THE DYNAMICS OF INDEXICAL INFORMATION IN SPEECH: CAN RECOGNIZABILITY BE CONTROLLED BY THE SPEAKER?

VOLKER DELLWO, ELISA PELLEGRINO, LEI HE,
THAYABARAN KATHIRESAN

ABSTRACT

Human voices are individual and humans have elaborate skills in recognizing speakers by their voice, phenomena that are deeply rooted in the evolution of human behavior. To date, the mechanisms of speaker recognition are not well understood because of the high variability of the acoustic cues to a speaker's identity. We wondered what role the speaker plays in making his/her voice more or less well recognizable. While it is evident from the literature that humans can control vocal properties to enhance their intelligibility, it is unclear whether speakers can and/or do control vocal characteristics to be better recognizable and whether such control mechanisms play a role in the communication process. In this paper, we reviewed results from the literature supporting the view that speaker idiosyncratic information is dynamic and that humans have the ability to control how well they can be recognized. We suggest possible experimental setups by which the control over identity in voice can be tested and present pilot acoustic characteristics of speech that was produced to be either targeted at being (a) intelligible (clear speech) and (b) suitable for person recognition (identity marked speech). Results revealed that there is reason to believe that speakers apply different mechanisms when making their individuality identifiable as opposed to making their speech better understood. We discuss predictions that a control of recognizability and intelligibility has within major theories of speech perception.

Key words: indexical information, voice recognition, identity marked speech

1. Introduction

February 10th 2019: Eliza D. makes her way home through a dark subway when a masked attacker grabs her from behind and commands in a whispered, foreign-accented voice: "Give me your money, quick!". She pulls out her wallet and before she understands, the man disappears and leaves her with nothing but the memory of his voice. Months later, Eliza appears at court and identifies a suspect as her attacker based on his voice. Such scenes are common to law enforcement agencies around the world. In this

https://doi.org/10.14712/24646830.2019.18

particular case, the probability of Eliza performing a correct recognition of the suspect would be estimated as rather low, because of the short duration of the familiarization (Clifford, 1980; Kerstholt et al., 2004; Yarmey, 1995), the long time lag between familiarization and recognition (Papcun et al., 1989; Yarmey, 1995), and because the presence of a foreign accent is likely to have biased her decision (Ladefoged & Ladefoged, 1980; Stevenage et al., 2012; Yarmey, 1995). However, for Eliza this was not the first time that the recognition of an individual by his/her voice was crucial in her life, in fact, from the time she was born, there were many occasions when her survival depended on it (Kriengwatana et al., 2015; Petkov et al., 2009). She recognized her mother as a central caregiver before (Kisilevsky et al., 2003, 2009; Panneton Cooper et al., 1997) and after birth (Sullivan et al., 2011), and relied on being recognized by others to receive the right amount of attention (Locke, 2006). Eliza's remarkable voice recognition skills are an ability she shares with numerous animal species (e.g. Belin, 2006; Larranaga et al., 2015; Molnár et al., 2009; Perrodin et al., 2011, 2015). Her individual voice became part of her overall personality (e.g. McAleer et al., 2014), it supports her in building up and position herself in social groups (Schegloff, 1979), it contains information about her fertility (Fisher et al., 2011; Raj et al., 2010), it attracts the right mating partner (Bruckert et al., 2010; Collins, 2001; Collins & Missing, 2003) and contributes to the trust that others have in her (Belin et al., 2017; O'Connor & Barclay, 2017; Oleszkiewicz et al., 2017). Her voice supports listeners in paying attention to her in the environment of other voices (Johnstrude et al., 2013) and it contributes to her esthetic appearance in casual or artistic activities like singing (Doscher, 1993; Sundberg, 1977). Losing her vocal identity cues (Kurowski et al., 1996) or her ability to recognize voices (Roswandowitz et al., 2014) – for example as a result of neurological malfunction – can drive her into social isolation.

Given the significance of her voice for her social life and the consequences of a loss or change in voice identity, it is not surprising that Eliza became a frequent motive in many fictional scenarios, for example as the flower girl *Eliza Doolittle* in George Bernard Shaw's *Pygmalion* (Shaw, 1916). It is surprising, however, that theories of speech and language processing have typically treated the vocal information about her identity (henceforth: idiosyncratic cues) as information that is unwanted acoustic variability, some form of noise that needs to be cancelled out to arrive at the underlying linguistic communicative message (see discussion in Creel & Bregman, 2011). As a form of noise, idiosyncratic cues have typically been understood as static information that is given away rather involuntarily and is not under the control of the speaker. However, considering Eliza's capacity to encode an extremely rich and multidimensional amount of information in her voice, it seems implausible that she and other speakers have no control over this information. In the present article, we investigated to what degree idiosyncratic information is a by-product of the articulation process (section 2). We reviewed results from speech and speaker information processing to suggest possible control mechanisms of speakers over their idiosyncratic information (section 3), and provide reasons for why it is plausible that control mechanisms of idiosyncrasy should exist (section 4). We then provide an experimental framework and first empirical evidence revealing that idiosyncratic and linguistic information may be controlled differently, when either speaker identity or linguistic intelligibility is at stake (section 5). As a conclusion, we outline predictions that a control of idiosyncratic properties has on information processing in major theories (abstractionist and exemplar models) of speech perception (section 6).

2. How invariable is speaker idiosyncratic information?

Two types of information are typically distinguished in a speech signal: linguistic (content of the message, i.e. what is being said?) and indexical (who said what in which way? Abercrombie, 1967; Dellwo et al., 2007; Levi and Pisoni, 2007). Indexical information can be manifold. The examples about Eliza in the previous section reveal that it can serve as cues to recognize a speaker (speaker idiosyncratic information) and/or to interpret his/her situational state (speaker state information). The term 'indexical' was probably introduced by David Abercrombie (1967) to the phonetics community and goes back to the semiotic theory of Peirce (Peirce et al., 1965). In this theory, indexicality is information that specifies an object further in the context in which it occurs. For example, smoke can be indexical for the presence of a fire and, in analogy, strong de-nasalisation in speech can be indexical for the speaker suffering of a cold or a low voice can be indexical for a male gender. This usage suggests that indexical information is treated as a mere by-product of the speech production process, i.e. involuntary information without a motivated communicative intent. It consequently implies that indexical information is not controlled by the speaker.

This view might initially seem plausible for speaker idiosyncratic information, as it should support the recognition of a speaker, independent of any situational variability. Idiosyncratic information is often categorised in inborn and acquired information (Nolan, 1997), the former being a result of anatomic shapes on dimensions of the articulatory apparatus, the latter the result of acquired characteristics through exposition to particular phonetic/phonological realisations of a certain social and/or geographical environment. While acquired idiosyncrasies can to some degree be reacquired, the nature of inborn information might appear particularly static and involuntary as the anatomic dimensions of the vocal apparatus can not easily be changed and if, then only to some degree. For this reason, inborn information has been understood as a strong invariable cue to the identity of the speaker (Belin, 2006; Nolan, 1997), even though there is a general awareness that also the inborn characteristics can underlie considerable within-speaker variability. It is also well known that within-speaker variability in either inborn or acquired information, probably poses the strongest problem on most recognition scenarios. In experimental settings, this variability is referred to as 'session variability', i.e. within-speaker variability that occurs when speakers produce speech during different recording sessions between which their cues to speaker idiosyncrasy may vary naturally or as a result of environmental influences (Hansen & Hasan, 2015). Within-speaker session variability might occur from a complex interaction between speaker idiosyncratic and speaker state information (e.g. varying emotional states), it might also occur as a result of external influence (e.g. accommodation to background noise or convergence between speakers).

Within a session, little attention has been paid to the variability of idiosyncratic information. This is also true in formal speaker recognition domains. In speaker recognition technology, for example, the most recent approach – so called i-vectors or x-vectors (Dehak et al., 2011; Garcia-Romero & Espy-Wilson, 2011) – idiosyncratic information of the entire speaker is reduced to a vector of about 200 dimensions, irrespective of session variability. In forensic phonetics, a sub-field of phonetics concerned with idiosyncratic information for the purpose of solving crime, a typical task is to decide whether a speech sample of a perpetrator (evidence) and a speech sample from a suspect (comparison) were produced by one and the same or different speakers (cf. discussion in Dellwo et al., 2018a). Also in such scenarios, the between session variability is often especially strong, as evidence and comparison recordings have typically been recorded under different speaker states and in different communicative situations (for example, shouting during a crime and relaxed telephone conversation during surveillance recording). The variability of a speaker within a session is typically not paid the same attention to.

The lack of attention to within-session idiosyncratic variability has recently been identified as some of the central problems in speaker recognition technology ("The speech signal is taken with uniformity", Sriram Ganapathy, personal communication & presentation at *Interspeech* conference 2018), thus a higher attention to selective detail within a session might enhance the recognition performance significantly. This view is supported by findings revealing that vowels and nasals are better suitable for automatic recognition compared to other consonants (Amino & Arai, 2009; Amino et al., 2009; Moez et al., 2016). Similar awareness is present in forensic speaker comparison, where vocal features such as fundamental (f_o) or formant frequency (F_1, F_2, etc.) characteristics are not seldom viewed as average statistics for a speaker in a session and are used to characterise this particular speaker (e.g. de Jong et al., 2007; Hudson et al., 2007). Here, the dynamics of formant characteristics have been pointed out to reveal a high amount of detail about the speaker at different points in time (He et al., 2019; McDougall & Nolan, 2007).

A novel view to within-person variability has been suggested by Burton et al. (2016) in the domain of facial identity cues. They argue that the acquisition of variability in a face is central to understanding how the face varies, which in return is central to the recognition process. It means, knowing more about the variability of a face helps a viewer to recognize this face under many different settings. This is highly plausible because obtaining data from a face under various different angles increases the probability that one can recognize the face under this particular position. For voices, this point of view has been taken up by Lavan et al. (2019). They argue that within-speaker variability in speech is an informative signal of individuality which means that obtaining a high amount of variability form vocalisations during an initialisation phase should support the speaker recognition process. This view is not readily in agreement with a forensic phonetic claim, which holds that variables best suitable for speaker recognition are those that offer high between-speaker variability and low within-speaker variability (Hansen & Hasan, 2015; Nolan, 1997). According to Burton et al. and Lavan et al., high within-speaker variability should provide necessary recognition information. While this approach seems highly relevant for understanding auditory speaker variability, the limitations for formal forensic scenarios are evident, as the amount of available speech data is typically too small to derive strong models about speakers' idiosyncratic cue variability. Regarding the question of the present article – whether speaker-specific information can be controlled – it seems plausible that the potential for control mechanisms is increased by an increased signal variability. Static information is typically of low entropy characterised by a low number of degrees of freedom for control. So if speakers have control over their vocal identity information, it seems plausible that they can control cues to the variability patterns of individuality. In the following section, we discuss theoretical frameworks with which such a control might be reached.

3. Possible control mechanisms of acoustic identity cues in speech

There is strong evidence from the niche field of forensic phonetics, revealing that speakers can deliberately change individuality cues to disguise their voice with more or less success (Eriksson, 2010; Eriksson & Wretling, 1997) by a sometimes seemingly random manipulation of their cues to individuality or by imitating cues to the individuality of other speakers (De Figueiredo et al., 1996; Eriksson & Wretling, 1997; Hirson & Duckworth, 1993; Hove & Dellwo, 2014; Kitamura, 2008; Růžičková & Skarnitzl, 2017; Wagner & Köster, 1999). This means that speakers can hide information relevant to their identity and they have some intuitive consciousness about which of the inborn information (e.g. fundamental frequency) and the acquired information (e.g. regional accent) needs to be chosen for this. Speakers can also imitate or caricature other speakers' voices more or less convincingly (Jansen et al., 2001; Klewitz & Couper-Kuhlen, 1999). This requires an awareness of speakers towards the idiosyncratic characteristics that are crucial to other speakers' identity. Professional actors can typically well control their vocal identity in acting a fictional character and maintain this constantly, sometimes over a variety of characters. Good examples are the German writer and actor Marc-Uwe Kling reading from his own books and changing voices between different characters or the actress Melissa Rauch playing the character 'Bernadette' in the US TV Show 'Big Bang Theory' which is distinctly different from the actress' non-acted voice. In summary, speakers can conceal their identity and they can imitate the identity of others. This demonstrates that speakers have some control over indexical cues. But can they also control cues to make themselves more recognizable?

Idiosyncratic information is sometimes at places that have been found to be less relevant for the encoding of linguistic information as, for example, coarticulatory parts of the signal between two segments. This view, however, is problematic, as coarticulatory transitions not seldom contain important cues to parse the linguistic message and idiosyncratic information is often part of the same cues that encode linguistic meaning (e.g. formant frequencies might vary relatively between speakers which is a cue to linguistics and speaker alike). Here we argue that the cues to idiosyncrasy can most likely be found intertwined with linguistic cues (Creel & Bregman, 2011); possibly a binary distinction between the cues is not even sensible. We find that there are two phenomena that play a role in controlling idiosyncrasy, (a) the choices over segments or prosodic patterns as idiosyncratic categories as well as within-segment acoustic variability and (b) variability in speaking styles that might make more or less use of the segmental/prosodic control mechanisms.

3.1. Choices and realisation of segments

As mentioned above, vowels and nasals reveal better speaker recognition performance compared to other speech segments (Amino & Arai, 2009; Amino et al. 2009; Moez et al., 2016). It thus seems plausible that a selective choice of segmental features might support recognition. This could be reached by a selective choice of words in which segments revealing stronger idiosyncrasies occur. It is unclear, however, whether the careful and intricate planning of linguistic information would allow such choices to a considerable degree. Given that vowels contain a higher amount of speaker-specific information, a more applicable mechanism is a long clear realisation of vowels as opposed to reduction or elision. Reducing vowels to schwa or even consonants is a technique that is widely distributed throughout the world's languages, in particular in unstressed syllables. It should also be possible to change vowel qualities to contain more or less amounts of idiosyncratic information. Techniques to make voices more or less recognizable in vocalic utterances could be viewed in a very similar way as the production of clear speech that is targeted at a higher signal intelligibility (Smiljanic & Bradlow, 2008; Hazan et al., 2012). The question of whether clear-speech and production targeted at idiosyncrasy should underlie the same mechanisms is therefore discussed in section 5.

Idiosyncratic information may also be distributed differently over time. He & Dellwo (2017) showed that within-syllable temporal information leading to the syllable nucleus is less variable compared to the temporal information between nucleus and offsets. They relate this to a lesser amount of articulatory control during the final part of the syllable that may reveal more system specific movement resonances (e.g. jaw movements). Such findings could also be replicated for the temporal development of formant frequencies (He et al., 2019). It seems probable that speakers should be able to control such characteristics by enunciating syllables in a more or less controlled way. Such temporal differences should be more salient in speech that is casually produced compared to speech in which temporal properties of syllables are more controlled (e.g. infant or child directed speech).

Given the results from facial recognition (Burton et al., 2016), Dellwo et al. (2018b) investigated whether more information about the human vocal tract aids recognition. Facial variability is transmitted through the visual channel and vocal variability through the acoustic channel. If facial variability contains cues to identity in the visual signal, then vocal tract variability – in analogy – should contain cues to identity in the speech signal. Dellwo et al. tested this hypothesis by comparing vowels with a sweeping f_o to vowels with steady state f_o. The latter leads to a sweeping of all harmonics in the vocal tract. In acoustic terms, this means that any fine detail of the vocal tract transfer function is sampled over a small period of time, while a steady state f_o predominantly samples the characteristics of the transfer function at the harmonic peaks. Consequently, this means that swept f_o in vowels should contain more fine speaker-specific detail about the vocal tract anatomy. Computers and human listeners were trained in this experiment with sentence utterances. Speaker recognition performance was tested with vocalic utterances of the test speakers that were either steady-state at low level (lvlo), mid-level (lvmd) or high-level (lvhi) pitch or with a sweeping fundamental at falling (fall), falling-rising (fari) or rising (rise) pitch. Results showed that the computer model as well as human listeners performed significantly poorer in speaker recognition for vowels with steady state compared to sweeping f_0, but the recognition performance for humans did not show such differences. The lack of an effect for humans was reasoned in a particular choice of a stimulus subset (15 speakers for computers compared to 4 speakers for humans), since humans cannot easily be tested on a large number of speakers while computers can. The analysis of the human data further revealed a high complexity as humans use multiple different time and frequency domain cues while machines rely predominantly on short term spectral information. Most importantly, humans pay strong attention to fundamental frequency which varies across low, mid and high tones and was often used as a cue to speakers' average f_0. In summary, there are plausible reasons to believe that particular realizations of vowels with more or less fundamental frequency variability contain more or less idiosyncratic speaker detail. While such effects still need to be shown for human listeners there is first evidence that computer recognition can profit from this variability.

3.2. Control of speaker-specific detail by controlling speaking styles

The control of segmental choices and the segmental realisations vary drastically with speaking styles. Some speaking styles contain more variability in f_0 than others which is why it seems plausible that maintaining certain speaking styles can have positive effects on recognizability. In an experiment with charismatic speech typical for politicians, Rosenberg and Hirschberg (2005) found that recognition performance of speakers is related to voice charisma. Other research argued that distinctive voices have recognition advantages (Foulkes & Barron, 2000; Skuk & Schweinberger, 2013). The effect of the speaker's voice characteristics extends also to word recognition (Goldinger, 1996; Kraik & Kirsner, 1974). Recognition memory for words has been shown to be increased by voice congruence between study and test (Campeanu et al., 2014) which implies that producing a charismatic or distinctive voice in public speaking has certain advantages for the content of utterances to be remembered. In other words, speakers wishing to increase the probability that their identity is remembered in connection with their verbal conten – for example politicians in a debate advertising for their ideas – should maintain a stable charismatic voice. Given the findings in Dellwo et al. (2018b; cf. discussion in the previous section), speaking styles containing high degrees of fundamental frequency variability might be particularly prone to contain a large amount of idiosyncratic detail about the vocal-tract. Such a speaking style is, for example, infant directed speech (IDS) and there is first evidence that there is a recognition advantage when speaker-specific detail is acquired through IDS (Kathiresan et al., 2019). The fact that infants are often addressed in IDS might thus support their ability to acquire the mother's voice with a high amount of variability as this variability contains highly salient cues to the speaker (Burton et al., 2016; Lavan et al., 2019).

4. Why should speakers control their acoustic cues to identity in speech?

The reasons for controlling identity in speech can be manifold. One obvious reason might be when identity is at stake in a forensic investigation or when speakers intend to imitate others for artistic reasons or for identity fraud. While such situations are interesting and in need of scientific clarification, they are possibly far from being part of everyday social communicative situations. The reason for identity control is likely to be a much more integral part of voice communication. We argue that one of the prime reasons to control idiosyncrasy lies in the fact that the information about who is speaking is crucial for structuring and understanding the linguistic message in speech. The identity of the speaker also allows many assumptions about the structure and content of the utterance, which provide abundant information relevant for top-down processing. For example, a speaker using the words ‘you know’ very frequently will not need to pronounce these words very clearly for listeners to understand them. In dialogue processing, the absence of voice identity cues might make the dialogue ambiguous at best. The following dialogue utterances (left) might have been carried out by two speakers (middle) or by three speakers (right):

Possible dialogue utterances:	**Interpretation I:**	**Interpretation II:**
How much is this?	**Buyer A:** How much is this?	**Buyer A:** How much is this?
Let’s say three dollars.	**Seller:** Let’s say three dollars.	**Seller:** Let’s say three dollars.
Oh, that’s expensive.	**Buyer A:** Oh, that’s expensive.	**Buyer A:** Oh, that’s expensive.
What about two?	**Seller:** What about two	**Buyer B:** What about two?
OK, let’s call it a deal!	**Buyer A:** OK, let’s call it a deal!	**Seller:** OK, let’s call it a deal!

Without voice processing abilities a listener could only make informed guesses about the speakers, e.g. based on the linguistic structure or cues to turn-taking. This means, the lack of speaker information makes a sensible processing of the dialogue impossible, in particular since there are several possible ways in which it could be read. In interpretation I (middle), it is most likely that A bought the item from the seller, in dialogue B it seems more plausible that B bought the item. Assuming that this dialogue was part of a radio play where speakers are not visible, listeners rely exclusively on vocal cues to identity for the correct processing.

Some circumstances make the present example very particular. In voice recognition, two major tasks are typically distinguished, first, the recognition of familiar voices and second, the discrimination of unfamiliar voices (cf. Stevenage, 2017 for a review). Both tasks might seem highly related but there is strong neurological evidence that they are separate processes (Belin & Zatorre, 2003; Latinus et al., 2011; von Kriegstein & Giraud, 2004; von Kriegstein et al., 2005) and it seems natural that the ability to discriminate should precede recognition but there is strong counter evidence (cf. discussion in Kreimann & Sidtis, 2011). The recognition of familiar voices requires previous exposure to a speaker during which other identity related features (e.g. name or face) are brought into relation with voice. In the discrimination of unfamiliar voices, the identity of the speaker is irrelevant, it only requires the ability to tell one voice from another. Phonagnostic listeners – i.e. listeners with impaired voice processing abilities (van Lancker et al., 1988; Roswandowitz et al., 2014) – also provide evidence for the view that voice recognition and discrimination are separate processes, as they are typically impaired in recognition but less in discrimination.

In the present example, it seems that voice recognition and discrimination are no longer easy to separate. To understand the dialogue, it is first of all essential to discriminate between voices to perceive the change in speaker. When arriving at the boundary between the third and the fourth utterance, discrimination is no longer sufficient. The listener will have to be able to remember, whether the voice from the fourth utterance is the same as the voice from the third utterance or not. This can only be solved by recognizing voices with which the listener had just been familiarized (henceforth: just-familiar voices). It requires that the listener has already created an abstract representation of the speaker the first time he/she listens to an utterance for each and every speaker in the dialogue. For the speakers – in return – it means that if they have the intention to be processed correctly in the dialogue they will have to find strategies to make themselves more recognizable, for example, by marking their voice more distinctive, i.e. use individuality cues to be better recognizable in the dialogue. While visible cues in natural dialogue situations might heavily support speaker recognition, in particular of just-familiar voices, there are numerous situations in which the visual attention of a listener is not directed towards each speaker, thus it must be assumed that audible cues play an equally important role.

The recognition of just-familiar voices will increase in difficulty with an increasing number of speakers. Interestingly, recent animal studies showed that indexical properties are related to population sizes: smaller populations have less need to distinguish themselves from each other compared to larger populations in voice recognizing animal populations (Pollard & Blumstein, 2011). While Pollard and Blumstein showed differences in idiosyncratic characteristics of unrelated populations of different animal species, such findings give rise to the idea that within populations the need for individualisation might increase with increasing numbers of participants, in particular in humans. Thus the need for individualisation in a dialogue situation as described above is even stronger with higher numbers of participants to maximise the chance that just-familiar voices can be reliably used for the processing of the dialogue. Such situations might occur in families with larger numbers of offspring and gives rise to the hypothesis that children growing up in larger families or environments with numerous peers (e.g. in an orphanage) should develop higher idiosyncratic, possibly more charismatic voices compared to children growing up individually. In analogy, children in smaller classrooms might be less idiosyncratic compared to children in larger classroom environments. Additionally, it might be that extrovert children in classrooms develop a particular amount of idiosyncrasy to make themselves more distinct and recognizable from their peers. Such situations might also occur situationally, e.g. in debates with varying numbers of speaker, in particular in the absence of visual cues. Politicians debating in a radio programme, for example, might produce voices in a particularly idiosyncratic way when they are debating with a larger number of others as opposed to being interviewed on their own or debating with one single peer.

5. An experimental design to study control mechanisms of acoustic cues to identity

Linguistic information is known to be highly dynamic. Speakers can choose to a high degree which words they use, otherwise encodings of messages would be problematic. To increase the success of linguistic information encoding, it has been demonstrated repeatedly that speakers can use mechanisms to make speech more intelligible, for example, under adverse listening conditions. This leads to a speaking style referred to as 'clear speech' (Hazan et al., 2012; Smiljanic and Bradlow, 2008). Clear speech is characterised by hyper-articulated segmental and prosodic characteristics. There is strong evidence showing that clear speech is more intelligible compared to so called conversational speech (Hazan & Baker, 2011) and that speakers can rapidly adapt their vocalisations to the particular needs of the listener (Burnham et al. 2002; Hansen & Hasan, 2015; Kemper et. al., 1998; Knoll et al. 2015). This means that speakers are aware of canonical acoustic forms that are essential to encode linguistic information and that they can control and adapt them depending on the situation. It seems to be the case that such control mechanisms could be identical to mechanisms described in (3) that make speech containing more or less idiosyncratic information. For this reason, we wanted to know whether speakers use identical mechanisms in increasing acoustic information to their identity when it is at stake (henceforth: identity marked speech) or when intelligibility is at stake (clear speech). We tested this with a mock speaker and speech recognition system. Speakers were asked to train either a speech or a voice recognition system by providing read utterances. They would hence need to test the system by reading a sentence from a screen and the system would either identify them (voice recognition) or recognize the linguistic message of the sentence on the screen (speech recognition). The system would randomly respond with an error to make the participant try to enunciate the utterance differently to obtain a higher speech or voice recognition success respectively. In the case of speech recognition, we expected typical clear speech realisations, for voice recognition it was unclear whether the identity marked speech that speakers would apply differs systematically from clear speech to make themselves better recognized.

5.1. Method

We recorded two male speakers at three different occasions. First, speakers were told that they would be recorded to train a speech technology system that we were developing. Speakers read 7 sentences into the system. Second, speakers were explained that part of the system was a speech recognizer which has problems recognizing speech correctly. Speakers would read sentences repeatedly into the system and the system would make them repeat sentences between 3 and 5 times before it would respond with the correct answer. Third, speakers were told that another part of the development was a speaker recognition system. Again, they read the sentences repeatedly until the system recognized them. The order of the experiment was balanced between the two speakers (i.e. one speaker carried out the speaker recognition first, the other vice versa). For the analysis we used the last repetition of the productions (i.e. n=42: 2 speakers * 3 styles [training, voice recognition, speech recognition] * 7 sentences). We carried out an acoustic analysis of the speech recordings in which we obtained measures of the total utterance durations, source signal characteristics (f_o mean and standard deviation) and vocal tract resonance characteristics (long-term F_1 and F_2 as well as F_1 standard deviation). Because of the small number of tokens (n=42) we refrained from using significance tests and based the analysis on a descriptive inspection of the variables analysed.

5.2. Results and discussion

Inspection of the data (Fig. 1) showed the typical acoustic differences between clear and conversational speech (here: read speech), for example, longer total duration of the utterance (i.e. slower speech rate in terms of syllable/seconds) and higher f_o. The standard deviation of F_1 is lower in clear speech, indicating a more stable formant frequency. For identity marked speech the f_o was higher than in the read training speech but lower than in clear speech, as was the total utterance duration (Fig. 2 top left and centre). From this result, it could be concluded that clear speech was just a stronger form of identity marked speech. However, looking at f_o variability (top right), we observed that this had a tendency to be higher than both read and clear speech. In fact, f_o variability in clear speech was comparatively low compared to its high f_o mean. This again confirms a typical low variability in some prosodic variables in clear speech. Looking at average long-time formants 1 and 2 (bottom left and center), we found that F_2 was comparatively high in identity marked speech, while formant variability of F_1 (standard deviation; bottom right) was lowest of all styles. This suggests that overall the vocal tract might have been shortened in identity marked speech compared to clear and read speech, leading to higher average long-term formants. An auditory analysis of the results additionally revealed that coarticulation in identity marked speech was stronger than in clear speech, where individual sounds were better identifiable as segments and which was rhythmically more staccato-like, putting emphasis on individual vowels. Such effects are difficult to quantify acoustically but it is plausible that speakers might want to maintain their coarticulation in ID marked speech as it contains rich information about individual articulation.

The tendency in identity marked speech to have a higher f_o variability might also be related to a possible mechanism by which individuals produce a larger amount of f_o variability to increase the information about their vocal tract characteristics (see section 3.1; Dellwo et al., 2018b).

In summary, the study provides first evidence of the acoustic characteristics of clear and identity marked speech based on a novel method that directly contrasts the two speaking styles in a human-computer interaction task. Given the small amount of data obtained thus far, it is difficult to draw safe conclusions but the data supports the view that speakers apply different techniques in counter acting situations in which their identity is at stake as opposed to situations in which they are not understood. The results motivate larger systematic studies to better understand the differences. It is unclear what influence the human-computer interaction can have on the realisation of the styles and whether human-human interaction would lead to similar results. It also seems plausible to involve participants in human-human interaction, e.g. over the telephone where in one case they are not being understood and in another case not recognized. The strong

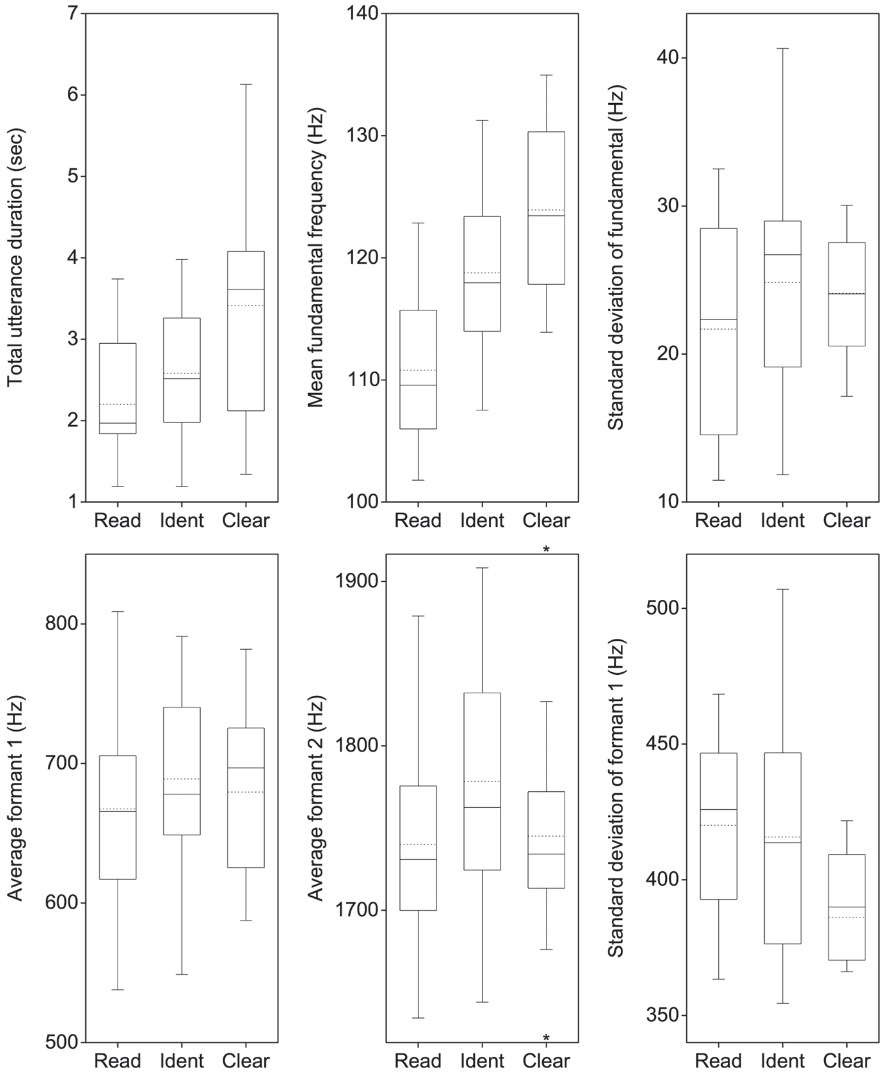

Figure 1. Distributions of six acoustic variables for each read, identity marked and clear speech: duration of utterance, fundamental frequency (f_0) mean, f_0 standard deviation, mean first formant (F_1), mean second formant (F_2) and F_1 standard deviation. All acoustic variables were calculated for each sentence utterance (n=7 in each box-plot).

advantage of the human-computer interaction is that it provides a plausible scenario in which speakers produce utterances of identical lexical content and structure by reading them. In human-human interaction, speakers would most likely have a spontaneous interaction on the telephone. This in return introduces a large amount of variability between utterances that needs to be counterbalanced by larger numbers of recordings and conversations.

6. An experimental framework for studying the dynamics of indexical information

In the previous section we saw that clear and identity marked speech are likely characterised by different acoustic features. These characteristics should help the intelligibility of the signal in the case of clear speech and they should support recognition of speakers in the case of identity marked speech. The effects of clear speech on intelligibility has been demonstrated repeatedly in the literature but the effects of identity marked speech on voice recognizability are unknown. Future research will show whether the mechanisms applied in situations in which speakers are not recognized can actually improve this situation. This could be tested by recognition experiments with humans and/or computers and the hypothesis would be that identity marked speech should lead to higher recognition rates of the speaker under conditions in which a listener has been trained on identity marked speech but possibly also under any recognition condition. Such experiments are interesting in respect to major theories of speech perception, which are probably divided by exactly the role of linguistic and indexical information in the speech signal: On one end of the scale are abstractionist theories (McClelland & Elman, 1986; Norris, 1994) which are mainly based on distinctions between language internal and external information (de Saussure, 1916: langue and parole; Chomsky, 1965: competence and performance). In these theories, indexical information is typically viewed as obstructing information (noise) that needs to be factored out of the signal to arrive at the abstract underlying linguistic forms (e.g. phonemes, words, utterances). Many phonetic theories are in line with this, viewing indexical information as a by-product of the articulation process which is an obstacle that listeners need to overcome to process the linguistic content (e.g. Fant, 1975; Liberman et al., 1967; Liberman & Mattingly, 1985). Vowels, for example, show varying formant frequencies depending on the length of the vocal tract (e.g. Peterson & Barney, 1952; Stevens, 1998) and it is argued that listeners need to normalize such speaker variability to arrive at the abstract vowel category (Adank et al., 2004). On the other end of the scale are exemplar models (Johnson, 1997) arguing that individual exemplars of speech are stored in human memory and aid linguistic processing such as recognizing phonological categories, syllables, words, etc. Thus, familiarity with a speaker's voice has a positive impact on linguistic processing which is typically measured in terms of speech processing abilities such as intelligibility. The hypothesis is that increased familiarity with a speaker leads to increased speech intelligibility (Creel & Tumlin, 2011; Nygaard et al., 1994; Theodore & Miller, 2010; Theodore et al., 2015). By now we know that there is a complex relationship between indexical and linguistic information. This relationship is also marked by studies showing that competence in a language enhances listeners' voice recognition ability (Bregman & Creel, 2014; Perrachione et al., 2015; for newborns: Fleming et al., 2014; Johnson et al., 2011; Perrachione et al., 2011). Thus, many models of speech recognition have been developed between the two poles of abstractionism and exemplarism and try to combine the advantages of each of the models (typically referred to as hybrid models, e.g. Kleinschmidt & Jaeger, 2015; see discussion in Smith, 2015).

Identity marked and clear speech allow different predictions regarding abstractionist and exemplar models of speech perception (see Fig. 2). In line with abstractionist models,

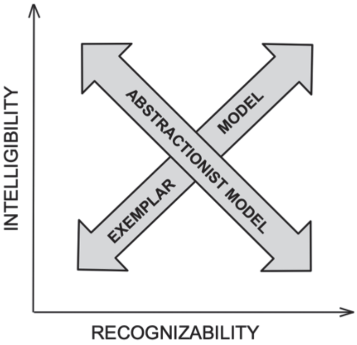

Figure 2. Predicted relationships about the intelligibility of speech and the recognizability of individuals in abstractionist and exemplar models.

it should be predicted that speakers become less recognizable with increasing intelligibility (i.e. increasing clarity) because, according to these models, speakers seem to suppress individual variability that obstructs intelligibility (Fig. 2, green arrow: negative correlation between intelligibility and recognizability). This would be in line with the findings under section 5 revealing that speakers might support two different acoustic modes for clear and identity marked speech. If the two were exclusive, it should be predicted that an increase in intelligibility automatically leads to a decrease in recognizability as speaker-specific information should be suppressed to warrant intelligibility. Given that individual variability is viewed as necessary to retrieve linguistic information from individual exemplars in exemplar models, it seems plausible that this relationship is reversed and that speakers become more recognizable with increasing intelligibility (Fig. 2, red arrow: positive correlation between intelligibility and recognizability). It will be interesting to test such predictions in the context of voice recognition experiments in future research.

ACKNOWLEDGEMENTS

Thanks to Sarah Lim for the graphical design of Fig. 2. The framework in this article will be funded by the Swiss National Science Foundation (grant number 100012_185399).

REFERENCES

Abercrombie, D. (1967). *Elements of General Phonetics*. Chicago: University of Chicago Press.

Adank, P., Smits, R. & Van Hout, R. (2004). A comparison of vowel normalization procedures for language variation research. *Journal of the Acoustical Society of America*, 116(5), 3099–3107.

Amino, K. & Arai, T. (2009). Speaker-dependent characteristics of the nasals. *Forensic Science International*, 185, 21–28.

Amino, K., Sugawara, T. & Arai, T. (2006). Idiosyncrasy of nasal sounds in human speaker identification and their acoustic properties. *Acoustical Science and Technology*, 27, 233–235.

Belin, P. (2006). Voice processing in human and non-human primates. *Philos Trans R Soc Lond B Biol Sci.*, 361(1476), 2091–2107.

Belin, P., Boehme, B. & McAleer, P. (2017). The sound of trustworthiness: Acoustic-based modulation of perceived voice personality. *PLoS One*, 12(10), e0185651.

Belin, P. & Zatorre, R. J. (2003). Adaptation to speaker's voice in right anterior temporal lobe. *Neuroreport*, 14(16), 2105–2109.

Bregman, M. R. & Creel, S. C. (2014). Gradient language dominance affects talker learning. *Cognition*, 130(1), 85–95.

Bruckert, L., Bestelmeyer, P., Latinus, M., Rouger, J., Charest, I., Rousselet, G. A., Kawahara, H. et al. (2010). Vocal attractiveness increases by averaging. *Current Biology*, 20(2), 116–120.

Burnham, D., Kitamura, C. & Vollmer-Conna, U. (2002). What's new, pussycat? On talking to babies and animals. *Science*, 296(5572), 1435–1435.

Burton, A. M., Kramer, R. S., Ritchie, K. L. & Jenkins, R. (2016). Identity from variation: Representations of faces derived from multiple instances. *Cognitive Science*, 40(1), 202–223.

Campeanu, S., Craik, F. I. M. & Alain, C (2013). Voice congruency facilitates word recognition. *PLoS One*, 8(3): e58778.

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.

Clifford, B. R. (1980). Voice identification by human listeners: On earwitness reliability. *Law and Human Behavior*, 4(4), 373–394.

Collins, S. A. (2001). Men's voices and women's choices. *Animal Behaviour*, 60(6), 773–780.

Collins, S. A. & Missing, C. (2003). Vocal and visual attractiveness are related in women. *Animal Behaviour*, 65(5), 997–1004.

Creel, S. C. & Bregman, M. R. (2011). How talker identity relates to language processing. *Linguistics and Language Compass*, 5(5), 190–204.

Creel, S. C. & Tumlin, M. A. (2011). On-line acoustic and semantic interpretation of talker information. *Journal of Memory and Language*, 65(3), 264–285.

De Figueiredo, R. M. & de Souza Britto, H. (1996). A report on the acoustic effects of one type of disguise. *Forensic Linguistics*, 3, 168–175.

de Saussure, F. (1916). *Cours de linguistique generale*. Laussane and Paris: Payot.

de Jong, G., McDougall, K., Hudson, T. & Nolan, F. (2007). The speaker-discriminating power of sounds undergoing historical change: A formant-based study. In: Proceedings of the 16th International Congress of Phonetic Sciences, 1813–1816.

Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P. & Ouellet, P. (2011). Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4), 788–798.

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

Dellwo, V., French, P. & He, L. (2018a). Voice biometrics for forensic speaker recognition applications. In: Frühholz, S. & Belin, P. (Eds.), *The Oxford Handbook of Voice Perception* (pp. 777–798). Oxford: Oxford University Press.

Dellwo, V., Kathiresan, T., Pellegrino, E., Schwab, S. & Maurer, D. (2018b). Influences of fundamental oscillation on speaker identification in vocalic utterances by humans and computers. In: *Proceedings of Interspeech 2018*, 3795–3799.

Doscher, B. (1993). *The Functional Unity of the Singing Voice*. Scarecrow Press.

Fant, G. (1975). Non-uniform vowel normalization. *STL-QPSR*, 2–3/1975, 1–19.

Fischer, J., Semple, S., Fickenscher, G., Jürgens, R., Kruse, E., Heistermann, M. et al. (2011). Do women's voices provide cues of the likelihood of ovulation? The importance of sampling regime. *PLoS One*, 6(9), e24490.

Fleming, D., Giordano, B. L., Caldara, R. & Belin, P. (2014). A language-familiarity effect for speaker discrimination without comprehension. *Proceedings of the National Academy of Sciences of the United States of America*, 111(38), 13795–13798.

Eriksson, A. (2010). The disguised voice: imitating accents or speech styles and impersonating individuals. In: Llamas, C. & Watt, D. (Eds.), *Language and Identities*. Edinburgh: Edinburgh University Press.

Eriksson, A. & Wretling, P. (1997). How flexible is the human voice? – A case study of mimicry. In: *Proceedings of Eurospeech 1997*, 1043–1046.

Foulkes, P. & Barron, A. (2000). Telephone speaker recognition amongst members of a close social network. *Forensic Linguistics*, 7, 180–198.

Garcia-Romero, D. & Espy-Wilson, C. Y. (2011). Analysis of i-vector length normalization in speaker recognition systems. In: *Proceedings of Interspeech 2011*.

Goldinger, S. D. (1996). Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22(5), 1166–1183.

Hansen, J. H. L. & Hasan, T. (2015). Speaker recognition by machines and humans: A tutorial review. *IEEE Signal Processing Magazine*, 32(6), 74–99.

Hazan, V. & Baker, R. (2011). Acoustic-phonetic characteristics of speech produced with communicative intent to counter adverse listening conditions. *Journal of the Acoustical Society of America*, 130(4), 2139–2152.

Hazan, V., Grynpas, J. & Baker, R. (2012). Is clear speech tailored to counter the effect of specific adverse listening conditions? *Journal of the Acoustical Society of America*, 132(5), EL371–EL377.

He, L. & Dellwo, V. (2016). The role of syllable intensity in between-speaker rhythmic variability. *International Journal of Speech, Language and the Law*, 23(2), 243–273.

He, L. & Dellwo, V. (2017). Between-speaker variability in temporal organizations of intensity contours. *Journal of the Acoustical Society of America*, 141(5): EL488–EL494.

He, L., Zhang, Y. & Dellwo, V. (2019). Between-speaker variability and temporal organization of the first formant. *Journal of the Acoustical Society of America*, 145(3): EL209–EL214.

Hirson, A. & Duckworth, M. (1993). Glottal fry and voice disguise: a case study in forensic phonetics. *Journal of Biomedical Engineering*, 15(3), 193–200.

Hove, I. & Dellwo, V. (2014). The effects of voice disguise on f0 and on the formants. In: *Proceedings of IAFPA 2014*.

Hudson, T., de Jong, G., McDougall, K., Harrison, P. & Nolan, F. (2007). F0 statistics for 100 young male speakers of Standard Southern British English. In: *Proceedings of the 16th International Congress of Phonetic Sciences*, 1809–1812.

Jansen, W., Gregory, M. L. & Brenier, J. M. (2001). Prosodic correlates of directly reported speech: Evidence from conversational speech. In: *ISCA tutorial and research workshop (ITRW) on prosody in speech recognition and understanding*.

Johnson, K. (1997). Speech perception without speaker normalization: An exemplar model. In: Johnson, K. & Mullennix, J. W. (Eds.), *Talker Variability in Speech Processing* (pp. 145–165). San Diego: Academic Press.

Johnson, E. K., Westrek, E., Nazzi, T. & Cutler, A. (2011). Infant ability to tell voices apart rests on language experience. *Developmental Science*, 14(5), 1002–1011.

Johnsrude, I. S., Mackey, A., Hakyemez, H., Alexander, E., Trang, H. P. & Carlyon, R. P. (2013). Swinging at a cocktail party: Voice familiarity aids speech perception in the presence of a competing voice. *Psychological Science*, 24(10), 1995–2004.

Kathiresan, T., Dilley, L., Townsend, S., Shi, R., Daum, M., Arjmandi, M. & Dellwo, V. (2019). Infant-directed speech enhances recognizability of individual mothers' voices. *Journal of the Acoustical Society of America*, 145(3), 1766.

Kemper, S., Finter-Urczyk, A., Ferrell, P., Harden, T. & Billington, C. (1998). Using elderspeak with older adults. *Discourse Processes*, 25(1), 55–73.

Kerstholt, J. H., Jansen, N. J., Van Amelsvoort, A. J. & Broeders A. P. A. (2004). Earwitnesses: Effects of speech duration, retention interval and acoustic environment. *Applied Cognitive Psychology*, 18(3), 327–336.

Kisilevsky, B. S., Hains, S. M. J., Lee, K., Xie, X., Huang, H., Ye, H., Zhang, K., & Wang, Z. (2003). Effects of experience on fetal voice recognition. *Psychological Science*, 14(3), 220–224.

Kisilevsky, B., Hains, S., Brown, C., Lee, C., Cowperthwaite, B. & Stutzman, S. (2009). Fetal sensitivity to properties of maternal speech and language. *Infant Behavior and Development*, 32, 59–71.

Kitamura, T. (2008). Acoustic analysis of imitated voice produced by a professional impersonator. In: *Proceedings of Interspeech 2008*, 813–816.

Kleinschmidt, D. F. & Florian Jaeger, T. (2015). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review*, 122(2), 148–203.

Klewitz, G. & Couper-Kuhlen, E. (1999). Quote-unquote? The role of prosody in the contextualization of reported speech sequences. *Pragmatics*, 9(4), 459–485.

Knoll, M. A., Johnstone, M. & Blakely, C. (2015). Can you hear me? Acoustic modifications in speech directed to foreigners and hearing-impaired people. In: *Proceedings of Interspeech 2015*, 2987–2990.

Craik, F. I. M. & Kirsner, K. (1974). The effect of speaker's voice on word recognition. *Quarterly Journal of Experimental Psychology*, 26(2), 274–284.

Kriengwatana, B., Escudero, P. & ten Cate, C. (2015). Revisiting vocal perception in non-human animals: A review of vowel discrimination, speaker voice recognition, and speaker normalization. *Frontiers in Psychology*, 5, 1543.

Kreiman, J. & Sidtis, D. (2011). *Foundations of Voice Studies: An Interdisciplinary Approach to Voice Production and Perception*. Hoboken: John Wiley & Sons.

Kurowski, K. M., Blumstein, S. E. & Alexander, M. (1996). The Foreign Accent Syndrome: A reconsideration. *Brain and Language*, 54(1), 1–25.

Ladefoged, P. & Ladefoged, J. (1980). The ability of listeners to identify voices. *UCLA Working Papers in Phonetics*, 49, 43–89.

Larranaga, A., Bielza, C., Pongrácz, P., Faragó, T., Bálint, A. & Larranaga, P. (2015). Comparing supervised learning methods for classifying sex, age, context and individual Mudi dogs from barking. *Animal Cognition*, 18(2), 405–421.

Latinus, M. & Belin, P. (2011). Anti-voice adaptation suggests prototype-based coding of voice identity. *Frontiers in Psychology*, 2, 1–12.

Lavan, N., Burton, M., Scott, S. K. & McGettigan, C. (2019). Flexible voices: Identity perception from variable vocal signals. *Psychonomic Bulletin & Review*, 26(1), 90–102.

Levi, S. V. & Pisoni, D. B. (2007). Indexical and linguistic channels in speech perception: Some effects of voiceovers on advertising outcomes. In: T. M. Lowrey (Ed.), *Psycholinguistics Phenomena in Marketing Communications* (pp. 203–219). Mahwah: Lawrence Erlbaum.

Liberman, A. M., Cooper, F. S., Shankweiler, D. & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, 74, 431–461.

Liberman, A. M. & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition*, 21, 1–36.

Locke, J. L. (2006). Parental selection of vocal behavior: Crying, cooing, babbling, and the evolution of language. *Human Nature*, 17(2), 155–168.

McAleer, P., Todorov, A. & Belin, P. (2014). How do you say 'Hello'? Personality impressions from brief novel voices. *PLoS One*, 9(3): e90779.

McClelland, J. L. & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, 18(1), 1–86.

McDougall, K. & Nolan, F. (2007). Discrimination of speakers using the formant dynamics of /u:/ in British English. In: Proceedings of the 16th International Congress of Phonetic Sciences, 1825–1828.

Moez, A., Bonastre, J. F., Kheder, W. B., Rossato, S. & Kahn, J. (2016). Phonetic content impact on forensic voice comparison. In: *IEEE Spoken Language Technology Workshop (SLT)*.

Molnár, C., Pongrácz, P., Faragó, T., Dóka, A. & Miklósi, Á. (2009). Dogs discriminate between barks: the effect of context and identity of the caller. *Behavioural Processes*, 82(2), 198–201.

Nolan, F. (1997). Speaker recognition and forensic phonetics. In: W. Hardcastle & J. Laver (Eds.), *A Handbook of Phonetic Science*. Oxford: Blackwell.

Norris, D. (1994). Shortlist: A connectionist model of continuous speech recognition. *Cognition*, 52(3), 189–234.

Nygaard, L. C., Sommers, M. S. & Pisoni, D. B. (1994). Speech perception as a talker-contingent process. *Psychological Science*, 5(1), 42–46.

O'Connor, J. J. & Barclay, P. (2017). The influence of voice pitch on perceptions of trustworthiness across social contexts. *Evolution and Human Behavior*, 38(4), 506–512.

Oleszkiewicz, A., Pisanski, K., Lachowicz-Tabaczek, K. & Sorokowska, A. (2017). Voice-based assessments of trustworthiness, competence, and warmth in blind and sighted adults. *Psychonomic Bulletin & Review*, 24(3), 856–862.

Panneton Cooper, R., Abraham, J., Berman, S. & Staska, M. (1997). The development of infants' preference for motherese. *Infant Behavior and Development*, 20(4), 477–488.

Papcun, G., Kreiman, J. & Davis, A. (1989). Long-term memory for unfamiliar voices. *Journal of the Acoustical Society of America*, 85(2), 913–925.

Peirce, C. S., Hartshorne, C., Weiss, P. & Burks, A. W. (1965). *Collected papers of Charles Sanders Peirce*. Cambridge, Mass: Belknap.

Perrachione, T. K., Del Tufo, S. N. & Gabrieli, J. D. E. (2011). Human voice recognition depends on language ability. *Science*, 333(July), 595.

Perrachione, T. K., Dougherty, S. C., McLaughlin, D. E. & Lember, R. A. (2015). The effects of speech perception and speech comprehension on talker identification. In: *Proceedings of ICPhS 2015*.

Perrodin, C., Kayser, C., Logothetis, N. K. & Petkov, C. I. (2011). Voice cells in the primate temporal lobe. *Current Biology*, 21(16), 1408–1415.

Perrodin, C., Kayser, C., Abel, T. J., Logothetis, N. K. & Petkov, C. I. (2015). Who is that? Brain networks and mechanisms for identifying individuals. *Trends in Cognitive Sciences*, 19(12), 783–796.

Peterson, G. E. & Barney, H. L. (1952). Control methods used in a study of the vowels. *Journal of the Acoustical Society of America*, 24(2), 175–184.

Petkov, C. I., Logothetis, N. K. & Obleser, J. (2009). Where are the human speech and voice regions, and do other animals have anything like them? *The Neuroscientist*, 15(5), 419–429.

Pollard, K. A. & Blumstein, D. T. (2011). Social group size predicts the evolution of individuality. *Current Biology*, 21(5), 413–417.

Raj, A., Gupta, B., Chowdhury, A. & Chadha, S. (2010). A study of voice changes in various phases of menstrual cycle and in postmenopausal women. *Journal of Voice*, 24(3), 363–368.

Rosenberg, A. & Hirschberg, J. (2005). *Acoustic/Prosodic and lexical correlates of charismatic speech*. Columbia University: Academic Commons.

Roswandowitz, C., Mathias, S. R., Hintz, F., Kreitewolf, J., Schelinski, S. & von Kriegstein, K. (2014). Two cases of selective developmental voice-recognition impairments. *Current Biology*, 24(19), 2348–2353.

Růžičková, A. & Skarnitzl, R. (2017). Voice disguise strategies in Czech male speakers. *Acta Universitatis Carolinae – Philologica 3*, 19–34.

Schegloff, E. A. (1979). Identification and recognition in telephone conversation openings. In: Psathas, G. (Ed.), *Everyday Language: Studies in Ethnomethodology* (pp. 23–78). New York: Irvington Publishers.

Shaw, G. B. (1916). *Pygmalion*. New York: Brentano.

Skuk, V. G. & Schweinberger, S. R. (2013). Gender differences in familiar voice identification. *Hearing Research*, 296, 131–140.

Smiljanić, R. & Bradlow, A. R. (2008). Temporal organization of English clear and conversational speech. *Journal of the Acoustical Society of America*, 124(5), 3171–3182.

Smith, R. (2015). Perception of speaker-specific phonetic detail. In: Fuchs, S., Pape, D., Petrone, C. & Perrier, P (Eds.), *Individual Differences in Speech Production and Perception* (pp. 11–38). Frankfurt a. M.: Peter Lang.

Stevenage, S. V., Clarke, G. & McNeill, A. (2012). The "other-accent" effect in voice recognition. *Journal of Cognitive Psychology*, 24(6), 647–653.

Stevenage, S. V. (2017). Drawing a distinction between familiar and unfamiliar voice processing: A review of neuropsychological, clinical and empirical findings. *Neuropsychologia*, 31(116), 162–178.

Stevens, K. (1998). *Acoustic Phonetics*. Cambridge, MA: MIT Press.

Sullivan, R., Perry, R., Sloan, A., Kleinhaus, K. & Burtchen, N. (2011). Infant bonding and attachment to the caregiver: Insights from basic and clinical science. *Clinics in Perinatology*, 38, 643–655.

Sundberg, J. (1977). The acoustics of the singing voice. *Scientific American*, 236(3), 82–91.

Theodore, R. M. & Miller, J. L. (2010). Characteristics of listener sensitivity to talker-specific phonetic detail. *Journal of the Acoustical Society of America*, 128(4), 2090–2099.

Theodore, R. M., Blumstein, S. E. & Luthra, S. (2015). Attention modulates specificity effects in spoken word recognition: Challenges to the time-course hypothesis. *Attention, Perception, and Psychophysics*, 77(5), 1674–1684.

Van Lancker, D. R., Cummings, J. L., Kreiman, J. & Dobkin, B. H. (1988). Phonagnosia: A dissociation between familiar and unfamiliar voices. *Cortex*, 24(2), 195–209.

Von Kriegstein, K & Giraud, A. L. (2004). Distinct functional substrates along the right superior temporal sulcus for the processing of voices. *Neuroimage*, 22(2), 948–955.

Von Kriegstein, K., Kleinschmidt, A. & Giraud, A. L. (2005). Voice recognition and cross-modal responses to familiar speakers' voices in prosopagnosia. *Cerebral Cortex*, 16(9), 1314–1322.

Wagner, I. & Köster, O. (1999). Perceptual recognition of familiar voices using falsetto as a type of voice disguise. In: Proceedings of the 14th International Congress of Phonetic Sciences, 1381–1385.

Yarmey, A. D. (1995). Earwitness speaker identification. *Psychology, Public Policy, and Law*, 1(4), 792–816.

RESUMÉ

Každý člověk má jiný hlas a lidé mají propracované schopnosti, jak mluvčí rozpoznávat po hlasu. Jedná se o jev, který je hluboce zakořeněn ve vývoji lidského chování. Mechanismy rozpoznávání mluvčích dodnes jsou dobře pochopeny, a to především kvůli vysoké míře variability akustických vodítek k individualitě mluvčího. Příspěvek se zaměřuje na otázku, jakou roli hraje mluvčí, když svůj hlas dělá více či méně rozpoznatelný. Zatímco je z literatury evidentní, že mluvčí jsou schopni ovládat vlastnosti svého hlasu za účelem snadnější srozumitelnosti, není zřejmé, jestli jsou mluvčí schopni tyto vlastnosti ovládat za účelem snadnější rozpoznatelnosti a jestli to opravdu dělají. Otázkou také je, jestli takové ovládací mechanismy hrají nějakou roli v komunikačním procesu. Článek shrnuje výsledky dosavadních studií, které podporují názor, že idiosynkratické mluvčí jsou dynamické povahy a že lidé dovedou ovládat, nakolik bude jejich hlas rozpoznatelný. Autoři naznačují možné podoby experimentálního výzkumu, které by umožnily ovládání hlasové identity ověřit, a představují pilotní studii akustických vlastností řeči, která byla produkována s cílem být (a) srozumitelná (zřetelná řeč) nebo (b) vhodná pro rozpoznání mluvčího (řeč obsahující vodítka k identitě). Výsledky podporují názor, že když mluvčí chtějí, aby byla jejich identita dobře rozpoznatelná, využívají odlišných mechanismů ve srovnání se situací, kdy chtějí, aby jejich řeči bylo dobře rozumět. Autoři diskutují předpovědi, které z ovládání rozpoznatelnosti a srozumitelnosti vyplývají v rámci nejvýznamnějších teorií percepce řeči.

Volker Dellwo, Elisa Pellegrino, Lei He, Thayabaran Kathiresan
Phonetics & Speech Sciences Group
Institute of Computational Linguistics
University of Zurich, Switzerland
E-mail: volker.dellwo@uzh.ch