FONETICKÝ ÚSTAV

DIPLOMOVÁ PRÁCE

JITKA VAŇKOVÁ

Spectral properties of the source signal as speaker-specific cues

Spektrální vlastnosti zdrojového signálu jako údaje o identitě mluvčího

PRAHA 2012

Vedoucí práce: Mgr. Radek Skarnitzl, Ph.D.

I would like to express my sincere gratitude to my supervisor Mgr. Radek Skarnitzl, Ph.D. for his patient and kind guidance and for arousing my interest in forensic phonetics.

I declare that the present MA thesis is my own work for which I used only the sources and literature mentioned. I also declare that the thesis has not been used as a part of any other university programme or in order to obtain a different or the same university degree.

Prohlašuji, že jsem diplomovou práci vypracovala samostatně, že jsem řádně citovala všechny použité prameny a literaturu a že práce nebyla využita v rámci jiného vysokoškolského studia či k získání jiného nebo stejného titulu.

Praha, 17 January 2012 ……..…….

ABSTRACT: Despite a continuous development in computer sciences and related disciplines, speaker identification remains one of the most challenging tasks in forensic phonetics. The reason for this is the fact that our knowledge of how identity is reflected in the acoustic signal is still limited. The present study aims to contribute to the search of speaker-specific cues by examining spectral properties of the source signal. Specifically, it examines to what extent three short-term measures of spectral tilt, namely H1-H2, H1-A1 and H1-A3, can discriminate 16 Czech female speakers. It also addresses the influence of vowel quality, syllable status with respect to stress and position of stress group in the utterance on the values of these measures. The results show that these parameters do have some discriminative power, though the contribution of individual parameters differs. The study indicates that discrimination of speakers is the most successful in stressed syllables and argues that individual vowels could differ in their usefulness for speaker identification. The results of LDA based on these short-term measures of spectral tilt were complemented with long-term measures, namely alpha index, Kitzing index and Hammarberg index which quantify the slope of the LTAS. The present study suggests that phonatory modifications convey some speaker-specific information and could enhance speaker identification.

Key words: voice, long-term average spectrum, spectral slope, speaker identity, forensic phonetics

ABSTRAKT: Identifikace mluvčího zůstává i přes neustálý vývoj počítačových technologií jedním z nejsložitějších úkolů forenzní fonetiky. Důvodem je skutečnost, že naše znalosti akustické reprezentace identity mluvčího jsou omezené. Tato studie se zabývá spektrálními vlastnostmi zdrojového signálu a její snahou je zjistit, zda spektrální doména skýtá nějaké informace, které by mohli k identifikaci přispět. Těžištěm této studie jsou tři parametry vyjadřující krátkodobý spektrální sklon, H1-H2, H1-A1 and H1-A3 a to, jak jsou schopny rozlišit 16 českých ženských mluvčí. V souvislosti s tím je zkoumán vliv vokalické kvality, přízvučnosti slabiky a pozice taktu v promluvě na diskriminační schopnosti těchto parametrů. Výsledky ukázaly, že mluvčí vykazující statisticky významné odlišnosti v hodnotách těchto parametrů, i když užitečnost jednotlivých parametrů se liší. Ukázal se také vliv přízvučnosti slabiky; mluvčí jsou nejlépe rozpoznány v přízvučných slabikách. Studie poukazuje na možnost, že jednotlivé vokály jsou užitečnější pro identifikaci mluvčího, než vokály jiné. Výsledky diskriminační analýzy založené na krátkodobém spektrálním sklonu byly doplněny a srovnány s údaji o dlouhodobém spektrálním sklonu vyjádřeném alpha indexem, Kitzingovým indexem a Hammarbergové indexem, která kvantifikují dlouhodobé spektrum. Tato studie naznačuje, že spektrální vlastnosti zdrojového signálu by mohly přispět k identifikaci mluvčího.

Klíčová slova: hlas, dlouhodobé spektrum, spektrální sklon, identita mluvčího, forenzní fonetika

TABLE OF CONTENTS

1 INTRODUCTION

Speech undoubtedly carries some information about its producer. Thus on hearing a person speaking, we receive not only some message, i.e. *what* is said, but also some indication about the identity of the speaker, such as his or her sex, age, social and regional background, education, physical and psychological state, etc., i.e. *how* it is said. Speaker identification, however, remains one of the most challenging tasks that a forensic phonetician faces, which is caused to a great extent by the fact that our knowledge of how identity is manifested in the acoustic signal is still limited.

Some researchers have searched for speaker-specific cues in the information contained in segments (e.g., Nolan, 1983; Amino & Arai, 2009), others on the suprasegmental level - in temporal structuring of speech (van Dommelen, 1987) or its melodic patterns (Lindh & Eriksson, 2007). Recently, a claim has been made to include another prosodic domain, namely phonatory modifications or voice quality. It has been shown to be exploited for paralinguistic purposes as fundamental frequency is, but independently of it (Campbell & Mokhtari, 2003). The applicability of voice quality for forensic purposes has not been, however, thoroughly examined.

The most reliable tool for quantifying voice quality appears to be the long-term average spectrum (LTAS) (Harmegnies & Landercy, 1988; Tanner et al., 2004; Master et al., 2006). The LTAS reflects the distribution of energy across different frequencies averaged over a longer period of time, thus providing some information on the contribution of both the source signal and the vocal tract to voice quality. Several parameters, which can be automatically computed from the LTAS, have been claimed to reflect differences in voice quality by quantifying the spectral tilt, namely *alpha index* (Frøkjær-Jensen & Prytz, 1976), *Hammarberg index* (Hammarberg et al., 1980) and *Kitzing index* (Kitzing, 1986). Long-term parameters are considered one of the most powerful indicators of individual voice quality (Hollien, 1990, pp. 239-240) precisely because of the fact that they factor out the contribution of individual sounds and yield an overall value for a speaker that is independent of the contribution of individual sounds to the parameters (Rose, 2002, p. 59). Consequently, they are important for forensic purposes.

Another set of parameters which has been claimed to reflect differences in voice quality was suggested by Hanson (1997). These parameters, namely *H1-H2, H1-A1* and *H1-*

A3, likewise quantify spectral tilt, but in this case a short-term one and can be derived directly from the acoustic spectra of vowels.

The present study aims to examine to what degree can spectral information, as quantified by the above-mentioned parameters, offer speaker-specific cues of Czech female speakers. Specifically, it will observe the robustness of the three parameters suggested by Hanson (1997) to discriminate 16 Czech female speakers. It will also examine the possible influence of individual vowels, stress and stress group position in the utterance on the discriminative power of the parameters. This information will be supplemented with data obtained by the LTAS.

The present paper would like to contribute to the study of speaker-specific cues by focusing on parameters reflecting phonatory modulations or voice quality.

The theoretical introduction to the study will discuss two main areas related to the present research, namely forensic phonetics and voice quality. The first part will present forensic phonetics as a discipline and its history; specifically, how did this branch of phonetics come into existence. Afterwards, it will discuss the fields that can be distinguished within forensic phonetics in order to specify the position of speaker identification which is central to the present paper. It will also cover some comments on how individuality is reflected in a voice, what kinds of variability are present and what we mean by *speaker space*. The approaches used in speaker identification and the complications that a forensic phonetician faces when undertaking a speaker identification task will be the topic of the next section. It will be terminated by a summary of what previous research has discovered when searching for speaker-specific cues. The second part will focus on voice quality as another possible domain which could convey some information about speaker identity. Specifically, it will provide a definition of what is meant by voice quality and briefly mention subjective methods which have been used to describe it. The theoretical introduction will be finished by discussing objective methods which are used to complement the subjective evaluations. The focus will be on measures of both long-term and short-term spectral tilt (derived from the LTAS and the spectra of vowels, respectively), but other methods which relate to voice quality, such as jitter, shimmer and harmonics-to-noise ratio (HNR) will likewise be mentioned.

2 FORENSIC PHONETICS

Forensic phonetics is an applied discipline which uses the knowledge, theories and methods of general phonetics to solve practical tasks in police investigations or in court. Some tasks which arise out of these contexts require, in addition, strong ties to other disciplines, specifically speech technology and general acoustics (Jessen, 2008).

It is a relatively young discipline; the term 'forensic phonetics' has arguably received its official status at the foundation of the 'International Association for Forensic Phonetics' (IAFP) [1] in York, United Kingdom in 1991 (Jessen, 2008), and a specialist academic and professional journal for the field, *Forensic Linguistics: the International Journal of Speech, Language and the Law*, was established as recently as 1994 (Foulkes & French, 2001). In this journal, many contributions and conference reports from annual meetings of IAFP are published (Jessen, 2008). In 2004, by adding 'acoustics' to its name IAFP changed to IAFPA in order to acknowledge the contributions by specialists in acoustic signal analysis and speech technology to this field (Jessen, 2008).

One can likewise encounter an alternative term to 'Forensic Phonetics and Acoustics', namely 'Forensic Speech and Audio Analysis'. While the two terms denote the same in practice, the latter one is less specific regarding the expected education of its practitioners. Thus in some countries, forensic speaker identification or related tasks are considered more technical domains which can be undertaken by police officers or engineers trained in the use of a particular hardware or software. Knowledge of phonetics and some background in linguistics is, however, essential (Jessen, 2008).

2.1. History[2]

The use of phonetics as a forensic tool has developed mainly over the last two decades, being speeded up by the advancement of computers and the increased need of specialist analyses of speech samples in criminal trials. However, various kinds of speaker identification have been going on for thousands of years.

Even before spoken language was well organized, simple forms of signal processing were used, evidence of which comes with the development of a system of writing. First interpretable and valuable references date back to ancient Greece and Rome. One of the

[1] *International Association for Forensic Phonetics and Acoustics*, http://www.iafpa.net/ (Last accessed: October 5, 2011).

[2] Unless indicated otherwise, the introductory part dealing with forensic phonetics is based on Hollien, H. (2002) *Forensic Voice Identification*. San Diego: Academic Press, Chapter 2.

pioneers of speaker identification was the Roman philosopher Quintilian, claiming "The voice of the speaker is as easily distinguished by the ear as the face is by the eye" (Quintilian, 1899, in Hollien, 2002, p. 18).

Nevertheless, though scarce references appear earlier as, for instance, in the case of William Hulet in 1660 in Great Britain, more numerous references to speaker identification are documented only from the nineteenth century onwards. A curious case which happened in New York involves a dog being accused of killing sheep as their owner recognized its unusual bark. This accusation was accepted on the basis that "some people have such peculiar voices that they can be identified by acquaintances who hear them talk without seeing them" (Wilbur vs. Hubbard, 1861, cited in McGehee, 1937, in Hollien, 2002, p. 19). The nineteenth century also saw first discussions whether or not voices could be recognized over the telephone. At that time, though, it was not considered possible.

The turn of the century brought another important case in the history of speaker identification in Florida. The defendant, not having been seen by the victim during the crime of rape, was identified by the victim solely on the basis of his voice. Her testimony was accepted by the judge who believed that being in such a state of terror and alarm, all senses and faculties are at their most receptive and under such circumstances the voice can photograph itself indelibly and vividly in one's memory; thus enabling future recognition.

It was only a few decades later, when first modern experiments on aural-perceptual speaker identification were undertaken. A psychologist called Frances McGehee investigated what can be expected of a lay witness and what happens to his or her identification rates over time. The significance of her work lays in the fact that hers were one of the first insights into the nature of aural-perceptual speaker identification, and contemporary research still supports many of her conclusions.

The number of important speaker identification cases rose significantly when World War II began. One of them followed an assassination attempt on Adolf Hitler. A group of university phoneticians and engineers in Indiana was confronted with the task to determine whether the person making speeches after the assassination attempt was Hitler himself, or if a double had taken his place. To do so, they asked both phoneticians and panels of auditors to do aural-perceptual analysis, which they complemented by using every processing system and device available. This case was one of the first cohesive and multivector efforts in speaker identification.

Other important attempts at speaker identification were taking place at Bell Telephone Laboratories in New Jersey where a sonagraph, a machine which was supposed to make

speech visible, was developed. It was an advanced development for that time and can still be of use today. However, the use of its output, a sonagram, or 'voiceprint' as it was called, was misunderstood by some scientists. Thus while the post-war years of 1950s and 1960s saw little research in Europe as it was recovering from war, in the USA the excitement of voiceprint was spreading. Lawrence Kersta, an engineer in Bell Laboratories, believed that a reliable speaker identification system can be provided by sonagraph. He tried to find idiosyncratic patterns on sonagrams on basis of which it would be possible to identify individual speakers. The voiceprint method, the name being a desired analogy on fingerprint, is thus based on a mere comparison of sonagrams and, depending on the degree of their match, the decision is made whether the two samples come from the same speaker or from different ones. Though Kersta claimed that his method was highly accurate and many people believed this false impression, his studies lacked transparency and further research discredited this methodology. Despite that, some people working in speaker identification still use this method nowadays (Hollien, 2002, p. 64).

Another trend of the 1950s and 1960s was the excessive use of speaker identification methods by police without any reasonable guidelines and relevant research. An example thereof is the increased use of earwitness line-ups, i.e. identifying the culprit from a set of recorded voices (Foulkes & French, 2001), erroneously believed to be as effective as visual identification.

The development of modern approaches towards speaker identification started after World War II and is connected mainly to three countries: the United States, the United Kingdom and Germany. Though the 'voiceprint period' in the USA slowed down the progress for some time, it also encouraged additional research and, consequently, many early investigations were undertaken in the United States, for instance, at the Bell Laboratories, MIT, and the University of Florida. German research in this field matured later than the other two, which brought the advantage of avoiding many early errors from previous research. Due to their different conception of phonetics in general, the United Kingdom and the United States differed in their philosophies toward speaker identification; the former favouring segmental approaches while the latter favoured suprasegmental ones. In mid 1980s, the focus of research seems to have shifted from speaker identification to speaker verification (Hollien, 2002, pp. 70, 74).

An advancement of forensic phonetics was speeded up by the development of computers and relevant disciplines as well as further findings in phonetics and related disciplines, which brought new possibilities of speaker identification and verification. Even

though automatic speaker identification remains an elusive goal even in the 21st century, this method, mostly based on Gaussian mixture modelling, has been successfully used in speaker verification (Broaders, 2001). For some more comments on the development of methods in speaker identification, see Section 2.4.

In order to understand the more detailed discussion related to speaker identification, it is first important to give a preliminary overview of the fields which can be distinguished within forensic phonetics and to delimit the position of speaker identification among them.

2.2. Fields

Forensic phonetics encompasses several areas involving analysis of the recorded human voice. Authors differ in their classification depending on the criteria employed and perspective taken. As a guideline, Butcher's (2002) classification will be used for its comprehensiveness and complemented by other approaches.

According to the author, the four most frequent areas in which a forensic phonetician might be asked to prepare reports are speaker identification, disputed utterances, tape authentication and voice line-ups. Some authors single out another category, that of speaker profiling (Nolan, 1999; Foulkes & French, 2001; Rose, 2002, p. 18). In contrast, Baldwin and French distinguish only two main categories, i.e. "evidential" and "investigative" (Baldwin & French, 1990, p. 64), differing in whether a potential group of suspects has been already delimited or preliminary investigations are carried out without no particular suspects in mind, respectively. These two categories subsume the above mentioned areas nonetheless. A comprehensive approach is used by Jessen (2008), who focuses on speaker identification as the most central aspect of forensic phonetics subsuming voice comparison, voice profiling and speaker classification, and speaker identification by victims and witnesses.

Since most cases in forensic phonetics involve speaker identification and it is at the same time the most relevant field to the topic of this paper, it will be discussed as the last one in the list, and in more detail than the remaining fields.

2.2.1. Disputed utterances

Analysing the content of speech samples in which intelligibility is reduced constitutes a significant application of phonetic skills in forensics (Nolan, 1999; Jessen, 2008). Foulkes and French (2001) include it under a broader category of 'difficult recordings'. The content of recordings, which can provide useful information to law enforcers, can be difficult to decipher for technical, e.g. due to the presence of noise, or behavioural reasons, e.g. nonstandard

patterns of pronunciation. In case of the presence of noise, bearing in mind their potential negative consequences, selective filtering or other signal processing methods developed by speech engineers might prove helpful to enhance the audibility of the speech signal and determine what was actually said. Phonetic and linguistic skills are necessary especially if the unwanted noise masks the crucial frequencies or if an unusual accent is concerned (Jessen, 2008).

One famous example of a disputed utterances case is Hirson and Howard (1994) (in Foulkes & French, 2001) who analysed the content of a black box flight recording recovered from a wreck of an aircraft lost in mid-flight in 1987. As the tape spent more than a year under water, it was highly degraded. Nevertheless, phonetic analysis enabled to transcribe most of its content and this case triggered improvement in the structure and positioning of flight recorders (Foulkes & French, 2001).

The term disputed utterances sometimes refers only to those instances of problematic interpretation that are more localized. Such a case usually involves a single word, hence called 'disputed word' that is compared with undisputed examples of words in the speech of the respective talker, which the particular disputed word can represent (Foulkes & French, 2001). Examples of such analyses can be found in Baldwin and French (1990). As disputed utterances provide a defendant with the possibility to challenge the prosecution's version of what was said in the course of the recording, specialists' opinion should be sought (Butcher, 2002).

2.2.2. Tape authentication

A forensic phonetician might be consulted to examine the authenticity of a speech recording. Artificial changes to a speech sample are no longer as financially and technically demanding as Baldwin and French (1990) describe in their work; on the contrary, current software enables almost seamless editing of a signal. Therefore, as digital editing and signal manipulation becomes more widespread and physical cues of some tampering more elusive, the only kind of evidence of editing left might be of linguistic nature in the form of, for instance, unnatural changes in rhythm, intonation or tempo (Baldwin & French, 1990; Nolan, 1999; Butcher, 2002).

2.2.3. Voice line-ups

A forensic phonetician might also be asked to construct a 'voice line-up' or 'voice parade', or to analyse its contents. A voice line-up is used in those cases of speaker

identification where no permanent record of the voice involved in a crime exists, but the voice of a perpetrator has been heard by the victim or a lay witness. The earwitness then has to prove a recall of the voice, which is most often tested by means of a voice line-up (Foulkes & French, 2001).

A voice line-up is a procedure analogous to a visual line-up with the difference of the line-up being auditory instead of visual, and submitted sequentially rather than simultaneously. It consists of a set of recorded voices of people who are unrelated to the crime, so-called 'foils', and that of the suspect (Jessen, 2008). There are several criteria which need to be observed during the construction of a line-up to ensure that there is no feature of any of the voices or the recording which would stand out unfairly in relation to the others, for instance, a markedly different accent or recording fidelity (Foulkes & French, 2001; Butcher, 2002). Several guidelines for constructing voice line-ups exist; the most specific of which has been published by Broaders and van Amelsvoort (1999; in Jessen, 2008) and is followed by many forensic laboratories in Europe.

Due to the fact that the memory for voice identities decays rapidly[3], it is essential for the earwitness to be interviewed as soon as possible after the incident in order to elicit any characteristics of the voice of the offender that the witness can remember. These can help in finding the suspect and selecting foils with similar voice characteristics. When suitable foils have been selected, recordings are made of all speakers. The most important criterion is for all recordings to be technically and stylistically similar to that of the suspect. A voice line-up is correctly constructed once it is proved that every voice has an equal chance of being selected (Jessen, 2008).

This recording is then played to the witness who is asked to identify the voice of the perpetrator (Butcher, 2002). After hearing each voice, the witness has to make a decision whether this is the voice from the crime or not. Though each voice can be played several times, it is not possible to go back to previously played voices. In addition, it is advisable to record the process on a video or make observations as the witness might also react nonverbally (Jessen, 2008). Though selection of the suspect does not solely suffice as a proof that he or she is a criminal, it is powerful evidence for investigational or trial purposes (Hollien, 2002, p. 12).

[3] The first systematic research on voice memory was undertaken by Frances McGehee, as was already outlined in section 2.1. The results of her experiment show how quickly recognition level drops. Later studies elaborated on her research and discovered that it does not drop only as a function of time but a number of interacting factors is involved (Jessen, 2008). Cf. Section 2.3.1.

2.2.4. Speaker profiling

In cases where a recording of the voice of a culprit is the only clue to his or her identity, phonetic and sociolinguistic knowledge can help to define a target population and thereby narrow the search for the culprit (Foulkes & French, 2001). In such situations, when there is nothing to compare an unknown sample with, a forensic expert is requested to deliver a voice profile (Jessen, 2008).

Establishing a speaker profile is, therefore, regularly requested in the early stages of investigation of, for instance, telephone threats or kidnappings. A specialist can provide various kinds of information about the speaker's identity such as the speaker's sex, age, regional and social background and possible idiosyncratic features (Foulkes & French, 2001). Furthermore, the sample might contain other aspects which catch one's attention, such as an unusually fast speech, pausing, etc. This fact is precisely what distinguishes voice profiling from speaker classification, which is used, for instance, for constructing voice line-ups. While the main purpose of the former is to provide any information about a speaker which might be important for finding a suspect, the latter is defined in more theoretical terms; it tries to infer from the patterns of a speech recording classes or categories, such as age, social and regional background (Jessen, 2008).

Nevertheless, the strength of the conclusions that one can make in speaker profiling is highly variable. It is dependent on some aspects of the recording; for instance, its length, quality and the extent of voice disguise, but also on the dialectological and sociolinguistic information available (Foulkes & French, 2001). Due to the high demands which speaker profiling poses on a forensic expert, cooperation of specialists in different areas is often necessary (Jessen, 2008). A more detailed account of speaker profiling with numerous examples of cases can be found in Baldwin and French (1990).

2.2.5. Speaker identification

Up to 90 per cent of forensic cases involve speaker identification, i.e. identifying, by means of comparative phonetic testing, a person speaking in a criminal recording (Foulkes & French, 2001; Butcher, 2002). For this reason and the fact that aim of this study is to investigate certain speaker-specific cues, it will be discussed in more detail. Firstly, a few related terms will be addressed as these are often used interchangeably or differently by different authors.

There are three terms which are closely related to one another, i.e. speaker recognition, speaker verification and speaker identification. For these, parallel expressions exist; thus one might likewise encounter their respective synonyms where ‘speaker’ is substituted by ‘voice’ or ‘verification’ by ‘authentication’ (Hollien, 2002, p. 5). According to Hollien, the relation between them is in a way hierarchical; speaker recognition being a more general concept which subsumes the other two. Speaker identification and speaker verification then both involve identifying a person from their speech but differ in their methodology and motivation (Hollien, 2002, p. 5). Yet other authors class forensic speaker recognition tasks as speaker identification (Nolan, 1999). In the next paragraphs, the categories of speaker verification and speaker identification will be preserved and their similarities and differences pointed out.

In case of speaker verification, it is not of importance what is being said but rather who is talking. Its potential uses abound; they include, for instance, the possibility to access secure areas by a ‘voice command’, verification of identity for telephone banking (Hollien, 2002, p. 5), or for other purposes where the truth of an identity claim has to be assessed (Nolan, 1999). In contrast to speaker identification, this task is relatively straightforward since the speaker actually wants to be recognized (Hollien, 2002, p. 6). As a consequence, he or she is likely to be cooperative and willing to produce or even repeat a chosen utterance for comparison. This is ordinarily done automatically by a computer by means of comparing the voice in question to a stored reference sample of the speech of the person whose identity is being claimed. This sample is usually a rather extensive reference set which can be further developed in order to accommodate day-to-day variation in voice (Nolan, 1999). Such a variation is a natural and inevitable phenomenon since, for instance, emotions and temporary health conditions can modify the signal (Hollien, 2002, p. 8). Another factor which makes the task relatively easy is the high quality of equipment used.

Speaker identification is, according to Hollien (2002), the more difficult one of the two fields subsumed under speaker recognition. The usual task is for a forensic phonetician to compare the questioned voice in a criminal recording with that of a suspect and assess the likelihood of their being the same person (Foulkes & French, 2001) on the basis of both acoustic and perceptual properties. However, the circumstances can be, in comparison to those in speaker verification, more complicated (Nolan, 1999).

Most importantly, even if the suspect is cooperative, it is hard to obtain from him or her a sample of speech equivalent to the one which occurred during the crime (Nolan, 1999). Firstly, as in the case of speaker verification, a person’s voice is subject to day-to-day variation. However, in speaker identification, extensive reference sets, with which suspect’s

voice could be compared, are less likely to be available. In addition, the criminal sample may involve disguise as the suspect is likely to try to mask his or her identity. Secondly, the signal can be in some way degraded. This may refer to reduced frequency response when speaking on the telephone or due to low quality recording devices or microphones. Moreover, elements useful for identification may be masked by presence of noise in the environment. Though some kind of filtering can be used, one has to be always cautious as possible idiosyncratic features might thus be eliminated. These two causes which might hinder the process of speaker identification are referred to as speaker distortion and system distortion, respectively (Hollien, 2002, p. 8).

Another factor which, according to Hollien (2002) distinguishes the process of speaker verification from that of speaker identification is whether the respective trials are “closed” or “open” (Hollien, 2002, pp. 6-7). Speaker verification, the author claims, always involves closed trials; that is, the speaker belongs to some group. This group can be formed, for instance, by employees of a company who enter company buildings by a voice command. In contrast, speaker identification involves open sets as the unknown speaker must be identified within a large population of possibilities (Hollien, 2002, p. 6). However, as Nolan (1999) comments, closed sets are rather rare in real legal cases. Usually, two samples are compared; one from a known speaker and the other one from an unknown source. A forensic phonetician then assesses whether the two are similar enough to belong to the same speaker (Nolan, 1999).

Despite the complications which these tasks face, a considerable amount of speaker identification is already possible, which is enabled by what is known about the relationship between people and their voices. Those relationships which allow identification of an individual result from an integration of one’s anatomical features with his or her habitual speaking patterns. This will be the topic of the following section.

2.3. Individuality in voice

Undoubtedly, speech carries information about its producer. Thus one is often able to tell who is speaking without seeing the speaker, e.g. on the telephone or on the radio. Researchers have been trying to discover how a speech signal encodes information about its producer, i.e. which are the cues which enable us to make a judgement, and how reliably these can be recovered (Nolan, 1999). There are two main problems which complicate this process; firstly, a person’s voice is by no means constant, and secondly, it is not known whether every single person’s voice unique is to him or her and different from those of all other people, i.e.

"whether intraspeaker variability is always smaller than interspeaker variability in all situations and under all conditions" (Hollien, 2002, p. 7). Arguably, however, adding dimensions for discriminating speakers could result in the fact that speakers' ranges of variation cease overlapping (Rose, 2002, p. 31).

2.3.1. Sources of variability

Traditionally, interspeaker variability is divided into two categories, i.e. 'organic' and 'learned', though this division seems to be simplistic. Organic variability subsumes all kinds of variation which can be explained by differences in one's physique; for instance, resonance frequencies and the rate of vocal folds vibration are dependent on the dimensions of a vocal tract. Learned variability, on the other hand, is a product of one's linguistic environment; in other words, by living in a certain environment, people acquire some regional and social variety. A very simple model of the information contained in a voice would thus be that it consists of two parts, namely individual and linguistic. The point to stress is that this division is not absolute but these two phenomena combine together in a single manifestation (Nolan, 1999). To give an example, the range of an individual's fundamental frequency reflects both the organic aspect, i.e. the structure of the larynx, and the learned aspect, i.e. features of the particular language (Nolan, 1999). The individual and linguistic aspects are thus convolved in its acoustic representation, which is an important fact for speaker identification as it is necessary to understand how this happens in order to interpret the variation inherent in speech (Rose, 2002, p. 60).

Incidentally, this interplay of organic and learned features in one's voice provides an explanation why the once acknowledged parallel between fingerprinting and speaker identification is invalid. While a fingerprint reflects an organic (and invariable) difference only, in a recording of speech, the organic aspect (which is variable) is combined with the effects resulting from one's linguistic environment together with the choices made at the given moment (Nolan, 1999).

The source of this variability is the high 'plasticity' of the mechanism producing speech. Though a speaker's physique poses some limits on, for instance, the range of fundamental frequency, within these limits speakers dispose of wide scope for controlled variation. This plasticity can serve linguistic purposes, e.g. realizing elements of phonology, paralinguistic purposes to convey anger or affection, or can be exploited for non-linguistic purposes, such as voice disguise. In addition, apart from volitional uses of the plasticity, there are other factors as a result of which one's speech varies. These include temporary conditions

such as cold or other states of ill health which affect vocal organs, stress, fatigue or intoxication (Nolan, 1999). Moreover, different voices can be affected in different ways; for instance, when talking on the telephone, most people, but not all, speak more loudly, which results in a rise of average fundamental frequency (Foulkes & French, 2001). Some indications of how anatomy is reflected in one's voice will be discussed below.

2.3.2. "Speech as anatomy made audible"[4]

In order to illustrate how anatomy is manifested in the acoustic signal, it is useful to adopt the source-filter model of speech production (Fant, 1960); that is, larynx being the source of acoustic energy that is shaped in the supralaryngeal vocal tract, which thus functions as a filter (Nolan, 1999).

As for the 'source', the relation between anatomy and acoustics can be demonstrated on the vocal folds. Their length and mass determines the range of frequencies at which they can vibrate, which is reflected in the shape of the glottal source wave. Possible anatomical irregularities, such as uneven cycles, are manifested in the acoustic signal, too. The vocal tract 'filter' likewise differs in size and shape, which is reflected in formant frequencies. Importantly, the source and the filter interact. To give one example, the impression of a high voice can be caused by high fundamental frequency due to small vocal folds as well as high formant frequencies as a result of a small vocal tract (Nolan, 1999).

The vibration of vocal folds is not the only source of acoustic energy in speech; it can also be generated by air turbulence. Some claims have therefore been made for the usefulness of fricatives as speaker-specific cues since their precise acoustic properties depend on the shape and size of the relevant place of articulation (Künzel, 1987, pp. 93-4, in Nolan, 1999). However, even the articulation of fricatives is susceptible to volitional changes (Nolan, 1999). This fact led some researchers to focus on nasal sounds as the best manifestations of anatomic individuality since the shape of the nasal cavity varies among individuals and is not volitionally changeable. Nevertheless, there are two main reasons why even nasals have to be approached with caution; firstly, nasal resonance is not available in isolation but rather combined with other resonances of the vocal tract and secondly, as was already mentioned, it varies as a result of temporary health conditions, such as cold (Nolan, 1999).

[4] Quoted from Nolan, F. (1999). Speaker Recognition and Forensic Phonetics. In: William, J. and Laver J. (Eds.) *The Handbook of Phonetic Sciences*. Hardcastle: Blackwell Publishing, Blackwell Reference Online. 28 December 2007. Available online from: http://www.blackwellreference.com/subscriber/tocnode?id=g9780631214786_chunk_g978063121478625. (Last accessed: December 18, 2011)

Though considerable research on finding speaker-specific cues has been undertaken (see Section 2.5 for a more detailed account of areas to which the search of speaker-specific cues has led), our knowledge of how identity is encoded in the acoustic signal is limited.

Apart from anatomy, speech conveys information also about a given communicative situation. The following section will briefly examine what is meant by 'learned' variability.

2.3.3. Speech as a tool

Speech is not only a physical event, but it is also shaped by the environment. As a consequence, every fully competent speaker of a given community has a mastery of several registers which he or she uses as situation requires by exploiting the potential of a language on all its levels (Nolan, 1999). The fact that speakers do not have an invariant accent has consequences for forensic phonetic comparisons as sociolinguistic factors have to be accounted for (Rose, 2002, p. 62).

In addition, between organically determined and learned variability, there seems to be a space for potential individuality. By combining all these resources, that is, social, economic, geographical and educational factors but also sex, intelligence, etc. (Hollien, 2002, pp. 9-10), people create for themselves "a linguistic phonetic system" (Nolan, 1999, p. 3) which marks them as members of a subgroup of population which might, in extreme cases, consist of a single individual (Nolan, 1999). Thus sometimes it can be the use of idiosyncratic phrases which makes one recognize the speaker (Hollien, 2002, p. 2); the use of lexis and grammar is also one of the parameters used in casework (Foulkes & French, 2001). However, the focus here will be on phonetic parameters. Much research tries to determine which phonetic parameters of a voice are the most useful in identifying an individual speaker (Foulkes & French, 2001) (see Section 2.5 for a more detailed discussion of speaker-specific cues). According to Hollien (2002), there seem to be two main "schools of thought" (p. 13) in speaker identification, namely segmental and suprasegmental. While the former favours the segmental level, the latter stresses the importance of the suprasegmental one as it is considered more stable and speaker specific than phonemes themselves (Hollien, 2002, p. 14).

Though there seems to be no single feature of a person's voice which would allow him or her to be differentiated from all other people, a combination of several robust parameters provides a reasonable characterization of an individual (Hollien, 2002, p. 10). Research shows that despite the inherent variability of a voice, some elements of a person's voice are idiosyncratic enough for the purposes of speaker identification provided that the approach taken is well structured and robust techniques are employed (Hollien, 2002). The

following subchapter and its sections will discuss the development of approaches towards characterizing an individual speaker.

2.4. Development of approaches

As mentioned in Section 2.1, all speaker identification which had occurred until the twentieth century was done by human beings only. By listening to a voice, people carried out some kind of aural-perceptual analysis and stored relevant features of the heard voice in their memory. On hearing a voice, they attempted to link it to a particular individual; that is, one whose speech they have already heard and for whom they had stored some set of features. It should be pointed out that the process of voice perception and recognition differs substantially for familiar and unfamiliar voices. According to Kreiman & Sidtis (2011), recognizing a familiar voice resembles *pattern recognition* where "top down" processes dominate over "bottom up" processes and the process of voice recognition is very fast. In contrast, recognition of unfamiliar voices relies much more on *featural comparison* and is thus primarily driven by "bottom up" processes (p. 187). Kreiman & Sidtis (2011) call it "Fox and Hedgehog" model of voice perception (p. 187-8), thus referring to the Greek poet Archilochus who wrote that "the fox knows many things but the hedgehog one big thing" (Kreiman & Sidtis, 2011, p. 188). As Kreiman & Sidtis comment,

> For voices, the little things, or features, are utilized more successfully in unfamiliar voice perception, whereas the familiar voice is one big thing, in which "features" appear in idiosyncratic combinations cohering and/or "emergent" to yield a complex, integrated pattern.
>
> Kreiman & Sidtis (2011, p. 188)

It was only recently that these features have been formalized and two more organized approaches supplemented the previous unstructured one, i.e. earwitness identification and analysis undertaken by professionals who are specifically trained for this purpose (Hollien, 2002, p. 11). An advancement in speaker identification by experts occurred in the first half of the last century when the tape recorder and sound spectrograph made it possible to capture, replay and visually represent speech (Broaders, 2001). With a further technological advancement and the invention of a computer, an auditory method was supplemented with an acoustic one. Nowadays, speaker identification tasks include both an auditory and an acoustic analysis, though there has been a debate on the relative merits of the two approaches (Foulkes & French, 2001) (see Section 2.4.2).

Thanks to a continuous advancement in the development of computers and related disciplines, such as audio engineering but also acoustics, etc., new, automatic techniques of speaker recognition became possible. However, speaker identification remains a challenging field even in the 21st century and the promise held by technological advances remains largely unfulfilled. Automatic methods, mostly using Gaussian mixture modelling, remain to be limited to relatively low-risk applications in the area of speaker verification.

In speaker identification, the most successful system of this kind seems to be SAUSI, i.e., a semiautomatic system of speaker identification, which has been developed over the last several decades by Hollien and his colleagues of various professions, e.g. phoneticians and forensic phonetician, audio-engineers, computer scientists, psychologists and linguists. Already ten years ago, SAUSI presented a well-developed and successful system of speaker identification as it has been reflecting new findings in phonetics and related disciplines.

There are several reasons why fully automatic methods cannot be used for speaker identification. Firstly, there is a considerable possibility for two speakers to be, in some respect, identical as there seems to be no feature of the voice which would be unique to every speaker. Secondly, it is not known whether there are some features of the voice which cannot be consciously changed by the speaker (or some other conditions, such as cold in case of nasals). Another fact which complicates the process is that there is not enough data to quantify the chances for two speakers to be similar or even identical with respect to a certain features. In addition, acoustic parameters vary as a consequence of different recording conditions or differences in the voice itself and automatic techniques are not yet able to separate these two sources of variation (Butcher, 2002).

The focus here will be on auditory and acoustic approaches towards speaker identification. The following section will discuss some general aspects of auditory approaches. After that, the focus will shift to the analysis by professionals both auditory (Section 2.4.2) and acoustic (Section 2.4.3).

2.4.1. General aspects in speaker identification

There are some factors which can influence the accuracy of speaker identification. They will be divided into three groups, namely those relating to the listener, the sample and the speaker.

As mentioned in Section 2.2.3, an important variable in auditory approaches is voice memory. Frances McGehee's experiment on this topic was important not only because it showed the degree of decay over time, but it also opened a discussion as to other possible

factors which can play a role in the process of speaker identification, such as the relevance of gender, foreign dialect and voice disguise (Hollien, 2002, pp. 28-9), which will be addressed later in this chapter. Moreover, it provided some insights into the process of auditory speaker identification itself and contemporary research still supports some of McGehee's views. Nevertheless, our knowledge of voice memory is still sketchy; it is not yet possible to specify exactly the shape of a decay curve or what can be expected of any particular individual in speaker identification tasks as there are many variables which can affect it (Hollien, pp. 30-1).

Non-contemporary speech, i.e. samples separated by some period of time, had been claimed to pose as difficult a challenge to speaker identification as voice memory decay (Rothman, 1977, in Hollien, 2002, pp. 31-2). Rothman reports a drop to 42% of accuracy when non-contemporary samples are included, which would be detrimental to auditory approaches (Hollien, 2002, pp. 31-2). However, he included in his research so-called 'sound-alikes', e.g. brothers, father and son, etc. (Hollien, 2002, p. 35), which strongly affected his results. Later research by Schwartz (1995) (in Hollien, 2002, pp. 32-4), has shown that judgements become unstable only after a period of 20 years. Thus using non-contemporary samples has just little effect on speaker identification unless a very long period has passed (Hollien, 2002, pp. 32-4).

To explain McGehee's results in more detail, a set of experiments was undertaken to study the possible effects of gender and training. Though McGehee also reported the performance of her male listeners to be better than that of females, recent studies do not show any effect of gender either for the speaker[5] or for the listener. However, the effect of training is apparent. Trained phoneticians perform generally better in speaker identification tasks than people without similar background and experience. This effect is even more apparent when conditions are unfavourable, that is, the sample is short or a speaker unknown, etc. (Hollien, 2002, p. 34). Additional training in forensics and well-structured systematic approaches result in yet better performance in speaker identification tasks (Hollien, 2002, p. 39).

Another factor which influences the speaker identification process is familiarity with the speaker. In general, familiar speakers are easier to recognize than unfamiliar ones (e.g., Nolan, 1999, pp. 677-681; Hollien, 2002, pp. 43-6). In his research, Hollien sought to assess how familiarity with the speaker affects judgement accuracy. The results show that if the listener is very familiar with the speaker, he or she can be recognized even when conditions

[5] This task is usually an easy one since by listening to speaker's fundamental frequency or vowel formants, identification of the speaker's gender tends to be straightforward. However, problems might occur if the values of his or her fundamental frequency lie between those for men and women. Yet there is evidence that judgements of gender can by likewise made by listening to speaker's consonants (Hollien, 2002, p. 37).

are unfavourable; in this case stress and voice disguise. However, lack of familiarity, even after some training in recognition, results in the decrease in accuracy, especially under difficult conditions (Hollien, 2002, pp. 44-5).

Factors relating to a speech sample, particularly its size or duration and acoustic quality, can also influence the performance of a listener, be it an earwitness or a professional, in speaker identification.

As for the size, researchers differ in what they consider a sufficient sample (Hollien, 2002, pp. 39-40). Künzel (in Hollien, 2002, p. 40), for instance, claims that the German Bundeskriminalamt, i.e. Federal Criminal Police Office of Germany, requires at least a 30-second sample. Hollien, on the other hand, argues that shorter samples can be sufficient (Hollien, 2002, p. 41), depending on what is the subject of investigation. This topic will be further examined in relation to LTAS in Section 3.1. Unfavourable acoustic conditions, especially noise and limited bandwidth, present a further complication to speaker identification as they mask or distort the speech signal. A noise can be broad band or narrow band; steady or intermittent. Furthermore, it does not have to be aperiodic in nature. A term which is often encountered in forensic context is ‘forensic noise’ (Hollien, 2002, p. 41). It subsumes all competing signals which interfere with the speaker identification process, such as music, other speakers, etc. The most troublesome kind of noise is a loud, broad band, steady noise. Nevertheless, various filtering and other methods exist to mitigate these effects (Hollien, 2002, p. 41). The speech signal can be further degraded by limited bandwidth, the most common cause of which is the telephone. Though some more complicated remedy methods exist, reasonable speaker identification can be carried out even on a speech sample obtained over the telephone (Hollien, 2002, p. 41).

Lastly, there is a group of factors originating with the speaker, such as disguise, presence of stress or emotion, or the language being spoken (Hollien, 2002, p. 46), which can all hinder speaker identification tasks. However, before discussing them, two other important factors which make speaker identification problematic will be mentioned. First of these is the size of population. As a general rule, selecting a speaker out of a smaller number of possibilities is faster and easier than from a larger one. Nevertheless, the problem one faces in real cases is that the number of possibilities cannot be controlled (Hollien, 2002, p. 46). The second complicating factor involves uniqueness of a speaker’s voice; that is, a voice that has few distinctive features will be harder to recognize than a voice which exhibits more idiosyncratic features (Hollien, 2002, pp. 46-7). Similarly, Nolan claims that speakers who are the most distinctive lie outside the ‘normal’ range (Nolan, 1999, p. 6). According to him,

these are, for instance, speakers with different kinds of speech pathologies and impediments, or non-native speakers of a particular language. Though speech phenomena which lie outside the normal range are very valuable for speaker identification purposes due to their rarity, it is important not to overemphasize their significance (Nolan, 1999). One example which illustrates this is a so-called 'Brother-My-Brother' case (Hollien, 2002, pp. 47-48).

Disguise is another important factor in speaker identification because, if a speaker is good at it, its effects can be detrimental. However, any attempt at voice disguise poses problems for speaker identification. One of the most challenging types of disguise is whisper as all information about fundamental frequency and heard pitch is reduced or eliminated. Similarly, it affects information about vocal intensity, voice quality, but also prosody and speech timing (Hollien, 2002, p. 49); that is, most features or parameters important for recognizing a voice. Only little research has been done on how to detect and counteract voice disguise (e.g., Perrot et al., 2007). The first step, according to Hollien, is always to determine whether the speaker is attempting voice disguise or not. The fact that it is very difficult to consistently disguise one's voice makes this decision easier (Hollien, 2002, p. 50). However, in real cases, one often has to work with only limited samples. Yet effective speaker identification is sometimes possible even if voice disguise is involved (Perrot et al., 2009).

Stress and emotions likewise present obstacles to speaker identification tasks. There has been substantial research on the effects of psychological stress, such as anxiety, fear, anger or fatigue (Hollien, 2002, p. 53) on a voice, and some general trends have been discovered; for instance, a rise in fundamental frequency, a less marked increase in vocal intensity and speaking rate, etc. There are thus some predictable patterns which a voice under stress tends to follow and knowing them allows one to compensate for these effects. Evidence shows, however, that a small number of people under stress do not show these characteristics; they might even reverse them (Hollien, 2002, pp. 51-3). Nevertheless, though research has been done on the effects of stress and, recently, also emotions, little is known about how they affect speaker identification process (Hollien, 2002, p. 51). Notwithstanding, a forensic phonetician must be able to identify these factors and should be able to counteract them (Hollien, 2002, p. 52).

Lastly, effects of dialects, accents or foreign languages will be briefly mentioned. Identifying dialects of different speech samples to be the same is often of only limited importance unless the number of speakers of such a dialect is very small. In contrast, identifying two dialects as distinct is rather significant as it suggests that the speakers involved are two different people, unless a different accent is adopted with the view of

concealing one's identity (Baldwin & French, 1990, pp. 66-9). As for foreign languages, some researchers report that identifying a speaker of a different language than their own complicates the process. Though Hollien believes that the knowledge and methods a forensic phonetician has at his or her disposal can counteract the possible effects (Hollien, 2002, pp. 54-5), it should be stressed that speaker identity is signalled differently in different languages and a forensic phonetician must, therefore, have a specialist knowledge of the language under investigation (French, 1994, p. 174, in Rose, 2002, p. 47).

Before getting to analyses by professionals, features of voice which people use when recognizing voices will be mentioned. These include heard pitch, articulation, voice quality, i.e. the signal produced by the vocal folds and modified by the resonances of the supralaryngeal vocal tract, prosody, i.e. temporal patterning of speech and melody, vocal intensity and possible idiosyncrasies. To recognize a speaker, a listener may use all or only some of these features and relationships discussed above (Hollien, 2002, pp. 59-61).

Professionals, who are prevailingly phoneticians with some background in linguistics and computer or audio-engineering (Hollien, 2002, p. 12), employ these elements when doing analyses, too. They combine them into an appropriate system after which they examine the samples acoustically, extracting relevant parameters and thus performing a composite speaker identification task (Hollien, 2002, p. 62). This combined approach is, however, only one of three different philosophies held among forensic phoneticians. Some researchers argue that auditory approach is relevant on its own (e.g., Baldwin and French, 1990, p. 9), while others claim the opposite; that is, the auditory analysis is not necessary as the acoustic one can provide sufficient data. Today both auditory and acoustic analyses are generally recognized as indispensable and of equal importance; thus the combined approach is used by most forensic phoneticians (Rose, 2002, pp. 48-9).

The auditory analysis will be discussed in the following section and the acoustic analysis in the subsequent one.

2.4.2. Auditory analysis

As has been just mentioned, professionals in the area of speaker identification are prevailingly phoneticians, though there is a considerable degree of heterogeneity of people working in this field regarding their background, training, talent but also opinions about speaker identification in general (Hollien, 2002, p. 63).

There is an array of people ranging from private detectives through linguists to forensic phoneticians. The first group is very numerous and involves also some members of

an official organization called International Association of Identification (see http://www.theiai.org/), who still use some form of voiceprint method. The second group includes specialists of various related disciplines, such as speech pathologists or audio engineers, who lack background in other fields necessary for speaker identification tasks. Many engineers work in speaker verification, which seems better suited for their skills and knowledge. Importantly, progress in both domains is supported by their cooperation. The core group working in speaker identification is formed by forensic phoneticians. The essential areas in which one has to be skilled include acoustics, physiological and perceptual phonetics, psychoacoustics, linguistics and statistics. Furthermore, computer skills, grounding in behavioural sciences, electrical engineering and forensic sciences as well as understanding of how the legal system works are likewise desirable (Hollien, 2002, pp. 63-6). A list of requirements which are supposed to assess the capabilities of a forensic phonetician are included in Hollien (2002, p. 69).

An auditory analysis compares voices as to their auditory features, that is, how voices and speech sounds sound. It logically precedes an acoustic analysis as it is necessary first to listen to a sample to assess whether its quality enables a further analysis and to identify parameters which could be used to compare the samples. It should provide an overview of both similarities and differences between the speech samples compared (Rose, 2002, pp. 48-50).

As mentioned in Section 2.2, two main approaches towards characterizing an individual speaker are segmental and suprasegmental.

A segmental approach has been subjected to criticism for various reasons. Some people argue that relevant segmental elements for speaker identification have not been identified and tested in different conditions. Furthermore, they have not been organized into any model tested for its effectiveness. An exception to this seems to be Nolan's well-structured model based on the Framework of British pronunciation (Hollien, 2002, pp. 71-3). Another source of criticism derives from the fact that segmental approaches usually require a large amount of data and the system of transcription, namely narrow phonetic transcription using International Phonetic Alphabet, is too complex. However, this complex system is at the same time its advantage as it allows a phonetician to describe any sound they hear by means of an established system. Other phoneticians can thus easily check and verify the assessment. (Rose, 2002, p. 51). Despite the mentioned weaknesses, a segmental approach can be very useful in, for instance, assessing regional dialects. Its strength is further enhanced when combined with suprasegmental or acoustic approaches (Hollien, 2002, pp. 73-4).

Other approaches have derived their procedures rather from experimental phonetics than traditional one, and use parameters on the suprasegmental level (Hollien, 2002, pp. 74-5). In casework, these two are usually combined. Descriptive and classificatory frameworks exist also for these aspects of voice. For voice quality, for instance, a thorough guideline is provided in Laver (1980) *The Phonetic Description of Voice Quality* (Rose, 2002, pp. 51-2).

Hollien argues that if an aural-perceptual approach is well structured, it can be sufficient on its own[6], though effectiveness is increased by complementing it with an acoustic analysis (Hollien, 2002, p. 70), and provides an example of a highly structured approach based on the assessment of up to twenty scaled comparisons. These features involve information about pitch, voice quality, intensity variation, dialect, segments, prosody and possible disorders. To assess them, it is recommended to construct a so-called 'pairs tape' (Hollien, 2002, p. 78) which allows better comparisons of the known and unknown speaker. Then, using a 10-point scale, which expresses the confidence level as to match or non-match, the parameters are considered one at a time, listening as many times as necessary to permit judgement about a single relationship. Only then, the next parameter is considered. The entire set of judgements must be completed in one sitting and repeated a number of times, preferably on different days. Then, individual means for each parameter are obtained. Sometimes it is necessary to 'weight' the values, that is, to check relative importance of the features. For instance, as has been already noted, a match in dialect is rather insignificant if it is spoken by a large number of people. A resulting polarized score should be likewise approached with caution if it is caused by differences in the situation or the environment. Therefore, for this approach to be a robust one, it requires some rigor in its administration (Hollien, 2002, pp. 78-85).

Hollien's structured approach is thus designed in such a way so that it could be used as a stand-alone procedure. However, an auditory analysis is usually complemented with an acoustic one. An auditory analysis in general thus involves careful and repeated listening by a specialist during which he or she assesses relevant features of voice by means of agreed frameworks that enable analysing fine differences and a comparison between subjects and studies. According to Butcher (2002), features of voice which are subject to an auditory analysis and may convey information about the identity of a speaker fall into four main categories.

[6] This view was already held by Baldwin and French: "… there is a certain amount of disagreement in the academic world at the present time as to which method should be employed in forensics, and I will summarize my position on that matter by saying that I have found the auditory approach to be fully adequate for the task" (Baldwin and French, 1990, p. 9).

Firstly, the expert has to ascertain voice quality[7] , that is, the sound created by the vibration of vocal folds, regardless the contribution of resonances created in supralaryngeal cavities. There are several description frameworks which allow its quantification. These include terms such as 'strain', 'breathy', 'creaky' (Butcher, 2002), which are supposed to describe auditory impressions of a listener as accurately and objectively as possible.

Secondly, articulatory settings, i.e. characteristics of a voice which are not produced by the larynx, are commented on. In practice, this means evaluating the effects of the long-term setting of the throat, tongue, lip or nasal resonances. As in the case of voice quality, established descriptive frameworks are available for articulatory settings, too. They rate the speech as to, for instance, 'hypernasality', 'pharyngalisation' and 'labialisation' (Butcher, 2002).

Thirdly, a set of parameters is used to describe articulation patterns which could provide clues to speaker's geographical and social background (Butcher, 2002). In assessing these features, therefore, cooperation with a dialectologist is advisable (Nolan, 1999). Nevertheless, the usefulness of this approach depends on a particular linguistic community. While some languages exhibit an array of sociolinguistic and regional variation, others do not.

Lastly, the expert undertaking an auditory analysis listens for a possible presence of idiosyncratic features of any kind. These might concern the articulation of consonants, stuttering, various kinds of dysfluency, etc. (Butcher, 2002).

Though some authors argue the sufficiency of sole auditory analysis for characterizing an individual speaker (e.g., Baldwin & French, 1990, p. 9; Hollien, 2002, 70), nowadays an approach that combines an auditory and acoustic analysis is required because both analyses, when applied on their own, exhibit significant shortcomings. An auditory approach is not sufficient on its own as, due to how our perceptual mechanism works, two voices can sound similar despite their significant differences in the acoustics. An acoustic approach lacks adequacy when used on its own for the same reason, i.e. two speakers may not be effectively distinguishable acoustically and yet differences can be spotted by the trained ear. Moreover, only an auditory analysis enables identification of linguistically relevant data. A previous auditory analysis is necessary to indicate what is comparable and to select appropriate parameters for both auditory and acoustic analyses (Rose, 2002, pp. 49-50). The following chapter will discuss acoustic techniques.

[7] For a more detailed discussion of voice quality, see Chapter 3.

2.4.3. Acoustic analysis

An acoustic analysis allows for speaker-related aspects of speech assessed in an auditory analysis to be quantified. Furthermore, it reveals information which our auditory system may obscure due to its being engaged with extracting linguistic information. The reliability of results gained by an acoustic analysis for speaker identification tasks has, however, likewise been questioned, mainly due to our lack of understanding of 'speaker space' (Nolan, 1999, p. 2). If one imagines a multidimensional space comprised of all parameters along which speakers are differentiated; for instance, mean fundamental frequency or mean second-formant frequency, then each speaker occupies a region in this space, hence speaker space, which covers the variability of his or her speech (Nolan, 1999). The problem with finding acoustic cues which would be reliable indicators of one's identity is, as was already stated, that we do not know whether intraspeaker variability is always smaller than interspeaker variability. Thus a considerable amount of research focuses on finding the most useful phonetic parameters for identifying an individual speaker as well as determining the degree of variability along various phonetic and sociolinguistic dimensions (Foulkes & French, 2001).

Ideally, the acoustic parameters should exhibit large interspeaker variability and low intraspeaker variability and should be extractable even from short samples, i.e. they should have a high frequency of occurrence. They should also be easy to extract, accurately measurable, and resistant to disguise and other kinds of distortion (Nolan, 1983, p. 11, in Rose, 2002, pp. 65-6). There seems to be no single parameter that would satisfy all criteria. In addition, parameters employed should be independent of one another (Rose, 2002, p. 66). According to Hollien, the most robust parameters for expressing individuality should be based on 'natural' speech features which are ordinarily used by humans in everyday processes of identifying people from their voices, such as "the pitch level of voice, pitch patterns and variability, vocal intensity patterns, dialect, voice and speech quality, prosody (the timing and/or melody of speech), articulation, and so on" (Hollien, 2002, p. 10). There is a wide array of different parameters which allow comparison of speech samples for forensic purposes and their choice should be determined by "a linguistically informed analysis" (Rose, 2002, p. 67) of the speech material. Thus, although there are some preferred parameters which should be compared if possible, there is no predetermined set of parameters to use. The choice always depends on the specific case and is partly language-dependent as intraspeaker

variation and interspeaker variation is reflected in different features in different languages (Rose, 2002, p. 47).

The following chapter will provide an overview of the results of the search for acoustic cues characterizing individual speakers, which will be subsumed under four categories: segmental information, temporal structuring, melodic parameters and phonatory modulations.

2.5. Parameters relevant for characterizing an individual speaker

This section will consist of four parts, each dealing with one domain which is considered to provide, to some extent, some indication of speaker's individuality. The discussion will start on the segmental level and will move to assessing three domains on the suprasegmental level with an emphasis on the last parameter, namely phonatory modulations or voice quality.

2.5.1. Segmental information

The first area to which the search for speaker-specific cues has led is the information contained in segments. As previous research has shown, speaker individuality interacts with phonological information in the speech signal and researchers have been thus trying to find segments which would allow reliable discrimination of speakers. If acoustical correlates of one's individuality contained in speech sounds were found, they would allow not only more efficient speaker identification by focusing on those sounds but would also contribute to automatic speaker recognition. After several decades of research in this area, it has been shown that both consonantal and vocalic characteristics can provide some information about a speaker.

Firstly, consonantal cues will be discussed. Traditionally, nasals have been considered robust indicators of one's identity. The search for speaker-specific cues in this direction was motivated by the fact that nasal sounds appear to be reliable manifestations of one's anatomy as the shape and volume of nasal cavity is not volitionally changeable and varies among individuals (Nolan, 1999). Their contribution to the process of speaker identification was shown already several decades ago (see, e.g., Glenn and Kleiner, 1968; Wolf, 1972, in Nolan, 1983, pp. 75-6), though the earlier research on nasals studied them either in isolation or in a single environment (Nolan, 1983, p. 76). Su et al. (1974) (in Nolan, 1983, pp. 76-7) carried the investigation of nasal parameters further by considering the phonetic environment of the nasals, too. Their results show that speakers can be characterized by the extent of their

coarticulation, and they claim nasal coarticulation to be even a better clue to one's identity than spectra alone. Amino & Arai (2009) examined interspeaker and intraspeaker differences of nasals as opposed to non-nasal sounds by means of the parameter of energy transitions which appear to capture both abrupt temporal changes (in this case, the timing of velar closure in nasal-vowel sequence) and articulatory idiosyncrasies. Their results confirm previous findings and show that energy contours of nasals are speaker-dependent and differ significantly among individuals[8] (Amino & Arai, 2009). Since nasals exhibit low intraspeaker and high interspeaker variation and the shape of nasal cavity is not changeable at will, they are considered indicators of one's identity. However, it is important to be aware of the factors which can influence acoustic properties of nasals and as a result of which intraspeaker variability can be increased (see Section 2.3.2),

Lateral and rhotic consonants appear to be further indicators of one's identity. Nolan examined the spectral properties of /l/ and /r/ in English (Nolan, 1983, p. 77). /l/ was chosen on the grounds of its acoustic similarities with nasals, such as the interaction of antiresonances with formant structure and undergoing articulatory changes depending on its phonemic environment; /r/ by its virtue of being likewise a liquid, and having a range of possible secondary articulations which can be exploited by speakers. Moreover, /l/ and /r/ meet several criteria for a robust parameter; for instance, frequent occurrence and robustness in transmission (concentration of spectral energy lies within the telephone band, i.e. 300 – 3500 Hz, which makes them better candidates for speaker-specific cues than, for instance, fricatives). His results show that both /l/ and /r/[9] convey speaker-specific information and are useful for speaker identification purposes, though of lower value than nasals (Nolan, 1983, p. 115).

Other researchers have argued for the robustness of fricatives (Künzel, 1987, in Nolan, 1999) as speaker-specific cues. Nolan (1999) gives an example of the sound [s], the acoustic properties of which depend on the size and shape of person's teeth and which should be preserved even in whispered speech. If that were confirmed, it would be a powerful tool for forensic purposes since whisper presents one of the most serious ways of speech signal degradation as most information is lost or eliminated. Amino & Arai (2009) likewise studied

[8] Their perceptual speaker identification points to coronal nasal sounds as the most effective in perceptual speaker identification. In addition, their results show interaction of other factors, such as the phonetic environment (Amino & Arai, 2009).

[9] F2 and F3 of both consonants show marked variation between speakers. As for the influence of the following vowel, /l/ exhibits a higher degree of coarticulation than /r/ as a consequence of which it is better suited to cases where a reference corpus covering more vowel environments is available (Nolan, 1983, p. 115).

the acoustic properties of fricatives, particularly [s], [z] and [ʃ], as they scored high in perceptual speaker identification. Though their energy contours showed speaker-dependent shapes, the interspeaker variation was not significant.

As for vocalic cues, vowel formants have been studied for their values and dynamic properties within the course of vocalic articulation (see, e.g., Goldstein, 1976; McDougall, 2006). Since previous research on vowel formants has shown that elements residing within provide cues to speaker identity, Hollien included the information about vowel formant frequencies into one of the four vectors in his semi-automatic speaker identification system (SAUSI). The vowel formant tracking vector, the parameters of which derive from vowel formant frequency distribution in voiced speech – specifically, both the centre frequencies of the first three formants and their ratios F1/F2 and F2/F3 (Hollien, 1990, p. 242) - is one of the most powerful ones as it is very sensitive to speaker-specific differences and resistant to all kinds of distortion (Hollien, 2002, p. 169). Other researchers mention the F2/F1 ratio for particular vowels to be, to a certain extent, speaker-specific (e.g., Skarnitzl, 2012, in print). Vowel formant ratios are considered to convey speaker-specific information since the ratios are probably not changeable at will.

2.5.2. Temporal structuring

Speech timing appears to be another important cue to one's identity (van Dommelen, 1987; Hollien, 2002, p. 167), yet research focusing on temporal parameters is scarce. Two parameters expressing speech timing are speech rate (SR) and articulation rate (AR), differing in what is included in the calculations. While articulation rate measures the rate of speaking with all pauses being excluded from the calculation; speech rate takes both the contribution of the rate of articulation and the contribution of pausing and other fluency interruptions into account (Jessen, 2007). Henze (1953) for German and Goldman-Eisler (1968) for English (both in Jessen, 2007) consider AR to have more speaker-discriminating power than SR because it has considerably lower intraspeaker variation than SR does. This finding was later confirmed by Künzel (1997) (in Jessen, 2007) who studied both AR and SR together with several pausing parameters in German for forensic application. Butcher's study (1981, p. 148) (in Rose, 2002, p. 180), on the other hand, has shown that AR differs significantly between spoken and read speech; in other words, it exhibits high intraspeaker variation. The reliability and validity of AR for forensic purposes thus still needs to be investigated from the Bayesian point of view (Rose, 2002, p. 180). The most promising parameter in terms of conveying speaker-specific information seems to be local articulation rate (Jessen, 2007).

Hollien (2002, p. 167) includes information about temporal structuring in one of the four vectors in his SAUSI system, namely time-energy distribution vector, though it is not considered as robust as other vectors. The importance of temporal patterning in speech has also been claimed for automatic speaker recognition. Instead of traditionally used mel-frequency cepstral coefficients (MFCCs), Bocklet et al. (2007) used temporal patterns analysing different frequency bands over a longer period of time, which increased recognition rate by 12%. Other researchers suggested measuring speech rhythm by durational characteristics of consonantal (ΔC) and vocalic (%V) intervals as they appear to exhibit low intraspeaker and high interspeaker variation, though their validity as carriers of speaker-specific information has been questioned (Dellwo & Koreman, 2008).

2.5.3. Melodic parameters

Another suprasegmental parameter which is considered an indicator of speaker's identity is fundamental frequency (F0) contour. Fundamental frequency, being the acoustic representation of the rate of vocal fold vibration, is an important measure for forensic purposes since considerable information is available about its distribution amongst the adult population at large (Butcher, 2002). All speakers have a range of fundamental frequency which they habitually use and within which they feel most comfortable. F0 does not, however, reflect only anatomical differences but is also exploited for linguistic purposes (Rose, 2002, pp. 53-5), which results in constant fluctuations of its values. To avoid influences of local, short-term factors, researchers have frequently focused on a long-term mean fundamental frequency, also referred to as speaking fundamental frequency (SFF) (Hollien, 1990, p. 240), which appears to be a better representation of individual characteristics (Rose, 1991, in Rose, 2002, p. 59). Hollien (2002) likewise considers SFF a reliable indicator of speaker identity, and employs it together with other related parameters, such as F0 geometric mean, standard deviation or phonation-time ratio in his SAUSI system (Hollien, 2002, p. 165).

Even though the extraction of F0 tends to be reliable under optimal conditions, the frequency of extraction errors increases in real life conditions, such as different speaking styles or recording quality. Researchers have been trying to make the value of SFF less susceptible to extraction errors (see, e.g., Lindh & Eriksson, 2007). The measure proposed by Lindh and Eriksson (2007), alternative fundamental frequency baseline, seems to be one of the most promising achievements in this area as it appears to be more robust to different sources of variation, such as channel distortion or an emotional attitude of a speaker, than

traditionally used mean, median, or standard deviation of F0. It thus offers a more reliable representation of the neutral[10] fundamental frequency of a given speaker (Lindh & Eriksson, 2007).

2.5.4. Phonatory modulations

Apart from the three prosodic parameters, i.e. temporal, melodic and dynamic modulations, claims have been made to include a fourth one, namely phonatory modulations or voice quality. Evidence shows that it is exploited for paralinguistic purposes[11] as fundamental frequency is, but independently of it (Campbell & Mokhtari, 2003). It can signal, for instance, emotional states (Gobl & Ní Chasaide, 2003), reflect attitude towards the content of the message or the interlocutor (Campbell & Mokhtari, 2003), but it also reflects personal idiosyncrasy (Rose, 2002, p. 59). The following paragraph will give an overview of different approaches towards studying voice quality, and since it is the central topic of this paper, a more detailed account of voice quality and its acoustical correlates will be presented in the chapter to follow.

Several methods and acoustical analyses have been proposed to quantify voice quality, the most reliable of which, at least for forensic purposes, seems to be the long-term average spectrum (LTAS) (Harmegnies & Landercy, 1988; Hollien, 2002, p. 164). Investigators have been attempting to find spectral moments or relations which would reflect differences in voice quality the best. Previous research has pointed to several parameters which relate to the overall slope of the spectrum (e.g., Frøkjær-Jensen & Prytz, 1976; Hammarberg at al., 1980; Kitzing, 1986; Sundberg & Nordenberg, 2006) (see Section 3.2.1). Other researchers focused on acoustic spectra of vowels, comparing the amplitude of the fundamental with another spectral peak (Hanson, 1997) (see Section 3.2.2). Yet others studied the acoustic properties of the source signal by parameters reflecting the stability of phonation, namely jitter, shimmer and harmonics-to-noise ratio (HNR) (e.g., Schoentgen & de Guchteneere, 1995; Qi & Hillman, 1997; Kreiman & Jody, 2003; Brockmann et al., 2008) (see Sections 3.2.3 and 3.2.4).

[10] Lindh and Eriksson consider vocal fold vibration to function in a similar way as human gestures or movements, that is, there is "a point of departure, a resting position or baseline… resulting in a neutral mode and frequency of vibration to which they return after prosodic or other types of excursions in frequency" (Lindh & Eriksson, 2007).

[11] It can also have a phonological function as, for instance, in Northern Vietnamese (Rose, 2002, p. 59).

3 VOICE QUALITY

Voice quality is of interest not only in the field of phonetics itself but plays an important part in numerous disciplines. It is in fact one of the primary means by which speakers project their identity to the world; that is, their "physical, psychological, and social characteristics" (Laver, 1980, p. 2, in Kreiman & Sidtis, 2011, p. 1) or their "auditory face" (Belin, Fecteau & Bedard, 2004, in Kreiman & Sidtis, 2011, p. 1). This is reflected in the many definitions of voice quality which one can encounter, each approaching it from a perspective which is the most central for a given purpose (Kreiman, Vanlancker-Sidtis & Garrett, 2003). For the purposes of phonetics, it is relevant to investigate its physiological, perceptual and acoustic aspects.

3.1. Voice quality: definitions

In physiological terms, the approach to voice quality has traditionally been twofold; some investigators define it in its narrow sense (e.g., Gobl & Ní Chasaide, 1992; Campbell & Mokhtari, 2003), i.e. "the sound produced by the vibration of the vocal folds" (Kreiman, Vanlancker-Sidtis, & Garrett, 2003, p. 115), while others include the long-term effects of vocal tract settings, too (Master et al., 2006). Perceptually, voice quality reflects how a voice sounds. Depending on the stance taken, voice quality may thus refer to the perceptual impression created by the vocal fold vibration only or by the contribution of both the glottal source and vocal tract resonances (Kreiman, Vanlancker-Sidtis & Garrett, 2003), described by means of frameworks established for this purpose (see, e.g., Laver, 1980). However, perceptual judgements are necessarily subjective as they reflect socioeconomic and cultural aspects as well as individual preferences (Master et al., 2006). This resulted in the search for relevant acoustic methods which would supplement subjective evaluations by objective data (Hammarberg et al., 1980), and would thus allow better comparison between subjects and studies.

3.2. Voice quality: measurements

As has been mentioned above, one way of evaluating voice quality is to describe it by means of descriptive labels which reflect the perceptual impression of a listener. However, as there is no unified approach for doing so, a highly variable number of different terms is in use. Hammarberg et al. (1980), for instance, mention a study which revealed that for describing voice quality and pitch, logopedists and students of logopedy used 88 different

terms. To solve this problem, some attempts have been made to find interrelationships between these features in order to reduce them into a limited number of clusters or "factors" (p. 441). Though factor analysis reduces the redundancies, it is still subjective. Another method of assessing voice quality is called "functional" or "creative listening" (Kitzing, 1986, p. 478). In functional/creative listening, a person who assesses voice quality imitates the voice sample in question and thus experiences the voice function by one's own apparatus, which should enhance its description. However, this method is likewise subjective as it is dependent on a personal evaluation of the vocal function, and appears to be suited rather for clinical purposes.

Therefore, investigators have been trying to find an acoustic method which would reflect and quantify differences in voice quality of both subjects with organic and functional disorders as well as subjects without any disorder. The most reliable tool – at least for forensic purposes - appears to be the long-term average spectrum (LTAS) (e.g., Harmegnies & Landercy, 1988; Tanner et al., 2004). The following sections will provide an overview of methods which are used for an objective assessment of voice quality. Section 3.2.1 will examine LTAS and some common ways of its quantification. Section 3.2.2 will focus on parameters derived from acoustic spectra of vowels, which arguably reflect individual differences in glottal characteristics and voice quality. Both LTAS and the vowel spectra parameters will be mentioned in more detail since these methods will be also employed in our study as those indicators of voice quality which convey speaker-specific information. Section 3.2.3 will comment on jitter and shimmer, two parameters reflecting fluctuations in glottal cycles, and, lastly, Section 3.2.4 will discuss a parameter expressing the ratio between harmonic and noise components in voice, namely harmonics-to-noise ratio (HNR).

3.2.1. Long-term average spectrum

The LTAS is an acoustic analysis (fast Fourier transform-generated power spectrum) which "provides information on the spectral distribution of the speech signal over a period of time" (Löfqvist, 1986, p. 471). As the speech signal is the product of the sound source and the transfer function of the vocal tract which varies for different segments, using a longer sample allows the short-term variations due to phonetic structure to be averaged out (Löfqvist, 1986). The resulting spectrum is thus not influenced by the differences in speech samples but is considered to reflect the contribution of both the glottal source and the vocal tract to voice quality (Nordenberg & Sundberg, 2003; Master et al., 2006).

The LTAS has proved sensitive to different voice qualities and has thus frequently been used as an objective tool for complementing a perceptual analysis of voice quality. Moreover, it has a considerable advantage of not requiring a periodic or quasiperiodic signal to provide a reliable analysis (Tanner et al., 2005). Firstly, it was used primarily for clinical purposes in case of both organical and functional voice disorders (see, e.g., Hammarberg et al., 1980; Kitzing, 1986, respectively). Later, it was used to study acoustic differences between normal and pathological voices (e.g., Löfqvist, 1986). So far, it has been reported to reflect differences between gender (e.g. Mendoza et al., 1996; White, 2001; Nordenberg & Sundberg, 2003), age (Linville, 2002, in Master et al., 2006; da Silva et al., 2010), to be able to discriminate professional voices from untrained ones (Leino, 1993, in Master et al., 2006), to assess various features of dysphonic voices (Hammarberg et al., 1980) and reflect voice improvement after therapy (Kitzing & Åkerlund, 1993; Tanner et al., 2005). As for the use for forensic purposes, long-term spectra vector forms a part of Hollien's SAUSI system (2002) and is considered sensitive to speaker identity even under unfavourable conditions, such as noise, limited passband or speaker stress (Hollien, 2002, p. 162).

Previous research has suggested several parameters which allow quantification of the LTAS. These indicate the overall slope of the spectral envelope since spectral tilt has been directly related to voice quality (e.g., Hammarberg et al., 1986; Leino, 1993, in Master et al., 2006) by comparing particular spectral peaks or by the ratio of the amount of energy in certain frequency bands as discussed in the following paragraph.

From a psychoacoustic point of view, a sonorous voice as opposed to a dull or husky voice should be reflected in the LTAS in higher energy in the harmonics, that is, less steep spectral tilt (Löfqvist, 1986). This is what numerous studies indeed show. Hammarberg et al. (1980), for instance, found a significant correlation between voices perceived as breathy and the slope of LTAS; more precisely, a steep decrease of spectral level between bands 0-2 kHz and 2-5 kHz. The so-called *Hammarberg index* which expresses the difference between the maximum energy in the 0-2 kHz region and the 2-5 kHz region has been claimed to distinguish not only different voice qualities (Hammarberg et al., 1980) but also different speech styles (Monzo et al., 2007). A lower concentration of energy above the first-formant region and a higher concentration of energy above 5 kHz has also been claimed to relate to breathy voice quality or a hypofunctional voice by other studies (Soyama 2005, in Master et al., 2006).

Many researchers have been since trying to find relations of peaks or regions in the spectral tilt which would allow quantification of perceived voice quality. One of the most

commonly employed ones is *alpha index* (α) suggested by Frøkjær-Jensen & Prytz (1976), which is defined as the ratio of intensity above and below 1000 Hz (1-5 kHz/ 0-1 kHz). Alpha index is claimed to be a potent criterion for distinguishing voice quality by numerous researchers (see, e.g., Löfqvist, 1986; Sundberg & Nordenberg, 2006; Leino, 2008). Kitzing (1986) used α as a basis for a new parameter, which we therefore call *Kitzing index*. Firstly, he found inverted α, i.e. ratio of intensity below and above 1000 Hz, likewise a reliable tool for differentiating of voice qualities. However, since not all voice qualities were discriminated (arguably due to the non-systematic variations of energy in the 3-4 kHz range of the spectra influenced by the resonances of the vocal tract rather than the source signal), *Kitzing index* expresses the ratio of energy in the 1-2 kHz frequency range of spectra as opposed to the 0-1 kHz range (0-1 kHz/ 1-2 kHz). Apart from these, several other measures of spectral slope are in use, such as the ratio of 0-1 kHz/1-6.5 kHz which accounts for data of all voice-source characteristics but excludes possible higher frequency effects of, for instance, plosives or fricatives, and the ratio of 0-1 kHz/1-20 kHz which includes all auditory data that a listener would be likely to hear (Sergeant & Welch, 2008).

Measuring the difference between the f0 and the first-formant amplitude, referred to as L1-L0, has been likewise shown to provide information about the phonation mode (Sundberg, 1987, in Master et al., 2006; Kitzing, 1986). L0 stronger than L1 indicates a breathy or weak intensity voice, while a more tense, strong intensity voice is reflected in L1 being stronger than L0 (Master et al., 2006).

As has been mentioned above, not only the ratio of the amount of energy in certain frequency bands or of particular peaks, but also individual peaks have been related to perceived voice quality. For instance, Sundberg (1987) (in Master et al., 2006) identified a peak between 2.8 – 3.4 kHz when studying voices of lyric singers by the LTAS, hence "singer's formant". This peak forms by grouping of F3, F4 and F5 and is related to the perception of projected voices. This finding has been later confirmed also by, for instance, Leino (1993) (in Master et al., 2006) who studied voices of male actors and identified a "speaker's/ actor's formant".

Ternström (2008) used the LTAS to investigate the distribution of spectral energy in the regions above 5 kHz, which has traditionally been the threshold in speech research as it is in this region that most energy is concentrated. However, though disregarding spectral energy in frequencies above 5 kHz does not impair intelligibility of speech, this range is audible and important for perception. Ternström's data reveal that the LTAS contour (after removing voiceless sounds) is "quite personal" (Ternström, 2008, p. 3) even in the frequency range 5-20

kHz. Lu & Dang (2008) likewise report speaker discriminative information in higher frequencies and have employed it for enhancing speaker recognition. A better understanding of distribution of energy in higher frequencies could also, for instance, enhance naturalness of synthetic speech (Ternström, 2008).

Though the usefulness of long-term measures dwells in factoring out the contribution of individual sounds to acoustic parameters, which results in gaining an overall value for a speaker, long-term measures are just like all other measures "never totally inert to real-world factors" (Rose, 2002, p. 59). For instance, Nordenberg & Sundberg (2003) in their study showed that comparing data which were produced in different degrees of loudness can be questioned due to the fact that the frequency response for the same intensity increase is not linear. Specifically, an increase of vocal loudness causes a greater gain in higher frequencies (1500-3000 Hz) than in lower frequencies. This motivated the authors to study the effects of vocal loudness on the LTAS in more detail. Sundberg & Nordenberg (2006) demonstrated that an increase of the level *can* be approximated by certain functions but interindividual variation exists. The question of reliability of the LTAS has also been addressed by Löfqvist (1986) who points out that intraspeaker variability in the LTAS, which was in his study quantified by the ratio of energy between 0-1 kHz and 1-5 kHz, can be considerable. His subjects were people using their voice constantly during the day as a result of which the spectral tilt of their LTAS was markedly steeper at the end of the day.

The results obtained by the LTAS in our study will be discussed in Section 6.3.

3.2.2. Vowel spectra parameters

Another group of parameters which are considered to reflect individual differences in voice quality is derived from the acoustic spectra of vowels. These parameters quantify variations in glottal characteristics which lead to different voice qualities. Previous research has studied glottal characteristics by inverse filtering (e.g., Gobl & Ní Chasaide, 1992, who measured the differences between the fundamental and the first four formants), or by visual inspection of vocal folds (Södersten et al., 1991, in Hanson, 1997), which is necessarily invasive. Ní Chasaide & Gobl (1993, in Hanson, 1997) extracted glottal parameters both from the glottal waveform and vowel spectra. In their study, an increase in glottal adduction has been found to result in a steeper spectral slope of the vowel spectrum and in lower formant amplitudes (especially F1). A further study by Holmberg et al. (1995, in Hanson, 1997) has shown that the relative amplitudes of the first two harmonics (computed from the glottal

waveform) relate to adduction quotient and that the relative amplitude of the first and third formant peaks tends to reflect the speed of glottal closure.

Hanson (1997) suggested a set of parameters which take the effect of the vocal tract filter into consideration, which allows in addition for examining the effect of the glottal source on filter (its bandwidths), thus providing further information about the glottal configuration (Hanson, 1997). In contrast to previous research (e.g. Hillebrand & Cleveland, 1994), the vowels studied are derived from carrier sentences, which offers a more natural data than a sustained vowel production (Löfqvist, 1986) .

In the following paragraphs, the parameters suggested by Hanson (1997) will be explained in more detail since this study aims to test them as possible indicators of speaker identity.

These measures are made directly on the acoustic spectra of vowels and thus give some indication of the vocal-fold and glottal configuration during a vowel production. The parameters compare amplitudes of spectral peaks; the fundamental, that is, the first harmonic (referred to as H1), which has been shown to correlate with the vibratory amplitude of the glottis (Gauffin & Sundberg, 1980 in Kitzing, 1986), and another spectral moment. These include the amplitude of the second harmonic (H2), the amplitude of the first formant (A1), and the amplitude of the third formant, (A3) (Hanson, 1997). The comparison of these peaks thus yields threes parameters; namely, H1-H2, H1-A1 and H1-A3.

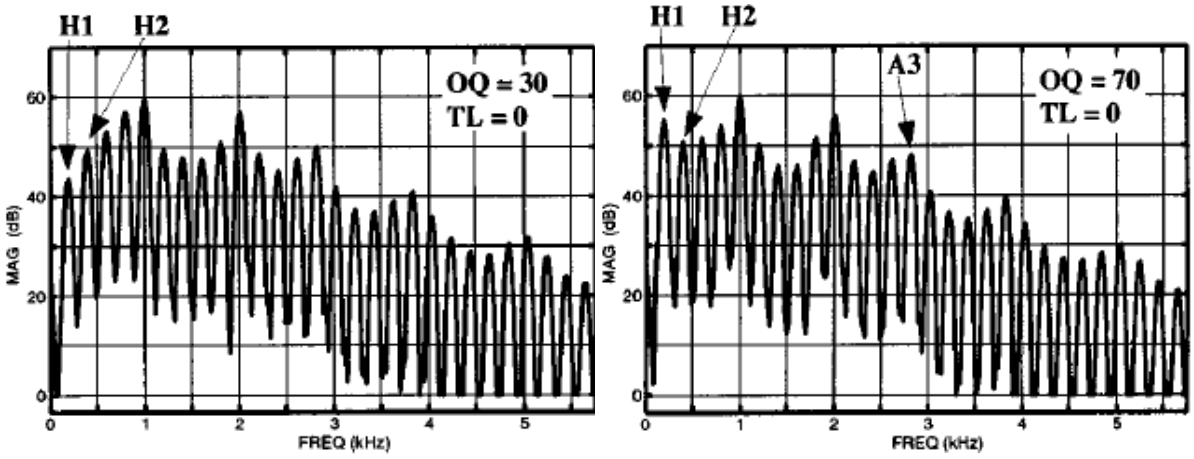


Fig. 1 Spectra of a synthesized vowel /æ/ using different glottal configurations. On the left, the open quotient is 30%, while on the right, it is 70%. The relative amplitude of the first and the second harmonics (marked H1 and H2, respectively) changes as a result of a different glottal configuration, specifically, a different open quotient.

(Adapted from Hanson, 1997, p. 468).

The parameter H1-H2, by comparing the amplitude of the first and the second harmonics, has been shown to provide an indication of open quotient (Hanson, 1997). Its computation is illustrated on a synthesized vowel /æ/ in English in Figure 1. Both pictures show a spectrum of the vowel /æ/ but they differ in the open quotient (OQ). On the left, the open quotient is 30% while on the right, it is 70%. We can see that the relative amplitude of the first two harmonics changes by about 10 dB. If the differences across vowels are desirable to be minimized, corrections of both H1 and H2 can be made (Hanson, 1997). Hanson (1997) reports a range of 10 dB for this parameter.

H1-A1 is considered an indication of the first-formant bandwidth, and can also provide some information about the degree to which the glottis fails to close completely during the closing phase of the glottal cycle (Hanson, 1997). Formant bandwidths relate to acoustic energy losses in the vocal tract coming from several sources, such as the resistance of the walls of the vocal tract. When the glottis is open even during the closing phase and air flows it, further energy losses are introduced, especially in lower frequencies. Since the amplitude should be proportional to the inverse of the bandwidth, that is, a larger bandwidth results in a reduced peak, the first-formant peak amplitude relative to that of the first harmonic is considered to provide an indication of the first-formant bandwidth and hence the degree to which the glottis fails to close completely. Figure 2 shows two spectra of the English vowel /æ/ derived from waveforms differing in the first-formant bandwidth. The spectrum on the left is derived from a waveform with a narrower first-formant bandwidth; the spectrum on the right is derived from a waveform with a wider one.

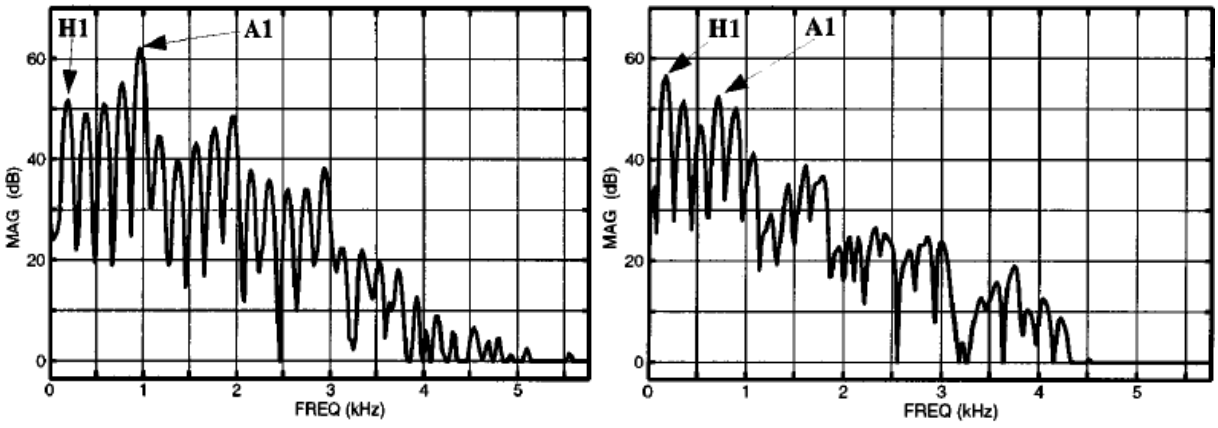


Fig. 2 Spectra of the English vowel /æ/ derived from waveforms differing in the first-formant bandwidth. The spectrum on the left has been derived from a waveform with a narrower bandwidth, while the spectrum on the right has been derived from a waveform with a wider first-formant bandwidth. H1 = the amplitude of the first harmonic, A1 = the amplitude of the first-formant peak. While a narrower bandwidth is reflected in a more prominent A1 (on the left), a more damped first-formant peak (A1) corresponds to a wider bandwidth (on the right).

(Adapted from Hanson, 1997, p. 471)

However, since the amplitude of the first formant is compared to that of the first harmonic, the value of this parameter will be influenced also by the variation in the amplitude of the first harmonic across speakers (Hanson, 1997). Hanson (1997) reports a range of 16 dB for this parameter, the lowest value being -11 dB and the highest 5 dB.

Lastly, the parameter H1-A3 is considered to provide some information on spectral tilt. The spectrum at middle and high frequencies is influenced by the abruptness with which the flow of air is cut off during the closing phase of a vibration cycle. A more gradual cut off due to, for instance, a non-simultaneous vocal fold closure, results in an additional downward tilt. This is illustrated in Figure 3. It shows spectra of a synthesized vowel /æ/ with the same open quotient (70%) but synthesized using different glottal characteristics as to the abruptness of glottal closure. The spectrum on the left shows a more abrupt glottal closure which is reflected in the higher amplitude of the third-formant peak (marked as A3) in comparison to the spectrum on the right. There are two main causes of this effect. A glottal closing may not be simultaneous at all points along the anterior-posterior length of the vocal folds. This kind of 'zipper' closing leads to a more gradual cut-off, which results in steeper spectral tilt. The effect on the spectral tilt in the third-formant region is still higher when the closure is incomplete due to, for instance, a glottal chink. Since the amplitude of the third formant is dependent on both the location of F1 and F2 and on F3 bandwidth, if the results are to be comparable across vowels and speakers, normalization of the values is recommended (Hanson, 1997). The maximal value measure by Hanson (1997) is 35 dB and the minimal 8.6 dB.

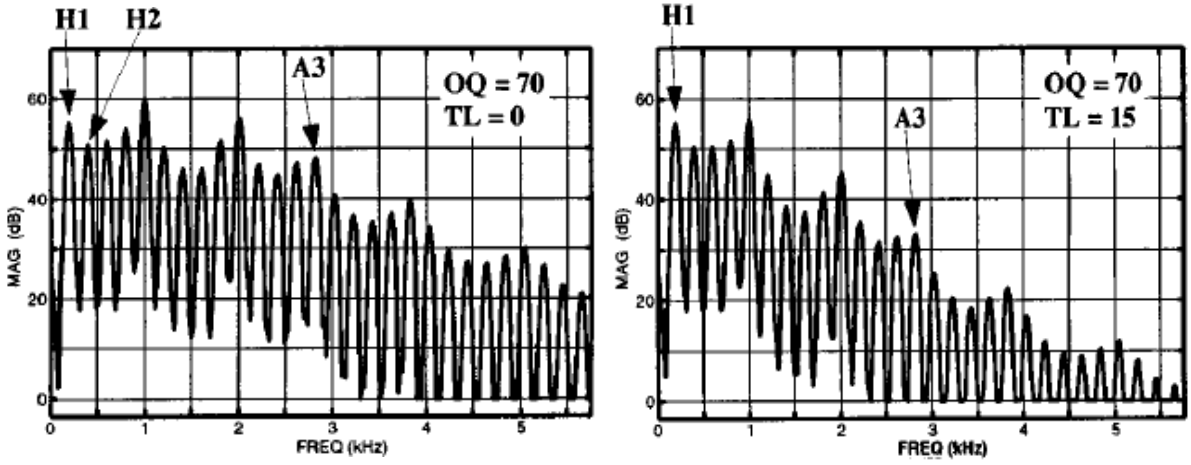


Fig. 3 Spectra of a synthesized vowel /æ/ with the same open quotient, i.e. 70%, (OQ = 70) but synthesized using different glottal characteristics as to the abruptness of glottal closure. H1 = the amplitude of the first harmonic, A3 = the amplitude of the third-formant peak. A less abrupt closure (on the right) introduces additional downward spectral tilt (TL = 15 in contrast to the spectra on the left, where TL = 0), which is reflected in a lower amplitude of the third-formant peak (A3).

(Adapted from Hanson, 1997, p. 468).

Some relationship between these parameters can be predicted by theory. For instance, if the glottis does not close completely during a vibration cycle, the airflow causes an increase in both F1 bandwidth (quantified by H1-A1) and spectral tilt (quantified by H1-A3). Though a larger open quotient (quantified by H1-H2) is expected to lead to greater losses (i.e. an increase in both H1-A1 and H1-A3), it has not been found to correlate with either of the measures (Hanson, 1997) suggesting that open quotient is independent of other glottal parameters. Table 1 displays correlations of the three parameters for all the three non-high vowels combined, namely /æ, ʌ, ɛ/, which Hanson (1997) used in her study. An asterisk indicates that the values have been normalized for the effect of vowel quality.

	$H1^*-H2^*$	$H1^*-A1$	$H1^*-A3^*$
$H1^*-H2^*$	1		
$H1^*-A1$	0.53	1	
$H1^*-A3^*$	0.46	0.68	1

Table 1 Pearson product moment correlation coefficients (*r*) for the acoustic parameters for the three vowels /æ, ʌ, ɛ/ combined. An asterisk indicates a normalized value.

(Adapted from Hanson, 1997, p. 478)

Since a strong correlation is a correlation with ($r > 0.70$), the correlation between H1*-A3* and H1*-A1 can be considered good. Other correlations are weaker and the parameters thus seem to be independent of one another. By plotting H1*-A1 and H1*-A3* against each other, Hanson (1997) distinguished two groups of speakers, namely those supposedly having abrupt glottal closure, which is reflected in lower values of the two parameters, and those having non-simultaneous and/or incomplete glottal closure, which is reflected in higher values. Her assumptions were confirmed by direct observations of vocal folds. These parameters were thus proposed to reflect individual differences in glottal characteristics.

In the present study, it will be examined to what extent they can be considered speaker-specific cues (see Sections 6.1 and 6.2).

3.2.3. Jitter and shimmer

As Frøkjær-Jensen & Prytz (1976) commented, since voice quality is an auditory property, just as important as the acoustic structure of the speech spectrum are cycle-to-cycle variations in pitch. Since the LTAS does not reflect the time domain, other parameters are often used to complement it. Two methods which quantify irregular vocal fold vibrations are

jitter and shimmer; the former measuring the fluctuations in glottal cycle lengths and the latter in its amplitudes (Brockmann et al., 2008). These are typically measured on sustained vowels as in connected speech voluntary perturbations are exploited to produce voicing and prosodic cues, and their relative contribution to jitter or shimmer cannot be factored out (Schoentgen & de Guchteneere, 1995). Jitter and shimmer have been related to perceived voice quality of roughness and hoarseness (Yumoto et al., 1982; Dejonckere et al., 1996, in Brockmann et al., 2008), though their reliabiliy and validity for clinical purposes has been questioned as they require periodic or quasiperiodic signal and are measured from sustained vowels which tend to be mildly affected in contrast to connected speech (Tanner et al., 2005, who reports a perceived change after a therapy to be the best reflected by the LTAS, particularly spectral mean and standard deviation). Jitter and shimmer has been studied also in healthy adults (e.g., Brockmann et al., 2008) and for its applicability in automatic affect recognition (Fernandez & Picard, 2005), but the extent to which jitter and shimmer convey speaker-specific information remains relatively unexplored.

3.2.4. Harmonics-to-Noise Ratio (HNR)

Apart from the above mentioned measures which relate to the stability of phonation, the amount of noise components in voice has likewise proved relevant for the perception of voice quality (Kreiman & Garrett, 2003). The amount of noise relative to harmonic components is quantified by harmonics-to-noise ratio (HNR) (Qi & Hillman, 1997). Though HNR has been claimed to be a sensitive index of vocal aging (Ferrand, 2002) and a useful tool for clinical purposes as an acoustical correlate of perceived hoarseness (Yumoto, Sasaki & Okamura, 1982), the degree to which it expresses individual differences and its applicability for forensic purposes has likewise not been addressed.

Michaelis et al. (1997) proposed an alternative to HNR, namely glottal-to-noise excitation ratio (GNR). This parameter indicates whether a voice signal originates from the vibrations of the vocal folds or a turbulent noise; thus being related to the degree pf breathiness. Their study which is based on artificial signals suggests that in contrast to HNR, GNR is almost independent of jitter and shimmer. Its use, however, also appears to be limited to clinical purposes (Michaelis, Gramss & Strube, 1997).

4 AIMS OF THE PRESENT STUDY

Before proceeding from the theoretical to the experimental part, the aims of the present study should be repeated. From the above discussion it is apparent that the use of voice quality; specifically, the parameters relating to voice quality or phonatory modifications, for forensic purposes remains relatively unexplored. These parameters are likewise missing in the procedures used in forensic investigations in the Czech Republic (personal communication of Mgr. Radek Skarnitzl, Ph.D. with Marie Svobodová, Ph.D). The present study therefore aims to examine to what extent spectral properties of the source signal can be considered an indication of speaker identity. This will be done by focusing on short-term measures of spectral tilt suggested by Hanson (1997) (see Section 3.2.2) which will be complemented by long-terms measures, i.e. the LTAS parameters, namely, alpha index, Hammarberg index and Kitzing index (see Section 3.2.1).

The primary aim is to examine how these parameters expresssing spectral tilt discriminate 16 Czech female speakers. To do so, we will explore in more detail the parameters reflecting short-term spectral tilt suggested by Hanson (1997), as they are claimed to reflect individual differences in glottal characteristics. Hanson (for English) included in her material only non-high vowels, which were embedded in the phrase "Say bVd again" (Hanson, 1997, p. 475); 'V' standing for 'vowel'. The phonemic environment was thus kept constant. Her subjects read this utterance fifteen times; that is, five times for each vowel quality. Her sample therefore consisted of stressed syllables in an utterance-non-final stress group. In addition, the values were normalized to minimize differences across vowels. In the present study, we used a continuous text instead of a carrier phrase to obtain for each speaker as natural a sample as possible. We also used a more extensive data set (15 instances of each vowel quality for each speaker) and took vowel quality, syllable status with respect to stress and stress group position in the utterance into account, which enabled us to study their effect on the parameters. These results will be at the end complemented by and compared with the results of the LTAS, i.e. the long-term measures of spectral tilt.

The present paper would like to contribute to the ongoing research of speaker-specific cues by focusing on parameters reflecting phonatory modifications or voice quality.

Based on the discussion in Section 3.2.2, we can expect vowel quality to have some effect on the parameters due to the different location of formants for individual vowels. The effect of stress and stress group position in the utterance can likewise be expected due to varying vocal effort. Yet this study is designed as exploratory in nature and no specific

working hypotheses are going to be tested. A possible null hypothesis would claim that there are no differences in the parameter values for individual speakers. If the null hypothesis is falsified, the study will attempt to point out relations which could be further examined in a future study.

5 METHOD

In this chapter we will first describe the sample which was used for the present research (Section 5.1) and then explain how the measures of parameters that we intend to study were obtained (Section 5.2).

5.1. Speakers and speech material

The material for the analyses was taken from one part of the Prague phonetic corpus (Skarnitzl, 2010), which contains 80 short read dialogues constructed to convey specific phonetic phenomena, but applicable for various phonetic analyses. These dialogues consist of three to five turns read by 25 pairs of speakers with each pair recorded twice so that every speaker reads all turns. Before the recording, speakers were asked to read the dialogues in order to become acquainted with them. During the second reading, they were asked to "act it out". The speech material was recorded in a sound-treated recording studio. Our subjects were also instructed not to change their loudness to avoid its interference with data (Hammarberg et al., 1980; Sundberg & Nordenberg, 2006, for effects of vocal loudness variation on the LTAS).

For the present study, 8 pairs reading the same set of dialogues were selected at random but with the condition that both speakers are female. This decision was motivated by the considerable gender differences in glottal configuration; specifically, the fact that during normal phonation, females are more likely to have an incomplete closure of the vocal folds than males (e.g., Linville, 1992, in Hanson, 1997). Extracting data for an analysis from the same text for all speakers allowed us to account for an additional variation which would be introduced by different phonemic contents. All the sixteen female speakers were in their first years of study of linguistic programs at the Faculty of Arts and native speakers of Czech.

In order to separate the utterances of the two speakers in each pair, the beginning and the end of each turn has been labelled using the Praat software (Boersma & Weenink, 2010), and cut into individual turns with the help of scripts written for this purpose. Each acquired item thus involved one speaker only. Subsequently, by means of the Prague Labeller (Pollák et al., 2007), boundaries of all segments were automatically detected using Hidden Markov Models (HMMs). The boundaries of target segments necessary for the vowel spectra analysis were then adjusted manually following the suggestions presented in Machač & Skarnitzl (2009).

As for the material for the vowel spectra analysis, the target segments consisted of 75 vowels for each speaker, i.e. 15 instances of each of the five short vowels in Czech, /a, e, ɪ, o, u/. Several criteria had been observed before the final set of 75 vowels for each speaker was selected. Firstly, only autosemantic words were considered since synsemantic words are more likely to undergo reductions (Johnson, 2004). Since one of the objectives was to compare the robustness of the above mentioned parameters for individual vowels, quality reductions would hinder such a comparison. Secondly, the phonemic environment was considered. Vowels followed by a palatal consonant or a liquid were automatically disregarded due to the influence of these consonants on vowel formant frequencies; specifically, lowering the frequency of the first formant. Vowels followed by /ɦ/ were likewise taken out of consideration as the following glottal fricative could introduce additional breathiness which would interfere with the observed data. The last criterion was syllable status with respect to stress. The total number of 15 instances of each vowel quality was balanced for the position of stress. To give an example, 5 instances of the vowel /a/ were in a stressed syllable, other five in a post-stress syllable and the last 5 instances appeared two or more syllables after stress. This was motivated by the fact that in Czech, stress is always on the first syllable of autosemantic words, but certain prosodic features are realized on the following, post-stress syllable, such as f0 movement (a decrease on the first syllable, an increase on the second). Since all our parameters measure the difference of some spectral peak from the fundamental, the two types of unstressed syllables were differentiated (Palková & Volín, 2003). In addition, we observed whether the selected vowel appears in an utterance-final stress group or not, though it was not possible to balance these two groups; the number of vowels in utterance-non-final stress groups is higher (654 cases) than in utterance-final stress groups (539). The decision of distinguishing the two was motivated by the fact that in utterance-final stress groups, vocal tension tends to decrease and creaky or breathy phonation is often present, which could interfere with the measurements. All 75 vowels, i.e. 15 instances of each of the 5 vowels in Czech, selected for the analysis were the same for all 16 speakers. The complete text of the read dialogues used for the present study is enclosed in the Appendix.

As for the material for the LTAS analysis, it was necessary to create a long enough string of utterances for each speaker to yield a mean spectrum which is not greatly affected by any differences in the speech material. The frequencies of the first two formants, which exhibit a larger variation between the vowels, thus become represented by an average,

"evidencing the formants with less variable values - F3, F4 and F5 - that are related to the voice quality" (Sundberg, 1987, in Master et al., 2006).

Though the duration of samples used in previous studies ranged from mere few seconds up to 3 minutes (Sergeant & Welch, 2008), the generally accepted sample length in most recent studies seems to be between 20 and 40 seconds (Kitzing, 1986; Löfqvist, 1986; White, 2001; Master et al., 2006). A speech sample shorter than 20 seconds has been reported to yield a spectrum which is text-dependent (Harmegnies & Landercy, 1988).

To create this string, as many turns spoken by a single speaker and the same for each speaker to further minimize variations due to the phonemic content (cf. Löfqvist, 1986) were concatened by Praat so that even the shortest string would be at least 40 seconds long (Fritzell, 1974, in Nordenberg & Sundberg, 2003). 16 strings, one for every speaker, ranging from 40 to 55 seconds (depending on the speaking rate of individual speakers) were thus obtained.

There has been a debate on whether or not to exclude voiceless sounds from an LTAS analysis. Those who recommend excluding them (e.g., Hammarberg et al., 1980; Löfqvist, 1986; Kitzing, 1986) claim that voiceless sounds might "corrupt the averaging of data of voiced segments and mask information of the voice source" (Löfqvist & Mandersson, 1987, in Sergeant & Welch, 2008, p. 660) as the high-frequency noise might be undistinguishable from the noise of the voice source (Hammarberg et al., 1980). The opposite view is that removal of unvoiced sounds would yield an incomplete analysis (Sergeant & Welch, 2008). Undoubtedly, this decision depends on the subject of study. For studying the phonation mode for clinical purposes, removal of voiceless sounds appears to be desirable (Hammarberg et al., 1980). However, for our present purposes, since the text for all speakers was the same (White, 2001) and voiceless sounds seem to increase the LTAS level only above about 5000 Hz (Löfqvist, 1986; Tanner et al., 2004; Sundberg & Nordenberg, 2006; Ternström, 2008), we decided to include all sounds in the analysis.

5.2. Measurements

In the section to follow (Section 5.2.1), the extraction of the above mentioned vowel spectra parameters will be described and complemented with illustrative examples, after which the LTAS parameters will be considered (Section 5.2.2). All analyses were undertaken in Praat.

5.2.1. Extracting vowel spectra parameters

Out of the total number of 1200 vowels (75 vowels, i.e. five instances of each of the five Czech short vowels in 3 different conditions, namely in a stressed, post-stress and unstressed syllable, multiplied by 16 speakers), 7 had to be eliminated from the sample. The total number of vowels for our analysis was thus 1193. The reason for removing the vowels from the analysis was either that the vowel was not pronounced (either omitted or a syllabic consonant was pronounced instead) or the formants were not discernible from either the spectrogram or the spectrum.

The three acoustic parameters described in Section 3.2.2, that is, H1-H2, H1-A1 and H1-A3, were then manually extracted from all vowels chosen for each speaker using Praat software. With the help of a script, the middle 20 milliseconds of each vowel were selected to obtain stable formant values and from this selection all data obtained. The following paragraphs will describe the extraction of values in more detail, for each parameter separately.

H1-H2

The parameter H1-H2 expresses the difference between the amplitudes of the first and the second harmonics (H1 and H2, respectively; see Section 3.2.2). Its measurement was usually rather straightforward since all data were drawn directly from the spectra of the middle 20 milliseconds of each vowel as illustrated in Figure 4, where the first peak corresponds to the first harmonic and the second peak to the second harmonic. The value of this parameter was thus obtained by subtracting the amplitude of the second peak from that of the first peak in decibels.

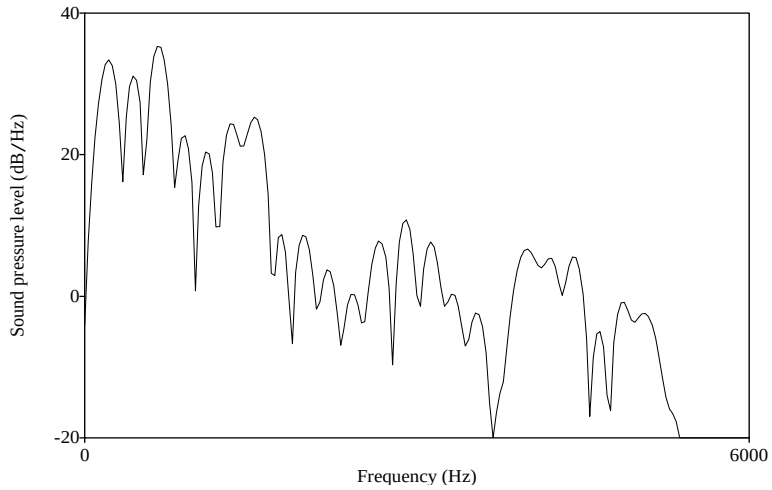


Fig. 4 A spectrum of the vowel /a/ with clearly separated peaks of the first and the second harmonics. The first peak of the spectrum corresponds to the amplitude of the first harmonic and the second peak corresponds to the amplitude of the second harmonic (the second peak is of a double frequency of the first peak).

In a small number of cases, the first two harmonics (or, alternatively, the second and the third) created a single peak which was apparent from the fact that what appeared to be the second harmonic, that is, the second peak, was of a triple frequency than the first peak. This was mostly solved by shifting a boundary of the relevant vowel by a half of the glottal cycle which resulted in a separation of the two peaks. In those instances where this correction did not help, the difference between the first two discernible peaks was counted. However, since this was the case in no more than 5 instances out of the total number of 1200 measurements, the possible influence on the results is negligible.

H1-A1

The parameter H1-A1 expresses the difference between the amplitude of the first harmonic and the first-formant peak (A1). While obtaining the value of the first harmonic was straightforward as discussed above, identifying A1 was more complex. Prior to extracting the values from the spectra, the spectrogram was always inspected. Firstly, the automatically extracted first-formant value from the selection (the middle 20 milliseconds) by Praat was checked. If it was in agreement with the perceived vowel quality and the spectrogram, the peak of such a harmonic that was the closest in frequency to the automatically extracted F1, was taken for A1. To give an example, if an automatically (and correctly) extracted first-formant frequency was 550 Hz and there were two peaks in the spectra, one around 400 Hz and the other one around 600 Hz, the second one, by virtue of its being closer to the measured value, was selected as the first-formant peak. To get the value for this parameter, its amplitude was then subtracted from the amplitude of the first harmonic.

In many cases, the first-formant peak, defined in this way, was identical with H2 since the second harmonic lay the closest in frequency, and seldom even with H1. This happened especially in closed vowels since their F1 is the lowest. Hanson (1997) avoided this by considering only non-high vowels since “the first formant is well separated from the first harmonic, simplifying the acoustic measurements” (p. 475). However, in a considerable number of instances, the second harmonic was the closest one to F1 even for the Czech open vowel /a/, which is in agreement with the findings of Skarnitzl & Volín (submitted) who have shown that nowadays, the values of F1 only seldom reach the ‘expected’ values as defined by Palková (1997, p. 174) but rather tend to be lower. Due to this fact, unlike Hanson (1997), who used for her measurements “the amplitude of the strongest harmonic of the F1 peak” (p. 475) because it was well separated from the first harmonic, we took the amplitude of the closest harmonic in frequency to the automatically extracted value unless this value was

detected in a wrong way. If the detection was faulty, we extracted A1 manually on the basis of visual inspection of the spectrogram.

H1-A3

The last parameter, H1-A3, expresses the difference between the amplitude of the first harmonic and the third-formant peak (A3). As in the case of obtaining A1 discussed above, the value of the third formant was firstly automatically extracted from the given selection after which (if the extraction was correct) the closest harmonic was identified in the spectrum, its peak measured and the value subtracted from that of the first harmonic.

The complications with obtaining the value of A3 were threefold. Firstly, in some cases, no F3 was detected by Praat. This was mostly solved by changing the default settings of the maximum formant frequency for 3 formants from 3300 Hz to 3500 Hz. If this modification did not result in the detection of F3, the third formant value was manually extracted on the basis of visual inspection of the spectrogram. Secondly, some automatically detected third formant values fell right in between two harmonics in the spectra, that is, both harmonics had the same chance of being selected as A3 by their virtue of having the same frequency distance from the third formant. If this happened, the selection was slightly extended, i.e. to include more than the 20 milliseconds, which resulted in the fact that one of the harmonics was a better candidate than the other one.

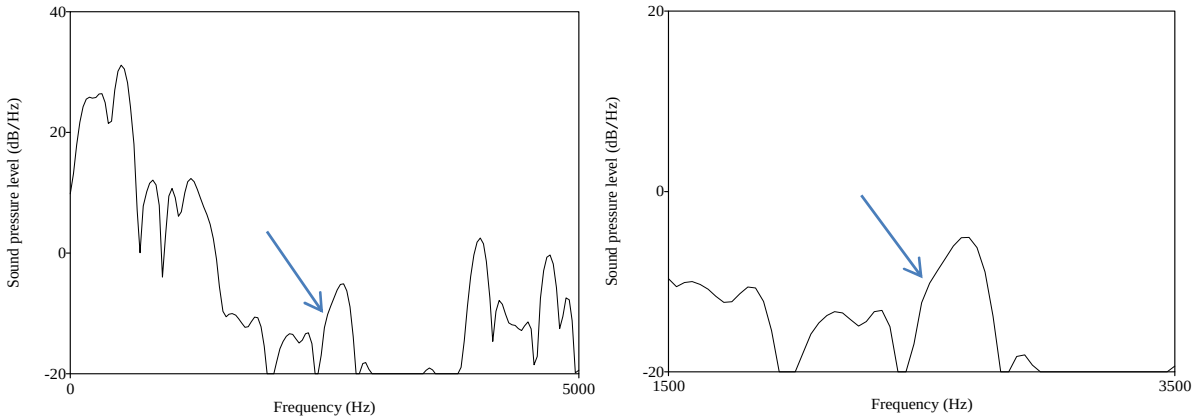


Fig. 5 Spectra of the vowel /u/ showing two harmonics 'melted' in a single peak. On the left: a frequency range of 0 - 5000 Hz; on the right, the same spectrum but with a smaller frequency range, 1500 – 3500 Hz. The amplitude of the harmonic closest in frequency to the measured F3 (marked with an arrow in both pictures) is 'melted' with the next harmonic.

Lastly, as was already mentioned in relation to the other parameters, sometimes two harmonics created a single peak, which is illustrated in Figure 5 and marked with an arrow. On the left, we can see a spectrum with a range of 0 – 5000 Hz, where the harmonic closest to F3 is in its centre. The picture on the right offers a detailed view of the same peak. Since the

harmonic closest to the third formant value (where the arrow points) creates a single peak with the following harmonic, the value of the whole peak (corresponding to the amplitude of the following amplitude of higher intensity) was considered. The general rule in cases like these was that if the amplitude of the sought harmonic is not discernible due to its being subsumed under another peak, the other peak is considered and taken for A3.

5.2.2. LTAS parameters

The values of the LTAS parameters for individual speakers were obtained from the strings of utterances described in more detail in Section 5.1. Firstly, the LTAS for each speaker was obtained by Praat and visually inspected for any interesting features, such as an unusually high or low amount of energy in some frequency band. Secondly, by means of scripts written for this purpose, the three parameters discussed in Section 3.2.1, namely alpha index, Hammarberg index and Kitzing index, were computed. For each of the 16 speakers we thus obtained three values, one for each parameter.

6 RESULTS

The results of our study will be provided in this section. Our primary aim was to examine the robustness of the short-term measures of spectral tilt as speaker-specific cues. We did so by first examining the influence of individual independent variables on the parameters by means of analysis of variance (ANOVA), and subsequent linear discriminant analysis (LDA). Section 6.1 will present the most general results of ANOVA and an overview of variables. The following three sections will discuss the influence of the independent variables on the parameters. These findings were then used for conducting LDA, the results of which will be provided in Section 6.2. Section 6.3 will then compare the results of LDA with the results of long-term measures of spectral tilt obtained by LTAS.

6.1. Short-term measures of spectral tilt

All our parameters were subjected to the analysis of variance and were found to reflect statistically highly significant differences between speakers. A one-way ANOVA for the factor SPEAKER has shown its effect on all three parameters, i.e. H1-H2, H1-A1 and H1-A3 (F (15, 1177) = 11.083; $p < 0.001$; F (15, 1177) = 13.953; $p < 0.001$, and F (15, 1177) = 11.758; $p < 0.001$), respectively). Since speakers were found to differ in the values of our parameters, the influence of individual independent variables and their possible interaction was then studied in more detail in order to discover under which conditions the discriminative power of these parameters is the strongest.

To sum up, we had three dependent numerical variables, namely *H1-H2*, *H1-A1* and *H1-A3*, and four independent categorical variables, namely *speaker*, *vowel*, *syllable status with respect to stress* and *stress group position in the utterance.* As mentioned in Section 5.1, there were 16 speakers, 5 vowel qualities (/a, e, ɪ, o, u/), 3 syllable statuses with respect to stress (stressed, post-stress, unstressed) and two stress group positions in the utterance (final and non-final). Since we were not interested in individual values of the parameters but in their robustness as speaker-specific cues, that is, their applicability for discrimination of speakers, the individual values will not be mentioned here.

The following section will discuss the influence of the above mentioned independent variables on H1-H2. Section 6.1.2 will examine their influence on the values of H1-A1 and Section 6.1.3 on H1-A3; thus providing the results for each parameter separately.

6.1.1. H1-H2 (an indication of open/adduction quotient)

Speakers were found to differ in their values of H1-H2 and these differences proved statistically highly significant (F (15, 1177) = 11.083; $p < 0.001$). The H1-H2 values for individual speakers are illustrated in Figure 6.

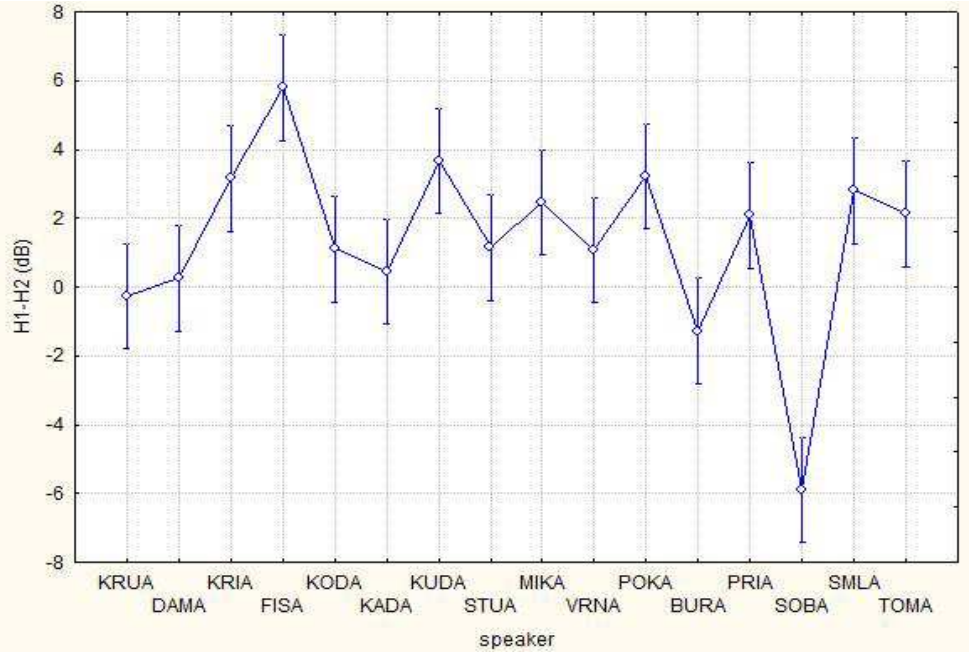


Fig. 6 H1-H2 values in decibels for individual speakers. (Error bars indicate 95% confidence intervals.)

We can see that speaker SOBA differs in her H1-H2 values from other speakers the most. A subsequent post-hoc test has shown that all these differences are statistically highly significant (Tukey HSD post-hoc test: $p < 0.001$), apart from a comparison with BURA, where $p < 0.05$. There are also other speakers who contribute to this effect; for instance, FISA who differs significantly from 8 speakers ($p < 0.001$ for 5 comparisons, for the remaining three $p < 0.05$) and BURA from 6 speakers ($p < 0.001$ for two 2 comparisons; $p < 0.05$ for the remaining four).

We were then interested whether *syllable status with respect to stress* has some influence on the values of H1-H2 for individual speakers; in other words, whether speakers differ in stressed, post-stress as well as unstressed syllables. The results of the three analyses are presented in Table 2.

H1-H2	ANOVA
stressed	F (15, 383) = 5.6644; p < 0.001
post-stress	F (15, 381) = 2.5950; p = 0.001
unstressed	F (15, 381) = 4.5708; p < 0.001

Table 2 The influence of different syllable statuses with respect to stress on the values of H1-H2 for individual speakers.

As the table shows, the differences between speakers are statistically highly significant in stressed, post-stress as well as unstressed syllables; the effect size being the largest in stressed syllables and the lowest in unstressed syllables. This may be caused by the fact that vocal effort is higher in stressed than in unstressed syllables. The values in stressed syllables are thus more stable. In post-stress syllables, additional variability is present.

Since stressed, post-stress and unstressed syllables were found to behave differently in utterance-final and utterance-non-final stress groups (see Figure 7), we likewise examined the influence of *stress group position in the utterance* on H1-H2 values for individual speakers.

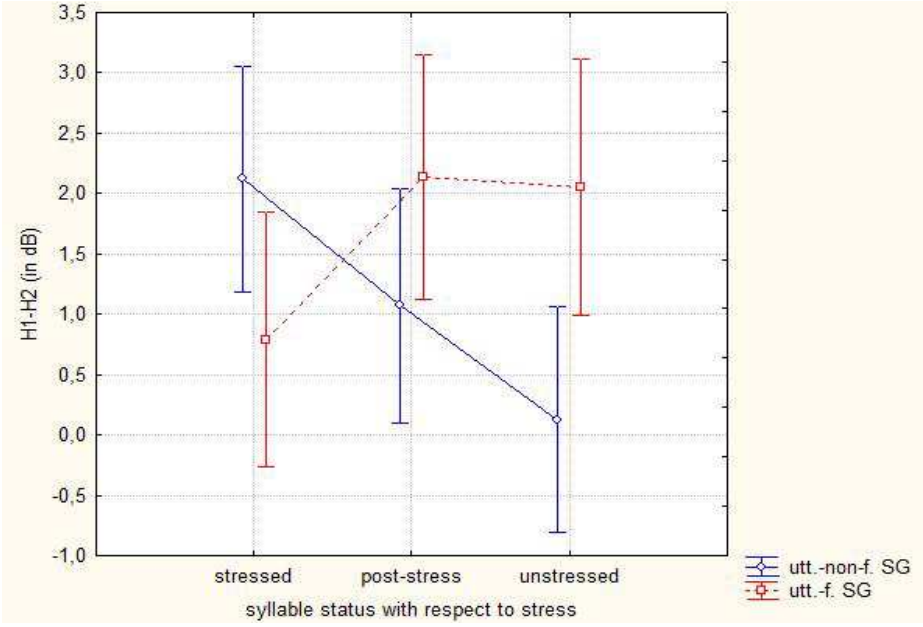


Fig. 7 The values of H1-H2 in stressed, post-stress and unstressed syllables in both utterance-non-final stress groups (utt.-non-f. SG) and utterance-final stress groups (utt.-f. SG). (Error bars indicate 95% confidence intervals.)

The results of ANOVA concerning the values of H1-H2 for individual speakers in utterance-final and utterance-non-final stress groups are presented in Table 3. It shows that the differences between speakers are statistically highly significant in utterance-non-final as

well as utterance-final stress groups; the effect size being larger in utterance-non-final stress groups. Since H1-H2 is considered to be an indication of an open or adduction quotient (see Section 3.2.2), a possible explanation could be that utterance-final stress groups are more susceptible to variability due to a decrease of vocal effort and a more frequent presence of breathy or creaky phonation than in utterance-non-final stress groups.

H1-H2	ANOVA
utterance-non-final stress group	F (15, 638) = 7.5856; $p < 0.001$
utterance-final stress group	F (15, 523) = 5.0957; $p < 0.001$

Table 3 Statistical significance of differences in the values of H1-H2 between speakers in utterance-final and utterance-non-final stress groups.

Individual vowels were also found to differ in H1-H2 values (F (4, 1188) = 2.8065; $p < 0.05$). Specifically, a significant difference has been found between the vowels /u/ and /e/ (Tukey HSD post-hoc test: $p < 0.05$). We were therefore interested whether some vowels reflect differences in H1-H2 values between speakers better than others. The results of the five analyses of variance are summarized in Table 4.

H1-H2	ANOVA
/ɪ/	F (15, 223) = 3.3243; $p < 0.001$
/e/	F (15, 191) = 3.6314; $p < 0.001$
/a/	F (15, 192) = 2.7612; $p < 0.001$
/o/	F (15, 222) = 2.0608; $p = 0.01$
/u/	F (15, 221) = 2.1231; $p = 0.01$

Table 4 The effect of individual vowels on H1-H2 values of individual speakers.

The table shows that all vowels reflect statistically significant differences between speakers in their H1-H2 values, though the degree of their significance varies. The differences between speakers are the most significant for the front vowel /e/, followed by the other front vowel /ɪ/. The back vowels /o/ and /u/ show a lower degree of significance.

6.1.2. H1-A1 (an indication of first-formant bandwidth)

Our speakers were also found to exhibit statistically highly significant differences in their values of H1-A1 (F (15, 1177) = 13.953; $p < 0.001$). The values of H1-A1 for individual speakers are illustrated in Figure 8.

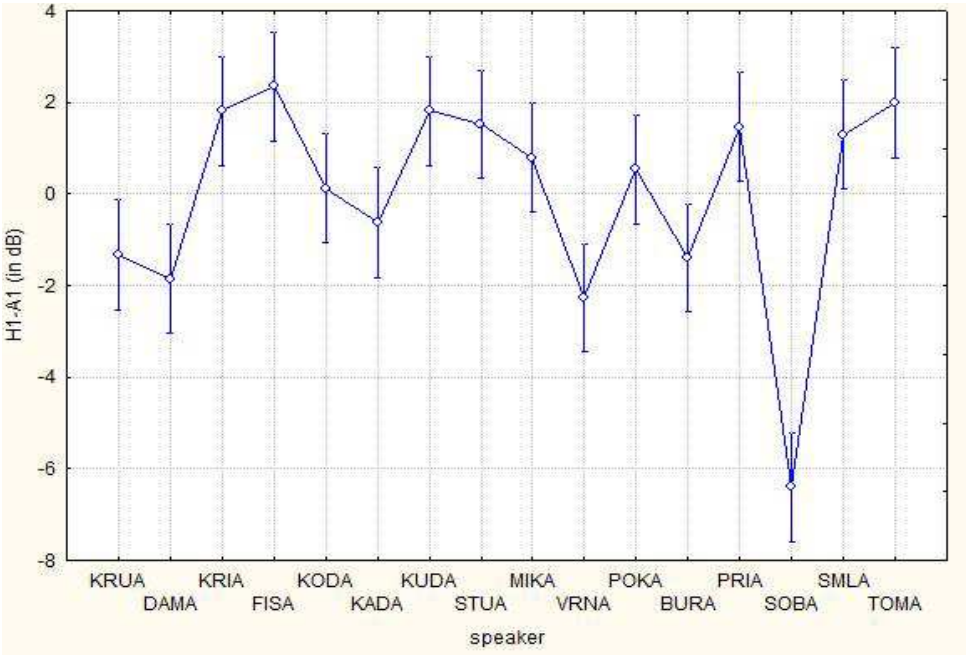


Fig. 8 H1-A1 values in decibels for individual speakers. (Error bars indicate 95% confidence intervals.)

A subsequent post-hoc test revealed that more speakers contribute to this effect than in case of H1-H2. Speaker SOBA again significantly differs from all other speakers (Tukey HSD post-hoc test: $p < 0.001$; for all comparisons). Speaker VRNA, who differed significantly in her H1-H2 values only from FISA and SOBA, differs in her H1-A1 values from 9 speakers ($p < 0.001$ for 5 comparisons, $p < 0.05$ for the remaining four). Speaker DAMA, who differed significantly only from 2 speakers in her H1-H2 values, differs significantly in her H1-A1 values from 8 speakers, FISA from 6 speakers; KRUA, KRIA, KUDA, BURA and TOMA from 5 speakers.

We were interested whether speakers differ in stressed, post-stress as well as unstressed syllables. As Table 5 shows, they do and the differences are statistically highly significant in all cases. The effect size is again the largest for stressed syllables.

H1-A1	ANOVA
stressed	F (15, 383) = 8.5349; $p < 0.001$
post-stress	F (15, 381) = 4.3430; $p < 0.001$
unstressed	F (15, 381) = 4.1942; $p < 0.001$

Table 5 The influence of different syllable statuses with respect to stress on the values of H1-A1 for individual speakers.

A statistically highly significant difference has also been found between H1-A1 values in utterance-final and utterance-non-final stress groups (F (1, 1191) = 21.839; $p < 0.001$). As in the case of H1-H2, we examined whether speakers differ in H1-A1 values in utterance-final as well as utterance-non-final stress groups. As Table 6 shows, the differences between speakers are statistically highly significant in both cases, though the effect size is again slightly larger in utterance-non-final stress groups.

H1-A1	ANOVA
utterance-non-final stress group	F (15, 638) = 8.6626; $p < 0.001$
utterance-final stress group	F (15, 523) = 7.3114; $p < 0.001$

Table 6 The statistical significance of differences in H1-A1 between speakers in utterance-non-final and utterance-final stress groups.

Individual vowels were likewise found to differ in their values of H1-A1 (F (4, 1188) = 5.5724, p = 0.001). It is caused mainly by the vowel /a/, which differs from all other vowels (Tukey post-hoc test: $p < 0.001$ for /e/; $p < 0.05$ for /ɪ/ and /u/) apart from /o/. As Figure 9 shows, individual vowels also behave differently in stressed, post-stress and unstressed syllables (F (8, 1178) = 7.4773; $p < 0.001$); the differences being the most marked in unstressed syllables and the most reduced in stressed syllables.

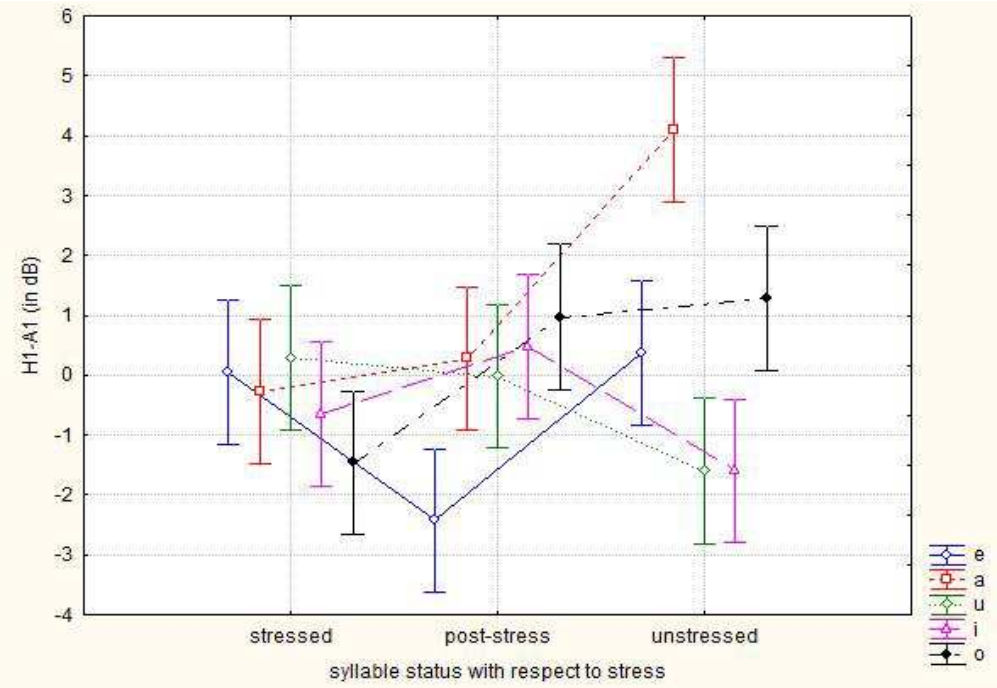


Fig. 9 The H1-A3 values in decibels for individual vowels in stressed, post-stress and unstressed syllables. (Error bars indicate 95% confidence intervals.)

As in the case of H1-H2, we wanted to discover whether speakers differ in their H1-A1 values more for some vowels than for others. The results of ANOVA are presented in Table 7.

H1-A1	ANOVA
/ɪ/	F (15, 223) = 5.1383; $p < 0.001$
/e/	F (15, 223) = 6.5986; $p < 0.001$
/a/	F (15, 224) = 3.8066; $p < 0.001$
/o/	F (15, 222) = 1.7304; $p < 0.05$
/u/	F (15, 221) = 2.0647; $p = 0.01$

Table 7 The influence of individual vowels on H1-A1 values of our speakers.

The table shows that all vowels reflect statistically significant differences between speakers also in the values of H1-A1. As for the two front vowels /e/ and /ɪ/, and the central vowel /a/, these differences are statistically highly significant, while the differences between speakers in the back vowels /o/ and /u/ are of lower significance. The vowel /e/ exhibits again the largest effect size and /o/ the lowest.

6.1.3. H1-A3 (an indication of spectral tilt)

Lastly, speakers were found to differ also in their values of H1-A3 (F (15, 1177) = 11.758; $p < 0.001$). This parameter seems to differentiate the highest number of speakers and, in addition, different ones than the previous two. It has a range of about 20 dB (see Figure 10), indicating a wide variation in spectral tilt between the subjects. H1-A3 values for individual speakers are illustrated in Figure 10.

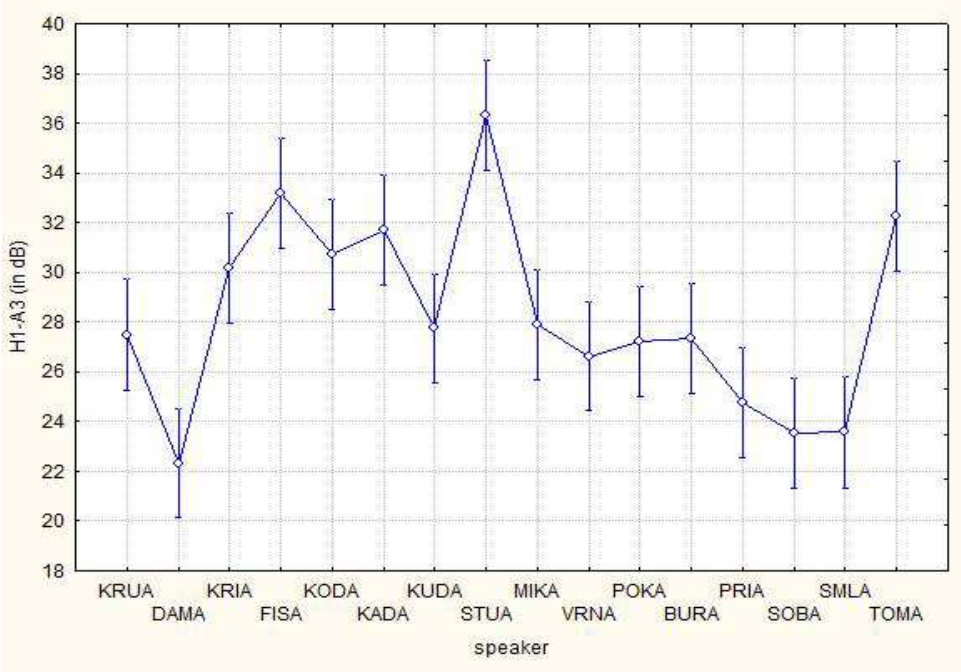


Fig. 10 H1-A3 values in decibels for individual speakers. (Error bars indicate 95% confidence intervals.)

Speaker STUA differs in her H1-A3 values significantly from all speakers apart from three (Tukey HSD post-hoc test: $p < 0.001$ for 10 comparisons; $p < 0.05$ for the remaining two), which is considerably more than for the other parameters since in case of H1-H2, she differed significantly from 2 speakers and in case of H1-A1 from three. Speaker FISA differs from 8 speakers ($p < 0.001$ for 4 comparisons; $p < 0.05$ for the remaining four) and DAMA from seven ($p < 0.001$ for six comparisons; $p < 0.05$ for two). SOBA, who was discriminated from almost all speakers by H1-H2 and H1-A1, and SMLA differ significantly from 6 speakers; DAMA, PRIA and TOMA from 5 speakers.

Again we examined whether speakers differ in stressed, post-stress as well as unstressed syllables. As it is apparent from Table 8, in case of H1-A3, the differences between speakers in stressed syllables are very similar to those in post-stress and unstressed syllables as the effect size for all the three conditions is comparable.

H1-A3	ANOVA
stressed	F (15, 383) = 4.2935; $p < 0.001$
post-stress	F (15, 381) = 4.1913; $p < 0.001$
unstressed	F (15, 381) = 4.2444; $p < 0.001$

Table 8 The influence of different syllable statuses with respect to stress on the values of H1-A3 for individual speakers.

Also H1-A3 values were found to differ in utterance-non-final as opposed to utterance-final stress groups (F (1, 1191) = 25.925; $p < 0.001$). The values in utterance-non-final stress groups are significantly lower as Figure 11 shows; that is, the spectral tilt is less steep than in utterance-final stress groups.

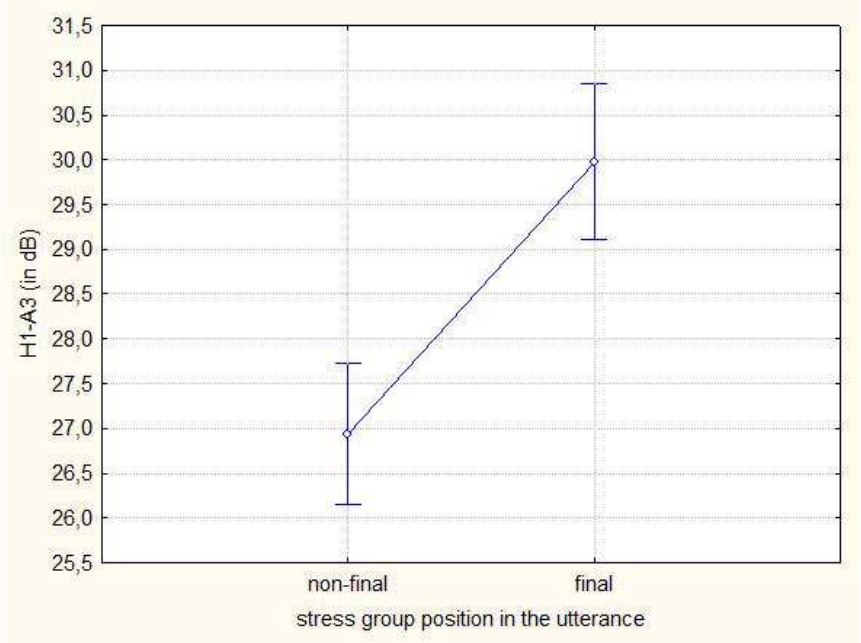


Fig. 11 H1-A3 values in decibels in utterance-final and utterance-non-final stress groups. (Error bars indicate 95% confidence intervals.)

We again examined whether speakers exhibit differences in both utterance-final and utterance-non-final stress groups. The results of ANOVA are presented in Table 9. As the table shows, the differences between speakers are statistically highly significant in both cases though the effect size is larger for utterance-non-final stress groups.

H1-A3	ANOVA
utterance-non-final stress group	F (15, 638) = 9.5087; $p < 0.001$
utterance-final stress group	F (15, 523) = 4.0697; $p < 0.001$

Table 9 The effect of stress group position in the utterance on H1-A3 for individual speakers.

Lastly, we were interested whether also in case of H1-A3 speakers differ more in their values for some vowels than for others. As Table 10 shows, the differences between speakers are statistically highly significant for all vowels and the effect size is again the highest for the front vowel /e/ and the lowest for the back vowel /o/.

H1-A3	ANOVA
/ɪ/	F (15, 223) = 4.1351; $p < 0.001$
/e/	F (15, 223) = 5.7098; $p < 0.001$
/a/	F (15, 224) = 4.8101; $p < 0.001$
/o/	F (15, 222) = 3.7423; $p < 0.001$
/u/	F (15, 221) = 5.4954; $p < 0.001$

Table 10 The influence of individual vowels on the differences in H1-A3 between speakers.

6.2. The discriminative power of H1-H2, H1-A1 and H1-A3

To assess the robustness of the three parameters for discriminating speakers we used linear discriminant analysis (LDA). Meloun, Militký & Hill (2005) (in Volín, 2007, p. 276), consider a sample to be large enough when it contains more than 20 cases for each predictor or each category of a dependent variable, depending on which one is more numerous. Since we had 3 predictors, namely H1-H2, H1-A1 and H1-A3, and 16 categories of a dependent variable, that is, 16 speakers, we needed at least 320 cases (16 times 20) for our results to be reliable. Though our sample was large enough as it consisted of 1193 cases (see Section 5.2), we randomly divided our data into a training set and a testing set, as Volín recommends. The training set consisted of ⅔ of the sample and the remaining ⅓ formed the testing set. The percentage of correctly assigned cases in both sets is presented in Table 11. Since the classification success rate in both sets is comparable, we joined the two sets again and subjected the whole sample to LDA.

Classification success rate (%)	
Training set	Testing set
17.02	15.34

Table 11 Classification success rate of linear discriminant analysis in a training and a testing set.

As the results of ANOVA in previous sections (Sections 6.1.1 - 6.1.3) have shown, our speakers exhibit statistically significant differences in the values of H1-H2, H1-A1 and H1-A3 for all vowel qualities, in stressed, post-stress as well as unstressed syllables and in

both utterance-final and utterance-non-final stress groups. Therefore, we examined the whole sample by means of LDA.

The total classification success rate for the whole sample using the three parameters was 15.84%; that is, 15.84% of all cases were correctly assigned to a respective category, i.e. speaker, which is more than would be caused by mere chance. Considering the number of categories, chance would enable to correctly assign approximately 6% of cases (100 divided by 16). A higher classification success rate thus indicates that the combination of the three parameters accounts for the differences between speakers and has some discriminative power.

As the classification matrix in Table 12 shows (the column '% correctly assigned'), the contribution of individual speakers to the total classification success rate differs considerably. The highest score was reached by SOBA, who was correctly recognized in 57.33% of cases. Speaker STUA was correctly recognized in almost half of the cases (49.33%). Other speakers who scored high are FISA, who has been correctly assigned about a third of cases, and DAMA and PRIA one fifth of cases. KRUA, KODA, MIKA and VRNA were, on the other hand, correctly recognized in only 1 of 75 cases.

Rows: Observed classifications Columns: Predicted classifications	% correctly assigned	KRUA	DAMA	KRIA	FISA	KODA	KADA	KUDA	STUA	MIKA	VRNA	POKA	BURA	PRIA	SOBA	SMLA	TOMA
KRUA	1.35	1	9	1	3	1	3	3	12	0	6	0	4	5	18	5	3
DAMA	20.00	0	15	0	5	0	1	1	3	1	2	5	2	10	24	3	3
KRIA	5.41	0	2	4	7	2	5	3	13	0	1	2	7	8	5	7	8
FISA	29.73	0	6	1	22	1	0	5	15	1	0	1	3	3	3	6	7
KODA	1.33	0	3	0	9	1	5	1	18	1	0	1	6	11	11	4	4
KADA	14.86	0	5	1	6	1	11	0	14	0	3	1	11	2	9	2	8
KUDA	5.33	0	7	1	10	1	3	4	10	0	0	4	6	6	11	8	4
STUA	49.33	0	1	0	1	2	5	3	37	0	1	2	4	4	4	1	10
MIKA	1.33	0	8	0	5	1	2	7	13	1	1	3	7	5	7	8	7
VRNA	1.33	0	9	1	14	0	1	1	9	0	1	2	5	7	21	4	0
POKA	4.00	0	14	0	12	0	2	0	8	0	1	3	3	11	6	12	3
BURA	13.33	0	14	1	2	4	5	2	10	0	0	2	10	6	12	3	4
PRIA	20.00	0	16	0	5	1	2	5	3	1	0	4	5	15	6	7	5
SOBA	57.33	0	6	1	1	1	1	0	11	0	1	1	2	5	43	1	1
SMLA	16.44	0	11	2	5	0	1	3	6	0	1	2	8	13	7	12	2
TOMA	12.16	0	3	1	2	6	5	5	21	1	0	1	4	12	3	1	9
Total	15.84	1	129	14	109	22	52	43	203	6	18	34	87	123	190	84	78

Table 12 Classification matrix for the whole sample; the numbers refer to individual cases. The highlighted diagonal line shows the number of correctly assigned cases for each speaker.

The classification matrix offers also a more detailed view of our data; the observed categories being in rows and the results of classification in columns. The numbers refer to individual cases; therefore, their sum in a row is either 74 or 75, depending on the number of vowels available for each speaker. The diagonal line shows how many cases were correctly assigned to each speaker; the numbers above and below express how many times a respective speaker has been mistaken for another one, thus showing how speakers overlap.

Several things can be commented on. Firstly, we can see that our categories differ considerably in how many cases they were assigned. STUA and SOBA are the two most numerous categories; they were each assigned about ⅙ of all cases (203 and 190, respectively). It was also these two speakers who scored the highest classification success rate (see Table 12) and at the same time thus those who were most frequently assigned other cases. Other numerous categories are DAMA, FISA and PRIA, who were each assigned about one tenth of all cases. A parallel with the total classification success rate can likewise be observed (with the exception of PRIA, who scored rather low; cf. Figure 13 which shows that the classification success rate of SOBA increases when post-stress and unstressed syllables are removed from the analysis). In contrast, there are categories of a very low number, such as KRUA (1 case) and MIKA (6 cases).

Secondly, we can see how many times each speaker has been mistaken for another speaker. KRUA, who was correctly recognized only in one case, has been thus assigned to SOBA in 18 cases, to STUA in 12 cases, to DAMA in 9 cases, to VRNA in six, to PRIA and STUA in five, etc. In contrast, SOBA, who scored the highest classification success rate, has been mistaken for STUA in 11 cases, for DAMA in 6 cases, for PRIA in 5 cases and for other few speakers in one or two cases. The fact that speaker SOBA is rather distinct on the basis of the three parameters has been also shown by its generally high values of squared Mahalanobis distances, while the values for KRUA are very low; in other words, while SOBA is easily distinguishable from other speakers, the opposite is true for KRUA. The overview of squared Mahalanobis distances is presented in Table 13; the highest and the lowest value are in bold. It expresses how easily two categories, i.e. speakers, can be distinguished. The higher the number, the better can be the two speakers distinguished. The lower the number, the more similar they are (on the basis of the variables used in the analysis, that is, H1-H2, H1-A1 and H1-A3). The most different speakers in our study are thus SOBA and FISA as their value of squared Mahalanobis distances is the highest, i.e. 3.75. The most similar speakers are KRUA and BURA together with KODA and KADA as the value is the lowest, i.e. 0.04.

	KRUA	DAMA	KRIA	FISA	KODA	KADA	KUDA	STUA	MIKA	VRNA	POKA	BURA	PRIA	SOBA	SMLA	TOMA
KRUA	0.00	0.36	0.40	0.95	0.15	0.19	0.45	1.04	0.21	0.22	0.30	**0.04**	0.47	1.03	0.61	0.54
DAMA	0.36	0.00	0.94	1.61	0.86	1.03	0.65	2.53	0.48	0.25	0.39	0.48	0.45	1.15	0.36	1.46
KRIA	0.40	0.94	0.00	0.24	0.15	0.33	0.09	0.66	0.08	0.72	0.19	0.50	0.32	2.71	0.48	0.12
FISA	0.95	1.61	0.24	0.00	0.49	0.64	0.37	0.88	0.43	1.01	0.43	1.20	0.95	3.75	1.04	0.44
KODA	0.15	0.86	0.15	0.49	0.00	**0.04**	0.35	0.44	0.17	0.51	0.33	0.19	0.55	1.81	0.79	0.15
KADA	0.19	1.03	0.33	0.64	0.04	0.00	0.61	0.40	0.36	0.53	0.53	0.22	0.87	1.61	1.14	0.26
KUDA	0.45	0.65	0.09	0.37	0.35	0.61	0.00	1.22	0.05	0.64	0.07	0.60	0.13	2.74	0.18	0.39
STUA	1.04	2.53	0.66	0.88	0.44	0.40	1.22	0.00	1.03	1.81	1.40	0.99	1.65	3.24	2.15	0.25
MIKA	0.21	0.48	0.08	0.43	0.17	0.36	0.05	1.03	0.00	0.40	0.05	0.33	0.16	2.10	0.24	0.33
VRNA	0.22	0.25	0.72	1.01	0.51	0.53	0.64	1.81	0.40	0.00	0.29	0.41	0.77	1.10	0.73	1.13
POKA	0.30	0.39	0.19	0.43	0.33	0.53	0.07	1.40	0.05	0.29	0.00	0.49	0.23	2.18	0.22	0.57
BURA	0.04	0.48	0.50	1.20	0.19	0.22	0.60	0.99	0.33	0.41	0.49	0.00	0.54	0.94	0.75	0.55
PRIA	0.47	0.45	0.32	0.95	0.55	0.87	0.13	1.65	0.16	0.77	0.23	0.54	0.00	2.42	0.05	0.64
SOBA	1.03	1.15	2.71	**3.75**	1.81	1.61	2.74	3.24	2.10	1.10	2.18	0.94	2.42	0.00	2.56	2.86
SMLA	0.61	0.36	0.48	1.04	0.79	1.14	0.18	2.15	0.24	0.73	0.22	0.75	0.05	2.56	0.00	0.96
TOMA	0.54	1.46	0.12	0.44	0.15	0.26	0.39	0.25	0.33	1.13	0.57	0.55	0.64	2.86	0.96	0.00

Table 13 Squared Mahalanobis distances for all speakers; the highest and lowest values are in bold.

Having discovered that the combination of our three parameters has some discriminative power, we were interested how individual parameters contribute to the overall model; that is, whether some parameter contributes to the discrimination of speakers more than others. This can be inferred from the comparison of Wilks' lambda with the value of Wilks' lambda for the situation when one variable would be removed from the model. Before doing so, the terminology will be clarified.

Wilks' lambda (λ) expresses the ratio of within-group variance and the total variance in the data set, which in turn consists of between-group and within-group variance. Consequently, the larger between-group variance is in comparison to within-group variance, the lower λ. The limit values of Wilks' lambda are 0 and 1. If $\lambda = 1$, the categories are indistinguishable on the basis of the variables used because within-group variance equals between-group variance. The lower the value, the higher degree of the total variance is explained by the combination of the independent variables; in our case H1-H2, H1-A1 and H1-A3. Wilks' lambda for the situation when one variable would be removed from the model informs us how the efficacy of the whole model would change if a respective variable was removed from it. A high value in this case therefore signals an importance of the variable for the model while a value close to (the total) Wilks' lambda means that a removal of the variable would cause only a minor decrease in the efficacy of the whole model. Let us have a look at the values of Wilks' lambda in our study.

Wilks' lambda (λ) for the overall results is 0.712. Though the value is rather high, it says that the groups *can* be distinguished on the basis of our three parameters (F (45, 3491) = 9.3858; $p < 0.001$). We compared Wilks' lambda with Wilks' lambda after removing one variable from the analysis. The values are as follows:

for H1-H2 λ = 0.755
for H1-A1 λ = 0.768
for H1-A3 λ = 0.808

The most important parameter for discriminating our speakers is thus H1-A3 since if we removed it, Wilks' lambda for the overall results would increase the most; in other words, the largest amount of variation in the data would be left unaccounted for. Removing H1-H2 from our analysis, on the other hand, would have the smallest impact on the efficacy of our model and thus seems to be the least useful parameter. Partial lambda confirmed these results from the opposite perspective. Since partial lambda expresses the contribution of a respective variable to the efficacy of the model, the higher the value, the less useful the variable is for distinguishing the categories. The values of partial lambda for individual independent variables are as follows:

for H1-H2 λ = 0.944
for H1-A1 λ = 0.928
for H1-A3 λ = 0.882

H1-H2 is thus the least useful parameter for discriminating speakers as it accounts for the lowest amount of variation in the data, which is reflected in its highest value of partial lambda. H1-A3 alone accounts for the highest amount of variation, which is reflected in its lowest value, and thus appears to be the most useful parameter.

We can conclude this section by saying that the three parameters have been found to have some discriminative power as they correctly assigned 15.84% of all cases to a respective category, which is more than would be cause by chance. As a next step, we were interested how classification success rate will change if we remove certain data from the analysis.

6.2.1. The influence of *syllable status with respect to stress* on classification success rate

Even though the results of ANOVA have shown that differences between speakers are statistically highly significant in stressed as well as post-stress and unstressed syllables (see Section 6.1.1 for H1-H2, Section 6.1.2 for H1-A1 and Section 6.1.3 for H1-A3), especially in case of H1-A1, the effect size for stressed syllables was considerably larger than for the other two. The three types of syllable statuses with respect to stress have also been found to behave differently in utterance-final and utterance-non-final stress groups, the values being the most stable in stressed syllables, as Figure 12 shows.

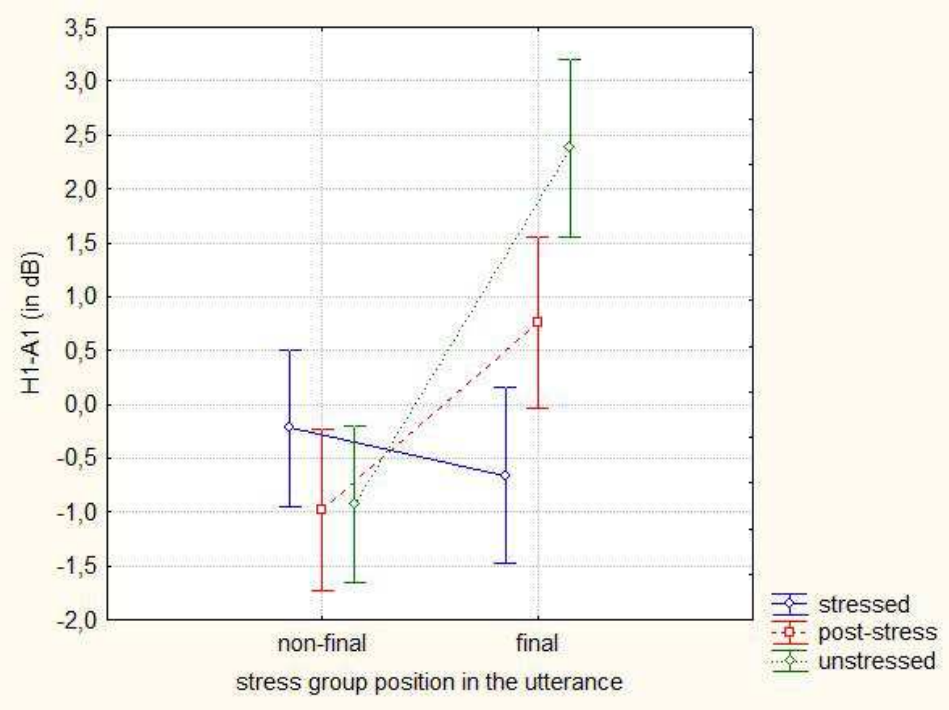


Fig. 12 H1-A1 values in decibels in stressed, post-stress and unstressed syllables in both utterance-final and utterance-non-final stress groups. (Error bars indicate 95% confidence intervals.)

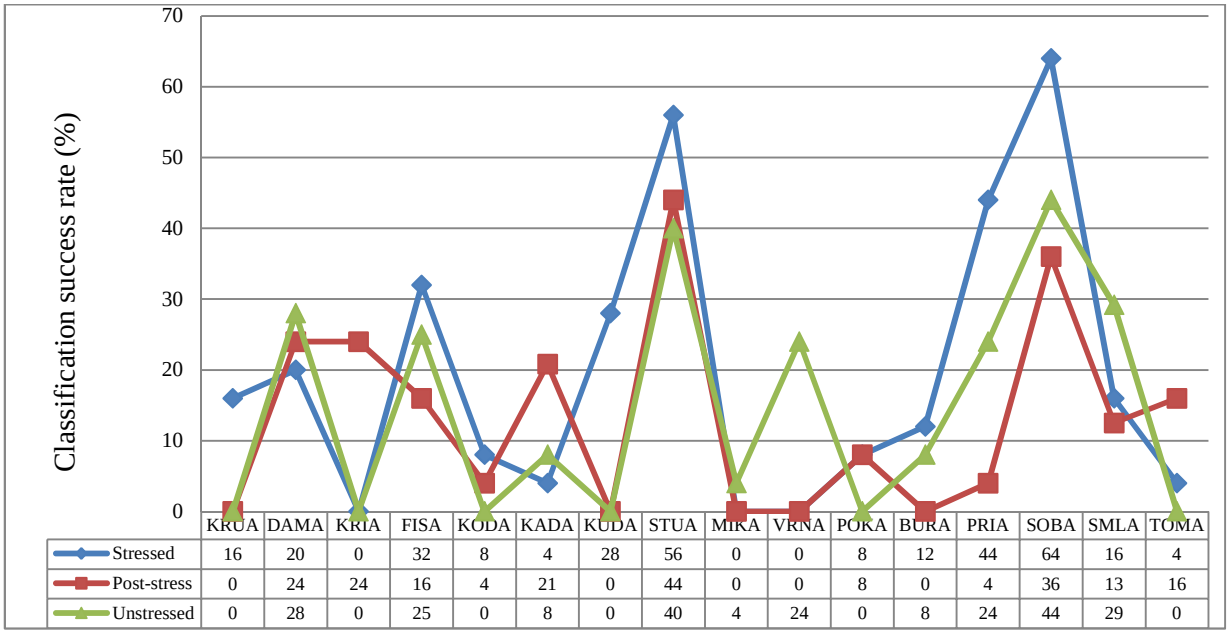
We therefore conducted LDA for stressed, post-stress and unstressed syllables separately in order to examine whether our speakers are better discriminated in stressed, post-stress or unstressed syllables. Classification success rates for these three cases are presented in Table 14.

Syllable status with respect to stress	Classification success rate (%)
stressed	19.55
post-stress	13.10
unstressed	14.61

Table 14 Classification success rate in stressed, post-stress and unstressed syllables.

As we can see, the classification success rate exceeds the 6% threshold in all three cases, but the individual scores differ. The classification success rate increases up to 19.55% when post-stress and unstressed syllables are removed from the analysis, i.e. when only stressed syllables are considered. In unstressed syllables, the success rate decreases to 14.61% and it is the lowest in post-stress syllables (13.10%). The values in stressed syllables thus appear the most stable (cf. Figure 12) and therefore the most reliable for discriminating speakers, while in post-stress syllables, additional variability is introduced. The values of Wilks' lambda (see Table 15, the column 'Wilks' lambda') likewise show that within-group variance is the smallest in stressed syllables, while in the other two cases, additional variability is present, which is reflected in the higher values of Wilks' lambda.

As a next step, we wanted to discover whether this applies for all speakers. The classification success rates for individual speakers in stressed, post-stress and unstressed syllables are presented in Figure 13; the numbers below the graph express the classification success rates (in %) for the three cases. Though most speakers are recognized the best in stressed syllables, we can see that some speakers reach higher classification success rate in post-stress or unstressed syllables; for instance, KRIA and VRNA, respectively.



	KRUA	DAMA	KRIA	FISA	KODA	KADA	KUDA	STUA	MIKA	VRNA	POKA	BURA	PRIA	SOBA	SMLA	TOMA
Stressed	16	20	0	32	8	4	28	56	0	0	8	12	44	64	16	4
Post-stress	0	24	24	16	4	21	0	44	0	0	8	0	4	36	13	16
Unstressed	0	28	0	25	0	8	0	40	4	24	0	8	24	44	29	0

Fig. 13 Classification success rate in stressed, post-stress and unstressed syllables for individual speakers.

We again examined, which of the parameters is the most useful for discriminating speakers in the three cases. The values of Wilks' lambda and Wilks' lambda after removing one variable from the model are presented in Table 15.

Syllable status with respect to stress	Wilks' lambda	λ for H1-H2	λ for H1-A1	λ for H1-A3
stressed	0.625	0.669	**0.719**	0.702
post-stress	0.672	0.735	0.759	**0.800**
unstressed	0.677	0.748	0.730	**0.783**

Table 15 The values of Wilks' lambda in stressed (N = 399), post-stress (N = 397) and unstressed (N = 397) syllables (the first column) and the values of Wilks' lambda after removing one of the variables from the model (the second, third and fourth column). The values in bold signal which variable is the most useful in the analysis as its removal would decrease the efficacy of the whole model.

We can see that removing the variable H1-A3, which was the most useful when the whole sample was considered (Section 6.2), would be the most detrimental for discriminating our speakers in post-stress and unstressed syllables. However, if only stressed syllables are considered, the most important parameter in our study appears to be H1-A1.

6.2.2. The influence of *stress group position in the utterance* on classification success rate

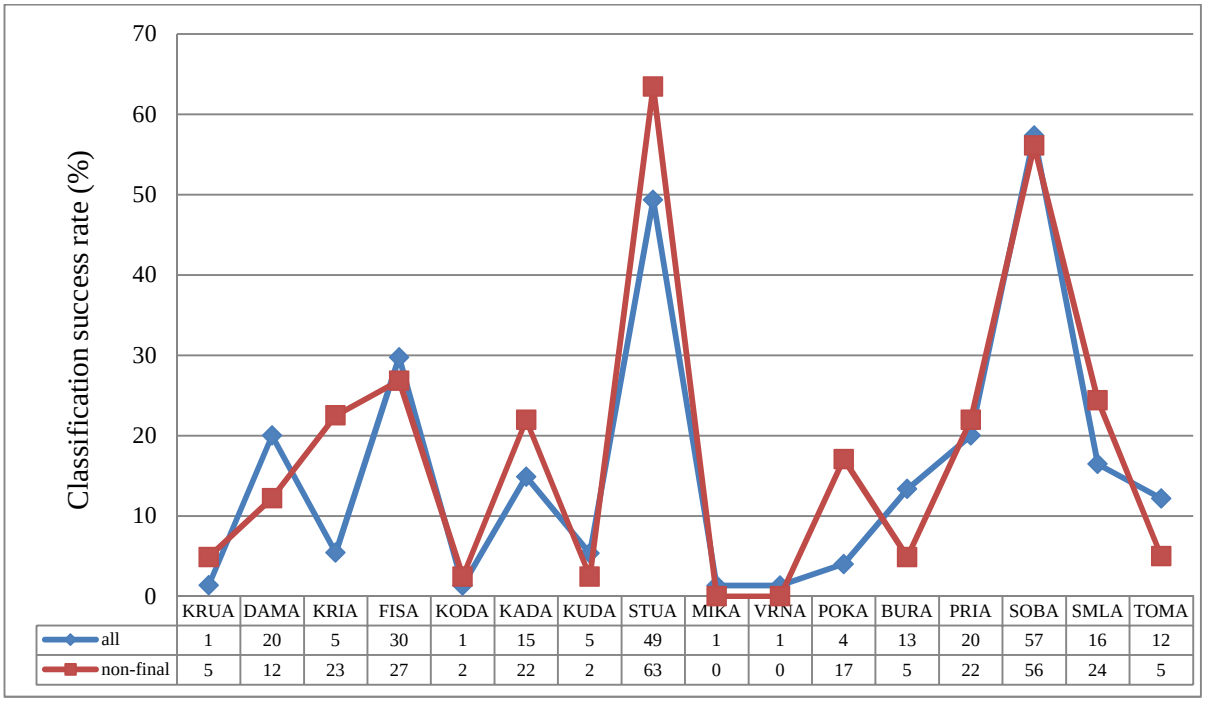
Sections 6.1.1, 6.1.2 and 6.1.3 have also shown that speakers exhibit statistically highly significant differences in the values of all three parameters in both utterance-final and utterance-non-final stress groups, though in all cases the effect size was larger for the latter. We were therefore interested how the classification success changes after removing utterance-final stress groups from the analysis due to possible higher variability of values in utterance-final stress groups. The results are presented in Table 16, which compares the total classification success rate and Wilks' lambda for the whole sample ('all') and for utterance-non-final stress groups ('non-final').

	Stress group position in the utterance	
	all	non-final
Classification success rate (%)	15.84	17.89
Wilks' lambda	0.712	0.646

Table 16 Classification success rate and Wilks' lambda for the whole sample ('all'; N = 1193) and after removing utterance-final stress groups from the analysis ('non-final'; N = 654).

We can see that after removing utterance-final stress groups from the analysis, the classification success rate increases from 15.84% to 17.89%. The improvement of the model is also reflected in the values of Wilks' lambda: λ = 0.712 for the whole sample and 0.646 in utterance-non-final stress groups. The lower value for utterance-non-final stress groups thus signals that within-group variance is smaller; that is, speakers exhibit a lower degree of

variability. A comparison of classification success rate for the whole sample (‘all’) and after removing utterance-final stress groups from the analysis (‘non-final’) for individual speakers is presented in Figure 14.



	KRUA	DAMA	KRIA	FISA	KODA	KADA	KUDA	STUA	MIKA	VRNA	POKA	BURA	PRIA	SOBA	SMLA	TOMA
all	1	20	5	30	1	15	5	49	1	1	4	13	20	57	16	12
non-final	5	12	23	27	2	22	2	63	0	0	17	5	22	56	24	5

Fig. 14 Classification success rate in the whole sample (‘all’, *N* = 1193) and in utterance-non-final stress groups (‘non-final’, *N* = 654).

We can see that the differences in classification success rates for the two cases vary among speakers. While for some speakers, the difference is negligible (e.g., FISA, MIKA and SOBA), other speakers exhibit a larger difference (e.g., KRIA, STUA and POKA). The latter group of speakers can therefore be expected to exhibit a larger variability of values in utterance-final stress groups as opposed to utterance-non-final stress groups since it is in utterance-non-final stress groups that they are better discriminated. Some speakers, in contrast, score a higher classification success rate when the whole data set is considered (e.g., DAMA and BURA); in other words, in their case, it is values in utterance-final stress groups that exhibit lower variability and are more speaker-specific.

To conclude, referring back to Table 16, classification success rate in general increases after removing utterance-final stress groups from the analysis. Utterance-non-final stress groups therefore appear better suited for discrimination of speakers in our study. An even higher classification success rate is, predictably, reached when also unstressed and post-stress syllables are removed from the analysis; that is, when only stressed syllables in utterance-non-final stress groups are considered. The success rate reached is 21.97%, though

the low number of cases (N = 223) does not yield stable results any more. However, under real conditions or even just for different speakers, for instance, male speakers, it may happen that vowels in utterance-final stress groups are more speaker-specific (as it has been also shown for some of our speakers, see Figure 14). In forensic casework it is not possible to omit utterance-final stress groups and to base an analysis on utterance-non-final stress groups only. As a result, stressed syllables (in both utterance-final and utterance-non-final stress groups) appear to be the most reliable for discrimination of speakers (cf. Table 12).

6.2.3. The influence of *vowel* on classification success rate

Lastly, we conducted LDA for each vowel quality separately since Sections 6.1.1 to 6.1.3 have shown that some vowels express the differences between speakers better than others. LDA has confirmed the findings obtained by ANOVA; specifically, the highest classification success rate was scored in the values of /e/ (20.92%) and the lowest in /o/ (14.71%). The number of cases (N = 237 to 240) is, however, too low to yield a reliable analysis. It can be expected that the classification rate would again increase when all unstressed and post-stress /e/ were removed from the analysis, but the amount of data available for the present study does not suffice to prove this.

The following section will shortly comment on the ranges of parameters values measured in our study after which the relation between the three parameters will be examined.

6.2.4. Ranges of parameter values and their relations

Since all three parameters quantify spectral tilt by comparing different amplitudes in the spectra, the lower is the value of a parameter, the less steep is spectral tilt and vice versa.

H1-H2 has a range of about 16 dB (see Figure 6), the maximum value being 8 dB and the minimum value -8 dB; that is, while H1 is more prominent for some speakers than H2, for others it is the other way round. H1-A1 has a range of 12 dB (see Figure 8), the minimum value being -8 dB and the maximum 4 dB, suggesting that while the first-formant peak is quite prominent for some speakers, it is rather damped for others. H1-A3 ranges from 39 dB to 20 dB (see Figure 10), which indicates a considerable variation in spectral tilt among our subjects. Such high values of H1-A3, i.e. arguably steep spectral tilt, could also suggest that in case of some subjects the vocal folds do not close simultaneously or completely (or both) during the closing phase (see Section 3.2.2).

As has been mentioned in Section 3.2.2, the relationship between these measures can be predicted by theory in some cases; especially in situations when the glottis does not close completely (for H1-A1 and H1-A3). For the purposes of the present study, a high correlation is not desirable since parameters expressing speaker identity should be ideally independent of one another. The correlations between the three parameters for all vowels combined are provided in Table 17.

	H1-H2	*H1-A1*	*H1-A3*
H1-H2	1	0.68	0.3
H1-A1	0.68	1	0.28
H1-A3	0.3	0.28	1

Table 17 Pearson product moment correlation coefficients (*r*) for the three parameters for all five vowels /a, e, ɪ, o, u/ combined (*N* = 1193).

Considering a correlation with r ≥ 0.70 to be strong, this threshold is almost reached by the correlation between H1-H2 and H1-A1; other correlations are low. We inspected the scatterplots and present the scatterplot of the correlation between H1-H2 and H1-A1 in Figure 15. It portrays all vowels spoken by all speakers, i.e. 1193 data points.

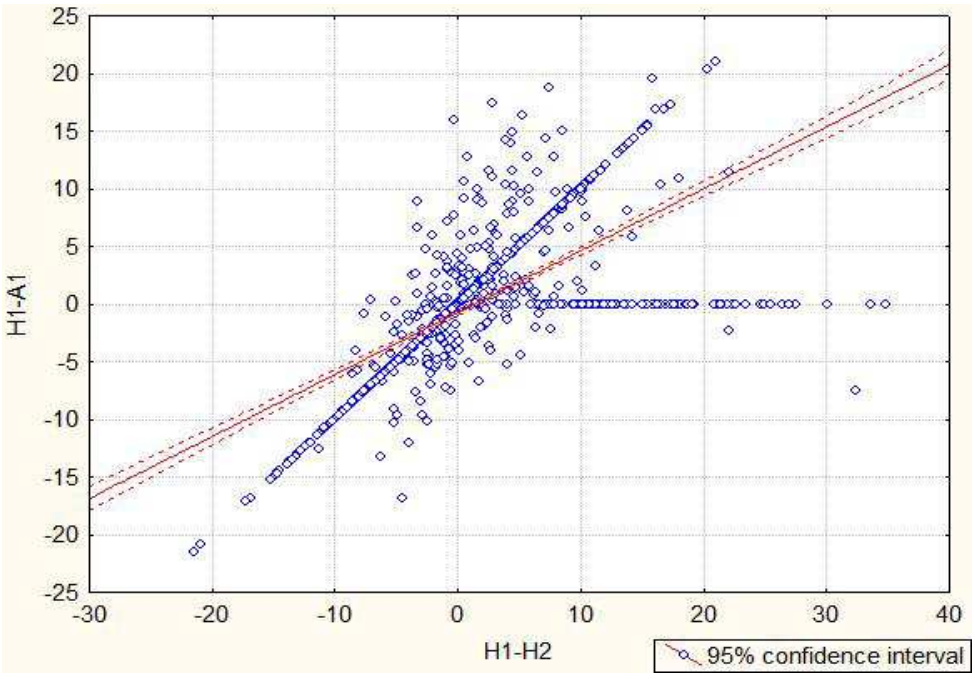


Fig. 15 The scatterplot of a correlation between the parameters H1-H2 and H1-A1 (*r* = 0.68). Each point represents one realization of a vowel by one speaker, hence 1193 data points.

The scatterplot shows regularly arranged data points into a diagonal line. These are cases when A1 equals H2. This happened rather often (in 826 cases, which forms ⅔ of the

whole sample); especially in high vowels, where F1 is the lowest and the closest harmonic in frequency consequently tended to be the second one or in some cases even the first one, which can likewise be seen in the figure. If A1 equals H1, then H1-A1 is zero. The data points forming a horizontal line on the level of zero represent these cases.

Due to the finding that these two parameters overlap to a considerable extent, we removed H1-H2 from the analysis because it proved the less useful of the two (see Sections 6.2.1 and 6.2.2), and conducted LDA again, this time only with two predictors, namely H1-A1 and H1-A3, and then compared the results.

6.2.5. LDA with 2 predictors (H1-A1 and H1-A3)

We again conducted LDA for the whole sample, then for stressed, post-stress and unstressed syllables separately and, lastly, for utterance-non-final stress groups only (but for stressed, post-stress as well as unstressed syllables as otherwise the number of cases would be too low to yield stable results; see Section 6.2.2). As in the previous analysis, the highest classification success rate was achieved in stressed syllables. All results, both for individual speakers and total, are presented in Table 18.

	Classification success rate (%)				
	all	stressed	post-stress	unstressed	non-final
KRUA	0.00	16.00	0.00	0.00	9.76
DAMA	25.33	28.00	16.00	24.00	17.07
KRIA	2.70	0.00	24.00	0.00	20.00
FISA	14.86	4.00	0.00	12.50	0.00
KODA	2.67	8.00	0.00	0.00	0.00
KADA	16.22	20.00	20.83	8.00	26.83
KUDA	9.33	24.00	4.00	0.00	2.44
STUA	40.00	44.00	36.00	40.00	51.22
MIKA	0.00	0.00	4.00	0.00	2.44
VRNA	4.00	4.00	4.00	0.00	0.00
POKA	2.67	20.00	4.00	0.00	12.20
BURA	0.00	4.00	4.00	8.00	4.88
PRIA	13.33	40.00	16.00	24.00	7.32
SOBA	57.33	68.00	40.00	48.00	53.66
SMLA	24.66	12.00	16.67	25.00	36.59
TOMA	1.35	8.00	12.00	0.00	0.00
Total	13.41	18.80	12.59	11.84	15.29

Table 18 Classification success rate of the LDA with 2 predictors (H1-A1 and H1-A3). 'All' = for the whole sample (*N* = 1193), 'stressed' = only for stressed syllables (*N* = 399), 'post-stress' = only for post-stress syllables (*N* = 397), 'unstressed' = only for unstressed syllables (*N* = 397) and 'non-final' = for utterance non-final stress groups (*N* = 654).

This time we will not comment on the results in such a detail as in LDA with 3 predictors but will only summarize the outcome and use the data for a comparison with the previous analysis. The total classification success rate is 13.41%, which is above the chance classification rate (6%). Thus even only two predictors have some discriminative power.

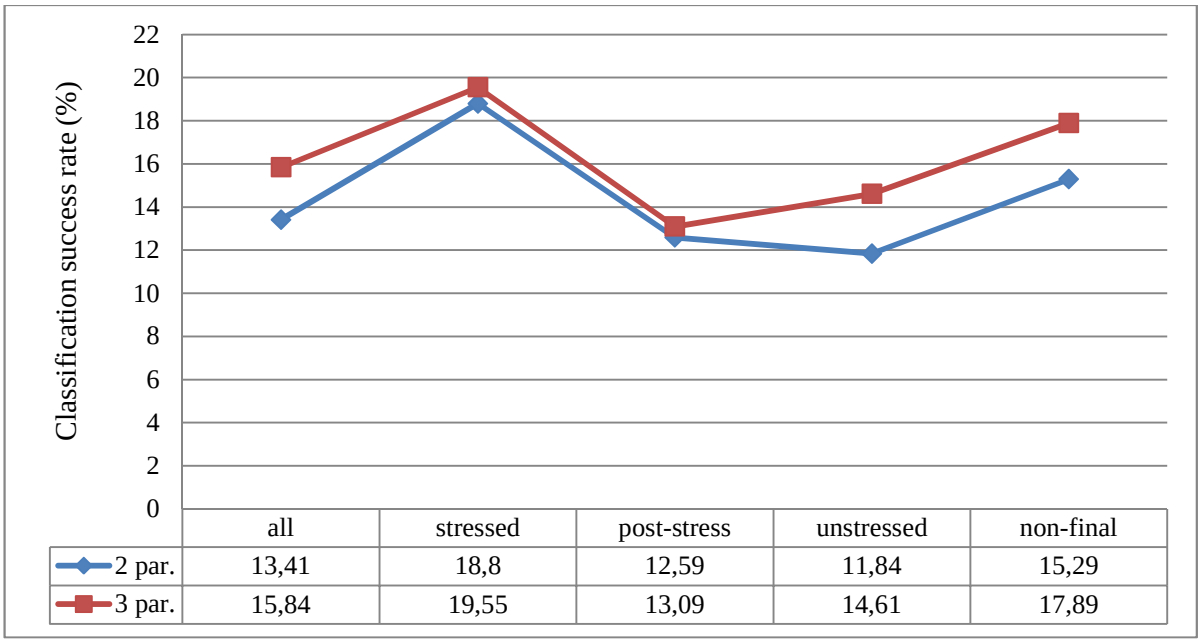
We can again see that individual speakers differ in their contribution, which ranges from 0% (KRUA, MIKA and BURA) to 57.33% (STUA). The overall highest classification success rate increases in stressed syllables, where also most speakers score the highest, though other speakers are the best discriminated in post-stress syllables (the most markedly KRIA) or unstressed syllables (SMLA). The total classification success rate again increases when utterance-final stress groups are removed from the analysis ('non-final'), though half of the speakers scores in utterance-non-final stress groups lower than in utterance-final ones (cf. Figure 14). For these speakers, vowels in utterance-final stress groups appear more speaker-specific.

To see which parameter contributes to the discrimination of speakers more, we inspected Wilks' lambda. Its values for LDA with 2 predictors for the whole sample and for stressed, post-stress and unstressed syllables are presented in Table 19. The figures in bold signal which parameter is the more useful one in the whole model as its removal would have a greater impact on its efficacy.

Syllable status with respect to stress	Wilks' lambda	λ for H1-A1	λ for H1-A3
all	0.755	**0.870**	0.849
stressed	0.669	**0.856**	0.750
post-stress	0.735	**0.858**	0.854
unstressed	0.748	0.857	**0.858**

Table 19 The values of Wilks' lambda in the whole sample ('all', *N* = 1193) and in stressed (*N* = 399), post-stress (*N* = 397) and unstressed (*N* = 397) syllables (the first column) and the values of Wilks' lambda after removing one of the variables from the analysis (the second and third column). The values in bold signal which parameter is more useful in the analysis as its removal would decrease the efficacy of the whole model.

As Table 19 shows, within-group variance is the lowest in stressed syllables, which is also where classification success rate is the highest (see Table 18). The more important variable of the two is in all cases apart from unstressed syllables, though even there it is very close, H1-A1. After removing H1-H2, H1-A1 thus gains on importance.



	all	stressed	post-stress	unstressed	non-final
2 par.	13,41	18,8	12,59	11,84	15,29
3 par.	15,84	19,55	13,09	14,61	17,89

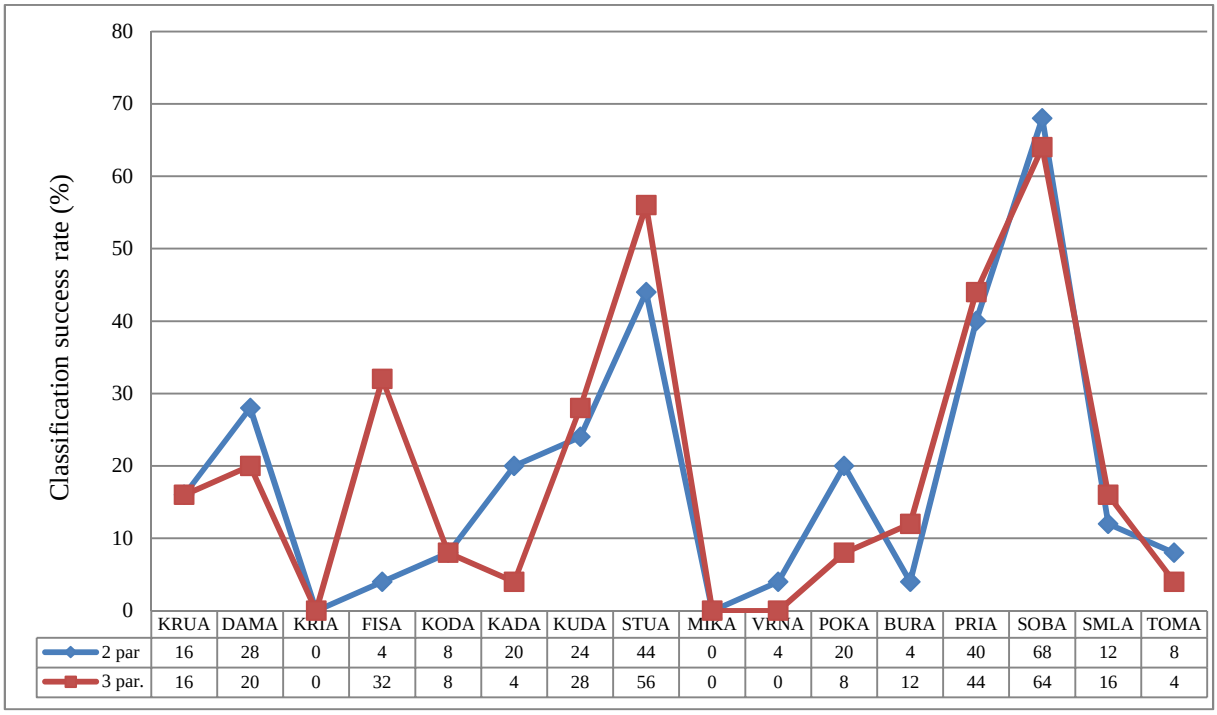
Fig. 16 A comparison of classification success rates for LDA with two (2 par.) and three (3 par.) predictors. The success rate is given in percentages. 'total' = for the whole sample (*N* = 1193), 'stressed' = only for stressed syllables (*N* = 399), 'post-stress' = only for post-stress syllables (*N* = 397), 'unstressed' = only for unstressed syllables (*N* = 397) and 'non-final' = for utterance non-final stress groups (*N* = 654).

Let us compare the classification success rates of LDA with 2 and 3 predictors, the illustration of which is presented in Figure 16. As we can see, classification success rates of LDA with 2 predictors are generally lower than LDA with 3 predictors, but the differences are not linear. While the removal of H1-H2 has only a minor impact on the efficacy of the model in stressed and post-stress syllables, the differences are more marked in unstressed syllables, where additional variability seems to be present, for which the combination of the two parameters after removing H1-H2 fails to account.

In stressed syllables, the classification success rate of LDA with 2 predictors thus increased up to 18.80%, which is very close to the success rate achieved with all three predictors (19.55%). In case of post-stress syllables, the success rate of the two analyses is likewise very similar; 12.59% with 2 parameters and 13.10% with 3 parameters. The differences are more marked in unstressed syllables; 11.84% with 2 parameters as opposed to 14.61% with 3 parameters. The total classification success rate again increases by approximately 2% when utterance-final stress groups are removed from the analysis.

Another interesting fact to notice is that in stressed syllables, the classification success rate of LDA with only 2 predictors exceeds that of 3 predictors elsewhere (apart from stressed syllables). In other words, if only two parameters were used, the classification success rate in stressed syllables would be higher than if 3 parameters would be used in post-stress or unstressed syllables. Furthermore, as Figure 17 (for stressed syllables only) shows, certain

speakers reach a higher classification success rate when only 2 parameters are used; for instance, DAMA, KADA or POKA. For these speakers, the parameter H1-H2 does not appear to yield speaker-specific values; in contrast, it increases the variance of values. The opposite is also true; the inclusion of H1-H2 significantly increases the score of FISA. For other speakers (KRUA, KRIA, KODA and MIKA) the results are the same with or without H1-H2.



	KRUA	DAMA	KRIA	FISA	KODA	KADA	KUDA	STUA	MIKA	VRNA	POKA	BURA	PRIA	SOBA	SMLA	TOMA
2 par	16	28	0	4	8	20	24	44	0	4	20	4	40	68	12	8
3 par.	16	20	0	32	8	4	28	56	0	0	8	12	44	64	16	4

Fig. 17 Comparison of classification success rate in stressed syllables by LDA with 2 and 3 parameters for individual speakers. The numbers below the graph express the classification success rate in %.

We can therefore conclude that removing the parameter H1-H2 decreases the overall classification success rate; specifically, the success rate in unstressed syllables. In stressed syllables, the results of LDA with 2 and 3 parameters are comparable. Furthermore, as Figure 17 has shown, individual speakers differ in classification success rates reached with 2 and 3 predictors; some speakers score even better when only 2 predictors are considered.

As a next step, we used the information about how speakers overlap as it seemed that our three parameters could distinguish types of speakers. We therefore tried to match the least successful speakers to that speaker to whom they were most frequently assigned by classification and expected that this would lead to a higher classification success rate. The results are summarized in the following section.

6.2.6. “Types of categories/speakers”

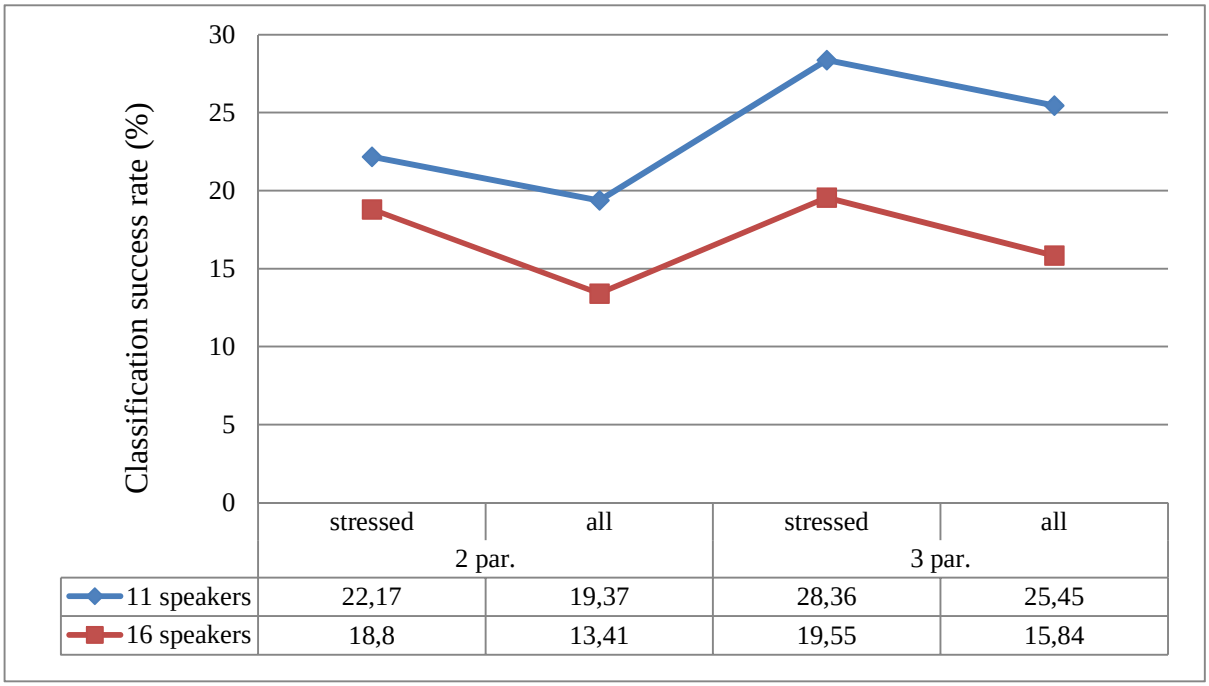
If we have a look back at the classification matrix in Table 12 (p. 66), and as we also commented on in Section 6.2, our categories differ in how numerous they are. There are, on the one hand, categories of a very small number, for instance, KRUA, MIKA, KRIA and VRNA, and, on the other hand, very numerous categories such as STUA, SOBA, FISA and PRIA which are - for some reason (on the basis of certain similarities or overlap of values) - assigned also other cases apart from their own. We could therefore speak of “types of categories” instead of individual categories. Looking at the classification matrix, speaker KRUA, for instance, would be “type SOBA” since she has been the most frequently assigned to this speaker, and speaker KODA would be “type STUA”. Therefore, on the basis of a classification matrix, we tried to manually match speakers with the lowest classification success rate to that speaker to whom they were the most frequently assigned by classification and subjected such a modified sample to LDA again. Surprisingly, the resulting classification success rate was lower than before. Yet we believe that this method, i.e. creating “types of categories” on the basis of a classification matrix would be worth examining in more detail in a future study as understanding why some speakers are assigned to other speakers could bring improvement to the whole model.

6.2.7. LDA for a limited number of speakers

Since the above mentioned manual “recategorization” has not brought the expected improvement, we decided to remove the least successful speakers from the sample entirely and conducted LDA without them. The expectation is again that removal of the least successful speakers will lead to an improvement of the model. The basis for the analysis was the outcome of the classification in stressed syllables (with all 3 predictors) because classification success rate in stressed syllables proved higher than elsewhere (see Section 6.2.1). The threshold of chance classification success rate (6%) was chosen as the criterion for removing speakers. Five speakers were thus removed from the analysis, namely KRIA (0%), KADA (4%), MIKA (0%), VRNA (0%) and TOMA (4%). The numbers in brackets indicate the classification success rate for individual speakers in LDA for stressed syllables (see Figure 13, ‘stressed’).

The analysis for the remaining 11 speakers was conducted in the whole sample and in stressed syllables, for both 2 and 3 predictors because Section 6.2.5 has shown that the

difference between the two in stressed syllables in negligible and some speakers are even better discriminated when only 2 parameters are used. The whole sample was included both for comparison and also for the reason that the number of cases after removing post-stress and unstressed syllables dropped considerably; 275 cases were left. Considering the number of categories, which is 11, this number should, however, still provide stable results. The general results of the analyses are illustrated in Figure 18, which compares the results of LDA for 11 and all 16 categories for both 2 and 3 predictors in stressed (‘stressed’) syllables and in the whole sample (‘all’). The removal of the 5 least successful speakers leads to the increase of classification success rate as predicted.



	2 par.		3 par.	
	stressed	all	stressed	all
11 speakers	22,17	19,37	28,36	25,45
16 speakers	18,8	13,41	19,55	15,84

Fig. 18 Comparison of classification success rate of LDA with 11 and 16 categories (after removing 5 least successful speakers from the analysis). The figure provides results of LDA with 3 predictors in stressed syllables (‘stressed’, *N* = 275 for 11 speakers, *N* = 399 for 16 speakers) and in the whole sample (‘all’, *N* = 821 for 11 speakers and *N* = 1193 for 16 speakers) as well as with 2 predictors for the same data.

If we have a look at the classification success rate of LDA after removing the 5 speakers (Figure 18, the upper line), we can see that the discrimination of speakers is again the most successful in stressed syllables, when all three predictors are included (28.36%). This result needs to be compared to the success rate which would be caused by chance. For 11 speakers, chance would enable to correctly assign approximately 9% of the cases. The score achieved in our study can thus be considered high.

Figure 18 shows one more interesting fact, namely the difference in classification success rates between LDA with 2 and 3 predictors for the two compared analyses. Both in

the LDA with 11 and 16 categories, 3 predictors yield a higher classification rate than 2 predictors (in stressed syllables as well as in the whole sample). However, the difference is more marked in the analysis with 11 categories. The removal of the 5 least successful speakers thus leads to a more significant improvement in the analysis with 3 predictors, which could mean that those speakers who were removed were those whose variance of H1-H2 values was the highest. It is also what the values of Wilks' lambda in Table 20 suggest as the decrease of within-speaker variance in LDA with 3 predictors as opposed to 2 predictors is more marked in the analysis with 11 speakers than with all 16.

The improvement of the model after removing the 5 least successful speakers is thus also apparent from the values of Wilks' lambda, which are summarized and compared with the values for all 16 speakers in Table 20.

11 speakers			
3 predictors		2 predictors	
stressed	all	stressed	all
0.562	0.674	0.620	0.724

16 speakers			
3 predictors		2 predictors	
stressed	all	stressed	All
0.625	0.712	0.669	0.755

Table 20 The values of Wilks' lambda for LDA with 2 and 3 predictors for stressed syllables ('stressed') and the whole sample ('all') for both 11 and 16 categories. The number of cases is as follows: 'stressed', $N = 275$ for 11 categories, $N = 399$ for 16 categories; 'all', $N = 821$ for 11 categories and $N = 1193$ for 16 categories.

The table shows that the removal of the 5 least successful speakers results in lowering within-group variance in comparison to the total variance in the data, which is reflected in lower values of λ, for all comparisons. Interestingly, the value of Wilks' lambda in LDA with 2 predictors and 11 categories in stressed syllables is even slightly lower than in LDA with 3 predictors and 16 categories in the same data. This is also reflected in Figure 18 which shows that classification success rate in LDA with 11 categories is higher with only 2 predictors than the success rate in LDA with 16 categories with all 3 predictors.

Let us have a short look how the removal of the 5 speakers from the analysis affects the results of individual speakers. This information is provided in Table 21 on the following page.

As we can see, nothing changes in the results of the two discriminant analyses with 3 predictors in stressed syllables apart from the classification rate of KRUA and KODA, whose score increased after removing the 5 categories by 4%.

	11 speakers				16 speakers (adapted)			
	3 parameters		2 parameters		3 parameters		2 parameters	
	stressed	all	stressed	all	stressed	all	stressed	all
KRUA	20.00	6.76	20.00	0.00	16.00	1.35	16.00	0.00
DAMA	20.00	21.33	28.00	26.67	20.00	20.00	28.00	25.33
FISA	32.00	32.43	4.00	14.86	32.00	29.73	4.00	14.86
KODA	12.00	1.33	16.00	8.00	8.00	1.33	8.00	2.67
KUDA	28.00	6.67	24.00	10.67	28.00	5.33	24.00	9.33
STUA	56.00	61.33	44.00	45.33	56.00	49.33	44.00	40.00
POKA	8.00	4.00	20.00	2.67	8.00	4.00	20.00	2.67
BURA	12.00	13.33	4.00	6.67	12.00	13.33	4.00	0.00
PRIA	44.00	21.33	40.00	14.67	44.00	20.00	40.00	13.33
SOBA	64.00	57.33	68.00	58.67	64.00	57.33	68.00	57.33
SMLA	16.00	17.81	12.00	24.66	16.00	16.44	12.00	24.66
Total	28.36	22.17	25.45	19.37	19.55	15.84	18.80	13.41

Table 21 The results of LDA with 2 and 3 predictors in stressed syllable and the whole sample for all speakers and after removing 5 speakers with the lowest classification success rate.

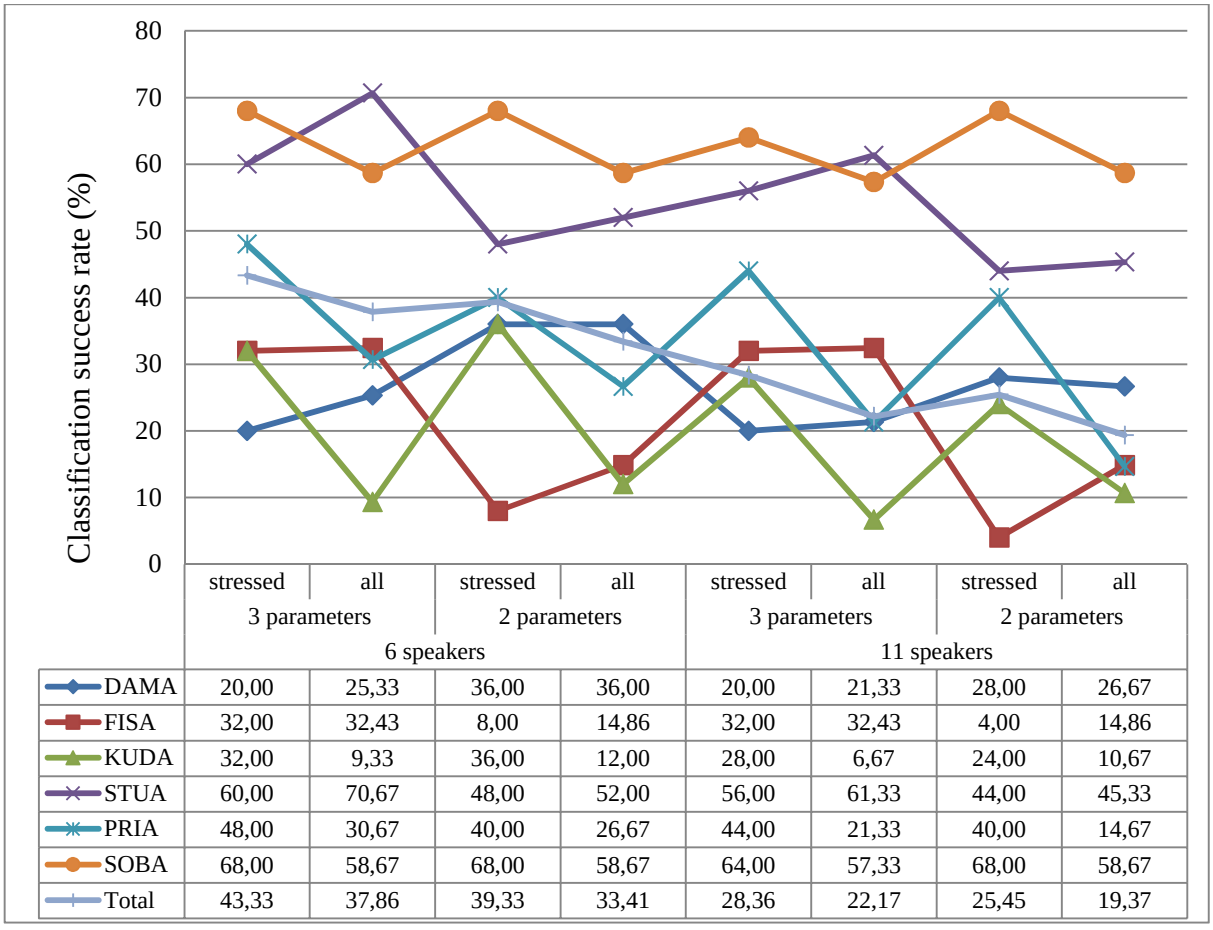
However, if we compare the results of LDA with 3 predictors in the whole sample (‘all’), an increase of classification success rate after the removal of 5 speakers can be observed in most speakers. Only in case of 4 speakers (KODA, POKA, BURA and SOBA), the classification success rate remained the same. The increase of classification success rate ranges from 1.33%, i.e. one case/vowel, (e.g., DAMA) to 12% (STUA). From the observed fact that the removal of 5 categories does not affect the results of individual speakers in stressed syllables but increases them considerably in the whole sample (i.e. stressed, post-stress as well as unstressed syllables), we can infer that in post-stress and unstressed syllables, the 11 speakers were more frequently assigned also to those 5 which we afterwards decided to remove. In stressed syllables, in contrast, our 11 speakers do not seem to be assigned to them, which is reflected in the fact that their removal does not change the results (apart from KRUA and KODA in 3 cases each, i.e. 4%, see above). This is in agreement with the finding, that within-speaker variance is smaller in stressed syllables (cf. Table 20).

As for LDA with 2 predictors, we can observe a larger improvement in the analysis of the whole sample (‘all’), while in stressed syllables the increase of classification success rate concerns again only KRUA and KODA as in the LDA with 3 predictors.

Some speakers are better recognized when only 2 parameters are used, namely, DAMA, KODA, POKA and SOBA, this discrepancy being the most marked for POKA. Apart from KODA, these speakers are the same whose score was higher for 2 predictors than for 3 predictors also in LDA with 16 categories (cf. Figure 17). The other three speakers who

were better recognized when only 2 predictors were used (LDA with 16 categories), namely KADA, VRNA and TOMA were included in the five speakers who were removed.

To terminate our discussion of LDA, we considered a yet smaller number of speakers to see how the results change and to better illustrate how parameter values are distributed for individual speakers. We chose 6 speakers who scored the highest classification success rate; specifically, 20% and more in LDA with all 3 predictors and 16 categories in stressed syllables (see Figure 13). Having 6 categories, a chance classification would cause a success rate of 16%. LDA for the 6 speakers, namely DAMA, FISA, KUDA, STUA, PRIA and SOBA, with 2 predictors reached a classification success rate of 33.41% ($\lambda = 0.633$; $N = 449$) and with all 3 predictors yet a little higher, 37.86% ($\lambda = 0.581$). An expected increase of classification success rate (up to 39.33% and 43.33%, respectively) and decrease of Wilks' lambda (down to 0.510 and 0.488, respectively) was observed in stressed syllables ($N = 150$) for both analyses. The overview of the values for individual speakers as well as their comparison with the previous analysis (LDA with 11 categories) is presented in Figure 19.



	6 speakers				11 speakers			
	3 parameters		2 parameters		3 parameters		2 parameters	
	stressed	all	stressed	all	stressed	all	stressed	all
DAMA	20,00	25,33	36,00	36,00	20,00	21,33	28,00	26,67
FISA	32,00	32,43	8,00	14,86	32,00	32,43	4,00	14,86
KUDA	32,00	9,33	36,00	12,00	28,00	6,67	24,00	10,67
STUA	60,00	70,67	48,00	52,00	56,00	61,33	44,00	45,33
PRIA	48,00	30,67	40,00	26,67	44,00	21,33	40,00	14,67
SOBA	68,00	58,67	68,00	58,67	64,00	57,33	68,00	58,67
Total	43,33	37,86	39,33	33,41	28,36	22,17	25,45	19,37

Fig. 19 The comparison of the classification rate of LDA for 6 and 11 categories with 2 and 3 predictors in stressed syllables ($N = 150$ for 6 speakers and 275 for 11 speakers) and the whole sample ($N = 449$ for 6 speakers and 821 for 11 speakers) both for individual speakers and the total.

The removal of 5 more speakers leads to a further increase of the overall classification success rate as could be expected ('total', signalled by a light blue line). As can be seen in Figure 19, speakers again differ in what enhances their recognition. For KUDA, PRIA and SOBA, it is stressed syllables, which is reflected in the zigzag line rising in stressed syllables ('stressed') and falling for the whole sample ('all'), this difference being especially marked for KUDA and PRIA. Their values in post-stress and unstressed syllables thus seem more variable and less speaker-specific. A reverse of this tendency can be observed in STUA, DAMA and FISA who, in contrast, score a higher classification success rate in the whole sample. However, the differences are much smaller and for FISA only in LDA for 2 predictors. Our speakers likewise again differ in success rates reached in LDA with 2 and 3 predictors. Though the overall trend is for the score in LDA with 3 predictors to be higher (the most markedly in FISA), in some cases it is the other way round, i.e. speakers score higher in LDA with 2 parameters (see DAMA, cf. Table 21).

To illustrate the distribution of parameter values for individual speakers in stressed syllables (as success rate is generally the highest there), Figures 20, 21 and 22 plot the parameters against each other, namely H1-A1/H1-A3, H1-H2/H1-A3 and H1-H2/H1-A1, respectively. The data points are differentiated by 6 types of symbols, each belonging to one speaker.

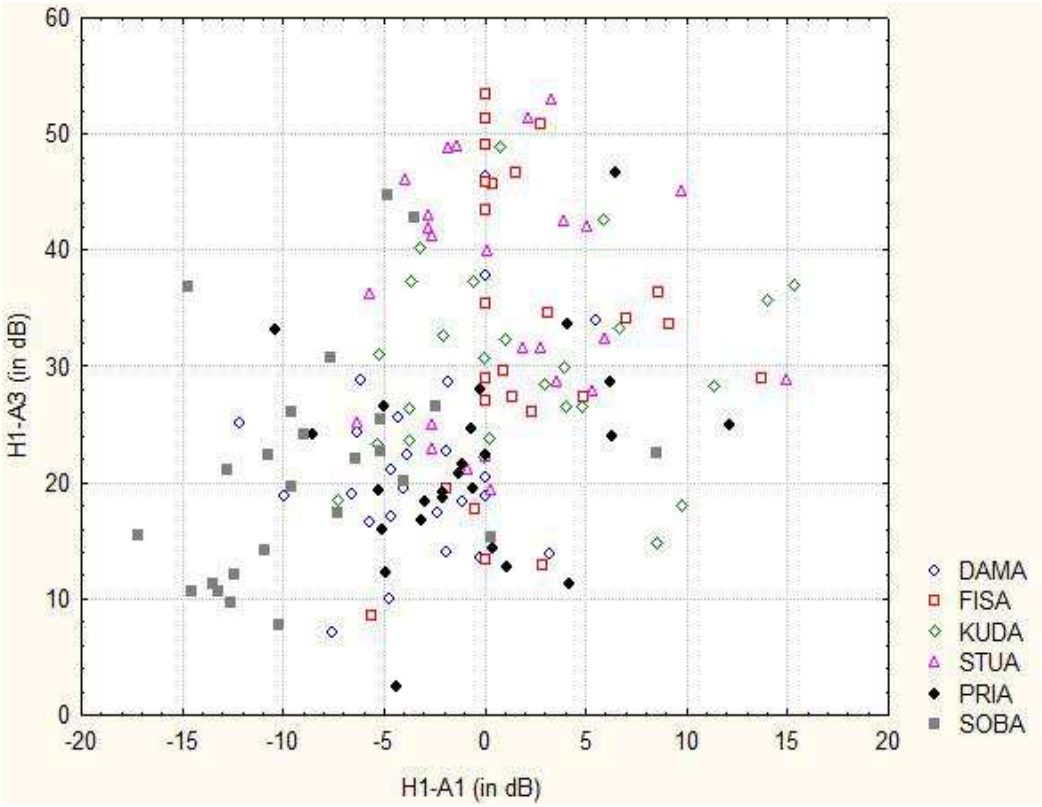


Fig. 20 Discrimination of 6 speakers by their H1-A1 and H1-A3 values in stressed syllables (*N* = 150).

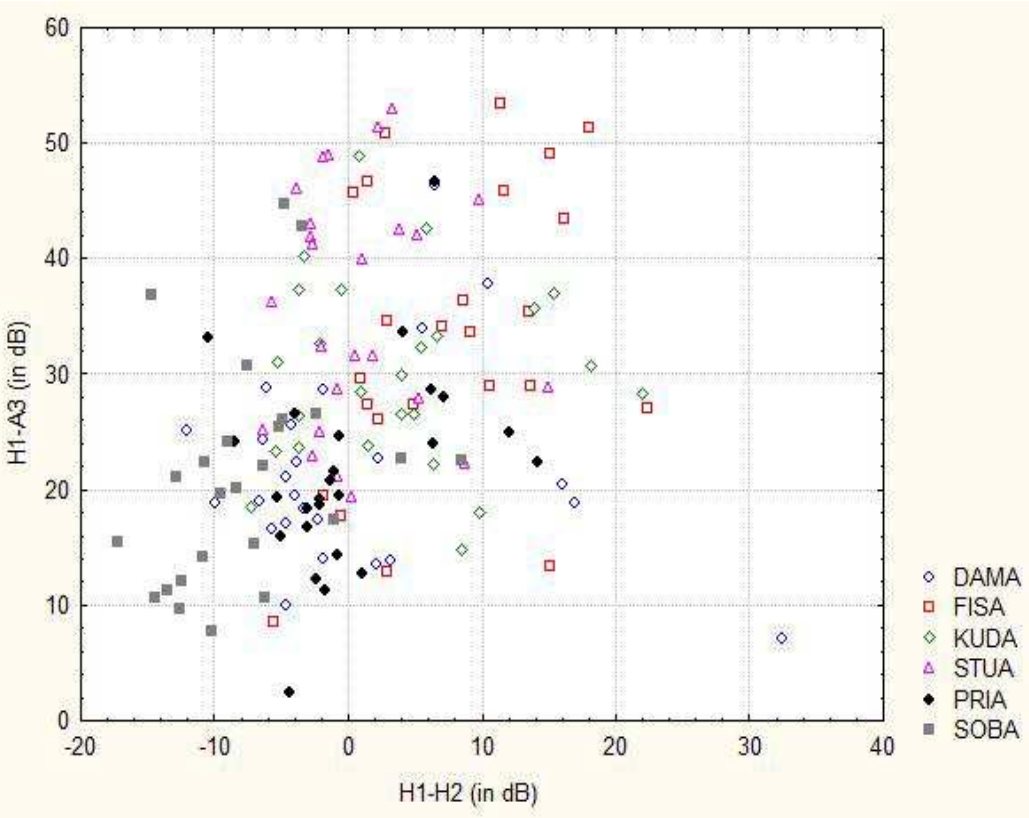


Fig. 21 Discrimination of 6 speakers by their H1-A3 and H1-H2 values in stressed syllables (*N* = 150).

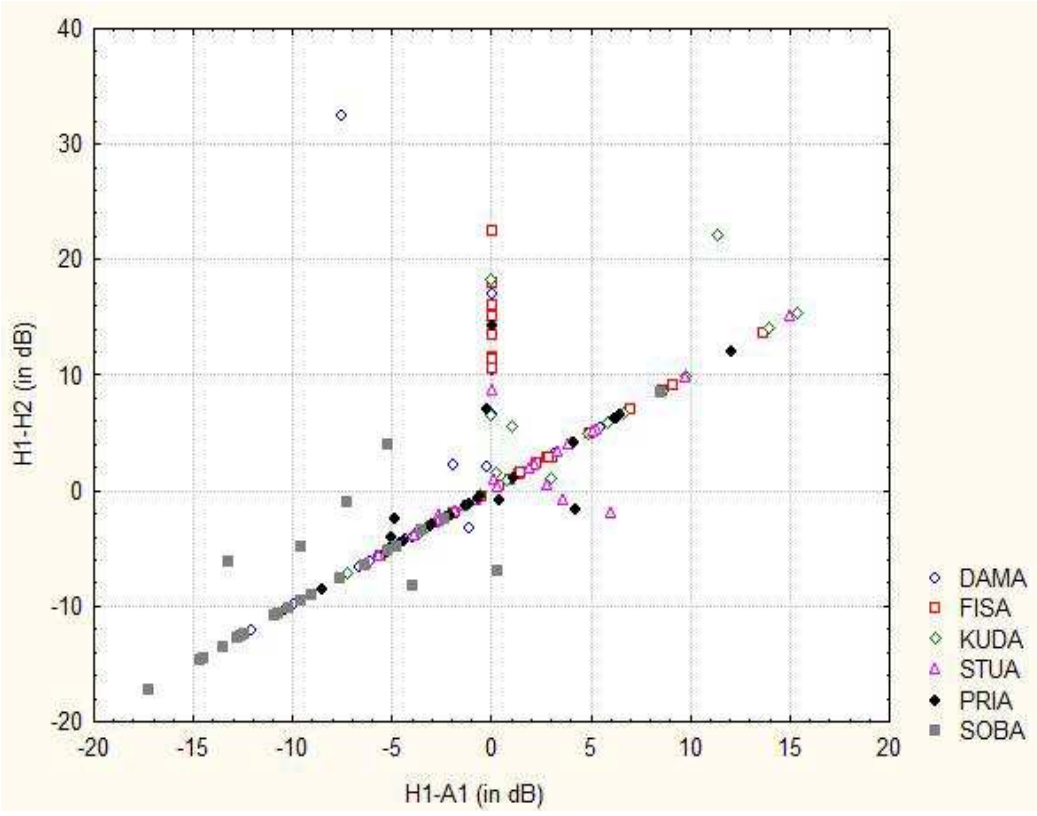


Fig. 22 Discrimination of 6 speakers by their H1-A1 and H1-H2 values in stressed syllables (*N* = 150).

All the figures reveal, though Figure 20 and 21 better than Figure 22 due to the overlap of values, that speaker SOBA is the most distinct due to her generally lowest values of all three parameters. According to squared Mahalanobis distances, the most different speakers are SOBA and FISA (for LDA with 6 categories in stressed syllables as well as for all 16 categories, see Table 13, p. 68). As the figures also show, their overlap is indeed minimal. They seem to be distinguished the best by the parameters H1-A1 (see Figure 20 where SOBA has lower values than FISA) and H1-H2 (see Figure 21). Their overlap on the horizontal axis of the same figures, i.e. in H1-A3, appears to be slightly larger. In contrast, the parameter H1-A3 would help discriminate, for instance, PRIA and STUA better than H1-A1 or H1-H2 would (see Figure 20 and 21, respectively). It thus appears that different speakers are discriminated by different parameters, as might be expected. The figures also reveal that speakers who reached the lowest classification success rate, i.e. DAMA, FISA and KUDA (Fig. 19) are more difficult to discriminate due to their higher extent of overlap with other speakers (DAMA) or larger variance of values (FISA and KUDA).

6.3. Long-term measures of spectral tilt

To complement the outcome of LDA which was based on parameters expressing short-term spectral tilt, we used the results of the LTAS, namely alpha index, Hammarberg index and Kitzing index (see Section 5.1 and 5.2 for a description of how they were obtained), which quantify long-term spectral tilt. The results will be first presented and discussed, and then compared with the results obtained by LDA, by which we will terminate Chapter 6.

Before we provide the values of the three indices for individual speakers, it should be shortly reminded what they express (for a more detailed discussion, see Section 3.2.1). *Alpha index* is the ratio of energy above 1 kHz as opposed to energy below 1 kHz (1-5 kHz/0-1 kHz), from which it follows that the higher its value, the less steep spectral tilt. *Kitzing index* is ‘an inverted alpha’ but with a narrower range, i.e. the ratio of energy below 1 kHz as opposed to energy above 1 kHz (0-1 kHz/1-2 kHz). In this case, a high value therefore signals steeper spectral tilt. Lastly, Hammarberg index expresses the difference between the maximal energy in two frequency bands; specifically, 0-2 kHz and 2-5 kHz; a higher value, i.e. a higher difference between the two amplitudes, therefore again expresses a steeper slope.

Based on the mathematical background, the closest relation is between alpha index and Kitzing index since both express a ratio of energy below and above 1 kHz. Alpha index and Hammarberg index are more distinct since one (alpha index) expresses a ratio and the other (Hammarberg index) a difference. However, they both consider the amount of energy up

to 5 kHz. The most distinct two are thus Kitzing index and Hammarberg index; one being a ratio (Kitzing index) and the other a difference (Hammarberg index) and differing also in which frequency bands they involve. As Table 22 shows, this is reflected in the respective correlation strengths: the strongest correlation is between alpha index and Kitzing index (-0.85) and the weakest between Kitzing index and Hammarberg index (0.54), though this correlation is still moderate.

	alpha	Hammarberg	Kitzing
alpha	1.00	-0.60	-0.85
Hammarberg	-0.60	1.00	0.54
Kitzing	-0.85	0.54	1.00

Table 22 Pearson correlation coefficients (*r*) for correlations between alpha index, Hammarberg index and Kitzing index (*N* = 16).

To allow a comparison of the three indices between the subjects, all values were converted to a *z*-score; the data was normalized against the average value of the respective index for all speakers. Figure 23 illustrates the *z*-scores of all indices for individual speakers.

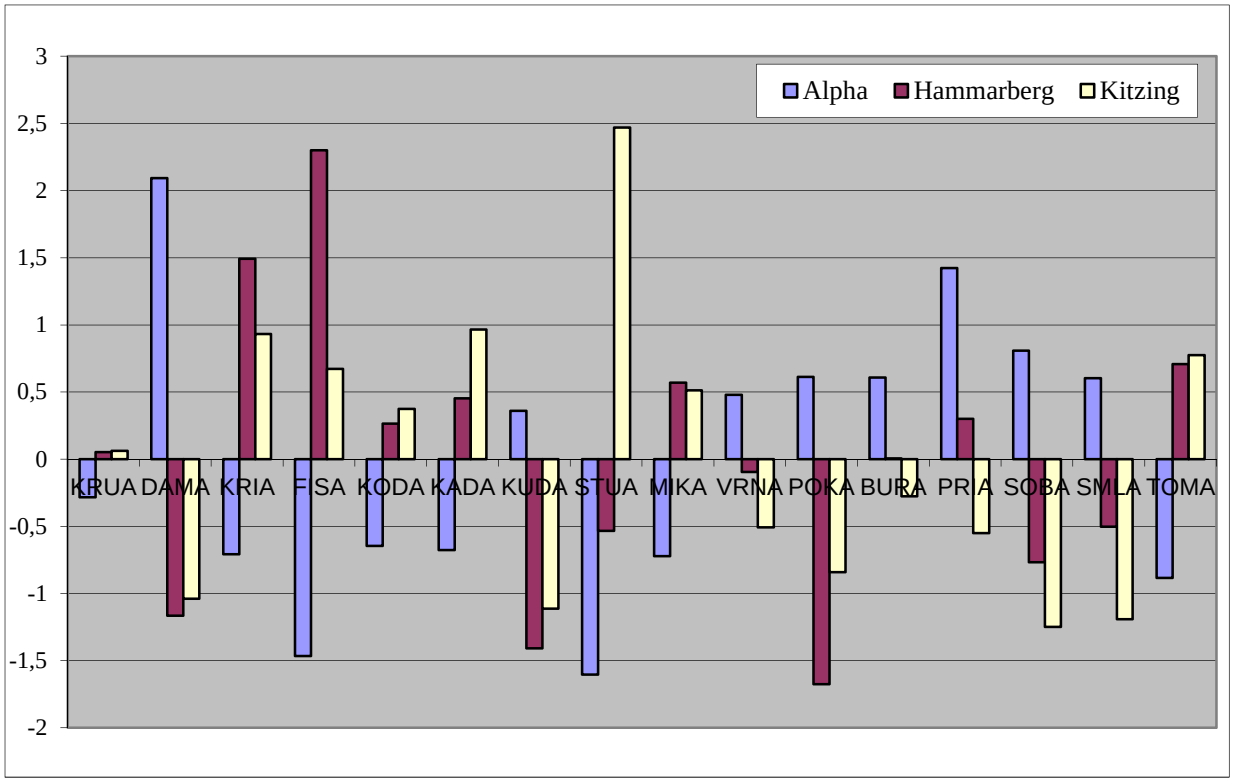


Fig. 23 The values of alpha index, Hammarberg index and Kitzing index for individual speakers converted to *z*-score.

We can see that our speakers exhibit considerable differences in the values of the three indices. While some speakers do not differ much from the average value of any index (KRUA), others exhibit values up to the distance of 2.5 standard deviation (SD) from the average (STUA for Kitzing index). We shall limit our discussion to the speakers who differ the most from the average value for some index and those who differ the least. After that, these results will be compared with the scores of classification success rate reached in LDA.

Figure 23 shows that speakers DAMA, FISA and STUA differ from the average the most: DAMA in her value of alpha index, FISA in Hammarberg index, and STUA in her value of Kitzing index. In all cases, these values exceed the distance of 2 SD from the average; that is, only less than 2% of values would be higher.

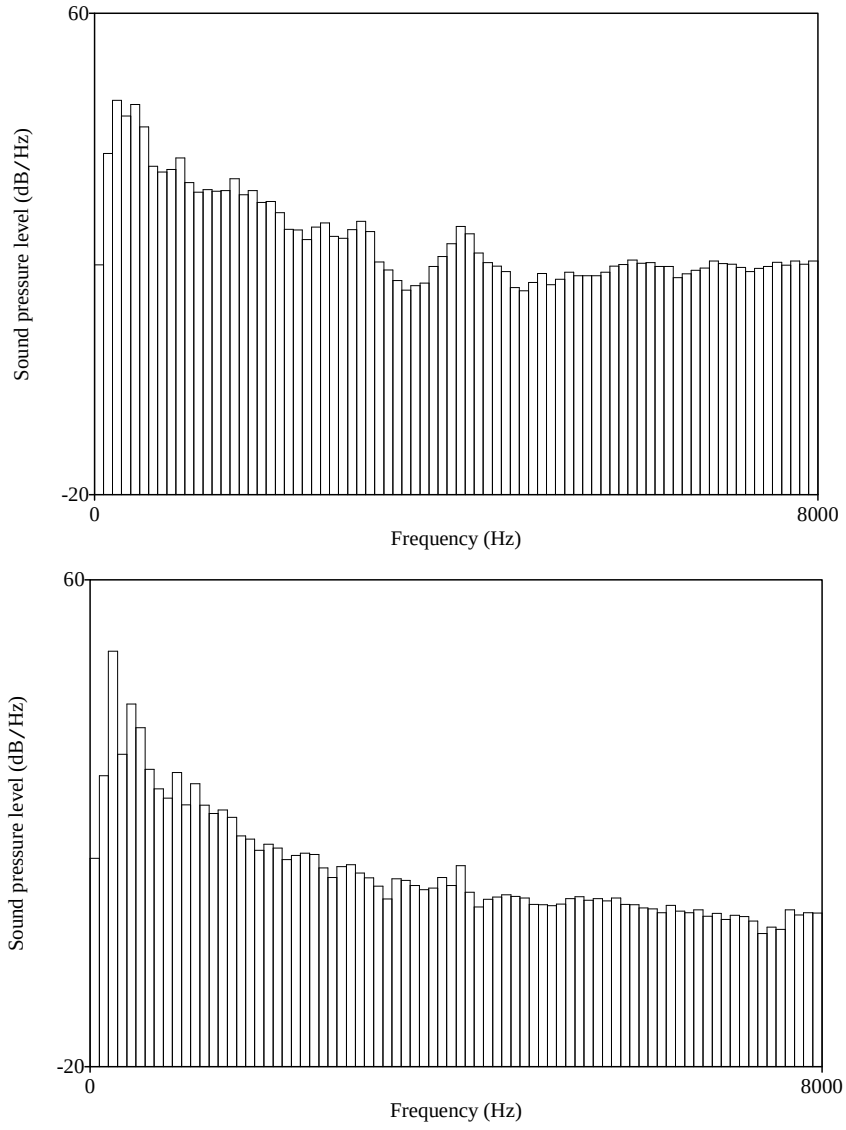


Fig. 24 The LTAS of two speakers showing a frequency range 0-8000 Hz. Speaker DAMA (on the left) has a less steep spectral tilt reflected in her highest value of alpha index (see text above). Speaker FISA (on the right) has one of the lowest values of alpha index and the highest value of Hammarberg index, hence steep spectral tilt.

According to these results, DAMA should have the least steep spectral tilt (since the higher alpha, the less steep spectral tilt) and STUA and FISA the steepest (reflected in the highest value of Kitzing and Hammarberg index, respectively, as well as one of the lowest values of alpha index). Figure 24 (on previous page) illustrates the LTAS of these two speakers; DAMA is above and FISA below. The frequency range is 0-8000 Hz. If we compare the two LTAS, the smaller amount of energy in the range from 1 to 5 kHz as opposed to the range 0-1 kHz in case of FISA is clearly visible. As for the range above 5 kHz, the higher amount of energy in case of DAMA might come from aspiration noise. These two speakers exhibit significant differences also in the other two indices, which is also apparent from Figure 24.

As for the speakers who are the closest to the average, these include KRUA, VRNA and BURA (see Figure 23), but there are also several other speakers within the distance of 1 SD from the average for all three parameters, such as KODA, KADA, MIKA and TOMA.

We were interested whether those speakers who reached the highest classification success rate in LDA; in other words, who were discriminated the best on the basis of the short-term measures of spectral tilt, also differ the most in their LTAS as quantified by the three indices. As a last step, we therefore compared the results of LDA with these long-terms measures. For a better comparison, the results of LDA with all three predictors, i.e. H1-H2, H1-A1 and H1-A3, for the whole sample are provided again in Table 23, which is adapted from Table 12 (p. 66), where the results of LDA for 16 categories and 3 predictors for the whole sample were introduced.

Classification success rate (%)			
SOBA	57.33	TOMA	12.16
STUA	49.33	KRIA	5.41
FISA	29.73	KUDA	5.33
DAMA	20.00	POKA	4.00
PRIA	20.00	KRUA	1.35
SMLA	16.44	KODA	1.33
KADA	14.86	MIKA	1.33
BURA	13.33	VRNA	1.33

Table 23 Classification success rate of LDA with 3 predictors and 16 categories for the whole sample, i.e. 1193 cases.

According to Table 23, the 5 most successful speakers in LDA were SOBA, STUA, FISA, DAMA and PRIA. The distinctiveness of STUA, FISA and DAMA in their values of

the three long-term measures (Figure 23) has been already commented on - it is these three speakers who have the most distinct values of the indices, each of a different one. Speaker PRIA reached in LDA the same classification success rate as DAMA, 20%, and is also the fourth most distinct speaker in her alpha index after the three mentioned. As for SOBA, who reached the highest classification success rate in LDA, her long-term measures do not reflect her distinctiveness so clearly.

In contrast, there are speakers who appear distinct in their values of the long-term measures but who scored rather low in LDA (around 5%), such as POKA, KRIA and KUDA, who exceed or are very close to the distance of 1.5 SD from the average in their values of Hammarberg index. However, if we compare it with the results of LDA after removing utterance-final stress groups (Figure 14, p. 73), we can see that in utterance-non-final stress groups, KRIA and POKA do score high, 23% and 17%, respectively. Similarly, KUDA scores a low overall success rate but in stressed syllables it increases up to 28% (Figure 13, p. 71).

Let us also have a look at speakers who scored the lowest in LDA (Table 23). There are 4 speakers, namely VRNA, MIKA, KODA and KRUA, who were recognized in only one case out of 75 (or 74; see Section 6.2), hence classification success rate is around 1.5%. If we compare it with the results of the long-term measures (Figure 23), we can see that these speakers are also those who scored the most average values in all the three parameters. KRUA has not reached even the distance of 0.5 SD from the average for any of the three parameters, KODA only for alpha index, VRNA for alpha and Kitzing index and MIKA for all three. A slight discrepancy has been found for speaker BURA, who scored one of the most average long-term values (especially of Kitzing and Hammarberg index), but in her short-term measures was recognized not much worse than KADA and SMLA, who exhibit more distinct long-term values.

Lastly, we wanted to compare whether those speakers who differed the most and those who differed the least in short-term measures (as reflected in squared Mahalanobis distances for 3 predictors, 16 categories and the whole sample) exhibit comparable differences in the long-term measures. Table 24 presents squared Mahalanobis distances again (first presented in Table 13, p. 68); the lowest and the highest values are in bold.

	KRUA	DAMA	KRIA	FISA	KODA	KADA	KUDA	STUA	MIKA	VRNA	POKA	BURA	PRIA	SOBA	SMLA	TOMA
KRUA	0.00	0.36	0.40	0.95	0.15	0.19	0.45	1.04	0.21	0.22	0.30	**0.04**	0.47	1.03	0.61	0.54
DAMA	0.36	0.00	0.94	1.61	0.86	1.03	0.65	2.53	0.48	0.25	0.39	0.48	0.45	1.15	0.36	1.46
KRIA	0.40	0.94	0.00	0.24	0.15	0.33	0.09	0.66	0.08	0.72	0.19	0.50	0.32	2.71	0.48	0.12
FISA	0.95	1.61	0.24	0.00	0.49	0.64	0.37	0.88	0.43	1.01	0.43	1.20	0.95	3.75	1.04	0.44
KODA	0.15	0.86	0.15	0.49	0.00	**0.04**	0.35	0.44	0.17	0.51	0.33	0.19	0.55	1.81	0.79	0.15
KADA	0.19	1.03	0.33	0.64	0.04	0.00	0.61	0.40	0.36	0.53	0.53	0.22	0.87	1.61	1.14	0.26
KUDA	0.45	0.65	0.09	0.37	0.35	0.61	0.00	1.22	0.05	0.64	0.07	0.60	0.13	2.74	0.18	0.39
STUA	1.04	2.53	0.66	0.88	0.44	0.40	1.22	0.00	1.03	1.81	1.40	0.99	1.65	3.24	2.15	0.25
MIKA	0.21	0.48	0.08	0.43	0.17	0.36	0.05	1.03	0.00	0.40	0.05	0.33	0.16	2.10	0.24	0.33
VRNA	0.22	0.25	0.72	1.01	0.51	0.53	0.64	1.81	0.40	0.00	0.29	0.41	0.77	1.10	0.73	1.13
POKA	0.30	0.39	0.19	0.43	0.33	0.53	0.07	1.40	0.05	0.29	0.00	0.49	0.23	2.18	0.22	0.57
BURA	0.04	0.48	0.50	1.20	0.19	0.22	0.60	0.99	0.33	0.41	0.49	0.00	0.54	0.94	0.75	0.55
PRIA	0.47	0.45	0.32	0.95	0.55	0.87	0.13	1.65	0.16	0.77	0.23	0.54	0.00	2.42	0.05	0.64
SOBA	1.03	1.15	2.71	**3.75**	1.81	1.61	2.74	3.24	2.10	1.10	2.18	0.94	2.42	0.00	2.56	2.86
SMLA	0.61	0.36	0.48	1.04	0.79	1.14	0.18	2.15	0.24	0.73	0.22	0.75	0.05	2.56	0.00	0.96
TOMA	0.54	1.46	0.12	0.44	0.15	0.26	0.39	0.25	0.33	1.13	0.57	0.55	0.64	2.86	0.96	0.00

Table 24 Squared Mahalanobis distances for all speakers (LDA with 3 predictors and 16 categories, *N* = 1193); the highest and lowest values are in bold.

Certain parallels can again be observed. One of the two pairs of speakers who are the most difficult to distinguish in short-term measures is KODA and KADA (Table 24; the value of squared Mahalanobis distances is 0.04). Their long-term values are also very similar (Figure 23). In contrast, the two speakers who are distinguished the best in short-term measures are SOBA and FISA, whose long-term values are also very different, though not the most. The most different speakers in long-term measures appear to be DAMA and FISA or DAMA and STUA. If we have a look at Table 24, it is FISA and STUA whose squared Mahalanobis distances from DAMA are the largest.

However, in other cases the short-term and the long-term results differ as, for instance, in case of MIKA and KUDA as well as MIKA and POKA whose squared Mahalanobis distances are very low (0.05 for both comparisons); that is, these two speakers are difficult to distinguish. The long-term measures, however, do appear to reflect the differences between them. The opposite is also true. If we have a look at the long-term values of, for instance, SOBA and SMLA, their values are very similar. The short-term measures (Table 24), however, discriminate the two speakers well.

As in the case of the short-term measures, we include for illustration a whole picture of how the three long-term measures of spectral tilt discriminate our 16 speakers. It is provided in Figure 25.

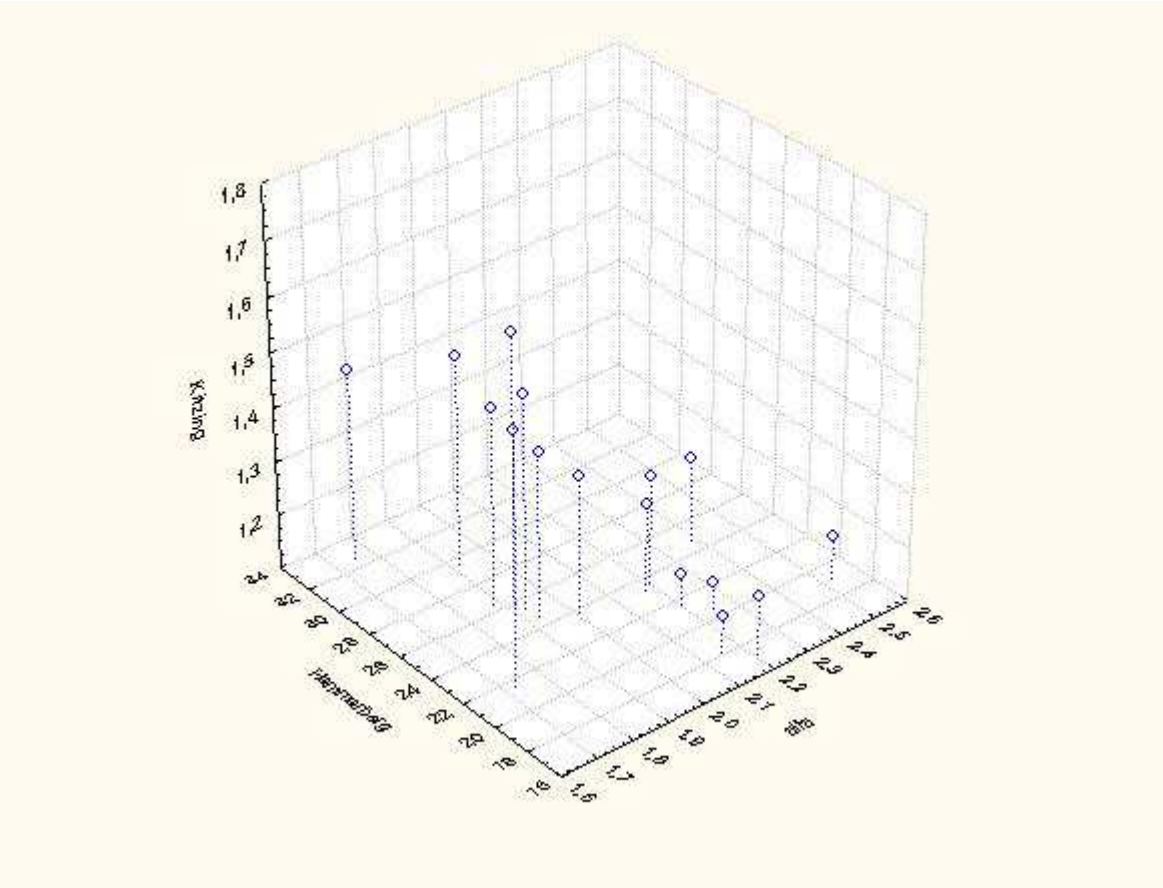


Fig. 25 The values of Kitzing index, Hammarberg index and alpha index for all speakers.

We can see that in a three-dimensional space created by the three indices some speakers are discriminated from other speakers very well, though others occupy a similar region within. However, it must be pointed out that we had only one value for each index per speaker. To be able to assess the long-term measures in more detail, more values for each speaker would be necessary.

7 DISCUSSION

7.1. Short-term measures of spectral tilt

Our speakers have been found to exhibit statistically significant differences in the values of all three parameters, i.e. H1-H2, H1-A1 and H1-A3, which were suggested by Hanson (1997) as acoustical correlates of glottal characteristics. Since these measures have not been examined for their discriminative power before, we shall first compare the values obtained in the present study with the values obtained in previous research and after that discuss the results of our study and its implications in more detail. Let us therefore have a look at how our measured values relate to the values in Hanson's study (1997) and how the results contribute to the aim of the study, i.e. to examine spectral properties of the source signal as possible speaker-specific cues.

7.1.1. Ranges of parameter values

Gobl & Ní Chasaide (1992), for instance, likewise examined the differences of spectral peaks in acoustic spectra vowels, specifically, the peak of the fundamental (L0) and the first four formants (L1-L4). However, as mentioned in Section 3.2.2, they did so by inverse filtering and these data are therefore not directly comparable. Moreover, the material was a single word uttered in different voice qualities to observe how it is reflected in the measures. Since Hanson's (1997) study is the most closely related to ours, we will limit the comparison to her data. For convenience' sake, Table 25 summarizes the ranges obtained by Hanson (1997), which were mentioned in Section 3.2.2, and those of our study.

	Hanson (1997)		present study (2012)	
	Min	max	min	max
H1-H2	-3 dB	7 dB	-8 dB	8 dB
H1-A1	-11 dB	5 dB	-8 dB	4 dB
H1-A3	9 dB	35 dB	20 dB	39 dB

Table 25 Comparison of ranges of the three parameters measured in the present study and in the study of Hanson (1997). The ranges are expressed by their extreme values, i.e. the minimum (min) and the maximum (max).

If we have a look at the ranges, we can see that some are very similar *(H1-A1*) but others differ to a greater extent (*H1-A3*). We can hypothesize that the differences come from

three sources. Firstly, from the material used, secondly, from a slightly different way of measurement and, thirdly, from the differences in glottal configurations for individual speakers.

As for the material, Hanson (1997) inspected only non-high vowels in stressed syllables of utterance-non-final stress groups in a carrier sentence with constant phonemic environment. In addition, she corrected the values for the effect of vowel quality to minimize the differences across vowels (see Section 3.2.2). This decision was motivated by the aim of her study – to examine these parameters as acoustic correlates of glottal characteristics. Since our aim was their applicability for forensic purposes, we used a continuous text to obtain a more natural sample. It also allowed us to take the effects of vowel quality, syllable status with respect to stress and stress group position in the utterance into account to observe their influence on parameter values. Since the influence of these variables on the parameters has been proved by ANOVA (Section 6.1), different ranges can be expected.

The slight differences in the way of measurement, discussed in more detail in Section 5.2, were motivated by the differences in material. As has been pointed out (Chapter 4), Hanson inspected only non-high vowels since their first-formant peak is well separated from the first harmonic. Our study inspected also high vowels, where F1 is lower. Consequently, the first-formant peak was not so well separated. Moreover, it was observed that vowels do not always reach the expected formant values as Skarnitzl and Volín (submitted) pointed out. Due to this fact, unlike Hanson, for whom A1 was the amplitude of the strongest harmonic of the F1 peak, for us it was an amplitude of that harmonic, which lay the closest in frequency to automatically extracted F1 values (in case the automatic extraction was correct), i.e. not necessarily the strongest. Yet the range of H1-A1 in our study and Hanson's study appears surprisingly small. Other factors which may play a role are the differences in H1 across speakers or how well is A1 centered on a harmonic.

Lastly, different ranges arise naturally as a result of different speakers and their diverse glottal characteristics. Some differences could have been also introduced by possible differences in the segmentation technique.

The parameter *H1-A3* exhibits the largest difference in our and Hanson's study. As has been mentioned in Section 5.2, measuring A3 was more complicated. Sometimes no F3 was detected by Praat, which tended to be solved by changing the default settings to a higher value. If such a change did not result in detecting F3, the value was derived by visual inspection of a spectrogram. The generally higher value in our study could also suggest that our speakers have considerably steeper spectral tilt than in Hanson's. This might mean that

some of our subjects have non-simultaneous or incomplete (or both) glottal closure. However, this does not seem to be the case since the correlation between *H1-A1* and *H1-A3* in our study is low. The correlations between the individual parameters in our study are generally lower than in Hanson's study apart from the correlation between *H1-H2* and *H1-A1*, which has been discussed in Section 6.2.4. The comparison of the ranges is given in Table 26. The correlation of values in Hanson's study was introduced in Table 1 (p. 44).

Hanson (1997)				present study (2012)			
	H1*-H2*	H1*-A1	H1*-A3*		*H1-H2*	*H1-A1*	*H1-A3*
H1*-H2*	1	0.53	0.46	*H1-H2*	1	0.68	0.3
H1*-A1	0.53	1	0.68	*H1-A1*	0.68	1	0.28
H1*-A3*	0.46	0.68	1	*H1-A3*	0.3	0.28	1

Table 26 Comparison of correlations as expressed by Pearson correlation coefficient (*r*) in Hanson (1997) and the present study.

An alternative explanation of this difference could be that a carrier sentence "Say bVd again" (Hanson, 1997, p. 475) could result in the fact that speakers gave more emphasis on the word in question and could have pronounced it with an increased loudness, which would be reflected in lowering spectral tilt. Nordenberg & Sundberg (2003) studied the effect of increased vocal loudness on the long-term spectral tilt and found out that the increase of the level is larger at 3 kHz (which is the area of the third formant) than 0.5 kHz.

The comparison of the measured values in our study with other studies is hindered for several reasons. Since the purpose of most studies has been the relation of the physiological function of the vocal folds and/or the perceived voice quality (Gobl & Ní Chasaide, 1992; Holmberg et al., 1995) with the spectra, researchers tend to rule out the factors which can interfere, such as the vocal tract filter, which is solved by inverse filtering. Another strategy is to preserve constant phonemic environment in the form of sustained vowel productions or carrier sentences. Since the purpose of our study was to test the applicability of parameters derived from the spectra of vowels for forensic purposes, these factors needed to be included and examined. Let us therefore have a look how these factors were found to influence the parameter values of individual speakers.

7.1.2. The effect of independent variables on parameter values

Considering syllable status with respect to stress, our speakers have been found to exhibit statistically highly significant differences in the values of the three parameters in all three cases, i.e. not only in stressed but also in post-stressed and unstressed syllables. The effect size for all parameters was the largest for stressed syllables. This could be expected since vowels in stressed syllables are the most stable due to the highest vocal effort and, therefore, the most speaker-specific. However, the fact that speakers exhibit statistically significant differences even in post-stress and unstressed syllables indicates that these parameters do convey some speaker-specific information.

As for stress group position in the utterance, both utterance-final and utterance-non-final stress groups have been found to reflect statistically highly significant differences between speakers; the effect size being larger in the latter for all parameters. This could be explained by the fact that in utterance-final stress groups, vocal effort is more likely to fluctuate and decrease, as a result of which vowels are less stable and exhibit more variability.

Lastly, we examined the effect of vowel quality on parameter values of individual speakers. As Table 25 summarizes (adapted from Table 4, p. 58, Table 7, p. 61 and Table 10, p. 64, where the results for each parameter were first presented), all 5 short vowels have been found to reflect statistically siginificant differences between speakers in the three parameters, though the effect size for individual vowels differs.

	H1-H2	*H1-A1*	*H1-A3*
[ɪ]	F (15, 223) = 3.3243; $p < 0.001$	F (15, 223) = 5.1383; $p < 0.001$	F (15, 223) = 4.1351; $p < 0.001$
[e]	F (15, 191) = 3.6314; $p < 0.001$	F (15, 223) = 6.5986; $p < 0.001$	F (15, 223) = 5.7098; $p < 0.001$
[a]	F (15, 192) = 2.7612; $p < 0.001$	F (15, 224) = 3.8066; $p < 0.001$	F (15, 224) = 4.8101; $p < 0.001$
[o]	F (15, 222) = 2.0608; $p < 0.05$	F (15, 222) = 1.7304, $p < 0.05$	F (15, 222) = 3.7423, $p < 0.001$
[u]	F (15, 221) = 2.1231; $p = 0.01$	F (15, 221) = 2.0647; $p = 0.01$	F (15, 221) = 5.4954; $p < 0.001$

Table 27 The effect of individual vowels on parameters values of individual speakers.

As the table shows, for all three parameters, the effect size was the largest for the front vowel /e/ and the lowest for the back vowel /o/. The second largest effect size was for the other front vowel, /ɪ/, and the second lowest for the other back vowel, /u/. An exception is the parameter H1-A3, where the effect of /u/ was the second largest. The results of our study therefore suggest that front vowels could be more useful for discriminating speakers than back vowels, with the exception of H1-A3. We could hypothesize that this

discrepancy might be caused by a possible interference with the degree roundedness for individual speakers. If a vowel is more rounded, it lowers its F2. In relation to H1-A3, Hanson comments:

> "The amplitude of the third formant is also influenced by other factors, one being the location of F1 and F2. Another is that the bandwidth of F3 is affected by the radiation characteristic to a greater extent than are the lower formants, and the degree of this influence varies with the configuration of the vocal tract for the vowel."
>
> Hanson (1997, p. 469)

It is therefore possible that the larger effect of /u/ for H1-A3 could be caused by conveying some information about the degree to which speakers round this back vowel. However, this is hypothesis is not supported by /o/ , the effect size of which is the lowest for all parameters. To derive any conclusion of the usefulness of individual vowel qualities as carriers of speaker-specific information, a more focused study would have to be conducted.

7.1.3. Linear discriminant analysis

Since all three parameters were found to reflect statistically significant differences between speakers for all vowels in all positions, the whole sample was subjected to LDA, which confirmed the results obtained by ANOVA and offered more insight into the discriminative power of the parameters.

The classification success rate based on the combination of the three parameters was 15.84%, which needs to be compared with classification success rate that would be caused by change, i.e. 6%. The combination of the three parameters has just proved to have some discriminative power. Out of the three parameters, H1-A3 has been found to contribute to the discrimination of speakers in our study the most as its values have been found to yield the smallest intraspeaker variance.

Recognition of speakers proved to improve in stressed syllables (from 15.84% to 19.55%), which confirmed our expectation that in stressed syllables, vowels are the most stable and speaker-specific, and therefore most suitable for discrimination of speakers. Unstressed and post-stress syllables scored considerably lower, 14.61% and 13.10%, respectively. The lower classification success rate in unstressed syllables can be explained by a lower vocal effort which results in higher within-speaker variance. An additional variability appears to be present in post-stress syllables, which scored the lowest. This might be related

to a varying degree of f0 movement on the second, i.e. post-stress, syllable. Yet for certain speakers, it was vowels in post-stress syllables what yielded the most speaker-specific values.

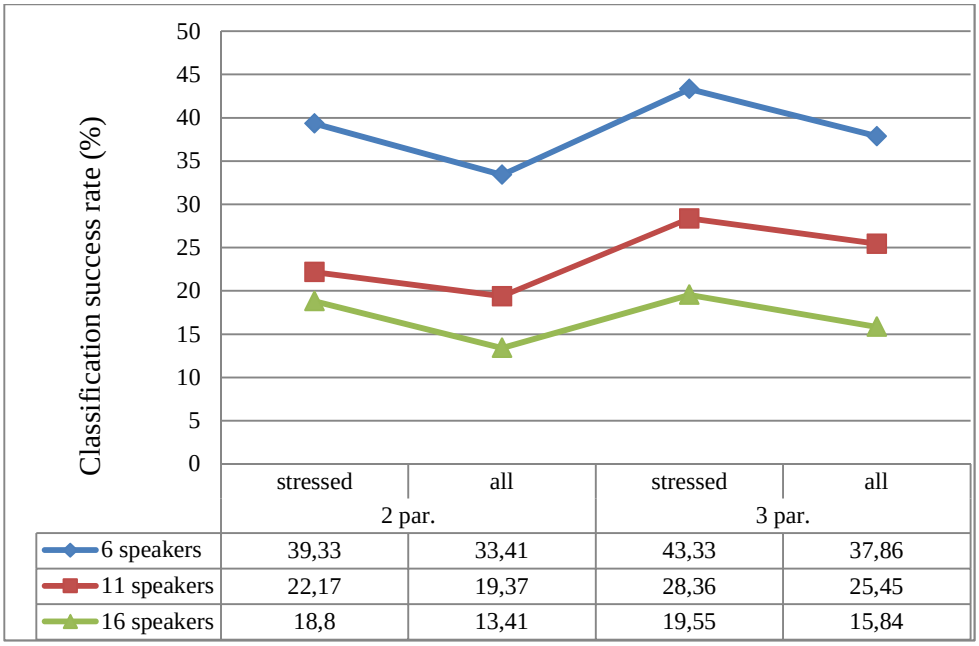
However, our expectation that utterance-final stress groups are less suitable for discrimination of speakers due to decreasing vocal effort and resulting creaky or breathy phonation has been proved only partly. The overall results of our study have shown that removing utterance-final stress groups from the analysis leads to an improvement of the model by about 2%. When inspecting the scores of individual speakers, it has been observed that about half of our speakers exhibits higher classification success rate in utterance-final stress groups. We consequently cannot exclude the possibility, that for other speakers or in real recordings which are used in forensic casework, vowels in utterance-final stress groups would be more speaker-specific. Both should be included in the analysis.

The finding by ANOVA that front vowels could discriminate speakers better than back vowels could not be confirmed by LDA since the sample was too small to yield stable results. Yet we conducted LDA with these limitations on mind. We believe that the considerable difference between the classification success rates of /e/ (20.92%) and /o/ (14.71%) holds some promise for enhancing the discriminative power of these parameters. Moreover, the classification success rate might be again expected to increase when only stressed syllables are considered. For this assumption to be confirmed, a future study would need to examine the discriminative power of individual vowels in more detail and their possible interaction with other factors.

Another interesting finding of our study was the fact that the parameters H1-H2 and H1-A1 overlap to a great extent, specifically, in ⅔ of the cases. Its removal from the analysis proved that the efficacy of the whole model indeed does not change much. The classification success rate decreased from 15.48% to 13.41%, which is still well above the chance classification success rate. It has also been found that the differences between 2-predictor and 3-predictor LDA for different data are not linear; the difference being negligible in stressed and post-stress syllables and the highest in unstressed syllables. Since we concluded that stressed syllables are the most suitable for discrimination of speakers, the usefulness of H1-H2 appears to be questioned in the present study. This is supported by the fact that some speakers were discriminated better by only H1-A1 and H1-A3 than by the combination of all three parameters. However, others exhibited considerably better results with all three parameters. The parameter H1-H2 could also prove more useful with other speakers, for instance, for male speakers. Since f0 of male speakers is approximately a half of f0 of female speakers, we could expect that H2 would be better separated from A1 and the parameter H1-

H2 would contribute to the discrimination of speakers more than in the present study. Yet though the discriminative power of both H1-A1 and H1-A3 has been demonstrated, the contribution of H1-H2 appears less clear.

Another interesting finding by our study was the fact that the three parameters appear to distinguish types of speakers. A parallel has been found between classification success rate of a category and the number of cases it contains. Following the presupposition that if a speaker is frequently assigned to another speaker, there has to be something which "makes him more alike another speaker" than himself or herself, we manually assigned the least successful speakers to those to whom they were most frequently assigned by classification. This has surprisingly led to a decrease in the overall success rate. Yet we believe that this would be worth examining in more detail by a future study.



	2 par.		3 par.	
	stressed	all	stressed	all
6 speakers	39,33	33,41	43,33	37,86
11 speakers	22,17	19,37	28,36	25,45
16 speakers	18,8	13,41	19,55	15,84

Fig. 26 Comparison of classification success rate of LDA with 2 and 3 predictors in stressed syllables and the whole sample for 6, 11 and 16 speakers. The success rate below the graph for individual cases is in %.

The following step was to remove these speakers entirely and observe its effect on the results. Since the improvement of LDA with 3 predictors was larger than of LDA with 2 predictors both in stressed syllables and the whole sample (see Figure 18, p. 80), it appears that the five speakers who were removed were those, whose H1-H2 exhibited the most variable values or, alternatively, those whose H1-H2 and H1-A1 values overlapped the most. Concerning the effect on the success rate of individual speakers, it has been shown that while the improvement was only minor in stressed syllables, the increase of classification success rate was more marked in the whole sample, i.e. for stressed, post-stress and unstressed

syllables together. From that we surmise that while in post-stress and/or unstressed syllables the remaining 11 speakers were also assigned to those 5 which we later decided to removed, this overlap was not present in stressed syllables. This again supports the view that in stressed syllables, the parameter values exhibit smaller within-speaker variance.

Removing 5 more speakers predictably resulted in a further improvement of the model. As Figure 26 shows, the difference between 2-predictor and 3-predictor LDA became more leveled, which would again point to a low significance of H1-H2. However, in the LDA with 6 categories, H1-H2 proved more important for discrimination of speakers than H1-A1, though only in the whole sample. We can therefore conclude that even though H1-A1 and H1-H2 overlap to a great extent, removal of any would lower the efficiency of the whole model. This is also supported by the fact that each parameter distinguishes different speakers. Therefore, though they differ in their importance for the whole model, they all contribute to its efficacy. One more interesting fact to observe in Figure 26 is that the removal of the second five speakers caused a higher improvement than the removal of the first five speakers. Since in real casework only two speakers are compared, it could be expected that the efficacy of the model would be considerably higher. Yet there would be other factors which would hinder the analysis, such as different material, different recording conditions, etc.

7.2. Long-term measures of spectral tilt

Since only one value for each long-term measure of spectral tilt has been obtained in the present study, these results could be used for complementation and comparison with the short-term measures, but not for any conclusion as to their discriminative power.

The values converted to *z*-scores (Figure 23, p. 87) offered us an overview of how the values are distributed and suggested that our speakers significantly differ in their long-term spectral tilt as defined by the three indices. These results indicate considerable differences in the contribution of voice source and vocal tract to voice quality in glottal configurations for our speakers, though, as has been mentioned, their discriminative power could not be assessed. The reason for that is that we had only one value of each index per speaker and research suggests that intraspeaker variability of the LTAS can be considerable (see Section 3.2.1). However, we tried to enhance comparability by long enough samples to factor out the contribution of individual sounds and, in addition, by using the same text for all speakers. Our speakers were likewise instructed to keep constant loudness, though its effect on the LTAS could not be avoided entirely. Having these limitations in mind, we presented a three-dimensional picture of distribution of the values for the three parameters among our

speakers (Figure 25, p. 92). Certain speakers thus appear to be well distinguished from others. However, if more values for each speaker were available - which would be necessary in real casework – it can be expected that the speaker spaces would overlap considerably, though for some speakers more than for others.

When we compared the results of the long-term measures (Figure 23, p. 87) with the classification success rate of LDA, parallels have been found. Specifically, speakers who reached the highest success rates in LDA based on short-term measures of spectral tilt were also those who exhibited the most distinct values of the indices expressing long-term spectral tilt. Similarly, speakers who scored low in LDA yield very average values in long-term measures. In other cases, these two approaches appear to complement each other. For instance, some speakers who were found difficult to distinguish by their short-term measures (according to their squared Mahalanobis distances), appeared distinct by their long-term measures. Long-term measures could thus complement short-term ones by their virtue of factoring out the differences between vowels and providing an average value for a speaker. Their usefulness could be expected to increase in real conditions where obtaining comparable samples is rather exceptional. However, also in case of long-term measures, comparability and the strength of conclusion which can be in a given situation made must be considered.

Nevertheless, this study demonstrates that spectral tilt as quantified by the three short-term measures conveys some speaker-specific information, and that spectral properties of the source signal could thus be possible indicators of one's identity. Yet its applicability for speaker identification purposes would have to be addressed directly in a separate study which would examine its robustness in real-life conditions which are ordinarily encountered in forensic casework.

7.3. Limitations of the present study and suggestions for future research

Since this is a pioneering study, it undoubtedly has many limitations and only opened some of the questions which need to be addressed in order to assess spectral properties of the source signal as speaker-specific cues.

Already in the measurement section, it has been pointed out that H2 often equals A1. The degree of this overlap has been later quantified by correlations which revealed that the measures H1-H2 and H1-A1 overlap in about two thirds of cases, especially in high vowels, where F1 is the lowest. This presents a considerable drawback to our study and possible corrections or alternatives might be hypothesized. One option could be to stick to Hanson's methodology, i.e. that A1 is the amplitude of the strongest harmonic of the first-formant peak.

To do so consistently, expected (and extended) frequency ranges would have to be stated in advance. Another option would be to test other spectral relations as possible indicators of speaker identity. Some researchers suggest that ratios of formant frequencies, such as F1/F2 and F2/F3 (Hollien, 1990, p. 42; Skarnitzl, 2012, in print) convey some speaker-specific information as the ratios are not changeable at will. Since there is some relation between frequency and bandwidth as well as bandwidth and amplitude, these findings could thus provide a basis for the search of other possible parameters.

Though this study addressed the influence of vowel quality, syllable status with respect to stress and stress group position in the utterance, the conclusion drawn were not definite and other studies could examine the influence of these variables on the parameter values in more detail. This is suggested by the observation that despite the fact that overall success rate was reached in stressed syllables, some speakers scored higher in post-stress or even unstressed syllables. Similarly, the overall classification success rate was higher in utterance-non-final stress groups but half of our speakers exhibited more speaker-specific values in utterance-final ones. The assumption that individual vowel qualities could differ in their discriminative power could not be tested due to a low number of instances of individual vowels. We consider this possibility worth examining in more detail. Importantly, the possibility has not been falsified that front vowels are simply those where H1-H2 and H1-A1 overlapped the least. This should have been attended to. Other variables which could be studied for their effect on parameters and have not been addressed in the present research could be, for instance, the position of intonation phrase boundary.

The possibility that the three parameters could discriminate types of speakers could likewise be addressed in more detail. This is what the classification matrix appeared to indicate. Importantly, the fact that these parameters do not discriminate our 16 speakers should not diminish their potential. The results need to be seen in context. Since all the three parameters quantify the short-term spectral tilt, it could not be expected that they will discriminate 16 speakers alone. Its power will be enhanced when combined with other parameters or indicators of speaker identity mentioned in the theoretical part, such as formant frequencies, temporal structuring, and others. The intra- and interspeaker variability should also be examined in more detail to show how these parameters relate to the concept of speaker space mentioned in the theoretical part of this study.

Furthermore, if the parameters were to be used for forensic purposes, their robustness in real-life conditions normally encountered in forensic casework, such as the presence of speaker distortion or system distortion, would have to be examined. Other limitations likewise

need to be considered; for instance, the fact that often the material available for comparison is very diverse and many variables cannot be controlled.

As for the long-term measures, the main drawback was that only one value for each index per speaker has been obtained. Having more values of the indices would allow us to make a better picture of how the values for our speakers are distributed in a multidimensional space created by these parameters. The significant advantage of the long-term measures of spectral tilt is that they can be measured automatically. If these indices proved to convey some speaker-specific information, its application would also be beneficial for speaker verification or speech synthesis.

Lastly, let us quote Rose (2002) who comments:

> Irrespective of the type of the parameter, the ultimate question remains the same. Given these speech samples, what is the probability of observing this difference for this parameter assuming the samples come from the same speaker, and what is the probability of observing the difference assuming different speakers?
>
> Rose (2002, p. 65)

Nevertheless, before its usefulness for forensic purposes can be discussed, these parameters need to be examined in both studies that carefully control any undesirable variables, and in studies that use more real-life samples. Finally, as has been noted in the theoretical part, researchers have to be aware of the limitations of acoustic analysis which should be always complemented by auditory one.

8 CONCLUSION

As has been discussed in the theoretical part, speaker identification remains one of the most challenging tasks of forensic phonetics due to the fact that our knowledge of how individuality is reflected in a voice is still limited. This is complicated by the fact that a human voice is far from constant. It is susceptible both to speaker- (volitional as well as non-volitional) and system distortion. Consequently, it cannot be said with certainty whether intraspeaker is always smaller than interspeaker variation, in all situations and under all conditions. The conclusions that one can make are therefore always tentative, never absolute. The strength of conclusions depends on each individual case and on reference information from the field of phonetics, sociolinguistics, and other.

Finding speaker-specific cues enhances speaker identification because it adds dimensions to a space within which each speaker occupies certain space; it makes the space more definite. Adding dimensions could arguably result in the fact that speakers stop overlapping. The present study aims to add to such a research.

Apart from the use in speaker identification, finding reliable indicators of speaker's identity would be beneficial also in speaker verification. Nowadays it is possible to access personal information, such as a bank account, by voice command. Examining speaker-specific cues and their robustness, such as its resistance to voice disguise, can therefore bring an improvement into this area, too.

Previous research has pointed out several domains which hold promise as conveying some speaker specific information. These include the acoustical properties of certain segments, both consonants and vowels. Other studies focused on the suprasegmental level and revealed that temporal structuring or melodic patterns can also offer some clues to speaker's identity. Since recent research in prosody suggests that voice quality or phonatory modulations is used for paralinguistic purposes such as fundamental frequency but independently of it, it is possible that some personal idiosyncrasies can likewise be discovered in phonatory modifications or voice quality. The present study sought to examine this possibility.

The LTAS has been considered the most relevant tool for quantifying voice quality; long-term spectral tilt has been directly related to perceived differences in voice quality by previous research. The parameters used for its quantification express the ratio of the amount of energy in certain frequency bands, such as alpha index and Kitzing index, or the difference between two spectral peaks, such as Hammarberg index. Another set of parameters which

reflects differences in glottal configuration is derived from the acoustic spectra of vowels and quantifies short-term spectral tilt. Our study focused on examining the latter group of parameters; specifically, their usefulness for discriminating 16 Czech female speakers. The results of the analysis were then compared with the results obtained by the long-term measures.

Our study has shown that speakers exhibit statistically significant differences in the values of these parameters for all vowels in stressed, post-stressed as well as unstressed syllables of both utterance-final and utterance-non-final stress groups. A subsequent LDA allowed us to examine the data in more detail.

The effect of stress on discrimination of speakers has been confirmed. Classification success rate was generally the highest in stressed syllables, though certain speakers exhibited more speaker-specific values in post-stress syllables. Similarly, though utterance-non-final stress groups in general yield higher recognition rate, a half of the speakers diverges from this trend. An interesting fact pointed out by this study is the possibility that certain vowels are more suitable for discriminating speakers than others. The results of ANOVA for every parameter revealed the largest effect size for /e/ and the lowest for /o/. With the exception of the parameter H1-A3, the results appear to indicate that front vowels could be more speaker-specific than back vowels. However, this hypothesis could not be confirmed or disproved by LDA due to an insufficient sample, and would need to be further tested.

Another question which this study has opened is whether these parameters would distinguish types of speakers. Future research could thus focus on why certain speakers are assigned to others. One more interesting finding is the degree of overlap of two parameters, namely H1-H2 and H1-A1, and the fact that all parameters contribute to speaker identification to a different degree. H1-H2 has been found the least useful, in case of some speakers it even caused a decrease in classification success rate. However, it would need to be further tested on different material, or under different conditions to make any conclusion about its usefulness.

Certain parallels have been found between short-term and long-term measures of spectral tilt suggesting that these measures do reflect individual differences in glottal configuration and voice quality. Though many questioned remained unasked and those asked give only tentative answers, this study will meet its purpose if it opens another area where speaker-specific cues could be found, and motivates further research in finding them.

REFERENCES

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

Baldwin, J. & French, P. (1990). *Forensic Phonetic*s. London: Pinter Publishers.

Bocklet, T., Maier, A. & Nöth, E. (2007). Text-independent speaker identification using temporal patterns. In: Matoušek, V. & Mautner, P. (Eds.), TSD 2007, LNAI 4629, pp. 318-325. Berlin: Springer Verlag.

Boersma, P. & Weenink, D. (2010). Praat - Doing phonetics by computer (Version 5.1.31.). Downloaded in November, 2010, from http://www.praat.org/.

Brockmann, M., Storck, C., Carding, P. N. & Drinnan, M. J. (2008). Voice loudness and gender effects on jitter and shimmer in healthy adults. *Journal of Speech, Language and Hearing Research*, 51, pp. 1152-1160.

Butcher, A. R. (2002). Forensic Phonetics: Issues in speaker identification evidence. *Proceedings of the Inaugural International Conference of the Institute of Forensic Studies*, Italy.

Campbell, N. & Mokhtari, P. (2003). Voice quality: the 4th prosodic dimension. *Proceedings of the 15th ICPhS, Barcelona*: Conference Organizers, pp. 2417-2420.

da Silva, P. et al. (2010). Acoustic and Long-Term Average Spectrum Measures to Detect Vocal Aging in Women. *Journal of Voice*, 25, pp. 411-419.

Dejonckere, P. (1998). Effect of louder voicing on acoustical measurements in dysphonic patients. *Logopedics Phoniatrics Vocology*, 23, pp. 79–84.

Dellwo, V. & Koreman, J. (2008). How speaker idiosyncratic is measurable speech rhythm? *Proceedings, IAFPA 2008, Swiss Federal Institute of Technology Lausanne (EPFL).*

Fernandez, R. & Picard, R. W. (2005) Classical and Novel Discriminant Features for Affect Recognition from Speech. Interspeech 2005, September 4-8, Lisbon, Portugal, pp-. 1-4.

Ferrand, C. (2002). Harmonics-to-Noise Ratio: An Index of Vocal Aging. *Journal of Voice*, 16, pp. 480-487.

Foulkes, P. & French, J. (2001). Forensic Phonetics and Sociolinguistics. In: Mesthrie, R. (ed.) *Concise Encyclopaedia of Sociolinguistics*. Amsterdam: Elsevier Press, pp. 329-332. http://www-users.york.ac.uk/~pf11/foulkes&french.pdf. (Last accessed: September 15, 2011).

Frøkjær-Jensen, B. & Prytz, S. (1976). Registration of voice quality. *Technical Review*, 3, pp. 3-17.

Gobl, C. & Ní Chasaide, A. (1992). Acoustic measurements of voice quality. S*peech Communication*, 11, pp. 481-490.

Gobl, C. & Ní Chasaide, A. (2003). The role of voice quality in communicating emotion, mood and attitude. *Speech Communication*, 40, pp. 189-212.

Hammarberg, B. et al. (1980). Perceptual and acoustic correlates fo abnormal voice qualities. *Acta otolaryngol*, pp. 441-451.

Hanson, H. M. (1997). Glottal characteristics of female speakers: Acoustic correlates. *Journal of Acoustical Society of America*, 101, pp. 466-481.

Harmegnies, B. & Landercy, A. (1988). Intra-speaker variability of the long term speech spectrum. *Speech Communication*, 7, pp. 81-86.

Hillebrand, J. & Cleveland, R. A. (1994). Acoustic correlates of vocal quality. *Journal of Speech and Hearing Research*, 37 (4), pp. 769-778.

Hollien, H. (1990). *The acoustics of crime*. New York: Plenum Press.

Hollien, H. (2002). *Forensic Voice Identification*. San Diego: Academic Press.

International Association for Identification. http://www.theiai.org/. (Last accessed on October 10, 2011).

Jessen, M. (2007). Forensic reference data on articulation rate in German. *Science and Justice*, 47, pp. 50-67.

Jessen, M. (2008). Forensic Phonetics. *Langauge and Linguistics Compass* 2/4, pp. 671-711.

Johnson, K. (2004). Massive reduction in conversational American English. In: K. Yoneyama & K. Maekawa (Eds.), *Spontaneous Speech: Data and Analysis* (pp. 29-54). Tokyo: The National Institute for Japanese Language.

Kitzing, P. (1986). LTAS criteria pertinent to the measurement of voice quality. *Journal of Phonetics*, 14, pp. 477-482.

Kitzing, P. & Åkerlund, L. (1993). Long-time average spectrograms of dysphonic voices before and after therapy. *Folia Phoniatrica*, 45, pp. 53-61.

Kreiman, J. & Garrett, B. R. (2003). Jitter, Shimmer, and Noise in Pathological Voice Quality Perception. *VOQUAL'03, Geneva, August 27-29, 2003*, pp. 57-61.

Kreiman, J., Vanlancker-Sidtis, D. & Garrett, B. (2003). Defining and measuring voice quality. *VOQUAL'03 Geneva, August 27-29, 2003*, pp. 115-119.

Kreiman, J. & Sidtis, D. (2011). *Foundations of Voice Studies: An interdisciplinary approach to voice production and perception*. Chichester: Wiley-Blackwell.

Leino, T. (2008). Long-term average spectrum in screening of voice quality in speech: untrained make university students. *Journal of Voice*, 23, pp. 671-676.

Lindh, J. & Eriksson, A. (2007). Robustness of Long Time Measures of Fundamental Frequency. In: *Proceedings of the 8th Annual Conference of the International Speech Communication Association (Interspeech 2007), Antwerpen: ISCA*, pp. 2025-2028.

Löfqvist, A. (1986). The long-time-average spectrum as a tool in voice research. *Journal of Phonetics*, 14, pp. 471-474.

Machač, P. & Skarnitzl, R. (2009). *Fonetická segmentace hlásek*. Praha: Epocha.

Master et al. (2006). The long-term average spectrum in research and in the clinical practice of speech therapists. *Pró-Fono Revista de Atualzacao Científica,* 18, pp. 111-120.

McDougall, K. (2006). Dynamic features of speech and the characterization of speakers: towards a new approach using formant frequencies. *Speech, Language and the Law*, 13, pp. 89-126.

Michaelis, D., Gramss, T. & Strube, H.W. (1997). Glottal-to-Noise Excitation Ratio – a New Measure for Describing Pathological Voices, *Acustica*, 83, pp. 700-706.

Monzo , C., Alías, F., Iriondo, I., Gonzalvo, X. & Planet, S. (2007). *Discriminating expressive speech styles by voice quality parametrization. ICPhS XVI Saarbrücken, 6-10 August 2007*, pp. 2081-2084.

Nolan, F. (1983). *The phonetic bases of speaker recognition*. Cambridge: Cambridge University Press.

Nolan, F. (1999). Speaker Recognition and Forensic Phonetics. In: William, J. and Laver, J. (eds) *The Handbook of Phonetic Sciences*. Hardcastle: Blackwell Publishing. Blackwell Reference Online. 28 December 2007 http://www.blackwellreference.com/subscriber/tocnode?id=g9780631214786_chunk_g978063121478625. (Last accessed on October 12, 2011).

Nordenberg, M., & Sundberg, J. (2003). Effect on LTAS of vocal loudness variation. TMH-QPSR, KTH, 45, pp. 93-100.

Palková, Z. & Volín, J. (2003). The role of F0 contours in determining foot boundaries in Czech. In: *Proceedings of the 15th ICPhS*, pp. 1783-1786. Barcelona: Organizing Committee.

Perrot, P., Aversano, G. & Chollet, G. (2007). Voice Disguise and Automatic Detection: Review and Perspectives. In: Stylianou, Y., Faundez-Zanuy, M. & Esposito, A. (Eds.) *Progress in nonlinear speech processing*. Springer-Verlag Berlin Heidelberg, pp. 101–117.

Perrot et al. (2009). Vocal Forgery in Forensic Sciences. http://razik.univ-tln.fr/perrot09.pdf. Last accessed on September 20, 2011.

Qi, Y. & Hillman, R. E. (1997). Temporal and spectral estimations of harmonics-to-noise ratio in human voice signals. *Journal of the Acoustical Society of America*, 102, pp. 537-543.

Rose, P. (2002). *Forensic Speaker Identification*. London: Taylor & Francis.

Sergeant, D. & Welch, G. (2008). Age-related changes in long-term average spectra of children's voices. *Journal of Voice*, 22, pp. 658-670.

Schoentgen, J. & de Guchteneere, R. (1995). Time series analysis of jitter. *Journal of Phonetics*, 23, pp. 189-201.

Skarnitzl, R., & Volín, J. (submitted). *Referenční hodnoty českých vokalických formantů*.

Sundberg, J. & Nordenberg, M. (2006). Effects of vocal loudness variation on spectrum balance as reflected by the alpha measure of long-term-average spectra of speech. *Journal of the Acoustical Society of America*, 120, pp. 453-457.

Tanner, K. et al. (2005). Spectral moments of the long-term average spectrum: sensitive indices of voice change after therapy? *Journal of Voice*, 19, pp. 211-222.

Ternström, S. O. (2008). Hi-Fi voice: observations on the distribution of energy in the singing voice spectrum above 5kHz. *Acoustics'08, Paris, June 29-July 4, 2008*.

van Dommelen, W.A. (1987). The contribution of speech rhythm and pitch to speaker recognition. *Language and Speech,* 30, pp. 325-338.

Volín, J. (2007). *Statistické metody ve fonetickém výzkumu.* Praha: Epocha.

White, P. (2001). Long-term average spectrum (LTAS) analysis of sex- and gender-related differences in children's voices. *Logopedics Phoniatrics Vocology*, 26, pp. 97-101.

Yumoto, E., Sasaki, Y. & Okamura, H. (1982). Harmonics-to-Noise Ratio and Psychological Measurement of the Degree of Hoarseness. *Journal of the Acoustical Society of America,* 71, pp. 1544-1549.

APPENDIX

TEXT

Já ti řeknu, co uděláš. Nejdřív najdeš Hučku a Atamana.
Dobře. A až je najdu?
Řekneš jim, co si myslíš.
To asi budou dost nadávat, co?
To bych se nebál.

Začni opatrně. A o další spolupráci bych se nezmiňoval.
To by mě ani nenapadlo.
Řekneš jim, co si myslíš?
Nevím, snad nebudu muset.
Dobře, nezapomeň, hlavní heslo: *opatrnost*.

Nejlepší bude, když zatroubíte a počkáte, až vylezou.
Hmm, no a potom?
Řeknete jim, co si myslíte.
A vy na nás počkáte?
Jasně, ani se nehneme z místa.

Možná budou problémy. Promysleli jste si to pořádně?
Jo, krok za krokem.
Řeknete jim, co si myslíte?
Určitě, hnedka zkraje.
No, připravte se, že budou prskat.

Takže se posadíš a budeš se usmívat. Žádnou paniku.
Jasně. A co mám dělat, až přijde ten jejich šéf?
Zeptáš se, co bude dál.
To je všechno? Nemám mu říct, že už to víme?
Ne, nic nevíš. Ani slovo.

Už se někdo ozval?
Ne. Všichni to vědí, ale vesele se předstírá, že nic.
To je teda situace. Zeptáš se, co bude dál?
Budu muset. Jinak budeme předstírat a předstírat, až už bude pozdě.
Hmm, to ti nezávidím.

Když budou chtít napřed vidět peníze, řekněte jim, že jsou na cestě.
Jasně. A čím máme začít, až nás vezmou dovnitř?
Zeptáte se, co bude dál.
Myslíš, že v tom jedou všichni?
Určitě. Je to jejich priorita.

Máte nějakou představu, jak to bude probíhat?
Zhruba. Od nich přijdou taky tři.
To by mohlo zaskřípat. Zeptáte se, co bude dál?
Zeptáme, ale ne hned zkraje. Musíme pomalu.

Až začnou o tom transportu, nastražíš uši.

Vždyť mě znáš. Jen mě trochu nasměruj. O co jde?
Zjistíš, v kolik to pojede.
No, to je snad samo sebou.
Jo, ale potřebujeme to přesně. Hodně přesně.

Tak, ještě kartáček na zuby a je to. Snad mám všechno.
Dobře, chceš ještě nějak pomoct?
Zjistíš, v kolik to pojede?
No, na ty informace jsem volal, ale pořád bylo obsazeno.
Na internetu to není?

Trochu se bojím, že je tím rozčílíte.
Aha, my asi potřebujeme, aby byli v klidu. O co půjde?
Zjistíte, v kolik to pojede.
No jo, ale to je to poslední, co by nám chtěli prozradit.
Právě proto žádný provokace. Klid.

Přestaňte se mi tu motat pod nohy. Nevím, co dřív.
Ale my jsme přišli pomáhat. A na to nádraží taky dojdeme.
Zjistíte, v kolik to pojede?
Dobře, něco po poledni, jo?
Jo. Určitě před druhou.

Moc se neošívej, ale dělej, že to je pro tebe novinka.
Proč? Hrozí něco?
Nemáš ponětí, kdo pojede vzadu.
Jasně, chápu.
Hlavně buď bez obav, už jsme zvládli horší věci.

Hodilo by se trochu víc informací.
Já už nemám čas na to myslet.
Nemáš ponětí, kdo pojede vzadu?
Nevím. Doufám, že ne ten idiot Kukla.
To by byl kolosální malér. Jen ne Kukla.

Začíná mě mrazit v zádech. Vy jste v pohodě?
Jasně. Ty myslíš, že by se to mohlo zvrtnout?
Nemáte ponětí, kdo pojede vzadu.
No, to nemáme. Snad ne posily.
Když tak radši nic nezkoušejte. Ještě bude spousta šancí.

Kam bych to měl soustředit? Máte nákresy?
Kousek za půlku. A nebo prostředek.
Nemáte ponětí, kdo pojede vzadu?
Krmivo, pomocná síla a sanitka.
Jo, takže ne moc za půlku.

A pamatuj: o místě určení ani slovo.
Takže se nemám zapojovat?
Nedávej na sobě nic znát. Netušíš, kam to přesouvají.
Jasně. Mám je k tomu nějak nasměrovat?
Ne. Až se o tom začnou bavit, tvař se překvapeně.

Kdo ví, kam až se s tím povlečeme.

A to se to ještě může zkomplikovat kvůli ostraze.
Netušíš, kam to přesouvají?
Ne. Ale dělají s tím zbytečně tajnosti.
Jako by to bylo bůhví co.

Takže všechno to teď závisí na vás.
No jo, ale jak se k tomu máme stavět.
Nesmíte ani mrknout. Netušíte, kam to přesouvají.
Jo, to se lehce řekne. Nám se klepou ruce už teď, že jo, Frede?
Klid copak jste začátečníci?

To je teda náklaďáků. To snad nemá konce.
Hmm. Na tohle si čas a prostředky najdou.
Netušíte, kam to přesouvají?
Někam za Rudnou. Ale tam se teď nikdo nedostane.
A ze vzduchu nic vidět není?

Tentokrát nemají šanci. Máš to v kapse.
Proč myslíš, že jsem ve výhodě?
Víš, kdy to dostaneš.
No, to mi pár minut získá. Ale jinak nevíme nic.
Já ti říkám, že je to v suchu.

Měl bys mít všechno po ruce. Bude zmatek.
No jo. Už jsem si to kontroloval nejmíň pětkrát.
Víš, kdy to dostaneš?
Hned jak Batulka zavře bránu.
Teda tam bych se nechtěl přimotat.

Máte jeden trumf, a na tom se dá vydělat.
Jak to?
Víte, kdy to dostanete.
No jo, to je pravda.
A právě s tím nikdo nepočítá.

Jen abyste měli dost peněz, až to přijde.
Bez obav. S tím se počítá.
Víte, kdy to dostanete?
No, někdy po druhý hodině. Až zmizí první směna.
Musíte hrát hodně opatrně. Nepřehánět sázky.