Universität Zürich[UZH]

Master Thesis

Autumn Semester 2020

Uncovering the role of prosody in voice discrimination

Author : Debora Beuret
Matricule number: 16-717-951

Supervisors: Volker Dellwo & Elisa Pellegrino
Department of Computational Linguistics

Date: November 30, 2020

Contents

Abstract

Voice is highly idiosyncratic and humans are good at identifying voices, but it isn't clear which features of voice they rely on to perform voice perception tasks. This study is aimed at disentangling the effects of timbre and prosody in voice perception and understanding their contribution towards voice individuality. To this end, we used a voice conversion algorithm that, given a source and a target voice, outputs a voice that imitates the timbre of the target voice but retains the prosody of the source voice. As a result, the participants heard voice pairs that either shared their prosody, their timbre, or no common feature. This is a relatively new methodology in this field and it allows to focus on prosody, the role of which is not yet well understood. The results were analyzed with a two-way repeated ANOVA focusing on the common feature and the duration of the stimuli. The most important findings are that a longer exposure to the voices increases the likelihood to assimilate two voices sharing a common feature and that listeners rely equally on both timbre and prosody to discriminate between voices. The latter result is surprising: although prosodic cues are known to show speaker individuality, no study yet had proved they are important in voice perception.

1 Introduction

1.1 What are timbre and prosody, and why are they relevant?

This study aims at disentangling the respective effects of timbre and prosody in voice discrimination tasks. The question is highly relevant because although the role of spectral features in voice perception has been known for a long time [Matsumoto et al., 1973], [Walden et al., 1978], [Singh and Murry, 1978], [Van Lancker et al., 1985], [Van Dommelen, 1990], [Nolan et al., 2011], [Perrachione et al., 2019], the same isn't true for prosody.

But what are timbre and prosody?

Timbre commonly refers to the spectral features of voice, usually the factors in the voice that are influenced by the speaker's physiology. This encompasses the F0 and the mel-cepstral coefficient [Kobayashi and Toda, 2018]. Prosody, in the other hands, refers to the temporal features of speech. It comprises things such as speech rhythm, intonation, speech rate and stress. The prosodic cues are behavioral features, as they are influenced by the speakers themselves [Dellwo et al., 2018]. A more detailed explanation of the distinction between timbre and prosody follows in section 2.1.

There is a broad body of research over the past decades that has focused on understanding the role of timbre in voice perception, with a lot of interesting results. The role of F0, for example, has been described in [Matsumoto et al., 1973], [Walden et al., 1978], [Singh and Murry, 1978], [Van Lancker et al., 1985], [Van Dommelen, 1990], [Nolan et al., 2011] and [Perrachione et al., 2019]. There is little research about the perceptual role of prosody, which is precisely the problem we tackle in this study.

One of the reasons that the role of prosody is under-researched is because it is more complicated to study its effects. Timbre can be taken out of a sound file frame by frame, but the same isn't possible for a feature that reveals over time [Kobayashi and Toda, 2018].

In recent years, the rise of efficient voice conversion (VC) systems has expanded the opportunities. VC

algorithms are used to make a person's voice sound different, in most cases like that of another person [Kinnunen et al., 2012]. A lot of voice conversion algorithms manipulate only the spectral features of voice while leaving the temporal features intact [Kobayashi and Toda, 2018]. In other words, the algorithm receives two voices, the source and the target, and modifies only the timbre of the source voice to match that of the target voice. The final result is a voice that kept the prosody from the source, but shares its timbre with the target.
Using VC technology opened up the possibility to create a discrimination task where the listeners would hear a pair of voices that would share either their timbre, their prosody, or no common feature, which would help study the effect of each feature separately. Voice conversion is explained in section 2.6 and the precise algorithm we used is described in section 3.2.3.
We expected the result to reveal relevant information about the strategies used by listeners when they have to discriminate between two voices : were they more likely to judge based on the spectral features or the prosodic cues?
Using converted voices for a perceptual experiment is a relatively new process and something that has not been experimented with a lot. It is possible that the future will bring more such experiments, as manipulating voices is becoming more accessible and shows better results than ever before.

1.1.1 Outline of the study

We worked with a set of five male speakers, on which we performed voice conversion. One speaker served as a source voice while the four others provided target voices. With this strategy, we were able to create hybrid pairs of voice that would match in either their prosody or their timbre. In the condition where both voices shared their prosody, we created voice pairs that shared the same source speaker but had different target speakers. As the source voice provides the prosody, it would match. In the condition where both voices shared their timbre, we used synthetic replicas of the target voice. The pair consisted in two voices with the same target but different prosodies : once that of the source speaker and once that of the natural target speaker's voice. Stimuli in which the two voices were completely different were also tested. A more detailed overview of our study can be found in section 3.
We performed a two-way repeated ANOVA on our results to investigate the effect of a common feature as well as the effect of duration. Interaction effects were also examined. The results of our analysis will be outlined in section 4.

1.1.2 Structure of this work

This paper will be structured as follows: the next section will focus on laying the bases of a common understanding about all the important concepts mentioned in this work. We will go over the definition of voice and clarify timbre and prosody further. The focus will then shift to the role of voice as a feature of individuality and the characteristics that make voices unique as well as other factors influencing voice

perception. Finally, we will define the task of voice discrimination and explain voice conversion.
The subsequent sections will focus on the specificities of the study : the research question and our hypotheses, our methodology including a description of the material, our listeners, and the stimuli. Finally, we will discuss the results and share some concluding remarks.

2 What is voice?

According to the Merriam-Webster Dictionary [Merriam-Webster.com, 2020], voice is the "sound produced by vertebrates by means of lungs, larynx, or syrinx". Another source, Lexico.com [Oxford-University-Press, 2020] defines voice as being the " sound produced in a person's larynx and uttered through the mouth, as speech or song."
More precise definitions take the vocal tract resonances into account [Kreiman et al., 2003]. According to [McKinney, 2005], resonation is "the process by which the basic product of phonation is enhanced in timbre and/or intensity by the air-filled cavities through which it passes on its way to the outside air" [McKinney, 2005, p. 120]. Various terms are used in relation to resonation : filtering, amplification, enrichment, intensification, etc [McKinney, 2005]. The resonation is influenced by all of the involved physiological parts of the body : the chest, the tracheal tree, the larynx itself, the pharynx, the oral cavity, the nasal cavity, and the sinuses [Greene and Mathieson, 2001]. Thus, the sound coming out of one's voice is the result of the interaction between all these factors.
Voice, as its main function, is the carrier of speech. For this purpose, humans have the ability to use and modify their voices to carry meaning. When the word *voice* is used as a synonym for *speech*, it encompasses a number of features: : pitch, amplitude variation, temporal patterning, phonatory quality, etc [Kreiman et al., 2003].

2.1 Timbre and prosody

As explained above, this research focuses on two common features of voice : timbre and prosody. These could also be described as being the segmental and the supra-segmental (or temporal) aspects of speech. In [Dellwo et al., 2018], the distinction is explained as the difference between physiological features and behavioural features.

> [The] physiological biometric features are all characteristics that enter the acoustic speech signal as a result of individual dimensions of the articulators (in particular, the vocal tract) and the movement behaviour of the articulators that results out of their particular design (size, weight, etc.) [Dellwo et al., 2018, p. 3].

2.1.1 Timbre

Timbre is a very complex aspect of the voice and it is often described by what it is not, rather than by what it is.

> Informally, the standard definition of timbre is regarded with considerable amusement. You might expect the definition of timbre to tell you something about what timbre is, but all the

> definition tells you is that there are a few things that timbre is not. It is not pitch, it is not loudness, and it is not duration. It is everything else. [Patterson et al., 2010, p. 223]

Another description of timbre that is often made is that it is a voice's tone color [Howell, 2016], tone color, spectral color or spectral timbre [Łetowski, 2014].
Although timbre seems to be described in similar ways by researchers, there seems to be no definitive compromise on its definition, which is why it is important to define what timbre will stand for in this research.
Timbre encompasses several dimensions of voice but relates to short-term spectral features, as in [Wu and Li, 2014]. The most important features in this study will be the F0 and the mel-cepstral coefficient (MCC), as they are the features modified by our voice conversion algorithm [Kobayashi and Toda, 2018].
Prosody is related to the speaker's physiology and influenced by the shape of the different organs in its vocal tract. However, it is noted that unlike DNA or fingerprint, voice goes through both short-term (e.g. when having a sore throat) and long-term (ageing) changes [Dellwo et al., 2018]. These changes are related to the fact that voice is a performance metric, as already mentioned in [Hansen and Hasan, 2015] and generally in the section 2.2.

2.1.2 Prosody

Prosody, on the other hand, refers to the temporal characteristics of a voice when speaking. Things like rhythm, intonation, speech rate, unique pronunciation of different phonemes or stress are all part of a person prosody. We could describe it as "the way one speaks". Behavioural features that are learned, not innate, are usually understood as being part of prosody [Dellwo et al., 2018].

> [These behavioural features] are, to a large extent, the product of a speaker's linguistic socialization and psychosocial orientations (i.e. speakers learn and adopt the social, regional, and ethnic norms of the speech communities in which they live, identify with, and aspire to). [Dellwo et al., 2018, p. 4]

Prosody has a clearer definition than timbre and encompasses a number of different features : intonation, rhythm pattern, rate of speaking, stress, pronunciation, etc. These factors can also carry individuality and depend a lot on variables such as the social environment, the linguistic environment while growing up, the personality and others. Importantly, prosodic cues can also be influenced by physiology : people in fear or panic have a higher speech rate [Schweinberger et al., 2018]. Age and gender influence the behavioural features as well [Dellwo et al., 2018], and older people tend to speak more slowly, for example [van Brenk et al., 2009]. With this in mind, it is important to mention that a mere separation of prosody and timbre is an over-simplification, as learning and physiological characteristics both interact in intricate ways [Nolan, 1997].

Generally speaking, humans have a great deal of control over their prosodic cues. Speakers can control the speed at which they speak, the language they speak, the way they pronounce certain words. Voice prosody can be changed by the speaker much easier than its F0. This is also visible when humans try to manipulate their voice to sound like someone else, something that can be done maliciously in case of spoofing attacks [Wu and Li, 2014].

> Partial evidence [...] suggests that humans are most effective in mimicking speakers with "similar" voice characteristics to their own, while impersonating an arbitrary speaker appears challenging [Lau et al., 2004]. Professional voice mimics, often voice actors, tend to mimic prosody, accent, pronunciation, lexicon, and other high-level speaker traits, rather than spectral cues [...]. [Wu and Li, 2014, p. 1]

This present study will aim at disentangling the effect of timbre and prosody in voice discrimination. Both these features have shown to be important in speaker individuality and will be examined in more details in the sections below.

2.1.3 The source-filter model

One of the most common voice model is called the source-filter model [Fant, 1981], and it corresponds to the distinction we drew between timbre and prosody. The model assumes that voice production depends on two factors. The source here refers to the source of the sound, the creation of the raw material. This happens when the vocal folds vibrate, which defines the pitch frequency of the voice [Fant, 1981], [Kuwabara, 1995]. The source can be assimilated to the timbre. The filter, on the other hand, describes how the raw sound is being shaped by the vocal tract, including the the mouth and nasal cavities. The vocal tract controls the formant frequencies, which create resonances for both voiced and unvoiced sounds [Ramakrishnan, 2013] [Kuwabara, 1995]. The control that humans exert on their vocal tract to shape language is part of the filter, thus there is a clear parallel with prosody.
These two dimensions of speech production are equally important, however, humans have the ability to exert control only on part of the filter, positioning their jaws, tongue and lips in configurations allowing to utter the sounds they wish to. The vocal folds and the shape of the vocal tract are determined by genetic, age, weight and health factors and cannot be influenced voluntarily. [Kuwabara, 1995] makes a distinction between what they call the hardware and the software of the voice. The hardware is the physiological apparatus, while the software is the control one exerts over the apparatus and can be "programmed". Trying to copy someone's voice (or modify one's own voice) amounts to trying to modify one's own software to look like someone else's.

2.2 Voice as a biometric feature

Voice carries individuality and as such, it is can be considered a biometric feature of humans [Hansen and Hasan, 2015]. However, there are some challenges inherent to finding which features are the most relevant to individuality. This section will give an overview of voice as a biometric feature and the challenges attached to defining its individuality. It also serves as introduction to the research about the speaker idiosyncratic features in the next section (2.3).

Unlike fingerprints or irises, voice is a performance biometric, only accessible when someone speaks. It is represented by the way speech is performed, not what is said [Hansen and Hasan, 2015], which means it is subjected to a great deal of variability. This, in turn makes it difficult to have reliable metrics [Hansen and Hasan, 2015]. Humans exert a great deal of control over their voice. They are able to utter an important variety of different sounds and can control the amplitude (loudness) of their voice, their intonation, their rhythmic pattern and many other factors. They can shout, whisper or sing. They can also control their pitch, albeit they are limited by their physiology. Pitch, which is described perceptually by how high the voice sounds, depends on many of the physiological features previously mentioned and varies with a speaker's age, gender, health state. Men, for example, have lower voices because their vocal folds are longer and thicker. The male pitch is typically around 120Hz, while the female one is around 200Hz [Nishio and Niimi, 2008]. Significant changes can be observed during the aging process. [Nishio and Niimi, 2008] found that these changes are biggest in women. In their sample, young women typically had a pitch around 220Hz, while older women were closer to 180Hz. Not only pitch, but many other parameters contribute to one's voice's variability, as shown below.

One source of variability is stemming from the speaker itself. A speaker can choose to express themselves with different voice qualities. Voice quality refers to the way the vocal folds are vibrating and the laryngeal setting, as different settings are possible [Ladefoged and Maddieson, 1996]. [Ladefoged and Maddieson, 1996] lists these different settings, with some examples listed below.

Modal voice is used to describe the normal speaking voice with a regular vibration of the vocal folds. The voiceless setting refers to the absence of vibration of the vocal folds. It is common for a glottal stop, where the airflow is obstructed, such as in the British pronunciation of the word *better*. Aspirated voice, in turn, is used when there is a greater rate of airflow than normally, typically when someone is out of breath. Breathy voice (or murmur) describes vocal folds that do vibrate but without appreciable contact, with a higher rate of airflow than in modal voice. Other voice quality settings are creaky voice, slack voice and stiff voice [Ladefoged and Maddieson, 1996]. The voice quality settings do not carry meaning in English, but this is not the case in every language : in Danish, for example, there is a prosodic feature called "stød" which refers to creaky voice. Minimal pairs have shown that the stød indeed carries meaning [Fischer-Jørgensen, 1989].

The voice can also be altered by the stress the speaker is feeling, for example if they have to perform another task, like driving, while speaking [Hansen and Hasan, 2015]. The vocal effort performed, for example when someone is screaming or singing, can also affect the voice. Voice can also carry infor-

mation about the speaker's emotional state, for example if they are angry or scared. The same happens in case of altered physiological state, e.g. the speaker is intoxicated or suffers from a flu [Hansen and Hasan, 2015]. The communication situation the speaker is in can also influence his voice : they will not behave in the same way if they are doing a monologue, a two-person conversation, a prompted speech or a spontaneous public speech. The language or dialect spoken can also differ [Hansen and Hasan, 2015]. All these sources of variability can make a speaker very difficult to recognize or discern. However, humans as well as machines can perform speaker recognition and speaker discrimination tasks with good results. This is made possible because each speaker has idiosyncratic characteristics.

2.3 Speaker idiosyncratic characteristics

Speaker idiosyncratic characteristics are important to this study because they are the features that allow listeners to perform discrimination. A lot of research has been conducted overtime to know which features contribute to voice perception and to what extent. This research will be reviewed in this section.
As explained in [Wu and Li, 2014], human voices carries individuality on three levels : segmental, supra-segmental, and linguistic information.

> The segmental information relates to the short-term feature representations, such as spectrum and instantaneous fundamental frequency (F0). The supra- segmental information describes prosodic features such as duration, tone, stress, rhythm over longer stretches of speech than phonetic units. It is more related to the signal but spanning a longer time than the segmental informa- tion. The linguistic information is encoded and expressed through lexical words in a message. Since each speaker has his/her own lexical preference, the choice of words and sen- tence structures, the same linguistic information can be conveyed by different people in different ways. [Wu and Li, 2014, p. 2]

In other words, the segmental features correspond to single points in time. The measurements like instantaneous F0 are made at this specific point in time and these measurements on single points are repeated over the whole duration of the speech sample. On the other hand, supra-segmental features are temporal : they relate to changes in speech that happen overtime. The intonation, which is nothing more than the trajectory of the F0 over time, is one such example. The rate of speech or the rhythmic pattern are other examples [Wu and Li, 2014], [Kobayashi and Toda, 2018], [Hansen and Hasan, 2015].
The linguistic information refers to the words spoken and the lexical meaning they carry. They too can be markers of individuality, as different people will choose different words or sentence structures to express themselves [Hansen and Hasan, 2015].
For the past decades, research has focused on trying to find what acoustic parameters influence voice individuality the most.
[Wolf, 1972], [Kuwabara, 1995], [Fernández Gallardo, 2016] and [Matsumoto et al., 1973] among others mention the fundamental frequency of voice (also referred to as F0) as one major factor that influences

voice individuality.
[Perrachione et al., 2019] mentioned different acoustic features as potential influences for the perception of voice dissimilarity, one of them being the mean fundamental frequency.

> Across both listener groups, both talker languages, and both forward and time-reversed speech, differences in talkers' mean fundamental frequency were most strongly related to listeners' dissimilarity judgments. For a pair of recordings in which talkers exhibited greater differences in mean f0, listeners were more likely to rate that pair of voices as sounding dissimilar, all other factors notwithstanding. Other acoustic factors were also related to listeners' judgments of talker dissimilarity, including HNR and formant dispersion. [Perrachione et al., 2019, p. 3396]

The abbreviation HNR in the previous quote refers to "harmonics-to-noise" ratio and measures the amount of additional noise in a voice signal [Ferrand, 2002]. This measure tends to go up as the speakers age [Ferrand, 2002].
[Walden et al., 1978] also studied the effect of acoustic features on perception of talker similarity. They asked twenty males between 20 and 55 of age to utter several monosyllabic words. The word "bean" was selected out of all the words because it showed the biggest variance across all twenty talkers. Eleven adults were then asked to rate the similarity between all possible voice pairs. They performed a four-dimensional INDSCAL analysis[1] of the similarity ratings and found that two acoustic features, the fundamental frequency and the word duration, correlated moderately with two psychological dimensions. These two dimensions corresponded to acoustic measurements, while the two others represent talker age and voice quality [Walden et al., 1978].
The study by [Singh and Murry, 1978] is particularly interesting because it uncovers fundamental differences between males and females. They too worked with an INDSCAL and found that three dimensions showed the best correlations with their data. The first dimension clearly separates male and female speakers in a dichotomous rather than continuous way, leading the authors to state an inherent difference betweeen males and females [Singh and Murry, 1978]. Their second and third dimensions could only be interpreted for males respectively females. The second dimension correlated highly with pitch judgements, hoarseness judgements and F0 measurement for male speakers. The third dimension, on the other hand, correlated with total duration for females. This result, along with the fact that all judges clustered closely and thus seemed to follow the same perceptual strategies, indicates that listeners could rely on different strategies when listening to different groups [Singh and Murry, 1978]. Another possible interpretation is that male and female speakers could use different voice production strategies. Males, for

[1]The INDSCAL model is particularly suited for dissimilarity judgement tasks. It " assumes that an individual subject's judgment of stimulus similarity is a decreasing linear function of the interstimulus distance (assuming a modified Euclidean metric) in a hypothetical underlying perceptual space [...] The net outcome of an INDSCAL analysis is an overall group perceptual space reflecting the perceptual structure of the stimulus set. as well as a dimensional weight vector for each subject indicating the relative saliency of each dimension for that subject"[Howard and Silverman, 1976, p. 2].

example, used F0 and F2 to create hoarseness in their voice, while the number of pitch shifts predicted breathiness for females but not for males. These strategies could be learned to comply with culturally acceptable or desirable voice features [Singh and Murry, 1978]. In an older study, [Matsumoto et al., 1973] examined the relationships between the acoustical parameters of eight voices and their configuration on a three-dimensional psychological auditory space [Matsumoto et al., 1973]. Their main findings are shown below.

> The relative contribution of the mean fundamental pitch frequency to the perception of the personal quality of voice is the largest among all parameters, and its contribution to the perceptual dimension is almost independent of those of other acoustical parameters. [...] Among the voice samples with same fundamental pitch frequency, the vocal tract characteristics (the deviation of the formant frequencies) and the glottal source characteristics (the slope of the glottal source spectrum and the rapid fluctuation of the fundamental pitch period) contribute to different perceptual dimensions from each other. [Matsumoto et al., 1973, p. 435]

The importance of non-prosodic factors could also be demonstrated in a study of backwards speech. [Van Lancker et al., 1985] created an experiment in which participants had to recognize famous voices. There were two types of speech samples : 2-second forward or 4-second backward speech. The results showed that the performance was relatively low (26.6% of correct identification) in the forward condition when the subjects were not offered response alternatives. However, when they were presented with six possible choices, their results improved drastically, as they identified 69.9% of the voices correctly. This performance dropped to 57.5% when the voices were played backwards, which represents a 12% decrease. This suggests that listeners could be successful using only pitch, pitch range, speech rate and voice quality, but this does not need to be the case for every speaker. The paper shows that not all voices were affected by backward speech in the same way : some voices were almost equally recognizable when played backwards, while others saw a bigger drop in their recognition. This also indicates that voices could be perceived as unique patterns and not necessarily basing on a single acoustic characteristics.

> How this is achieved neuropsychologically remains to be understood. Whatever the exact mechanisms may be, our study supports the view of voice recognition as operating on a loosely structured constellation of cues, any of which or any combination of which can evince recognition - rather than a linear model which would posit one or two or three acoustic parameters adding up to fully specify the identity of all voices. The findings here, then, present a picture of each perceived voice as a unique pattern made up of elements taken from a large pool of possible characteristics. [Van Lancker et al., 1985, p. 33]

The role of F0 was also confirmed in [Nolan et al., 2011], where they asked listeners to rate the similarity between voices and applied a multi-dimensional scaling, ending up with five dimensions. They found

correlations with F0 as well as the first three formants [Nolan et al., 2011].
Next to the spectral cues listed above, prosodic cues have been shown to carry individuality too [Van Dommelen, 1990], [Dellwo et al., 2012], [Leemann et al., 2014], [Dellwo et al., 2015]. However, it is crucial to understand that while most of the research cited above was based on voice perception by human listeners, the literature below is based on purely acoustic parameters [Van Dommelen, 1990], [Dellwo et al., 2012], [Leemann et al., 2014], [Dellwo et al., 2015]. It means that even though it is demonstrated that prosodic cues vary strongly from individual to individual, there is no indication of how this affects listeners' perception.
[Van Dommelen, 1990], for example, focused on the three factors of identification F0, F0 contour and speech rhythm and used synthetic speech, manipulating these variables. They conducted tests with modified F0 contour (original, monotonously declining or averaged)and speech rhythm (original or equalizing the original segments' duration) and manipulated F0 height, thus being able to separate the effects of these factors to a certain extent [Van Dommelen, 1990]. The results showed that F0 was important when the speakers displayed particularly high or low value, while F0 contour was only of secondary importance, although they mention it may be due to a ceiling effect, as the recognition was already relatively high when F0 contour was introduced [Van Dommelen, 1990]. They mention that speech rhythm has a consistent influence, albeit not very big [Van Dommelen, 1990].
More recent studies have focused more exclusively on speech rhythm features. The idea behind an idiosyncratic speech rhythm lays in the fact that the articulators have individual properties, which leads to the movements of the muscles, ligaments, cartilages, etc to be different for each individual. As a consequence, there would be differences in speech as well [Dellwo et al., 2015]. There is a strong parallel between the idea of speaker-specific speech rhythm and the fact that other movements, such as gait or typing, are highly individual [Dellwo et al., 2015], [Leemann et al., 2014]. These studies usually examine a wide variety of temporal measures to investigate their effect on speaker individuality. The most important are listed hereafter. The %V represents the percentage over which speech is vocalic, meaning that all vocalic intervals are summed up and divided by the duration of the complete utterance [Leemann et al., 2014, p. 61]. The %VO is similar to the %V but represents voiced segments instead of vowels. VarcoV and VarcoC stand for the rate-normalized standard deviation of vocalic respectively consonantal interval durations, in other words the variability of vocalic and consonantal durations [Leemann et al., 2014, p. 61]. nPVI_v and nPVI_c represent the rate-normalized average differences between consecutive vocalic respectively consonantal interval durations[Leemann et al., 2014, p. 61]. Instead of taking the general variability like Varco measures, PVI measures examine the duration of successive segments, with a high value indicating a high variability between successive segments [Ferragne and Pellegrino, 2004]. nPVI_vo, the nPVI value for voiced segments, is another important measure [Leemann et al., 2014]. While all the previous measures mentioned are based on vocalic and consonant intervals, there are measures based on syllable peaks and the interval between them. VarcoPeak and nPVI_Peak, which stand for the rate-normalized standard deviation of syllable-peak-to-syllable-peak interval durations and the the rate-normalized average differences between consecutivesyllable-peak-to-syllable-peak interval

durations respectively [Leemann et al., 2014, p. 61], [Dellwo et al., 2012].

[Dellwo et al., 2012] mentions the creation of the TEVOID (Temporal Voice Idiosyncracy) corpus to study the effects of speech rhythm on voice identity. The speakers contributing to the corpus are highly homogenous, as they speak the same dialectal variety of Zurich German and are from the same age group, ensuring that temporal variability stems only from individual factors [Dellwo et al., 2012]. The results demonstrated that %V and %VO had a high between-speaker variability, even though both didn't correlate, i.e. a speaker who has high %V doesn't necessarily have a high %VO [Dellwo et al., 2012]. Furthermore, VarcoPeak was highly significant for between-speaker variability [Dellwo et al., 2012]. Small effects were found for the measures of variability between vocalic intervals such as nPVI_v and to a lesser extent VarcoV. No effects were found for consonantal variability [Dellwo et al., 2012].
[Leemann et al., 2014] analyzed all the measures cited above on speakers with two different conditions : read speech and spontaneous speech. They also analyzed the channel of communication, with both hi-fi recorded sentences and phone-transmitted sentences [Leemann et al., 2014]. They found that the %V and %VO had the strongest speaker effects and were also robust to speaking style and channel variability, which means that speaker-specific rhythm remains even when prosodic variability is introduced [Leemann et al., 2014].
[Dellwo et al., 2015] designed two experiments on High German and Zurich German to rate the effects of speech rhythm on between-speaker individuality. Although the linguistic and prosodic within-speaker variability was strong, they found consistent differences between speakers [Dellwo et al., 2015]. In the first experiment, speakers were asked to read sentences at different rates : normal, slow, very slow, fast and as fast as possible [Dellwo et al., 2015]. The results showed that even though the speech rate was manipulated, there was a significant variability of acoustic measures between speakers. There was also a strong variability between the sentences that was consistent despite the important prosodic changes caused by the different speaking rates, e.g. introducing pauses and silences between clauses [Dellwo et al., 2015]. The second experiment aimed at understanding how free choice of words and grammatical structures could influence an individual's rhythmic signature [Dellwo et al., 2015]. They recorded 16 individuals producing spontaneous speech and then chose one sentence by individual, which every speaker then had to read out loud. The person who'd originally uttered it also read it, so that a comparison between the spontaneously produced sentence and the reading aloud of this same sentence could be compared [Dellwo et al., 2015]. The results of this experiment indicated that sentences vary in the complexity of their vocalic and consonantal intervals, causing changes in the rhythm pattern. However, comparing sentences constructed by the speaker themself with sentences constructed by others could not explain the between-speaker variability [Dellwo et al., 2015]. In other words, speaker-specific choice of words and grammatical structures is not the source of between-speaker variability. The overall results and conclusions are seen below:

In both experiments we found strong effects of speaker and sentence but little to no influence

> of prosodic variability on speaker-specific results. Experiment 2 showed clearly, linguistic structural characteristics of a speaker were not responsible for idiosyncratic rhythm. [...] . It is thus increasingly likely that individual ways of operating the articulators should influence speaker-specific temporal variability. [Dellwo et al., 2015, p. 1525-1526]

To sum up, [Dellwo et al., 2015] found that there was a high variability in consonantal and vocalic interval durations between speakers. Within-speaker changes in prosody (first experiment) or linguistic content (second experiment) had little effect on the between-speaker variability. The authors conclude that the articulators and the way of controlling them are likely responsible for the differences between speakers [Dellwo et al., 2015].

While the literature mentioned above gives important pointers towards factors that affect voice individuality and what makes a voice unique, it is also important to see that so far, there is no consensus on the final features that represent speaker individuality. [Schweinberger et al., 2018] mentions that some speaker characteristics could be defined consistently by acoustic features. For example, F0 and formant frequencies are good indicators of gender, while emotions such as fear and panic are characterized by a high F0 and a fast speech rate and sadness is expressed by low energy [Schweinberger et al., 2018]. However, the same has never been achieved for speaker individuality.

> By contrast, and despite considerable efforts, it seems that no acoustic parameters could be identified that signal speaker identity consistently. In retrospect, this failure may simply indicate that whereas many communicative social signals are characterized by very systematic acoustic correlates, in order to recognize a speaker's identity, the perceptual system uses whatever acoustic cue is particularly salient or characteristic for that particular voice. [Schweinberger et al., 2018, p. 2]

It is important to understand that speaker individuality probably has a number of dimensions and it is likely that not voice individuality can hardly be defined by a single feature [Kuwabara and Sagisak, 1995], as it is likely to be multidimensional, covering aspects such as F0, but also pitch, voice quality, rate of speech and many others.

2.4 Other factors influencing voice perception

Acoustic featurese and factors relying purely on people's voice are not the only ones that have been examined. Some researchers have explored different ideas as well.

2.4.1 Frequency range

Do different frequency ranges influence voice individuality? [Furui, 1986], for example, found that the frequencies between 2.5 and 3.5 kHz contributed the most to voice individuality. This range almost integrally falls into the standard narrow-band (NB) range, which goes from 300 to 3kHz. This narrow-band width corresponds to the initial standard telephone bandwidth. As channel bandwidth used to be much more costly, this standard was chosen because it included most of the energy of speech signals in an efficient way [Fernández Gallardo, 2016]. An extended bandwidth ranging from 50 to 7 kHz (termed wideband or WB), however, improves the quality and intelligibility of the speech signal. In particular, the high frequency contributed to a better differentiation of fricatives, i.e. sounds such as [f], [ʃ], [s] or [ʒ] [Fernández Gallardo, 2016]. This could be an important point for voice discrimination, because studies indicate listeners, next to focusing on acoustic factors such as F0 and tract resonance characteristics, rely on the pronunciation of phonemes. While vowels and nasals provide the best discrimination, fricatives also have a big contribution[Fernández Gallardo, 2016]. [Hansen and Hasan, 2015] also mentions technology as having an important impact on voice-related tasks. The quality of the recording could be a challenge for both humans and machines when having to perform voice discrimination or voice recognition. Not only the quality of the microphone and the file, but also the presence or absence of background noise were mentioned [Hansen and Hasan, 2015].

2.4.2 Language familiarity effect

Another phenomenon that has been studied with interesting results is the ”language familiarity effect” [Wester, 2012], [Bregman and Creel, 2014]. This effect implies that listeners are more capable to discriminate speakers in languages that they are familiar with.
In [Wester, 2012], native English listeners had to discriminate between bilingual speakers in the language pairs English-German, English-Finnish and English-Mandarin. The material was provided by native speakers of either German, Finnish or Mandarin who spoke English as a second language. English was the listeners' first language and they had no experience in either of the other languages. The study was aimed at evaluating how they performed when presented with stimuli in their native language and other languages.
Listeners were subjected to stimuli from one language pair and separated by speaker gender (e.g. they would hear female speakers in English and German). They listened to equal numbers of monolingual (Eng-Eng and Eng-Ger) and cross-lingual speech pairs (e.g. Eng-Ger). The results showed that listeners were performing well above chance level in all three situations. Their accuracy was always above 90% when hearing Eng-Eng pairs and almost as high for monolingual pairs in other languages. However, it decreased when they were exposed to cross-lingual speech pairs, with a remarkably low accuracy for English-Mandarin pairs uttered by females. [Wester, 2012].
In [Bregman and Creel, 2014], the focus was a little different. Their study was aimed at understanding

"the relationship between speech and talker processing". For this, they chose two classes of participants: bilingual Korean-English speakers and monolingual English speakers. The stimuli consisted of 15 sentences in each language, spoken by four female native Korean speakers and four female native American English speakers. Participants learned to associate each talker with cartoon figures. After each stimuli, they were asked to chose the right cartoon figure, with two choices given.
The results showed that speaker learning happened more quickly in one's native language : the Korean speaking listeners were quicker to identify Korean speakers than English speakers, with the opposite being true for English speaking listeners. Additionally, a second effect was observed : the rate at which the Korean-English bilingual speakers learned to identify talkers in their second language correlated with the age at which they'd acquired it [Bregman and Creel, 2014].
A different research was performed by [Perrachione et al., 2019], as listeners performed a voice dissimilarity judgement, deciding how similar two voices sounded. In [Perrachione et al., 2019], natives of either English or Mandarin language were presented with speech samples from both languages, played backwards and forwards. The idea was to replicate an older study ([Fleming et al., 2014]) where the results implicated that listeners of different language backgrounds would focus on different low-level acoustic features when they had to rate the similarity level between two languages [Fleming et al., 2014], [Perrachione et al., 2019]. Backwards speech was used because it retains a lot of acoustic information while disrupting the prosodic cues [Perrachione et al., 2019], [Fleming et al., 2014]:

> Time-reversal was chosen because it is a simple procedure that compromises intelligibility while preserving some of the information present in the natural speech signal. For example, time-reversal disrupts the temporal attributes of speech segments, such as onsets and decays, and reverses pitch curves. Conversely, reversed speech is identical to natural speech in amplitude, duration, and mean fundamental frequency. Furthermore, the formant transition structure of many phonemes (e.g., fricatives and long vowels) is approximately mirrored in the reversed signal, and important indexical cues to speaker identity are also retained. [Fleming et al., 2014, p. 13797]

The results in [Fleming et al., 2014] showed that the language familiarity effect seemed to be retained in backwards speech.

> We found that listeners rated pairs of speakers of their own language as more dissimilar on average than pairs of speakers of a different language, even though all stimuli were rendered unintelligible by time-reversal. This result implies that the LFE is not rooted in language comprehension per se, but rather is based on familiarity with the acoustical fingerprint of one's language, in a manner analogous to the "Other-Race Effect" (ORE) in face recognition. [Fleming et al., 2014, p. 13797]

[Perrachione et al., 2019] mentions that different low-level acoustic features are still present under reversed speech conditions : the syllable organization, with some languages being syllable-timed and other

stress-timed, leads to different patterns in the duration of the syllable and their amplitude. These features would be preserved, for example. In [Perrachione et al., 2019], while the cross-language speech samples had higher dissimilarity ratings, the effect found in [Fleming et al., 2014] was not confirmed.

> Finally, unlike the prior report of [Fleming et al., 2014], we did not observe a consistent difference in listeners' perceptual dissimilarity judgments for native- vs foreign-language talker pairs. [Perrachione et al., 2019, p. 3395]

All these researches show that while language familiarity may have an effect on speaker recognition ([Bregman and Creel, 2014]) and speaker dissimilarity judgement ([Fleming et al., 2014], these results were not confirmed by other research ([Wester, 2012], [Perrachione et al., 2019]). Presenting listeners with cross-lingual speech samples seemed to alter the results significantly ([Wester, 2012], [Perrachione et al., 2019]), but when the speech samples were of the same language, the task is performed with a high accuracy even in foreign languages [Wester, 2012].
In our research, we don't expect the LFE to affect the results, as the German speech samples will be presented to listeners with a knowledge of this language since childhood.

2.4.3 Stimulus duration

Another factor commonly taken into consideration when doing research on both voice discrimination and voice duration is the duration of the stimulus. Studies indicated that a longer duration often lead to a better performance of the listeners, due to the bigger number of phonemes heard [Fernández Gallardo, 2016]. However, [Schweinberger et al., 1997a] showed that the duration effect is limited in time. In a voice recognition setting, the performance did increase with the duration of the stimulus, but only until the duration reached about one second. After that, the listeners seemed to have reached a threshold and their performance stabilized [Schweinberger et al., 1997a]. Note that the paper referred to language recognition and as the two tasks differ and results could vary when listeners are confronted with a discrimination task.
[Schweinberger et al., 2018] also mentions duration as an important factor, although the underlying reason is still not very clear.

> There has been some debate about whether the benefit of longer samples originates from increased phonetic variability or from increased exposure duration per se. This is not an easy question, as both variables are typically confounded in natural speech. Although some researchers have argued for the importance of phonetic variety rather than duration for recognizing once-heard unfamiliar voices [Roebuck and Wilding, 1993], it is likely that both variability and exposure duration contribute to familiar voice recognition [Cook and Wilding, 1997]. [Schweinberger et al., 2018, p. 3]

Duration thus seems to matter. It is however hard to disentangle the effects of the bigger phonetic sample that is made available to the hearer from those of the mere duration.
In our case, duration will be one of the variables that will be manipulated, with longer and shorter stimuli. More on this can be found in Sections 3 and 4.

2.5 Voice discrimination

The task in our study is a typical voice discrimination task. It seems important to lay down the basis of a voice discrimination task and how it differs from a voice recognition task.

2.5.1 Voice discrimination and Voice recognition

Voice discrimination describes the task of differentiating two unfamiliar voices. Participants typically listen to pairs of audio stimuli and decide if both stimuli were uttered by the same person, or if two different people were speaking. Discrimination between unfamiliar voices is not the same as recognition of familiar voices.
In the case of the latter, the participants are exposed to samples of familiar voices and asked to identify who the speaker is. Sometimes, the choice is completely open and other times, participants have to choose between a selection of alternatives [Van Lancker et al., 1985].
These two tasks are considered to be different tasks for the human brain. One of the biggest arguments in this direction is that selective impairment can occur in the case of brain damage. In their paper, [Van Lancker and Kreiman, 1987] tested both tasks on patients with either left brain damage, right brain damage or bilateral damage. They found that voice recognition was difficult for left-brain impaired people, while voice discrimination impairment occurred after damage in either hemisphere [Van Lancker and Kreiman, 1987]. Furthermore, voice recognition could be achieved without voice discrimination, which shows that not only are the tasks separate, but also unordered : discrimination is not a preliminary step for recognition.
Voice recognition seems to be a task of pattern recognition and it is performed in a holistic way, while unfamiliar voices are processed in a terms of features [Schweinberger et al., 2018]. These features depend heavily on the individual, meaning the listener could focus on the feature that makes the voice special [Van Lancker et al., 1985].
In terms of automatic voice processing, both options exist. In case of voice-based access, the common process is similar to voice discrimination. The system disposes of a sample of the correct voice and compares it with the voice that is demanding access. The same features are extracted and a similarity score is calculated, the result of which will determine if access is granted [Hansen and Hasan, 2015].
In this study, we will focus only on the first task, voice discrimination. In other words, the listeners will hear stimuli made of two sentences and their task will be to decide if both were uttered by the same person or not.

2.6 Voice conversion

Voice conversion has been known in the literature for a few decades. It refers to the modification of one's voice to sound different, typically like another speaker's voice [Kinnunen et al., 2012].

> [A] voice conversion system only modifies the speaker-dependent characteristics of speech, such as formants, fundamental frequency (F0), intonation, intensity and duration, while carrying over the speaker-independent speech content. [Sisman et al., 2020, p. 3]

Voice conversion aims at modifying the segmental and suprasegmental features of voice, but without changing its linguistic contents, i.e. the words that are being uttered [Wu and Li, 2014]. Most algorithms follow a similar flow : they consist of three important functions : analysis, mapping and reconstruction [Sisman et al., 2020]. All the following equations have been taken from [Sisman et al., 2020, p. 3] and explain the process of voice conversion mathematically. Given x for the speech features of the source speaker and y for the speech features of the target speaker, the mapping function can be described as follows:

$$y = F(x) \tag{1}$$

F transforms the vocal features x into the y.
In the analysis process, the source voice's features x that are transformed in the function (1) are extracted. Considering X for the source speech signal, the analysis function is:

$$x = A(X) \tag{2}$$

The reconstruction of the target speech signal Y can then be put as :

$$Y = R(y) \tag{3}$$

In other words, and as shown in [Sisman et al., 2020], the whole process can be represented as a combination of functions :

$$Y = (R \circ F \circ A)(X) = C(X) \tag{4}$$

In words, the source voice X is analyzed and its features are extracted. These features are transformed to imitate the features of the target voice Y. Once that step has been completed, the vocal features are reconstructed into a speech signal.
These three steps describe the pipeline of most voice conversion systems although the specific steps are carried differently by every algorithm, some using statistical methods, others using deep learning networks [Sisman et al., 2020]. In practice, most voice conversion systems so far have focused on modifying the spectral features of voice, while prosody remained largely untouched [Şişman et al., 2017]. Sprocket, the voice conversion algorithm we used for the experiment is one of them [Kobayashi and Toda, 2018].

It will be described in more detail in section 3.2.3.

2.6.1 Applications of voice conversion

Voice conversion can have many real-life applications that range from personal speech synthesis and voice dubbing for movies to voice mimicry.
Voice conversion has many applications beneficial to users and society. An overview of those can be found in [Stylianou, 2009].

> [Voice conversion can be used for] for creating target or virtual voices, but also to model various effects (e.g., Lombard effect), synthesize emotions, to make more natural the dialog systems which use speech synthesis etc. Besides speech synthesis, however, Voice Transformation has other potential applications in areas like entertainment, film, and music industry, toys, chat rooms and games, dialog systems, security and speaker individuality for interpreting telephony, high-end hearing aids, vocal pathology and voice restoration. [Stylianou, 2009].

However, voice conversion can also be used maliciously. Voice conversion algorithms could be used to mimic another person's voice and enter their account, a process referred to as a spoofing attack [Alegre et al., 2013], [Wu and Li, 2014]. These types of attacks do not necessarily rely on voice conversion; they can also voice mimicry or replay previously recorded speech samples [Wu and Li, 2014]. Spoofing attacks can happen on any kind of system that performs speaker verification.
Such automatic systems are already in use, for example for banking applications. In Switzerland, PostFinance uses algorithms to either grant or deny access to a client's account over the telephone [2]. Such systems could be at risk if voice conversion were to be efficient enough to confuse their verification algorithm.
The rise of spoofing attacks has lead to an increase in anti-spoofing research [Wu and Li, 2014], [Alegre et al., 2013] and the two lines of research may be closer than they appear.

> The voice conversion and anti-spoofing studies can improve one another. For example, the techniques/features developed for anti-spoofing might be used to identify the weakness of voice conversion, which could be investigated to improve voice conversion techniques. On the other hand, the improved voice conversion techniques will drive the improvement of speaker verification. [Wu and Li, 2014, p. 12]

While the direct link between automatic voice identification and this present research may not be clear at first, it is in fact very relevant : the most important features that anti-spoofing systems rely on are likely

[2] https://www.postfinance.ch/de/privat/support/persoenliche-daten/authentifizierung-stimmerkennung.html

to be the biggest carriers of individuality in voice.

Another important application field for voice conversion and voice recognition is the forensic field. In some cases, perpetrators of crimes leave auditory traces. For example, they may be heard on a 911 call. This may have been the case when Trayvon Martin was killed during an altercation with George Zimmerman. A 911 call captured a scream for help in the background, which the parents of both parties (Zimmerman and Martin) identified as being their son [Hansen and Hasan, 2015]. In other cases, perpetrators could have made phone calls, e.g. death threats or sexual harassment calls [Dellwo et al., 2018]. Being able to identify the actual culprit has critical consequences.

In forensics, speaker identification can be performed both by human experts, with automatic speaker recognition systems or a combination of both. Systematic analysis methods are used especially when there is sufficient speech material, and speaker characteristics can be extracted [Hansen and Hasan, 2015]. Speaker features are typically used in the forensic field to identify a speaker. Usually, the sound evidence from the crime is compared with a sample provided by a suspect, as was the case in the Zimmerman case [Hansen and Hasan, 2015]. These features are chosen on the basis of different criteria, for example showing high between-speaker and low within-speaker variability, but they also have to be resistant to disguise or mimicry and robust to transmission among other factors [Hansen and Hasan, 2015]. [McDougall and Duckworth, 2018] offers an overview of the common features used in forensic speaker identification. He mentions formant frequencies of vowels as well as long-term formant frequencies, but also the fundamental frequency, properties of consonants, features of intonation and speaking rate as well as rhythmic properties [McDougall and Duckworth, 2018]. All these mirror the parameters that were discussed previously in the sections about voice individuality. This is normal, since some of this research did specifically mention forensic applications for their research [Dellwo et al., 2018], [Leemann et al., 2014].

Voice conversion, anti-spoofing studies and voice identification are all closely related and have real-life consequences. Having reliable systems to identify voices is of crucial importance: as more and more applications use automatic speaker verification to grant access, the outcome of a wrong decision can prove costly. In the worst case in the forensic field, it could send someone innocent in prison, or prevent the arrest of a real perpetrator [Hansen and Hasan, 2015].

Understanding the role of different voice features for voice individuality can make a great contribution in the understanding of voice recognition and discrimination. If humans and machines were to identify speakers with a better level of certainty, mistakes could be avoided in the future.

3 The study

3.1 Research question and Hypotheses

This study, as explained before, aims at understanding the respective effects of timbre and prosody in voice perception, more precisely in a voice discrimination task. What features are human listeners more likely to rely on and by extension, what features carry voice individuality more? A lot of studies have tried to figure out the most important acoustic features in voice individuality, but very few of them have tried to computationally modify the voices before perception tasks [Matsumoto et al., 1973], [Van Dommelen, 1990].
We used voice conversion to computationally modify the voices we used as stimuli : voice pairs matched in either their timbre (when the voice conversion algorithm had used the same target speaker), their prosody (when the voice conversion algorithm had used the same source speaker), or none of these features. The purpose of this manipulation was to separate the effects of timbre and prosody, which would in turn allow to disentangle them in later analyses. Voices we hear are usually unique or similar in every aspect, which means that correctly identifying them does not help knowing which features we consciously or unconsciously relied on to make a choice. Voice conversion gave us a chance to compare voices with only some features matching and see how they would be received.
During the voice discrimination experiment, participants were presented with the three types of stimuli pairs (same prosody, same timbre, no common feature). The goal was to examine in which condition listeners were likely consider two voices as being the same. Would two voices with the same timbre be considered more similar than two voices with the same prosody? The presence or absence of common variable was called “Feature combination” for the purpose of this research and it had three possible expressions : “timbre”, “prosody” and “none”. Our second variable of interest was duration, as we expected this variable to influence the performance. The duration was dichotomous, with either longer or shorter sentences.

Based on the amount of research that mentioned segmental features to be the most important contributors to voice individuality [Matsumoto et al., 1973], [Walden et al., 1978], [Singh and Murry, 1978], [Van Lancker et al., 1985], [Van Dommelen, 1990], [Nolan et al., 2011] and [Perrachione et al., 2019], we expected timbre to be the main factor used by the listeners to discriminate between two voices. On the other hand, based on [Leemann et al., 2014], [Dellwo et al., 2015] and [Dellwo et al., 2018] among others, we also expect supra-segmental cues to play a role, albeit smaller than that of timbre. The corresponding hypotheses H1a and H1b are outlined below.

H1 : Main effect of feature combination
H1a : If timbre characteristics play the major role in voice discrimination, the %same is higher for stimuli having “same timbre/same target” than for stimuli having “same prosody/same source” and “none”.
H1b : If prosodic cues also count for speaker discrimination, the %same is higher for stimuli “same

prosody" than for stimuli "none".

Note that according to the literature about voice discrimination, we expect timbre to play the most important role, with prosody being a more marginal factor.
Our second hypothesis H2 was based on the research that shows different performance with a longer duration [Fernández Gallardo, 2016], [Schweinberger et al., 2018].

H2 : Main effect of duration
For sentences with a longer duration, the %same will be different than in sentences with a short duration.
Our last hypothesis addressed the idea of an interaction between both variables. Literature suggests that a longer stimulus (up to a certain threshold) can help discriminate or recognize speakers more accurately [Schweinberger et al., 1997b], [Schweinberger et al., 2018]. More specifically, our hypothesis targeted the idea that if a longer duration impacts perception because of an increased phonetic variability [Schweinberger et al., 2018], the %same of prosody in particular would be impacted.

H3 : Interaction between the feature combination (prosody) and the duration
If listeners pay attention to prosodic cues, the %same for stimuli "same prosody/same source" is higher in longer utterance than in short utterances.

3.2 Method

3.2.1 Listeners

The initial target was set to 80 participants with a two-way recruitment strategy.
Half of the participants were to be recruited on-site at the University of Zurich. They would participate locally on the OpenSesame software that was installed on the main author's computer. These participants were approached directly on the university campus in the Summer of 2020 and were offered to take part if they had the time and fitted the criteria. They received a 10CHF compensation for their participation. It was ensured that they participated in a calm environment, allowing them a good hearing quality of the material. All the recruited people were students at either the University of Zurich or at the Swiss Federal Institute of Technology. The on-site recruitment was successful and the initial target of 40 participants was reached within two weeks. The success is mostly imputable to the relatively high remuneration for the short participation time, with many students hesitating to agree before being informed about the contribution.
The other forty participants were to be recruited through Prolific. Prolific is a relatively recent platform to recruit participants for online experiments. It is similar to Amazon's Mechanical Turk, but has the advantage of being built specifically for the scientific community [Palan and Schitter, 2018]. Prolific has standards for recruitment and paying and aims at fixing some of the problems common on MTurk, as for example the fact that many participants multi-task while taking part or that there is no verification of

the participants' identities. When creating an experiment, Prolific allows to screen for participants that correspond to certain pre-defined demographic criteria (e.g. native language, age, etc). Only participants that fit the criteria are recruited and notified by Prolific, they participate with a link provided through the platform and receive payment on the platform directly[Palan and Schitter, 2018]. The Prolific participants were rewarded with 5 CHF upon completion. The recruitment on Prolific started at a normal pace but participation slowed down drastically after a few days. After two months, only about half the participants had taken part in the study and it was eventually terminated.
In both cases, the participants had to fit in a similar set of criteria :

- no hearing- or speech-related disorder
- aged between 18 and 35 years old
- **on-site participants:** Swiss German as one of their native languages (having a second mother tongue was not a criterion of exclusion), with knowledge of High German coming from primary school or earlier
- **Prolific participants:** High German as one of their native languages (having a second mother tongue was not a criterion of exclusion).

Due to the fact that we did not have complete data for the Prolific participants, it was decided to only work with the data from the on-site listeners.

3.2.2 Voice material

The used voice material stemmed from five different male speakers, one of which was a professional speaker. These speakers are identified either by short name or by their ID number (in parenthesis): Co (1), Ma (2), Jo (4) and Al (6), while the professional speaker is referred to as ProfSpeak (7).
The speakers were recorded while reading 137 different sentences. There were three different types of sentences : 44 longer 5-word sentences, 83 shorter 2-word sentences and finally, 10 yes/no questions. The latter were not used for our purposes. Examples of short and long sentences are found in Table 1.

Table 1: Sentence examples

Type	Sentence
Long	*Die Vorsitzende fährt ein Auto.*
Long	*Die Künstlerin malt ein Gemälde.*
Long	*Der Esel blockiert den Wanderweg.*
Short	*Er giesst.*
Short	*Er joggt.*

3.2.3 Sprocket voice conversion

All the speech material we used was computationally altered, meaning that every voice heard in the experiment was a synthetic voice.
The voice conversion algorithm "sprocket" was used for this task [Kobayashi and Toda, 2018]. The algorithm uses a parallel speech dataset, which means that the set consists of two different voices producing the same linguistic content. i.e. the same words and sentences. One of these voices will act as the source voice, while the second will be the target. In a five-step process, the source voice will be altered to resemble the target voice [Kobayashi and Toda, 2018]. First, the data must be prepared and aligned. Then, important acoustic features including F0, aperiodicity (???), and mel-cepstrum (???) are extracted from the speech signals in both the source and the target voice [Kobayashi and Toda, 2018]. In a third step, it the algorithm deals with the necessity to have frame-aligned feature vectors. Even if two speakers utter the exact same word, they are highly unlikely to have the exact same speaking style. One could speak faster, rush certain words, exaggerate a stress, etc. This leads to discrepancies in time and the voices need to be aligned frame by frame. The alignment process used was based on dynamic time warping. Dynamic time warping refers to a process that allows to see the similarities between two objects moving at different speeds. For example, if you had two people walking at different paces, this process would be able to still detect the similarities in their walks [Olsen et al., 2016], [Kobayashi and Toda, 2018]. The final step was the actual conversion of the voice. The features converted were the F0 and the mel-cepstrum, while the aperiodicity, speaking rate, temporal structure of the F0 and the power structure of the speech were retained. These later factors can be summed as belonging to a voice's prosody, which thus remained that of the source voice. These variables remained untouched because the algorithm does not dispose of features to modify the temporal structure. Rather, it maps spectral features segment by segment from the source to the target voice [Kobayashi and Toda, 2018].
To sum up, the end result is a synthetic voice that has the prosody of the source voice and the timbre of the target voice.
The first set of voices were thus voices that were modified to resemble that of other speakers. A second dataset was created later, the self-voice conversion dataset. Like the name suggests, it was created by using the sprocket algorithm with the same voice as source and target. This would result in a synthetic voice that would be an exact replica of the human voice used as source and target. The choice was made to only work with synthetic voices to have a better consistency in the stimuli heard by the user as well as to avoid a stark contrast between the stimuli where the voice was altered versus when it was natural.

3.2.4 Stimuli

The listeners heard 56 sentence pairs. Each pair consisted in the same sentence spoken twice, with a silence of 0.7 second in-between. Background noise was added on the sentences to combat the effect of having a perceptible robotic tone in the synthetic voices, especially those using different source and

target speakers. The level of all samples was equalized to 70 dB to ensure consistency. That way, once the listener had adjusted the hearing level to their preferred setting, they were ensured it would remain the same throughout the experiment.
For more clarity, the naming convention we used is explained as follows:

- VCD_speakername : Voice Conversion with different speakers, the given speakername refers to the target speaker, as the source is always the same professional speaker
- VCS_speakername : Voice Conversion with the same speaker, the given speakername refers to the speaker whose voice is being replicated by the algorithm.

To be able to disentangle the effects of timbre and prosody in voice discrimination, the participants were presented with three different types of sentence pairs.
condition A, both voices had the same timbre, but a different prosody. In other words, the target speaker was the same, but they had a different source speaker. In **condition B**, the two voices had a different timbre, but the same prosody; meaning that they shared the same source speaker. In **condition C**, none of these factors overlapped.
These sentences were combined in three different conditions:

- **condition A** : same timbre, e.g. VCS_Al combined with VCD_Al. The two sentences share the same target, Al. In the first one, Al is also the source while in the second, the source is the ProfSpeak.
- **condition B** : same prosody, e.g. VCD_Al combined with VCS_ProfSpeak. In that case, the two sentences share the same source, ProfSpeak. In the first, the target is Al and in the latter, the target is the ProfSpeak himself.
- **condition C** no common feature, e.g. VCS_Al combined with VCS_Co. For this category, all VCS voices were combined with one another and no factor overlapped. As there were voice combinations possible, there were a bigger number of this type of sentences in the experiment, as shown in Figure 1.

An overview of all the different combinations can be seen in Figure 1.

Figure 1: Speaker combinations used for the experiment

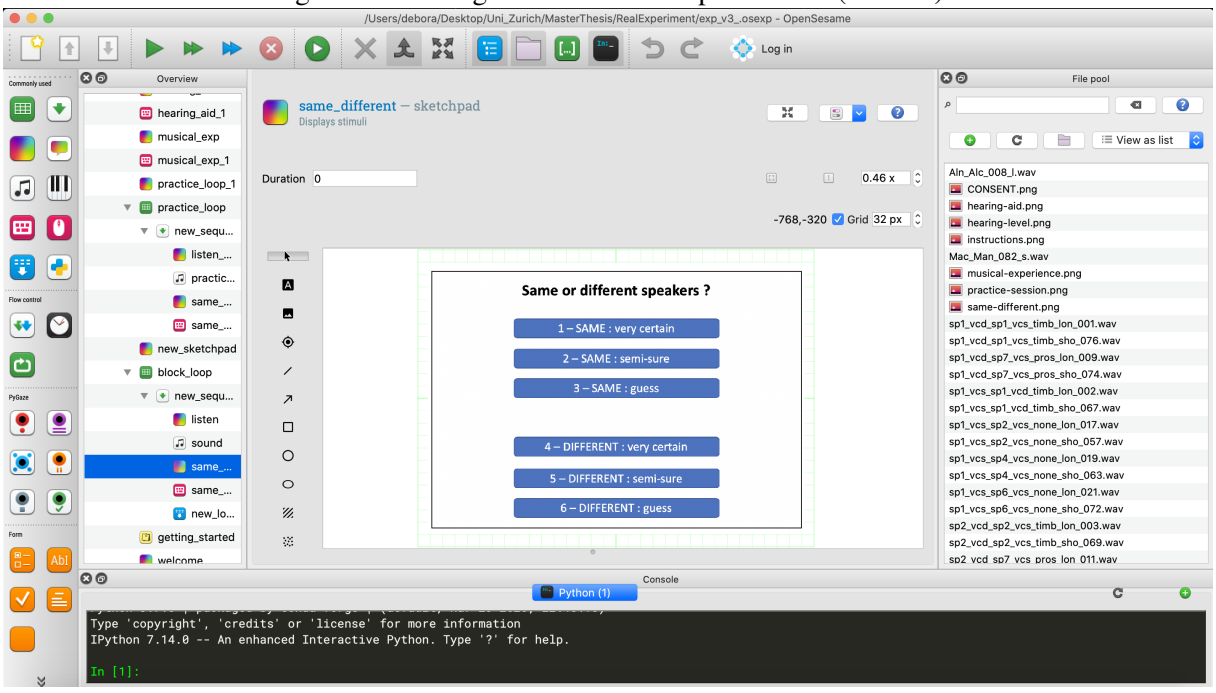
Condition A = same timbre	Condition B = same prosody	Condition C = diff timbre/prosody
1) VCD_Co + VCS_Co	1) VCD_Co + VCS_ProfSpeak	1) VCS_Co + VCS_Ma
2) VCD_Ma + VCS_Ma	2) VCD_Ma + VCS_ProfSpeak	2) VCS_Co + VCS_Al
3) VCD_Jo + VCS_Jo	3) VCD_Jo + VCS_ProfSpeak	3) VCS_Co + VCS_Jo
4) VCD_Al + VCS_Al	4) VCD_Al + VCS_ProfSpeak	4) VCS_Ma + VCS_Al
		5) VCS_Ma + VCS_Jo
		6) VCS_Al + VCS_Jo

For each speaker combination and each condition, two long sentences and two short sentences were heard. The total amounted to 56 sentences : 16 trials for the condition A (same timbre), 16 for the condition B (same prosody) and 24 for the condition C (different timbre and prosody).

3.2.5 Experimental setup

The experiment was setup on OpenSesame, which is a program specifically designed to create experiments for psychology, linguistics, neuroscience and other social sciences [Mathôt et al., 2012]. OpenSesame is free, cross-platform and supports Python scripting [Mathôt et al., 2012]. It was ideal for our purpose because it supports audio files and it is well-documented with an active forum for questions.
A screenshot of the OpenSesame environment can be seen in the Figure 2. Although the software looks a little old-school, it is powerful.

Figure 2: Working environment in OpenSesame (on Mac)



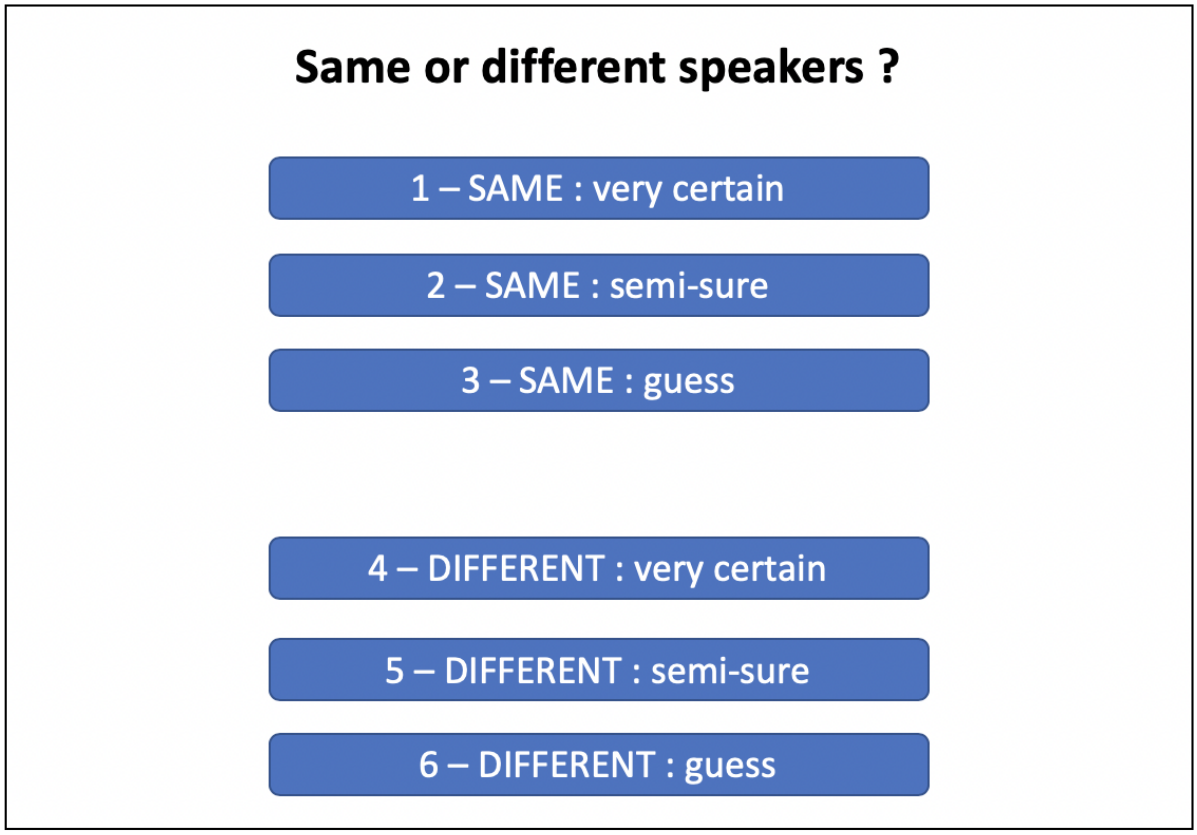
OpenSesame enables the creation of experiments that run both on the computer and online. As running it locally requires the installation of the relatively heavy program (1.35GB), it was crucial to provide the opportunity for participants to take part online. We did this using JATOS (Just Another Tool for Online Studies). JATOS is a server specifically designed to help run studies online, and it has high compatibility with OpenSesame, as the latter provides a built-in option to export the studies for JATOS. However, it is to be noted that not all the features that are available on the Desktop version can be exported to JATOS. To ensure that all listeners (on-site as well as on Prolific) participated under the same

conditions, we chose to work only with functionalities that were available for the web-supported version. For instance, we used keyboard-generated responses instead of multiple choice forms.
The JATOS server, once initialized, can host a number of different experiments. Different batches of workers can be created for these experiments, for which different types of participation links can be generated. Single links grant access to the experiment only once, while multiple links allow access an unlimited number of times [Lange et al., 2015]. JATOS also allows to generate link that can be used for recruitment through a platform, and relate both platforms to collect the results, which is what we used in our research.

3.2.6 Discrimination task

Participants were subjected to a classic discrimination task.
They were asked to answer the sociolinguistic questionnaire about their age, their linguistic background and their studies first. This order was chosen to avoid having situations in which the participants forgets to fill the questionnaire after having taking part.
The listeners heard every sentence pair once with no possibility to listen to it again and had to decide whether the two sentences were spoken by the same person or not. They were also asked for their level of confidence in the answer. They were advised to answer as quickly as possible without thinking, but they were given no time limit : the next sentence pair was only heard once they had answered the previous one.
The participants received precise instructions at the very beginning of the experiment. The first few screens also verified that they had no speech or hearing disorder and that they agreed to their responses being recorded and stored. The listeners went through a trial run with two sentence pairs, the results for which weren't recorded. This was to make sure they understood how the experiment worked.
An image of their multiple choice screen is shown in Figure 3.

Figure 3: Participants' view of the multiple choice



Filling the socio-linguistic questionnaire lasted about 5 minutes in average. Taking part to the experiment took between 5 and 10 minutes, depending on the individual. The total duration of interaction was thus approximately 15 minutes. Some listeners expressed that they found the experiment really difficult and confusing, while others seemed to find it easy.

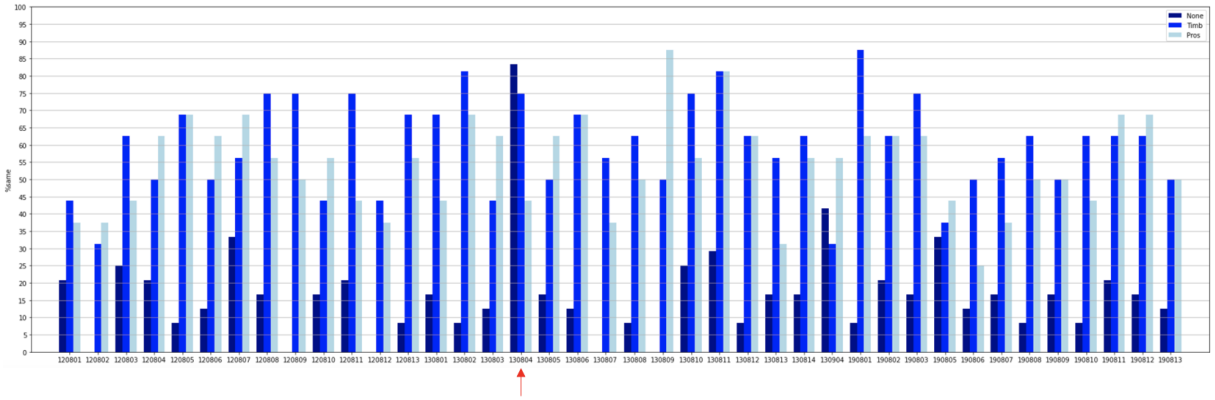
4 Data Analysis and statistics

The data was processed with both R and Python to perform statistical tests and generate visualizations of the results. Packages such as ggplot, rstatix, tidyverse, ez and tibble were useful with R, while packages such as numpy, pandas, matplotlib and seaborn were used in Python.
The dependent variable in our data analysis is the percentage of "same" responses reached in every condition. The other option would have been to rate their accuracy, but this would not make sense in this setting, as there is no right or wrong answers in the conditions A and B, where the voices are not identical but do share some common features. It is also not possible to tell if voices in condition A are more identical than voices in condition B. In fact, knowing which feature influences listeners more was the purpose of this study, so we could not define what an accuracy measure would represent. It would have possible to have an accuracy score for the condition C because the voices are completely different. In that case, a perfect accuracy score would have meant that the participants perceived the voices as different every time.
For the sake of consistency, the %same will be analyzed in every condition.

4.1 General overview

A first visualization in Figure 4 shows the global results for the %same by listener and feature combination.

Figure 4: %same by listener and feature combination



As a visual observation, Figure 4 shows that there is a strong tendency for the %same to be much lower in the condition C (no common feature). There are some listeners that show relatively high %same in that condition, but none as much as listener 130804, which is an extreme outlier. The figure does not allow to see whether timbre or prosody scored higher in the %same, as there is a lot of variability between the different participants.

4.2 ANOVA

The next step was to run an ANOVA analysis. ANOVA allows to reveal the factors that affect the data. In this case, a two-way repeated ANOVA analysis was made, as two factors were tested on the same subjects : feature combination and duration. The extreme outlier, 130804, was removed for the ANOVA analysis and all further statistical tests.
Table 2 shows the result of the two-way repeated ANOVA analysis.

Table 2: Two-way repeated ANOVA analyss

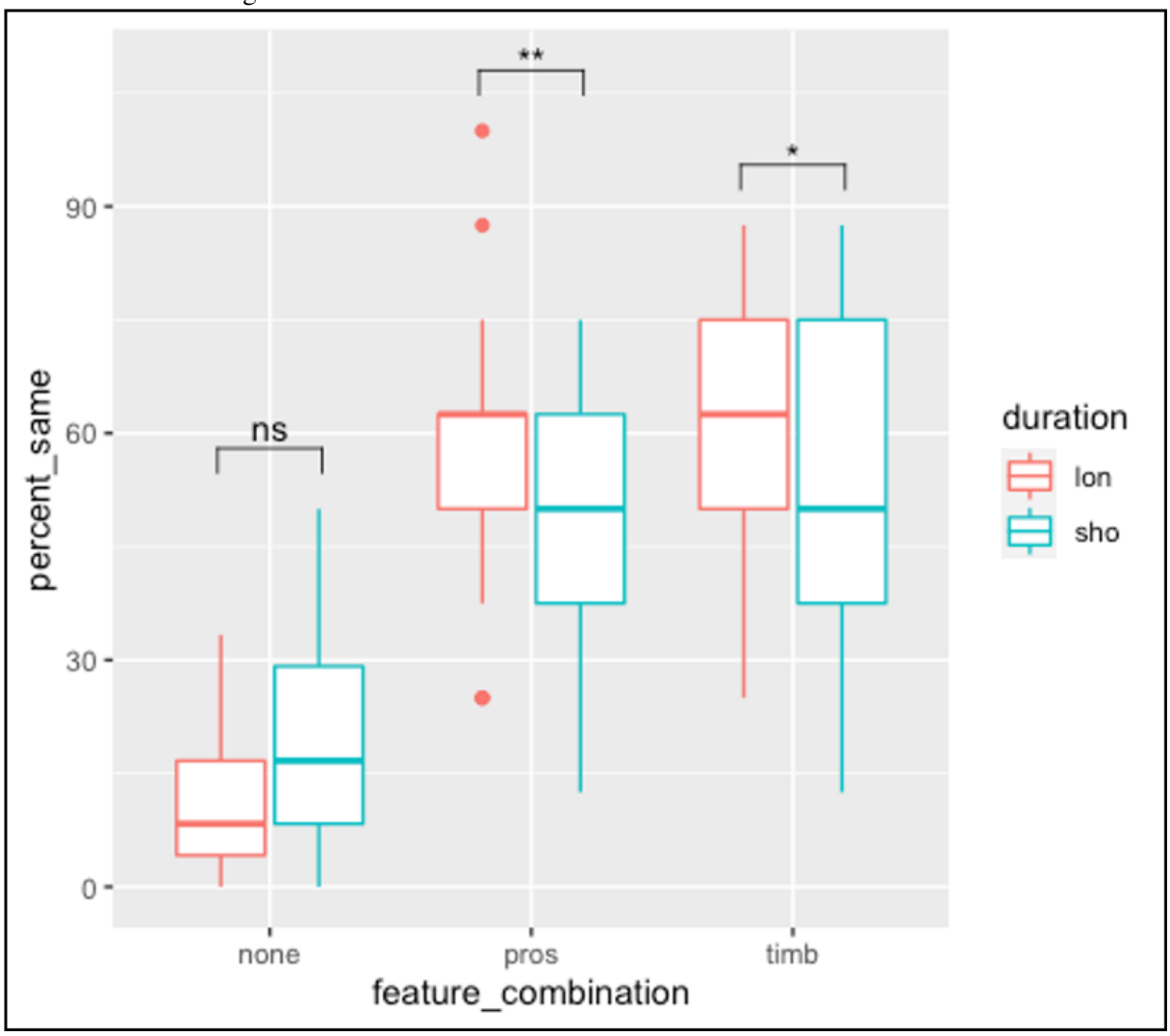
Effect	DFn	DFd	F	p	p<.05	ges
Duration	1	38	4.009	5.20e-02		0.016
Feature Combination	2	76	170.624	7.87e-29	*	0.615
Duration:Feature Combination	2	76	6.707	2.00e-03	*	0.039

The ANOVA analysis shows no main effect of Duration and a main effect of Feature Combination. However, there is also a proven interaction effect.The presence of an interaction effect means that the main effects can't be interpreted on their own, as they depend on the other feature [Frost, 2017]. The effect of feature combination depends on duration and vice-versa, so that the main effects can be ignored.

4.3 Interaction effect between Feature Combination and Duration

Figure 5 shows the interaction between Feature Combination and Duration and where the two variables come into play together.

Figure 5: Interaction between Feature Combination and Duration



The data corresponding to the Figure 5 can be seen in the Table 3.

Table 3: Interaction between Feature Combination and Duration

feat_comb	.y.	group1	group2	df	p	p.adj	p.adj.signif
none	%same	lon	sho	38	0.053	0.053	n.s.
pros	%same	lon	sho	38	0.01	0.01	**
timb	%same	lon	sho	38	0.031	0.031	*

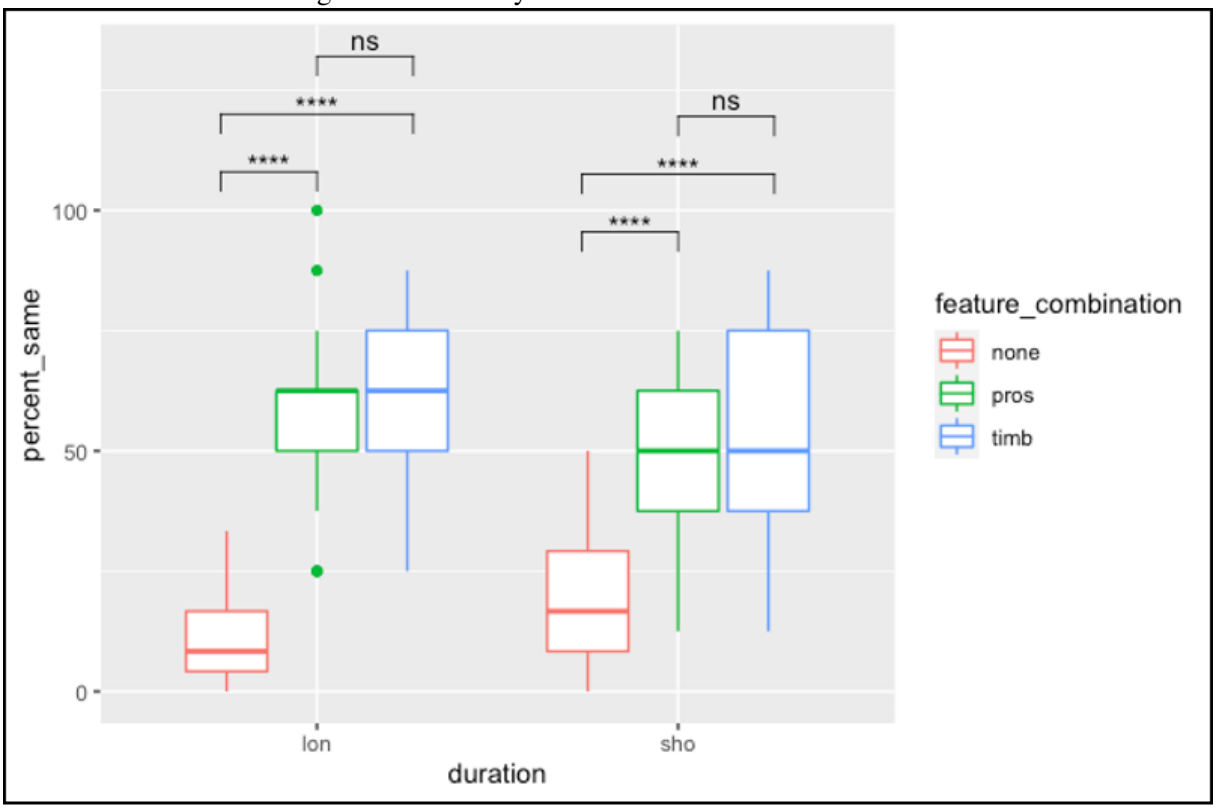
As shown in the plot and the table, the difference in duration is significant for the conditions A and

B. The pairwise comparison between a short and a long duration is significant in both case, with p≤ 0.01 for prosody and p≤ 0.05 for timbre. It is not significant, however, for the condition C, when the voices have no common feature.
It is also the only condition in which the %same is higher when the duration is short, meaning that with a longer exposure, the participants actually tended to be better at discriminating the two voices, although the difference is not significant statistically. The effect of duration is only statistically significant when the voices share the same timbre or the same prosody.
These results allow us to examine some of our hypotheses.
Our second hypothesis stated that with a longer duration, the %same would rise. As previously explained, main effects have to be disregarded to the profit of the interaction effect. Thus, we have to abandon this hypothesis.
Our third hypothesis stated that if listeners paid attention to prosodic cues, the %same for the stimuli sharing the same prosody would be higher in longer sentences, as they would be provided with a wider repertoire of phonemes and more prosodic material. This hypothesis is confirmed with a significance level under 0.01. We can also see that a longer duration lead to a higher %same when the timbre was identical, meaning that exposure time mattered for this feature as well.

4.4 Prominence of timbre and prosody

Our first hypothesis mentioned that timbre would be the most important factor affecting speaker discrimination (H1a) with prosody playing a possible role as well (H1b). As an interaction effect was shown, it only makes sense to analyze the data with a duration split. Figure 6 shows a comparison of the role of the different feature combinations, by duration. Note that the data shown is the same as in Figure 5, but the arrangement of the data allows for a clearer view of the result.

Figure 6: %same by feature combination and duration



The statistical analysis with the precise numbers corresponding to the Figure 6 is shown in Table 4. The first column shows the duration, while the second and third columns show the two feature combinations that are being compared. The last three columns express the statistical significance of the analysis. It is made clear that in both long and short sentences, the difference between pros and timb is not significant, meaning that none has a significantly bigger impact on the %same.

Table 4: Role of timbre, prosody and none by duration

duration	group1	group2	statistic	df	p	p.adj	p.adj.signif
long	none	pros	-16.4	38	8.30e-19	2.49e-18	****
long	none	timb	-16.2	38	1.12e-18	3.36e-18	****
long	pros	timb	-1.31	38	1.97e-1	5.91e-1	n.s.
short	none	pros	-9.99	38	3.52e-12	1.06e-11	****
short	none	timb	-9.02	38	5.49e-11	1-65e-10	****
short	pros	timb	1.47	38	1.49e-1	4.47e-1	n.s.

The plot in Figure 6 shows that in both longer and shorter sentences, the %same for both “prosody”

and "timbre" are significantly higher than for "none". The significance level is remarkably high with $p \leq 0.0001$. A higher %same when the voices are partially equivalent in contrast to when they are completely different was expected, and the high significance levels show that these two factors were highly relevant for the participants. The surprising result was that there is no significant difference in the %same between the conditions "timbre" and "prosody". In other words, our hypothesis H1a is rejected, while H1b is confirmed. This result suggests that participants relied on prosody and timbre equally and that none of these two features is more important than the other for speaker discrimination.

This finding was most surprising because it contradicts most of the body of literature, which usually mentions spectral features as the foremost important factor for both language discrimination and language recognition [Wolf, 1972], [Kuwabara, 1995], [Fernández Gallardo, 2016], [Matsumoto et al., 1973], [Walden et al., 1978], [Singh and Murry, 1978], [Van Lancker et al., 1985], [Van Dommelen, 1990], [Nolan et al., 2011] and [Perrachione et al., 2019].

5 Discussion

The results yielded by the statistical analysis brought both expected and unexpected results.
First, it was confirmed that voices that share a common attribute were perceived the same much more often that voices that had nothing in common. This was an expected result and it is normal that voices that are the same to a certain degree will be recognized as such by listeners.

We had also expected to have an effect of duration, given that the exposure to the voice increases. It was mentioned in the literature that it isn't clear if a longer exposure leads to better perception because of the increased duration itself or if it because of a more complete phonetic sample [Schweinberger et al., 2018]. Our hypothesis 3 mentioned that a higher %same for longer duration would indicate that listeners rely on prosodic cues. This hypothesis was confirmed, but voices with the same timbre also received significantly higher %same in the longer duration. Our results did not uncover a single factor, and it is likely that both the longer exposure and the bigger phonetic variability play a role in the impact of longer duration.
It would interesting to create a new study design aiming at disentangling the two effects. For example, it could be possible to increase the duration of the stimulus, both with and without added phonetic variability, e.g. by repeating the short sentence twice. If the % were to increase even in the absence of broader phonetic sample, it would be a clear indication that longer stimuli help for the perception of timbre as well.

We expected timbre to be an important factor in voice discrimination, and that assumption was correct. We can thus confirm the findings of a broad body of literature before us. Spectral features such as F0 or formant frequencies were mentioned in a lot of papers, such as [Matsumoto et al., 1973], [Walden et al., 1978], [Singh and Murry, 1978], [Van Lancker et al., 1985], [Van Dommelen, 1990], [Nolan et al., 2011] and [Perrachione et al., 2019].

As mentioned, the surprising result was the absence of a significant difference between the importance of timbre and prosody. Based on the literature cited above, there was a strong expectation to see timbre take a prominent role in voice discrimination, but this expectation was contradicted by our results. There was no statistical proof that voices with the same timbre were rated as the same voice more than the voices with the same prosody.
This is particularly interesting because the role of prosody isn't understood as well as that of timbre. Indeed, prosody is challenging to analyze because it is affected by both short-term and long-term changes in voice [Ming et al., 2016]. This difficulty is salient not only in the analysis of voice itself, but also in the research of voice conversion, where prosody has started to attract attention only in recent years [Ming et al., 2016], [Şişman et al., 2017]. In the past few years, researchers have tried to manipulate not only the spectral features, but also prosodic cues [Ming et al., 2016], [Şişman et al., 2017], but before that,

voice conversion systems focused only on converting spectral features. Sprocket, the algorithm we used for this study, falls into this category [Kobayashi and Toda, 2018]. This so-to-speak deficiency, however, is precisely what allowed us to separate timbre from prosody and thus, to test the perceptual role of the latter.

Using Voice Conversion for perception tests is a very new method and it could gain prominence in future researches, as it allows to test the effects of different voice features separately. This is an example of how new methods and technologies can help researchers overcome challenges, typically those related to disentangling the effects of features that are hard to examine separately.

But what do these results mean exactly? There are a few possible explanations.

The first and most likely one is that prosody indeed plays a crucial role in voice discrimination. The fact that prosody varies from speaker to speaker has been known and described, for example in [Dellwo et al., 2018] or [Leemann et al., 2014], where it was shown that speech rhythm is a carrier of individuality. Although the presence of these individual prosodic cues was proved, it wasn't very clear how listeners used these cues when perceiving voices. This present study could be a first step towards understanding the significance of prosody's perceptual role, which may well have been underestimated until now. It opens the way for more similar researches and a promising new line of studies that could focus exclusively on the role of prosodic cues. Just like segmental features, prosody is made of a lot different elements, such as the intonation, the rhythmic pattern, the rate of speech, the stress, unique pronunciation of some phonemes, etc. It would be deeply interesting to dive into these elements and examine them individually. It is likely that very interesting findings could result from such studies. It is one of the first times that prosody proves to matter in voice perception tasks and it is a promising first step for future research in this domain.

Even though it is most likely that prosody matters, alternative explanations for our results, involving both the participants or the speakers, can be put forward.

It is important to keep in mind that the sample size of the speakers was relatively low, as there were only five of them. That could have affected the result, if for example two of the speakers had similar timbres. If these two speakers were used as target speakers by the voice conversion algorithm, they would end up with very similar synthetic voices, as their prosody was identical and their timbre near-identical. Their %same would be considered in the "prosody" feature combination, as that was the manipulated feature, but the listener may have relied on the timbre to make their choice. It was also mentioned in the literature that the strategies used by the listeners depend on the voice [Van Lancker and Kreiman, 1987]. Given our sample size, these different strategies could have affected the results much more than if we had worked with a bigger number of speakers.

A second important factor may reside in the listeners themselves. The participants were all students and thus, they were familiar with research processes. They sometimes expressed an interest towards the "real underlying" matter, trying to understand what was really asked of them instead of answering without giving it much thought. This cognitive process of looking further than the obvious questions may have altered their answers in ways that are hard to assess.

Replicating this test with more voices, or voices that have more distinguishable timbres would be a first step in trying to uncover the actual importance of prosody in voice discrimination.
One influence we can dismiss in our case is the language familiarity effect, which should not have had any impact in our study, given that all listeners were familiar with High German to a similar extent.
All in all, our research was successful in demonstrating the use of Voice Conversion for perceptual tasks such as voice discrimination or voice recognition, which something that can be replicated in the future with more voice samples or a more in-depth analysis of the different features that constitute voice.
Additionally, we could uncover the potential role of prosody in voice perception. This is a crucial achievement, as the role of prosody isn't understood well, and this result opens up a new range of possibilities for prosody research.

6 Summary and Conclusion

This section will briefly summarize and conclude this work with indications for future research.
In this study, we have been investigating the role of timbre and prosody for tasks such as voice discrimination.
A classic voice discrimination task was created. Participants heard sentence pairs and had to decide whether the two voices belonged to the same person or if they were two different speakers. These voice pairs came in three different combinations : different voices, voice sharing the same timbre and voice sharing the same prosody. These combinations were made possible through the technology of voice conversion, which was performed with an algorithm called Sprocket. Two voices are inputted into the algorithm : the source voice and the target voice. The output of the program is a synthetic voice that has preserved the prosody of the source voice but imitates the timbre of the target speaker. Five male speakers provided the initial voice material. Thanks to the voice conversion technique, it was possible to test the effects of timbre and prosody separately, which was the focus of this research. Working with voice conversion is a relatively new method and was proved to be successful in this experiment.
Timbre was expected to be the most important characteristic involved in voice discrimination, with prosody playing a less important role. The prediction for the role of duration was that a longer duration would lead to the listeners hitting "same" more often.
The experiment was conducted with 40 listeners and the data was analyzed after having removed one extreme outlier. The performed two-way repeated ANOVA measure established an interaction effect between the feature combination, i.e. the common feature, and the duration. It was shown that duration had a significant effect on the feature combinations "timbre" and "prosody" but not of "none".
Additional analyses showed that contrary to all expectations, timbre did not seem to influence the listeners' perception more than prosody. The difference between the two feature combinations was not significant regardless of the duration. Listeners relied on prosody as much as they did on timbre to decide whether two voices are distinct or not. None of the literature that has been reviewed mentioned such a prominent role of prosody and it will be interesting to see if this result will be confirmed in the future. Is timbre the most important factor, as hinted by the previous literature, or will new studies validate that prosody may be as crucial as timbre in voice discrimination, something that had been overlooked until now?
Further work should focus on replicating this experiment, ideally with a bigger sample of both voices and sentences. It may be interesting to try and have voices that have distinguishable voice timbres as a means to avoid having voices post-conversion that are very similar without sharing the same target, supposing a similar algorithm is used.
As mentioned before, it would be interesting to focus on prosody more deeply in the future and try to unravel the effect of the individual elements that make up prosody.
The future promises many exciting novelties in the fields of voice conversion and voice discrimination. It will be interesting to understand in details how humans perform this task and how these results can

contribute to the automatic recognition of voice. Voice conversion algorithms are becoming increasingly convincing at imitating their human targets. Deep fakes are already on the rise and although they could pose threats to the democracy, for example by making prominent personalities appear as saying something they never actually did, it is likely that the technologies to counter them will be improving in parallel. We hope that our contribution in understanding the role prosody could hold in such applications will be useful.

7 Aknowledgements

I would like to thank Dr. Elisa Pellegrino for her availability, her guidance and her patience, not to mention her expertise in the field. I could not have wished for a better person to accompany me through this project. Similarly, many thanks to Prof. Dr. Volker Dellwo for the important contributions, especially during the writing process; as well as the opportunity to write my thesis under his mentoring. A special mention to Dr. Thayabaran Kathiresan for his relevant notes and help in the understanding of the technical sides of this research.
Finally, many thanks to my husband and my family, who patiently put up with me for the whole duration of my studies, especially during the most stressful phases.

References

[Alegre et al., 2013] Alegre, F., Amehraye, A., and Evans, N. (2013). Spoofing countermeasures to protect automatic speaker verification from voice conversion. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3068–3072. IEEE.

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

[Cook and Wilding, 1997] Cook, S. and Wilding, J. (1997). Earwitness testimony: Never mind the variety, hear the length. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, 11(2):95–111.

[Dellwo et al., 2018] Dellwo, V., French, P., He, L., Frühholz, S., and Belin, P. (2018). Voice biometrics for forensic speaker recognition applications.

[Dellwo et al., 2012] Dellwo, V., Leemann, A., and Kolly, M.-J. (2012). Speaker idiosyncratic rhythmic features in the speech signal. In *Thirteenth Annual Conference of the International Speech Communication Association*.

[Dellwo et al., 2015] Dellwo, V., Leemann, A., and Kolly, M.-J. (2015). Rhythmic variability between speakers: Articulatory, prosodic, and linguistic factors. *The Journal of the Acoustical Society of America*, 137(3):1513–1528.

[Fant, 1981] Fant, G. (1981). The source filter concept in voice production. *STL-QPSR*, 1(1981):21–37.

[Fernández Gallardo, 2016] Fernández Gallardo, L. (2016). Human and Automatic Speaker Recognition over Telecommunication Channels. (December):5–34.

[Ferragne and Pellegrino, 2004] Ferragne, E. and Pellegrino, F. (2004). A comparative account of the suprasegmental and rhythmic features of british english dialects. *Modelisations pour l'Identification des Langues*.

[Ferrand, 2002] Ferrand, C. T. (2002). Harmonics-to-noise ratio: an index of vocal aging. *Journal of voice*, 16(4):480–487.

[Fischer-Jørgensen, 1989] Fischer-Jørgensen, E. (1989). Phonetic analysis of the stød in standard danish. *Phonetica*, 46(1-3):1–59.

[Fleming et al., 2014] Fleming, D., Giordano, B. L., Caldara, R., and Belin, P. (2014). A language-familiarity effect for speaker discrimination without comprehension. *Proceedings of the National Academy of Sciences*, 111(38):13795–13798.

[Frost, 2017] Frost, J. (2017). Understanding interaction effects in statistics.

[Furui, 1986] Furui, S. (1986). Research of individuality features in speech waves and automatic speaker recognition techniques. *Speech communication*, 5(2):183–197.

[Greene and Mathieson, 2001] Greene, M. and Mathieson, L. (2001). The voice and its disorders, 6th edn. london & philadelphia.

[Hansen and Hasan, 2015] Hansen, J. H. and Hasan, T. (2015). Speaker recognition by machines and humans: A tutorial review. *IEEE Signal processing magazine*, 32(6):74–99.

[Howard and Silverman, 1976] Howard, J. H. and Silverman, E. B. (1976). A multidimensional scaling analysis of 16 complex sounds. *Perception & Psychophysics*, 19(2):193–200.

[Howell, 2016] Howell, I. (2016). *Parsing the spectral envelope: Toward a general theory of vocal tone color*. New England Conservatory of Music.

[Kinnunen et al., 2012] Kinnunen, T., Wu, Z.-Z., Lee, K. A., Sedlak, F., Chng, E. S., and Li, H. (2012). Vulnerability of speaker verification systems against voice conversion spoofing attacks: The case of telephone speech. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4401–4404. IEEE.

[Kobayashi and Toda, 2018] Kobayashi, K. and Toda, T. (2018). sprocket: Open-source voice conversion software. In *Odyssey*, pages 203–210.

[Kreiman et al., 2003] Kreiman, J., Vanlancker-Sidtis, D., and Gerratt, B. R. (2003). Defining and measuring voice quality. In *ISCA Tutorial and Research Workshop on Voice Quality: Functions, Analysis and Synthesis*.

[Kuwabara, 1995] Kuwabara, H. (1995). of speaker individuality : Control and. *Science*, 16:165–173.

[Kuwabara and Sagisak, 1995] Kuwabara, H. and Sagisak, Y. (1995). Acoustic characteristics of speaker individuality: Control and conversion. *Speech communication*, 16(2):165–173.

[Ladefoged and Maddieson, 1996] Ladefoged, P. and Maddieson, I. (1996). *The sounds of the world's languages*, volume 1012.

[Lange et al., 2015] Lange, K., Kühn, S., and Filevich, E. (2015). " just another tool for online studies"(jatos): An easy solution for setup and management of web servers supporting online studies. *PloS one*, 10(6):e0130834.

[Lau et al., 2004] Lau, Y. W., Wagner, M., and Tran, D. (2004). Vulnerability of speaker verification to voice mimicking. In *Proceedings of 2004 International Symposium on Intelligent Multimedia, Video and Speech Processing, 2004.*, pages 145–148. IEEE.

[Leemann et al., 2014] Leemann, A., Kolly, M.-J., and Dellwo, V. (2014). Speaker-individuality in suprasegmental temporal features: Implications for forensic voice comparison. *Forensic science international*, 238:59–67.

[Łetowski, 2014] Łetowski, T. (2014). Timbre, tone color, and sound quality: concepts and definitions. *Archives of Acoustics*, 17(1):17–30.

[Mathôt et al., 2012] Mathôt, S., Schreij, D., and Theeuwes, J. (2012). Opensesame: An open-source, graphical experiment builder for the social sciences. *Behavior research methods*, 44(2):314–324.

[Matsumoto et al., 1973] Matsumoto, H., Hiki, S., Sone, T., and Nimura, T. (1973). Multidimensional representation of personal quality of vowels and its acoustical correlates. *IEEE Transactions on Audio and Electroacoustics*, 21(5):428–436.

[McDougall and Duckworth, 2018] McDougall, K. and Duckworth, M. (2018). Individual patterns of disfluency across speaking styles: a forensic phonetic investigation of standard southern british english. *International Journal of Speech, Language & the Law*, 25(2).

[McKinney, 2005] McKinney, J. C. (2005). *The diagnosis and correction of vocal faults: A manual for teachers of singing and for choir directors*. Waveland Press.

[Merriam-Webster.com, 2020] Merriam-Webster.com (2020). "voice". `https://www.merriam-webster.com/dictionary/voice`. Accessed: 2020-11-29.

[Ming et al., 2016] Ming, H., Huang, D., Xie, L., Zhang, S., Dong, M., and Li, H. (2016). Exemplar-based sparse representation of timbre and prosody for voice conversion. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5175–5179. IEEE.

[Nishio and Niimi, 2008] Nishio, M. and Niimi, S. (2008). Changes in speaking fundamental frequency characteristics with aging. *Folia phoniatrica et logopaedica*, 60(3):120–127.

[Nolan, 1997] Nolan, F. (1997). Speaker recognition and forensic phonetics. *The handbook of phonetic sciences*, pages 744–767.

[Nolan et al., 2011] Nolan, F., McDougall, K., and Hudson, T. (2011). Some acoustic correlates of perceived (dis) similarity between same-accent voices. In *ICPhS*, pages 1506–1509.

[Olsen et al., 2016] Olsen, N. L., Markussen, B., and Raket, L. L. (2016). Simultaneous inference for misaligned multivariate functional data. *arXiv preprint arXiv:1606.03295*.

[Oxford-University-Press, 2020] Oxford-University-Press (2020). "definition of voice". `https://www.lexico.com/en/definition/voice`. Lexico.com. Accessed: 2020-11-29.

[Palan and Schitter, 2018] Palan, S. and Schitter, C. (2018). Prolific. ac—a subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, 17:22–27.

[Patterson et al., 2010] Patterson, R. D., Walters, T. C., Monaghan, J. J., and Gaudrain, E. (2010). Reviewing the definition of timbre as it pertains to the perception of speech and musical sounds. In *The Neurophysiological Bases of Auditory Perception*, pages 223–233. Springer.

[Perrachione et al., 2019] Perrachione, T. K., Furbeck, K. T., and Thurston, E. J. (2019). Acoustic and linguistic factors affecting perceptual dissimilarity judgments of voices. *The Journal of the Acoustical Society of America*, 146(5):3384–3399.

[Ramakrishnan, 2013] Ramakrishnan, A. (2013). What is the difference between pitch and formant frequency?

[Roebuck and Wilding, 1993] Roebuck, R. and Wilding, J. (1993). Effects of vowel variety and sample lenght on identification of a speaker in a line-up. *Applied Cognitive Psychology*, 7(6):475–481.

[Schweinberger et al., 1997a] Schweinberger, S. R., Herholz, A., and Sommer, W. (1997a). Recognizing famous voices: Influence of stimulus duration and different types of retrieval cues. *Journal of Speech, Language, and Hearing Research*, 40(2):453–463.

[Schweinberger et al., 1997b] Schweinberger, S. R., Herholz, A., and Stief, V. (1997b). Auditory long term memory: Repetition priming of voice recognition. *The Quarterly Journal of Experimental Psychology: Section A*, 50(3):498–517.

[Schweinberger et al., 2018] Schweinberger, S. R., Zäske, R., Schweinberger, S. R., and Zäske, R. (2018). Perceiving Speaker Identity from the Voice. *The Oxford Handbook of Voice Perception*, (April):538–560.

[Singh and Murry, 1978] Singh, S. and Murry, T. (1978). Multidimensional classification of normal voice qualities. *The Journal of the Acoustical Society of America*, 64(1):81–87.

[Şişman et al., 2017] Şişman, B., Li, H., and Tan, K. C. (2017). Transformation of prosody in voice conversion. In *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1537–1546. IEEE.

[Sisman et al., 2020] Sisman, B., Yamagishi, J., King, S., and Li, H. (2020). An overview of voice conversion and its challenges: From statistical modeling to deep learning. *arXiv preprint arXiv:2008.03648*.

[Stylianou, 2009] Stylianou, Y. (2009). Voice transformation: a survey. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3585–3588. IEEE.

[van Brenk et al., 2009] van Brenk, F., Terband, H., van Lieshout, P., Lowit, A., and Maassen, B. (2009). An analysis of speech rate strategies in ageing. *Interspeech 2009*, pages 792–795.

[Van Dommelen, 1990] Van Dommelen, W. A. (1990). Acoustic parameters in human speaker recognition. *Language and speech*, 33(3):259–272.

[Van Lancker and Kreiman, 1987] Van Lancker, D. and Kreiman, J. (1987). Voice discrimination and recognition are separate abilities. *Neuropsychologia*, 25(5):829–834.

[Van Lancker et al., 1985] Van Lancker, D., Kreiman, J., and Emmorey, K. (1985). Familiar voice recognition: patterns and parameters part i: Recognition of backward voices. *Journal of phonetics*, 13(1):19–38.

[Walden et al., 1978] Walden, B. E., Montgomery, A. A., Gibeily, G. J., Prosek, R. A., and Schwartz, D. M. (1978). Correlates of psychological dimensions in talker similarity. *Journal of Speech and hearing Research*, 21(2):265–275.

[Wester, 2012] Wester, M. (2012). Talker discrimination across languages. *Speech Communication*, 54(6):781–790.

[Wolf, 1972] Wolf, J. J. (1972). Efficient acoustic parameters for speaker recognition. *The Journal of the Acoustical Society of America*, 51(6B):2044–2056.

[Wu and Li, 2014] Wu, Z. and Li, H. (2014). Voice conversion versus speaker verification: an overview. *APSIPA Transactions on Signal and Information Processing*, 3.