Univerzita Karlova v Praze

Filozofická fakulta

Fonetický ústav

Diplomová práce

Bc. Michaela Rabanová

Intra- and inter-speaker variability of Czech noise segments

Variabilita českých šumových hlásek v rámci mluvčího a mezi mluvčími

Praha 2024 Vedoucí práce: doc. Mgr. Radek Skarnitzl, Ph.D.

Acknowledgments

I would like to express my sincere gratitude to doc. Mgr. Radek Skarnitzl, Ph.D., for his kind guidance, insightful feedback, and support throughout the process of working on this thesis. I would also like to thank the International Association for Forensic Phonetics and Acoustics (IAFPA) for providing me with a research grant.

Prohlášení:

Prohlašuji, že jsem bakalářskou 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.

V Praze, dne 19. srpna, 2024 Michaela Rabanová

Abstract:

This thesis aims to provide population statistics for the spectral moments of four Czech voiceless fricatives: alveolar fricative [s], postalveolar fricative [ʃ], velar fricative [x], and the voiceless allophone of the Czech fricative trill [r̝̊]. The goal was to improve the efficiency of assessing typicality of spectral moments values within the Czech male population and to examine how the spectral moments are affected by two types of telephone transmission— using narrowband and wideband codecs. The study is divided into theoretical and practical parts. The theoretical part introduces forensic phonetics, focusing on voice comparison and key concepts like similarity and typicality. It also provides a theoretical background for fricatives, emphasizing their articulatory and acoustic descriptions. The practical section describes the methodology for obtaining the population statistics. The results included detailed population statistics of spectral moment derived from analysing recordings of semi-spontaneous speech from 60 male speakers. The hypothesis that spectral moments would be significantly altered in narrowband codec simulations was supported, while the hypothesis for wideband codec changes was not fully confirmed due to the inability to determine statistical significance from the analysis. Future research is required to address this gap.

Abstrakt:

Cílem této práce je poskytnou populační statistiku pro spektrální momenty čtyř českých neznělých frikativ: alveolární frikativy [s], postalveolární frikativy [ʃ], velární frikativy [x] a neznělého alofonu české frikativní vibranty [r̝̊], aby šlo lépe určit jejich typičnost ve forenzním kontextu. Dále se také snaží zjistit, jak jsou spektrální momenty daných frikativ ovlivněny dvěma typy telefonních přenosů, a to přenosů s nízkopásmovým a širokopásmovým kodekem. Práce je rozdělena na teoretickou a praktickou část. V teoretické části je popsána forenzní fonetika a klíčové pojmy s ní spojené jako např. podobnost a typičnost. Dále teoretická část poskytuje frikativy a zaměřuje se především na jejich artikulační a akustický popis. Praktická část popisuje metodu, která byla použita k získání populační statistiky. Detailní populační statistika spektrálních momentů pro [s], [ʃ], [r̝̊] a [x] je prezentována ve výsledcích. Hypotéza, která předpokládala, že spektrální momenty budou v simulaci úzkopásmového přenosu silně ovlivněny, se potvrdila. Statistickou významnost změn v hodnotách spektrálních momentů v simulaci širokopásmového přenosu nebylo možno potvrdit, jelikož ve studii nezbyl prostor a čas na komplexní statistickou analýzu.

Keywords

[spectral moments, fricatives, forensic phonetics, telephone signal]

Klíčová slova

[spektrální momenty, frikativy, forenzní fonetika, telefonní signál]

Table of contents

List of tables

List of figures

1. Introduction

In current practice, forensic phoneticians mostly agree that a combination of auditory and acoustic analysis is needed to perform voice comparison effectively. This methodology is referred to as the phonetic-acoustic approach (Rose, 2002, p. 48). It involves carefully listening to the speech samples in question and filling up some type of protocol, e.g. the Vocal Profile Analysis (VPA; Laver, 1980), focusing on various dimensions of the speaker's voice and speech. The auditory analysis serves to identify the parameters that may subsequently be analysed and quantified in a more objective manner using acoustic measurements.

It is important to note that here the research on voice comparison has been focusing predominantly on analysing vocalic parameters rather than consonantal parameters (Schindler & Draxler, 2013, p. 2793). However, regardless of whether vocalic or consonantal parameters are analysed in a forensic context, it is crucial to assess their similarity and typicality. To be able to assess the latter of the two, information about the behaviour of the examined parameters in the relevant population must be available. Obtaining such population statistics is possible by accessing an already existing database. However, the number of parameters for which such a database is available remains limited.

This thesis aims to provide population statistics for one of the consonantal parameters whose assessment does not seem to be clear, specifically spectral properties of noise segments of Czech voiceless fricatives. The study will rely on parametrizing the acoustic spectrum by calculating the first four spectral moments: the center of gravity (COG), standard deviation, skewness, and kurtosis. The voiceless fricatives that will be the target of the analysis are the alveolar fricative [s], the postalveolar fricative [ʃ], the velar fricative [x], and the voiceless allophone of the famous Czech fricative trill ř, [r̝̊]. The reason for selecting these segments is their frequent presence in ordinary speech and their high potential of carrying highly idiosyncratic properties, which has been informally documented by the forensic casework experience of the supervisor of this thesis. The mentioned fricatives will be extracted from 60 approximately one-minute-long recordings of semi-spontaneous speech from male speakers, and subsequently, they will be analysed. Considering there is tentative evidence that the acoustic parameters of fricatives retain their speaker-discriminating potential in telephone speech (Smorenburg & Heeren, 2020; Christensen, 2023), the 60 recordings will undergo two types of telephone simulation – one simulating a telephone using a narrowband codec and the other simulating a telephone using a wideband codec. The acoustic parameters of the target fricatives in the simulated conditions will then be compared with those in the original recordings.

The hypothesis is that the acoustic parameters of the target fricatives will be significantly altered in the narrowband codec simulation. This is expected because the narrowband codec implements a bandwidth of 300 Hz to 3,400 Hz (Christensen, 2023, p. 54), which results in a loss of much of the high-frequency information. While the acoustic parameters of the fricatives will most likely be altered in the wideband codec simulation as well, the extent of the change cannot easily be predicted. However, the changes are expected to be less extreme compared to the narrowband codec simulation.

The theoretical part of this study will firstly introduce forensic phonetics. A description of the field and its areas of interest will be given, as well as an explanation of voice comparison, the methods associated with it, and key concepts such as similarity. It will also address how results are expressed in forensic phonetics, and what common difficulties the experts encounter in voice comparison. Secondly, a chapter dedicated to fricatives will be included. It will cover their articulatory and acoustic description, and it will also explore fricatives as idiosyncratic cues. Lastly, the effect of telephone transmission on fricatives will be explained. The practical part of this study will describe the methodology that was used when analyzing the acoustic properties of the four target fricatives present in the recordings of the 60 speakers. Finally, population statistics on the spectral moments of the four target fricatives will be presented in the results section. A comparison of the spectral moments across three different conditions: original recordings, telephone narrowband simulations, and telephone wideband simulations will also be given.

2. Theoretical background

2.1 Forensic phonetics

As already stated, this study aims to provide more detailed information about the distribution of values of acoustic parameters of Czech voiceless fricatives. Having such population statistics would be beneficial for forensic phonetics, as it could help to determine what values are typical or, on the contrary, rare for local demographics. Hoping to enrich the field of forensic phonetics, it is first necessary to provide a theoretical background for the field. Firstly, this chapter will contain a definition of what forensic phonetics is and what it is concerned with. Next, there will be a section dedicated to a central subfield of forensic phonetics called voice comparison. Within this section, important concepts of similarity, typicality, between-speaker and within-speaker variation, and Likelihood Ratio will be introduced and explained as well as the methodology typically used when conducting voice comparison. Lastly, there will be a section describing some of the common difficulties that forensic phoneticians may encounter in their analyses.

2.1.1 Definition and areas of interest

Forensic phonetics can be defined as “the application of the knowledge, theories, and methods of general phonetics to practical tasks that arise out of a context of police work or the presentation of evidence in court” (Jessen, 2008, p. 671). It is said that the term “forensic phonetics” itself has been used approximately since the foundation of the International Association for Forensic Phonetics (IAFP) in 1991. In 2004, one word was added to the name of the association, and it is currently known as the International Association for Forensic Phonetics and Acoustics (IAFPA).

It was stated in the 1990s already that more and more legal cases include recorded samples of speech (Nolan, 1991, p. 483). Even a decade later, Rose observed that opinions of forensic phoneticians are increasingly being sought (2002, p. 17). The recorded materials typically presented in legal processes include “hoax calls to emergency services, obscene calls, fraudulent deals negotiated over the telephone, or ransom demands” (Nolan, 1991, p. 483). When there is recorded material from the crime scene, the next step in the analysis depends on whether there is a known suspect or not. If there is a suspect, we are talking about **voice comparison**. This approach is key in this study and therefore there will be an entire chapter dedicated to it. It is also the most common type of analysis in forensic phonetics (Christensen, 2023, p. 63). If there is no suspect, we are talking about **voice profiling**. This method is usually used at the beginning of the police investigation. The voice profile of the recorded person is done by the forensic expert to narrow down the range of possible suspects. Depending on the quality and length of the recording, the voice profile can provide information about the person's region of upbringing, age, sex, level of education, social background, native language (in case there is a foreign accent present in the speech) and even medical condition that affects speech (Jessen, 2008, p. 674). Another possible scenario may happen when there is no recorded material, and the only available evidence is the witness of the crime. Such method is called **speaker identification by victims and witnesses.** Since there is no recording, the forensic experts are working with the witness's or the victim's perception of the perpetrator's voice only. In such cases, it is crucial to interview the witness or the victim as soon as possible after the criminal act took place because as Jessen states "memory for voice identities decays rapidly" (2008, p. 579). All these three approaches are subcategories of **forensic speaker identification.**

Forensic speaker identification, however, is not the only area of interest of forensic phoneticians. In more recent years, expert opinions have also been required to address whether a recording was manipulated by either inserting, deleting, or changing a certain passage (Jessen, 2008, p. 671). Naturally, such tasks require the experts to have strong interdisciplinary ties namely to speech technology and general acoustics. Knowledge in these disciplines may also be used in cases where the recordings in question are of poor quality and the experts are asked to enhance it with various filtering techniques to e.g. improve the intelligibility. Another area where speech technology expertise has become important is e.g. automatic speaker identification (Jessen, 2008, p. 671).

In other cases, forensic phoneticians may also be asked to analyse non-speech acoustic events e.g. gunshots, background noises such as birds singing, or even sounds from the cockpit of an aircraft (Jessen, 2008, p. 671). Lastly, forensic phoneticians (just like regular phoneticians) are expected to have fundamental expertise in psychology and physiology (Hollien, 2012, p. 27).

As it was described in this brief overview, the field of forensic phonetics has many areas of interest, and it is crucial for the analysts to acquire expertise in multiple varying fields. However, it is undeniable that speaker identification is the central aspect of forensic phonetics (Jessen, 2008, p. 673). As already mentioned, one of the subcategories of speaker identification, voice comparison, is of great importance for this study and therefore it will be discussed in detail in section 1.2 below.

2.2 Voice comparison

In order to use the method of voice comparison, two conditions must be met. Firstly, there must be a recording of an unknown speaker who is associated with a crime. Some examples of what type of crimes these might be were already stated in section 1.1. Jessen offers a few more examples e.g. drug dealing arrangements over the phone or stalking (2008, p. 673). Such evidential recording is referred to as a **disputed sample.** Secondly, there must be a person who is suspected to be involved in the crime and is suspected to be the same speaker as the one from the disputed sample. When a suspect is caught, typically the police must obtain recorded speech material which can then be compared with the disputed sample. Such recording obtained from the suspect is called a **known sample**. If the suspect is not cooperative, depending on the legal system of each country, a tapped telephone conversation may be used as a known sample (Jessen, 2008, p. 673). When these two conditions are met and both disputed and known samples are available, voice comparison can be conducted. When performing voice comparison, a wide variety of speech features are compared, and multiple methods are used (Jessen, 2008, p. 673). The main ones will be briefly described below.

2.2.1 Auditory and acoustic analysis

According to Rose, when speech samples are compared forensically, the phonetic parameters that are used are typically categorized based on two main distinctions: whether they are acoustic or auditory, or whether they are linguistic or non-linguistic (2002, p. 47). For the sake of relevance, however, only the distinction between acoustic and auditory analysis will be described in this paper.

Rose states that comparing voices with regards to their acoustic properties which are extracted by computer is probably what comes to mind first when we think of forensic voice comparison. However, he adds that another method that should not be forgotten is describing and comparing voices with respect to their auditory features – that is how the speech sounds and voices sound to an expert who is trained in recognizing and transcribing auditory features (2002, p. 47). Historically, there have been three opposing opinions about the difference between auditory and acoustic parameters and one will likely encounter all of them when voice comparison evidence is presented. The first view is that the auditory analysis is sufficient on its own. On the other hand, people who share the second opinion say that auditory analysis is not necessary at all, and it is only the acoustic parameters that are needed. Lastly, the third opinion is that auditory analysis is necessary, however, it must be combined with acoustic methods. It is the third, combined, approach that is widely accepted in today's time and it is referred to as the **phonetic-acoustic approach** (Rose, 2002, p. 48). Acoustic and auditory approaches on their own, as the first two opinions suggest, might exhibit significant shortcomings. An auditory approach alone is not adequate simply because of the human perceptual mechanism. In other words, “it is possible for two voices to sound similar even though there are significant differences in the acoustics“ (Rose, 2002, p. 48). On the other hand, an acoustic approach alone is inadequate because “without some kind of control from an auditory analysis indicating what is comparable, samples can differ acoustically to any extent“ (Rose, 2002, p. 48).

It is the auditory analysis that is conducted first. One of the reasons for this is that at the beginning it must be determined whether the quality of the sample is good enough to even proceed with the analysis. If for example, it is not possible to understand what the person on the recording said, it is not possible to compare it with any other speech sample. Rose adds that it is crucial to evaluate the understandability of recordings without any transcriptions because people generally tend to hear what they expect to hear (2002, p. 48). Furthermore, listening to samples first is done to identify the parameters that can be compared in further auditory or acoustic analysis. The aim of an auditory analysis is to provide an overview of the similarities and differences between the samples of the sound system that is used and the way it is realized (Rose, 2002, p. 48). An example of an auditory analysis statement could be that the *s* sound in one sample is realized with a noticeable lisp (speech impairment where one misarticulates sibilants), whereas the lisp is not present in the second sample.

Once the parameters that are to be compared are selected in the auditory analysis, the experts can proceed with the acoustic analysis. According to Rose, “It is in the acoustic analysis, perhaps, where the border between linguistic and individual information is most clearly crossed.“ (2002, p. 51).” This is because the shape and size of a vocal tract determine the acoustic output. Therefore, when individuals with differently shaped vocal tracts pronounce the same linguistic sounds, the output will differ acoustically. On the other hand, individuals whose anatomy of the vocal tract does not differ significantly tend to differ less in overall acoustics (Rose, 2002, p. 51). A typical example of the use of acoustic analysis in voice comparison is the measurement of the average fundamental frequency (f0), which is the acoustic correlate of the vibration frequency of the vocal folds in voice production. Voices that differ in average *f0* are typically described on a scale ranging from low-pitched to high-pitched. However, such perceptual description (which would be part of the auditory approach) would not be as accurate and objective as a description based on *f0* values (Jessen, 2008, p. 691). Measuring the average *f0* is typically mentioned in sources about speaker comparison because according to Jessen, it is one of the few areas in which we have population statistics available (2008, p. 691). However, there are many areas in which population statistics are not yet fully available. One such area, as already mentioned in the introduction, is the spectral properties of noise segments of Czech voiceless fricatives.

2.2.2 Similarity and typicality

Jessen mentions that in voice comparison there are two important aspects. The first one is the **similarity** aspect. It is, as the name suggests, "how similar or how different the two voices are with respect to the phonetic dimensions on which they are compared" (2008, p. 682). Jessen further explains that the more similar the voices are the more probable it is (everything else being equal) that they originate from the same speaker (2008, p. 682).

However, it is a common misconception in all branches of forensic science that whenever a strong degree of similarity is observed it automatically means that the two compared samples originate from the same source. Jessen quotes Rose who describes that such misconceptions, especially among laypeople, stem from various criminal genre TV shows where forensic detection is shown (2008, p. 682). In such shows, the supposed forensic experts often proclaim that given samples either are or are not a match. If it were to be proven that an absolute categorical match was made, there would have to be evidence that the matching pattern does not occur elsewhere in different sets of samples. So rather than making such absolute and categorical decisions, the current forensic sciences advocate for expressing conclusions with probabilities (Jessen, 2008, p. 682). Perhaps surprisingly, this principle applies to types of analyses such as the one of fingerprints. There has been an assumption for a long time that the conclusion of the fingerprint analysis ends in a categorical decision. However, this assumption proved to be problematic and today it is encouraged to express the conclusion in a probabilistic way as well (Jessen, 2008, p. 682).

The second important aspect of voice comparison is **typicality**. When phonetic characteristics are analysed in the disputed and known samples, it can be discovered the said characteristics are either relatively typical in the entire relevant population of speakers or on the other hand they are relatively rare. Jessen states that "Evidence for the identity of two speakers is stronger – everything else being equal – when typicality is low than when it is high" (2008, p. 282). However, to assess whether certain values are typical or rare, population statistics about the given characteristics that are used in a voice comparison must be available. To illustrate this better, I will use a specific illustrative example given by Rose (2002, p. 314).

In this example, Rose mentions that assuming the acoustic parameter compared in the disputed and known sample is long-term F0 (LTF0), we must know the average LTF0 for the given relevant population and the standard deviation (2002, p. 314). Such data could be obtained by measuring a large number of relevant speakers (in Rose's example it would be male speakers with a Broad Australian accent) and calculating the mean and standard deviation of those means. Such a large set of measurements is called the **reference sample** because it is with reference to them that the differences between the disputed and known samples have to be evaluated (Rose, 2002, p. 314). Rose stated the average LTF0 is 120 Hz, the values are normally distributed, and the standard deviation is 20 Hz. Now let us assume that the values of the disputed and the known samples are 80 Hz and 85 Hz. These values are clearly located at one extreme of the distribution. In other words, the males in the disputed and known samples have a much deeper voice than the average (120 Hz) for the given population. Rose explains that "In probability terms, it would now be highly unlikely for two values of this size to be drawn at random from the population. Now the evidence against the suspect is much stronger." (2002, p. 315). This example confirms what was mentioned above already - the evidence for the identity of the two speakers is stronger when typicality is low. Finally, Rose concludes that "the magnitude of the difference between questioned and suspect samples is not enough: the typicality of the two samples measured against a reference sample needs also to be taken into account." (2002, p. 315). Lastly, it must be noted that the reference sample is not invariant and naturally it changes depending on the specific case (Rose, 2002, p. 318).

When conducting any form of voice comparison activities, including those performed for commercial or purely research purposes, similarity, typicality, and population statistics are, as Jessen states, "necessary ingredients" (2008, p. 683). If all the ingredients are available and generally if the conditions are ideal, it is assumed that speakers can be identified by their voice reasonably easily, according to Rose (2002, p. 24). In such cases, the experts are dealing with a so-called **between-speaker variation** (on inter-speaker variation) which presupposes that different speakers indeed do possess different voices. The extent to which this can be assumed forensically, as Rose mentions (2002, p. 24), is a separate and more complicated issue, which is interesting to mention, however, it extends beyond the scope of this study, therefore it will not be discussed in detail.

It is crucial to understand that it is not only different speakers who have different voices. There will always be variation in the voice of the same speaker as well. In other words, it is impossible for one person to say the same thing in the same exact way. Such variation in one speaker's voice is called **within-speaker variation** (or intra-speaker variation) (Rose, 2002, p. 24). Within-speaker variation has many forms. Some of them are simply a necessary part of a given linguistic system. A clear illustration of such a case is the pitch of the voice and its acoustic correlate fundamental frequency (*f0*), which does not stay constant but varies depending on the intonation or tonal patterns of the given language (Jessen, 2008, p. 684). Other sources of within-speaker variation include affective states such as stress or anger, reading in contrast with speaking spontaneously, being under the influence of intoxicating substances, or being healthy versus having an illness affecting the larynx, etc. Interestingly, in some cases, even the speaker's dialect does remain invariant. Rose mentions an example of speakers of the Chinese dialect of Pudong, which is spoken near Shanghai, who tend to shift more towards the Shanghai standard dialect by changing the word-initial *h* to *f* when a more formal style is required. Similar phenomenon happens in American and British English, which both have *r* as a sociolinguistic variable. With rising formality, the percentage of rhotic *r* forms in American English increases, while the opposite happens in British English, where the rhotic *r* decreases (2002, p. 61). The most extreme case of within-speaker variation is voice disguise which is typically done with the intention to conceal one's identity (Jessen, 2008, p. 684). Taking everything that was mentioned in the paragraph above into account it is possible to generally state that an individual's voice varies in response to different situations. Rose explains that "depending on the circumstances, this can lead to samples from two different speakers having similar values for certain dimensions; or samples from the same speaker having different values for certain dimensions." (Rose, 2002, p. 34)

Jessen states that frequently there is a mismatch between the speech sample of the anonymous speaker and the caught suspect with respect to the linguistic and behavioural influences mentioned above (2008, p. 684). To give an example, in the disputed sample the anonymous individual may speak loudly in an emotionally agitated manner whereas, in the known sample, the speaker may be generally calmer and speak in a quieter manner. There are ways to address this problem such as "selecting portions in the recordings where the mismatching influence is minimal or by applying knowledge about the phonetic influences of these factors in an effort to compensate (technically or conceptually) for the mismatching influence" (Jessen, 2008, p. 684). Another problem, which Jessen mentions is that high within-speaker variety can result in a low similarity between the two samples and can lead to false rejection despite the fact the same speaker is involved. This behaviour-based variation is, of course, adding difficulties to the process of forensic speaker comparison, and it is a problem other forensic fields such as DNA analysis or fingerprint analysis do not have to contend with (2008, p. 683).

In conclusion, the most important consequence of the within-speaker variation is the fact that there will always be differences between speech samples, despite the fact they come from the same speaker. In order for speaker comparison to work, these differences have to be measured, quantified, and evaluated correctly (Rose, 2002, p. 24). Rose adds that forensic speaker comparison "involves being able to tell whether the inevitable differences between samples are more likely to be within-speaker differences or between-speaker differences" (2002, p. 24). Logically, in order for forensic speaker comparison (or in fact any other identification system) to be feasible, generally the variation between speakers must be bigger than the variation within a speaker. According to Rose, this indeed appears to be the case, given that the experts analyze the right things (2002, p. 24). Finally, Rose concludes that "it is intuitively obvious that the greater the ratio of between-speaker to within-speaker variation, the easier the identification" (2002, p. 24).

2.2.3 Expressing the outcome and Likelihood Ratio

When the analysis of the voices is finished, the experts are asked to provide a conclusion and an answer to the question of whether the compared samples may have originated from the same speaker or two different speakers. It is also worth mentioning that the conventions and conceptual frameworks of how the conclusions are expressed can differ across countries (Jessen, 2008, p. 673).

It must be noted that the conclusion of the voice comparison given by the experts "should be an informed opinion to aid the legal process" (Rose, 2002, p. 67). The conclusion may help the court decide whether the suspect is guilty or not. In other cases, it can help the police or other given authorities decide whether prosecution is sensible based on the strength or the lack of strength of the voice evidence. The defense can also use the conclusion of the experts to question the strength of the voice evidence against their client (Rose, 2002, p. 67).

Rose mentions that expressing such an informed opinion might seem straightforward. However, one must be careful not to make an inference about the identity of the speaker.

Another action that should be avoided is trying to make conclusions about the guilt of the suspect. Rose (2002, p. 68) provides a quote by Aitken (1995) which summarizes these two points:

> It is very tempting when assessing evidence to try to determine a value for the probability of guilt of a suspect, or the value for the odds in favour of guilt and perhaps even reach a decision regarding the suspect’s guilt. However, this is the role of the jury and/or judge. It is not the role of the forensic scientist or statistical expert witness to give an opinion on this. It is permissible for the scientist to say that the evidence is 1000 times more likely, say, if the suspect is guilty than if he is innocent. (p. 4)

To summarize, the opinions provided by the forensic experts are only about how likely it is for the evidence (the similarities and differences between the samples) to occur if both samples were from the same speaker, compared to the likelihood if they were from different speakers (Kavanagh, 2012, p. 33). In other and more general words “Instead of trying to state the probability of the hypothesis given the evidence, which is the job of the court, forensic experts must attempt to quantify the probability of the evidence given the two hypotheses (same speaker or two different speakers)” (Rose, 2002, p. 69).

There is a statistical approach that allows the experts to quantify the strength of the evidence given that similarity and typicality are known. It is known as the **Bayesian approach** and a core concept in it is the **Likelihood Ratio** (LR) (Jessen, 2008, p. 682**.** It can be expressed by a formula shown below in Figure 1 (Rose, 2002, p. 70).

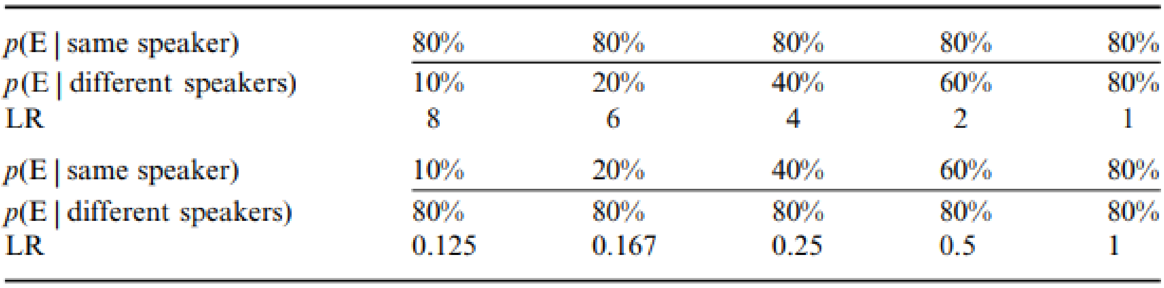
$$\mathrm{LR} = \frac{p(\mathrm{E} \mid \mathrm{H_p})}{p(\mathrm{E} \mid \mathrm{H_d})}$$

Figure 1: Formula of the Likelihood Ratio

Jessen explains that “the expression in the numerator refers to the probability (*p*) of obtaining the given speech evidence E if the prosecution hypothesis ($\mathrm{H_p}$) is correct (the two probes having the same origin)“ (2008, p. 683). On the other hand, “the denominator expresses the probability of obtaining the same evidence if the defense hypothesis ($\mathrm{H_d}$) is correct (the two probes having a different origin)“ (Jessen, 2008, p. 683).

If the value of the LR is larger than one, it means that there is relatively more evidence that the disputed and the known sample originate from the same source. On the other hand, if the LR value is less than one, there is relatively more evidence that the disputed and the known sample have a different source (Jessen, 2008, p. 683).

Rose further explains that "the numerator represents the degree of similarity between the questioned and suspect samples, and the denominator represents how typical they are: the probability that you would find measurements like the questioned and suspect samples by chance in the relevant population (2002, p. 70). If similarity is high, the likelihood that the two recordings have the same source is relatively high, and if similarity is low, then the likelihood that the disputed and the known sample have the same source is also relatively low (Jessen, 2008, p. 683). If typicality is high, then the likelihood that a different speaker is the source of the disputed sample is also relatively high. On the other hand, if typicality is low then the likelihood that some other than the suspected speaker is involved is relatively low (Jessen, 2008, p. 683). In Figure 2 we can see a simple example of how the likelihood ratio works (Rose, 2002, p. 71).



p(E | same speaker)	80%	80%	80%	80%	80%
p(E | different speakers)	10%	20%	40%	60%	80%
LR	8	6	4	2	1
p(E | same speaker)	10%	20%	40%	60%	80%
p(E | different speakers)	80%	80%	80%	80%	80%
LR	0.125	0.167	0.25	0.5	1

Figure 2: Illustration of the likelihood ratio

In this particular case, we can notice that in the top half of the table, the similarity between the disputed and the known samples (the numerator) is always 80% which indicates a high similarity. The probability of observing the difference between the known and the disputed samples by chance in the relevant population (the denominator) is increasing from 10% up to 80%. When similarity between the two samples is high (80%), and typicality is low, (10%), we can see that the Likelihood Ratio is large (80/10 = 8). As it was already mentioned before in this chapter, Likelihood Ratio values larger than one indicate that there is relatively more evidence that the two samples originate from the same speaker. As typicality increases from 10% to 80%, the Likelihood Ratio value decreases. Rose explains that if the Likelihood Ratio value is 1, the evidence is useless because such value indicates that "you are just as likely to observe the difference between the questioned (disputed) and suspect (known) samples if they come from the same speaker as if they are chosen at random from two different speakers in the population“ (2002, p. 71).

The bottom half of the table shows that this time it is the denominator which is held at 80% constantly. It means that the probability of finding the difference between the disputed and known samples by chance is high. The values of the numerator are increasing from 10% to 80%, indicating the increase of similarity. As the numerator increases so does the value of the Likelihood Ratio. The lowest LR value is 0.125 and it is supporting the hypothesis that the two samples come from two different speakers (Rose, 2002, p. 71). Rose states that interpretations of LR values less than 1 are best done in terms of their reciprocal. In his example he explains that "given a LR of 0.125, the odds in favour of the defence are 8 to 1, since the reciprocal of 0.125 is (1/0.125 =) 8“ (2012, p. 71). In other words, it would be eight times more likely to observe the difference between the known and the disputed samples if they came from different speakers (Rose, 2002, p. 71). The highest LR value is 1, which indicates that the evidence is of no use as already mentioned in the paragraph above.

Even though Likelihood Ratio is not included in the practical part of this study, a section was dedicated to it here to highlight the importance of the aspect of typicality. This inclusion underscores how crucial typicality is in forensic phonetics, as it is a fundamental component of statistical methods such as LR.

Although it is never possible to identify a speaker with absolute certainty because of the inherent plasticity of the human voice, it is possible to significantly limit the number of possible candidates and express this based on Likeluhood Ratio (Christensen, 2023, p. 63). To conclude, in order to calculate the Likelihood Ratio and express the outcome of a forensic voice comparison, both similarity and typicality must be known. As previously mentioned, determining whether a certain value is relatively typical or relatively rare is only possible when population statistics are known and available. This paper aims to provide such population statistics and to facilitate future forensic work in the Czech environment.

2.3 Common difficulties in voice comparison

The following chapter will discuss relevant issues which forensic phoneticians have to face when performing voice comparison.

Chapter 1.2.2 already briefly discussed difficulties caused by inherent withing-speaker variation in speech, which forensic experts must face. However, since the chapter was mostly focused on similarity and typicality, for the sake of relevance there was not much space given to the difficulties that within-speaker variation causes in voice comparison, therefore, it will be discussed in more detail here. Rose mentions two hypothetical scenarios that showcase the difficulty connected with within-speaker variation and the lack of control the experts have over the situations where forensic materials are captured. The first scenario is of an offender during an armed hold-up. The perpetrator is trying to be deliberately intimidating and therefore their voice pitch is high with a narrow range. Then, a suspect is interviewed at the police station in the early morning hours. The suspect feels tired, intimidated, cowed and reticent. The suspect is mostly silent but when they already speak their pitch is low. The pitch difference between the unknown sample from the armed hold-up and the known sample from the interrogation room will almost certainly be bigger than the difference between some other two speakers. However, just because the difference between the two samples is big and therefore the similarity is low, the suspect cannot be excluded from the investigation, since the circumstances involved were so different (2002, p. 34). The second scenario presents an offender whose voice is normally high-pitched. He is perpetrating a telephone fraud. The suspect who is caught normally has a low-pitched voice. However, when their telephone was monitored and recorded by the investigator they were talking to their daughter in an animated, congenial, happy, and very high-pitched voice which is very similar to that of the offender. In this case, although the pitch difference between the two samples is lower than that which would sometimes be found within the same speaker, it cannot automatically be seen as incriminatory because once again, the circumstances involved were very different (Rose, 2002, p. 34). It is apparent then that within-speaker variety and generally different circumstances of capturing the samples create difficulties when trying to arrive at a conclusion. Rose emphasizes that in order to evaluate the differences between samples that arise from inevitable variation, the experts must know the nature of situational effects of the voice dimension under question. In other words, it is necessary to know how voices differ in response to different circumstances (Rose, 2002, p. 34).

Forensic practitioners are, however, confronted not only with speaker-induced variability but also with the discrepancies between the good-quality speech material (the known sample) and often the low-quality, authentic forensic material (the disputed sample) (Fecher, 2011, p. 71). However, it must be noted that the quality of the known sample is not always that satisfactory either. Whether the quality of the sample is great, average, or even poor, of course, depends on the equipment that is available to the investigators. Jessen, for example, mentions that forensic phoneticians can be limited by poor transmission or recording equipment (2008, p. 685) while Fecher mentions that known samples can be of studio quality. Overall, it can be assumed that the quality of the recording equipment will depend on the finances available to the investigators. This likely differs from country to country. For example, in countries such as the Netherlands, Germany, Sweden, Austria, Spain, or Switzerland, the governments employ full-time forensic phoneticians whereas in other countries such as Australia or the United Kingdom, a lot of forensic work is done by academic phoneticians, who of course have other responsibilities. It is therefore likely that in countries where the governments place more importance on the work of forensic phoneticians their environment and equipment will be of higher quality. Unfortunately, no relevant sources that could describe the quality of the equipment of the Czech investigators were found. However, the supervisor of this thesis, who has experience in the field of forensic phonetics, reported that the quality of the known samples is often poor and taken by what appears to be a voice recorder of substandard quality (R. Skarnitzl, personal communication, August 12, 2024). Furthermore, the quality of the forensic material can be affected by practical factors such as noisy background, two people speaking at the same time, echoing rooms, etc (Rose, 2002, p. 35). Such factors can cause a part of the recording that could potentially provide useful information, to become useless.

Another type of technical issue, mentioned by Jessen, is a reduced quantity of the recording, in other words, the forensic sample is not long enough. It is a common misconception of laymen that it is possible to identify a voice or conduct a voice comparison from just a very short utterance, perhaps even a single vowel. Of course, phoneticians and people with a background in linguistics know that such an assumption is incorrect (2008, p. 686). This is because, as Jessen says, “the shorter the speech recording, the less representative the speech patterns in the recording are in relation to the whole range of the speech patterns of the speaker.” (2008, p. 686). If the recording is too short, there is a high possibility that the types of sounds rich in speaker-specific characteristics will not be present. The question of how short is too short for a forensic recording naturally arises. Jessen claims that there is not a necessarily fixed limit below which a voice comparison cannot be conducted. However, he adds that around eight seconds of speech from the anonymous speaker and at least around double that time for the suspect is recommended. He concludes the point by stating that this does not exclude the possibility of even a two-second recording containing very specific and distinct information about the speaker (2008, p. 686).

The last issue, and perhaps the most important one in the context of forensic phonetics, is the distortion of the recordings by telephone transmission. Christensen explains that forensic phonetics is “one of the fields most directly impacted by digital transmission of speech” (2023, p. 62). This is because it is common for the disputed sample to originate from telephone conversations. As previously stated in chapters 1.1 and 1.2, forensic phoneticians commonly deal with crimes such as drug dealing arrangements over the phone, stalking, hoax calls to emergency services, obscene calls, fraudulent deals negotiated over the telephone, or ransom demands. All the crimes mentioned involve speaking on the telephone, therefore the process of voice comparison in such cases will inevitably be affected by the telephone transmission. Detailed information about how telephone transmission affects fricatives, which is relevant to this paper, will be provided in section 2.5 later in the paper.

2.3 Fricatives

This paper focuses on spectral properties of noise segments of Czech voiceless obstruents, specifically four voiceless fricatives in a forensic context. Before it is possible to explain what role fricatives play in forensic phonetics, specifically in voice comparison, it is of course necessary to provide a theoretical background for them. This section will therefore include a definition of what fricatives are, how they differ in voicing, place of articulation, and most importantly how they differ in their acoustic parameters.

Fricatives are a category of obstruents that are produced by creating a narrowing in the vocal tract, specifically in or above the larynx. In order to create such narrowing, two articulatory organs must come together close enough so that the air passing through is turbulent, creating friction (Skarnitzl et al., 2016, p. 56).

2.3.1 Articulatory description of fricatives

This part of the study will describe fricatives based on their voicing and their place of articulation.

Fricatives are divided into two subcategories based on whether their production is accompanied by the vibration of the vocal folds. If the vibration of the vocal cords is present during the production of fricatives, they are defined as **voiced fricatives.** Skarnitzl et al. describe the perception of voiced fricatives as a “buzzing noise” (2016, p. 56). On the other hand, if the fricatives consist solely of turbulent airflow and the vocal cords are not vibrating during their production, such fricatives are defined as **voiceless** (Strevens, 1960, p. 32). The perception of voiceless fricatives is described as a “hissing noise” (Skarnitzl et al., 2016, p. 56).

As we already know constriction is crucial for the production of fricatives. The exact location of such constriction within the vocal tract is called the place of articulation. For fricatives, we distinguish nine places of articulation in total. They are **bilabial, labiodental, dental, alveolar, post-alveolar, palatal, velar, uvular,** and **laryngeal** (**glottal**). However, since this paper is focused on Czech fricatives (obstruents) the only places of articulation which will be described in more detail will be those that are relevant to the Czech language.

2.3.1.1 Place of articulation of Czech fricatives

The Czech language contains labio-dental, alveolar, post-alveolar, velar, and laryngeal (glottal) fricatives.

Labio-dental fricatives are produced by creating a constriction between the lower lip and the upper teeth, incisors to be precise (Skarnitzl et al., 2016, p. 60). The Czech inventory has a voiceless labio-dental fricative /f/ and a voiced labio-dental fricative /v/. Both have the status of a phoneme.

Alveolar fricatives are produced by raising the tip or the blade of the tongue toward the alveolar ridge (the bony ridge in the area where upper teeth are anchored and which extends inwards into the oral cavity) (Skarnitzl et al., 2016, p. 24) and creating a narrowing there (Strevens, 1960, p. 34). Czech contains both a voiceless alveolar fricative /s/ and a voiced alveolar fricative /z/. Once again, they both have the status of a phoneme.

Post-alveolar fricatives are produced similarly to alveolar fricatives as the name itself suggests. The tip or the blade of the tongue is also raised but it creates the narrowing at the end of the alveolar ridge where it merges into the hard palate (Strevens, 1960, p. 34). The Czech inventory contains both a voiceless post-alveolar fricative /ʃ/ and a voiced post-alveolar fricative /ʒ/. Just as in the previous cases, they are both phonemes.

The Czech language also contains a (post)-alveolar fricative trill /r̝/. As its name suggest, it cannot be strictly defined as a “pure” fricative because it is a segment with a combined manner of articulation. It is rather rare in world languages; however, it is as Skarnitzl et al. say “notoriously known in Czech” (Skarnitzl et al., 2016, p. 59). It is also known for being difficult to pronounce not only for foreigners but also for some native speakers of Czech (Isačenko, 2013, p. 1). For the purpose of this study, it will be included with the other fricatives. When producing the (post)-alveolar trill the tip of the tongue is raised towards the alveolar ridge and the airstream causes it to vibrate while the sides of the tongue touch the upper molars. Simultaneously, there is a turbulent noise that persist throughout the entire articulation (Skarnitzl et al., 2016, p. 59). The Czech inventory contains both a voiced (post)-alveolar fricative trill /r̝/ which is a phoneme, and a voiceless (post)-alveolar fricative trill /r̝̊/ which is only an allophone of /r̝/. The voiceless variant appears because of a process called the assimilation of voicing. In the word [**dr̝**ɪ] (meaning "toil" singular, second person, imperative) the trill is influenced by the neighbouring voiced segment, and therefore it is voiced. However, in the word [**tr̝̊**ɪ] (meaning "three") the trill loses its voicing because of its neighbouring voiceless segment. The phoneme /r̝/ is unique because it undergoes both regressive and progressive voicing assimilation, unlike other Czech phonemes which only undergo regressive assimilation.

Velar fricatives are articulated by raising the back of the tongue towards the soft palate (velum) and creating a narrowing there (Strevens, 1960, p. 34). The voiceless velar fricative /x/ is a phoneme in the Czech inventory unlike the voiced velar fricative /ɣ/ which is only an allophone of /x/ and once again occurs because of a process called the assimilation of voicing. In such a process /ɣ/ is influenced by a neighbouring voiced segment and gains voicing itself. An example of this process can be seen when we compare two Czech sentences [ja: bɪ**x** **t**am jɛl] (meaning "I would go there") and [ja: bɪ**ɣ** **d**o praɦɪ nɛjɛl] (meaning "I would not go to Prague"). We can see that in the first sentence, the velar fricative is realized as voiceless when it is followed by a voiceless segment, in this specific case a voiceless alveolar plosive /t/. However, the velar fricative is realized as voiced when the following segment is voiced as well, which is the case of a voiced alveolar plosive /d/ in the second sentence.

Next, there are laryngeal (also known as glottal) fricatives. According to Strevens, their exact place of articulation is a controversial subject (1960, p. 34). He states that some researchers believe the constriction is created "somewhere in the larynx" while "others believe cavity-friction to be generated throughout the vocal tract" (Strevens, 1960, p. 34). According to Laufer, some researchers do not believe laryngeal fricatives to be fricatives at all. However, he argues that the term "laryngeal fricatives" is correct and that the constriction happens in the glottis, as there is no other apparent constriction in the oral cavity tract (1991, p. 91). The Czech inventory only contains the voiced laryngeal fricative /ɦ/ and it has the status of a phoneme.

According to Skarnitzl et al., the voiced variant is relatively rare among the world languages compared to the voiceless one (2016, p. 62).

2.3.2 Acoustic description of fricatives

In the previous sections, it was already determined how to differentiate between fricatives based on their voicing and their place of articulation. Lastly, it must be described how fricatives differ acoustically. However, before providing details about how the acoustic characteristics of fricatives are described, it will be beneficial to first mention the theory behind how the shape of the vocal tract and the place of articulation cause spectral differences among the fricatives.

Strevens described the vocal tract as a highly mobile tube having a cylindrical cross-section (1960, p.33). This tube functions as a resonator for speech sounds. Naturally, the shape of the vocal tract changes when articulatory organs are moved so that different fricatives (or any other segments) can be produced. When the overall shape and length of the vocal tract change, so do the frequencies at which the tract naturally prefers to resonate. In other words, the resonant frequencies of each fricative are determined by the configuration of the vocal tract (Shadle & Mair, 1996, p.1).

Skarnitzl et al. specify that the length of the resonating tube between the constriction and the opening of the lips is significant (2016, p. 56). Generally, if the tube is shorter and the overall shape of the oral cavity is smaller, the resonant frequencies are higher (Johnson, 2003, p. 125). This is the case when for example an alveolar fricative /s/ is pronounced because the space between the alveolar ridge and the lips is relatively short (Skarnitzl et al., 2016, p. 56). On the other hand, when the constriction is located further back in the oral cavity and the space between it and the opening of the lips is longer, the resonant frequencies are lower (Johnson, 2003, p. 12.). An example of that would be a velar fricative /x/ (Skarnitzl et al., 2016, p. 56).

Having discussed how “the complex upper vocal tract geometry plays a significant role in shaping the speech spectrum” (Mohapatra et al., 2022, p. 764)”, the next step in this study is to introduce different ways of describing the spectral characteristic of fricatives. In the following sections, three such approaches will be explained: the analysis of the main resonant frequency, the computation of spectral moments, and the application of the Discrete Cosine Transform (DCT). Each of these methods provides unique insights into the acoustic properties of fricatives.

2.3.2.1 Main resonant frequency

It was previously stated that the main resonant frequency of each fricative is determined by the shape of the vocal tract. This frequency can then be observed in a spectrogram image and based on its value; fricatives can be differentiated from each other. Therefore, before the description of the main resonant frequencies of Czech voiceless fricatives is provided, it is necessary to briefly describe how to read a spectrogram image, since such images will be used below. A spectrogram is a visual representation of sound that allows us to observe three parameters - frequency, amplitude, and time all in one graph. In the spectrograms presented here, frequency can be observed on the y-axis, time on the x-axis, and lastly, amplitude can be observed in the shades of grey colour. The darker the colour is the higher the amplitude.

The first segment which will be discussed is once again a voiceless labio-dental fricative /f/. It is characterized by having the constriction at the very front of the oral cavity. In fact, the constriction is so far in the front that there is hardly any vocal tract in front of it to filter the sound. This is why the spectrum is flat and there is no apparently visible resonant frequency (which would appear as a horizontal band of dark grey colour) in the frequency range typically displayed in a spectrogram. It is possible for the main resonant frequency to be at approximately 10kHz. Below in Figure 3, we can see a visual representation of /f/ in the spectrogram.

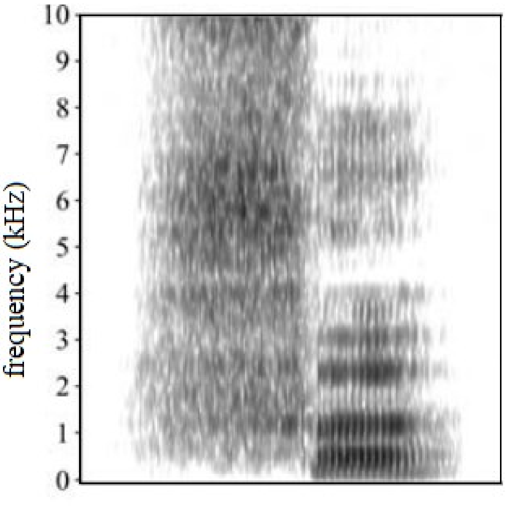


Figure 3: Spectrogram of /f/ followed by a vowel

The next segment to be described is one of the target fricatives of this research, a voiceless alveolar fricative /s/. The space between the constriction at the alveolar ridge and the opening of the lips is relatively short and it is said to be approximately 2 centimeters (Konsonanty, n.d.). This rather short space causes the resonant frequency to be high at approximately 4.5 kHz. Unlike in the case of a labio-dental fricative /f/, the resonant frequency of /s/ can be clearly identifiable in the spectrogram, which can be seen in Figure 4. It must be noted here that coarticulation, namely labialization, naturally affects the values of the resonant frequency. For example, /s/ would have a lower resonance frequency if it was followed by a rounded vowel /u/ compared to if it was followed by a non-rounded vowel /a/ (Smorenburg & Heeren, 2020, p. 950). This is because the anticipatory lip-rounding which is associated with lip protrusion lengthens the anterior cavity. This means that the acoustics of fricatives vary systematically as a function of phonetic context (Smorenburg & Heeren, 2020, p. 949). Therefore, it can be summarized that coarticulation processes such as labialization will naturally cause within-speaker variability in the acoustic parameters of fricatives (and other segments as well of course).

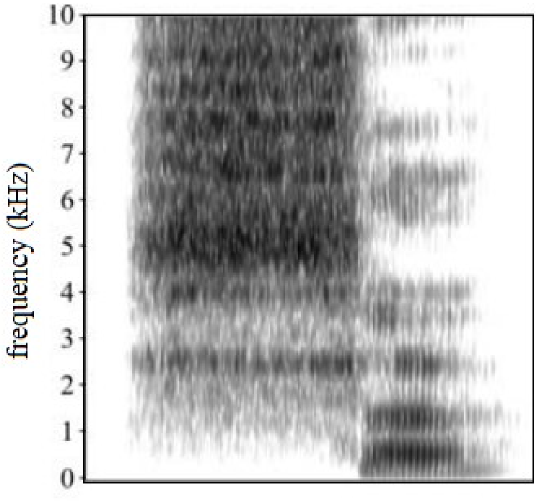


Figure 4: Spectrogram of /s/ followed by a vowel

Voiceless post-alveolar fricative /ʃ/ which is another of our target research segments, has the constriction further in the back compared to /s/, therefore the space between the constriction and the lips is longer. Furthermore, most people engage in labialization when pronouncing /ʃ/. That is why the resonant frequency of /ʃ/ is lower compared to /s/ at approximately 2.5-3.5 kHz. It must also be mentioned that the resonant frequency of /ʃ/ is caused by two resonances at two different places - one being at the palate and the second one being at the cavity under the tongue (Fonetický ústav, 2016). A spectrogram of /ʃ/ can be seen in Figure 5.

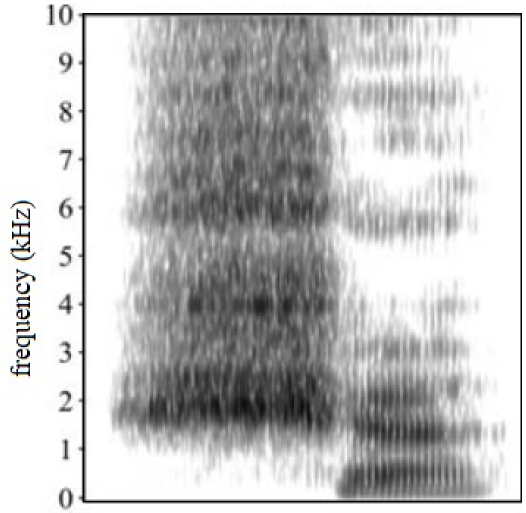


Figure 5: Spectrogram of /ʃ/ followed by a vowel

Voiceless (post)-alveolar fricative trill is acoustically the most similar to post-alveolar fricative /ʃ/. The resonant frequency of /r̝̊/ is approximately 2.5-3.5 kHz. A spectrogram of /r̝̊/ can be seen in Figure 6.

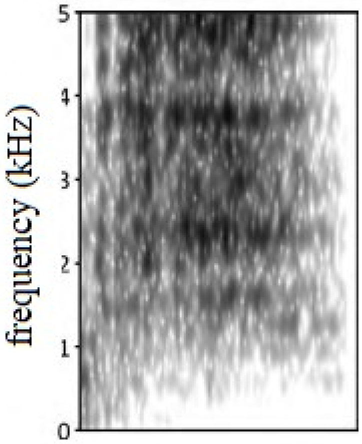


Figure 6: Spectrogram of /r̝̊/

Voiceless velar fricative /x/, which is the last target segment, has the obstruction much further in the back of the oral cavity compared to previously mentioned fricatives. The relatively large space between the velum and the opening of the lips causes the resonant frequency to be low at approximately 1-1.5 kHz. Once again, however, these values will vary depending on what segment follows the voiceless velar fricative. It can be expected that when /u/ follows, the resonant frequency will be lower in comparison with when /ɪ/ follows. We can observe the resonant frequency of /x/ depicted by a dark shade of grey in Figure 7.

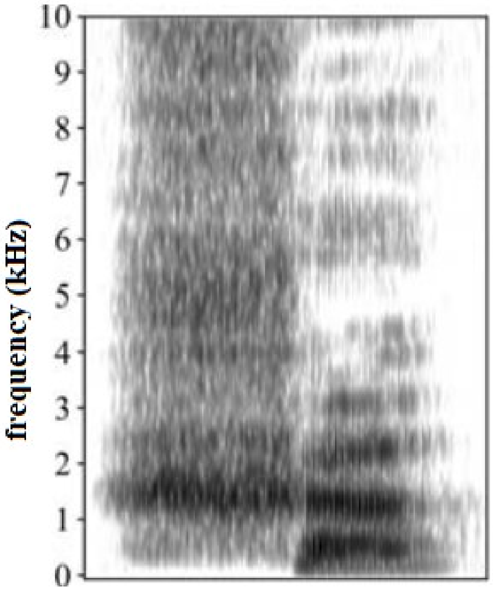


Figure 7: Spectrogram of /x/ followed by a vowel

Despite the fact that voiced fricatives were not discussed here as they are not crucial for this study, it is worth mentioning that they share many acoustic parameters with their voiceless counterparts. If we observe spectrograms of e.g. voiced and voiceless alveolar fricatives (/z/, /s/), we will notice that their resonant frequency has approximately the same value (since their place of articulation is the same), that is as we already know approximately 4.5 kHz. However, as we can see in Figure 8 the voiced fricative has a visible fundamental frequency - *f0* (circled in red), which signifies the activity of the vocal folds. On the other hand, there is no *f0* present when voiceless fricatives are pronounced since the vocal folds are inactive as it was already explained in section 1.1.

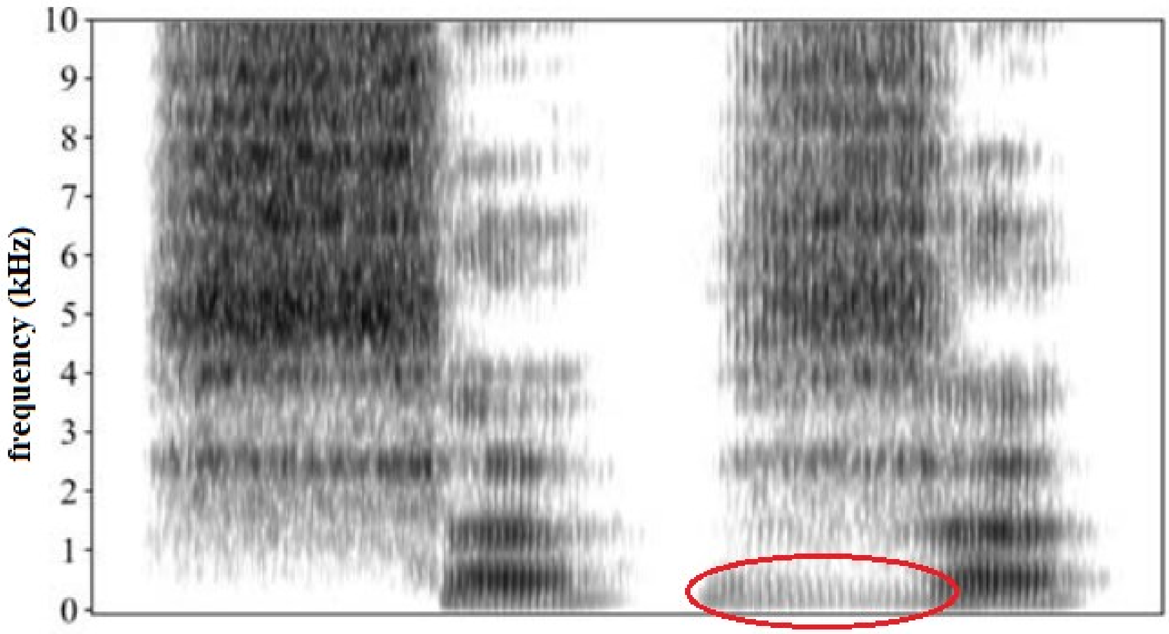


Figure 8: Spectrograms of /s/ on the left vs. of /z/ on the right followed by vowels

2.3.2.2 Spectral moments

As we already know from section 2.3.2.1, it is possible to use the resonant frequencies to describe the spectral shape of friction noises. However, the resonant frequencies by themselves do not provide a complete picture of the overall spectral shape (Nittrouer, 1995, p. 522). That is the reason why this section is dedicated to spectral moments analysis, which allows us to describe the shape in more detail by using just a few numbers (Nittrouer, 1995, p. 522). Furthermore, spectral moments will be of great importance in later chapters because their potential to carry idiosyncratic cues will be discussed.

When researchers examine spectral properties of obstruent consonants, they often do so by using different ways of parametrizing the acoustic spectrum. The most frequent approach proposed by Forrest, Weismer, Milenkovic, and Dougall (1988) is calculating the first four spectral moments (Skarnitzl & Nechanský, p. 23, 2025), which are **spectral mean**, **standard deviation**, **skewness,** and lastly **kurtosis**. This approach aims to "quantitatively describe the patterns of spectral energy within the band of noise characteristic of obstruent sounds" (Barrett, 2012, p. 1). Jongman et al. explain that "spectral moments analysis involves a statistical procedure for classifying obstruents, capturing both local (mean frequency) and global (spectral tilt and peakedness) aspects of speech sounds (2000, p. 1253). Nittrouer explains that spectral moments describe the distribution of frequencies in a spectral section (1995, p. 521). Lastly, Barett states that we can think of spectral moments as "statistical snapshots of the energy distribution within an acoustic spectrum because spectral moments analysis uses numerical values to describe the spectral energy of a speech sound within a static window or period of time (2012, p. 1). After the general definition of the spectral moments analysis was given, it is now beneficial to provide more details about each of the four spectral moments.

The first spectral moment, the spectral mean, also known as the **center of gravity** (COG), reflects the average energy distribution in a given section of the spectrum (Jongman, 2000, p. 1253).

The second spectral moment is the standard deviation. A general definition of the standard deviation in statistics is "a measure of the spread of values around the mean" (Rose, 2002, p. 259). In other words, and more specifically to phonetics, the standard deviation reflects how much the frequencies present in the spectrum vary in relation to the mean (Barrett, 2012, p. 1).

The third spectral moment is skewness. In statistics, "skewness is an indicator of a distribution's asymmetry" (Jongman, 2000, p. 1253). In phonetics specifically, the term skewness "refers to **spectral tilt**, the overall slant of the energy distribution" (Jongman, 2000, p. 1253). In other words, skewness concerns the degree of asymmetry of the spectrum around the center of gravity. Skewness of zero indicates that the distribution of values is symmetrical around the mean. "Positive skewness suggests a negative tilt with a concentration of energy in the lower frequencies" whereas "negative skewness is associated with a positive tilt and a predominance of energy in the higher frequencies" (Jongman, 2000, p. 1253).

The fourth and the last spectral moment is kurtosis. It is an indicator of the peakedness of the spectral distribution (Jongman, 2000, p. 1253). "Positive kurtosis values indicate a relatively high peakedness ~the higher the value, the more peaked the distribution, while negative values indicate a relatively flat distribution." (Jongman, 2000, p. 1253). Therefore, positive kurtosis indicates that the spectrum is clearly defined with well-resolved peaks, whereas negative kurtosis suggests that the spectrum is flat with no clearly defined peaks (Jongman, 2000, p. 1253).

Specific values of the spectral moments of each fricative will not be provided. The reason is that at least in the Czech population the available data is not straightforward, and it is the primary goal of this thesis to present these detailed population statistics. It is, however, possible that the distribution of values for fricatives is well documented in other languages. Providing the values here would not be relevant to this paper though.

2.3.2.3 The discrete cosine transformation

The last way of describing the spectral characteristic of fricatives that will be discussed is called the **discrete cosine transformation** (DCT). The DCT is a mathematical operation similar to the discrete Fourier transform (DFT). "It decomposes a signal into a set of sinusoids such that, when these are summed, the same signal is reconstructed" (Harrington, 2010, p. 206). There are a few differences between the two operations though. The first one is that in the DCT the sinusoids are at half-cycles, k = 0, 0.5, 1, 0.5…½ (*N* – 1), unlike the sinusoids in the DFT, which are at integer cycles k= 0, 1, 2.. *N*-1). Another difference is that the output of the DCT is sinusoids with no phase. However, since any sinusoid with no phase is a cosine wave, it is possible to say that the DCT decomposes a signal into a set of cosine waves at frequencies k = 0, 0.5, 1.0, 1.5…½ (*N* – 1) (Harrington, 2010, p. 206).

Harrington explains that the amplitudes of the cosine waves are called DCT coefficients. They are usually labelled from 0 to N-1. "The 0th coefficient, k0, is the amplitude of the k = 0 cosine wave; k1, the 1st coefficient, is the amplitude of the k = 0.5 cosine wave, and so on." (Harrington, 2010, p. 207). The DCT coefficients encode global properties of the signal's shape. Specifically, *k0, k1*, and *k2* are proportional to the signal's mean, slope, and curvature respectively. Therefore, they serve a very important function, just like spectral moments. Both spectral moments and the DCT coefficients reduce the quantity of information in a spectrum to just a small number of values and importantly, "in such a way that different phonetic categories are often quite well separated (assuming these categories have differently shaped spectra)" (Harrington, 2010, p. 207).

2.4 Fricatives as idiosyncratic cues

As we already know, speakers' voices convey idiosyncratic information. However, some speech sounds convey more speaker information than others (Smorenburg & Heeren, 2020, p. 949). Some sources state that the formants in vowels and sonorant consonants are regarded to be among the most informative cues for distinguishing between speakers at the segmental level (Skarnitzl & Nechanský, p. 23, 2025). Other sources oppose this view, arguing that measuring vowel formats is a difficult task because of the existence of numerous parameters that may alter the results. Instead, they suggest that using spectral moments of nasals and fricatives may be a more suitable approach for discriminating between speakers (Schindler & Draxler, 2013, p. 2793). Even though there are various scholarly opinions on this particular issue, the research on speaker discrimination has indeed been focusing predominantly on vocalic parameters rather than consonantal parameters (Schindler & Draxler, 2013, p. 2793). Nevertheless, there are numerous studies analyzing consonants in a forensic context. Specifically, calculating the first spectral moments: centre of gravity (COG), standard deviation, skewness, and kurtosis, has been used in numerous studies examining speaker-specific aspects of obstruents (e.g., Fecher & Watt, 2011; Kavanagh, 2012; Schindler & Draxler, 2013; Lo, 2018; Smorenburg & Heeren, 2020). Other studies relied on other ways of parametrizing the acoustic spectrum, for example, the DCT coefficients (e.g. Pingjai, 2019). It is crucial to note here that it is unclear whether the resulting measures of just mentioned approaches preserve their speaker-discriminating potential in a telephone signal. In the following section, some of the studies that examined speaker-specific aspects of fricatives using either the calculation of spectral moments or DCT will be introduced as well as the effect of the telephone signal on fricatives.

2.4.1 Spectral moments and DCT as a way of examining speaker-specific aspects of fricatives

The first study that will be discussed here is by Schindler and Draxler (2013). It aimed to investigate the difference between the speaker discriminating potential of fricatives, nasals, and vowels in German. As apparent from the title of this section, the spectral moments were used as the speaker-discriminating feature for both fricatives and nasals. The speaker-discriminating feature for vowels was formants. Apart from the speaker discriminating potential itself, the study also investigated whether the context of the nasals and fricatives and the speaking style (read speech vs. spontaneous speech) had any effect on their speaker discriminating power. The study used recordings of 49 male speakers from two different speech databases containing both spontaneous and read speech. The speech was segmented and labeled, the relevant phonemes were selected, and the midpoint of their spectrum was taken to calculate the spectral moments. Then, the F-ratio which calculates the ratio of the within-speaker and between-speaker or a certain parameter, was used as a statistical measure. The results showed high F-ratios for all the nasals and fricatives, whereas only certain vowels seemed to perform this well. All consonants reached a high significance with $p < 0.001$. While most vowels had significant F-ratios, not all of them did, as was the case with the fricatives and nasals. The results showed that the context of the nasals and fricatives did not improve the speaker discriminating power of the parameters. On the contrary, it appeared that only all contexts combined guaranteed satisfactory speaker discrimination. The results also showed that speaker-discriminating characteristics seem to be slightly stronger in spontaneous than in read speech. This may be because the speaker's individual characteristics are more present when the speaker decides what they wish to say instead of reading a given text. The study concluded that the spectral moments of both nasals and fricatives provide great speaker-discriminating potential whereas the vowel formants showed less satisfactory results. Although some vowels turned out to be useful, all of them had lower F-ratios than the consonants (Schindler & Draxler, 2013, p. 2795).

Another study worth mentioning in this section is by Smorenburg and Heeren (2020). Unlike numerous studies that examined how different speech sounds are more speaker-specific than others, this one examined speaker information of the same segment in different linguistic contexts. Specifically, the authors investigated whether Dutch fricatives /s/ and /x/ taken from telephone dialogues contain different speaker information as a function of syllabic position and labial co-articulation (Smorenburg & Heeren, 2020, p. 949). The first two spectral moments and a spectral tilt were used as speaker-specific information. Just like the previous study, the 66 recordings of spontaneous speech taken from the Spoken Dutch Corpus were from male speakers only. This is due to the overrepresentation of male speakers in forensic voice comparisons (Smorenburg & Heeren, 2020, p. 953). The target segments and their adjacent context were segmented and labeled. The first two spectral moments were once again computed from the spectrum at approximately the mid-50% point of the fricative. The statistical analysis then consisted of two parts. The first part used linear mixed-effect modelling to determine whether linguistic factors affected the acoustic properties of /s/ and /x/ in spontaneous telephone speech. The second one used multinominal logistic regression "to investigate whether the amount of speaker information in /s/ and /x/ varied as a function of syllabic position and labial co-articulation" (Smorenburg & Heeren, 2020, p. 954). The results showed that the telephone bandwidth captures the effects of perseverative and anticipatory labialization for the fricative /x/, as its spectral peaks fall within the telephone band (500-3400 Hz). However, the same does not apply to the fricative /s/ because its spectral peaks fall outside the telephone band. The results also showed that /s/ contains slightly more speaker information than /x/ in telephone speech. It was also discovered that speaker information is systematically distributed across the speech signal, "even though differences in classification accuracy were small, codas and tokens with labial neighbors yielded higher scores than onsets and tokens with non-labial neighbors for both /s/ and /x/" (Smorenburg & Heeren, 2020, p. 949). The authors concluded that linguistic contexts do affect fricative acoustics and that speaker information in the same speech sound is not the same across linguistic contexts. (Smorenburg & Heeren, 2020, p. 959).

The next and rather complex study that will be discussed in this paper is by Fecher (2011). It aimed to report on the acoustic-phonetic analysis of the voiceless fricatives /s, ʃ, f, θ/ under various face disguise conditions. This was done to account for the discrepancies between studio-quality speech material and the often low-quality, authentic forensic samples that forensic experts must work with, and simulate a more realistic forensic scenario, where perpetrators frequently use face-concealing garments (FCG) to hide their identity (Fecher & Watt, 2011, p. 664). The reasons why fricatives were chosen for this study were "their high perceptual confusability, their relevance as consonantal features in forensic phonetics, and an anticipated larger attenuation by certain FCGs of energy in higher frequency bands that are particularly discriminative for this phoneme class." (Fecher, 2011, p. 73). Six native British English speakers (this time 3 males and 3 females) were recorded reading phonetically controlled stimuli in a professional studio. The speakers were asked to repeatedly read aloud a list of 64 phonotactically legal /C1ɑC2/ syllables in a carried phrase He said <stimulus>. Each time the speakers would wear one of the face-concealing garments, which were: a flexible microporous surgical tape, surgical mask, motorcycle helmet containing cheek pads, hoodie and a bandana covering the mouth, knitted balaclava, lightweight polyester niqab (face covering garment for religious purposes), and finally a rubber mask covering the entire face (Fecher & Watt, 2011, p. 664). For comparison purposes, the speakers were also recorded wearing no FCG. The results of this study were complex and included more parameters than only spectral moments. The interpretation would be quite long, therefore for the sake of relevance, the exact results will not be included here, however, the conclusion will. As expected, there were shifts in the spectral properties of fricatives. The author concluded that the shifts might result from the acoustic damping effects of certain mask materials, which leads to energy being absorbed at higher frequencies (Fecher, 2011, p. 73). Articulatory behaviour and speech production are inherently affected because of the presence of the FCGs. Each of the speakers may adopt a way of compensating for the interference of the FCG, which may of course affect the acoustic properties of the fricatives as well.

The last study that will be mentioned in this paper is by Jannedy and Weirich (2017). The authors aimed to quantify the acoustic differences between German voiceless palatal fricative /ç/ and voiceless postalveolar fricative /ʃ/. For this purpose, both spectral moments calculations and the DCT were used. Data were collected from two speaker groups in Berlin and Kiel, where the contrast between the two fricatives is still being realized. The third group of speakers whose data were collected also came from Berlin, but they spoke what is known as Hood German, a multiethnolect used by youth in multilingual and multicultural neighborhoods. In this variety of German, the contrast between the two fricatives /ç/ and /ʃ/ weakened or got lost completely (Jannedy & Weirich, 2017, p. 395). The selected speakers were asked to read sets of minimal pairs embedded in carrier phrases. The minimal pairs were *fischte /fɪʃtə/* - fished, thirds person singular vs. *Fichte /fɪçtə/* - spruce, *misch /mɪʃ/* - mix vs. *mich /mɪç/* - myself, and *wischt /vɪʃt/* - wipe, thirds person singular vs. *Wicht /vɪçt/* - gnome. The full sentences then were for example *Ich habe fischte gesag (I said fished)* vs. *Ich habe Fichte gesagt (I said spruce)*. In total, 144 items were generated and they were played twice in a perception test given to listeners from Buxtehude, a city whose dialect has retained the difference between /ç/ and /ʃ/ (Jannedy & Weirich, 2017, p. 397). The results of the perception test showed that the listeners do differentiate between the two fricatives in minimal pairs produced by the speakers from Berlin and Kiel, but they do not differentiate between them in minimal pairs produced by speakers of Hood German. The acoustic analysis revealed that the fricatives are spectrally very similar in all varieties. However, the spectral moments in this case failed to reveal the differences between the fricatives which were apparent from visual inspection of the spectra and the perceived auditory differences. On the other hand, the discrete cosine transformation coefficients proved to better quantify these differences (Jannedy & Weirich, 2017, p. 397).

In conclusion, the provided studies demonstrate that fricatives and their spectral moments can be highly useful in forensic phonetics, particularly for speaker discrimination. The discrete cosine transform is also a valuable method, and in some cases, it yields even better results than spectral moment analysis. However, it is important to note that various factors can influence the measured parameters of fricatives. Within-speaker variability can be affected by linguistic context, articulatory strengthening (hyperarticulation) or weakening (hypoarticulation), speaking style, and in extreme cases even face-concealing garments and voice disguise. Differences in fricative parameters between speakers can be attributed to factors such as sex, individual anatomical differences, sociolinguistic factors, or even gender identity and sexual orientation (Smorenburg & Heeren, 2020, p. 949). Despite these factors, fricatives and their parameters have been proven to contain valuable speaker-discriminating information.

2.5 Impact of telephone signal on fricatives

It has already been established that recordings of telephone speech play a crucial role in forensic phonetics. Additionally, it has been proven that the acoustic parameters of fricatives contain valuable speaker-discriminating information. A natural question that arises is whether these measures retain their speaker-discriminating potential even in telephone speech. This chapter will briefly summarize the effect of telephone transmission on fricatives and how it might influence the values of their measured acoustic properties.

Humans perceive sounds between 20 Hz and 20 kHz. Speech usually consists of frequencies up to 10 kHz; however, the most relevant information is usually contained within frequencies below 8 kHz. Human speech is also reported to have significant energy between 200 Hz and 4 kHz. Telephones and mobile phones take advantage of this fact, and to save the data carrying capacity and storage capacity, they work within a limited bandwidth. The narrowband codecs (primarily used in 2G) implement a telephone bandwidth of 300 Hz to 3,400 Hz (Christensen, 2023, p. 54). The wideband codecs implement a telephone bandwidth that excludes speech sounds between 7-8 kHz and 10 kHz (Christensen, 2023, p. 55). Christensen explains that since the frequencies that contain speech-related acoustic information are excluded, the acoustic quality of the signal may be degraded (2023, p. 55). This information is crucial especially for forensic phonetics, because the speaker-discrimination information of the given segments might be affected as well. Christensen in her study concluded that digital transmission makes otherwise reliable acoustic measures less reliable. She further mentions that spectral measures, which are known to distinguish e.g. sibilants from non-sibilants, were found to be significantly lowered and make the fricatives almost identical (2023, p. 399). Since it was already established earlier in the study that some fricatives contain important information in their higher frequencies, it is now clear why they are particularly affected by telephone speech.

From a technical perspective, the frequencies are dependent on the sampling rate (number of sample points per second). In other words, “the sampling rate is directly related to the frequency range of the signal, as the upper limit of the bandwidth (in Hz/kHz) in a digital signal will always be half the sampling rate” (Christensen, 2023, p. 55). Therefore, it is evident that even in the newer technology that usually uses a sampling rate of 16 kHz, not all speech-related frequency information is included. In contrast, older or narrow-bandwidth technologies, such as those using an 8 kHz sampling rate, focus on preserving the most crucial components of speech intelligibility. However, while they maintain essential frequency ranges necessary for effective communication, they inevitably lose a significant amount of higher frequencies.

2.6 Hypotheses

This study primarily aims to obtain population statistics for the four spectral moments of the four Czech voiceless fricatives /s/, /ʃ/, /r̝̊/, and /x/. It is expected that the gathered data will provide clear, descriptive insights into the typical ranges, extreme values, and distribution characteristics of the spectral moments.

Another hypothesis is that the simulation of telephone transmission using a narrowband codec will alter the spectral moments of the fricatives more significantly and extremely than the simulation of telephone transmission using a wideband codec.

3. Practical part

Having established a theoretical foundation in the previous chapters, which provided information about forensic phonetics, fricatives, and how acoustic parameters of fricatives may be used in forensic phonetics, the study now transitions into its practical part. The methodological section will provide a detailed account of the procedures employed in handling the recordings and extracting data from them. It will also outline the methods used for data analysis. Lastly, the results will be presented, discussed and a conclusion will be drawn.

3.1 Method

3.1.1 Procedures concerning the recordings and data extraction

The recordings that were used in this study were kindly provided by the supervisor of this thesis. At first, I received 20 recordings of spontaneous speech from male speakers, which were all approximately one minute long. The first step in the process was to transcribe the content of the recordings. The completed text files were sent to my supervisor, who then provided me with automatically segmented text grids, containing three tiers – phone, word, and phrase. This significantly reduced the additional work that would have been required for manual segmentation. Nevertheless, the automatic segmentation is not perfect, and therefore manual adjusting of the phone borders was necessary. Adjusting the boundaries was carried out both visually and through careful auditory analysis in a computer software made for phonetic speech analysis called Praat (Boersma & Weenink, 2024). The recordings were all taken by the speakers themselves using their mobile phones. They were instructed to record themselves in a quiet environment ensuring there were no background noises or disturbances. In most cases, the recordings were of sufficient quality with no major disturbances. However, some of the recordings contained intrusive noises such as door banging, multiple people speaking at once, or persistent noise in the background which may be attributed to the use of lower-quality devices or a less-than-ideal environment. The visual and auditory analyses were more difficult in the recordings of lower quality and adjusting the phone boundaries took more time. Subsequently, I received an additional 40 recordings, also from male speakers, which were taken in the same conditions, and they were also approximately one minute long. These recordings required no further boundary adjustments, as they had already been processed by other students as part of their own projects.

Once the boundaries of phones were adjusted to be aligned with the sounds, the next step was to extract the spectral moments from all /s/, /ʃ/, /r̝̊/, and /x/ segments. For this purpose, creating a Praat script was essential, as manually extracting such a large volume of data would have been exceedingly time-consuming. My supervisor kindly provided me with a script, which was used for similar purposes in the past. Since the data from the previous project differed from those used in this study, the script needed to be adjusted accordingly. However, this modification was completed relatively quickly. The final script performed multiple tasks simultaneously. Firstly, the script went through all the segments and found the target ones. Then, it calculated a 30 ms window around the temporal midpoint of each target segment to ensure there were no residual effects from neighbouring sounds. The 30 ms sound window was then transformed into a spectrum, from which all four spectral moments were extracted. The extracted data were automatically written to an output text file. Lastly, it saved the 30 ms sound around the temporal midpoint of each target fricative as a WAV file for later use in Discrete Cosine Transform analysis. Although the script was capable of processing all the recordings and their corresponding text grids simultaneously, it could only extract and save data from one type of segment at once. Consequently, the script needed to be slightly modified and run four times and to ensure that data from all target segments were extracted. In total, spectral moments and 30 ms sounds were extracted from 1980 alveolar fricatives /s/, 526 post-alveolar fricatives /ʃ/, 378 velar fricatives /x/, and 278 (post)-alveolar fricative trills /r̝̊/.

In addition to obtaining population statistics on the spectral moment values of the four voiceless fricatives, this study also aimed to analyse how the spectral moments are affected by telephone transmission. In section 2.5 it was briefly discussed how certain frequencies get excluded from the signal depending on which telephone bandwidth is being used. Christensen (2023) only worked with wideband codecs which exclude speech sounds between 7-8 kHz and 10 kHz. She argued that the narrowband codecs that implement the telephone bandwidth of 300 Hz to 3,400 Hz are primarily used in 2G, and thus they were of no interest in her study (2023, p. 54). Although it is true that newer technologies often utilize wideband codecs and a 16 kHz sampling rate, it is still possible to encounter systems that employ narrowband codecs with an 8 kHz sampling rate. In a recent online article, Zerby mentions that narrowband codecs are sometimes still used in call centres, some meeting softwares, and even some phone calls (2024). For this reason, this study chose to simulate two types of phone calls—one using a wideband codec with a 16 kHz sample rate and a 12.65 Kbps bitrate, and the other a narrowband codec with an 8 kHz sample rate and a 12.20 Kbps bitrate —to examine how each affects the spectral moments of the target fricatives. For the telephone simulation, the computer program AVS Audio Converter was used. The program was user-friendly and efficiently converted all 60 recordings at once. However, it was not possible to save the recordings directly as WAV files, so they had to be re-uploaded in the new .amr format and then converted back to WAV to be compatible with Praat for further analysis. After the recordings were converted under the two conditions, they underwent the same process as the original 60 recordings. The same Praat script was used to exctract the spectral moments from all the target segments as well as the 30 ms sound excerpts. The three figures below present spectrograms of the same two words spoken by the same speaker in the mentioned conditions. Figure 9 shows the spectrogram of the original recording, taken in a quiet environment at home. Figure 10 displays the spectrogram after the original recording was converted using a wideband codec, while Figure 11 shows the spectrogram after the original recording was converted using a narrowband codec.

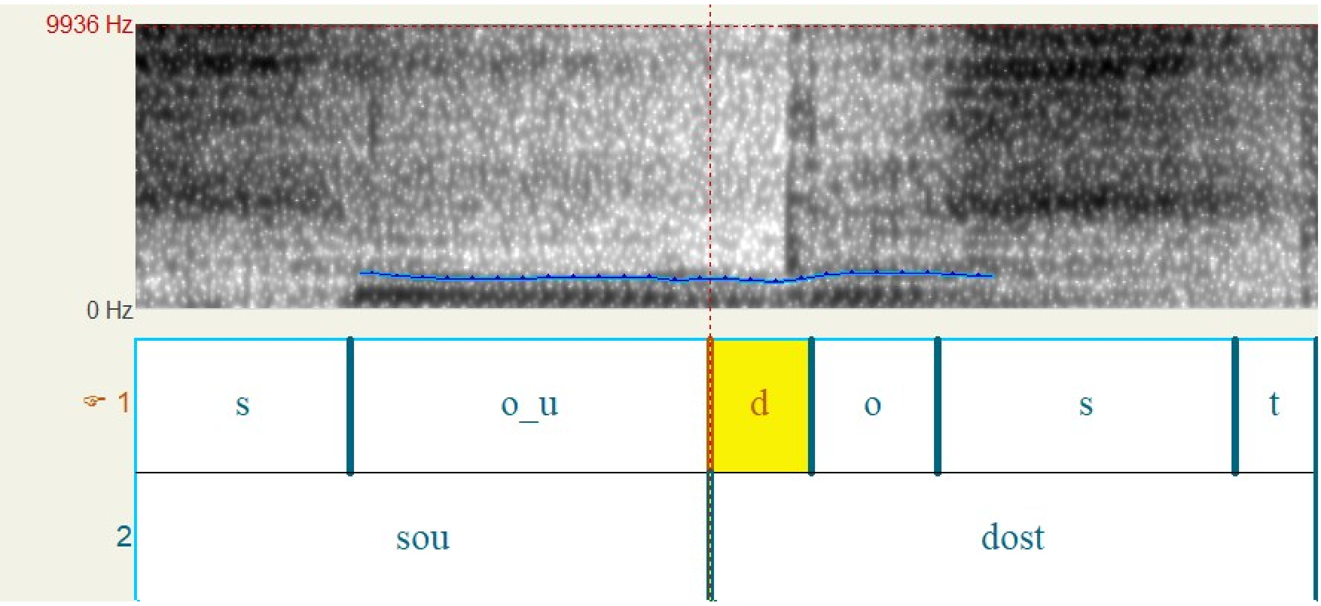


Figure 9: Spectrogram of the original recording

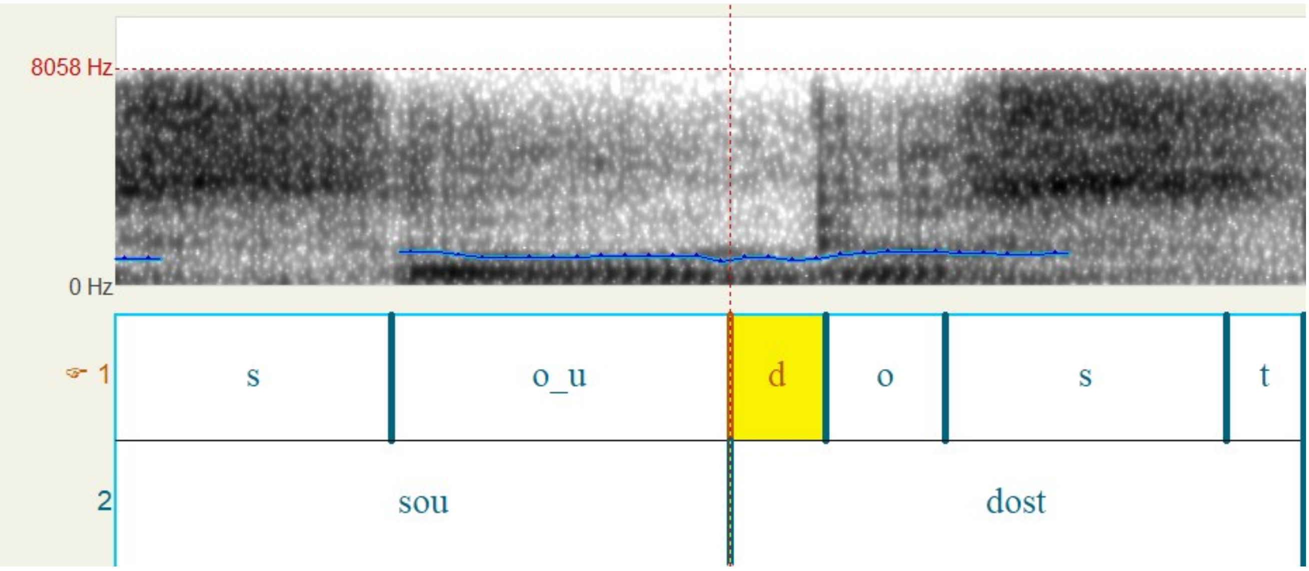


Figure 10: Spectrogram of the converted recording using a wideband codec

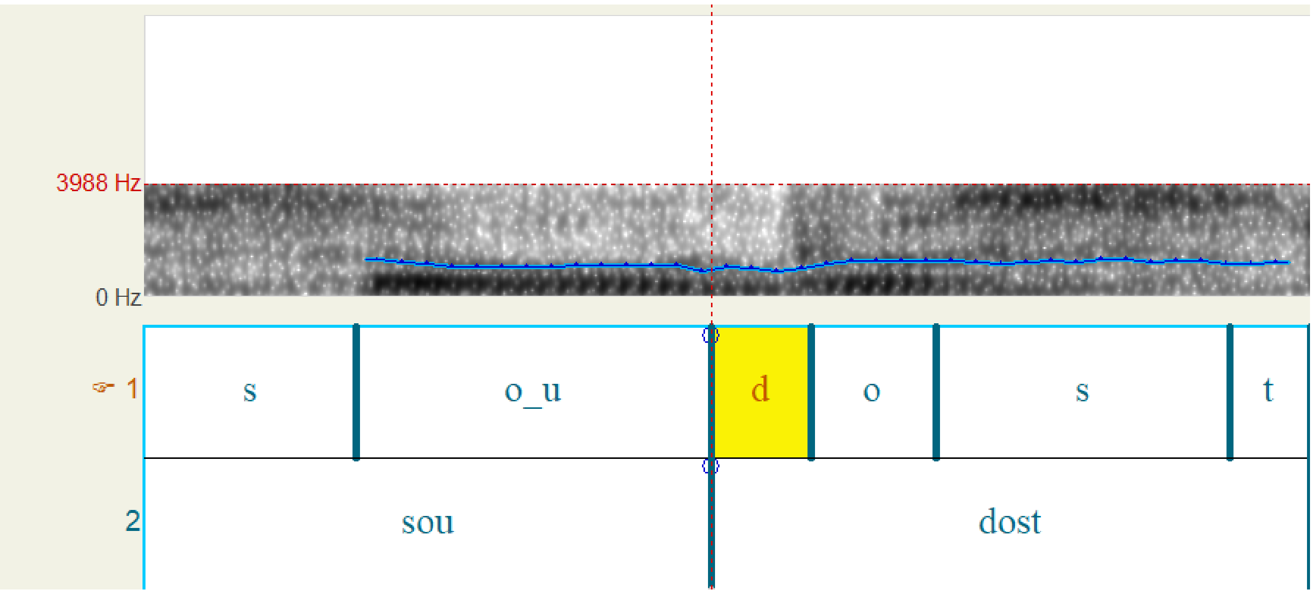


Figure 11: Spectrogram of the converted recording using a narrowband codec

The view range in Praat was set to 0-10 kHz in all three cases. As shown in Figure 9, the original recording displays frequencies up to 10 kHz. In Figure 10, it is apparent that frequencies above approximately 8 kHz were excluded. In Figure 11, all frequencies above 4 kHz were excluded.

3.1.2 Data analysis

Finally, after extracting all the spectral moments data from the original recordings and the two sets of converted recordings, the data were transferred from the text document to an Excel table. The table columns included the speaker code, segment ID, segment type (labeled as s, š, ř, and x for convenience instead of the IPA symbols s, ʃ, r̝̊, and x), all spectral moments (cog, sd, skew, kurt), and the recording type (original, wideband, narrowband). When the table was completed, the data analysis was performed using RStudio (R Core Team, 2023). It was necessary to write a script in R Studio to generate the graphs showing the results. The process of creating the script was significantly facilitated by the kind assistance of my supervisor. The data was imported using the readxl package (Wickham & Bryan, 2019), while data adjustments and analysis were conducted using the tidyverse package (Wickham et al., 2019).

Lastly, another R studio script was used to extract the four DCT coefficients from the 30 ms excerpts that reflect the mean amplitude of the spectrum (DCT0), the linear slope of the spectrum (DCT1), its curvature (DCT2), and the amplitude of the higher frequencies (DCT3) (Jannedy & Weirich, 2017, p. 399). The DCT coefficients were later transformed into graphs using a separate script. Both scripts, which exceeded my knowledge of the scripting language and therefore could not have been written by me, were kindly provided by a professor at the Institute of Phonetics, Faculty of Arts, Charles University.

Lastly, before presenting the results, it is important to note that the segment labels in the Excel table were represented as “s”, “š”, “ř”, and “x”, rather than the corresponding IPA symbols. Consequently, the graphs will also use these labels instead of the IPA symbols. This decision was made for convenience purposes: using graphemes like 'ř' is more simple than continuously copying IPA symbols such as 'r̝̊' when adjusting the script. Additionally, ensuring that the script ran smoothly was a priority. While R Studio generally handles IPA symbols well, there is a risk of errors, such as forgetting a diacritical mark in 'r̝̊', which could disrupt the script and require additional time for correction. Thus, using graphemes minimized potential issues in the entire process of generating graphs. Therefore, in the following chapters, only graphemes will be used.

3.1.3 DCT

Although Discrete Cosine Transform (DCT) was not officially part of the final analysis of the four fricatives, it was introduced in the study because it is used alongside the spectral moment analysis in related research. Therefore, I wanted to include at least examples of DCT graphs, showcasing all four different fricatives randomly selected from one speaker.

Figure 12 displays a DCT analysis of a randomly chosen fricative /s/ from the speaker BAJN.

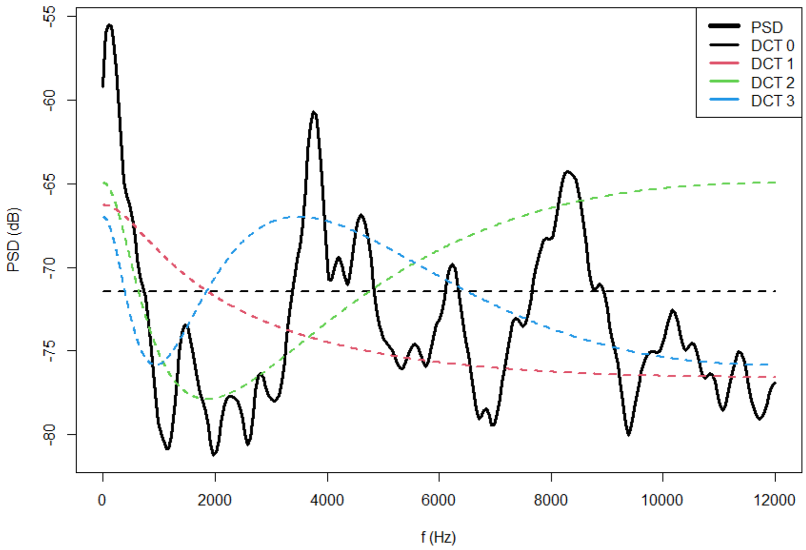


Figure 12: Discrete Cosine Transform (DCT) Analysis of the fricative /s/

The DCT coefficients are depicted by colours. DCT0 reflects the mean amplitude of the spectrum, the linear slope of the spectrum is reflected by DCT1, the curvature by DCT2, and the amplitude of the higher frequencies by DCT3.

Figure 13 displays a DCT analysis of a randomly chosen fricative /š/ from the speaker BAJN.

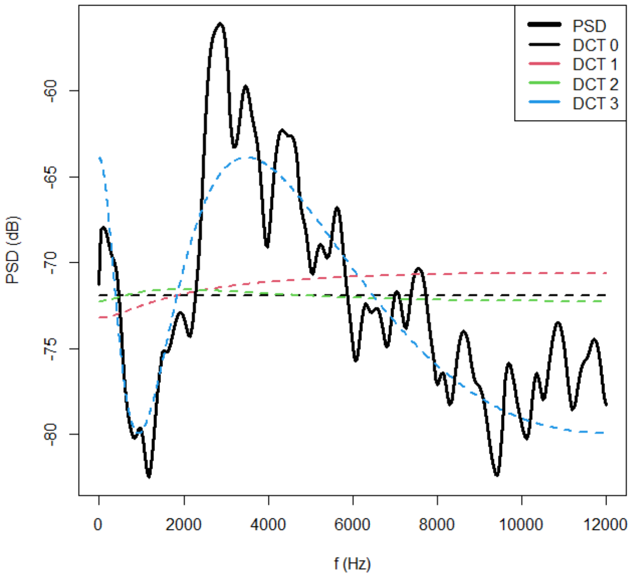


Figure 13: Discrete Cosine Transform (DCT) Analysis of the fricative /š/

Figure 14 displays a DCT analysis of a randomly chosen fricative /ř/ from the speaker BAJN.

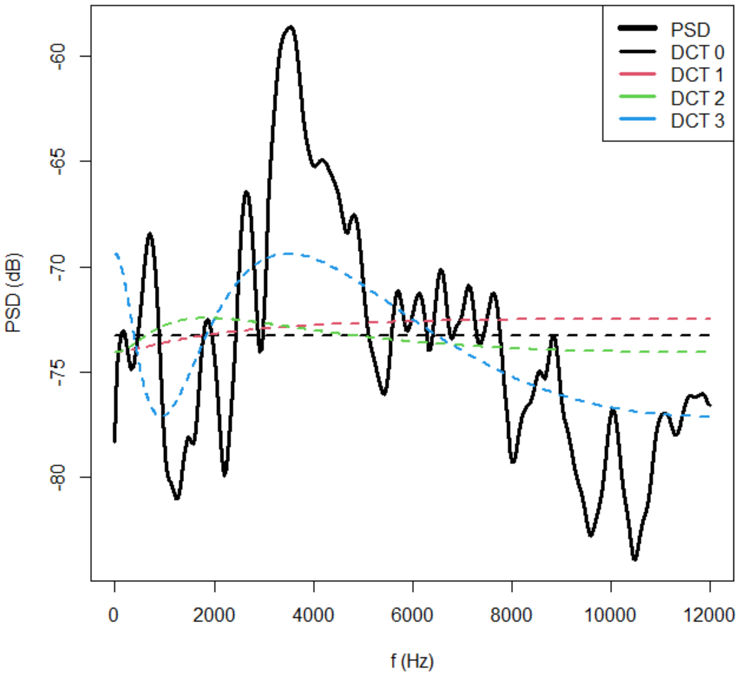


Figure 14: Discrete Cosine Transform (DCT) Analysis of the fricative /ř/

Figure 14 displays a DCT analysis of a randomly chosen fricative /ř/ from the speaker BAJN.

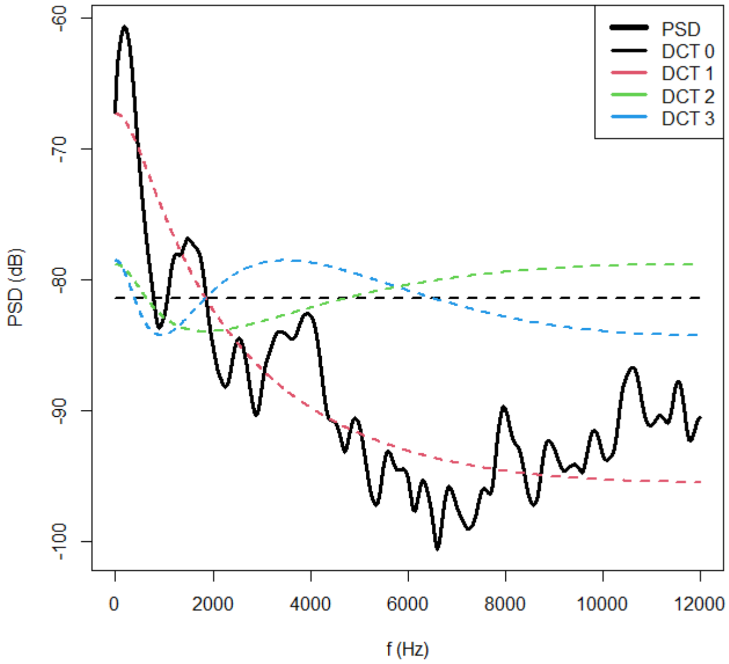


Figure 15: Discrete Cosine Transform (DCT) Analysis of the fricative /ř/

Although DCT is particularly useful for this study, as it focuses on analysing individual segments, it could be valuable in future research examining within-speaker variability and comparing different realizations of fricatives from the same speaker. Additionally, it may be used to compare realizations of the same fricative across different speakers, although this aspect is beyond the scope of this thesis. In both cases of within-speaker and between-speaker variation, it may provide more detailed information than spectral moments analysis.

4. Results and discussion

In the following chapter, the results of this study will be presented and discussed. The chapter is divided into three sections. The first section provides population statistics for the first four spectral moments of the four target Czech voiceless fricatives /s/, /š/, /ř/, and /x/ in original recordings. The second section examines within-speaker variability of the spectral moments across individual speakers. The third section analyzes the spectral moments of the target fricatives in recordings that were converted into two types of telephone signals—one using a wideband codec and the other a narrowband codec. The values of the spectral moments in these two conditions are then compared with the values in the original recordings.

4.1 Population statistics for the first four spectral moments of the four target fricatives

In this chapter, the population statistics for the first four spectral moments—center of gravity (cog), standard deviation (sd), skewness, and kurtosis—of the four target Czech voiceless fricatives: /s/, /š/, /ř/, and /x/, are presented. Having these data is crucial for identifying what is typical and what is rare in the population, which can be particularly useful in future analyses, especially within a forensic context.

4.1.1 Center of gravity

Figure 16 shows a boxplot depicting the distribution of center of gravity (COG) values for the four target fricative segments. The x-axis represents the four segments, each categorized and differentiated by color, allowing for a visual comparison of their respective COG distributions. The y-axis displays the COG values in Hertz (Hz).

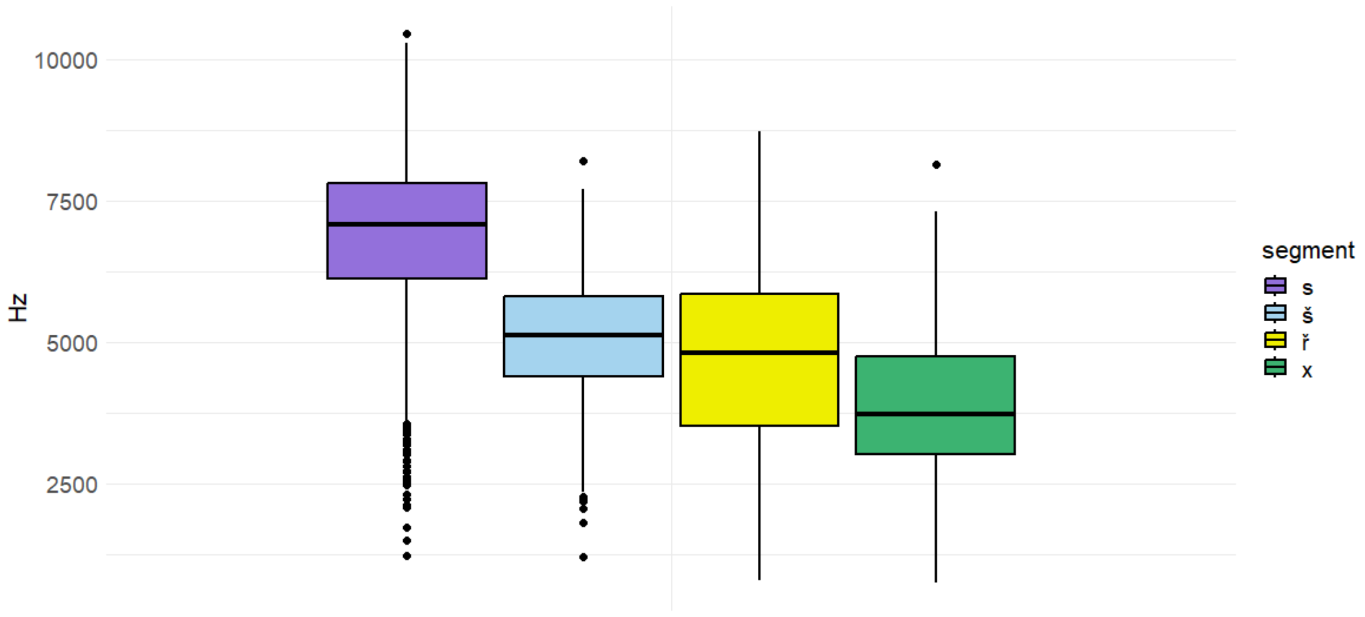


Figure 16: The mean and range of COG of /s/, /š/, /ř/ and /x/

The median COG values for the fricative segments are as follows: the median value for /s/ is **7073 Hz**, for /š/ is **5117 Hz**, for /ř/ is **4815 Hz,** and for /x/ is **3731 Hz**.

Additionally, Table 1 presents the Interquartile Ranges (IQR) for the center of gravity (COG) values across all segments. It provides a detailed view of the spread of the middle 50% of the data for each segment, complementing the visual representation in the boxplot. For example, the table indicates that 50% of all measured COG values for the alveolar fricative /s/ fall within the range of 6199 Hz to 7804 Hz.

Table 1: First quartile, third quartile, and Interquartile Range (IQR) of center of gravity (COG) for each target fricative

Segment	Q1	Q3	IQR
s	6119 Hz	7804 Hz	1685 Hz
š	4391 Hz	5805 Hz	1414 Hz
ř	3510 Hz	5860 Hz	2351 Hz
x	3020 Hz	4750 Hz	1730 Hz

The segment /ř/ has the highest IQR, which means that the middle 50% of its COG values are spread over the widest range compared to the other segments. However, it is important to note here that the number of /ř/ segments was the lowest out of all the target fricatives. In contrast, the segment /š/ has the lowest IQR, indicating less variability and a more concentrated range of the middle 50% of values.

The final observation in Figure 16 is that the COG values for /s/ exhibit the highest number of outliers, most of which are in the low frequencies, falling well below Q1 and outside the range of 1.5 times the IQR from Q1 and Q3. Once again though, it must be noted that the number of /s/ segments was the highest out of all target fricatives.

While the box plot effectively illustrates the median, quartiles, and the range of the data, it fails to convey information about the distribution's density. To make up for this limitation, Figure 17 displays a density plot, which provides a more detailed view of the data's distribution, highlighting areas where values are more concentrated and where they are less frequent.

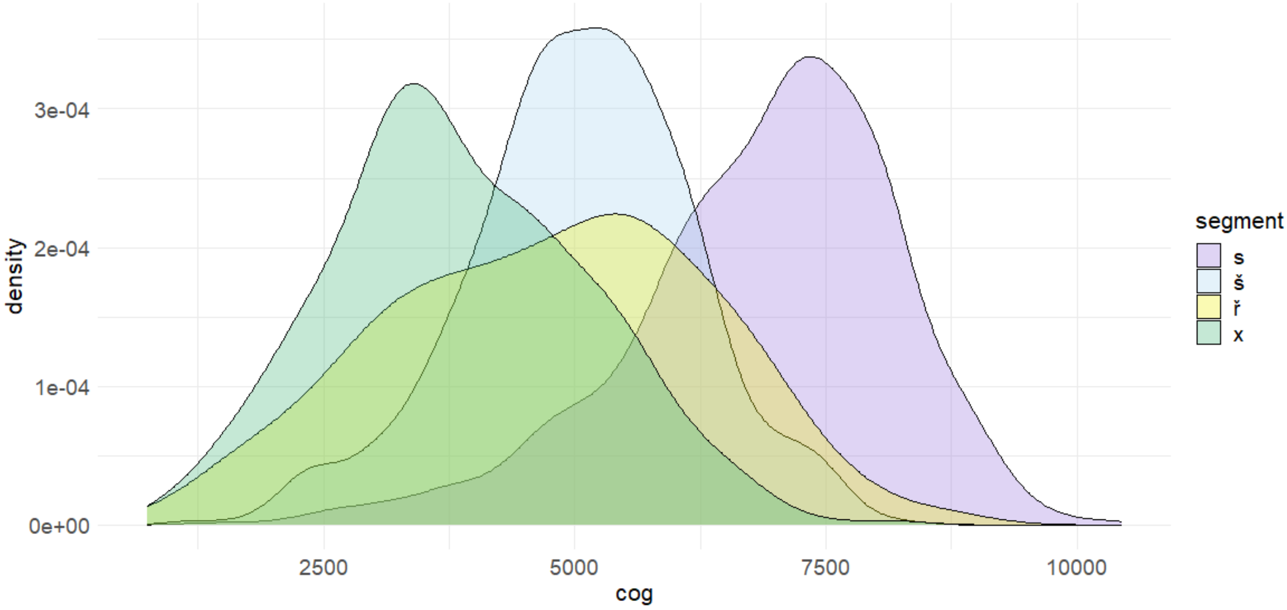


Figure 17: Density distribution of COG by segment

The density distribution for COG values of /s/ displays a negative skew. It indicates that there is a higher concentration of values in the higher frequency range, as the density curve extends more towards the lower frequencies while the peak of the distribution is shifted towards the higher frequencies. In contrast, /x/ displays a positive skew, which suggests that more COG values are concentrated in the lower frequency range, with the distribution extending more towards the higher frequencies. The density distribution for the COG values of /š/ appears to be normally distributed indicating a symmetric distribution of values around the mean. The density distribution for the COG values of /ř/ shows a more even spread of values and a less pronounced peak compared to the other segments.

Lastly, Figure 18 displays a violin plot that effectively combines both the box plot and the density plot features. While it does not introduce any new information, it provides a more nuanced view by integrating the summary statistics of the box plot with the distribution shape from the density plot.

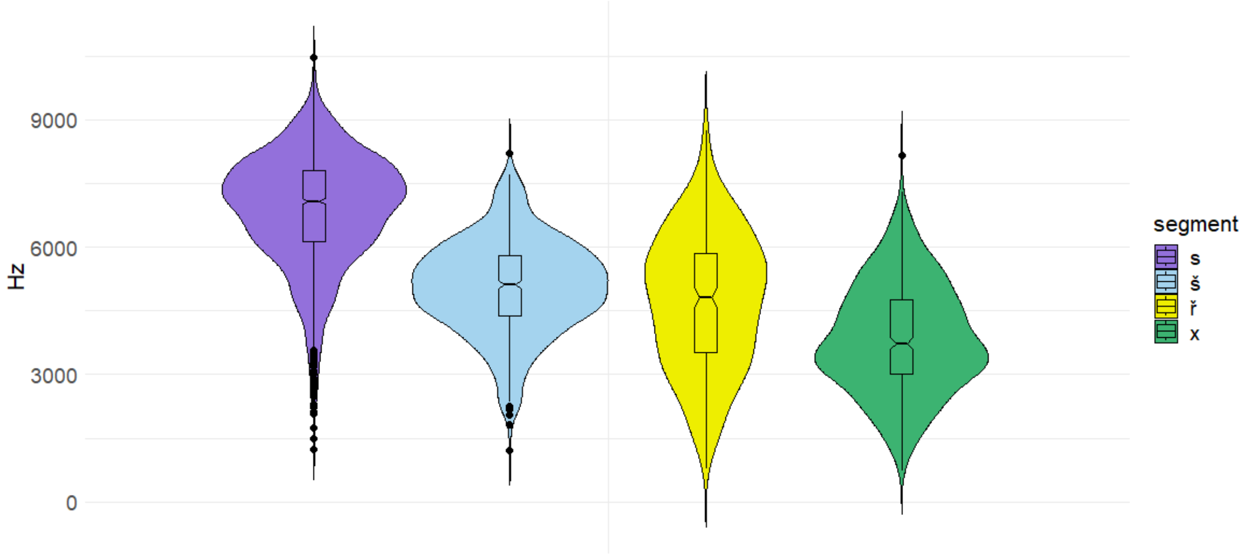


Figure 18: Distribution of COG by segment illustrated by a violin plot

So far, the analysis of center of gravity (COG) values was done by examining all the target fricatives in graphs together. To gain additional information about each of the target segments, in the following sections they will be briefly discussed individually.

4.1.1.1 /s/

Figure 19 showcases a histogram of the center of gravity (COG) values for the fricative /s/ only. Unlike the graphs above, the one presented here visually showcases percentiles, which allows us to see the distribution of values more efficiently.

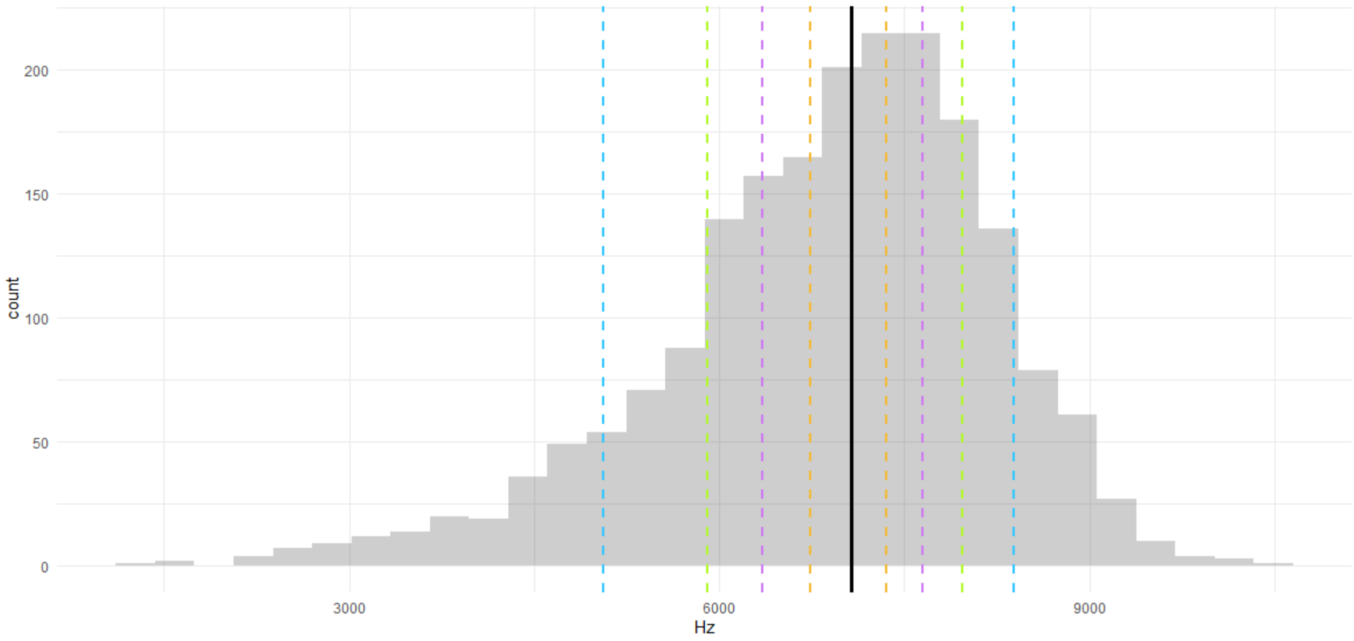


Figure 19: Distribution of COG for /s/ showcasing the median and percentiles

Firstly, the black line in the middle signifies the median of the COG values for /s/. As already stated in section 4.1.1, its value is **7073 Hz.** The blue dashed line on the left side of the median points to the 10th percentile, while the blue dashed line on the right side of the median points to the 90th percentile. This means that only 10% of the overall data lies below the lower blue line, and only 10% of the overall data lies above the upper blue line. In other words, 80% of all the values lay between the two blue lines. The green dashed lines represent the 20th and 80th percentiles. The purple dashed lines point to the 30th and 70th percentiles. Lastly, the orange lines point to the 40th and 60th percentiles. The exact values of each quantile will be given in the table below.

Table 2: Quantile values of the center of gravity in fricative /s/

Quantile	**Value (Hz)**
0.10 (= 10th percentile)	5057 Hz
0.20	5902 Hz
0.30	6350 Hz
0.40	6737 Hz
0.50 (Median)	7073 Hz
0.60	7354 Hz
0.70	7645 Hz
0.80	7974 Hz

0.90 (=90th percentile)	8387 Hz

Knowing the exact values of each quantile is a valuable piece of information because it allows us to approximately determine what values are typical and on the other hand what values are rare, which was one of the aims of this study. For example, it is now known that if the value of COG for /s/ was below 5057 Hz, it would be extremely rare. The sample applies for values above 8387 Hz. However, an exact threshold of what values can be considered typical and what can be considered rare is beyond the scope of this thesis and requires future analysis.

4.1.1.2 /š/

Figure 20 showcases a histogram of the center of gravity (COG) values for the fricative /š/.

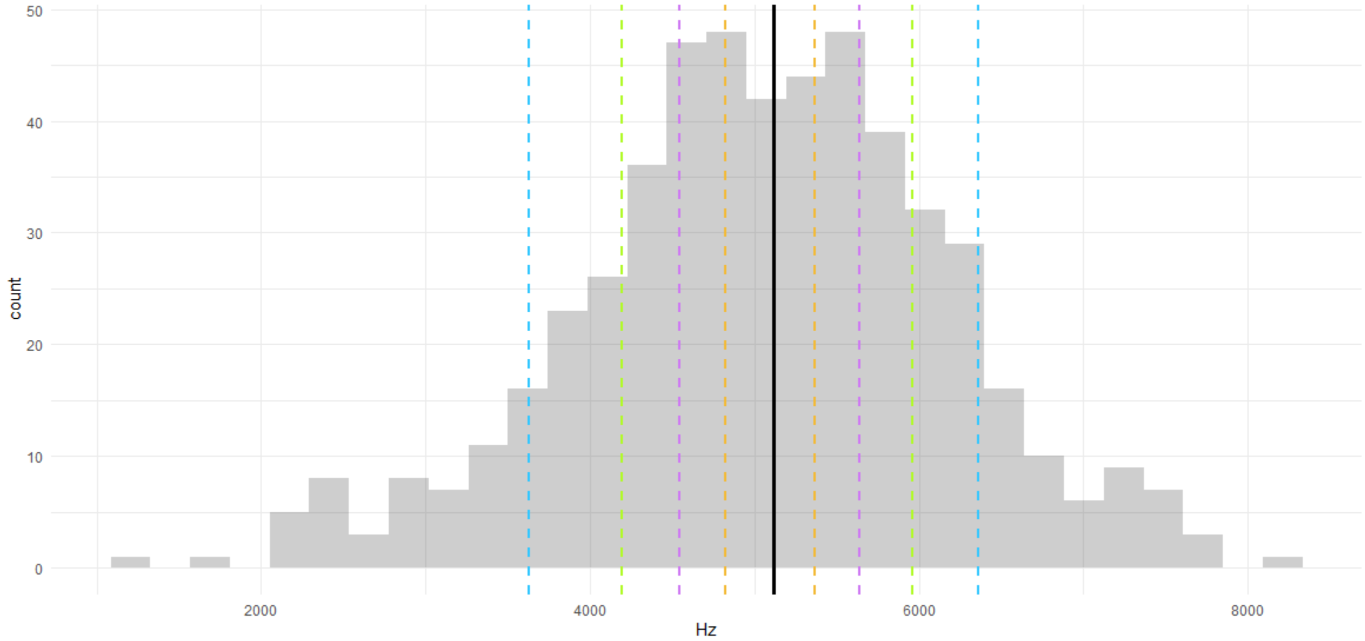


Figure 20: Distribution of COG for /š/ showcasing the median and percentiles

Just like in Figure 19, the black line in the middle shows the median of the COG values for /š/ and its value is **5117 Hz**. The coloured lines signifying the percentiles also remained the same as in Figure 15. The blue dashed line on the left side of the median points to the 10th percentile, while the blue dashed line on the right side of the median points to the 90th percentile. Once again, the green dashed lines represent the 20th and 80th percentiles, the purple dashed lines point to the 30th and 70th percentiles and the orange lines point to the 40th and 60th percentiles. The exact values of each quantile will be given in the table below.

Table 3: Quantile values of COG in fricative /š/

Quantile	Value (Hz)
0.10 (= 10th percentile)	3626 Hz
0.20	4194 Hz
0.30	4544 Hz
0.40	4821 Hz
0.50 (Median)	5177 Hz
0.60	5364 Hz
0.70	5641 Hz
0.80	5963 Hz
0.90 (=90th percentile)	6364 Hz

4.1.1.3 /ř/

Figure 21 showcases a histogram of the center of gravity (COG) values for the fricative /ř/.

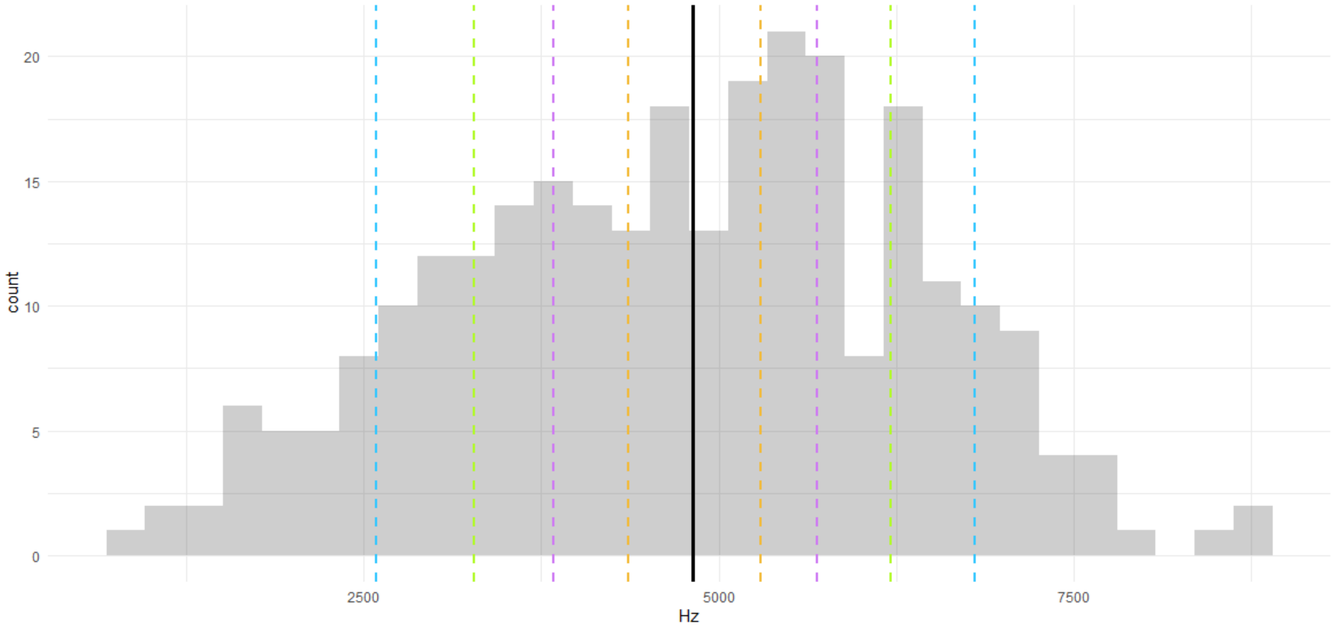


Figure 21: Distribution of COG for /ř/ showcasing the median and percentiles

Once again and just like in Figures 19 and 20, the lines signify the same values. The black line in the middle shows the median of the COG values for /ř/ and its value is **4815 Hz**. The blue dashed lines point to the 10th and the 90th percentile. The green dashed lines represent the 20th and 80th percentiles, the purple dashed lines point to the 30th and 70th percentiles and the orange lines point to the 40th and 60th percentiles. The exact values of each quantile will be given in the table below.

Table 4: Quantile values of COG in fricative /ř/

Quantile	Value (Hz)
0.10 (= 10th percentile)	2585 Hz
0.20	3727 Hz
0.30	3836 Hz
0.40	4365 Hz
0.50 (Median)	4815 Hz
0.60	5297 Hz
0.70	5689 Hz
0.80	6210 Hz
0.90 (=90th percentile)	6804 Hz

4.1.1.4 /x/

Figure 22 below displays a histogram of the center of gravity (COG) values for the fricative /x/.

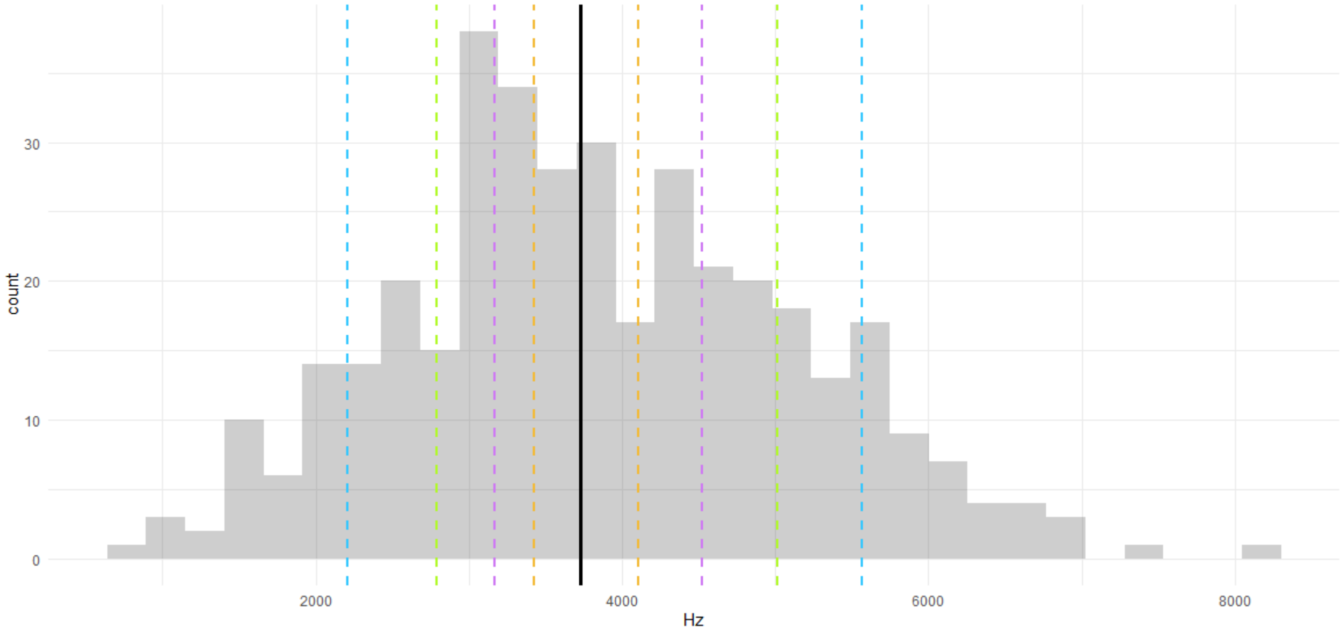


Figure 22: Distribution of COG for /x/ showcasing the median and percentiles

Once again and just like in Figures before, the lines signify the same values. The black line in the middle shows the median of the COG values for /x/ and its value is **4815 Hz**. The blue dashed lines point to the 10th and the 90th percentile. The green dashed lines represent the 20th and 80th percentiles, the purple dashed lines point to the 30th and 70th percentiles and the orange lines point to the 40th and 60th percentiles. The exact values of each quantile will be given in the table below.

Table 5: Quantile values of COG in fricative /x/

Quantile	**Value (Hz)**
0.10 (= 10th percentile)	2210 Hz
0.20	2792 Hz
0.30	3164 Hz
0.40	3425 Hz
0.50 (Median)	3731 Hz
0.60	4107 Hz
0.70	4525 Hz
0.80	5011 Hz
0.90 (=90th percentile)	5569 Hz

4.1.2 Standard deviation

Figure 23 displays a boxplot depicting the distribution of standard deviation (SD) values for the four target fricative segments. The x-axis represents the four segments, once again each differentiated by color for easy visual comparison The y-axis displays the SD values in Hertz (Hz).

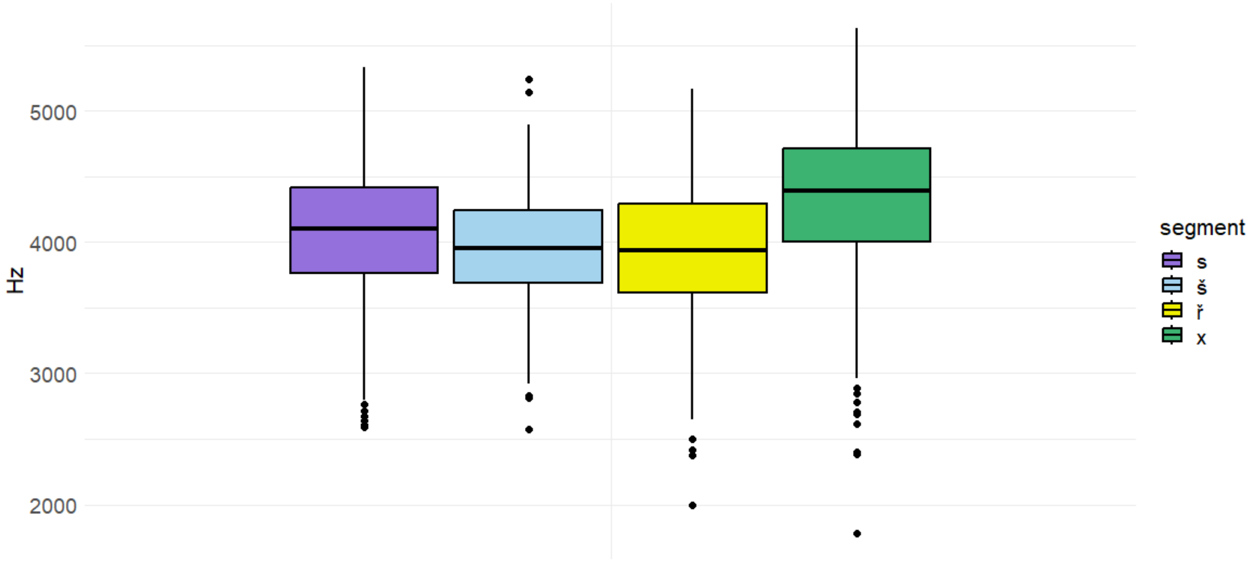


Figure 23: The distribution of SD for /s/, /š/, /ř/, and /x/

The median SD values for the target segments are as follows: the median value for /s/ is **4101 Hz**, for /š/ is **3954 Hz**, for /ř/ is **3939 Hz,** and for /x/ is **4388 Hz**.

Table 6: First quartile, third quartile, and Interquartile Range (IQR) of (SD) values for each target fricative

Segment	Q1	Q3	IQR
s	3761 Hz	4414 Hz	653 Hz
š	3688 Hz	4238 Hz	550 Hz
ř	3613 Hz	4290 Hz	678 Hz
x	4002 Hz	4709 Hz	707 Hz

The fricative /x/ has the highest IQR, which means that the middle 50% of its SD values are spread over the widest range compared to the other segments. On the other hand, the fricative /š/ has the lowest IQR, once again indicating a more concentrated range of the middle 50% of values.

It must be noted that /x/ is the only segment whose mean standard deviation value is higher than its mean center of gravity value. It signifies that the frequencies present in the spectrum of /x/ vary the most in relation to its mean.

Since the density plot for the standard deviation values of all target fricatives showed a significant amount of overlapping and appeared too cluttered, a violin plot was chosen instead. It is displayed in Figure 24.

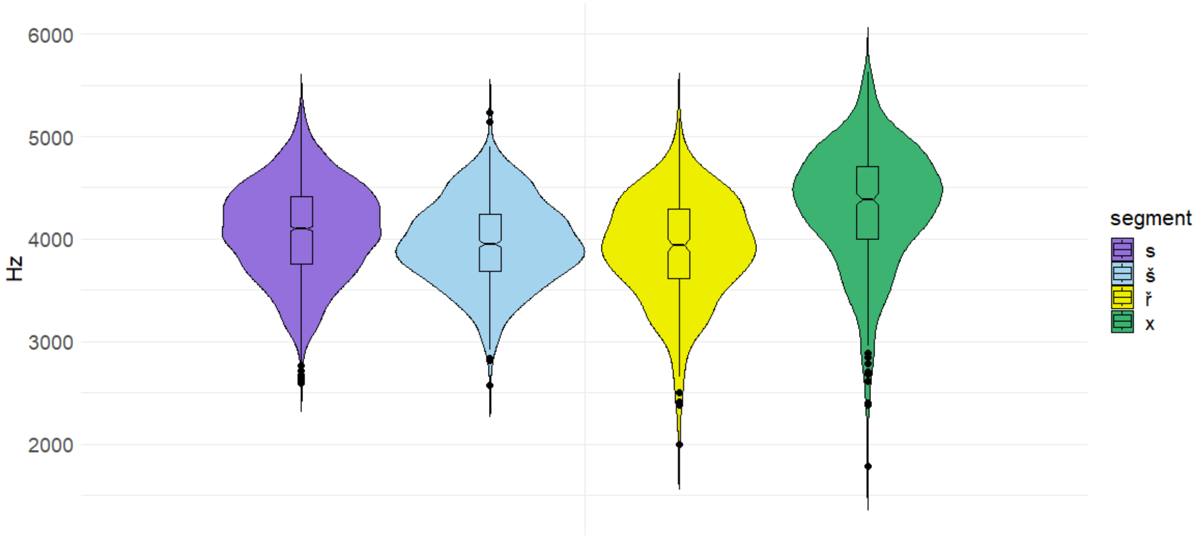


Figure 24: Distribution of SD by segment illustrated by a violin plot

Just like in section 4.1.1 the analysis of standard deviation (SD) values was so far done by examining all the target fricatives together. Once again to gain additional insight about each of the target segments, the following sections will discuss them individually.

4.1.2.1 /s/

Figure 25 displays a histogram of standard deviation (SD) values for the fricative /s/ alone. Once again, unlike the graphs in section 4.1.2, the one presented here visually showcases percentiles with lines, allowing us to see the distribution of values more efficiently.

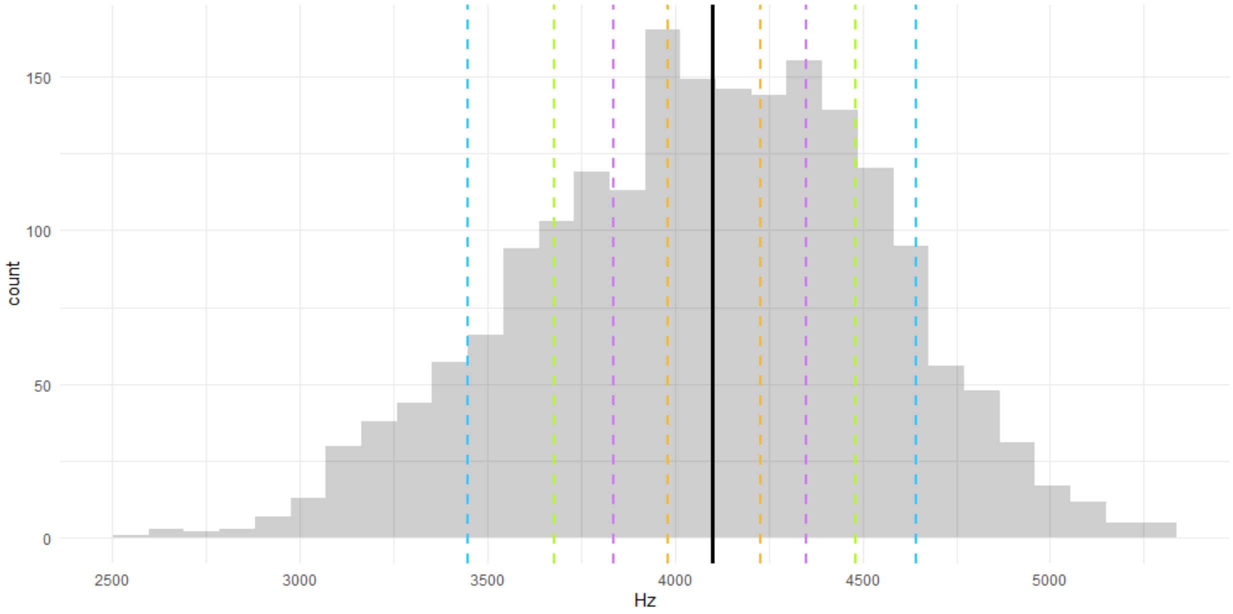


Figure 25: Distribution of SD for /s/ showcasing the median and percentiles

The system of showing the median and percentiles is the same as in earlier graphs concerning the values of center of gravity. Therefore, the black line in the middle signifies the median of the SD values for /s/. As already mentioned in section 4.1.2, its value is **4101 Hz.** The blue dashed line on the left side of the median points to the 10th percentile, while the blue dashed line on the right side of the median points to the 90th percentile. The green dashed lines represent the 20th and 80th percentiles, and the purple dashed lines point to the 30th and 70th percentiles. Finally, the orange lines point to the 40th and 60th percentiles. The exact values of each quantile will be given in the table below.

Table 7: Quantile values of the standard deviation in fricative /s/

Quantile	**Value (Hz)**
0.10 (= 10th percentile)	3449 Hz

0.20	3677 Hz
0.30	3837 Hz
0.40	3982 Hz
0.50 (Median)	4101 Hz
0.60	4229 Hz
0.70	4350 Hz
0.80	4483 Hz
0.90 (=90th percentile)	4644 Hz

4.1.2.2 /š/

Figure 26 displays a histogram of standard deviation (SD) values for the fricative /š/ alone. The median and percentiles are visually presented. To avoid repetitiveness, the colours of the dashed lines will no longer be described here, as the system remains the same as in all graphs above. Instead, the table of the exact quantile values will be given.

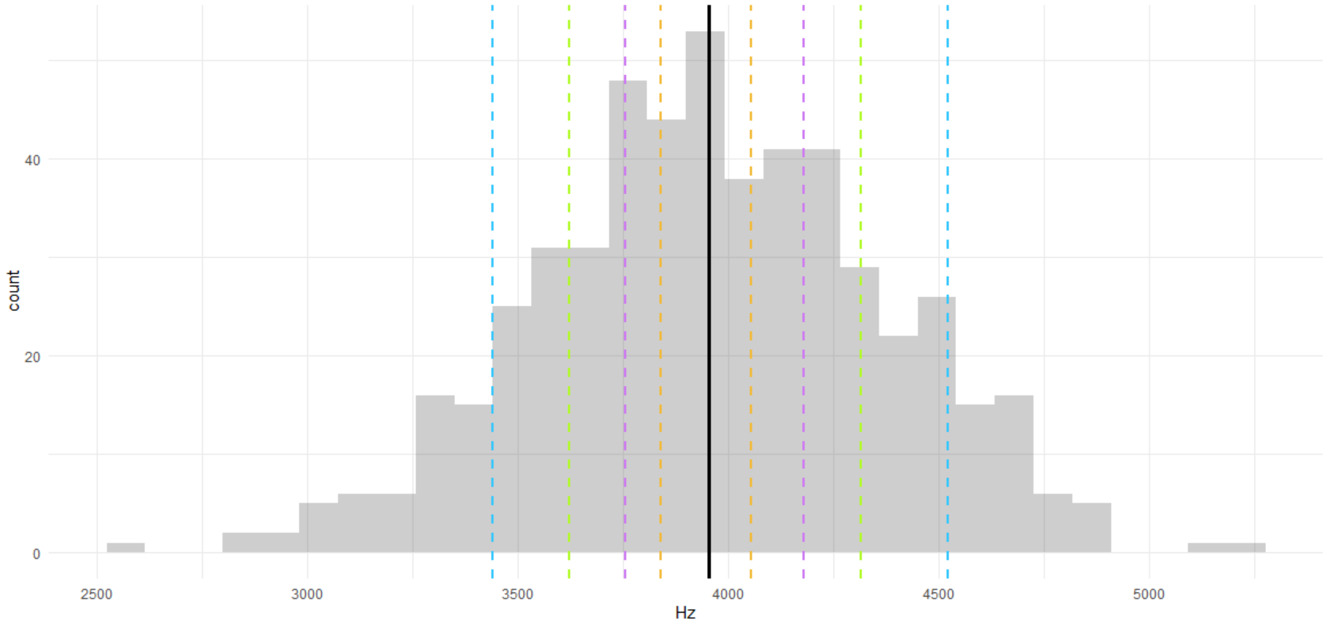


Figure 26: Distribution of SD for /š/ showcasing the median and percentiles

Table 8: Quantile values of the standard deviation in fricative /š/

Quantile	**Value (Hz)**
0.10 (= 10th percentile)	3411 Hz
0.20	3621 Hz

0.30	3756 Hz
0.40	3840 Hz
0.50 (Median)	3954 Hz
0.60	4055 Hz
0.70	4180 Hz
0.80	4314 Hz
0.90 (=90th percentile)	4521 Hz

4.1.2.3 /ř/

Figure 27 displays a histogram of standard deviation (SD) values for the fricative /ř/ alone. The median and percentiles are visually represented. The table of the exact quantile values is given below.

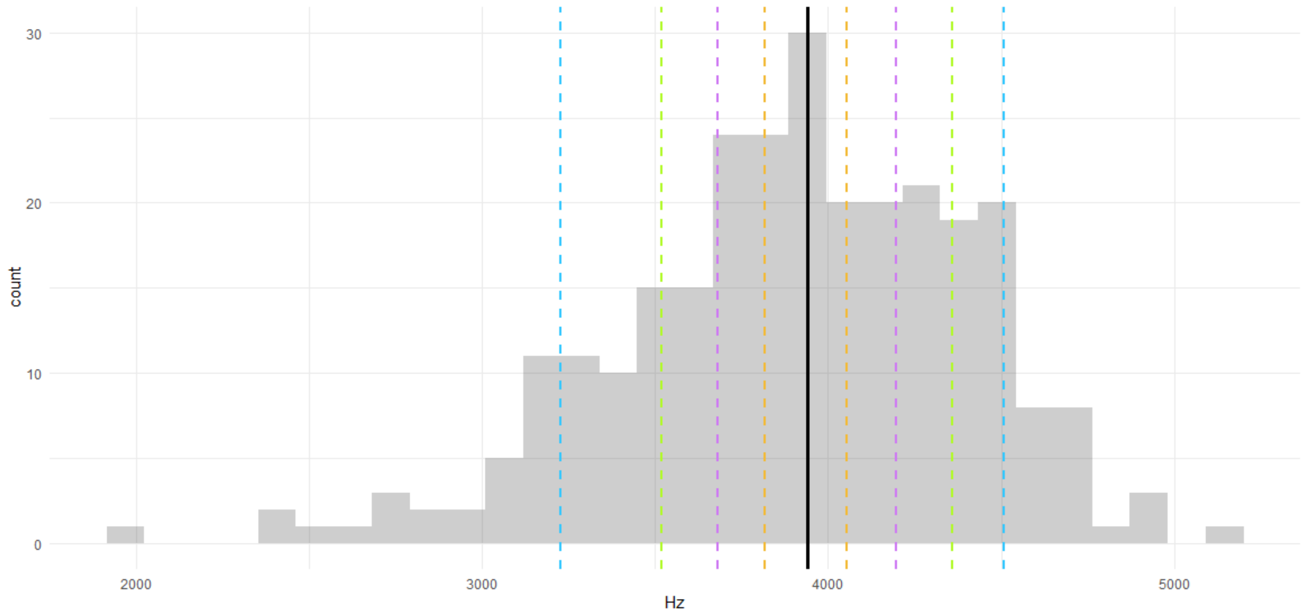


Figure 27: Distribution of SD for /ř/ showcasing the median and percentiles

Table 9: Quantile values of the standard deviation in fricative /ř/

Quantile	**Value (Hz)**
0.10 (= 10th percentile)	3227 Hz
0.20	3519 Hz
0.30	3682 Hz
0.40	3818 Hz

0.50 (Median)	3939 Hz
0.60	4053 Hz
0.70	4195 Hz
0.80	4357 Hz
0.90 (=90th percentile)	4506 Hz

4.1.2.4 /x/

Figure 28 displays a histogram of standard deviation (SD) values for the fricative /x/ alone. The median and percentiles are visually represented by lines. The table of the exact quantile values is given.

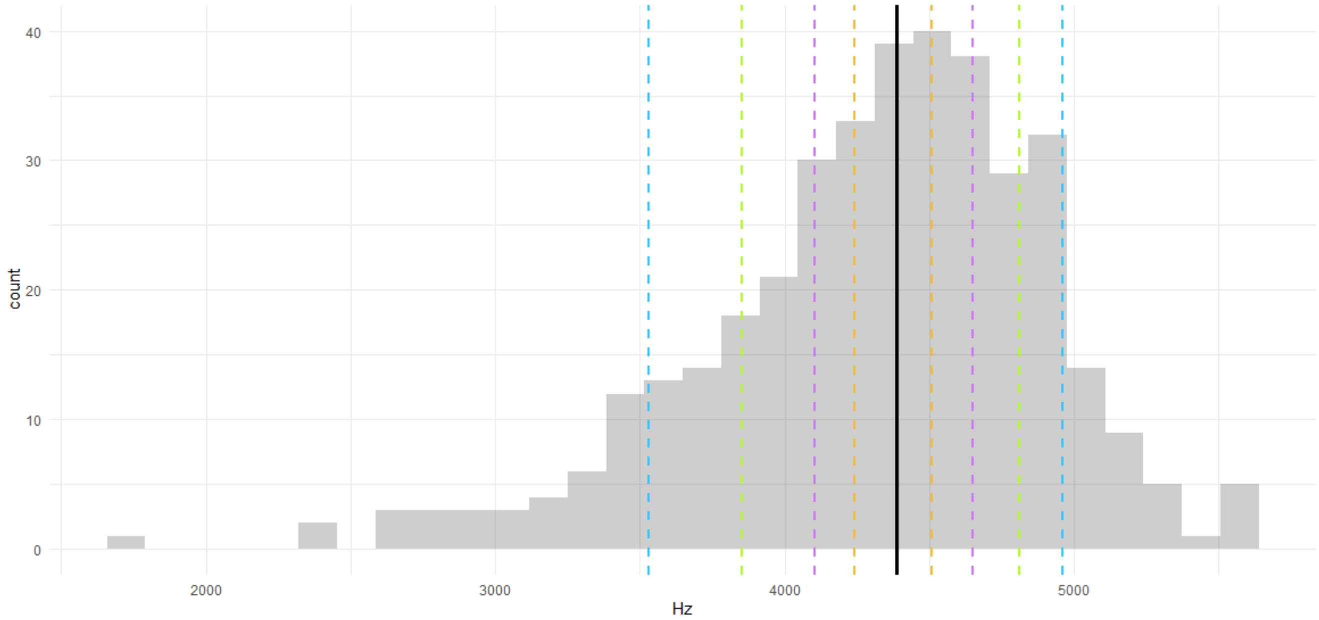


Figure 28: Distribution of SD for /x/ showcasing the median and percentiles

Quantile	**Value (Hz)**
0.10 (= 10th percentile)	3529 Hz
0.20	3854 Hz
0.30	4105 Hz
0.40	4242 Hz
0.50 (Median)	4388 Hz
0.60	4509 Hz
0.70	4648 Hz

0.80	4812 Hz
0.90 (=90th percentile)	4959 Hz

4.1.3 Skewness

Figure 29 showcases a boxplot depicting the distribution of skewness values for the four target fricative segments. The x-axis represents the four segments, each differentiated and categorized by color. The y-axis displays the values of skewness.

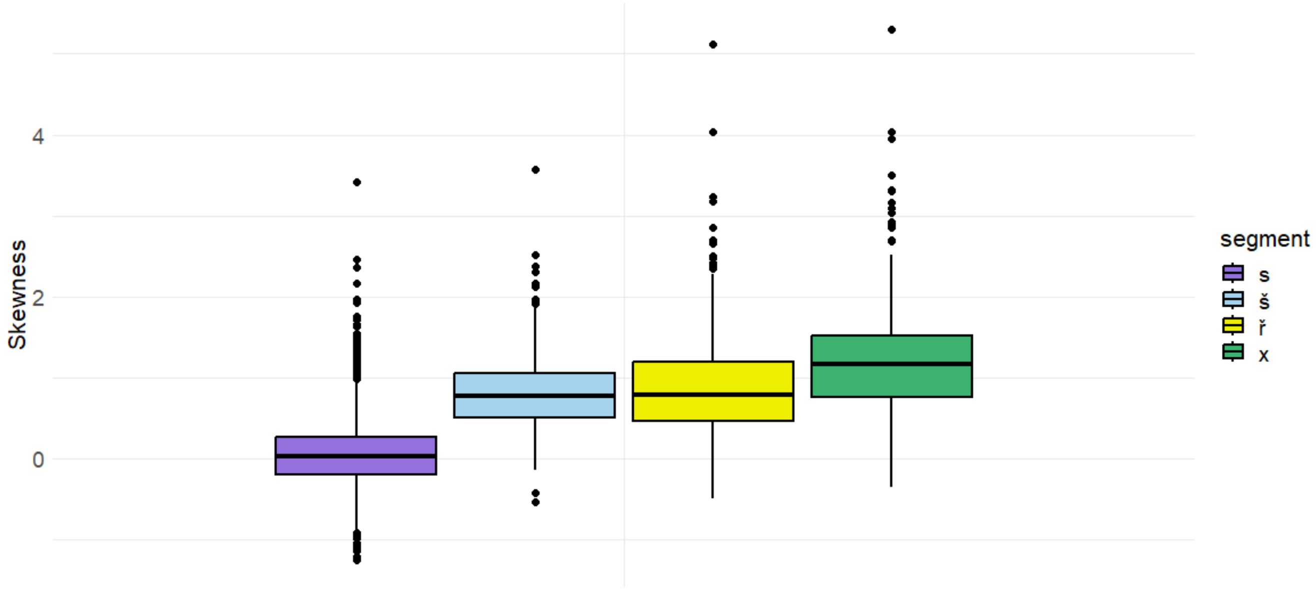


Figure 29: The distribution of skewness for /s/, /š/, /ř/ and /x/

The median skewness values for the target segments are as follows: the median value for /s/ is **0.02**, for /š/ is **0.075**, for /ř/ is **0.78,** and for /x/ is **1.17**.

Table 10: First quartile, third quartile, and Interquartile Range (IQR) of skewness values for each target fricative

Segment	Q1	Q3	IQR
s	-0.2	0.27	0.47
š	0.51	1.05	0.54
ř	0.46	1.20	0.737
x	0.76	1.52	0.758

The target fricative /s/ has the lowest IQR and 50% of its values are located between -0.2 and 0.027. Since skewness concerns the degree of asymmetry of the spectrum around the center of gravity, these values indicated that the distribution of energy is relatively symmetrical around the COG. However, as we can see in Figure 29 there are also numerous outliers located outside the range of 1.5 times the IQR from Q1 and Q3.

The target fricative /x / has the lowest IQR and 50% of its values are located between 0.76 and 1.52. This indicates a moderate positive skewness, with a concentration of energy in the lower frequencies. Moderate positive skewness can be observed in the remaining fricatives /š/ and /x/ as well. Even in the case of these three fricatives, we can observe several outliers, many of which exceed the value of 3, and some crossing the value of 4.

The density distribution of skewness values for all the target fricatives can be seen in Figure 30.

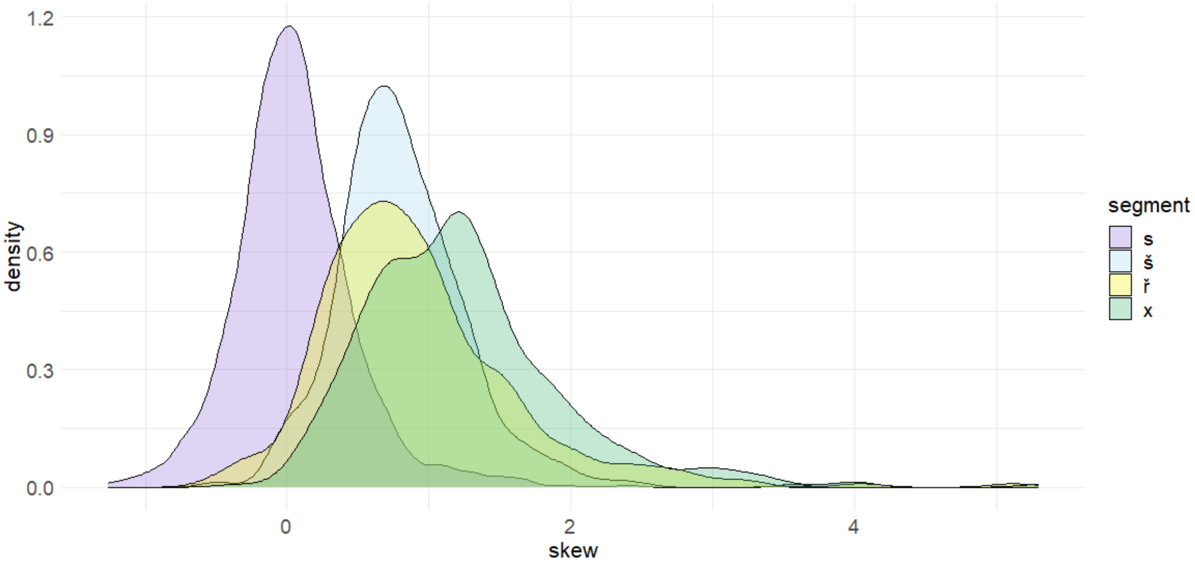


Figure 30: Density distribution of skewness values by segment

4.1.4 Kurtosis

Figure 31 displays a boxplot showing the distribution of kurtosis values for the four target fricatives. The x-axis represents the four fricatives, each differentiated and categorized by colour. The y-axis displays the values of kurtosis.

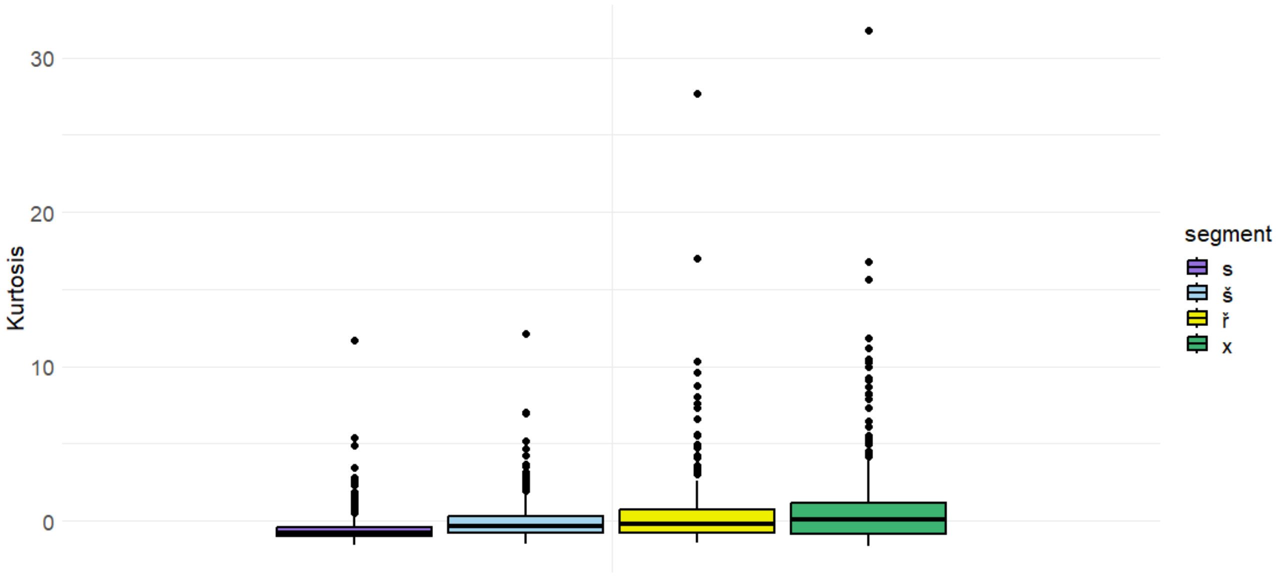


Figure 31: The distribution of kurtosis values for /s/, /š/, /ř/ and /x/

The median kurtosis values for the target segments are as follows: the median value for /s/ is **-0.76**, for /š/ is **-0.36**, for /ř/ is **-0.23,** and for /x/ is **0.06**.

Table 11: First quartile, third quartile, and Interquartile Range (IQR) of kurtosis values for each target fricative

Segment	Q1	Q3	IQR
s	-1.01	-0.47	0.59
š	-0.77	0.31	1.08
ř	-0.76	0.717	1.47
x	-0.827	1.12	1.95

The median kurtosis of all four segments appears to be relatively close to zero, with three out of four of the medians having a negative value. The interquartile range (IQR), of /s/, /š/, and /ř/ is quite narrow, suggesting that most of the 50% of the data points are close to the median. However, there are numerous outliers for each segment, especially for /x/ and /ř/, which indicates that there are instances where the kurtosis values are much higher than the rest of the data. The presence of many data points with negative kurtosis or kurtosis close to zero in all segments suggests that in many cases, the spectrum is relatively flat. On the other hand, the high values of the outliers in all segments point to instances where the spectrum has a sharp peak.

4.2 Within speaker variety

Until this point, the focus of the results section was to provide the population statistics of the four spectral moments of the target fricatives /s/, /š/, /ř/, and /x/. The shared results and values provided insight into the spectral characteristics of these four fricatives in a broader population of 60 male speakers. As already extensively mentioned in the theoretical part of this study, speech is inherently variable within individual speakers. This section of the study will therefore focus on within-speaker variability of two of the spectral moments (COG and SD) for the four target fricatives /s/, /š/, /ř/, and /x/ in the original (non-converted) recordings. Skewness and kurtosis are not included, because their overall value distribution is relatively narrow, and the graphs of within-speaker variety did not introduce any additional information.

4.2.1 Center of gravity

The following section will focus on within-speaker variety of center of gravity in all the target fricatives separately.

Figure 32 shows the distribution of center of gravity (COG) values of fricative /s/ in all speakers individually.

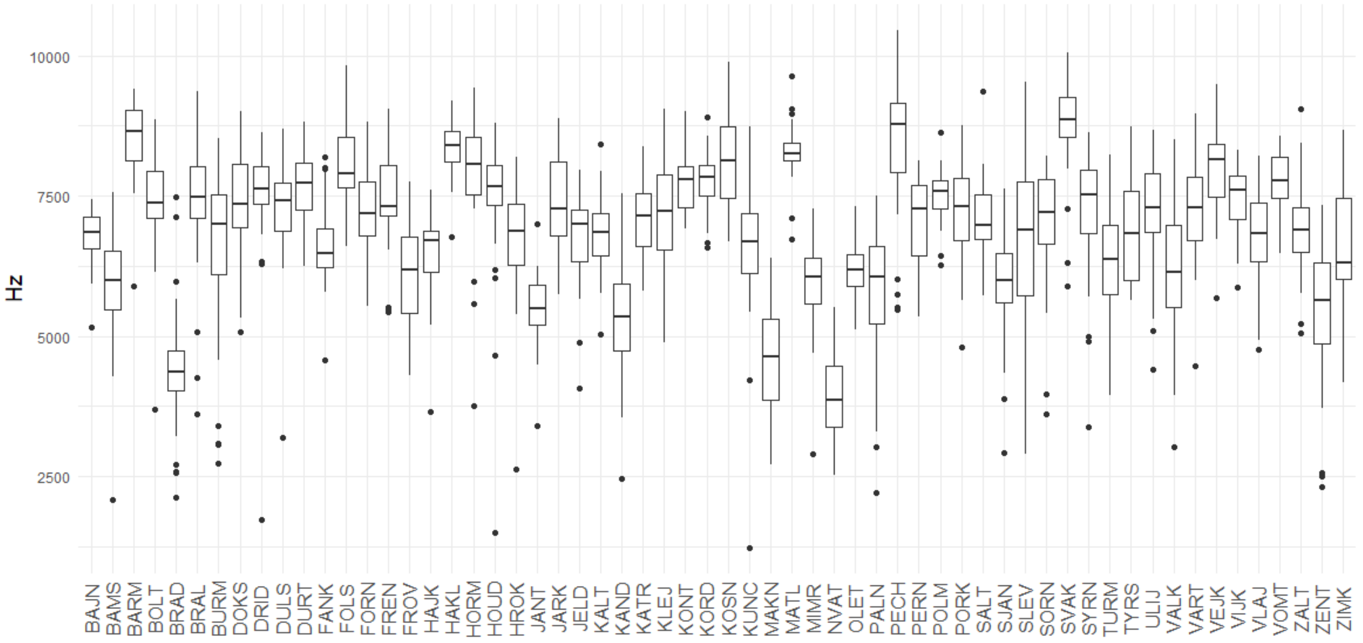


Figure 32: Mean and range of COG of /s/ by speaker

The speakers that are worth noting are those who have a very high within-speaker variety and on the other hand, those who have a low within-speaker variety. The speaker who stands out with his high within-speaker variety is SLEV. The noticeably large whiskers of the box plot indicate a wide range of COG values within this speaker's recording. On the other hand, speakers with a low within-speaker variety can be identified by the short whiskers of their box plot, indicating a narrow range of COG values within their recording. Such speakers are for example BAJN, HAKL, or MATL. Speakers, who are also worth pointing out, are those whose 50% of all realizations lie in extreme values on both ends. From Figure 19 which showcases a histogram of the center of gravity (COG) values for the fricative /s/, the values of the 10th and 90th percentile are known. The 10th percentile value is 5057 Hz and the 90th percentile value is 8387 Hz. If we observe speakers such as BRAD or NVAT, their mean COG and 50% of all their measured values of COG lie below 5000 Hz. This indicates that the COG values of these speakers are relatively rare and considered low in the measured population. The same could be said about speaker SVAK, whose values are considered high within the measured population.

Figure 33 displays the distribution of center of gravity (COG) values of fricative /š/ in all speakers individually.

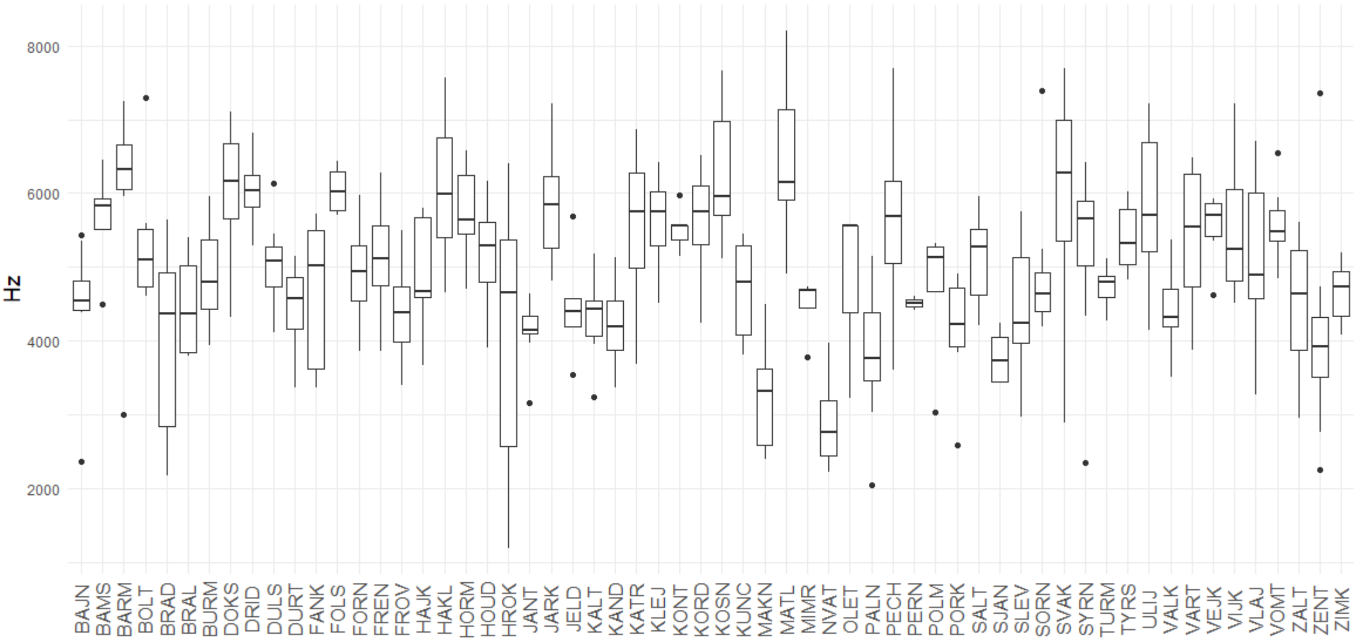


Figure 33: Mean and range of COG of /š/ by speaker

When it comes to the interpretation of within-speaker variety of any of the spectral moments in fricatives other than /s/, one must be careful. In the method section, it was mentioned that the number of extracted fricatives of each category was not the same. There was a disproportionately bigger number of /s/ segments extracted, and each speaker had approximately 30 realizations of /s/ in their recording. The same, however, cannot be said for the other fricatives. For example, if we observe Figure 33, it appears as if speaker PERN had an extremely narrow range of COG values within his recording. However, in his entire recording, there were only two realizations of /š/. Such a small number of /š/ realizations within one recording was relatively rare though. Other speakers, who display a narrow range of COG values and had a higher number of /š/ realizations are for example TURM, ZIMK, or JANT. A notably wide range of COG values can be observed in recordings of HROK and SVAK. Speakers, who are also worth pointing out, are MAKN and NVAK, since their values are considered rarely low within the measured population.

Figure 34 displays the distribution of center of gravity (COG) values of fricative /ř/ in all speakers individually.

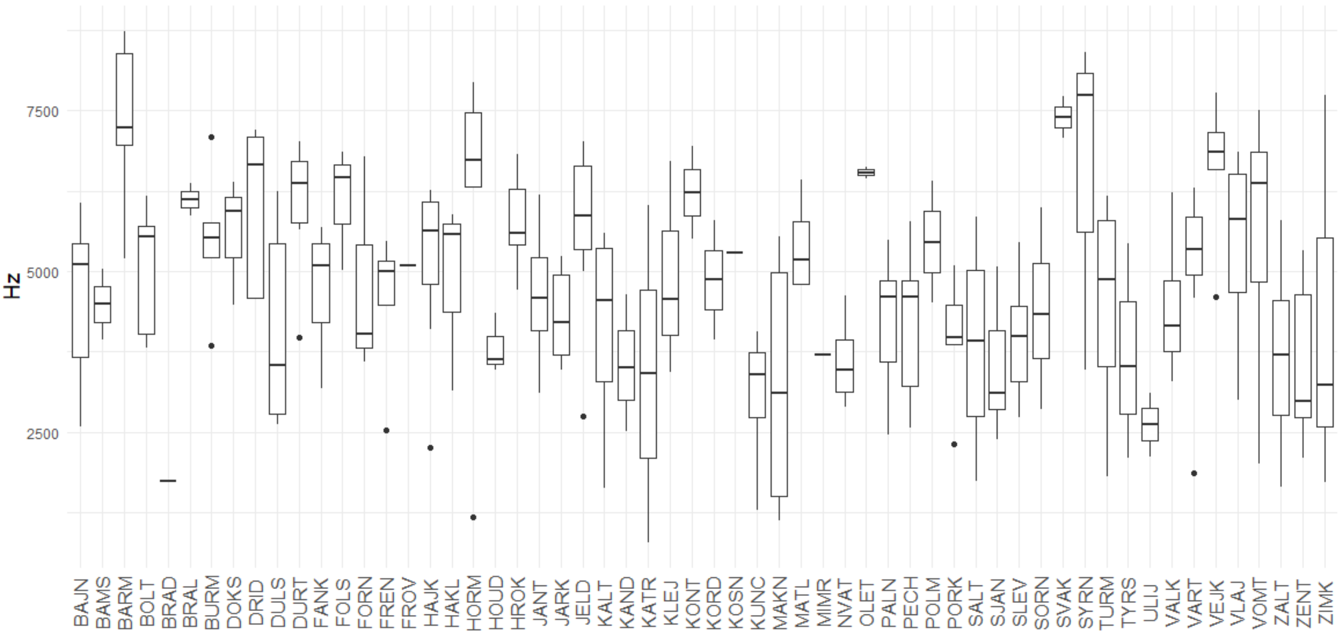


Figure 34: Mean and range of COG of /ř/ by speaker

A wide range of COG values can be observed in recordings of MAKN, VOMT, or ZIMK. The same can be said about the recording of speaker KATR, however, once again, he only had two realizations of /ř/ in his recording. Of course, this does not negate the fact that those two /ř/ realizations had to differ significantly in their COG and it still confirms that all speakers inherently do differ in the way they produce speech. However, more realizations would be necessary to make an objective conclusion about the degree of a within-speaker variety in such cases. A narrow range of COG values can be observed in speakers BURM and NVAT. In both cases, the number of realizations is higher.

Figure 35 displays the distribution of center of gravity (COG) values of fricative /x/ in all speakers individually.

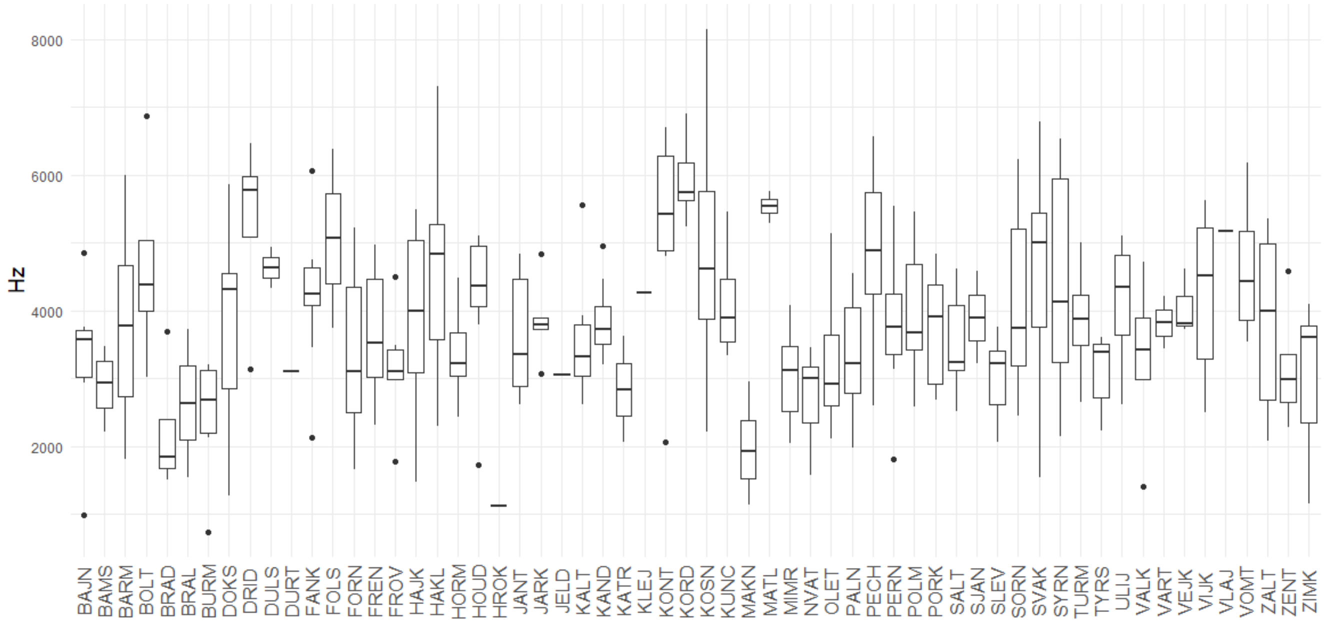


Figure 35: Mean and range of COG of /x/ by speaker

A wide range of COG values can be observed in speakers BARM, DOKS, HAJK, HAKL, KOSN, or SVAK. None of the recordings of the mentioned speakers have an abnormally low number of /x/ realizations. Conversely, a narrow range of COG values can be observed in speakers BAJN, KALT, KAND, MATL, or ZENT. Speakers who should also be mentioned are BRAD and MAKN, because their values are considered very low given the population statistics provided in section 4.1.1.4.

4.2.2. Standard deviation

The following section will focus on within-speaker variety of standard deviation in all the target fricatives separately.

Figure 36 shows the distribution of the standard deviation (SD) values of fricative /s/ in all speakers individually. A wide range of SD values can be observed in the recordings of speakers

BARM, DULS, FOLS, HORM or PERN. On the other hand, a comparably narrow range of SD values can be seen in BOLT, FROV, or OLET.

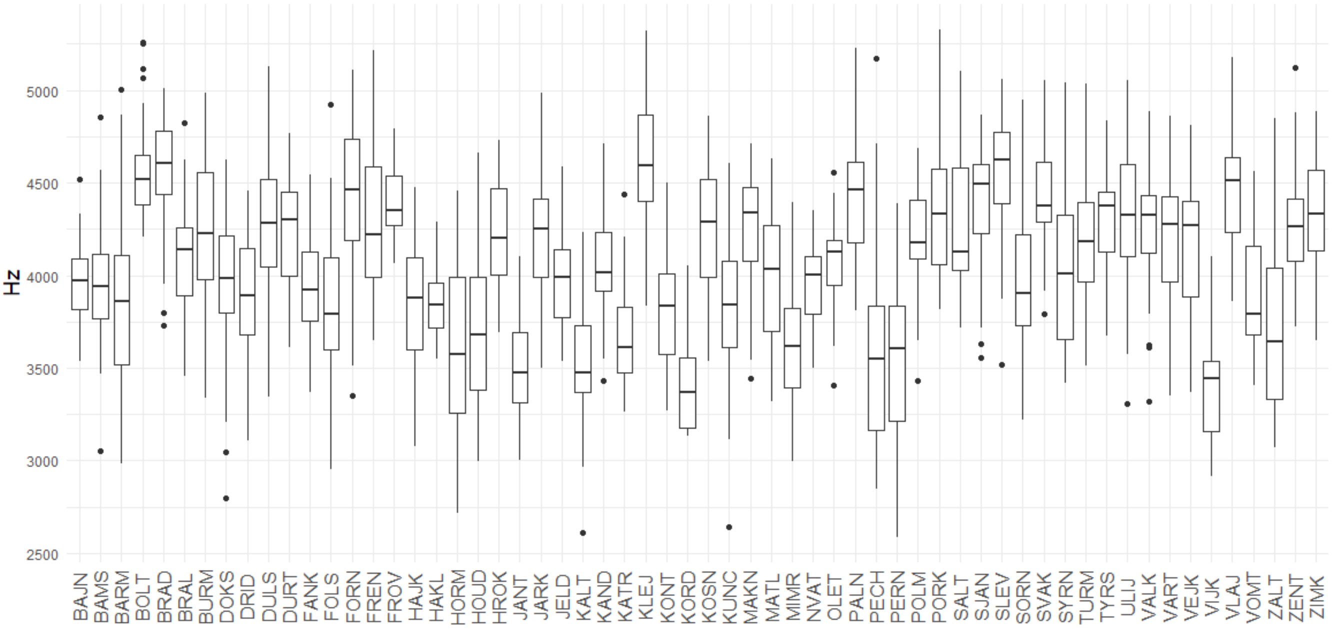


Figure 36: Mean and range of SD of /s/ by speaker

Figure 37 displays the distribution of the standard deviation (SD) values of fricative /š/ in all speakers individually.

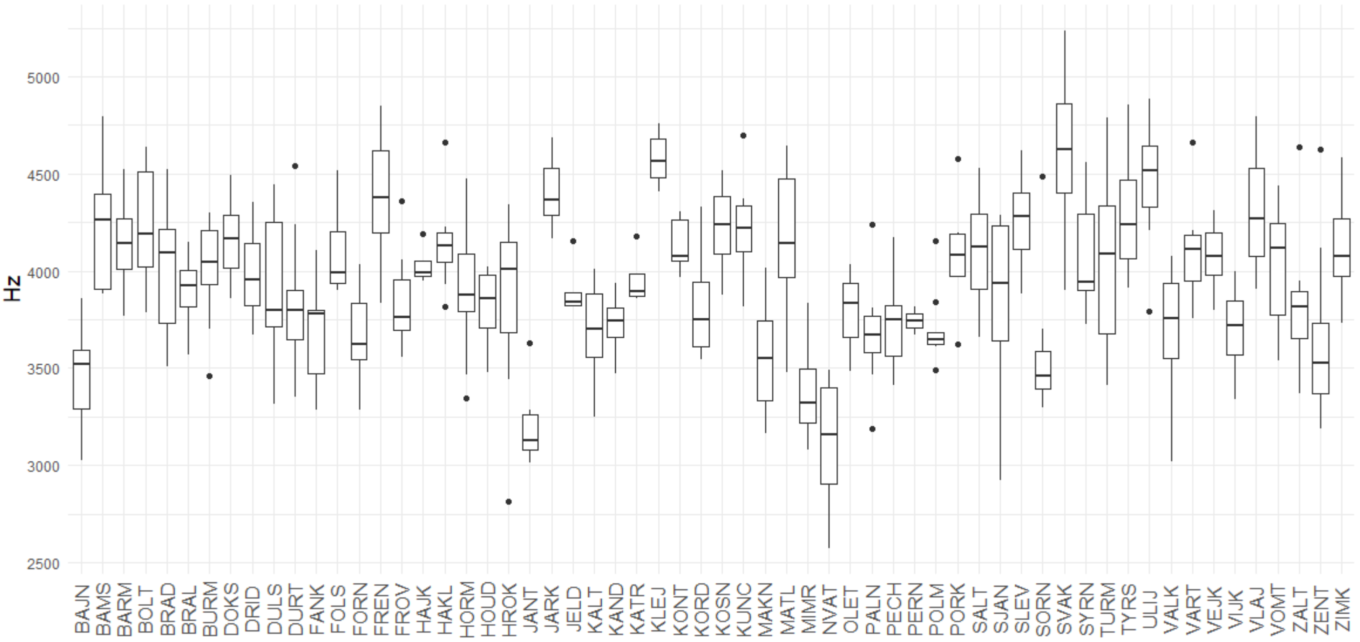


Figure 37: Mean and range of SD of /š/ by speaker

A notably wide range of SD values can be observed in SJAN or SVAK. On the other hand, a narrow range of SD values can be seen in HAJK, HAKL, JANT, JELD, POLM, or SORN. Speakers worth noting because of their unusually rare values are JANT, MIMR, and NVAT.

Figure 38 showcases the distribution of the standard deviation (SD) values of fricative /ř/ in all speakers separately.

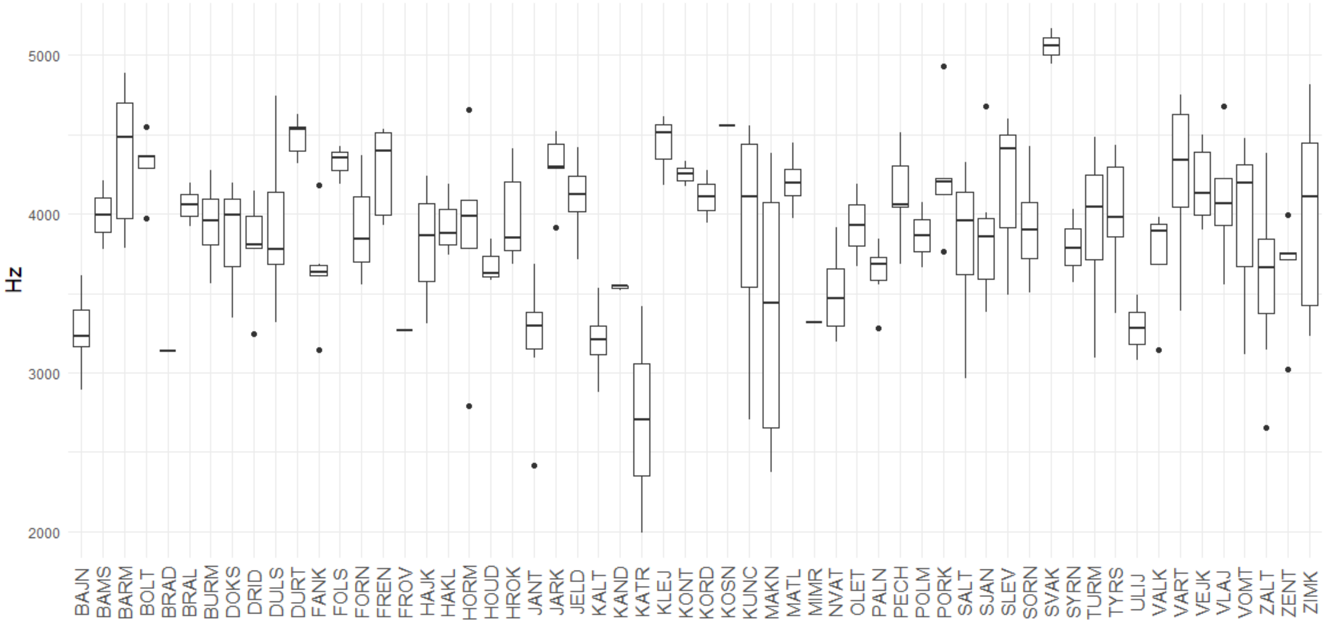


Figure 38: Mean and range of SD of /ř/ by speaker

A very noticeable wide range of SD values is seen in speakers KUNC and MAKN. Speaker KATR also displays a relatively wide range of SD values, however, as already mentioned he only had two realizations of /ř/ in his speech. Speaker SVAK shows very high values for the measured population and his range of values is very narrow, however, he as well only had two /ř/ realizations in his recording.

Lastly, Figure 39 displays the distribution of the standard deviation (SD) values of fricative /x/ in all speakers individually.

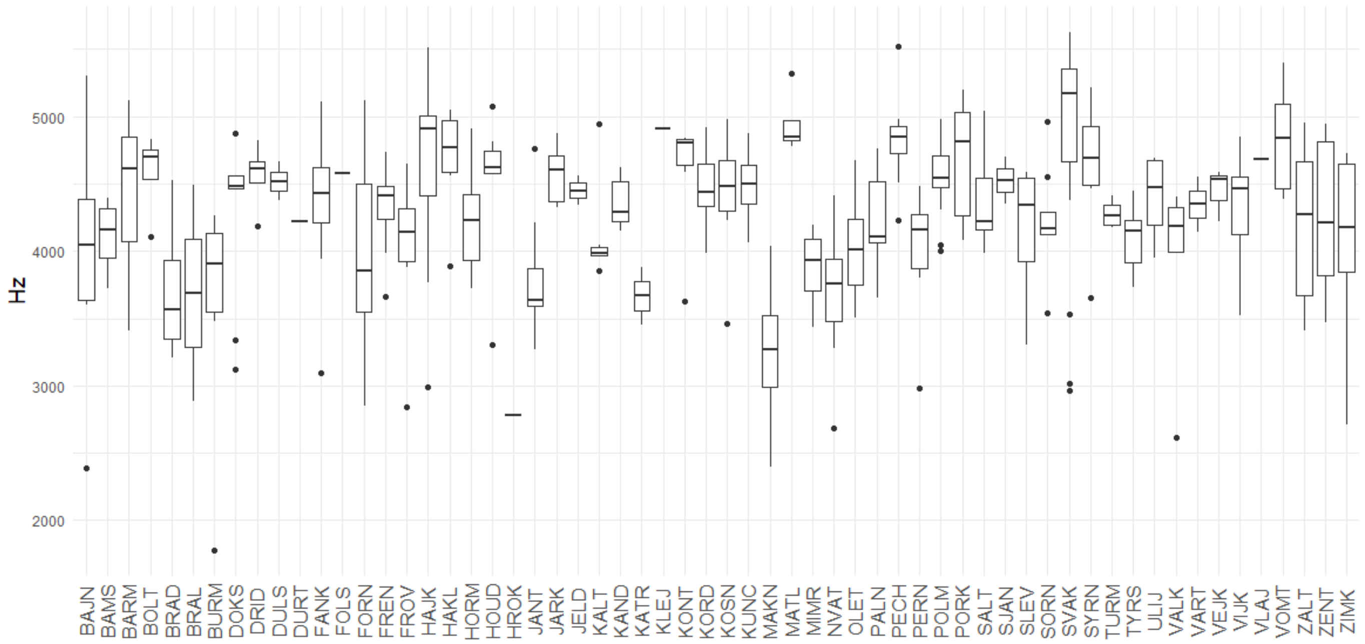


Figure 39: Mean and range of SD of /x/ by speaker

A relatively wide range of SD values can be observed in speakers FORN, HAJK, MAKN, and ZIMK. On the contrary, a low range of SD values can be seen in speakers BOLT, DRID, KALT, MATL, SORN, or TURM. Those whose range of values is narrow, but whose speech only contains a small number of the fricative /x/ realizations, were not mentioned again.

4.3 Comparison of recorded and telephone speech

The next section of this study analyses the extent of change in the spectral moments of /s/, /š/, /ř/, and /x/ under telephone-simulated conditions in comparison with the original recordings. The motivation for this comparison arises from the understanding that telephone transmission alters the acoustic properties of speech due to bandwidth limitation. Given the frequent use of telephone speech in forensic phonetics, it is important to evaluate how the spectral moments of the target fricatives are affected and whether they remain reliable as speaker-discriminating parameters.

4.3.1 Center of gravity

Figure 40 displays a box plot comparing the center of gravity (COG) values of the four target fricatives /s/, /š/, /ř/, and /x/ under three conditions - recorded speech, telephone with a narrowband codec (Telephone_NB) and telephone with a wideband codec (Telephone_WB). The x-axis shows the three data conditions. The y-axis represents the COG values measured in Hz.

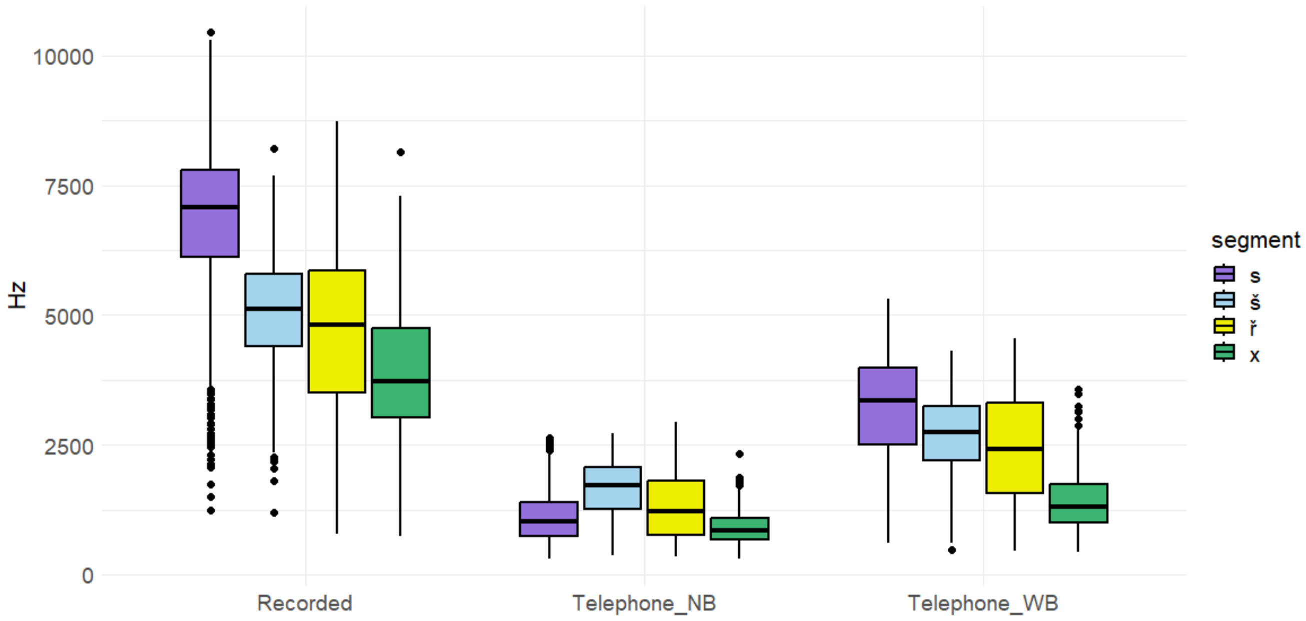


Figure 40: Comparison of center of gravity (COG) values and range of /s/, /š/, /ř/, and /x/ in recorded and telephone-simulated speech

Table 12: Median, first quartile, third quartile, and Interquartile Range (IQR) of center of gravity (COG) for each target fricative in three different conditions – recording, telephone NB, and telephone WB

Segment_condition	Median	Q1	Q3	IQR
s_rec	7073 Hz	6119 Hz	7804 Hz	1685 Hz
s_tel_NB	1021 Hz	731 Hz	1395 Hz	664 Hz
s__tel_WB	3353 Hz	2504 Hz	3981 Hz	1477 Hz
š_rec	5117 Hz	4391 Hz	5805 Hz	1414 Hz
š_tel_NB	1713 Hz	1267 Hz	2078 Hz	811 Hz
š__tel_WB	2752 Hz	2191 Hz	3253 Hz	1062 Hz
ř_rec	4815 Hz	3510 Hz	5860 Hz	2351 Hz
ř__tel_NB	1210 Hz	756 Hz	1817 Hz	1061 Hz
ř__tel_WB	2414 Hz	1568 Hz	3302 Hz	1735 Hz
x_rec	3731 Hz	3020 Hz	4750 Hz	1730 Hz
x_tel_NB	859 Hz	678 Hz	1083 Hz	405 Hz
x__tel_WB	1302 Hz	1004 Hz	1740 Hz	736 Hz

Table 12 presents a comparison of the median values, first quartiles (Q1), third quartiles (Q3), and interquartile ranges (IQR) of all the target fricatives in the three distinct conditions. The cells corresponding to each fricative and their values in the table are color-coded in the same manner as the box plots, which visually aids in the comparison and interpretation of the data.

There is a very noticeable and significant decrease in the COG values of all fricatives in the narrowband condition compared to the recorded speech. If we observe e.g. the fricative /s/, the median COG value changed from 7073 Hz to only 1021 Hz. The range of the COG values distribution also decreased significantly. The median values of /š/, /ř/, and /x/, although significantly decreased, retained their original order in the Telephone_NB condition. Specifically, the median COG of /š/ remained higher than that of /ř/, and the median COG of /ř/ remained higher than that of /x/. In contrast, /s/ is the only fricative that changed in this respect. In the original recordings, the median COG of /s/ was higher than that of /š/; however, in the Telephone_NB condition, this order is reversed, with /s/ having a lower median COG than /š/. In conclusion, the COG values of all the target fricatives decrease significantly in the telephone narrowband simulation, which indicates a substantial loss of high-frequency information due to the limited bandwidth as expected.

Fricatives in the telephone wideband condition appear to retain more high-frequency information in comparison with the narrowband condition, however, they still show a reduction in COG values. For example, the median center of gravity (COG) value of the fricative /s/ in the original recordings is 7073 Hz, whereas in the wideband telephone condition, it drops to 3353 Hz. This highlights the fact that while the wideband telephone condition offers some enhancement over the narrowband condition, the frequency range of fricatives still reduces noticeably compared to the original recordings. One more thing to note is that the median COG values for all fricatives retain their order, with /s/ having the highest median COG value, followed by /š/, /ř/, and /x/ with the lowest median COG value.

4.3.2 Standard deviation

Figure 41 displays a box plot comparing the standard deviation (SD) values of the four target fricatives /s/, /š/, /ř/, and /x/ under three conditions - recorded speech, telephone with a narrowband codec (Telephone_NB) and telephone with a wideband codec (Telephone_WB). The x-axis shows the three conditions of the recordings. The y-axis represents the SD values measured in Hz.

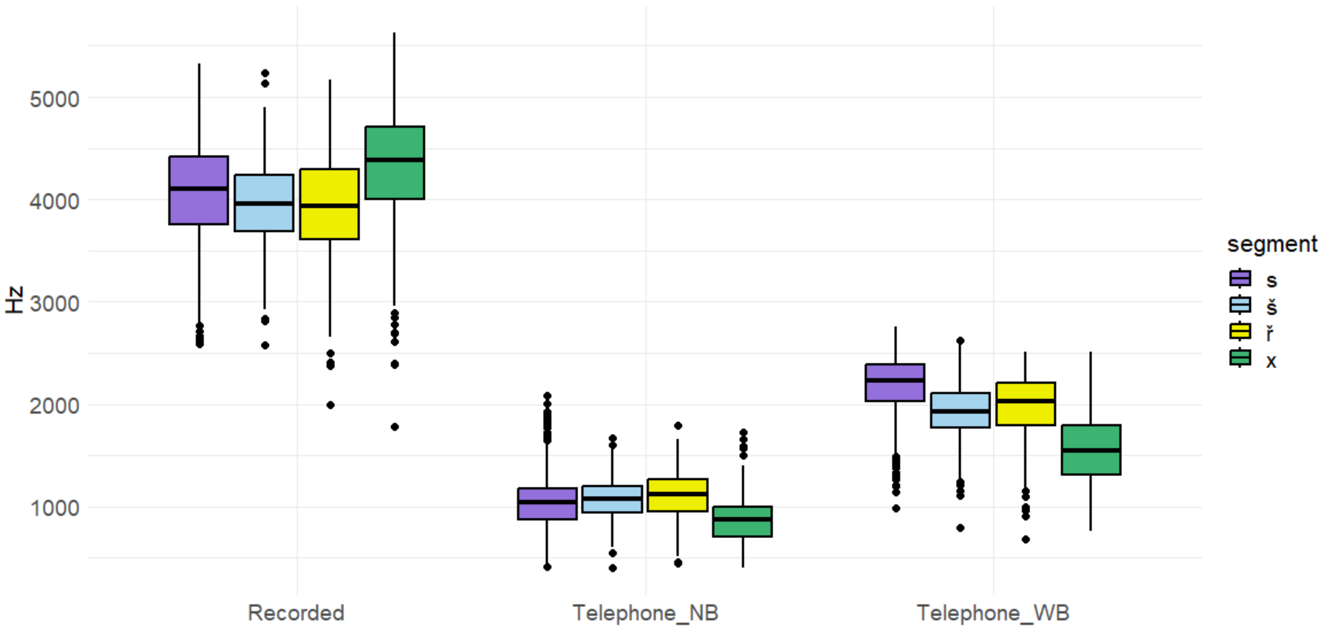


Figure 41: Comparison of standard deviation (SD) values and range of /s/, /š/, /ř/, and /x/ in recorded and telephone-simulated speech

Table 13: Median, first quartile, third quartile, and Interquartile Range (IQR) of standard deviation (SD) for each target fricative in three different conditions – recording, telephone NB, and telephone WB

Segment_condition	Median	Q1	Q3	IQR
s_rec	4101 Hz	3761 Hz	4414 Hz	653 Hz
s_tel_NB	1037 Hz	879 Hz	1182 Hz	303 Hz
s__tel_WB	2227 Hz	2029 Hz	2385 Hz	356 Hz
š_rec	3954 Hz	3688 Hz	4238 Hz	550 Hz
š_tel_NB	1074 Hz	944 Hz	1203 Hz	259 Hz
š__tel_WB	1930 Hz	1767 Hz	2104 Hz	337 Hz
ř_rec	3939 Hz	3613 Hz	4290 Hz	678 Hz
ř__tel_NB	1120 Hz	953 Hz	1262 Hz	309 Hz
ř__tel_WB	2033 Hz	1792 Hz	2204 Hz	412 Hz
x_rec	4388 Hz	4002 Hz	4709 Hz	707 Hz
x_tel_NB	870 Hz	709 Hz	997 Hz	288 Hz
x__tel_WB	1150 Hz	1307 Hz	1788 Hz	481 Hz

Table 13 again presents a comparison of the median values, first quartiles (Q1), third quartiles (Q3), and interquartile ranges (IQR) of all the target fricatives in the three distinct conditions.

The cells corresponding to each fricative and their values in the table are color-coded in the same manner as the box plots.

Similarly to COG values, there is a very noticeable and significant decrease in the SD values of all fricatives in the narrowband condition compared to the recorded speech. If we observe for example the fricative /x/ the median SD value changed from 4388 Hz to 870 Hz. The fricative /x/ also had the highest mean SD value in the original recordings out of all the fricatives, whereas in the NB condition, its mean SD value was the lowest. The fricative /ř/ has the highest median SD value in the NB condition followed by /š/, /s/, and lastly /x/. The range of SD values across all fricatives became narrower in the NB condition as well.

Once again, just like in the case of COG values, the standard deviation (SD) values of fricatives in the wideband telephone condition indicates better retention of high-frequency information compared to the narrowband condition. However, the SD values are still reduced compared to the original recordings. While the order of COG median values stayed the same in the original and WB conditions, the same is not the case for SD median values. In the recorded condition, the order of mean SD from highest to lowest is /x/, /s/, /š/, and /ř/. However, in the wideband telephone condition, the order shifts to /s/, /ř/, /š/, and /x/.

4.3.3 Skewness

Figure 42 displays a box plot comparing the skewness values of the four target fricatives /s/, /š/, /ř/, and /x/ under three conditions - recorded speech, telephone with a narrowband codec (Telephone_NB), and telephone with a wideband codec (Telephone_WB).

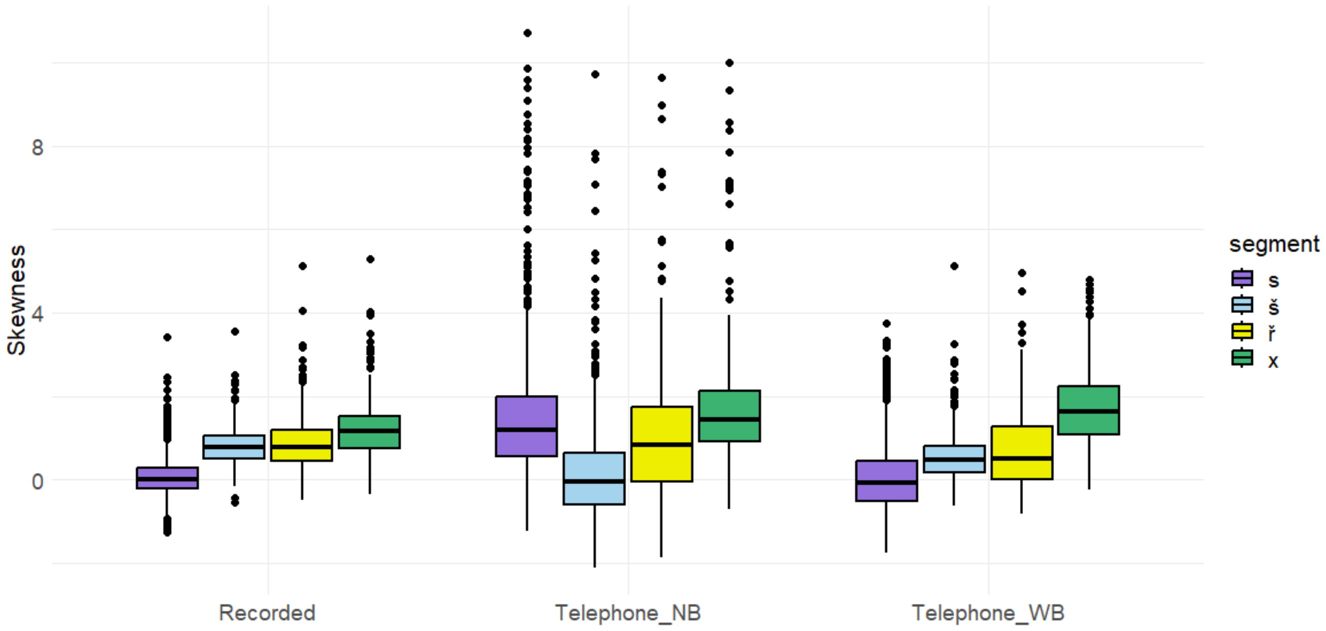


Figure 42: Comparison of skewness values and range of /s/, /š/, /ř/, and /x/ in recorded and telephone-simulated speech

Table 14: Median, first quartile, third quartile, and Interquartile Range (IQR) of skewness for each target fricative in three different conditions – recording, telephone NB, and telephone WB

Segment_condition	Median	Q1	Q3	IQR
s_rec	0.02	-0.2	0.27	0.47
s_tel_NB	1.2	0.57	2	1.43
s__tel_WB	-0.07	-0.51	0.46	0.97
š_rec	0.765	0.51	1.05	0.54
š_tel_NB	-0.05	-0.588	0.648	1.24
š__tel_WB	0.47	0.172	0.8	0.628
ř_rec	0.78	0.46	1.20	0.737
ř__tel_NB	0.84	-0.035	1.74	1.77
ř__tel_WB	0.51	0.015	1.28	1.27
x_rec	1.17	0.76	1.52	0.758
x_tel_NB	1.44	0.92	2.13	1.21
x__tel_WB	1.63	1.09	2.23	1.13

The range of skewness values is relatively narrow in the recorded condition. On the other hand, in both telephone NB and telephone WB conditions, the ranges increase and become wider.

Very noticeably, in the telephone NB condition, the outliers reach much greater skewness values than the other two conditions, suggesting a greater asymmetry in the spectrum after narrowband filtering. The order of median skewness values remains the same in the recorded and telephone WB conditions with /s/ being the lowest, followed by /š/, /ř/, and finally /x/. The order is different in the telephone NB condition with

4.3.4 Kurtosis

Figure 43 showcases a box plot comparing the kurtosis values of the four target fricatives /s/, /š/, /ř/, and /x/ under three conditions - recorded speech, telephone with a narrowband codec (Telephone_NB), and telephone with a wideband codec (Telephone_WB).

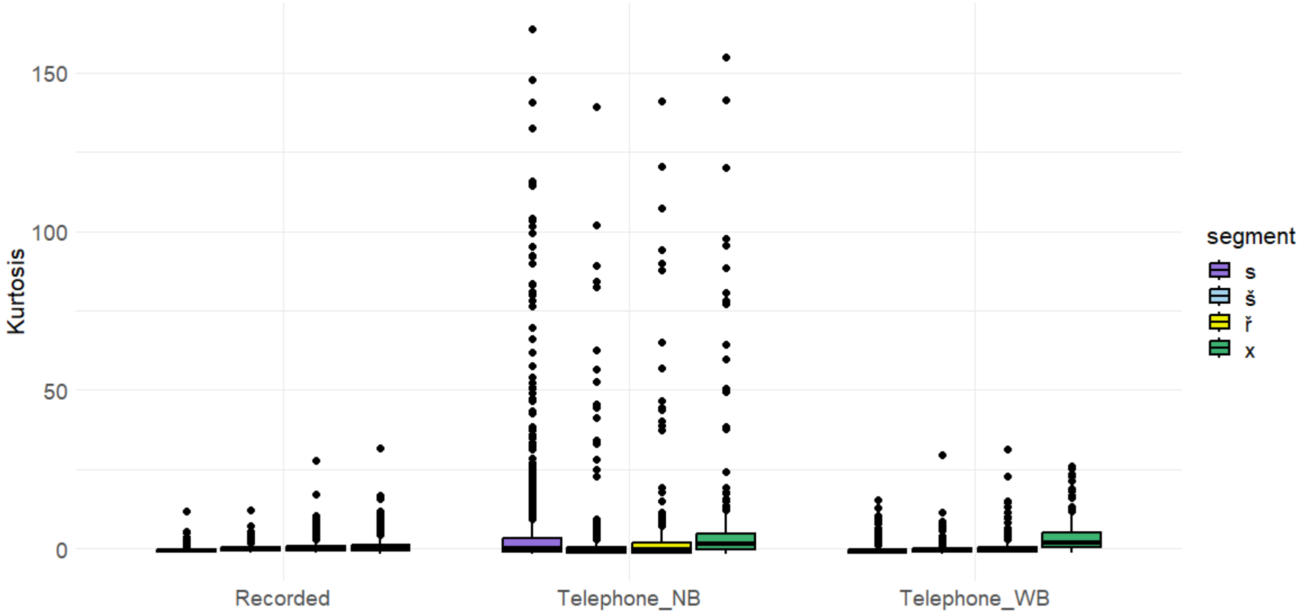


Figure 43: Comparison of kurtosis values and range of /s/, /š/, /ř/, and /x/ in recorded and telephone-simulated speech

Table 15: Median, first quartile, third quartile, and Interquartile Range (IQR) of kurtosis for each target fricative in three different conditions – recording, telephone NB, and telephone WB

Segment_condition	Median	Q1	Q3	IQR
s_rec	-0.76	-1.01	-0.42	0.59
s_tel_NB	0.165	-0.97	3.10	4.06
s__tel_WB	-1.08	-1.39	-0.44	0.95
š_rec	-0.36	-0.77	0.31	1.08
š_tel_NB	-0.825	-1.35	0.348	1.70
š__tel_WB	-0.525	-0.95	-0.07	0.88

ř_rec	-0.23	-0.76	0.712	1.47
ř__tel_NB	-0.435	-1.31	1.90	3.21
ř__tel_WB	-0.665	-1.09	0.428	1.52
x_rec	0.06	-0.827	1.12	1.95
x_tel_NB	1.4	-0.215	4.59	4.80
x__tel_WB	1.92	0.275	4.86	4.58

In all three conditions, the median kurtosis values remain relatively close to zero. The range of values is wider in the telephone NB condition in comparison with the recorded condition for all the fricatives. The range of values is narrower in the telephone WB condition in comparison with the telephone NB condition for all fricatives as well. However, the most noticeable aspect in Figure 39 is the extreme values of outliers in the telephone NB condition, some exceeding 100 or even 150. Although most of the data clusters around a low kurtosis value, the large number of outliers with high kurtosis values indicates that the spectrum under narrowband filtering can occasionally become highly peaked.

5. General discussion

As already explained in the theoretical background of this study, typicality is a crucial concept in forensic phonetics. Establishing typicality allows forensic phoneticians to create a reference for various phonetic features. It enables the use of various statistical methods to assess the probability that a given speech sample matches a particular speaker. Without typicality, the field of forensic phonetics would lack the necessary framework for making valid conclusions. However, as previously mentioned, to assess whether certain values are typical or rare, population statistics about the given characteristics must be available. The goal of this thesis was to obtain such population statistics about the four spectral moments of four Czech voiceless fricatives - /s/, /š/, /ř/, and /x/. While providing comprehensive population statistics of the spectral moments of the four segments was successful, there were some limitations. Although detailed information about medians, extreme values, quantiles, and generally the distribution of data was presented, the exact threshold between typical and rare values remains uncertain. While it is clear that values that are located e.g. below the 10th or above the 90th percentile are rare, defining a precise threshold for what is still considered typical and what is already rare proved to be beyond the scope of this thesis. Addressing this gap will require further research to establish more definitive criteria for distinguishing between typical and rare spectral moment values in all the target fricatives.

Another point in the study that proved to be slightly challenging was in interpretation of the within-speaker results. In case of fricatives /s/ and /š/ there were no problems as most of the 60 speakers had many realizations of both the fricatives in their speech. However, the same could not be said about /ř/ and /x/. For example, some speakers only had two realizations of /ř/, and the range of its e.g. COG values was very wide, which of course suggests a high within speaker variety. However, to make a conclusion about within-speaker variety with such a small number of realizations is not possible. The obvious question arises and that is what if one of the two realizations was just a mispronounced segment and it altered the median and the range of values? Overall, having at least approximately a similar number of each segment realizations for each speaker would surely be helpful. However, since the recordings were of semi-spontaneous speech and therefore each speaker was saying something else, ensuring a similar number of each segment realizations was not possible. In this case, and because of the nature of this study, the spontaneity of speech was a priority. It would of course be possible to obtain recorded speech of 60 speakers who would for example read the same text, therefore ensuring the same number of fricatives. As already mentioned in the theoretical part of this study though, speaking style may influence the values of spectral moments, as it is undeniable that people articulate differently when reading out loud while knowing they are being recorded versus speaking semi-spontaneously at home to their recording device.

Another point to mention is that a further analysis of individual speakers would be suitable. For example, recordings of speakers whose e.g. COG of /s/ was low should be examined in more detail. It would be valuable to investigate whether such low COG values result from specific articulation patterns of the speakers or if they are influenced by contextual factors, such as unusually frequent presence of neighboring rounded vowels in the recordings. Even though the spectral moments were extracted from around the midpoint of the fricative, coarticulation, especially labialization can still affect the values. Once again, this is something that would not have to be questioned in a read, controlled speech.

Lastly, the comparison of recorded and telephone speech must be mentioned. The recordings that were converted into simulated telephone speech using a narrowband codec showed significant alterations in the spectral moments in comparison with the original recordings. This was clear by observing the values of the spectral moments and the visual illustration of the distributions provided by graphs, and it partially confirmed the hypothesis of this study. These huge shifts in values point to the substantial effect of technology using narrowband codecs on spectral characteristics. The finding that such technology significantly alters the spectral moments of fricatives is of course concerning for the field of forensic phonetics. In contrast, while the telephone wideband condition also exhibited noticeable differences from the recorded condition, the changes were less pronounced. This suggests that while technology using wideband codecs does affect spectral moments, the extent of these effects is less clear. To understand these differences better, further investigation using detailed statistical testing is required. Specifically, employing mixed linear models to quantify and compare the changes in spectral moments between the original recording and the simulated telephone WB recordings should provide sufficient information. While I aimed to explore the effects of telephone speech using mixed linear models, the primary focus of this thesis was on obtaining comprehensive population statistics for the spectral moments of the fricatives. Due to the extensive analysis required to achieve this primary objective, there was limited space to fully address the complexities of telephone speech. Also, I wanted to avoid making the thesis even more complex and much longer than it already is. Extensive research, similar to one done by Christensen (2023) would be necessary.

6. Conclusion

The aim of this thesis was to provide population statistics for the spectral moments of four Czech voiceless fricatives: alveolar fricative [s], the postalveolar fricative [ʃ], the velar fricative [x], and the voiceless allophone of the famous Czech fricative trill ř, [r̝̊]. Furthermore, it aimed to determine how and to what extent the spectral moments of the four target fricatives change because of telephone transmission. The population statistics were obtained by extracting and analysing all the target fricatives from recordings of semi-spontaneous speech from 60 male speakers. The comparison of the spectral moments between the original and telephone speech was conducted by converting the original recordings into telephone simulation, using both narrowband and wideband codecs. This created two realistic conditions—one representing a poor-quality signal and the other representing a better-quality signal. It was expected that the acoustic parameters of the target fricatives would be altered significantly in the narrowband codec simulation. The acoustic parameters of the fricatives were expected to be altered in the wideband codec simulation as well but to a lesser extent.

The study was successful in providing population statistics for the spectral moments of the four target fricatives, thereby offering information for forensic phonetics. As a result, typicality can now potentially be determined more efficiently within the Czech male population. Valuable information about the spectral moments for each of the fricatives was provided, including the ranges of distributions, medians, quartiles, and within-speaker variability. It was accomplished by providing an abundant selection of graphs and tables. The hypothesis that the acoustic parameters of the target fricatives would be altered significantly in the narrowband codec simulation was supported by the findings of this study. This of course poses a challenge for forensic phonetics, as using such recordings for voice comparison would likely not be of much use. The second hypothesis could neither be fully confirmed nor entirely refuted based on the findings of this study. Although the spectral moments of all fricatives did differ in a telephone wideband condition from the original recordings, this study was unable to determine whether the difference was statistically significant. Analysis using mixed linear models was not performed in order to avoid increasing the length and complexity of an already extensive study. Future research should aim to address this gap.

7. Reference list

Barrett, J. (2012). *The Correlation Between Spectral Moment Measures and Electropalatometric Contact Patterns for /t/ and /k/.* [Master's thesis, Brigham Young University].

Boersma, P., & Weenink, D. (2024). *Praat: Doing phonetics by computer* (Version 6.3.23) [Computer software]. http://www.praat.org/

Christensen, K. V. (2023). *Understanding the acoustic implications of digital transmission on fricatives* [Doctoral dissertation, University of York].

Fecher, N. (2011). *Spectral properties of fricatives:a forensic approach.* Workshop on Experimental Linguistics, Paris.

Fecher, N., & Watt, D. (2011). *Speaking under cover: The effect of face-concealing garments on spectral properties of fricatives.* International Congress of Phonetic Sciences, Hong Kong.

Harrington, J. (2010). *Phonetic analysis of speech corpora*. Wiley-Blackwell.

Hollien, H. (2012). About forensic phonetics. *Linguistica*, *52*(1), 27-53. https://doi.org/10.4312/linguistica.52.1.27-53

Isačenko, A. V. (2013). O akustice české hlásky ř. *Linguistica Online, 1-6. https://www.phil.muni.cz/linguistica/art/issues/issue-015.pdf*

Jannedy, S., & Weirich, M. (2017). Spectral moments vs discrete cosine transformation coefficients: Evaluation of acoustic measures distinguishing two merging German fricatives. *The Journal of the Acoustical Society of America*, *142*(1), 395. https://doi.org/10.1121/1.4991347

Jessen, M. (2008). Forensic phonetics. *Language and Linguistics Compass, 2*(4), 671-711. https://doi.org/10.1111/j.1749-818X.2008.00066.x

Jongman, A., Wayland, R., & Wong, S. (2000). Acoustic characteristics of English fricatives. *Journal of the Acoustical Society of America, 108*(3), 1252–1263. https://doi.org/10.1121/1.1288413

Kavanagh, C. M. (2012). *New consonantal acoustic parameters for forensic speaker comparison* [Doctoral dissertation, University of York].

Konsonanty. (2016). Fonetický ústav. https://fonetika.ff.cuni.cz/wp-content/uploads/sites/104/2016/06/6_konsonanty.pdf

Laver, J. (1980). *The phonetic description of voice quality*. Cambridge University Press.

Mohapatra, D., Fleischer, M., Zappi, V., Birkholz, P., & Fels, S. (2022) Three-dimensional finite-difference time-domain acoustic analysis of simplified vocal tract shapes. Proc. Interspeech 2022, 764-768, doi: 10.21437/Interspeech.2022-10649

Nittrouer S. (1995). Children learn separate aspects of speech production at different rates: evidence from spectral moments. *The Journal of the Acoustical Society of America*, *97*(1), 520–530. https://doi.org/10.1121/1.412278

Nolan, F. (1991). Forensic phonetics. *Journal of Linguistics*, *27*(2), 483–493. doi:10.1017/S0022226700012755

R Core Team (2023). *R (4.3.1): A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org

Rose, P. (2002). Forensic speaker identification. Taylor & Francis.

Schindler, C., & Draxler, C. (2013). Using spectral moments as a speaker-specific feature in nasals and fricatives. In *Proceedings of Interspeech 2013* (pp. 2793-2796). https://doi.org/10.21437/Interspeech.2013-639

Shadle, C.H., & Mair, S.J. (1996). Quantifying spectral characteristics of fricatives. *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96, 3*, 1521-1524 vol.3.

Skarnitzl, R., Šturm, P., & Volín, J. (2016). *Zvuková báze řečové komunikace Fonetický a fonologický popis řeč.* Karolinum.

Skarnitzl, R., & Nechanský, M. (2025). Segmental cues. In: McDougall, K., Nolan, F. & Hudson, T. (Eds.), *Oxford Handbook of Forensic Phonetics*. Oxford University Press.

Smorenburg, L., & Heeren, W. (2020). The distribution of speaker information in Dutch fricatives /s/ and /x/ from telephone dialogues. *The Journal of the Acoustical Society of America*, *147*(2), 949. https://doi.org/10.1121/10.0000674

Strevens, P. (1960). Spectra of Fricative Noise in Human Speech. *Language and Speech*, *3*(1), 32-49. https://doi.org/10.1177/002383096000300105

Wickham, H., & Bryan, J. (2019). *readxl: Read Excel files*. R package version 1.3.1. https://CRAN.R-project.org/package=readxl

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., ... Yutani, H. (2019). *Welcome to the tidyverse*. Journal of Open Source Software, 4(43), 1686. https://doi.org/10.21105/joss.01686

Zerby, J. (2024, March 7). *What Are VoIP Codecs & How Do They Affect Call Sound Quality?* Nextiva Blog. https://www.nextiva.com/blog/voip-codecs.html