contrast, high-level information has been related to features that depend on the learned speaking style, idiolect, etc. In terms of speech, these features are usually — although not only— associated to voice source features, and characterise the source of voiced sounds, which are known as the glottal flow [Kinnunen and Alku 2009]. The fundamental frequency ($F0$), which is the vibration rate of the vocal folds, is the most characteristic voice source feature. Since prosody is conveyed through intonation — variation of $F0$ over time—, rhythm and stress, it is also considered as an element derived from the voice source [d'Alessandro 2006]. Some other voice quality parameters such as jitter and shimmer [Michaelis et al. 1998; Kreiman and Gerratt 2005], which can be used to identify the speakers' gender and age [Wittig and Mueller 2003], language styles [Li et al. 2005], and speaker recognition [Farrús et al. 2007; Farrús and Hernando 2009], can also be viewed as source-derived parameters.

Automatic speaker recognition systems have mostly relied on filter parameters — i.e. low-level characteristics— by means of using short-term features associated to the voice spectrum. They have even been used to detect some speaker characteristics such as ageing [Metze et al. 2007]. Nevertheless, other higher levels of information such as prosody or voice quality parameters also play an important role in human identification, as has been demonstrated in several studies [Carey et al. 1996; Sonmez et al. 1998; Doddington 2001; Andrews et al. 2002; Bartokva et al. 2002; Weber et al. 2002], so that they can provide complementary information. Therefore, these complementary features have been recently added to the traditional spectral-based systems, since they are of great value for the speaker recognition task [Peskin et al. 2003; Reynolds et al. 2003; Farrús et al. 2006b].

## 4. ROBUSTNESS OF AUTOMATIC SPEAKER RECOGNITION SYSTEMS TO VOICE DISGUISE

In the previous sections, the main characteristics of several types of disguise, together with a brief mention of the filter and source parameters encountered in the speech, have been described. The current section goes further this description and tries to analyse how speaker recognition systems react in front of the above-mentioned disguises. To this end, the most relevant and recent works on the vulnerability of the state-of-the-art systems are reviewed, by classifying them into four different sources of disguise: (i) natural disguise, (ii) channel degradation, (iii) voice imitation, and (iv) voice conversion, and further analysing the robustness of the systems against these disguises. A summary of the related literature is shown in both Table IV and Table V.

### 4.1. Natural Disguise

In what follows, we analyse the robustness of state-of-the-art speaker recognition systems tested with different types of non-deliberated and non-electronic voice disguise —i.e. *natural disguise*—. When identified, robustness to source and filter parameters is also shown.

*4.1.1. Pathological Voices.* Most of the works dealing with naturally disguised voices and using automatic speaker recognition techniques have been devoted to detect voice disorders and not specifically to analyse the robustness of those systems against voice disguise. However, their findings can also help to understand how these systems are affected by such disguises. Some examples can be found, for example, in [Tull and Rutledge 1996; Fredouille et al. 2005] and [Fezari et al. 2014].

[Tull and Rutledge 1996] was one of the first works in this respect, which aimed at understanding the voice characteristics of people having a cold, in order to provide to the automatic speaker recognition systems the capability of recognising individuals in both healthy and sick conditions. In other words, the aim was to provide enough information to the recognition systems so that they could be robust to voices having a cold.

In [Fredouille et al. 2005], a GMM-based automatic speaker recognition approach is adapted to dysphonic voices by computing 16 MFCC plus delta. The classification experiment was performed over a dysphonic corpus consisting of 80 female voices —from 17 to 50 years old—, from which 20 were normal voices, and the other 60 were equally balanced between dysphonic voices of grade 1, 2 and 3, respectively. Normal and dysphonic voices obtained a high classification rate when tested with their corresponding conditions —normal and disphonic, respectively—, which suggests that systems trained and tested with normal voices are not robust to dysphonia-related changes.

[Fezari et al. 2014] uses a German database containing healthy and pathological voices from 95 speakers —aged from 20 to 82— including chronic laryngitis, vocal fold nodules and dysphonia, among others [Putzer and Koreman 1997]. A GMM-based system using 12 MFCC plus delta, acceleration and energy, together with jitter and shimmer measurements, was used in an SVM classification experiment. As in the work of [Fredouille et al. 2005], normal and pathological voices obtained a high classification rate when tested with their corresponding conditions, suggesting again that those systems trained and tested with normal voices are not robust to pathological voices.

*4.1.2. Emotions.* A large amount of works can be found in literature on the effect of emotional changes in automatic speaker recognition. In [Ghiurcau et al. 2011], for instance, the authors explore the effect of six different emotions —happiness, fear, anger, boredom, sadness and neutrality— recorded from ten different speakers in a GMM-based speaker recognition system using 10–24 MFCC plus delta coefficients. When the system was trained with a neutral state and tested with different emotional states, the system was not able to achieve a correct performance above 60%, being anger and happiness the emotions that most affected the system performance, and boredom and sadness the ones that less affected the system. Instead, when training the system in different emotional states, the performance increased up to 98%. Other works such as [Chen and Yang 2011] go beyond the analysis of the emotion effects on automatic speaker recognition systems and try to overcome the emotion effects. The authors apply and compare several techniques —GMM-UBM, *i-vector* and Emotional Factor Analysis (EFA), (see also section 4.2)— in order to increase the system performance over the Mandarin Affective Speech Corpus (MASC) [Wu et al. 2006].

*4.1.3. Intoxication.* As stated above, and although it is usually deliberately induced, intoxication is one of the considered natural disguises, among others. An experiment

carried out by [Klingholz et al. 1988] over eleven male speakers in sober and alcohol intoxication conditions was performed in order to see whether it was possible to discriminate between both conditions by analysing frequency distributions of $F0$, signal-to-noise ratio (SNR), ratio of first- to second-formants ($F1/F2$), variation speed of $F0$, $F1$ and $F2$, and long-term average spectrum (LTAS). Only SNR and $F0$ frequency distributions —and LTAS with some reservation— were capable to discriminate both conditions, suggesting that spectral parameters are not altered in a significant degree in front of low and medium levels of alcohol in blood.

*4.1.4. Ageing.* The effects of ageing have been analysed in several works such as [Matveev 2013] and [Kelly et al. 2014]. The former work presents a brief overview on the degradation effects of ageing in automatic speaker recognition. In addition, the author uses a Russian conversational microphone speech database consisting of more than 200 speakers, and compares the different ageing ranges using pitch statistics, formant frequencies, and MFCC plus delta and acceleration coefficients. Nearest neighbour and GMM-SVM models were used for classification. In all cases, the system performance was degraded by 20% —in terms of EER— every 1–2 years. In the latter work, the authors show a series of experiments over the Trinity College Dublin Speaker Ageing Database, which consists of 15 males and 11 females with an age difference range of 28–58 years per speaker. The experiments, performed on both an *i-vector* and GMM-UBM systems using 19 and 12 MFCC plus delta and acceleration coefficients, respectively, showed that the performance of both systems drops significantly as the age range increases.

### 4.2. Channel Degradation

Source parameters, specifically when analysed in terms of fundamental frequency contour, have been shown to be more robust to acoustic degradations derived from channel and noise than the short-time spectral characteristics of speech [Atal 1972; Carey et al. 1996]. In [Atal 1972] —one of the first works in this respect— temporal variations of pitch were used to identify speakers. Over a database of 60 utterances spoken by 10 speakers, a 20-dimensional vector representing the pitch contour was computed, and the identification was determined in terms of the Euclidean distance between both test and reference, leading up to a 97% of correct performance. In the same way, formant frequencies were found to be robust to the frequency characteristics of the transmission system and the recording conditions. In the frame of the NIST 1995 Speaker Recognition Evaluation, [Carey et al. 1996] used prosodic characteristics based on both pitch and energy contours by using their first four statistics (mean, variance, skew and kurtosis), showing that prosodic features —particularly those based on pitch— were less vulnerable to handset variability than the spectral ones.

Contrarily, the performance of both automatic speech and speaker recognition systems that use MFCC is usually degraded in the presence of noise, since noise itself has an evident effect on these features [Narayana and Kopparapu 2009a]. In order to overcome the lack of robustness of spectral features, some recent works such as [Chougule and Chavan 2015] define a "robust spectral feature set NDSF" —which stands for Normalised Dynamic Spectral Features—, which is used for automatic speaker recognition in mismatch conditions. When compared with traditional MFCC and LPCC, and tested over a multi-variability speaker recognition and a multi-speaker Hindi speech database, the experiments show that NDSF are more robust to conventional cepstral features in both databases.

Prior to achieving results such as the ones in [Chougule and Chavan 2015], a lot of effort has been put on overcoming the lack of robust spectral features during the last decades. Recently, *i-vector*-based speaker recognition has been widely applied as

one of the state-of-the-art techniques in this direction. The *i-vectors* approach aims to model the speaker's long term prosodic and spectral characteristics by using continuous approximations of the prosodic and cepstral contours [Dehak 2009]. This approach evolves from the Joint Factor Analysis (JFA) [Kenny et al. 2008], which compensates for channel and session variability. Unlike JFA, *i-vector*-based speaker recognition does not require a separate estimation of speaker and channel spaces [Dehak et al. 2010; Shum et al. 2010], so that both spaces are modelled together in a low-dimensional total-variability space [Kanagasundaram et al. 2011]. Based of the similarity between the channel effect and the emotion effect, the *i-vector* approach is also used to overcome the emotion variability problem [Chen and Yang 2011].

### 4.3. Voice Imitation and Modification

The existing literature on intentional voice disguise in automatic speaker recognition is huge. The aim of the current section is to present a brief overview on voice imitation and also other intentional disguises and modifications, including those in which the speaker is not trying to imitate any specific person, as well as the resulting effects of trying to imitate a foreign accent or dialect.

*4.3.1. Voice Mimicking.* One of the first works concerning voice mimicry in automatic speaker recognition can be found in [Lummis and Rosenberg 1972], in which the authors tested a verification system that used F1, F2 and F3, as well as pitch and intensity level, against four well-trained impostors trying to imitate eight speakers, resulting in an acceptance rate of 27% —in comparison with a 1.2% rate for non-mimicking impostors.

Some decades later, complementary studies performed by [Farrús et al. 2008a] and [Farrús et al. 2008b] explored the ability of two male professional impersonators to approximate the prosody and source parameters of five well-known politicians. The recordings were taken from public radio interviews. An identification based on 12 source and prosody-related features was used to: (1) identify the target and natural voices for each of the following prosodic parameters: length of voiced and unvoiced segments, frames per word, $F0$ means, extrema and ranges; and jitter and shimmer features, and (2) distinguish between target and modified voices from the same speakers set. The results showed an increase of the identification error rate in all features when using the modified system instead of the baseline except for the $F0$ range.

More recent works such as [González-Hautamaki et al. 2013] have used both GMM-UBM and *i-vector*-based systems to analyse the effect of a professional Finnish imitator impersonating five well-known Finnish public figures, in which the results suggest that, although and increase in the false acceptance rate was observed in the *i-vector*-based system, the mimicry effects were less significant than those produced by voice conversion techniques. In a similar way, [Uzan and Wolf 2015] make use of both an *i-vector*-based system and Convolutional Neural Networks (CNN) to study the voice variability of professional actors when imitating specific characters, showing that CNN clearly outperform *i-vector*-based systems.

*4.3.2. Intentional Voice Modification.* One of the works that have been able to find highly robust features in deliberately disguised voices can be found in [Taseer 2005]. The author states that the glottal plosive "is one of the consonants that exist phonemically in all languages", and it is part of a natural phenomenon highly difficult to be controlled, so that the utterance of a glottal stop can be used as a unique quality of the pronunciation of the speakers. Based on these characteristics, [Taseer 2005] recorded a number of male and female speakers aged between 25 and 33 and analysed responses of both energy and frequency of the glottal plosives under a disguised condition resulting from stressing the vocal cords in Urdu. The results showed that the glottal pulse

information can be used to identify the speaker under both normal and disguised voice conditions.

Going beyond the work of [Kunzel 2000], the same author analyses the effects of natural voice disguise on the accuracy of a UBM-MAP adapted GMM forensic automatic speaker recognition system over 100 German speakers in order to evaluate the system performance degradation caused by increased voice pitch, lowered voice pitch and noise pinching [Kunzel et al. 2004]. A significant degradation was shown only when the reference population was assembled with normal speech, and highly mitigated when testing over reference populations containing the same type of disguise, which suggests a lack of robustness of this spectral-based system when dealing with such voice disguises. In the same line, [Kajarekar et al. 2006] analysed the effect of intentional voice modifications regarding *speaking style* —which turned out to be reflected by means of modifying pitch, duration or mimicking an accent by most of the speakers— against a GMM-based speaker recognition system using 13 MFCC on the FISHER database, as a part of the NIST 2003 Extended Speaker Recognition Evaluation (SRE[1]). The results showed an increase of the EER from 0.05% —tested with normal voices— to 7.46% —tested with disguised voices—, representing a 39% of false rejection of subjects disguising their voices.

Other similar works such as [Zhang and Tan 2008] and [Tan 2010] studied the effect of 10 kinds of voice disguise in a developed system called FASRS (Forensic Automatic Speaker Recognition System), together with normal voices recorded by 20 male students. The analysis showed that the performance of speaker recognition was highly degraded due to voice disguise, differing in several disguising types, except for the foreign accent, to which it was highly resistant —since FASRS was developed as a language and dialect independent system—, being whisper and masking on mouth the ones which had the greatest effect on the system. More recently, [González-Hautamäki et al. 2017] reported an increase of EER in an automatic speaker verification system due to significant differences in F0 and at least one of the formants in around 70% of utterances.

*4.3.3. Accent and dialect imitation.* In line to the findings of [Zhang and Tan 2008] and [Tan 2010], in which foreign accent was shown to be robust to the FASRS system, some other studies have investigated the vulnerability of speaker recognition systems to dialect and accent disguise. A GMM-based recognition system using 20 MFCC and including delta and acceleration coefficients was used in [Farrús et al. 2006a] over a database consisting of several movies excerpts spoken by two different American actors, one of them using British, Irish and Scottish English, apart from his own dialect. The experiments revealed that the recognition of the same dialect outscored recognition of the same speaker. So, it appeared that "accent-specific features dominate speaker-specific features, and that dialect imitation can confuse both the human and speaker recognition systems, yet in different ways".

[1] https://catalog.ldc.upenn.edu/LDC2010S03

Table IV. Summary of speaker recognition robustness to non-deliberate voice disguises: natural and channel degradation. The table shows whether the system is degraded when tested with the corresponding disguise, and which system features —if any— are found to be robust against such voice alterations.

| **Disguise** | **System characteristics** | **Database** | **System degradation** | **Robust features** | **References** |
|---|---|---|---|---|---|
| Natural | GMM, 16 MFCC + $\Delta$ | 80 female speakers:<br>20 normal voices +<br>60 dysphonic voices | yes | — | [Fredouille et al. 2005] |
| | GMM, 12 MFCC + $\Delta$ + $\Delta\Delta$<br>+ Energy<br>+ jitter, shimmer | 95 German normal<br>+ pathological voices | yes | — | [Fezari et al. 2014] |
| | GMM, 10-24 MFCC + $\Delta$ | 10 speakers, six emotions:<br>sadness, boredom, happiness,<br>anger, fear, neutrality | yes | — | [Ghiurcau et al. 2011] |
| | GMM-UBM, i-vector, EFA | Mandarin Affective<br>Speech Corpus | yes | — | [Wu et al. 2006] |
| | — | 11 alcohol-intoxicated<br>male speakers | yes | F1/F2<br>$\Delta$(F0,F1,F2) | [Klingholz et al. 1988] |
| | F0 statistics<br>formant frequencies<br>13 MFCC + $\Delta$ + $\Delta\Delta$ | $>$ 200 Russian speakers<br>(diff. age ranges)<br>convers. microph. speech | yes | — | [Matveev 2013] |
| | i-vector, 19 MFCC + $\Delta$ + $\Delta\Delta$<br>GMM-UBM, 12 MFCC+$\Delta$+$\Delta\Delta$ | Trinity College Dublin<br>Speaker Ageing | yes | — | [Kelly et al. 2014] |
| Channel degradation | 20-dim F0 contour vector | 10 speakers<br>60 utterances | no | F0 contour<br>formants | [Atal 1972] |
| | HMM, 12 cepstr. coeff.+$\Delta$<br>F0 and Energy statistics | Switchboard<br>21 male speakers | yes | F0 stats | [Carey et al. 1996] |
| | NDSF, MFCC, LPCC | multi-variability/speaker<br>continuous Hindi speech | yes | (NDSF) | [Chougule and<br>Chavan 2015] |

Table V. Summary of speaker recognition robustness to deliberate disguises: voice imitation and conversion. The table shows whether the system is degraded when tested with the corresponding disguise, and which system features —if any— are found to be robust against such voice alterations.

| Disguise | System characteristics | Database | System degradation | Robust features | References |
|---|---|---|---|---|---|
| Voice imitation | F0, intensity<br>F1,F2,F3 | 4 well-trained professionals<br>imitating 8 speakers | yes | — | [Lummis and Rosenberg 1972] |
| | 12 source/prosody parameters | 2 male professional impersonators<br>mimicking 5 well-known politicians | yes | F0 range | [Farrús et al. 2008a; 2008b] |
| | GMM-UBM<br>i-vector | Finnish imitator impersonating<br>5 well-known public figures | yes | — | [González-Hautamaki et al. 2013] |
| | LPCC, glottal stops<br>Energy and frequency | 5 male/female speakers with<br>vocal cords stressing | no | glottal pulses | [Taseer 2005] |
| | GMM-UBM-MAP<br>19 MFCC+Δ | 100 German speakers<br>faking kidnapper's calls | yes | — | [Kunzel et al. 2004] |
| | GMM, 13 MFCC | FISHER (pitch/duration<br>modification, accent mimicking) | yes | — | [Kajarekar et al. 2006] |
| | FASRS language/dialect independent | 20 male speakers<br>10 types of disguise | yes | foreign accent | [Zhang and Tan 2008; Tan 2010] |
| | GMM-UBM<br>12 MFCC+Δ+ΔΔ | 2 American actors using<br>American, British, Irish and Scottish accents | yes | — | [Farrús et al. 2006a] |
| Voice conversion | GMM, LPCC | 2 human + synthetic voices<br>word concat./diphone/re-synthesis<br>1 male, 1 female | (yes) | word concaten. | [Lindberg and Blomberg 1999] |
| | GMM, LPC+Δ | 10 human + HMM synthetic voices<br>6 males, 4 females | yes | — | [Masuko et al. 2000] |
| | GMM-UBM<br>16 LFCC+Δ | Eval05 NIST SRE corpus | yes | — | [Matrouf et al. 2006] |
| | GMM-UBM<br>12 MFCC+Δ+ΔΔ | 2 male and 2 female source voices<br>all of them converted to each other | yes | — | [Farrús et al. 2010] |

### 4.4. Voice Conversion

Several studies have been performed in order to analyse the speaker recognition systems vulnerability to voice disguise using synthetic voices. An early experiment reported in [Lindberg and Blomberg 1999] tried to spoof a verification system by means of diverse artificial voices created from client speech. The system, GMM-based and with LPCC parameterised speech, was tested over three types of synthesised voice: word concatenation, resynthesis and diphone synthesis from two clients: one male and one female, both Swedish from the same age range. The results showed word concatenation being an effective impostor technique, while resynthesis and diphone synthesis provided significant differences over the client voice, being not suitable as impostor voices.

Other works related to the robustness of these systems to synthetic speech can be found in [Masuko et al. 2000] and [Matrouf et al. 2006], in which the impostor acceptance rate increases when the impostor voice is modified. Specifically, [Masuko et al. 2000] used a HMM speaker verification system based on pitch and spectrum, through a feature vector consisting of 20 cepstral coefficients and delta parameters, over a database consisting of six male and four female speakers, in both human and HMM-synthetic modes. The results showed high false acceptance rates for synthetic speech, suggesting a lack of robustness in front of synthetic voices. [Matrouf et al. 2006] achieve high increase of the impostor acceptance rate of a GMM-UBM ASR system by modifying the impostor voice —-using a simple transformation method in a frame basis— in order to target the GMM of a specific speaker. The experiments are tested over the *Eva05* corpus of the NIST SRE evaluation campaign.

The work of [Farrús et al. 2010] aims at quantifying how good automatic converted voices can approximate other's voices by using a GMM-UBM speaker recognition system based of 12 MFCC features plus delta and acceleration. The system is used to test the quality of the converted voices, showing that the identification error rate increases over converted voices, especially in cross-gender conversions. Prosodic features were not modified between source and target voices, so they could not be used to test the speaker's identity. Another work worth mentioning in the scope of converted voices is the one by [Wu and Li 2013], which consists of an overview of spoofing attack and related techniques by focusing on voice conversion scenarios, as well as the one in [Kaur and Singh 2017], in which a MFCC-based SVM classifier is used to identify electronically disguised voices.

## 5. ISSUES FOR FUTURE RESEARCH

Voice disguise has been shown to be a significant threat for speaker recognition. However, there are still some open issues that would definitively help to improve the findings of the existing literature and overcome the lack of robustness that has been shown in most of the systems. Some of these pending issues for future research are presented next.

### 5.1. Automatic or Human Identification?

One of the central questions that arise when dealing with disguised voices in speaker recognition is whether automatic speaker recognition is better than human identification or vice versa. A series of experiments reported in [Sullivan and Pelecanos 2001] demonstrated that an automatic speaker verification system was less vulnerable to impersonators than those systems relying on human identification and verification. In contrast, [González-Hautamaki et al. 2015] showed that the human listeners widely outperformed three automatic verification systems. However, these results should be interpreted with caution since the listeners here were familiar to the target speak-

ers, which was a clear advantage over the automatic systems. Moreover, the database consisted of very short sentences, being another handicap for automatic systems. Although human identification is out of the scope of this survey, more effort should be put in this respect to better understand the idiosyncrasy and characteristics of automatic speaker recognition systems and to clearly analyse the differences between both types of recognition. This would help to understand why automatic speaker recognition is sometimes outperformed by human recognition, shedding light to further improvements to this respect.

Another different but related issue is the identification of text-to-speech (TTS) generated voices as a sort of disguise. As TTS improve over time, the generated voices are closer to human voice than some decades ago. Therefore, current studies should be performed, as the one pointed out in [Amino et al. 2018].

### 5.2. Source or Filter Parameters?

As stated in the Introduction section, systems using both source and filter parameters generally outperform those systems relying only on source or filter parameters. In [Lau et al. 2004], impersonators found it easier to approximate the source parameters than the filter ones. Contrarily, [Zetterholm 2006] showed that a professional impersonator was clearly able to target the filter parameters from a well-known target speaker.

Knowing which features are the most likely to be imitated is not an easy task. [Eriksson and Wretling 1997], for example, found that overall speech rate and mean $F0$ were easy to imitate, whereas segmental timing and formants were rather difficult. [Zetterholm et al. 2004] also found $F0$ easy to imitate and segmental durations more difficult. [Kitamura 2008], instead, showed in some experiments that spectrum and $F1$, $F3$, and $F4$ were successfully imitated, while $F2$ and $F0$ imitation was not that successful. In some other other experiments carried out by [Farrús et al. 2008a], it transpired that most prosodic parameters were affected, being $F0$ range an exception, since the imitators did not get to modify it.

Overall, it seems that prosodic features are quite easy to imitate, whereas the results on the imitation potentiality of spectral features are not very consistent. Nevertheless, most of the automatic speaker recognition systems rely on spectral characteristics, and since the success of a speaker recognition system depends on the features used to characterise speaker information [Espy-Wilson et al. 2006], this makes spectrum-based systems not robust to the prosodic information of the speaker. Moreover, imitation tends to exaggerate prosodic, idiosyncratic and lexical behaviour. Therefore, automatic speaker verification systems are usually not suitable to provide robustness to mimicry [González-Hautamaki et al. 2013]. In order to overcome this lack of robustness, source parameters should be clearly added as essential features when using automatic speaker recognition systems. As it has been seen in the reviewed works over this survey, robust features are usually those features that have been used as system characteristics.

### 5.3. Other Issues

The analysis automatic speaker recognition against voice disguise can lead to a sort of knowledge that can be applied to other fields. For example, [Sigmund 2008] proposes an interesting application to detect intoxication through the speech signal, in which the previously programmed ignition switch —controlled by a speaker recognition system— does not work under drug or alcohol intoxication of the driver. More insight into this kind of applied technology would be beneficial for any speech technology field.

Moreover, imitations can be highly related to those voices that are anatomically similar, such as those from twins or related persons. [Scheffer et al. 2001] reports that an

automatic speaker identification system was able to recognise a twin with a 85% of correct identification successfully performed in verification mode. However, there is not much work textcolorreddone in this area, and it would be doubtlessly a relevant field of study.

Many works in the literature focus on analysing which features are modified when a sort of disguise is applied. Many others focus on the system robustness in front of several disguises, without taking into account which human voice features were modified during the disguise. A more strong connection between both types of studies should be carried out in the future, so that a more understanding of the modified features could be applied in the designing of automatic speaker recognition systems dealing with voice disguises, especially the non-electronic ones, in which the affected characteristics tend to be unclear. Moreover, a greater deal of effort should be put on multi-disguise analysis; i.e., the understanding of how voice features are affected by several disguises at the same time, and how automatic recognition systems react to this effect.

## 6. CONCLUSIONS

This survey reviews some of the existing literature on automatic speaker recognition systems against voice disguise, based on the electronic/non-electronic and deliberate/non-deliberate dimensions, together with a previous review on several types of disguise and how voice features are modified.

The survey also addresses some future issues that are strongly linked to the content of the article. One of the main issues addressed is the poor connection between the features modified during a disguise and the features used in automatic speaker recognition systems, which affects the system robustness. A significant example is the one of voice conversions. As far as it is known —as most of the automatic speaker recognition systems do— voice conversion systems rely only the spectrum of the voice, without taking into account prosody and other linguistic dimensions [Duxans 2006], while voice mimicry makes a significant use of prosody alteration.

Dialectal changes and other disguises are based mainly on phonetic, prosodic and lexical alterations. Therefore, it would not be expected to find a priori a spectral automatic system capable of recognising one speaker's voice according to the spoken accent spoken or dialect —unless accent and dialectal characteristics are reflected in the spectrum of the voice, which is not usual—. Therefore, a wider understanding of which features are altered in mimicry —and in voice disguise in general— will certainly help to improve the design of the automatic speaker recognition system, and the addition of the disguise-based altered features into these systems will increase their performance, both in terms of accuracy and robustness.

## REFERENCES

Kanae Amino, Hisanori Makinae, and Toshiaki Kamada. 2018. Auditory discrimination of natural speech and synthetic speech used as voice disguise. *Acoustical Science and Technology* 39, 1 (2018), 48–50.

Walter D. Andrews, Mary A. Kohler, Joseph P. Campbell, John J. Godfrey, and Jaime Hernández-Cordero. 2002. Gender-dependent phonetic refraction for speaker recognition. In *Proceedings of the ICASSP*, Vol. 1. IEEE, Orlando, FL, USA, 149–152. DOI:http://dx.doi.org/10.1109/ICASSP.2002.5743676

Bishner Saroop Atal. 1972. Automatic speaker recognition based on pitch contours. *Journal of the Acoustical Society of America* 52, 6B (Dec. 1972), 16871697. DOI:http://dx.doi.org/10.1121/1.1913303

Katarina Bartokva, David Le-Gac, Delphine Jauvet, and Denis Jouvet. 2002. Prosodic parameter for speaker identification. In *Proceedings of the seventh International Conference on Spoken Language Processing*. Denver, Colorado, 1197–1200.

Jacob Benesty, Shoji Makino, and Jingdong Chen (eds.). 2005. *Speech Enhancement*. Springer.

Richard H. Bolt, Franklin S. Cooper, Edward E. David Jr., Peter B. Denes, James M. Pickett, and Kenneth N. Stevens. 1969. Identification of a Speaker by Speech Spectrograms. *Science* 166, 3903 (Oct. 1969), 338–342.

Markus Bruckl and Walter F. Sendlmeier. 2003. Aging Female Voices: an Acoustic and Perceptive Analysis. In *Proceedings of the VOQUAL03*. Geneva, Switzerland, 163–168.

Janet E. Cahn. 1990. The generation of affect in synthesized speech. *Journal of the American Voice I/O Society* 8 (1990), 1–9.

Joseph P. Campbell. 1997. Speaker recognition: A tutorial. *Proc. IEEE* 85 (Sept. 1997), 1437–1462. http://ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=628714

Michael J. Carey, Eluned S. Parris, Harvey Lloyd-Thomas, and Stephen Bennett. 1996. Robust prosodic features for speaker identification. In *Proceedings of the fourth International Conference on Spoken Language Processing*. Philadelphia, PA, 800–1803. DOI:http://dx.doi.org/10.1109/ICSLP.1996.607979

Rolf Carlson, Bjorn Granstrom, and Lennart Nord. 1992. Experiments with emotive speech, acted utterances and synthesized replicas. *Speech Communication* 11, 1 (March 1992), 347–355.

Li Chen and Yingchun Yang. 2011. Applying Emotional Factor Analysis and I-Vector to Emotional Speaker Recognition. In *Biometric Recognition. Proceedings of the 6th Chinese Conference (CCBR) (Lecture Notes in Computer Science)*, Zhenan Sun, Jianhuang Lai, and Xilin Chen Tieniu Tan (Eds.). Springer Berlin Heidelberg, Beijing, China, 174–179. DOI:http://dx.doi.org/10.1007/978-3-642-25449-9_22

Ling-Hui Chen, Zhen-Hua Ling, Li-Juan Liu, and Li-Rong Dai. 2014. Voice Conversion Using Deep Neural Networks with Layer-wise Generative Training. *IEEE/ACM Transactions on Audio, Speech and Language Processing* 22, 12 (Dec. 2014), 1859–1872. DOI:http://dx.doi.org/10.1109/TASLP.2014.2353991

Sharada V. Chougule and Mahesh S. Chavan. 2015. Robust Spectral Features for Automatic Speaker Recognition in Mismatch Condition. In *Procedia Computer Science (Second International Symposium on Computer Vision and the Internet (VisionNet15))*, Vol. 58. Elsevier, Kerala, India, 272–279. DOI:http://dx.doi.org/doi:10.1016/j.procs.2015.08.021

Jessica Clark and Paul Foulkes. 2007. Identification of voices in electronically disguised speech. *International Journal of Speech Language and the Law* 14, 2 (Dec. 2007). DOI:http://dx.doi.org/10.1558/ijsll.v14i2.195

Christophe d'Alessandro. 2006. Voice source parameters and prosodic analysis. In *Language Context and Cognition. Methods in Empirical Prosody Research*, Anita Steube (Ed.). Walter de Gruyter, Berlin, New York, 63–88.

Steven B. Davis and Paul Mermelstein. 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech and Signal Processing* 28, 4 (Aug. 1980), 357–366. DOI:http://dx.doi.org/10.1109/TASSP.1980.1163420

Najim Dehak. 2009. *Discriminative and Generative Approaches for Long- and Short-Term Speaker Characteristics Modeling: Application to Speaker Verification*. Phd dissertation. École de Technologie Supérieure, Montréal, Canada.

Najim Dehak, Reda Dehak, J. Glass, Douglas Reynolds, and Patrick Kenny. 2010. Cosine similarity scoring without score normalization techniques. In *Proceedings of the ODYSSEY10 – The Speaker and Language Recognition Workshop*. ISCA, Brno, Czech Republic, 71–75.

Véronique Delvaux, Lise Caucheteux, Kathy Huet, Myriam Piccaluga, and Bernard Harmegnies. 2017. Voice disguise vs. Impersonation: Acoustic and perceptual measurements of vocal flexibility in non experts. *Proceedings of the Interspeech 2017* (2017), 3777–3781.

George Doddington. 2001. Speaker recognition based on idiolectal differences between speakers. In *Proceedings of the Eurospeech*, Vol. 4. Aalborg Denmark, 2521–2524.

Helenca Duxans. 2006. *Voice Conversion applied to Text-to-Speech systems*. Phd dissertation. Universitat Politècnica de Catalunya, Department od Signal Processing and Communications, Barcelona, Catalonia.

Anders Eriksson and Par Wretling. 1997. How flexible is the human voice? - A case study of mimicry. In *Proceedings of the Eurospeech*. ISCA, Rhodes, Greece, 1043–1046. http://www.ling.gu.se/~anders/papers/a1008.pdf

Carol Y. Espy-Wilson, Sandeep Manocha, and Srikanth Vishnubhotla. 2006. A New set of features for text-independent Speaker Identification. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*. Pittsburgh, Pennsylvania, USA, 1475–1478. http://www.isr.umd.edu/Labs/SCL/publications/conference/espy_manocha_vish_icslp_06.pdf

Gunnar Fant. 1960. *Acoustic theory of speech production: with calculations based on X-ray studies of Russian articulations*. Mouton and Co., The Hague, Netherlands.

Mireia Farrús. 2008. *Fusing prosodic and acoustic information for speaker recognition*. Phd dissertation. Universitat Politècnica de Catalunya, Barcelona, Catalonia.

Mireia Farrús, Erik Eriksson, Kirk P.H. Sullivan, and Javier Hernando. 2006a. Dialect imitations in speaker recognition. In *Proceedings of the European IAFL Conference on Forensic Linguistics, Language and the Law*. Barcelona, 347–353.

Mireia Farrús, Ainara Garde, Pascual Ejarque, Jordi Luque, and Javier Hernando. 2006b. On the fusion of prosody, voice spectrum and face features for multimodal person verification. In *Proceedings of the ICSLP*. Pittsburgh, PA, 2106–2109.

Mireia Farrús and Javier Hernando. 2009. Using Jitter and Shimmer in speaker verification. *IET Signal Processing* 3, 4 (July 2009), 247–257. DOI: http://dx.doi.org/10.1049/iet-spr.2008.0147

Mireia Farrús, Javier Hernando, and Pascual Ejarque. 2007. Jitter and shimmer measurements for speaker recognition. In *Eighth Annual Conference of the International Speech Communication Association*.

Mireia Farrús, Michael Wagner, Jan Anguita, and Javier Hernando. 2008a. How vulnerable are prosodic features to professional imitators?. In *Proceedings of ODYSSEY08 The Speaker and Language Recognition Workshop*. Stellenbosch, South Africa.

Mireia Farrús, Michael Wagner, Jan Anguita, and Javier Hernando. 2008b. Robustness of prosodic features to voice imitation. In *Proceedings of the Interspeech*. Brisbane, Australia.

Mireia Farrús, Michael Wagner, Daniel Erro, and Havier Hernando. 2010. Automatic speaker recognition as a measurement of voice imitation and conversion. *The Intenational Journal of Speech. Language and the Law* 1, 17 (2010), 980–988.

Carole T Ferrand. 2002. Harmonics-to-Noise Ratio: An Index of Vocal Aging. *Journal of Voice* 16, 4 (Dec. 2002), 480–487. DOI: http://dx.doi.org/10.1016/S0892-1997(02)00123-6

Mohamed Fezari, Fethi Amara, and Ibrahim M. M. El-Emary. 2014. Acoustic Analysis for Detection of Voice Disorders Using Adaptive Features and Classifiers. In *Proceedings of the International Conference on Circuits, Systems and Control*. Interlaken, Switzerland, 112–117.

James L. Flanagan. 1972. *Speech Analysis, Synthesis and Perception*. Springer, Berlin-Heidelberg-New York.

Corinne Fredouille, Gilles Pouchoulin, Jean-Franois Bonastre, Marion Azzarello, Antoine Giovanni, and Alain Ghio. 2005. Application of Automatic Speaker Recognition techniques to pathological voice assessment (dysphonia). In *Proceedings of the Interspeech*. ISCA, 149–152.

Marius Vasile Ghiurcau, Corneliu Rusu, and Jaakko Astola. 2011. A study of the effect of emotional state upon text-independent speaker identification. In *Proceedings of the ICASSP*. IEEE, Prague, Czech Republic, 4944–4947. DOI: http://dx.doi.org/10.1109/ICASSP.2011.5947465

Herbert Gish and Michael Schmidt. 1994. Text-independent speaker identification. *IEEE Signal Processing Magazine* 11, 4 (Oct. 1994), 18–32. http://ieeexplore.ieee.org/xpl/articleDetails.jsp?tp=&arnumber=317924

Christer Gobl and Ailbhe Ní Chasaide. 2003. The role of voice quality in communicating emotion, mood and attitude. *Speech Communication* 40, 1–2 (April 2003), 189–212. DOI: http://dx.doi.org/10.1016/S0167-6393(02)00082-1

Rosa González-Hautamaki, Tomi Kinnunen, Ville Hautamki, and Anne-Maria Laukkanen. 2015. Automatic versus human speaker verification: The case of voice mimicry. *Speech Communication* 72 (May 2015), 13–31. DOI: http://dx.doi.org/10.1016/j.specom.2015.05.002

Rosa González-Hautamaki, Tomi Kinnunen, Ville Hautamki, Timo Leino, and Anne-Maria Laukkanen. 2013. I-vectors meet imitators: on vulnerability of speaker verification systems against voice mimicry. In *Proceedings of the Interspeech*. ISCA, Lyon, France, 930–934.

Rosa González-Hautamäki, Md Sahidullah, Ville Hautamäki, and Tomi Kinnunen. 2017. Acoustical and perceptual study of voice disguise by age modification in speaker verification. *Speech Communication* 95 (2017), 1–15.

Nate Halloran. 2003. *The acquisition of a stage dialect*. Master's thesis. Portland State University, Portland, OR.

David E. Hartman. 1979. The perceptual identity and characteristics of aging in normal male adult speakers. *Journal of Communication Disorders* 12, 1 (Feb. 1979), 53–61.

Hynek Hermansky. 1990. Perceptual linear predictive (PLP) analysis of speech. *Journal of the Acoustical Society of America* 87, 4 (Aug. 1990), 1738–1752. DOI: http://dx.doi.org/10.1121/1.399423

Harry Hollien, Gea DeJong, Camilo A. Martin, R. Schwartz, , and Kristen Liljegren. 2001a. Effects of ethanol intoxication on speech suprasegmentals. *Journal of the Acoustical Society of America* 110, 6 (Dec. 2001), 3198–206. DOI: http://dx.doi.org/10.1121/1.1413751

Harry Hollien, Kristen Liljegren, Camilo A. Martin, and Gea DeJong. 2001b. Production of intoxication states by actorsacoustic and temporal characteristics. *Journal of Forensic Sciences* 46, 1 (Feb. 2001), 68–73.

John Paul Hosom, Alexander B. Kain, Taniya Mishra, Jan P H Van Santen, Melanie Fried-Oken, and Janice Staehely. 2003. Intelligibility of modifications to dysarthric speech. In *Proceedings of the ICASSP*, Vol. 1. IEEE, Hong Kong, China, 924–927.

Mark Huckvale and Anne-Linn Kristiansen. 2012. Effectiveness of Electronic Voice Disguise between Friends. In *Proceedings of the 46th International Conference: Audio Forensics*. Stellenbosch, South Africa. http://www.aes.org/e-lib/browse.cfm?elib=16337

Tom Johnstone. 2001. *The effect of emotion on voice production and speech acoustics*. Phd dissertation. University of Western Australia, Psychology Department, Perth, Australia.

Tom Johnstone and Klaus R. Scherer. 1999. The effects of emotions on voice quality. In *Proceedings of the 14th International Conference of Phonetic Sciences*. San Francisco, USA, 2029–2032. http://www.keck.waisman.wisc.edu/~tjohnstone/0602.pdf

Sachin S. Kajarekar, Harry Bratt, Elizabeth Shriberg, and Rafael De León. 2006. A Study of Intentional Voice Modifications for Evading Automatic Speaker Recognition. In *Proceedings of the ODYSSEY06 – The Speaker and Language Recognition Workshop*. ISCA, San Juan, Puerto Rico, 1–6. DOI:http://dx.doi.org/10.1109/ODYSSEY.2006.248123

Ahilan Kanagasundaram, Robbie Vogt, David Dean, and Michael Mason. 2011. i-vector Based Speaker Recognition on Short Utterances. In *Proceedings of the Interspeech*. ISCA, Florence, Italy, 2341–2344.

Harleen Kaur and Ashutosh Guide Singh. 2017. *Speaker Identification of Disguised Voices Using MFCC Statistical Moment And SVM Classifier*. Ph.D. Dissertation.

Finnian Kelly, Rahim Saeidi, Naomi Harte, and David van Leeuwen. 2014. Effect of long-term ageing on i-vector speaker verification. In *Proceedings of the Interspeech*. International Speech Communication Association, Singapore, 86–90. http://www.mee.tcd.ie/~sigmedia/pmwiki/uploads/Main.Publications/finnian_interspeech14.pdf

Patrick Kenny, Pierre Ouellet, Najim Dehak, Vishwa Gupta, and Pierre Dumouchel. 2008. A study of interspeaker variability in speaker verication. *IEEE Transactions on Audio, Speech, and Language Processing* 16, 5 (2008), 980–988.

Lawrence G. Kersta. 1962. Voiceprint identification. *Nature* 4861 (Dec. 1962), 1253–1257.

Tomi Kinnunen and Paavo Alku. 2009. On separating glottal source and vocal tract information in telephony speaker verification. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing* . IEEE, Taipei, China, 4545–4548. DOI:http://dx.doi.org/10.1109/ICASSP.2009.4960641

Tatsuya Kitamura. 2008. Acoustic analysis of imitated voice produced by a professional impersonator. In *Proceedings of the Interspeech*. ISCA, Brisbane, Australia, 813–816.

Fritz Klingholz, R. Penning, and E. Liebhardt. 1988. Recognition of low-level alcohol intoxication from speech signal. *Journal of the Acoustical Society of America* 84, 3 (Sept. 1988), 929–935. DOI:http://dx.doi.org/10.1121/1.396661

Hisayoshi Kojima, Wilbur J. Gould, Anthony Lambiase, and Nobuhiko Isshiki. 1982. Computer Analysis of Hoarseness. *Acta Oto-laryngologica* 89, 3-6 (Jan. 1982), 547–554. DOI:http://dx.doi.org/10.3109/00016488009127173

Jody Kreiman and Bruce R. Gerratt. 2005. Perception of aperiodicity in pathological voice. *Journal of the Acoustical Society of America* 117, 4 (May 2005), 2201–2211. http://www.ncbi.nlm.nih.gov/pubmed/15898661

Hermann J. Kunzel. 2000. Effects of voice disguise on speaking fundamental frequency. *Forensic Linguistics* 7, 2 (Dec. 2000), 149–179. DOI:http://dx.doi.org/10.1558/sll.2000.7.2.149

Hermann J. Kunzel, Joaquín González-Rodríguez, and Javier Ortega-García. 2004. Effect of voice disguise on the performance of a forensic automatic speaker recognition system. In *Proceedings of the ODYSSEY04 – The Speaker and Language Recognition Workshop*. ISCA, Toledo, Spain, 153–156. http://www.isca-speech.org/archive_open/odyssey_04/ody4_153.html

Yee W. Lau, Dat Tran, and Michael Wagner. 2004. Vulnerability of speaker verification to voice mimicking. In *Proceedings of the International Symposium on Intelligent Multimedia, Video and Speech Processing*. Hong Kong, China, 145–148. DOI:http://dx.doi.org/10.1109/ISIMP.2004.1434021

Yee W. Lau, Dat Tran, and Michael Wagner. 2005. Testing voice mimicry with the YOHO speaker verification corpus. In *Proceedings of the International Conference on Knowledge-Based Intelligent Information and Engineering Systems (Lecture Notes in Computer Science)*, Vol. 3684. Springer, Melbourne, Australia, 15–20. DOI:http://dx.doi.org/10.1007/11554028_3

John Laver. 1994. *Principles of Phonetics*. Cambridge University Press, Cambridge.

Xi Li, Jidong Tao, Michael T. Johnson, Joseph Soltis, Anne Savage, Kirsten M. Leong, and John D. Newman. 2005. Stress and emotion classification using jitter and shimmer features. In *Proceedings of the ICASSP*, Vol. 4. Honolulu, Hawaii, 1081–1084.

Johan Lindberg and Mats Blomberg. 1999. Vulnerability in speaker verification. A study of technical impostor techniques. In *Proceedings of the Eurospeech*. Budapest, Hungary, 1211–1214.

Sue Ellen Linville. 2001. *Vocal Aging*. Singular Publishing Group, San Diego.

Robert C. Lummis and Aaron E. Rosenberg. 1972. Test of an automatic speaker verification method with intensively trained professional mimics. *Journal of the Acoustical Society of America* 51, 131 (Jan. 1972). DOI:http://dx.doi.org/10.1121/1.1981415

Evangeline Machlin. 1975. *Dialects for the stage*. Routledge/Theater Arts, New York.

John Makhoul. 1975. Linear Prediction: A Tutorial Review. *Proc. IEEE* 53, 4 (April 1975), 561–580. DOI:http://dx.doi.org/10.1109/PROC.1975.9792

Duncan Markham. 1997. *Phonetic Imitation, Accent, and the Learner*. Phd dissertation. Lund University, Lund, Sweden.

Judith A. Markowitz. 1996. *Using Speech Recognition*. Prentice Hall PTR, Upper Saddle River, N.J.

Judith A. Markowitz. 2007. The Many Roles of Speaker Classification in Speaker Verification and Identification. In *Speaker Classification I*, Christian Mueller (Ed.). Springer Berlin Heidelberg, Berlin, Heidelberg, 218–225.

Mikiko Mashimo, Tomoki Toda, Hiromichi Kawanami, Kiyohiro Shikano, and Nick Campbell. 2002. Evaluation of cross-language voice conversion using bilingual and non-bilingual databases. In *Proceedings of the ICSLP*. Denver, CO, 293–296.

Mikiko Mashimo, Tomoki Toda, Kiyohiro Shikano, and Nick Campbell. 2001. Evaluation of cross-language voice conversion based on GMM and STRAIGHT. In *Proceedings of the Eurospeech*. Aalborg, Denmark, 361–364.

Takashi Masuko, Keiichi Tokuda, and Takao Kobayashi. 2000. Imposture using synthetic speech against speaker verification based on spectrum and pitch. In *Proceedings of the ICSLP*, Vol. 2. Beijing, China, 302–305.

Driss Matrouf, Jean-François Bonastre, and Corinne Fredouille. 2006. Effect of speech transformation on impostor acceptance. In *Proceedings of the ICASSP*, Vol. 1. Tolouse, France, 933–936.

Yuri Matveev. 2013. The Problem of Voice Template Aging in Speaker Recognition Systems. In *Speech and Computer. Proceedings of the 15th International Conference, SPECOM 2013*, Miloš Železný, Ivan Habernal, and Andrey Ronzhin (Eds.). Lecture Notes in Computer Science, Vol. 8113. Springer International Publishing, Pilsen, Czech Republic, 345–353. DOI:http://dx.doi.org/10.1007/978-3-319-01931-4_46

Florian Metze, Jitendra Ajmera, Roman Englert, Udo Bub, Felix Burkhardt, Joachim Stegmann, Christian Muller, Richard Huber, Bernt Andrassy, Josef G. Bauer, and Bernhard Littel. 2007. Comparison of four approaches to age and gender recognition for telephone applications. In *Proceedings of the ICASSP*, Vol. 4. IEEE, 1089–1092. http://ieeexplore.ieee.org/xpl/abstractAuthors.jsp?arnumber=4218294

Dirk Michaelis, Matthias Frohlich, Hans Werner Strube, Eberhard Kruse, Brad Story, and Ingo R. Titze. 1998. Some simulations concerning jitter and shimmer measurement. In *Proceedings of the International Workshop on Advances in Quantitative Laryngoscopy*. Aachen, Germany, 744–754. http://www.dpi.physik.uni-goettingen.de/~micha/aachen98/aachen98.html

Seyed Hamidreza Mohammadi and Alexander Kain. 2014. Voice conversion using deep neural networks with speaker-independent pre-training. In *Proceedings of the IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 19–23.

Iain R. Murray and John L. Arnott. 1993. Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. *Journal of the Acoustical Society of America* 93, 2 (Feb. 1993), 1097–1108. DOI:http://dx.doi.org/10.1121/1.405558

Toru Nakashika, Ryoichi Takashima, Tetsuya Takiguchi, and Yasuo Ariki. 2013. Voice conversion in high-order eigen space using deep belief nets.. In *Proceedings of the Interspeech*. 369–372.

M. Laxmi Narayana and Sunil Kumar Kopparapu. 2009a. Effect of noise-in-speech on MFCC parameters. In *Proceedings of the 9th WSEAS International Conference on Signal, Speech and Image Processing, and 9th WSEAS International Conference on Multimedia, Internet and Video Technologies*. ACM, Stevens Point, Wisconsin, 39–43.

M. Laxmi Narayana and Sunil Kumar Kopparapu. 2009b. On the use of stress information in speech for speaker recognition. In *Proceedings of the IEEE Region 10 Conference — TENCON 2009*. IEEE, Singapore, 1–4. DOI:http://dx.doi.org/10.1109/TENCON.2009.5396003

Barbara Peskin, Jiri Navrátil, Joy Abramson, Doug Jones, David Klusáček, Douglas A. Reynolds, and Bing Xiang. 2003. Using prosodic and conversational features for high-performance speaker recognition: report from JHU WS02. In *Proceedings of the ICASSP*, Vol. 4. IEEE, Hong-Kong, China, 792–795. DOI:http://dx.doi.org/10.1109/ICASSP.2003.1202762

Jeff Pittam. 1994. *Voice in Social Interaction; an Interdisciplinary Approach*. SAGE Publications, Thousand Oaks.

Manfred Putzer and Jacques Koreman. 1997. A German database of patterns for vocal fold vibration. *Phonus 3, Institute of Phonetics, University of Saarland* (1997), 143–153.

Lawrence Rabiner and Biing-Hwang Juang. 1993. *Fundamentals of Speech Recognition*. Prentice Hall, Inc., Englewood Cliffs, New Jersey.

Alan R. Reich. 1981. Detecting the presence of vocal disguise in the male voice. *Journal of the Acoustical Society of America* 69, 5 (July 1981), 1458–1460. DOI: http://dx.doi.org/10.1121/1.385778

Douglas A. Reynolds, Walter D. Andrews, Joseph Campbell, Jiri Navrátil, Barbara Peskin, André Adami, Qin Jin, David Klusáček, Joy Abramson, Radu Mihaescu, Jack Godfrey, Doug Jones, and Bing Xiang. 2002. *Exploiting high-level information for high-performance speaker recognition*. SuperSID project final report. MIT Lincoln Laboratory, US Department of Defense, IBM, International Computer Science Institute, Oregon Graduate Institute, Carnegie Mellon University, Charles University, York University, Princeton University, Cornell University, Baltimore, MD,.

Douglas A. Reynolds, Walter D. Andrews, Joseph Campbell, Jiri Navrátil, Barbara Peskin, André Adami, Qin Jin, David Klusáček, Joy Abramson, Radu Mihaescu, Jack Godfrey, Doug Jones, and Bing Xiang. 2003. The SuperSID project: exploiting high-level information for high-accuracy speaker recognition. In *Proceedings of the ICASSP*, Vol. 4. IEEE, Hong-Kong, China, 784–787. DOI: http://dx.doi.org/10.1109/ICASSP.2003.1202762

Douglas A. Reynolds, Marc A. Zissman, Thomas F. Quatieri, and Gerald C. OLeary. 1995. The effects of telephone transmission degradations on speaker recognition performance. In *Proceedings of the ICASSP*. IEEE, Detroit, MI, 329–332. DOI: http://dx.doi.org/10.1109/ICASSP.1995.479540

Robert D. Rodman. 1998. Speaker recognition of disguised voices: A program for research. In *Proceedings of the Consortium on Speech Technology Conference on Speaker Recognition by Man and Machine: Directions for Forensic Applications*. Ankara, Turkey, 9–22.

Robert D. Rodman and Michael S. Powell. 2000. Computer Recognition of Speakers Who Disguise Their Voice. In *Proceedings of the International Conference on Signal Processing Applications & Technology 2000*. Dallas.

William J. Ryan and Kenneth W. Burk. 1974. Perceptual and acoustic correlates of aging in the speech of males. *Journal of Communication Disorders* 7, 2 (June 1974), 181–192. DOI: http://dx.doi.org/10.1016/0021-9924(74)90030-6

Nicolas Scheffer, jean François Bonastre, Alain Ghio, and Bernard Teston. 2001. Gémellité et reconnaissance automatique du locuteur. In *Proceedings of the XXVth Journées d'Etude sur la Parole (Lecture Notes in Computer Science)*. Fes, Morocco, 445–448. https://hal.archives-ouvertes.fr/hal-00134198

Klaus R. Scherer. 1986. Vocal affect expression: A review and a model for future research. *Psychological Bulletin* 99, 2 (March 1986), 143–65. http://www.affective-sciences.org/system/files/biblio/1986_Scherer_PsyBull.pdf

Klaus R. Scherer, Robert D. Ladd, and Kim E. A. Silverman. 1984. Vocal cues to speaker affect: Testing two models. *Journal of the Acoustical Society of America* 76, 5 (June 1984), 1346–1356. http://www.affective-sciences.org/system/files/biblio/1984_Scherer_JASA.pdf

Astrid Schmidt-Nielsen and Thomas H. Crystal. 2000. Speaker Verification by Human Listeners: Experiments Comparing Human and Machine Performance Using the NIST 1998 Speaker Evaluation Data. *Digital Signal Processing* 10, 1–3 (Jan. 2000), 249–266. DOI: http://dx.doi.org/10.1006/dspr.1999.0356

Susanne Schoetz. 2007. Acoustic Analysis of Adult Speaker Age. In *Speaker Classification I*, Christian Mueller (Ed.). Vol. 4343. Springer Berlin Heidelberg, Berlin, Heidelberg, 88–107.

Stephen Shum, Najim Dehak, Reda Dehak, and James R. Glass. 2010. Unsupervised speaker adaptation based on the cosine similarity for text-independent speaker verication. In *Proceedings of the ODYSSEY10 – The Speaker and Language Recognition Workshop*. ISCA, Brno, Czech Republic, 76–82.

Roger W. Shuy. 1990. Dialect as evidence in law cases. *Journal of English Linguistics* 23, 1 (April 1990), 195–208. DOI: http://dx.doi.org/10.1177/007542429002300116

Milan Sigmund. 2008. Automatic Speaker Recognition by Speech Signal. In *Frontiers in Robotics, Automation and Control*, Alexander Zemliak (Ed.). InTech. DOI: http://dx.doi.org/10.5772/6333

Kemal Sonmez, Elizabeth Shriberg, Larry P. Heck, and Elizabeth Weintraub. 1998. Modeling dynamic prosodic variation for speaker verification. In *Proceedings of the fifth International Conference on Spoken Language Processing*, Vol. 7. Sydney, Australia, 3189–3192.

Kenneth N. Stevens, Carl E. Williams, Jaime R. Carbonell, and Barbara Woods. 1968. Speaker authentication and identification: a comparison of spectrographic and auditory presentations of speech material. *Journal of the Acoustical Society of America* 44, 6 (Dec. 1968), 1596–1607. DOI: http://dx.doi.org/10.1121/1.1911302

Lucian Sulica. 2011. Hoarseness. *Archives of Otolaryngology - Head and Neck Surgery* 137, 6 (June 2011), 616–619. DOI: http://dx.doi.org/10.1001/archoto.2011.80

Kirk P.H. Sullivan and Jason Pelecanos. 2001. Revisiting Carl Bildts impostor: Would a speaker verification system foil him?. In *Proceedings of the International Conference on Audio- and Video-Based Biometric*

*Person Authentication (Lecture Notes in Computer Science)*, Vol. 2091. Springer, Halmstad, Sweden, 144–149. DOI: http://dx.doi.org/10.1007/3-540-45344-X_21

Bradford L. Swartz. 1992. Resistance of voice onset time variability to intoxication. *Perceptual and Motor Skills* 75, 2 (Oct. 1992), 415–424.

Tiejun Tan. 2010. The effect of voice disguise on Automatic Speaker Recognition. In *Proceedings of the 3rd International Congress on Image and Signal Processing (CISP)*. IEEE, Yantai, China, 3538–3541. DOI: http://dx.doi.org/10.1109/CISP.2010.5647131

Shahrukh K. Taseer. 2005. Speaker Identification for Speakers with Deliberately Disguised Voices using Glottal Pulse Information. In *Proceedings of the Pakistan Section Multitopic Conference*. IEEE, Karachi, Pakistan, 1–5. DOI: http://dx.doi.org/10.1109/INMIC.2005.334384

Oscar Tosi, Herbert Oyer, William Lashbrook, Charles Pedrey, Julie Nicol, and Ernest Nash. 1972. Experiment on Voice Identification. *Journal of the Acoustical Society of America* 51, 6B (June 1972), 2030–2043. DOI: http://dx.doi.org/10.1121/1.1913064

Renetta Garrison Tull and Janet C. Rutledge. 1996. Automatic speaker recognition based on pitch contours. *Acoustical Society of America 131st Meeting Lay Language Papers* (May 1996).

Lior Uzan and Lior Wolf. 2015. I know that voice: Identifying the voice actor behind the voice. In *Proceedings of the International Conference on Biometrics (ICB)*. IEEE, 46–51.

Ratree Wayland, Scott Gargash, and Allard Longman. 1995. Acoustic and perceptual investigation of breathy voice. *Journal of the Acoustical Society of America* 97, 5 (May 1995), 3364. DOI: http://dx.doi.org/10.1121/1.413011

Frederik Weber, Linda Manganaro, Barbara Peskin, and Elizabeth Shriberg. 2002. Using prosodic and lexical information for speaker identification. In *Proceedings of the ICASSP*, Vol. 1. IEEE, Orlando, FL, USA, 141–144. DOI: http://dx.doi.org/10.1109/ICASSP.2002.1005696

Carl E. Williams and Kenneth N. Stevens. 1972. Emotions and Speech: Some Acoustical Correlates. *Journal of the Acoustical Society of America* 52, 4B (March 1972), 1238–1250. http://www.ohio.edu/people/leec1/documents/sociophobia/williams_stevens_1972.pdf

Frank Wittig and Christian Mueller. 2003. Implicit feedback for user-adaptive systems by analyzing the user's speech. In *Proceedings of the Workshop on Adaptivität und Benutzermodellierung in interaktiven Softwaresystemen (ABIS)*. Karlsruhe, Germany.

Tian Wu, Yingchun Yang, Zhaohui Wu, and Dongdong Li. 2006. MASC: A Speech Corpus in Mandarin for Emotion Analysis and Affective Speaker Recognition. In *Proceedings of the ODYSSEY — The Speaker and Language Recognition Workshop*. IEEE, San Juan, Puerto Rico, 1–5. DOI: http://dx.doi.org/10.1109/ODYSSEY.2006.248084

Zhizheng Wu and Haizhou Li. 2013. Voice conversion and spoofing attack on speaker verification systems. In *Proceedings of the Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2013 Asia-Pacific*. IEEE, Kaohsiung, China, 1–9. DOI: http://dx.doi.org/10.1109/APSIPA.2013.6694344

Naoaki Yanagihara. 1967. Significance of Harmonic Changes and Noise Components in Hoarseness. *Journal of the American Speech-Language-Hearing Association* 10 (Sept. 1967), 531–541.

Eiji Yumoto. 1988. Quantitative assessment of the degree of hoarseness. *Journal of Voice* 1, 4 (Jan. 1988), 310–313. DOI: http://dx.doi.org/10.1016/S0892-1997(88)80003-1

Eiji Yumoto, Wilbur J. Gould, and Thomas Baer. 1982. Harmonics-to-noise ratio as an index of the degree of hoarseness. *Journal of the Acoustical Society of America* 71, 6 (June 1982), 1544–1549. http://www.ncbi.nlm.nih.gov/pubmed/7108029

Elisabeth Zetterholm. 2003. *Voice imitation: a phonetic study of perceptual illusions and acoustic success*. Phd dissertation. Lund University, Lund, Sweden.

Elisabeth Zetterholm. 2006. Same speaker—different voices. A study of one impersonator and some of his different imitations. In *Proceedings of the 11th Australian International Conference on Speech Science and Technology*. Auckland, New Zealand, 70–75.

Elisabeth Zetterholm, Daniel Elenius, and Mats Blomberg. 2004. A comparison between human perception and a speaker verification system score of a voice imitation. In *Proceedings of the Tenth Australian International Conference on Speech Science and Technology*. Australian Speech Science and Technology Association, Sydney, Australia, 393–397. https://lup.lub.lu.se/search/publication/52907e52-0553-4228-a120-addc5e1f9d24

Cuiling Zhang and Bin Lin. 2017. Acoustic analysis of whispery voice disguise in Chinese. *The Journal of the Acoustical Society of America* 141, 5 (2017), 3982–3982.

Cuiling Zhang and Tiejun Tan. 2008. Voice disguise and automatic speaker recognition. *Forensic Science International* 175, 2–3 (April 2008), 118–122. DOI: http://dx.doi.org/10.1016/j.forsciint.2007.05.019

Sue Anne Zollinger and Henrik Brumm. 2011. The Lombard effect. *Current Biology* 21, 16 (Aug. 2011), R614–R615. DOI: http://dx.doi.org/10.1016/j.cub.2011.06.003

Received May 2016; revised —; accepted —