BRNO UNIVERSITY OF TECHNOLOGY
VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

FACULTY OF INFORMATION TECHNOLOGY
FAKULTA INFORMAČNÍCH TECHNOLOGIÍ

DEPARTMENT OF COMPUTER GRAPHICS AND MULTIMEDIA
ÚSTAV POČÍTAČOVÉ GRAFIKY A MULTIMÉDIÍ

AGREEMENTS AND DISAGREEMENTS BETWEEN AUTOMATIC AND HUMAN SPEAKER RECOGNITION

SHODY A ROZDÍLY MEZI AUTOMATICKÝM A LIDSKÝM ROZPOZNÁVÁNÍM MLUVČÍHO

BACHELOR'S THESIS
BAKALÁŘSKÁ PRÁCE

AUTHOR **JAKUB VALENTA**
AUTOR PRÁCE

SUPERVISOR **Dr. JOHAN ANDREAS ROHDIN**
VEDOUCÍ PRÁCE

BRNO 2019

Brno University of Technology
Faculty of Information Technology

Department of Computer Graphics and Multimedia (DCGM) Academic year 2018/2019

Bachelor's Thesis Specification

21943

Student:	**Valenta Jakub**
Programme:	Information Technology
Title:	**Agreements and Disagreements between Automatic and Human Speaker Recognition**
Category:	Speech and Natural Language Processing

Assignment:

1. Get acquainted with the basics of speaker recognition (SR).
2. Study basic literature on forensic SR.
3. Prepare data for human and automatic SR, for example from NIST HASR evaluation, perform the human judgements.
4. Get familiar with and automatic SR system, for example from voicebiometry.org, and perform automatic judgements of the same data.
5. Compare the two and analyze agreements and differences, eventually use automatic techniques, such as attention models.
6. Devise at least one strategy to improve automatic SR, implement it and test.
7. Create a short video on your work.

Recommended literature:

- according to supervisor's advice.

Requirements for the first semester:

- Items 1 to 4, start of item 5.

Detailed formal requirements can be found at http://www.fit.vutbr.cz/info/szz/

Supervisor:	**Rohdin Johan A., Dr.**
Head of Department:	Černocký Jan, doc. Dr. Ing.
Beginning of work:	November 1, 2018
Submission deadline:	May 15, 2019
Approval date:	November 2, 2018

Abstract

This thesis deals with the problem of speaker recognition. The term is defined and individual methods related to it are described. The aim of the work is to point out the agreements and disagreements between human and automatic speaker recognition. At the beginning of the thesis, theoretical knowledge from both mentioned areas are described, i.e., what aspects of human speech do people, or automatic system respectively, focus on. Then, several experiments are performed to compare these two methods. These experiments are evaluated in such a way, that it is possible to observe which tasks can be better solved by people, in order to use the knowledge to improve the function of the automated system. At the end of the thesis, such an attempt to improve the automatic system is demonstrated and tested. Testing was successful and higher evaluation accuracy could be observed. Thus, such a result can be used in other research to allow further development in the field of automatic speaker recognition.

Abstrakt

Tato práce se zabývá problémem rozpoznáváním mluvčího. Uvedený pojem je definován a doplněn o jednotlivé metody, které s ním souvisí. Cílem práce je poukázat na shody a rozdíly mezi lidským a automatickým procesem rozpoznávání mluvčího. V úvodu práce jsou popsány teoretické poznatky z obou zmíněných oblastí, tj. na jaké aspekty lidské řeči se zaměřuje člověk, resp. automatický systém. Následně je provedeno několik experimentů, které mají za úkol srovnat tyto dvě metody. Tyto experimenty jsou vyhodnoceny tak, že je možné pozorovat, které testovací úlohy dokáže lépe vyřešit člověk, aby následně bylo možné tyto poznatky použít ke zlepšení funkce automatického systému. V závěru práce je takovýto návrh na zlepšení automatického systému předveden a otestován. Testování proběhlo úspěšně a byla zaznamenána vyšší přesnost při vyhodnocování. Takový výsledek tedy může být užitý v dalších výzkumech a umožnit tak další vývoj v oblasti automatického rozpoznávání mluvčích.

Keywords

recognition, speaker, human, automatic, voice, sound, fonetics, fonology, forensics, system, model, identification, verification, detection, idiosyncratic, analysis

Klíčová slova

rozpoznávání, mluvčí, lidské, automatické, hlas, řeč, fonetika, fonologie, forenzní vědy, systém, model, identifikace, verifikace, detekce, idiosynkratický, analýza

Reference

VALENTA, Jakub. *Agreements and Disagreements between Automatic and Human Speaker Recognition*. Brno, 2019. Bachelor's thesis. Brno University of Technology, Faculty of Information Technology. Supervisor Dr. Johan Andreas Rohdin

Rozšířený abstrakt

S rozvojem počítačových technologií se začaly rozvíjet taktéž systémy, které mají za úkol vykonávat procesy, které do té doby bylo mohli provádět pouze lidé. Mezi tyto procesy se řadí také rozpoznávání mluvčích. Dosud však automatické systémy, určené k řešení tohoto typu problému, nejsou dokonalé a nadále se využívá mimořádných schopností člověka. Cílem této práce je představit způsoby, jakými je proces rozpoznávání mluvčích prováděn lidmi, a naopak automatickými systémy, srovnat tyto dva rozdílné přístupy a navrhnout způsob, jakým by mohlo být dosaženo zlepšení funkce systému.

Problematika rozpoznávání mluvčího je založena na biometrické povaze lidské řeči. Jejím úkolem je najít a zmapovat unikátní znaky řeči tak, aby později mohly být použity pro rozeznání konkrétní osoby. Samotný proces rozpoznávání je postaven na otázce: „Patří tyto dva hlasové projevy stejnému člověku?“ V závislosti na tom, zda tuto otázku řeší člověk nebo počítačový software, je pak tato problematika rozdělena právě na:

- lidské rozpoznávání mluvčího
- automatické rozpoznávání mluvčího

První z uvedených využívá během procesu převážně znalosti vědních oborů fonetiky a fonologie, což umožňuje zachytit nejen informace o zvukových stránkách hlasu, ale také i o významu jednotlivých vyjadřovacích prvků (hlásky, slova, věty). Oproti tomu automatický systém různými způsoby extrahuje z hlasových projevů charakteristické znaky dané akustickým signálem. Obě součásti problematiky však vytváří ze získaných informací profily mluvčích, které následně srovnávají.

Ideálními charakteristickými znaky jsou takové, které mluvčí nemůže úmyslně maskovat a které jsou patrné bez ohledu na okolnosti. Takové znaky se pak nazývají idiosynkratické. Pro lidské rozpoznávání mluvčího jsou za charakteristické prvky považovány především pohlaví, přibližný věk, rychlost mluvy, zvláštní řečové chyby nebo časté opakování parazitních slov. Tyto všechny sice patří mezi fonetické znaky, nevykazují ale dostatečnou idiosynkratičnost. Mnohem větší míru idiosynkracie však vykazují základní hlasová frekvence, vokalické formanty a spektrální sklon. Ty jsou totiž z velké míry závislé na vlastním vokálním traktu člověka.

Automatické rozpoznávání mluvčího naopak využívá prvků z přenášeného hlasového signálu jakožto matematického modelu. Tento řečový signál je nejprve rozdělen do mnoha malých částí – rámců, které jsou následně uloženy do různých modelů. Proces automatického

rozpoznávání je rozdělen na dvě části, kdy v první z nich nejprve dochází k trénování systému trénovací/mi nahrávkou/ami a ve druhé k samotnému hledání podobností a určení skóre. Obě nahrávky však podstupují onu extrakci vlastností.

Pro vytváření modelů mluvčích je možné použít více způsobů, přičemž každý má své pozitiva i negativa. Mezi nejvýznamnější modely patří například Gaussovy modely směsí, lineární deskriptivní analýza (rozšířena také o pravděpodobnostní LDA), i-vektory a x-vektory. Nejpokročilejší z uvedených technologií (x-vektory) se pak věnovala část práce věnovaná experimentům.

Experimenty byly postavené tak, aby muselo být rozpoznání mluvčího provedeno automatickým i lidským způsobem a aby výsledky procesu přinesly srovnání těchto dvou procesů. Byly provedeny dva experimenty, jeden méně a druhý více obsáhlý, přičemž oba přinesly zajímavé poznatky. Těch bylo následně použito při poslední části této bakalářské práce – návrhu na zlepšení funkce automatického systému. Pro experimenty bylo využito

dvou datasetů. První z nich pochází z oficiální NIST 2010 SRE HASR evaluace, druhý je volně dostupný na internetu pod názvem VoxCeleb.

Experimenty ukázaly jistý vliv podobnosti fonémů na výsledek procesu. Tato vlastnost tedy byla zakomponována do procesu extrahování vlastností do x-vektorů s pozitivním výsledkem. Použitím takové změny systému bylo dosaženo nižší hodnoty chybovosti, tudíž se zde nabízí prostor pro další výzkum této oblasti.

Agreements and Disagreements between Automatic and Human Speaker Recognition

Declaration

Hereby I declare that this bachelor's thesis was prepared as an original author's work under the supervision of Dr. Johan Andreas Rohdin. The supplementary information was provided by doc. Dr. Ing. Jan Černocký. All the relevant information sources, which were used during preparation of this thesis, are properly cited and included in the list of references.

. . . .
Jakub Valenta
May 15, 2019

Acknowledgements

Here I would like to thank my supervisor, Dr. Johan Rohdin, for his friendly attitude, time, all the knowledge and help he gave me while writing this bachelor's thesis. My thanks also belong to doc. Jan Černocký, who provided me with a great deal of information, advice and thanks to whom the whole work could be created. I am very pleased that I could work with these people.

Contents

List of Figures

List of Tables

Introduction

From time immemorial, people have been using speech to express their feelings, thoughts, and opinions. That is mostly because it is easy to use spoken speech to communicate and to share any information or thought. Moreover, most of the population understands some sort of language. From a mathematical point of view, speech can be perceived as some kind of signal of different frequencies, which contains and transmits all of this information. There are some differences between the characteristic features of speech spoken by individual speakers. Some of them can be recognized by laics (e.g. language, dialect, emotion etc.), some are more difficult to recognize and require deeper knowledge of the problematics. Two related scientific disciplines – phonetics and phonology have evolved over time. Both of them are dealing with audible expressions of speech.

This knowledge can also be used for the phonetic speaker identification. The characteristic features of speech are examined for different utterances, and analysis of these features provides a possibility for comparison of the speakers. This can be used, for example, as an evidence during criminal proceedings. Unfortunately, the human factor and possible false conclusions of phoneticians and linguists have to be taken into account. With the evolution of technology, there came the possibility of recording the speech. This possibility led to attempts to automate solving these problems. The development of automated speaker recognition tools, which can be used both for the above-mentioned purposes (speaker recognition), such as voice-controlled devices, begins. A well-functioning speech recognition tool could also serve as real-time interpreters, which would definitely destroy language barriers.

The problem of automated systems paradoxically is that they are not people. This can be understood as that the automatic system does not understand the information which is transmitted by speech. Automatic systems work with the mathematical principle of speech – signals. For an automated speaker recognition system, only the characteristic features of each speaker are important, while all the other information transmitted in the signal (e.g. the meaning of words, ambient noise, transmission channel) affect the accuracy and the result in a bad way.

This work will describe the process of human speaker recognition, including the aspects, which a person focuses on during the process, and key identification features of speakers. Also, the automatic speaker recognition process, including the techniques used to create speakers' profile, and the results of these operations will be analyzed in the paper. The final part of the work is devoted to experiments, which demonstrate how the whole process of speaker recognition works and how people can co-operate with automatic systems to get better efficiency and accuracy. There was also an attempt to improve the automatic system.

Chapter 1

Speaker Recognition

Speech is undoubtedly the most common and natural way of communication between people. Basically, it is an audio signal that transmit a great deal of various data, including both language content and other relevant linguistic information as well as information about the speaker himself/herself. The latter includes, for example, information about the speaker's voice apparatus, his/her health, age, speech defects, as well as specific geographic features. People have the ability to distinguish the information and, despite the complexity involved, understand the content of the signal. Such ability to distinguish individual components of a speech signal is called *speech recognition*, i.e. recognition of what has been said. However, despite the area of speech recognition is scientifically researched on a large scale, this thesis will deal with another area.

In practice, it is not always necessary to know the content of the speech, i.e. the communicated idea. In a large number of cases, it is instead necessary to focus on identifying the particular speaker to which the speech belongs. This area is called „*speaker recognition*" and it rather focuses on the phonetic aspects of speech. As stated in [1], phonetics is a science dealing with the audio aspect of human speech. It deals with all aspects of speech and looks on the speech from the following points of view:

- speech formation in the speech apparatus,
- physical characteristics of the signal,
- speech signal perception,
- linguistic system (part of phonology).

This chapter will therefore deal with a more detailed definition of terms in the sphere of speaker recognition, a description of the basic tasks in this area, and how speakers can be recognized.

1.1 Definition of Speaker Recognition

Speaker recognition („SR") is (according to [25]) the intersection of two areas of science. The first is the natural language technologies, and the second is biometrics. It is defined as a method of identifying and confirming the identity of a person based on his voice. The question here is, if there are any features of speech, unique for every single person, which could be used for distinguishing people from each other. According to [16], there are some variabilities caused by the structure of speech organs, which are different to everyone (see 2.1). Even despite the fact, that people are able to deform their speech, there are always some information which can be used as an identifier.

During the process of SR, we are supposed to decide, whether the two utterances, one of which belongs to an unknown person, were both uttered by the same speaker. According to [23], there are two types of speaker recognition:

- **Text dependent** – tested person is asked to say a fixed phrase. The phrase is usually the same as the one in a disputed utterance[1], but there is no strict rule about the content (therefore it can also be randomly generated).
- **Text independent** – the logical content of both the disputed and the test[2] utterances does not matter during the verification process.

SR technology can be really helpful in the field of forensics (see 2.2, because is possible to extract information from telephone conversations, including between criminals. Due to this, there has been increasing interest to integrate SR to supplement auditory and semi-automatic analysis methods in recent years. Both the text-independent and text-dependent analysis can be used during the forensic analysis. The decision of which one of these types will be used depends on specific situation. Also the analysis can be performed in two different way, which are mentioned below.

1.1.1 Speaker Recognition Methods

The process of recognizing a speaker is not a simple matter. In the past, this could be only done by people, based on their experience and available knowledge of phonetics and phonology. With today's automation, however, semi-automatic or fully automatic systems designed for this purpose are also being used. So, the SR can now be divided into the *human speaker recognition* and the *automatic speaker recognition*. As the name implies, the human speaker recognition is a human-performed process of the SR (in practice, for example, by a forensic phonetician/phonologist); on the contrary, the automatic speaker recognition is performed by a computer system, virtually without human intervention (in practice, for example, in security devices). It is also common that both the methods are combined to obtain more accurate results and to eliminate errors.

The following paragraphs describe these two methods of speaker recognition in details.

Human Speaker Recognition

Human Speaker Recognition (hereinafter referred to as „HSR“) uses, as has been mentioned several times, scientific knowledge in the fields of phonetics and phonology. According to [26], HSR is based on two basic types of analysis:

- acoustic,
- listening.

Both of them have their advantages and/or disadvantages. The listening (also called prosodic) analysis captures those similarities, or differences, respectively, between the disputed and the test utterance concerning the language properties of the speaker in question.

[1] An utterance, which is available before testing and which we want to recognize speaker on.

[2] An utterance obtained from the tested person, which is later compared with the disputed utterance, in order to match the speaker in it.

It is primarily a dialect, an accent itself, speech defects or an idiosyncratic way of creating some of the speech sounds. On the other hand, the acoustic analysis distinguishes information unattainable by listening. These are, for example, the base frequency value, speech rate, values of formant frequencies, etc. Both of these analyzes usually take place simultaneously. During the listening analysis interesting elements of speech are selected, which are then processed by the acoustic analysis.

Since this is a human-performed process, phonological phenomena can also be observed, such as excessive use of some phrases, addendums, filler words, etc. This possibility is still an advantage of the analysis carried out by people. More details about the HSR and what is being examined will be described in the Chapter 2.

Automatic Speaker Recognition

In contrast to the HSR, in today's world of speaker recognition, it is common to use automated systems. Automatic Speaker Recognition (hereinafter referred to as „ASR"). This is a computer software that initially creates a profile for a given speaker (based on an enrollment utterance) in a variety of ways, by extracting important information from the speech signal and storing it as a mathematical (e.g. probability) model into its database. In the case of both speaker verification and identification (explained and defined in 1.2), the same process is then performed also for a test speech utterance. The two models thus created can be then compared and it can be determined how much these two models overlap. The result is then usually interpreted as a value from a given interval, which expresses the certainty of the outcome.

It should be borne in mind that the computer program works deterministically. The automatic system is not aware of the meaning of words, context, dialect, morphology etc. It is only referred to the mathematical interpretation of the speech signal and the information that can be read from such signal. It means that the computer software can work with various frequencies, tones and noises, using mathematical operations. So, we can say that the ASR will be just as perfect as the programmer of the given automatic system and the mathematical procedures for processing extracted information are. However, this perfection (same as for the HSR) never reaches 100%, because natural speech is a random phenomenon to some extent.

In practice, we often encounter the ASR and the HSR interconnected, when both the methods complement each other. The ASR can serve to verify judgment based on the HSR, or vice versa, the judgment of the automatic system can be verified and refined by a „human view" of the matter. For more details about the ASR, see the Chapter 3.

1.2 Tasks of Speaker Recognition

Scientific disciplines in the area of speaker recognition can be practiced in many industries. Since the interest in these sciences has begun to rise, their rapid development occurred and the areas of their possible use have been greatly expanded. They can be used both in the commercial sphere as well as in corporate and legal spheres.

Two basic types of tasks are distinguished in the field of speaker recognition. These are „*speaker verification*" and „*speaker identification*". Both of these problems consist of verifying a person based on a sample of his/her speech. Depending on the type of problem, different approaches are followed and different aspects are verified. These questions will be described in the following subsections.

1.2.1 Definition of Speaker Identification

The task, referred to as *speaker identification*, solves a problem in which the speaker's identity is found. The speaker's identity means a personal speech model of a particular person. The task of identifying is to compare the spoken test utterance of a speaker, whose identity needs to be confirmed, with the speech models of a whole group of speakers (see figure 1.1). The goal is then to correctly determine which speaker from this group is speaking in the test utterance. In everyday life, the role of identification is quite common. If a person hears the voice of someone he knows, he immediately tries to determine (i.e. to identify) the speaker the voice belongs to.

From mathematical point of view, it is a search for matching in the system „*1* to *N*“, where N is the number of possible speakers (for HSR) or the number of speaker models in the database (for ASR). From this it can be read out that the success of the correct solution depends directly on the size of the set N. Although biometrically no two speech models are the same, speech as such is a sort of random phenomenon, and if multiple speakers in the set N have similar voices, a certain degree of misclassification can occur. Also, the fact, that the speech model of the tested speaker may not be contained in N, can cause a complication.

The solution to this task has simplified, for example, the identity verification process in security devices (ASR), in which the speaker identification is used to narrow the number of possible speakers. This narrowed selection is then passed to resolve the task described below.

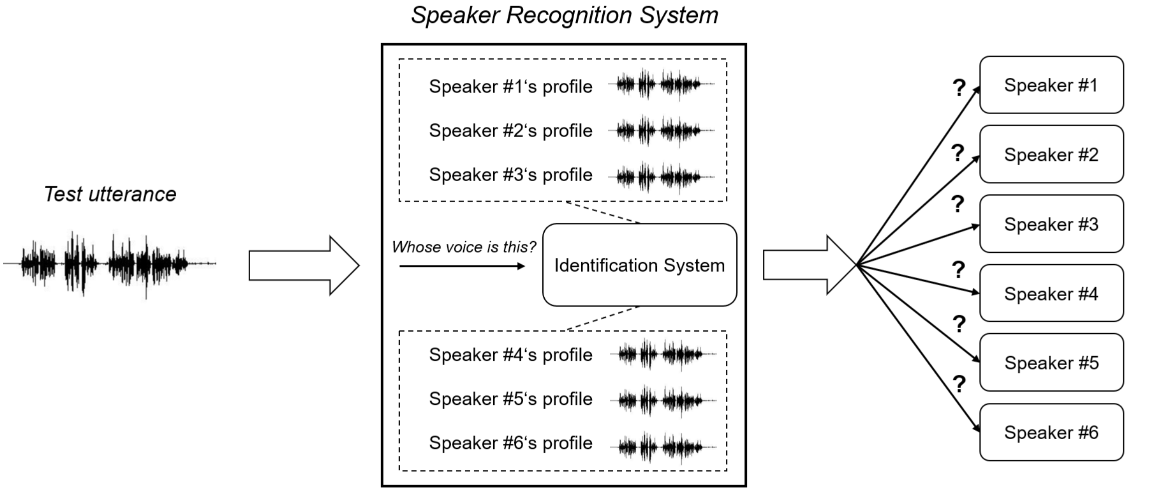

Figure 1.1: Process of speaker identification

1.2.2 Definition of Speaker Verification

Another basic task in speaker recognition is *speaker verification*. The goal in solving this task is to determine whether two speech utterances belong to the same person (see figure 1.2). Thus, it results in one of the following two statements:

- true – so both speeches come from the same speaker, or
- false – the tested speech does not belong to the person in the reference sample.

In some sources, for obvious reasons, the term „*detection*“ is also used for verification.

Solving this problem is also common in practical life. This task must also be addressed by, for example, security systems. The common situation is, that the person who wants to use the system do the „enrollment“ by submiting his/her referential speech model (e.g. by reading his/her ID card), and then later for gaining access, he/she provides a test sample of his/her speech. Such sample is then compared with the referential model (see [5]). Also, in the area of forensic applications, it is easier to identify a speaker from a smaller group of potential suspects than identifying one of many. This is possible because mathematically this task matches the „*1* to *1*“ model. From this it can be concluded that the success rate does not depend on the number of reference speakers.

This thesis, in its other parts, especially in the Chapter 4, will deal mainly with the task of speaker verification.

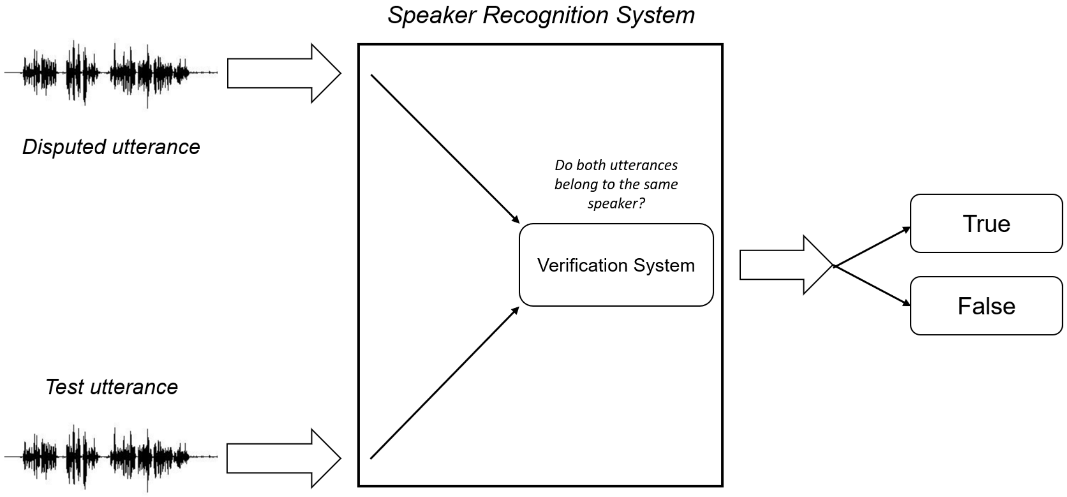
Speaker Recognition System
Disputed utterance
Do both utterances belong to the same speaker?
Verification System
True
False
Test utterance

Figure 1.2: Process of speaker verification

Chapter 2

Human Speaker Recognition

At the beginning of the scientific field of speaker recognition, there were people. People has a unique skill to analyze the speech of others, to recognize individual words and to capture the meaning communicated through speech. Although people use this ability every day, it gained the real meaning in practice with the beginning of the 20th century. After the Florida court accepted the testimony of a raped woman in 1907 that identified the assailant by voice, the speech started to be interesting for forensic sciences [9]. Forensic Phonetics is currently the most common practical use of this human ability and an entire section below in this chapter will be devoted to it (see 2.2).

Human Speaker Recognition (HSR) is based on the general concept of Speaker Recognition (SR). From the record of human speech, as much of the characteristic and other unusual characteristics are extracted during the listening analysis, which then describe the speaker's profile. Such speaker's profile is then compared to a test record of a person whose identity is supposed to be confirmed. Typical speaker's characteristics include:

- sex and approximate age,
- dialect,
- tempo of the speech,
- unusual pronunciation of some sounds,
- speech defects, or
- frequent use of some specific phrases.

HSR tries to find biometric characters in human speech, similar to fingerprints, cornea, or DNA. The forensic area of SR has found some elements of idiosyncrasy among individual speakers from the acoustic analysis:

- fundamental speech frequency,
- vocal formants,
- spectral slope.

These are partly included in the automatic speech analysis, because human senses are not able to precisely determine and differentiate them. Forensic analysis, however, would not have worked well without the help of automated systems today.

2.1 Phonetic analysis of human speech

In the field of HSR, phonetic sciences are used predominantly. Phonetics is a science, which examines sounds, especially the sound aspect of human speech, the way of its creation and perception. It examines speech as a sound signal and performs its analysis considering, how this signal is treated by a person in his/her communication. Phonetics deals with the construction of human vocal tract and its individual components, with elements of human speech and unique articulation characteristics of individual speakers. This chapter describes the different areas in which phonetics is involved, and specific indicators that distinguish the speaker from one another. The information mentioned in this section is taken from *Marie Svobodová's lecture* and from the book *Fonetika a fonologie češtiny* [21].

2.1.1 Vocal tract

Human speech is formed in the so-called vocal (phonational) system of a person. This phonation system is located inside the larynx, which is firmly attached to the trachea (respiratory tract). The larynx, as such, is composed of cartilages that are movably bound together by the ligaments. The individual components of the larynx are:

- cricoid cartilage,
- thyroid cartilage,
- epiglottis,
- pair of arytenoid cartilages,
- vocal cords.

The entire larynx structure can be seen in the figure 2.1. All parts of the larynx are them, either actively or passively, involved in the voice creation.

Speech creation is primarily formed by the pair of arytenoid cartilages and vocal cords. When breathing, the air passes through the open glottis and the vocal muscles are set loose. However, if there is a pulse from the brain that a voice should be created, the arytenoid cartilages will come close to each other and the vocal muscles will stretch. When the air passes, the pressure affecting the closed vocal cords causes the vocal cords to vibrate and to create a tone of a certain frequency. However, other aspects, especially the articulation mechanism, also affect the final voice.

By the term „articulation system“, a set of organs in the space above the larynx, involved in speech creation, is meant. This system includes:

- throat cavity,
- oral cavity,
- nasal cavity,
- lips,
- teeth,
- tongue,
- hard and soft palate.

These cavities and organs affect the base vocal tone as a filter or obstacle, thus helping to create different sounds. Because a large part of the articulation device is fixed and unique to each speaker, it is necessary to use the movable parts of the vocal tract to create speech. This is probably the main cause of the idiosyncratic features of human speech.

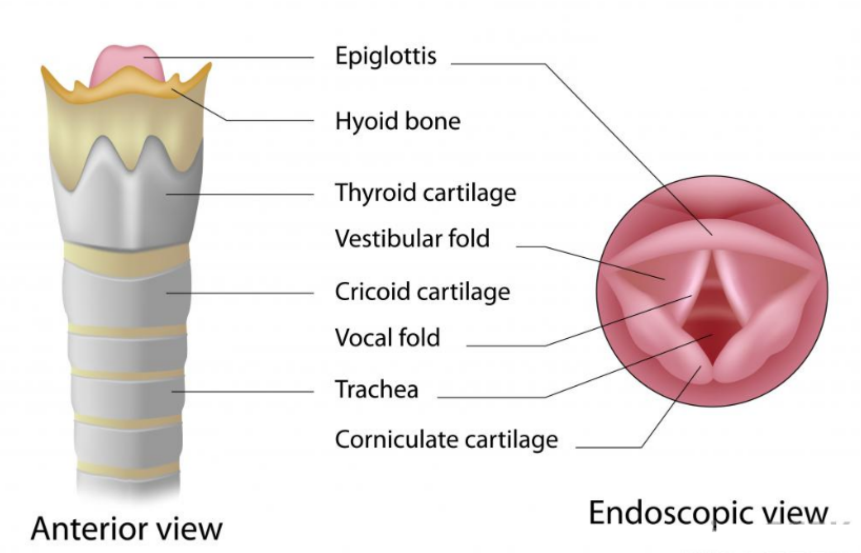

Figure 2.1: Structure of larynx (source: www.wisegeek.com).

2.1.2 Units of speech

Human speech can be divided into smaller parts, that can be pronounced, and their connection produces the final speech. Nowadays, the phonetics work with three units of human speech, depending on what aspects are being explored. These units are considered to be:

- a sound,
- a phoneme,
- a morphoneme.

All these units are described in the following paragraphs.

Sounds

The first elemental speech segment is a *sound*. The „sound“ is a term, which indicates of such a segment of human speech, which is separable as an elemental part. It is therefore possible to recognize and distinguish it from the others. There are 42 voices in Czech, which can be divided into vocals and consonants, according to the way in which they are created.

Phonemes

The second of the elementary speech segments is a *phoneme*. This is a more general unit than *sound*, it looks only for some of the acoustic features. The phoneme was defined, because individual sounds can be pronounced in different ways, therefore it has a distinctive function in language. A phoneme is referred to as such a acoustic segment, that is so obligatory in that language, that its change would also change the word's meaning. In [21], Palková writes, that a phoneme is a set of all similar tones (allophones), which in the language are a particular form of one voice and have the same function. Phonemes can be distinguished from one another by certain essential features, which at the same time distinguish the meaning of words. The science field of phonology deals with phonemes, their function, and a description of the differences between them.

Morphonemes

The last segment of speech is a *morphoneme*. Again, it is a generalization of the previous unit. Morphoneme is such a set of phonemes, which are most commonly alternated within the morpheme, while deriving words. However, this change does not lead to a change of word's meaning.

2.1.3 Idiosyncratic features of speech

Every person transmits certain information about himself/herself in his/her speech. Phonetics try to extract this information from the speech during the listening analysis. These information are such elements that are unusual or unique in speech, or elements perceived as speech disorders. Elements that are recognizable and which help to distinguish speakers sufficiently are for example:

- dialect,
- speech tempo (amount of sounds in time),
- shortening/lengthening of vowels,
- hypercorrect pronunciation,
- presence of speech disorders (dyslexia, voice „nodules“, stuttering...)

Of course, there are much more characteristics. During the HSR, people can also examine the phonological aspect of speech and draw attention, for example, to the use of unusual word verbs, word repetition etc.

2.2 Forensic speaker analysis

Forensic sciences are one of the most common HSR applications. Forensic analysis is the process of person identification depending on an evidence left at the crime scene. It uses those human characteristics that have a certain uniqueness and thus undoubtedly identify a single person. Forensic sciences include, for example:

- dactyloscopy,
- ballistics,
- psychology,
- genetics,
- linguistics,
- phonetics, etc.

This chapter deals only with forensic phonetics, as a science used to identify the criminal based on his/her speech. Since phonetics assume that each person's vocal tract is unique and that the vocal tract affects the quality of speech, human speech is then considered a biometric characteristic in forensic phonetics.

Forensic speech analysis involves both the listening and the acoustic parts. In the listening part, elements mentioned in the previous chapter, are examined. This part only describes the elements examined in the acoustic part.

2.2.1 Fundamental frequency

The first of the acoustic parameters of human speech, which tend to be unique for different speakers, is the *fundamental frequency.* According to Skarnitzl [26], it is an acoustic correlation of cords' frequencies during pronunciation of voiced sounds. In the literature, the fundamental frequency is referred to as Fo, F_0 or f_0. The fundamental frequency is more useful in forensic identification than, for example, the voice pitch. There are several factors affecting the voice pitch. In addition to the fundamental frequency, these are, for example, also the vocal formants (see below). If we are able to separate the fundamental frequency from the speech, we will extend a significant element of idiosyncrasy. The benefits of F_0 are, for example:

- easy to extract even from short samples,
- sufficiently resistant to distortion,
- high variable among speakers.

However, the fundamental frequency also shows one the significant drawback – high individual variability.

The fundamental frequency is one of the characteristic features of each speaker's speech, so there is an evident effort to calculate a certain mean value, which would distinguish the speaker from others more accurately. This value is called *speaking fundamental frequency* (SFF) and is usually determined as an arithmetic mean, median, or modus. Together with the SFF, its variability also show some uniqueness among the speakers (variability – the

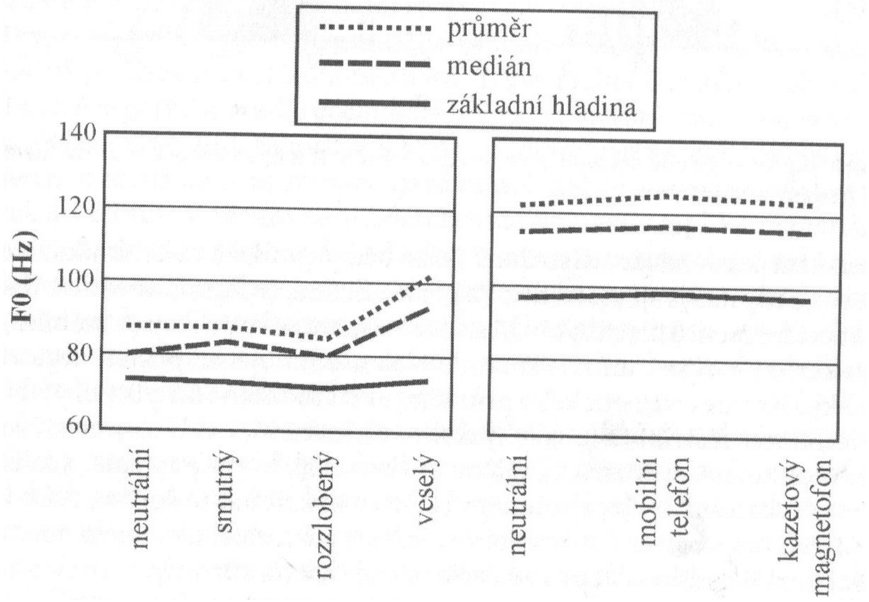

Figure 2.2: Variability of mean, median and baseline depending on different emotional stages (Source: Skarnitzl [26]).

deviation from the average value). Because the fundamental frequency is one of the most researched parameters, there are numerous national statistics for it. These make it easier to work with speech in forensic phonetics.

Due to the variability of F_0 within a single speaker, there came a search for its parameter, whose own variability would be minimal in different situations. According to [29], there is a certain *carrier frequency* that can be exactly this type of parameter. The carrier frequency of F_0 represents the neutral frequency of vocal cords to which the speaker returns after terminating his speech (however emotionally influenced). In the publication [14], a more accurate calculation of F_0 is mentioned, generally as a frequency that lies on the percentile of 7.64 in the set of fundamental frequencies. This was named *„the baseline“*. Compared to the average and the median, the mean value of the baseline values changes the least in most situations (see the figure 2.2.

2.2.2 Vocal formants

Next parameters that can be obtained by acoustic analysis of human speech and which are investigated by forensic phonetics are *vocal formants*. Vocal formants are defined as the *resonance frequencies* of the vocal tract when excised by the cords' signal. Their values vary depending on the spoken sounds. The quality of pronounced vowels is most determined by the first two formants, F_1 and F_2. The quality of is also (more or less) related to F_3. Formats F_4 and F_5 no longer determine the quality of spoken sounds, but are more dependent on the specific speaker (see [26]). Unfortunately, these values in the speech signal often do not occur or are very poorly recognizable.

Format F_1, also called *the throat formant*, along with the F_2 formant (*oral formant*, are able to change their values depending on how the vocal tract is deformed. Therefore, they do not carry much of idiosyncratic information themselves. Formant F_3 (also *nasal formant*), however, cannot be deformed as much as the two previous formants. It is therefore considered to be the main point of recognition of the speaker. Using all three formats at the same time, we can (again according to Skarnitzl) achieve a recognition accuracy of up to 95

Formants are measured at the *phonetic target of a vowel*, that is, where the vocal formants have a steady course. Their statistical values for czech language are listed in the table 2.1. However, according to Skarnitzl, to achieve higher accuracy of identification, the so-called *formant trajectory* is used. These consider not only the value of the formant in the phonetic target but also in several other points during the phoning (Skarnitzl sais another 4-9 points). Obviously, much more information about the speaker is captured by dynamic measurement.

Table 2.1: F_1, F_2 and F_3 formant value ranges for czech language according to Palková (for source see [21]).

vowel	F_1 [Hz]	F_2 [Hz]	F_3 [Hz]
i, í	300-500	2000-2800	2300-3500
e, é	480-700	1560-2100	2000-3000
a, á	700-1100	1100-1500	1500-3000
o, ó	500-700	850-1200	1500-3000
u, ú	300-500	600-1000	1900-2900

Formants undoubtedly exhibit idiosyncratic properties. They describe both the physiological properties of the vocal tract and also the speaker's articulation habits. This makes them very important parameters used in forensic practice, despite their certain instability (because of the voice deformations).

2.2.3 Spectral slope

Another important parameter studied in the field of HSR is the *spectral slope* of the speech signal. Spectral slope is, according to [30], meant to be a phenomenon resulting from the combination of the cords' signal and the filtering characteristics of the vocal tract (thus largely dependent on the speaker's vocal tract). It is reflected by the decrease of the amplitudes of the individual frequency components at an increasing frequency. For SR, it is useful because of its dependence on some unique speech properties, such as voice quality. As shown in [27], during phoning, the spectral slope of the cords' signal is approximately -12 dB per octave. However, when the speech spreads to the environment, the lips act as a filter and the resulting spectral slope of speech is around -6 dB per octave.

The spectral slope is generally considered to be an acoustic correlation of the voice color (see [26]). By this, the typical voice coloring and other articulation characteristics that a speaker uses to transmit speech information, are meant. In the case of a flatter speech spectrum, i.e. higher energy at higher frequencies, the speech signal shows a lower spectral slope, and the speaker's voice appears to be *strident* or *tense*. The lower spectral slope can be also seen if the speaker increased his/her speaking effort. This is directly related to the increase in volume, which results in higher energy increases. On the contrary, the highest spectral gradient can be observed in a spectrum of *calm*, *fine* or *silent* voice. Thus, the emotion also has an influence on the spectrum of human speech, as Banse and Scherer showed in [3]. This idea is expanded in the article with reference to the fact, that a person is able to recognize speaker's emotions from his voice, therefore this information must somewhere be encoded in a speech signal.

Spectral slope indicators are measured from two points of view. A long-term view, called LTAS, provides information about the spectral distribution of energy over a longer period of time. By including a longer time span, averaging short-term changes, that are not important for speaker identification, is achieved. The resulting spectrum is not affected by segmental differences. LTAS can be used to distinguish female and male voices, speaker's approximate age and voice training. On the other hand, the short-term point of view usually focuses on one sound, eventually on one its part. Short-term indicators describe, for example, a speaker's accent or a vocal cords' setting. This makes them interesting for forensic applications, as they also carry unique speaker information.

Chapter 3

Automatic Speaker Recognition

As has been mentioned several times, the development of technology brings requests for creation of an automated speaker recognition (ASR) system. Based on mathematical relationships, the automated system should classify individual speeches without difficulty and help in decision-making on identification and verification issues. The ideal system is one that can handle the tasks shown in Section 1.2 with 100 % success rate. However, creating such a system is not a simple matter, and state-of-the-art systems still make many mistakes.

The ASR is intended for removing the humans from the speaker recognition process, finding unique elements in each speaker's speech, and thus determining the whole process. Also, for this reason, the ASR system works differently than the HSR.

In this part of the thesis, the details and course of the ASR process will be described and the most commonly used procedures for creating personality speaker models will be detailed. This will also explain the principle of the ASR system function.

3.1 ASR process overview

Since the ASR is a subset of more general speaker recognition issues, we also can address two basic tasks – verification and identification. Both these tasks are solved, researched and refined in common practice. However, this chapter will only deal with the role of verification, because its concept is more descriptive for the ASR system description, which is, after all, the purpose of this thesis. The following paragraphs are mainly based on the document [17].

A prerequisite of the ASR process is the existence of so-called disputed utterance, which can be as long as several minutes. In applications, it can be, for example, a recording of a phone call of a blackmailer, a phone tapping or other voice recordings of the offender. To make the work with the recording easier, it is then divided into small sections of a length of tens of milliseconds. As mentioned in [8], signal variability in such a short section is minimal, so the voice in these sections can be considered as an almost stationary signal. From each section of the set thus formed, the necessary information characterizing the speaker in question is then extracted. This is done using the various methods described below in this chapter. The final result is the personal model of the speaker, containing a set of parameters that characterize the speaker in question. This phase is called *„system enrollment*“.

The second phase is the speaker verification itself. Another, usually shorter, recording of a speaker, whose identity needs to be verified, is assumed. Next, the same process as

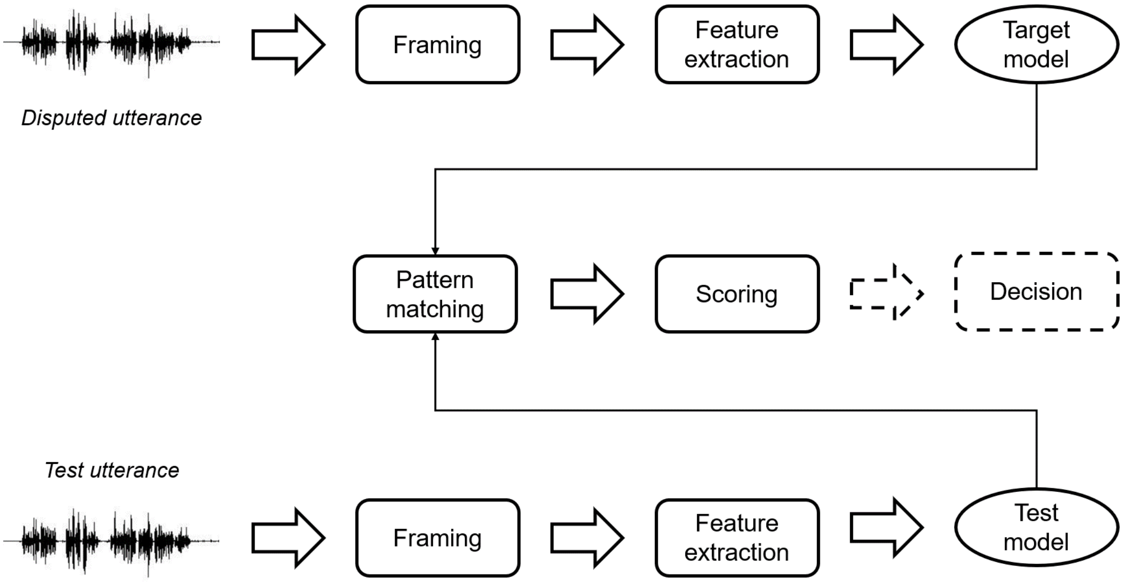

Figure 3.1: The entire process of automatic speaker verification

with the disputed utterance (division, extraction, model creation) is performed. During verification, the model created from the disputed utterance is compared with the model created by extraction from the test utterance. The result is usually a score on a defined scale, indicating how likely it is that the two recordings come from the same speaker. The whole process is illustrated in the figure 3.1

3.2 Closer description of automatic system

This section describes some of the parts of the ASR system, as suggested in the previous section. Specifically, it will be the processes of framing and extraction of features. As stated in [8], a sufficiently short signal segment can be considered stationary. Typically, segments of tens of milliseconds are considered „stationary“, usually 20-25 ms. Such samples are called *frames*. In order to ensure continuity, individual frames are extracted normally every 10 ms and thus overlap each other. The frames can be obtained by using so-called *window-function* (see [2]), usually by its modification *Hamming-window function*, which is bell shaped (as can be seen in the 3.2 figure), eliminating high-frequency distortion that occurs at the edges of the extracted frame. The second part of the process is extraction of features from the frames. With the help of various methods, the necessary information defining the speaker is read out of the frames and subsequently the information is exported into an easy-to-use form. The result of this process is a set of so-called *feature vectors*, each of which represents one speech frame. The final form and dimensionality of these vectors depends on the method used (see below).

For the sake of completeness, I will only mention that the resulting set of feature vectors all together form a personal speech model of the speaker.

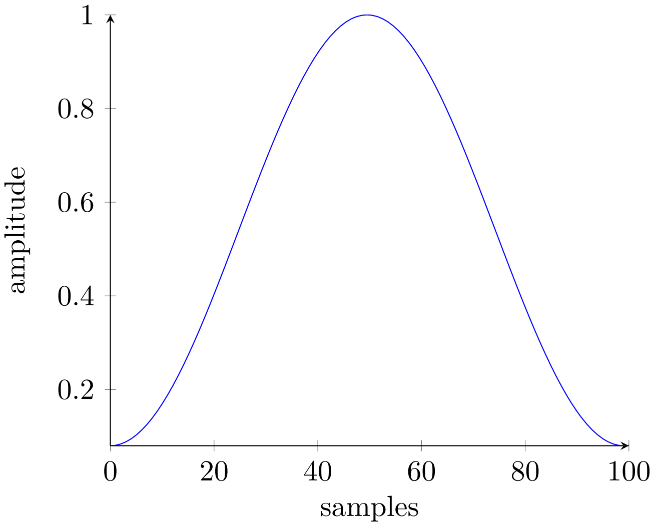

Figure 3.2: Course of the descrete Hamming-window function

3.2.1 Gaussian Mixture Models

Since human speech is, to some extent, a random phenomenon, probability models are usually used to describe it. When we extract data from a speech record during ASR, we get data samples – *feature vectors*. These samples are generated by a speaker's personal speech model and bear signs of a certain probability distribution. The most common and most natural probability model for these purposes is Gaussian distribution (or Normal distribution). The probability density function for GMM is given by the following formula:

$$P(x; \mu_k, \Sigma_k) = \frac{1}{(2\pi)^{\frac{D}{2}} |\Sigma_k|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1} (x-\mu_k)} q$$

where D is the dimensionality of the feature vector.

However, because speech is a very inhomogeneous phenomenon, it is usually better to describe the speaker with a greater number of Gaussian models. Such a set of multiple distributions can be merged into one unit, creating the so-called *Gaussian Mixture Model* („GMM“). According to [8], GMM is defined as a family of mixture models in which the probability density function of each of has Gaussian distribution. The image 3.3 illustrates a two-dimensional GMM that could be used as a speaker model. After extracting the feature vectors, speech samples of a speaker, whose model is illustrated, would show signs that they were generated by this model.

During the ASR, where the speaker's model described by GMM is considered, we look for the value of *likelihood-ratio*. It tells us, with what certainty the speech sample was generated by such model. The Likelihood ratio is calculated from the following relationship:

$$S(X) = \frac{P(X|H_{target})}{P(X|H_{impostor})} \geq \Phi \Rightarrow target$$

where $P(X|H_{target})$ and $P(X|H_{impostor})$ are the likelihoods of X with the assumption that X is uttered (or is not) by the speaker. If the computed score $S(X)$ is greater than the threshold Φ, the conclusion is then that the segment X was uttered by the tested speaker (viz [12]).

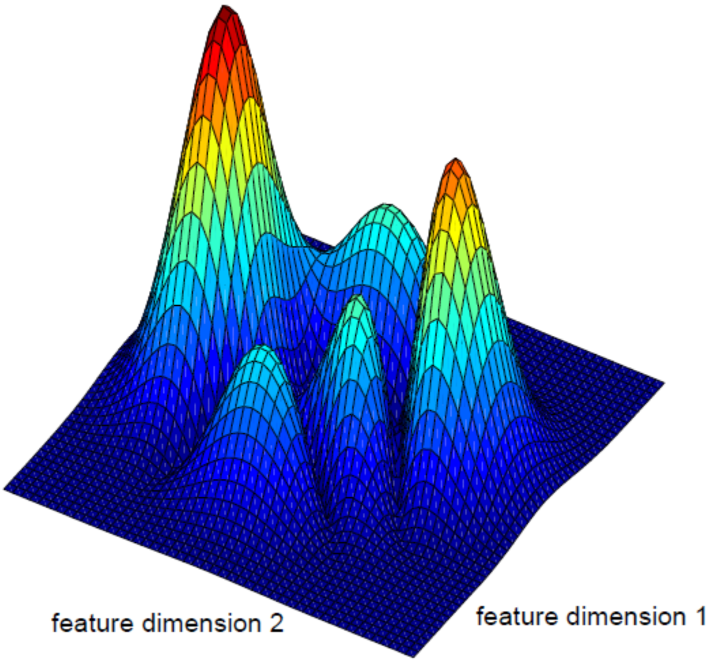

Figure 3.3: Illustration of possible two-dimensional GMM for a speaker (source: Glembek, 2012 [8]).

Linear Discriminant Analysis

The GMM system also brought the basics for so-called *Linear Discriminant Analysis* (LDA). This is a method that significantly improves classification by reducing dimensionality of models (see [24]). The LDA tries to project data in such a direction (to find a dimension) in which there is:

- maximum distance between the mean values of individual classes
- minimum average variance within classes

An illustration of the LDA function can be seen in the figure 3.4. LDA was also the precursor of Probabilistic LDA, which began to be used extensively with the arrival of i-vectors. A special section below will be devoted to the PLDA.

3.2.2 Modeling using i-vectors

Later on, Najim Dehak [7] came with so called i-vectors. The main idea of i-vectors is to provide a simple way of reducing high-dimensional input data to a low-dimensional feature vector, while retaining most of the relevant information. According to Dehak's work, the utterance supervector M of utterance r and of speaker s can be modelled as:

$$M = \mu + \mathbf{T}x_{r,s}$$

In this model, M is the supervector; μ stands for a supervector, which is independent on the speaker and the utterance; $\mathbf{T}$ is a rectangular, low rank matrix representing the variation

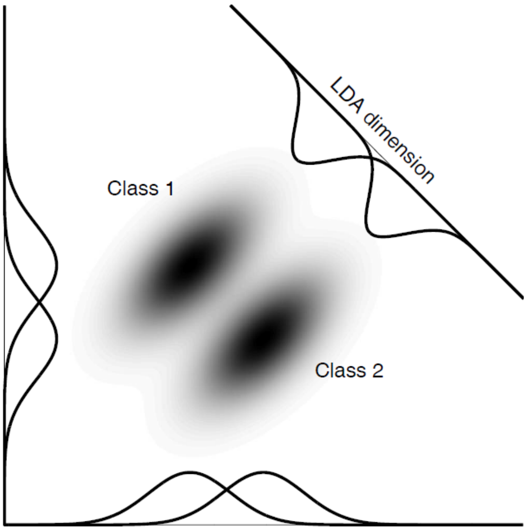

Figure 3.4: Illustration of the LDA function on two-dimensional GMM (source: https://bit.ly/2MyK2e4).

across a large collection of development data; and finally, $x_{r,s}$ is the *i-vector*, which is randomly distributed with Gaussian distribution. The $\mathbf{T}$ matrix is also sometimes reffered to as „total variability space“.

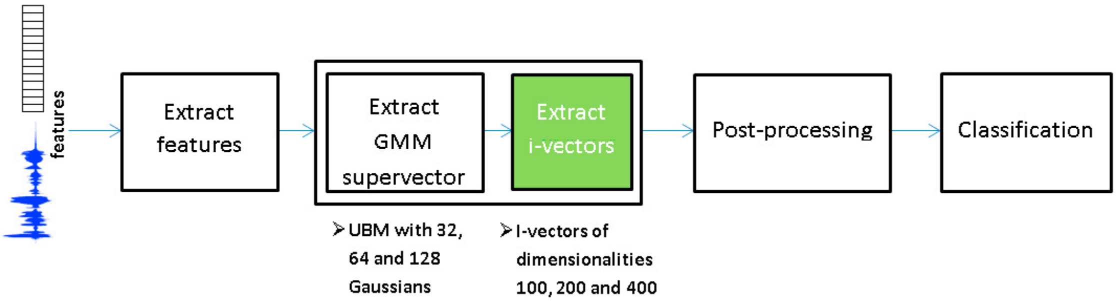

Figure 3.5: Diagram showing the entire utterance classification process using i-vectors (source: [10]).

Figure 3.5 shows the process of i-vector modeling. The process in the figure is described in [10]. First there is feature extraction, which means extracting energy features of the speech signal (using Hamming window function) every few miliseconds. Second, it is the i-vector extraction. For each utterance, the corresponding feature sequence is converted to an i-vector using an extractor based on a GMM. Before the scoring itself, the i-vector system has to perform a post-processing. Basically, the process involves some channel-compensation, so the data obtained from different channels can be properly scored; as well as Probability LDA which will be described below.

3.2.3 Modeling using x-vectors

Another way of modeling, that is nowadays really expanded to create speaker models, is x-vector modeling. It is based on the principle of deep neural networks (DNN). These DNNs are trained to differentiate speakers by mapping variable-length utterances to fixed-dimensional vectors. It takes the frame-level features of an utterance and directly produces an utterance level representation of it – embedding, also referred to as *x-vector*. To perform speaker verification using x-vectors, these embeddings are extracted and used in, for example, PLDA. Such system is powerful as it have been proven (see [28] and [20]) competitive for both short and long utterances in text-independent speaker verification .

That is an advantage of x-vectors. DNN's performance appears to be highly scalable with the amount of training data. This means, that x-vectors are able to use large-scale training data sets better than i-vectors. It was found out that although large-scale data behave beneficially in the PLDA classifier, they may be a big problem to work with i-vector extractor. This is the reason to use x-vectors as the DNN work efficiently with data augmentation, due to their supervised training. As a result, the x-vectors achieve really good performance during the evaluation.

3.3 Scoring

After the ASR system has been trained with the training utterance, the speaker model has been created and the test utterance has been analyzed, the system needs to calculate the output. This output is calculated as a value, that represents the likelihood of how much is the system „sure“ about its decision. The higher the value is, the more it is likely, that both the training and the test utterances belong to the same speaker. The most basic likelihood computation formula [13] can be used as the Bayesian probability formula:

$$\mathrm{LR} = \frac{p(\phi_1, \phi_2 | \mathcal{H}_s)}{p(\phi_1, \phi_2 | \mathcal{H}_d)}$$

where X is the evidence and H_s and H_d are hypotheses that both utterances came from the same or different speakers respectively.

Though, there are numerous ways of how to compute the likelihood ratio. One was mentioned above in the section of PLDA analysis. This way of computation, that is used within the PLDA scoring process, is commonly used as a standard, basically because the PLDA analysis is used very often during the process of ASR.

Probabilistic LDA

Probabilistic LDA (PLDA) is a probabilistic extension of LDA. In speaker recognition, PLDA factorizes the variability of the observation for a specific class (e.g. for one speaker), using an *identity variable*. According to the PLDA model [11], [15], [8], the i-vectors can be decomposed as

$$\phi = \mu + \mathbf{V}y + \mathbf{U}x + \epsilon$$

where $\mathbf{V}$ and $\mathbf{U}$ are low rank matrices which describe the speaker and channel subspace respectively, y and x are variables representing the speaker and channel, and finally ϵ stands for the residual noise and is normally distributed (Gaussian with zero mean and diagonal covariance). The $\mathbf{V}$ and $\mathbf{U}$ matrices help to maximize the likelihood of the observed i-vectors assuming that i-vectors from the same speaker share the same speaker factor.

Likelihood ratio is then calculated for the i-vector model processed with the PLDA. It is performed for the two i-vectors extracted from training and test utterances. According to [4], the likelihood ratio is defined by the following formulas:

$$\begin{aligned} LLR &= \log \frac{p(\phi_1, \phi_2 | \mathcal{H}_s)}{p(\phi_1, \phi_2 | \mathcal{H}_d)} \\ &= \log \frac{\int p(\phi_1 | \mathbf{y}) p(\phi_2 | \mathbf{y}) p(\mathbf{y}) d\mathbf{y}}{p(\phi_1) p(\phi_2)} \\ &= \phi_1^T \Lambda \phi_2 + \phi_2^T \Lambda \phi_1 + \phi_1^T \Gamma \phi_1 + \phi_2^T \Gamma \phi_2 + (\phi_1 + \phi_2)^T \mathbf{c} + k, \end{aligned}$$

where ϕ_1 and ϕ_2 are i-vectors, $\mathbf{y}$ is vector representing a speaker and calculations of $\Lambda, \Gamma, \mathbf{c}$ and k are given by mentioned paper [4]

Chapter 4

Prelimary experiment

To get familiar with the process of speaker recognition itself, in accordance with the assignment, both the human and automatic speaker recognitions were performed. I obtained a pack of official data used in NIST 2010 SRE HASR evaluation [18], which contained 30 utterances – 15 for enrollment, 15 for testing. Details of each process performed are mentioned below. My personal outputs were then compared with the official results.

Based on the outcome of this prelimary experiment and analysis in this Chapter, we will design a larger experiment in Chapter 5

4.1 Experiment with HSR

Not looking for the official reference results, I performed a judgement with all 15 pairs of utterances. All the samples were opened in a open-source software Audacity (source: https://www.audacityteam.org/). As a strategy I chose to listen to the whole enrollment utterance, looking for some specific words used by the speaker, tempo and melody of the speech, and also other prosodic aspects. Then I listened to the test utterance, usually to the longest speech segments, trying to catch the same speech features as I did for the enrollment utterance. I then performed a judgement, scoring the pair of utterances with a value from an interval of $[-5, 5]$, where:

- +5 means that I was 100 % sure that both utterances belong to the same speaker,
- –5 means that I was 100 % sure that each utterance belong to a different speaker,
- 0 means that I was not sure at all about given utterances.

While listening, I had to adjust the volume of utterances. Enrollment utterance was usually very quiet and I could not hear the speaker properly. The test utterance, on the other hand, was usually very loud. Some enrollment utterances were also really noisy and sometimes incomprehensible. I always managed to perform the judgement, however. My personal results were added to the table 4.1.

4.2 Experiment with ASR

For the second part of my personal experiment, this time with the ASR system, the automatic system was needed. I was recommended to use Kaldi, which is an open-source toolkit for speech and speaker recognition, available on http://kaldi-asr.org/. This toolkit was then installed into a virtual computer (using VirtualBox; source: https://www.virtualbox.org/, which I had chosen to be Ubuntu version 18.10 (64bit); source: https://www.osboxes.org/ubuntu/. Experiment was performed with pretrained models available on http://kaldi-asr.org/models/m3

4.2.1 Kaldi toolkit

As I mentioned, I had been recommended to use Kaldi toolkit to perform the experiment with ASR system. Kaldi is an open-source project, that is easy to use, written in C++ language. It can work both with i-vectors and x-vectors, though for my experiment I only decided to work with x-vectors.

Kaldi works as a normal ASR system. According to the script, that is used for the x-vector based speaker recognition, the steps performed are:

1. Prepare enroll and test data (utterances)
2. Extract features and make the Mel Frequency Ceptral Coefficients (MFCCs)
3. Perform the Voice Activity Detection
4. Train the Uniform Background Model
5. Prepare the x-vector extractor
6. Extract the x-vectors
7. Train the PLDA model
8. Score the utterances using the PLDA model.

Scores are then stored in a special text file. These represent the likelihood that the two tested utterances belong to the same speaker as it is usual within an ASR program.

4.2.2 Performance of the experiment

The experiment itself was then performed, as mentioned, with the pack of the official data used during NIST 2010 SRE HASR evaluation. The data pack itself consist of 15 enrollment utterances and 15 test utterances. To perform the evaluation, special script was used; it is attached to the thesis on a special medium. The script is slightly different from the one available in the Kaldi data base.

The process ran as it is mentioned above; the scoring was performed for each utterance against each other one. I then manually extracted only the relevant scores, which were then compared to the official results of NIST HASR evaluation. Each trial was scored with positive or negative score. Positive score means that the *same-speaker* hypothesis is probably true. On the other hand, negative score means the very opposite – probably, the two utterances came from two different speakers. The higher/lower the score is, the more is the system sure about its decision. Values close to zero mean, that the system is not sure at all. For the results, see 4.1.

4.3 Conclusions

After both the HSR and ASR experiments had been performed, my personal results were written into the table taken from [18] document. The final table itself with added results has the following appearance:

Table 4.1: Edited table from [18] comparing personal results of HSR and ASR results with official results from NIST evaluation.

Enroll	Test	SYS1	SYS2	Personal HSR	Personal ASR	Reference
01_ehzum-a	01_trevr-a	-1.27	1.25	3.5	10.06972	*target*
02_eibhn-a	02_thftl-b	0.57	-1.66	0.5	-42.98383	*nontarget*
03_ehymn-a	03_tcrte-b	-5.35	2.14	-3.0	-33.07984	*nontarget*
04_eqbwg-a	04_trify-a	0.34	-3.51	4.0	-44.34171	*nontarget*
05_ehjmk-a	05_trrkn-a	-0.14	-3.11	1.5	-11.75468	*target*
06_ejntb-a	06_tnbcy-b	-6.23	3.90	-2.0	-28.41536	*nontarget*
07_eftsd-a	07_tzocd-b	5.14	2.40	5.0	19.92282	*target*
08_elnji-a	08_tdaxi-a	-1.82	-1.05	-3.5	-32.12069	*nontarget*
09_eltwa-a	09_tpzez-a	-5.40	-7.48	-0.5	-38.31912	*nontarget*
10_euerk-a	10_tzvxk-a	5.12	6.44	4.0	17.95538	*target*
11_erjdr-a	11_tvndw-a	4.00	8.56	-2.5	-2.884458	*nontarget*
12_ezlsi-a	12_thpms-b	-0.32	-2.12	-2.0	-40.67811	*nontarget*
13_ekuzl-a	13_tkvay-a	-6.51	-3.30	-5.0	-37.08413	*nontarget*
14_ebgbw-a	14_tncns-b	6.35	9.27	4.5	38.78508	*target*
15_enjvn-a	15_tepkw-b	-2.49	-0.14	4.0	12.58519	*target*

In the table, columns „SYS1“ and „SYS2“ show values, according to [18], given by automatic systems HASR1-2012 and ABC 2010 respectively. Next two columns, „Personal HSR“ and „Personal ASR“, show output values of the personal experiments. For Personal HSR column, decision threshold is assumed to be 0 (in reality, this is much more complicated matter as the optimal threshold depends on other aspects).

As shown, the automatic system has always given the „correct“ value (according to the reference), except for the 5^{th} trial. This is probably caused by completely different channels in both utterances. For a human, it is possible to adjust the volume, to focus only on parts where only the desired speaker is heard and to find some matching signs of speech (parasite words, tempo). Also, in the 11^{th} pair, the system was not really sure about its judgement (most likely because there are two similarly women speaking in the enrollment utterance). The accuracy of the experiment with ASR comes to be 93 %. The HSR experiment's accuracy is 80 % as three mistakes occured in the 2^{nd} and the 9^{th} trials, which the speakers were questionable in, and in the 4^{th} trial, which had been given almost the highest score. After listening again to last trial mentioned, no change would be made, this trial's utterances still sound as a match.

Chapter 5

Main experiment

In the previous chapters, the theoretical background of the issue of *speaker recognition* was presented. Based on the data from the official NIST evaluation, both ASR and HSR were performed, which demonstrated how these processes work and what is their accuracy. However, to make the results more credible, these processes need to be done on a larger number of different data. This will bring better awareness of the differences between HSR and ASR, which can then be used to improve the efficiency of each method.

Therefore, another aim is to verify whether some of the facts and characteristics of human speech have a greater influence on the result of the speaker recognition process than others. Referring to previous theoretical knowledge, it is believed that these characteristics may be:

- similar/different vocabulary of speakers on both enrollment and test utterances,
- similar/different speech tempo,
- similar/different speech colour,
- similar/different phoneme pronunciation.

Thus, these aspects of speech, contained in utterances, need to be identified and analyzed. If this information is available, it can be used for further experiments to see their potential impact on speaker verification process results.

This chapter further describes the procedure used to obtain the above-mentioned necessary data and then to analyze it to improve ASR functionality.

5.1 VoxCeleb dataset

As mentioned above, it was necessary to use more data to perform further experiments. For other purposes, the *VoxCeleb* [19] dataset has been selected as an online dataset containing over a million utterances of about 7 000 speakers. These recordings have been extracted from videos available online on YouTube, and the speakers are mostly celebrities. The Database consists of two versions – VoxCeleb1 and VoxCeleb2. These two versions are complementary to each other. VoxCeleb 1 test set was used for testing. The Kaldi [22] SITW recipe (`sitw/v2`) was used for training. This recipe uses VoxCeleb 1 and 2 [19], [6] for training as training set. The recipe was used as it is, except for that VoxCeleb 1 test set was not included in the training set.

VoxCeleb1 data set

VoxCeleb1 contains over 100 000 utterances of interviews with 1 251 celebrities. The dataset is almost gender balanced, about 55 % of the speakers are male, the rest 45 % are female. Utterances are spoken by speakers of different ethnic groups, age, accent and nationality. The place, on which recordings were recorded, varies from silent studio interviews, speeches in front of a large audience to interviews in a noisy environment or with background noise. All utterances are at least 4 seconds long. These information are listed in the Table 5.1.

VoxCeleb2 data set

On the other hand, VoxCeleb2 contains over 1 million utterances for over 6 000 celebrities, also fairly gender balanced. The speakers span a wide range of different ethnicities, accents, professions and ages. Audio segments present in the dataset are also degraded with background noises, overlapping speech and varying room acoustics. The development set of VoxCeleb2 has no overlap with the identities in the VoxCeleb1 or SITW datasets.

VoxCeleb data sets are freely available to download on the website. For each speaker, individual utterances are divided into folders.

Table 5.1: Statistics of VoxCeleb1 and VoxCeleb2 datasets. (for source see [19] and [6]).

	VoxCeleb1	**VoxCeleb2**
Number of speakers	1 251	6 112
Number of male speakers	690	3 761
Avg number of videos per speaker	18	25
Avg number of utterances per speaker	116	185
Average length of utterances	8.2 s	7.8 s

5.2 Performance of automatic speech/speaker recognition

Other processes than speaker recognition need to be used to identify these aspects of the utterances. To obtain information about vocabulary, speech rate, and similarity of phonemes, *speech recognition* has to been done. The Kaldi toolkit, previously mentioned in Chapter 4.2.1, was used for both ASR and speech recognition.

5.2.1 Speech recognition process

During speech recognition, the system records changes in speech signal and then, using neural networks, determines which speech characteristics are present in the utterance. This is possible thanks to the previous system training by speech models, which are usually specific to each dataset. The output is a tree structure of all possible sequences (lattices) of words, each path is given score. For the purpose of this work, only the path with highest score from this tree structure was used.

Speech recognition was performed using the official Kaldi [22] Tedlium speech recognition recipe (`s5_r3`). This recipe uses a TDNN based acoustic model with i-vector adaptation and an RNN based language model. Phoneme posteriors were obtained from the decoding lattices using the forward-backward algorithm and then converted to hard labels. Speech recognition was performed on the VoxCeleb data sets. The results were written into a file (table) containing 3 columns, where each row represent one word in a given utterance. The columns are as follows:

1. Word start time
2. Word end time
3. Word

The word start and end times could be displayed in the amount of frames, or in the amount of seconds (amount of 10 ms frames for a word divided by 100 – amount frames per second). With this information, the *speech tempo* can be easily calculated as, for example, the amount of phonemes/words per time unit.

In order to get information about vocabulary similarity in each pair of tested utterances, a simple script in BASH was used. The above-mentioned files, speech recognition results, have been modified to contain only one column with the alphabetically ordered unique occurrences of each word in the utterance. Then, a new output file (table) was created with the `awk` language program. The table has following columns:

1. Name of enrollment utterance
2. Name of test utterance
3. Number of unique words in enrollment utterance
4. Number of unique words in test utterance
5. Number of overlapping words in two utterances
6. Phoneme pronunciation similarity (Jensen-Shannon divergence of phoneme frequencies in the two utterances).

These files thus contain information about the vocabulary similarity of speakers in the individual utterances. The column *Phoneme pronunciation similarity* contains the values of phoneme similarity in the enrollment and test utterances, these were determined by performed speech recognition. Collected information were then shown in graphs, that can be seen below (Graphs 5.1 and 5.2).

5.2.2 Speaker recognition process

Speaker recognition (*speaker verification* specifically) was also performed on the same data used for speech recognition, in order to obtain the score results of an automatic system. This score, along with the reference, was then added to the files mentioned above. Following columns have been added to the table:

7. Likelihood ratio (score)
8. Reference (target/nontarget)

Only rows with trials, in which each enrollment and test utterances contained between 15 to 20 unique words, were then used from the table to narrow the data range. Such a range selection seems to be reasonable, since the utterances therein contained are all about the same length, but not too long at the same time. This results in an effect, that the influence of utterance length on the SR process can be neglected. Thus, the selected data was plotted in the graphs of score – number of overlapping words Graph 5.1) and score – phoneme similarities in the recordings (Graph 5.2).

5.2.3 Results of the processes

The Graph 5.1 shows the score – number of overlapping words dependencies for both target and nontarget trials. In general, it can be seen that the speaker recognition system determines nontarget trials with very good accuracy (most nontarget trials have received very negative scores). In addition, this accuracy seems to be increasing with increasing numbers of overlapping words.

This fact shows, that the success rate of ASR may be influenced, to some extent, by the occurrence of the same words and phrases in both enrollment and test recording. If this statement were true, it would bring the possibility of improving the success rate of the process by selecting test utterances, that contain higher amount of the same words as in the enrollment utterance. This actually is why text-dependent ASR is so effective – evaluated trial utterances contain identical vocabulary (see Section 1.1). On the graph 5.2 it is possible to see the dependence of the score on the similarity of phonemes contained in the trial utterances, again for both types of target and nontarget trials. Before starting the experiment, for target trials, a decreasing trend of score with lower phoneme similarity was expected. This trend can really be seen in the graph. If we consider the lower phoneme similarity value to be more similar, we can see the highest score values are at the lowest values on the horizontal axis. This trend then really decreases with a lower phoneme similarity value.

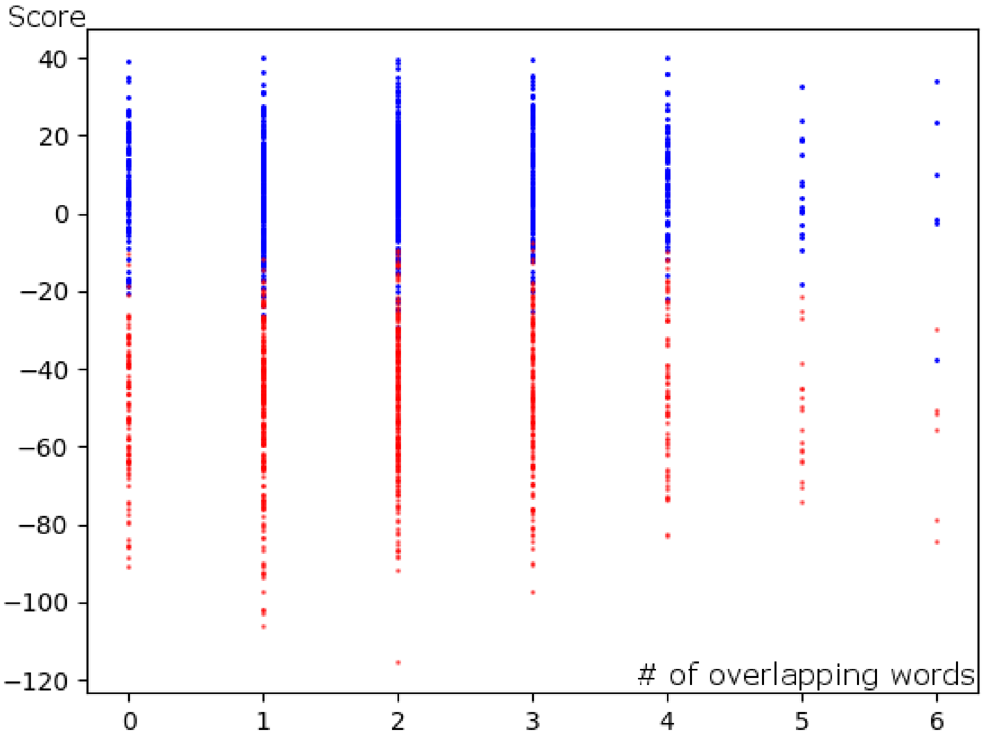

Figure 5.1: Score – Number of overlapping words graph. Blue dots are for target trials, red dots are for nontarget trials according to the reference.

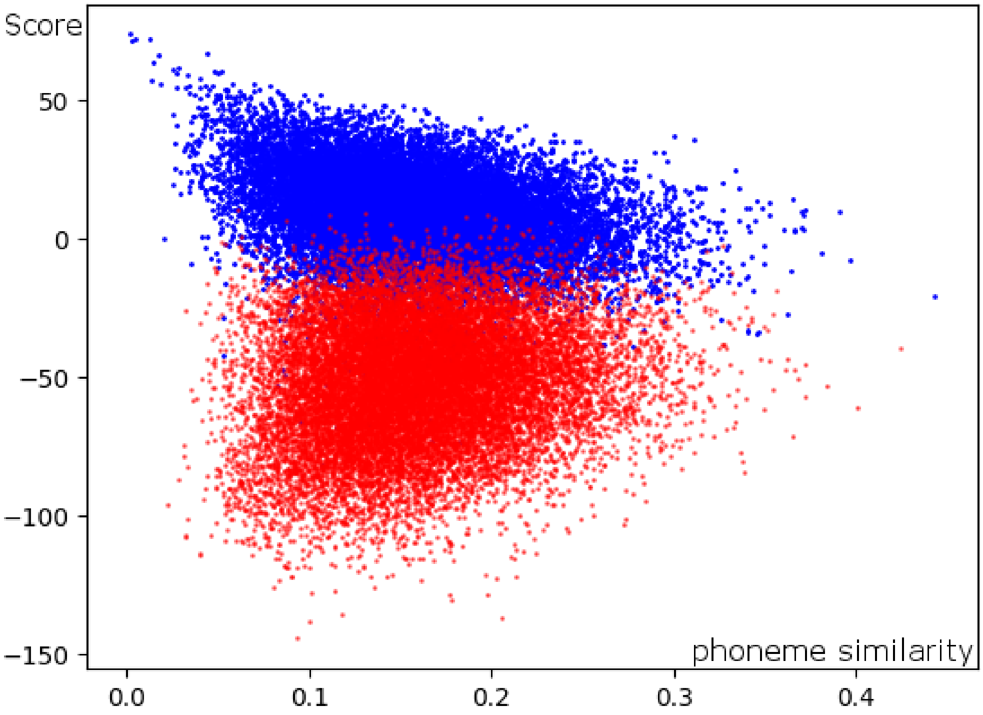

Figure 5.2: Score – Phoneme similarity graph. Blue dots are for target trials, red dots are for nontarget trials according to the reference. Lower value of phoneme similarity refers to higher similarity

In addition, a slightly increasing trend of plots for nontarget trials is also observed. There are two effects of phoneme similarity on the score that could be expected. First, higher phoneme similarity means that utterances sounds more similar to the system, which should increase the LLR score. And second, the more phonetically similar the scores are, the easier it should be for the system to tell the difference between target and nontarget scores. In other words, target scores should increase and nontarget scores should decrease.

This could also improve the ASR process. If it were proved that this aspect of speech has some influence on the result of the speaker recognition process, the values of phoneme similarity could be somehow incorporated into the ASR system. This would stabilize the score in more secure values.

5.3 Extensive data testing

Comparison of HSR and ASR processes has already been done in Chapter 4. However, this experiment is not very conclusive, as it was done by only one person, who have some knowledge of the problematics. The experiment had to be repeated in different conditions and with more carefully selected utterances.

5.3.1 Determination of dataset structure and testing conditions

The personal experiment mentioned in Chapter 4 was performed on one person. However, a larger number of test subjects was needed to get statistics, indicating the tendencies of people performing HSR. Therefore, a group of 8 people was selected. None of these people have ever made any similar speaker recognition tests, nor are they aware of the functions of the ASR system or the phonetic procedures of HSR. This ensures the same level of knowledge for all participants and minimizes variations caused by awareness of these processes. These people were randomly grouped into two groups of 4 people each. One test set was created for each of the two groups.

To see if in some cases, people can perform speaker recognition with better success than the automatic system, it was necessary to incorporate trials labeled as *target*, as well as *nonarget* trials in to the data set. From previous analyzes and data extraction, the above-mentioned table was created using automatic speech and speaker recognition (see Subsection 5.2.2). In this table, it is possible to find information about recordings (file names), number of overlapping words, score and reference. The test data were then selected into data sets according to the following criteria:

- Target trial
  - Likelihood ratio very high (HS)
  - Likelihood ratio very low (LS)
  - Likelihood ratio around zero (ZS)
  - High amount of overlapping words (HW)
  - No overlapping words (ZW)
- Nontarget trial
  - the same as for target trial

According to these criteria, trials were found in the table to meet these requirements. A total of 20 trials have been selected, ie. 40 utterances, and all of them were verified for that none of the criteria was evaluated incorrectly by the automatic system. This verification was done manually. Listening to chosen utterances, the correctness of the recognized words during automatic speech recognition was checked, the number of overlapping words, and finally, for the target trials, also whether the selected enrollment and test utterances were indeed correctly labeled as target. In some trials (including target trials) the man's voice was clearly heard in the enrollment utterance, while female voice could be clearly heard in test utterance. This error probably occurred because of wrong video extraction.

Checked trials were added into datasets. Each contains 10 trials, 5 of which are target and 5 are nontarget. Each of these 5 trials was selected according to a different criterion (see criteria above) to determine in which cases a person is able to judge match or mismatch better than an automated system, and which aspects of human speech do people focus on. According to these results, it may then be possible to improve the automatic system. The two datasets created are shown in tables 5.2 and 5.3.

Table 5.2: Final version of the first testing dataset used in the HSR process.

Data set #1					
Label	**Enrollment utterance**	**Test utterance**	**OW**	**Score**	**Ref**
1HS_T	id10280/v0Q-Vy04TjI/04	id10280/v0Q-Vy04TjI/01	5	52.21	T
1LS_T	id10304/HTL8iLI75TY/04	id10304/FJ0v0ooCIvs/05	1	-73.37	T
1ZS_T	id10272/olePnztkm6U/02	id10272/olePnztkm6U/14	1	0.045	T
1HW_T	id10306/tELtXMqh_xg/04	id10306/cc5WSHcgkl8/11	9	27.43	T
1ZW_T	id10270/5r0dWxy17C8/04	id10270/5r0dWxy17C8/24	0	22.54	T
1HS_N	id10278/shA_TsV-PFs/06	id10300/fAe7DXqrZls/04	3	9.35	NT
1LS_N	id10284/7yx9A0yzLYk/34	id10291/kt4P4cyTpWQ/10	0	-37.88	NT
1ZS_N	id10309/rxnN8thYzEQ/17	id10276/5YncBThNd5E/07	1	0.38	NT
1HW_N	id10300/8Sz2-IYJ2GA/20	id10283/GcHWzqveqyc/06	15	-82.09	NT
1ZW_N	id10304/hDBMV_0Vz4E/12	id10293/CvoEw4AJj6c/02	0	-28.58	NT

5.3.2 Human speaker recognition using datasets

After determining the datasets, 10 test subjects were asked for help. Dataset 1 was randomly assigned to 5 people, the remaining 5 people used Dataset 2. Each subject then listened to 10 trials one by one, and after each of the trials he was asked for the desired outputs. These

Table 5.3: Final version of the second testing dataset used in the HSR process.

Data set #2					
Label	**Enrollment utterance**	**Test utterance**	**OW**	**Score**	**Ref**
2HS_T	id10303/Zs8VK91yVMI/17	id10303/Zs8VK91yVMI/10	3	48.42	T
2LS_T	id10277/0rpfN7wThsg/05	id10277/tbh20gz_KRA/11	1	-56.09	T
2ZS_T	id10271/zWyD72sHVwA/01	id10271/PfcJLmkhGbk/08	1	0.51	T
2HW_T	id10271/OZcAiOZBNjQ/01	id10271/PfcJLmkhGbk/01	7	3.59	T
2ZW_T	id10270/5r0dWxy17C8/05	id10270/8jEAjG6SegY/01	0	9.40	T
2HS_N	id10288/mMHfMIiIQP0/01	id10308/ygGJs_Rh7ds/04	4	8.79	NT
2LS_N	id10298/6qFnVechX9o/10	id10304/zV7LlMi0eHo/13	0	-44.50	NT
2ZS_N	id10309/RkN41E3gRpc/04	id10276/X1Yr4eI2oRw/03	3	0.30	NT
2HW_N	id10286/ZAK9gLYNqPs/02	id10306/SBO7kM1IPaY/04	10	-47.52	NT
2ZW_N	id10305/nJbBcMdxQU4/03	id10290/kKQpvHl21Cg/07	0	-12.65	NT

outputs mean scores and additional knowledge. Scores were requested between –5 and 5 with standard tagging:

- -5 means: I am 100 % sure that both utterances belong to the same speaker
- 0 means: I am not sure at all about given utterances.
- -5 means: I am 100 % sure that each utterance belong to a different speaker

Additional knowledge then means score commenting. The subject was asked why he/she chose such a score, what led him/her to his/her decision and whether the utterance has some extraordinary parts. This will show what people are focusing on during their trial. Testing took place separately, independent of each other, and the individual trials were always played in random order. After testing, no information about the results of the automatic system was reported to the test subjects. Individual subjects' responses are attached in Appendice A.

Conclusions

The test results for data set #1 bring interesting conclusions of trials 1LS_T, 1HW_T and 1HW_N. All volunteers assigned a relatively high score to the 1LS_T trial, even though the automated system evaluated it negatively. According to verbal comments, it can be seen that the error in the automatic system may be caused by different channel or disturbing elements in the background of trial utterances, as most people, despite complaints, evaluated the trial correctly. Human tendency to judge the accent and the voice pitch can also be observed. Based on the results of the other two trials mentioned, it can be seen that the success of HSR is very high, with increasing occurrence of overlapping words. Comments suggest, that this success could actually be affected by the occurrence of the same words.

On the other hand, the results for data set #2 bring interesting outcomes for 2HW_T and 2HW_N. According to the verbal comments, the first mentioned trial was unambiguously evaluated correctly due to the occurrence of the same parasitic words. However, no respondent mentioned any occurrence of repetitive phrases in trial 2HW_N, they rather talked about different speech style and accent. From this it may be concluded that although people tend to listen to the words, they are not always the only clues for solving the task.

5.4 Improvement of ASR

The last part of this work is devoted to possible improvements of the automatic system for ASR. The aim was to bring at least one possible way to increase efficiency and success. This way is described in this section, along with its testing, evaluation of results and possible benefits it may bring.

5.4.1 Appending phoneme similarity to x-vectors

This improvement is built on a very simple basis. After phoneme similarity values were extracted from each trial (see 5.2), these coefficients could be used to create speaker profiles. Phoneme similarity labels were simply appended to the feature vector (as one-vectors) in the x-vector system. Thanks to this little improvement, the system becomes aware of which phoneme it process' in each frame and gains the ability to use that information during x-vector training and extraction as well as PLDA scoring. This improvement takes only few changes to append phonemes information in to the automatic script used for ASR during previous experiments.

After implementation, the required testing was performed. Testing was done with the data set VoxCeleb1 (see 5.1) and some little improvements can be seen on its results. Specifically, we may see an enhancement in error rate. An automatic system without awareness of phoneme similarity reached error rate of EER = 3.579 %, while the system with knowledge of this information reached EER = 3.245 %. Anyway, even while this shift is not large, it can be a step towards better results in the future.

Chapter 6

Summary

Speech is a common way of sharing information between people. Extensive studies have found that human language has signs of biometric properties and, to a certain extent, it is possible to recognize speakers according to their speech. This can be used, for example, for automatic annotation of lecture recordings, security systems, or criminal proceedings. The question, however, is which idiosyncratic features should be focused on. Speech is formed in the human vocal tract, specifically in the larynx together with arytenoid cartilages and vocal cords. Before the speech reaches the listener's hearing system, it is filtered through the vocal tract of the speaker in different ways. Because the vocal tract of each person is unique, the acoustic aspect of speech is also unique. Furthermore, if each person has different speech habits and defects, the phonetic aspect of the speech can be considered unique to every speaker. This paper describes what speech properties are considered unique and how these properties can be observed.

Speaker recognition can be performed in different ways. First, it is the natural way of human speaker recognition, second it is the modern way of automatic speaker recognition. Because people, unlike automated systems, can also capture the meaning of speech-transmitted information, both of these speaker recognition methods work differently. While the human method focuses on the acoustic and phonetic elements of speech, including, for example, speech defects, the recurrent use of some words, the tempo of speech and unusual speech habits; the automatic method extracts the important elements of speech into a mathematical model, which then represents the speaker. Therefore, the work explains how speech analysis is performed for the purpose of recognizing a speaker from both automatic and human views. At the end of the thesis, human and automatic speaker recognition experiments were performed and evaluated in a small range (1 test subject) and bigger range (8 test subjects).

Information learned from these experiments were then used in order to improve some functionality of the automatic system. These improvements were tested and outputs were evaluated. It was shown that there is a possibility of improvement by incorporating some typical aspects of speech into x-vector features.

Bibliography

[1] Co je to fonetika? online. (Visited on 20.01.2019).
Retrieved from: https://fonetika.ff.cuni.cz/o-fonetice/co-je-fonetika/

[2] Window funtion. online. (Visited on 21.01.2019).
Retrieved from: https://en.wikipedia.org/wiki/Window_function

[3] Banse, R.; Scherer, K.: Acoustic Profiles in Vocal Emotion Expression. *Journal of personality and social psychology.* vol. 70. 04 1996: pp. 614–636. doi:10.1037/0022-3514.70.3.614.

[4] Burget, L.; Plchot, O.; Cumani, S.; et al.: Discriminatively trained Probabilistic Linear Discriminant Analysis for speaker verification. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. May 2011. ISSN 2379-90X. pp. 4832–4835. doi:10.1109/ICASSP.2011.5947437.

[5] Campbell, J.: Speaker recognition: A tutorial. In *Proceedings of the IEEE*, vol. 89. IEEE. 9 1997. pp. 1437–1462.

[6] Chung, J. S.; Nagrani, A.; Zisserman, A.: VoxCeleb2: Deep Speaker Recognition. In *INTERSPEECH*. 2018.

[7] Dehak, N.; Kenny, P.; Dehak, R.; et al.: Front-End Factor Analysis for Speaker Verification. *Audio, Speech, and Language Processing, IEEE Transactions on.* vol. 19. 06 2011: pp. 788 – 798. doi:10.1109/TASL.2010.2064307.

[8] Glembek, O.: *Optimalizace modelování gaussovských směsí v podprostorech a jejich skórování v rozpoznávání mluvčího.* PhD. Thesis. Vysoké učení technické v Brně, Fakulta informačních technologií. Brno, CZ. 2012.
Retrieved from: http://www.fit.vutbr.cz/study/DP/PD.php?id=209

[9] Hollien, H.: *Forensic voice identification.* San Diego: Calif: Academic Press. 2002.

[10] Ibrahim, S.; Ramli, D.: I-vector Extraction for Speaker Recognition Based on Dimensionality Reduction. *Procedia Computer Science.* vol. 126. 01 2018: pp. 1534–1540. doi:10.1016/j.procs.2018.08.126.

[11] Ioffe, S.: Probabilistic Linear Discriminant Analysis. 01 1970. pp. 531–542. doi:10.1007/11744085_41.

[12] Kanagasundaram, A.: *Speaker Verification using i-vector Features.* PhD. Thesis. Queensland University of Technology, Science and Engineering Faculty. Queensland, Australia. 2014.

Retrieved from:
https://eprints.qut.edu.au/77834/1/Ahilan_Kanagasundaram_Thesis.pdf

[13] Kinoshita, Y.: Use of likelihood ratio and Bayesian approach in forensic speaker identification. *Proceedings of the 9th Australian International Conference on Speech Science and Technology*. 01 2002: pp. 297–302.

[14] Lindh, J.; Eriksson, A.: Robustness of long time measures of fundamental frequency. In *INTERSPEECH 2007*. 01 2007. pp. 2025–2028.

[15] Lu, L.; Renals, S.: Probabilistic Linear Discriminant Analysis for Acoustic Modelling. In *IEEE Signal Processing Letters*, vol. 10. IEEE. 2014.

[16] Mackenzie Beck, J.: *Organic Variation of the Vocal Apparatus*. 01 2010. ISBN 9781444317251. pp. 153 – 201. doi:10.1002/9781444317251.ch5.

[17] Markowitz, J. A.: Speaker recognition. *Information Security Technical Report*. vol. 3, no. 1. 1998: pp. 14 – 20.

[18] Matějka, P. e. a.: BUT HASR'12 Experience: Are Developers of SRE Systems Naïve Listeners? Technical report. Fakulta informačních technologií VUT v Brně. 2013.

[19] Nagrani, A.; Chung, J. S.; Zisserman, A.: VoxCeleb: a large-scale speaker identification dataset. In *INTERSPEECH*. 2017.

[20] Novotný, O.; Plchot, O.; Matejka, P.; et al.: On the use of X-vectors for Robust Speaker Recognition. 06 2018. pp. 168–175. doi:10.21437/Odyssey.2018-24.

[21] Palková, Z.: *Fonetika a fonologie češtiny*. Univerzita Karlova v Praze. 1994. ISBN 9788070888438.

[22] Povey, D.; Ghoshal, A.; Boulianne, G.; et al.: The Kaldi Speech Recognition Toolkit. In *Proceedings of ASRU 2011*. IEEE Signal Processing Society. 2011. ISBN 978-1-4673-0366-8. pp. 1–4.
Retrieved from: http://www.fit.vutbr.cz/research/view_pub.php?id=11196

[23] Rydin, S.: Text dependent and text independent speaker verification systems. Technology and applications. Technical report. Centre for Speech Technology KHT. Stockholm. 1 2001.

[24] Silovský, J.: *Generativní a diskriminativní klasifikátory v úlohách textově nezávislého rozpoznávání a diarizace mluvčích*. PhD. Thesis. Technický univerzita v Liberci, Fakulta mechatroniky, informatiky a mezioborových studií. Liberec, CZ. 2011.

[25] Singh, N.; Khan, P. R.; Pandey, R.: Applications of Speaker Recognition. *Procedia Engineering*. vol. 38. 12 2012: pp. 3122–3126.

[26] Skarnitzl, R. (editor): *Fonetická identifikace mluvčího*. Praha: Filozofická fakulta UK v Praze. first edition. 2014.

[27] Skarnitzl, R.; Šturm, P.; Volín, J.: *Zvuková báze řečové komunikace. Fonetický a fonologický popis řeči*. Univerzita Karlova v Praze. 2016.

[28] Snyder, D.; Garcia-Romero, D.; Sell, G.; et al.: X-vectors: Robust DNN Embeddings for Speaker Recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2018.
Retrieved from: http://www.danielpovey.com/files/2018_icassp_xvectors.pdf

[29] Traunmüller, H.: Conventional, Biological and Environmental Factors in Speech Communication: A Modulation Theory. *Phonetica*. vol. 51. 02 1994: pp. 170–83. doi:10.1159/000261968.

[30] Weingartová, L.; Volín, J.: Short-term Spectral Slope Measures and their Sensitivity to Speaker, Vowel Identity and Prominence. *Akustické listy*. vol. 20. 04 2014: pp. 5–12.

Appendices

Appendix A

Volunteers' responses to HSR

Here, the results of the experiment with 8 volunteers were written in the following table and points. Table A.1 shows the score values that the test subjects have assigned to each trial. The procedure of the experiment and the meanings of the values can be found in Section 5.3.2. Additional verbal comments, which were required, are listed in following points.

Table A.1: Table of scores volunteers have given to trials. Labels are the same as in tables 5.2 and 5.3

Data set #1					Data set #2				
Label	V1	V2	V3	V4	**Label**	V5	V6	V7	V8
1HS_T	4	4	-2	5	2HS_T	4	5	4	4
1LS_T	3	2	3	3	2LS_T	-0.5	-2	2	0
1ZS_T	5	-1	-3	4	2ZS_T	3	3	-2	1.5
1HW_T	4	3	5	4	2HW_T	4	4	5	4
1ZW_T	-1	3	3	2	2ZW_T	2	-4	-4	-1
1HS_N	2	-2	0	5	2HS_N	2	3	-2	-2
1LS_N	-5	-4	-4	-5	2LS_N	-4	-4	-3	-4
1ZS_N	3	4	-4	2	2ZS_N	0.5	-1	4	-1.5
1HW_N	-5	-3	-2	-5	2HW_N	-3	-5	-0.5	-5
1ZW_N	-2	-2	0	1	2ZW_N	-2	-5	-3	-5

Comments should give an account of the specific aspects of human speech that people focus on during their hesitation. Sentences were slightly modified, but their meaning has been preserved. There is always a label of the trial, followed by 4 comments (one for each volunteer) in the order they were assigned (V1 – V8) in the table above.

Data set #1

- 1HS_T

  1. Both have totally British accent. Pronunciations of „somebody“, „that“ and „so“ is really similar.
  2. Both have the same tone, accent and speaking style.
  3. They sound similar, although they speak about totally different things.
  4. The two sound very similar, but I feel like the girl in the second utterance speaks quicker, but the accent is unmistakably the same.

- `1LS_T`
  1. The background noise is really making it tough, they both have the same tempo and voice pitch.
  2. Same accent, both use lots of „you know“. Background is different, which is making it harder.
  3. Background noise is really annoying
  4. The intonation and colour of the voice sounds similar, but the speed and some letters do sound different.
- `1ZS_T`
  1. Sound like Hawaiian, the sound „M“ is pronounced with similar force.
  2. Man in enrollment utterance is just highlighting some words; the one in test utterance always puts accent on a same syllable. Maybe its because he's swearing.
  3. Speaking style and word expressions are different.
  4. The tempo is different, but the accent and colour sound alike.
- `1HW_T`
  1. Same tone, british english, same vocabulary
  2. The word „and“ is really interesting, bouth sound similar.
  3. Here im sure, many similar words, same pronounciation.
  4. I feel like it's the same woman. Accent sounds identical and the emphasis on certain letters is the same, but the intonation doesn't match all the time.
- `1ZW_T`
  1. Also british accent, I'm not really sure though. There is not much similar in these utterances.
  2. Vocabulary in enrollment utterance is very interesting. Both have the same accent, phoneme pronunciation is very open.
  3. Speaking styles of both are very similar.
  4. Both voices are female and have similar accents, though the second one sounds older, shakier and the emphasis on some letters does feel different.
- `1HS_N`
  1. They sound almost the same, but I'm not sure because of the background noises.
  2. Person in test utterance sounds lisping. Both are lowering the voice. Words like „thing“ and „first“ were very helpful.
  3. I really don't know.
  4. They sound the same, with the same accent. Also, he has similar pronunciation of certain letters.
- `1LS_N`
  1. Really easy, it is like old woman vs young lady.

2. Enrollment speaker is much older, the tempo is different.
3. The first speaker is older.
4. Obviously two people. First one sounds rough and older (probably a male), the second one is more feminine and younger.

- `1ZS_N`

1. Seems to be similar. Same tone, tempo, pronunciation of „actually“.
2. Special pronunciation of phonemes, „crying“ in particular
3. Meaning of both speakers' sentences are in a contradiction.
4. The second speaker has a slight lisp in his pronunciation. They have similar voice colour.

- `1HW_N`

1. Totally different, they are not similar.
2. Both are saying the same stuff words like „you know“, „okay“. Enrol sounds older though. Interesting words like marvelous.
3. Different speaking style and voice pitch.
4. Definitely different voices, different colour and accent.

- `1ZW_N`

1. This was really hard. Same colour, but the tempo is different.
2. Test speaker has different tempo and voice colour, both are repeating words but I'm not 100% sure.
3. I can't see anything similar nor different.
4. Although, they sound quite similar, there are some differences. I think the accent is different, but it is not that well distinguishable.

Data set #2

- `2HS_T`

5. Same tone, tempo, pronunciation of some words. Both are kind of lisping.
6. Accent, parasite words are the same.
7. Tone, tempo and accent seems to be the very similar. I'm almost sure.
8. Same tempo, and pronunciation of words.

- `2LS_T`

5. Different languages are making it tough. Could be the same, the first one doesn't sound like spanish guy.
6. It's hard due to different languages in utterances. Tempo is different.
7. It's hard because of different languages in utterances. Tempo is different but intonation is similar.
8. Bad english vs. fast spanish, I have no idea.

- `2ZS_T`

  5. Street vs. studio. They have the same tempo and voice colour.
  6. Both have the same accent and tempo, similar voice pitch.
  7. Enrollment is not native english speaker, has interesting accent. Test is stretching the words' length.
  8. Test speaks much quicker, but the voice pitch is similar.

- `2HW_T`

  5. Lots of „you know“ and „kind of“.
  6. „You know.“
  7. I hear the same word „you know“ and „others“.
  8. Same parasite words like „you know“.

- `2ZW_T`

  5. Different background but the tempo is quite similar.
  6. I dont like the second voice. They have different intonation and speech colour.
  7. Test is stretching the vowels' lengths. They have different tone, accent.
  8. Enrollment speaks quicker and have different accent.

- `2HS_N`

  5. Advertisement vs. noisy background. Age would be the same, tempo and pitch as well.
  6. Similar intonation, speech colour.
  7. Enrollment has higher voice pitch.
  8. Enrollment sounds younger, has different speech colour.

- `2LS_N`

  5. The age and nationality doesn't match. The first one is much older.
  6. Aged german vs. adult american. The accent is really different.
  7. Different tone. Enrollment is smooth while test is hoarsy.
  8. Enrollment is much older, different accent, intonation.

- `2ZS_N`

  5. The channel mismatch is only making it tough. I'm not really sure.
  6. I'm not sure due to the background noise.
  7. Both have no accent, same tone and tempo.
  8. Different accents, but the background is bad.

- `2HW_N`

  5. Different tempo and accent.

    6. Totally different accent, voice pitch.
    7. Accents and tempo are similar. Background noises are confusing though.
    8. Two different people absolutely. They have different speaking style.

- `2ZW_N`

    5. Enrollment is not a native English speaker, test is.
    6. Enrollment seems to be German. The speech style is different.
    7. Totally different accents.
    8. Enrollment have really bad accent and test's English is much better. Different speakers.