Linguistic and speaker variation in Russian fricatives

Natalja Ulrich

N° d'ordre NNT : 2022LYO20031

THÈSE de DOCTORAT DE L'UNIVERSITÉ LUMIÈRE LYON 2

École Doctorale : ED 484

Lettres, Langues, Linguistique et Arts

Discipline : Sciences du langage

Soutenue publiquement le 14 décembre 2022, par :

Natalja ULRICH

Linguistic and speaker variation in Russian fricatives.

Devant le jury composé de :

Christine MEUNIER, Directrice de recherche CNRS, Aix-Marseille Université, Présidente

Emmanuel FERRAGNE, Maître de conférences HDR, Université Paris Diderot, Rapporteur

Jalaleddin AL-TAMIMI, Maître de conférences, Université Paris Cité, Rapporteur

Martine GRICE, Professor, Université de Cologne, Examinatrice

Marc ALLASSONNIÈRE-TANG, Chargé de recherche CNRS, CNRS-MNHN Université de Paris, Co-Directeur de thèse

François PELLEGRINO, Directeur de recherche CNRS, CNRS, Co-Directeur de thèse

Contrat de diffusion

N° National de Thèse : XXX

THESE de DOCTORAT DE L'UNIVERSITÉ DE LYON

Opérée au sein de

L'UNIVERSITÉ LUMIÈRE LYON 2

Discipline : **Phonétique**

Laboratoire Dynamique Du Langage

ED 484 Lettres, Langues, Linguistique et Arts

Présentée et soutenue publiquement le 14 décembre 2022
par **Natalja Ulrich**

Linguistic and speaker variation in Russian fricatives

Directeurs/Directrices de Thèse : François Pelligrino, Marc Allassonnière-Tang

Devant la commission d'examen formée de :

M.	François Pellegrino	*CNRS*	CoDirecteur
M.	Marc Allassonnière-Tang	*CNRS-MNHN-Université de Paris*	CoDirecteur
Mme.	Martine Grice	*University of Cologne*	Examinatrice
Mme.	Christine Meunier	*Aix-Marseille Université*	Examinatrice
M.	Jalaleddin Al-Tamimi	*Université Paris Cité*	Rapporteur
M.	Emmanuel Ferragne	*Université Paris Cité*	Rapporteur

L'UNIVERSITÉ LUMIÈRE LYON 2
École doctorale ED 484 Lettres, Langues, Linguistique et Arts
Laboratoire Dynamique Du Langage

THESE

Opérée au sein de

DOCTORAT DE L'UNIVERSITÉ DE LYON 2

Discipline : Phonetics

Linguistic and speaker variation in Russian fricatives

par **Natalja Ulrich**

Directeurs/Directrices de Thèse : François Pelligrino, Marc Allassonnière-Tang

Présentée et soutenue publiquement le 14 décembre 2022

Devant la commission d'examen formée de :

M.	François Pellegrino	*CNRS*	CoDirecteur
M.	Marc Allassonnière-Tang	*CNRS-MNHN-Université de Paris*	CoDirecteur
Mme.	Martine Grice	*University of Cologne*	Examinatrice
Mme.	Christine Meunier	*Aix-Marseille Université*	Examinatrice
M.	Jalaleddin Al-Tamimi	*Université Paris Cité*	Rapporteur
M.	Emmanuel Ferragne	*Université Paris Cité*	Rapporteur

Acknowledgements

Writing a thesis is always challenging, especially in such chaotic times. The endless months of home office and unexpected difficulties made it an even bigger challenge than initially thought.

I want to thank all my colleagues, friends and family, who were supporting me and helped me reach my goals.

First of all, I am very grateful to my supervisors François Pellegrino and Marc Allassonnière-Tang. Thank you François P. for all your understanding, pragmatism and diplomacy. You were always supporting me and were available when it was needed. I am also extremely thankful to Marc A-T. who was not just an amazing supervisor, but also a true friend. I want to thank you for all your patience and all the hours you spent listening to me, you never gave up on me and were always pushing me forward with new ideas and suggestions.

Furthermore, I would like to thank Rémi Anselme. Rémi you are one of the most positive and inspiring people I have ever met. You helped me with so many aspects of organising my life in Lyon and solved numerous problems. Thank you very much for sharing with me your coffees, food, and flats and all the unmaterial things. It was a real pleasure to spend time with you.

Even if my time at the DDL was short, I would like to thank my colleagues and in particular, the PhD students, who made my stays in Lyon very enjoyable.

I wish also to thank our participants, the Phonetic Lab in St. Petersburg (and the sound engineer Tatiana Chukaeva in particular), the University of Zürich for financial support, technical support and help with the design of the experiment (to Volker Dellwo in particular). I am also grateful for the financial support of the University of Lyon and the French National Research Agency.

Finally, I would like to thank all the other people who contributed to the successful completion of my thesis, with all the ideas, discussions and inputs.

Abstract

This thesis represents an acoustic-phonetic investigation of phonetic details in Russian fricatives. The main aim was to detect acoustic correlates that carry linguistic and idiosyncratic information. The questions addressed were whether the place of articulation, speakers' gender and ID can be predicted by a set of acoustic cues and which acoustic measures represent the most reliable indicators. Furthermore, the distribution of speaker-specific characteristics and inter- and intra-speaker variation across acoustic cues were studied in more detail.

The project started with the generation of a large audio database of Russian fricatives. Then, two follow-up analyses were conducted. Acoustic recordings were collected from 59 native Russian speakers. The resulting dataset consists of 22,561 tokens including the fricatives [f], [s], [ʃ], [x], [v], [z], [ʒ], [sʲ], [ɕ], [vʲ], [zʲ].

The first study employed a data sample of 6320 tokens (from 40 speakers). Temporal and spectral measurements were extracted using three acoustic cue extraction techniques (full sound, the noise part, and the middle 30ms windows). Furthermore, 13 Mel Frequency Cepstral Coefficients were computed from the middle 30ms window. Classifiers based on single decision trees, random forests, support vector machines, and neural networks were trained and tested to distinguish between the three non-palatalized fricatives [f], [s] and [ʃ]. The results demonstrate that machine learning techniques are very successful at classifying the Russian voiceless non-palatalized fricatives [f], [s] and [ʃ] by using the centre of gravity and the spectral spread irrespective of contextual and speaker variation. The three acoustic cue extraction techniques performed similarly in terms of classification accuracy (93% and 99%), but the spectral measurements extracted from the noise parts resulted in slightly better accuracy. Furthermore, Mel Frequency Cepstral Coefficients show marginally higher predictive power over spectral cues ($<$ 2%). This suggests that both spectral measures and Mel Frequency Cepstral provide sufficient information for the classification of these fricatives and their choice depends on the particular research question or application.

The second study's dataset consists of 15812 tokens (59 speakers) that contain [f], [s], [s[j]], [ɕ], [x], [v], [z], [ʒ], [s[s]]. As in the first study, two types of acoustic cues were extracted including 11 acoustic speech features (spectral cues, duration and HNR measures) and 13 Mel Frequency Cepstral Coefficients. Classifiers based on single decision trees and random forests were trained and tested to predict speakers' gender and ID. Additional statistical methods were applied to understand the distribution of gender and speaker information across different fricatives and acoustic cues. The output shows gender and speaker characteristics in the acoustics of voiceless, voiced and palatal fricatives. Gender can be predicted with a good performance by both acoustic speech features (72%) and Mel Frequency Cepstral Coefficients (88%), whereby Mel Frequency Cepstral Coefficients clearly outperform acoustic speech features. Speakers' ID can only be predicted by Mel Frequency Cepstral Coefficients with a moderate performance of 64%. Acoustic speech features encoded speakers' idiosyncrasy in fricative sounds in a highly individual manner, and no set of cues can predict those idiosyncrasies. The findings suggest that Mel Frequency Cepstral Coefficients capture better speaker's idiosyncrasies than common acoustic speech features.

In conclusion, Russian fricatives contain linguistic and idiosyncratic information which can be extracted by Mel Frequency Cepstral Coefficients and acoustic speech features. A detailed exploration of several spectral, temporal, amplitude and harmonics-to-noise ratio measures suggests that the spectral cues are sufficient to distinguish between the place of articulation in voiceless fricatives. Acoustic differences between female and male speakers are also observed in

the spectrum, duration and harmonics-to-noise ratio measures in most of the investigated fricatives. In regard to the inter-and intra- speaker variation, no clear patterns could be identified across the eight fricatives and acoustic cues.

Résumé (French)

Cette thèse présente une investigation acoustico-phonétique des détails phonétiques des fricatives russes. L'objectif principal était de détecter des corrélats acoustiques porteurs d'informations linguistiques et idiosyncrasiques. Les questions abordées étaient de savoir si le lieu d'articulation, le sexe du locuteur ou son identité peuvent être prédits par des indices acoustiques et quelles mesures acoustiques représentent les indicateurs les plus fiables. En outre, la distribution des caractéristiques spécifiques au locuteur et à la variation inter et intra locuteur à travers les indices acoustiques a été étudiée plus en détail. Le projet a commencé par la création d'une grande base de données audio des fricatives russes. Des enregistrements acoustiques ont été obtenus auprès de 59 locuteurs russes natifs.

Le jeu de données résultant est composé de 22 561 occurrences comprenant les fricatives [f], [s], [ʃ], [x], [v], [z], [ʒ], [sʲ], [ɕ], [vʲ], [zʲ]. Deux analyses ont été menées à partir de cette base de données. Dans la première étude, un échantillon de données de 6320 occurrences (40 locuteurs) a été utilisé. Trois techniques d'extraction acoustisque (à partir du son complet, de la durée du bruit et des fenêtres centrales de 30 ms) ont été sollicitées pour extraire des mesures temporelles et spectrales. En outre, 13 coefficients cepstraux (Mel-Frequency Cepstral Coefficients, MFCC) ont été calculés à partir de la fenêtre centrale de 30 ms. Des classificateurs fondés sur des arbres de décision simples, des forêts aléatoires, des machines à vecteurs de support (Support-vector machine, SVM) et des réseaux neuronaux ont été entraînés et testés pour distinguer trois fricatives non palatalisées [f], [s] et [ʃ]. Les résultats montrent que les techniques d'apprentissage automatique réussissent très bien à classer les fricatives non voisées non-palatalisées russes [f], [s] et [ʃ] en utilisant le centre de gravité et la propagation spectrale, indépendamment des variations contextuelles et de celles du locuteur. Les trois techniques d'extraction d'indices acoustiques ont donné des résultats similaires en termes de précision (accuracy) dans la classification (93% et 99%), mais les mesures spectrales extraites de la durée totale du bruit de la frication ont donné une précision (accuracy) nettement supérieure. En outre, les coefficients cepstraux (MFCC) présentent un pouvoir prédictif légèrement supérieur à celui des indices spectraux ($< 2\%$). Cela suggère que les deux mesures spectrales et les coefficients cepstraux fournissent des informations suffisantes pour la classification de ces fricatives et que leur choix dépend de la question de recherche ou de l'apConcerningrticulière.

Dans la deuxième étude, 15812 occurrences (59 locuteurs) de huit fricatives russes ([f], [s], [ʃ], [x], [v], [z], [ʒ], [sʲ], [ɕ]) ont été analysés. Comme dans la première étude, deux types d'indices acoustiques ont été sélectionnés. Tout d'abord, 11 caractéristiques acoustiques de la parole comprenant des indices spectraux, des mesures de durée et de HNR ont été extraits, suivis de 13 coefficients cepstraux (MFCC). Des classificateurs fondés sur des arbres de décision simples et des forêts aléatoires ont été entraînés et testés pour prédire le sexe et l'identité des locuteurs. Des méthodes statistiques supplémentaires ont été appliquées pour comprendre la distribution des informations sur le genre et le locuteur à travers différentes fricatives et indices acoustiques. L'étude montre que les fricatives non voisées, voisées et palatales contiennent des informations spécifiques au genre et aux locuteurs. Les résultats montrent que le sexe du locuteur peut être prédit avec une bonne précision à la fois par les caractéristiques acoustiques de la parole (72%) et par les coefficients cepstraux (88%), les coefficients cepstraux étant nettement plus performants que les caractéristiques acoustiques de la parole. L'identité des locuteurs ne peut être prédite que par les coefficients cepstraux (64%). Les caractéristiques acoustiques de la parole ont encodé les singularités des locuteurs dans les sons fricatifs d'une manière très individuelle, et aucun ensemble d'indices ne peut prédire ces singularités. L'étude montre donc

que les coefficients cepstraux fournissent de meilleures informations sur le locuteur que les caractéristiques acoustiques courantes de la parole.

En conclusion, les fricatives russes contiennent des informations linguistiques et idiosyncratiques qui peuvent être extraites par les coefficients cepstraux et les caractéristiques acoustiques de la parole. Une exploration détaillée d'un certain nombre de mesures spectrales, temporelles, d'amplitude et de rapport harmoniques/bruit, suggère que, premièrement, les indices spectraux sont suffisants pour distinguer le lieu d'articulation des fricatives sans voix. Deuxièmement, des différences acoustiques entre les locuteurs femmes et hommes sont observées dans les mesures du spectre, de la durée et du rapport harmoniques/bruit pour la plupart des fricatives étudiées. Troisièmement, aucun modèle clair de variation inter et intra locuteur n'a pu être identifié parmi les huit fricatives et les indices acoustiques.

LIST OF PAPERS

1. Ulrich N. (2022), Database description: Russian fricatives (Submitted to JASA as a letter to Editors on the 14th of October 2022).

2. Ulrich, N., Allassonnière-Tang, M., Pellegrino, F., Dediu, D. (2021). Identifying the Russian voiceless non-palatalized fricatives [f], [s], and [ʃ] from acoustic cues using machine learning. The Journal of the Acoustical Society of America, 150(3), 1806-1820.

3. Ulrich, N., Allassonnière-Tang, M., Pellegrino, F., Inter- and intra- speaker variation in eight Russian fricatives. (Submitted to JASA, revisions due to the 6th of December 2022)

Contents

Chapter 1

Introduction

The perception of human speech involves the processing of linguistic content as well as the idiosyncratic characteristics of speakers. In linguistic and ASR (Automatic Speech Recognition) research, it is agreed that this dual coding is conveyed by acoustic-phonetic details in speech signals and that it can be measured by technological applications and evaluated by listeners (e.g. Blumstein and Stevens, 1981; Dellwo et al., 2007; He and Dellwo, 2014). On the one hand, these acoustic-phonetic details are assumed to encompass certain features that remain stable in different contexts and speaking conditions (Blumstein and Stevens, 1981) which allow the understanding of speech regardless of whether it is spoken aloud, softly, or whispered. On the other hand, research on idiosyncratic aspects of speech sounds has demonstrated that speakers can be recognised and distinguished according to specific acoustic characteristics (Dellwo et al., 2007; He and Dellwo, 2014). Thereby, various observations indicate that the acoustic cues extracted from speech sounds are not equally informative for all speakers (Kavanagh, 2012). And while some speakers could be recognised at a high rate, other speakers showed a poor performance (Gendrot et al., 2019). However, it has been reported that the speaker discrimination potential depends on a series of factors including speakers' properties, the communication situation and the ASR system used (Bonastre et al., 2015).

In acoustic-phonetic research, the understanding of how these different types of information are encoded in speech sounds and how they can be extracted is strongly influenced by the distribution of periodic and aperiodic energy across different sound categories. As a result, linguistic and speaker-specific aspects are better understood in sound categories consisting predominantly of periodic energy, such as vowels, than in sound categories such as fricatives containing aperiodic components.

Fricative sounds can consist either of only aperiodic energy, as in the case of voiceless fricatives, or of the interaction of periodic and aperiodic components, as in voiced fricatives. Existing linguistic research on fricatives is mostly focused on the place of articulation in voiceless fricatives. In the acoustic analysis of fricatives several temporal, amplitude, spectral and other measures were obtained and compared. The peak frequency and spectral moments have been reported as the most crucial acoustic cues to distinguish the place of articulation (e.g. Forrest et al., 1988; Hughes and Halle, 1956). Other investigations suggest that while the spectral moments carry important insights about fricatives, they cannot reliably distinguish the places of articulation of these fricatives (Shadle and Mair, 1996). Several surveys even argue that there is no set of properties that characterizes all fricatives, and that only a distinction between the sibilants and non-sibilants can be made (Ladefoged and Maddieson, 1996). Nevertheless,

a number of researchers have tried to define invariant acoustic cues in fricatives and confirmed the importance of the spectral domain for the identification of the place of articulation. Besides certain inconsistencies in defining the most crucial acoustic features, these analyses also show a main effect of the vowel context (Mann and Repp, 1980; Nirgianaki, 2014; Soli, 1981; Stevens, 1998), speaker and gender (e.g. Ghaffarvand Mokari and Mahdinezhad Sardhaei, 2020; Gordon et al., 2002; Hughes and Halle, 1956; Jongman et al., 2000; Kochetov, 2017; Nirgianaki, 2014) in the spectral domain. These and several other findings indicate that the spectral domain also carries speaker-specif characteristics (e.g. Newman et al., 2001; Schindler and Draxler, 2013; Smorenburg and Heeren, 2020). Generally, it is still debated to what extent fricative sounds exhibit speaker-specific acoustics. Several articles in forensic speaker comparison or speaker recognition, however, have concluded that fricatives carry enough idiosyncrasies to categorize and recognise speakers (Antal, 2008), suggesting their further exploration (Kavanagh, 2012; Schindler and Draxler, 2013). According to other investigations, fricatives are one of the sound categories that encode fewer speaker characteristics (Gendrot et al., 2020). In addition to these controversies, the acoustic analysis of fricatives has several limitations. The majority of phonetic studies focus on English voiceless fricatives, and only a few have been conducted in other languages. Most of those investigations considered a small set of fricatives with a focus on the voiceless sibilant fricatives. Also, the number of speakers rarely exceeds 10 speakers of the same gender.

These studies, however, show conflicting results as to whether the spectral domain provides enough linguistic details for distinguishing the place of articulation. Furthermore, it is unclear how stable these measures are across different sounds, speakers, and contexts.

The present dissertation aims to address some of these inconsistencies and limitations. The main goal is to provide a phonetic-acoustic description of Russian fricatives and to understand how linguistic and speaker information is encoded in Russian fricatives. In the first step of the current project, a large database of Russian fricatives was generated. Two follow-up experiments were then conducted. The first study investigated which acoustic features can distinguish the place of articulation of three voiceless fricatives. In the second study, the intra- and inter-speaker variation in eight fricatives was assessed. For these studies, two sets of measures were extracted. The first set represents common acoustic speech features (ASFs), including measures such as peak frequency, spectral moments, duration, amplitude and HNR (Harmonic to Noise Ratio) measures. The second set of measurements consists of 13 MFCCs (Mel-frequency cepstrum coefficients).

1.1 Research questions and Contributions

To approach the goals outlined in the introduction, the following research questions are asked.

- 1. How can machine learning techniques contribute to automatizing and standardising the segmentation of noise duration in voiceless fricatives?
- 2. What is the effect of window length (the entire sound, a fixed-duration window in the middle of the sound, or only the noise part) in extracting acoustic cues from voiceless fricatives?
- 3. Can the Russian fricatives [f], [s] and [ʃ] be correctly classified by a set of acoustic cues? And how does the performance of the models differ between using ASFs or/and

MFCCs?

- 4. Can speakers' gender be predicted by acoustic cues? And how does the performance of the models differ between using ASFs or/and MFCCs?
- 5. Can speakers' ID be predicted by acoustic cues? And how does the performance of the models differ between using ASFs or/and MFCCs?
- 6. How do speakers differ in their acoustic characteristics on the individual level?

The main contributions are:

- 1. Showing that machine learning can advance the automatic fricative noise segmentation in voiceless fricatives by the separation of the voiced and unvoiced parts.
- 2. Demonstrating that the acoustic extraction techniques applied to obtain acoustic measures have only a marginal influence on the prediction of the place of articulation, but that the entire noise duration gives the best results.
- 3. Providing evidence that the place of articulation can be predicted by machine learning using both sets of acoustic cues (ASFs and MFCCs). The results suggest that the mean distribution of energy *cog* and the spectral spread *sdev* of the Russian fricatives [f], [s], and [ʃ] can be reliably distinguished.
- 4. Confirming that speakers' gender can be predicted by ASFs and MFCCs and giving a description of gender variation in common ASFs as the peak frequency, spectral moments, duration and HNR measures.
- 5. Showing that in Russian fricatives speakers' ID can only be predicted by MFCCs but not by ASFs in the current database. Providing insights into the complexity of inter- and intra-speaker variation across the ASFs may explain the disability to identify speakers by these cues.
- 6. Exploration of the distribution of speaker information encoded across the ASFs in two sounds of three different speakers.

1.2 Outline of the Thesis

The dissertation is structured as follows:

Chapter 2 reviews the investigations conducted on fricative sounds in different fields. The section starts with a typological overview of the distribution of fricative inventories across a large set of languages. The basic mechanisms involved in fricative production are then described, followed by a survey of acoustic studies and their results on linguistic and speaker characteristics in fricatives. Then some challenges in understanding the acoustics of fricatives are outlined. Finally, the methods applied in the analysis of fricative sounds are addressed. The section finishes with an overview of the Russian consonant inventory, some phonotactic rules, and a survey of studies dealing with Russian fricatives.

Chapter 3 provides a description of the database, the data retrieval process, the experimental design, the resulting data files and possible reuse options The section is based on the data paper:

Ulrich N. (2022), Database description: Russian fricatives. The description is submitted to the Journal of the Acoustical Society of America as a letter to Editors on the 14th of October 2022.

In chapter 4 the segmentation and the analysis of three voiceless fricatives are themed and the first three research questions of the thesis are addressed. This section is represented by publications: Ulrich, N., Allassonnière-Tang, M., Pellegrino, F., Dediu, D. (2021). Identifying the Russian voiceless non-palatalized fricatives [f], [s], and [ʃ] from acoustic cues using machine learning. The Journal of the Acoustical Society of America, 150(3), 1806-1820.

Chapter 5 deals with the speaker information coded in fricative sounds. In this section, the last three research questions of the thesis are investigated. The chapter is based on the article: Ulrich, N., Allassonnière-Tang, M., Pellegrino, F., Inter- and intra- speaker variation in eight Russian fricatives. (Revisions due to the 6th of December 2022)

The last chapter 6 summarises the results and the contribution of the thesis. The output of the current project is compared with the research questions. The chapter ends with an outline of the limitations of the investigation and an outlook on further research.

Chapter 2

Background

Research in acoustic-phonetic, speech perception, forensic speaker comparison and application of ASR techniques shows that the detection of phonetic-acoustic features providing linguistic and speaker information represents a different challenge for each speech sound category. In this regard, one of the most fundamental aspects of understanding the acoustic nature of sounds is the distribution of periodic and aperiodic energy. Speech sounds with a high degree of periodic energy, such as vowels, are better understood than speech sounds with a high degree of aperiodic energy, such as fricatives.

Phonetic investigations on vowels and different consonant categories have identified, for instance, the importance of voice onset time and formants in stop consonants and vowels as stable acoustic and perceptual cues. Moreover, studies focusing on idiosyncratic information in speech sounds found speaker-specific characteristics in vowel formants (McDougall and Nolan, 2007; Rose, 2007), and nasals (Enzinger and Balazs, 2011; Kavanagh, 2012).

Despite the wide presence of fricatives in languages of the world and the extensive studies on fricatives, existing research does not fully explain how fricatives can be identified and classified efficiently using acoustic cues. In addition, it is still debated to which extent fricatives contain idiosyncratic information and by which cues this information can be measured and extracted.

2.1 Typology of fricatives

From a typological view, fricatives represent the second largest group of obstruents (after stop consonants) across the world's languages (Maddieson and Disner, 1984). They exist at various places and voice settings and can undergo several secondary articulation processes such as palatalisation or aspiration (Ladefoged and Maddieson, 1996; Maddieson et al., 2013). Fricative inventories can vary widely across the languages of the world, as reported in the LAPSYD database (Maddieson et al., 2013). There are languages as the Australian languages that arguably lack fricatives, (Butcher, 2003; Maddieson and Disner, 1984), and languages like Abkhaz, with 19 fricatives (Maddieson et al., 2013). Table 2.1 shows the number and percentage of languages by the size of fricative inventories.

In the phoneme inventories of 10% of the languages, phonological fricatives are absent. Interestingly, there are only a few languages with one fricative. A large number of languages employ between two and four fricatives. The modal number is three and accounts for 20 % of all languages. Languages have rarely more than four fricatives and even rarer is the occurrence of

Table 2.1: Number and percentage of languages by the size of fricative inventories across 683 languages from various areas of the world taken from the LAPSYD database (Maddieson et al., 2013).

Fricatives	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	19
Languages	68	34	116	138	88	53	51	44	29	26	6	10	5	4	5	2	3	1
Procentage	10	5	17	20	13	8	7	6	4	4	1	1	1	1	1	1	1	1

more than nine fricatives. The distribution of fricative inventory sizes by areas is visualised in Figure 2.1. The languages (683) are grouped into six areas (with different amounts of languages per group) and one group summarises languages with non-defined areas (NN).

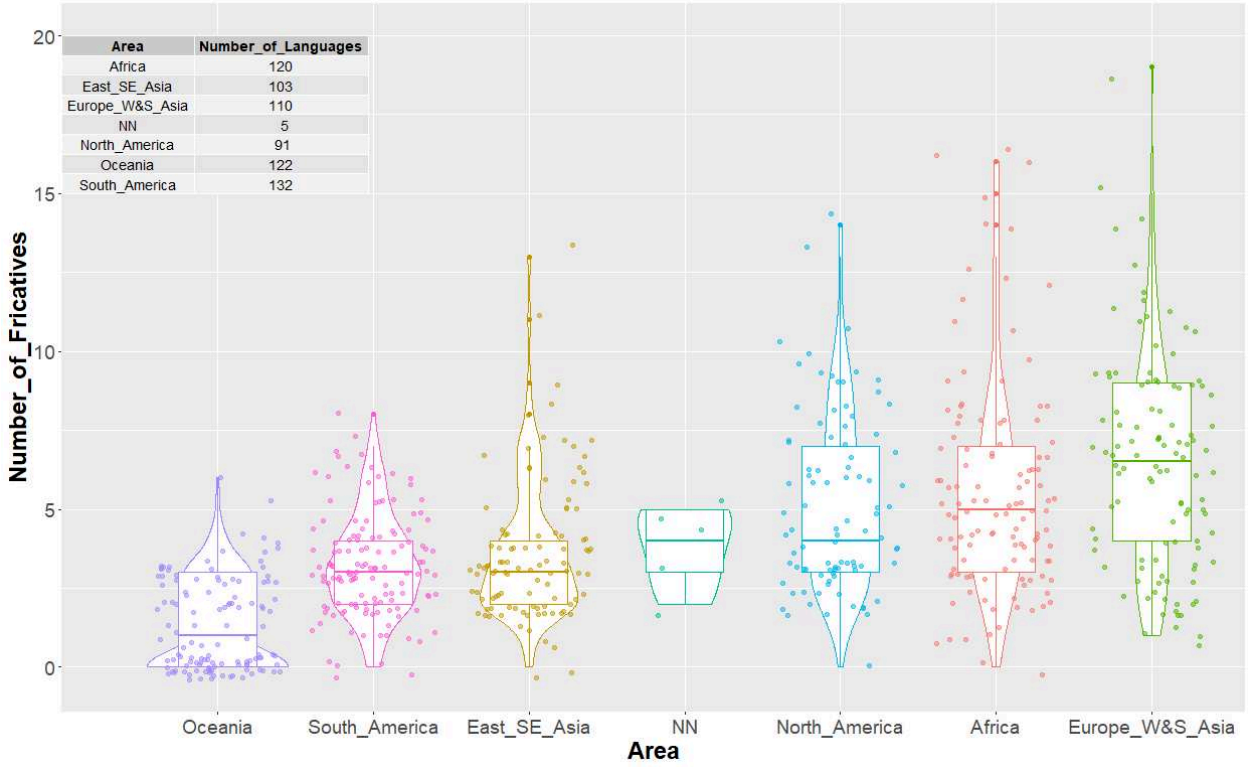

Figure 2.1: The areal distribution of fricatives in a sample of 683 languages from the LAPSYD database (Maddieson et al., 2013). The x-axis refers to the areas and the label NN refers to languages with non-defined areas. The y-axis shows the number of fricatives

The typological distribution of fricatives across the areas shows that Australian languages are not the only group lacking fricatives. Fricatives are also absent in some languages across the areas of South and North America, Oceania, East and Southeast Asia and Africa. The largest number of languages missing fricative phonemes is in Oceania. In general, it is striking that the Oceanic and South American languages have very small fricative inventories of a maximum of 6 and 8 fricatives. Some specific languages from the remaining areas show larger fricative inventories. As an example, Abkhaz (Area Europe_W&S_Asia) has 19 fricatives. Languages such as Tashlhiyt, Hobyot, and Ghomara Berber (from the area of Africa) also have 16 fricatives.

2.2 Production of fricatives

The basic production mechanisms of fricative sounds are well understood and described by multiple studies (Catford, 1997; Shadle, 1990; Stevens, 1998). In the production of both voiceless and voiced fricatives, two overlapping and interacting sources are involved. The most significant parameters defining the acoustics of noise sounds are the length of the front cavities, the flow rate and the presence of an obstacle.

The production of voiceless fricatives involves the interaction of two noise sources: the generation of a turbulent airflow in the pharynx and the oral cavities (Catford, 1997; Shadle, 1990; Stevens, 1998). The first source of frication noise production is a turbulent airstream resulting from the airflow passing through a narrow constriction of the vocal tract and producing random fluctuations (Catford, 1997; Shadle, 1990; Stevens, 1998). The second noise source is represented by different configurations of the front cavities (Shadle, 1990). In the production of sibilant fricatives, the upper and lower teeth are involved in noise generation by acting as an obstacle. In fricatives such as the voiceless velar, noise is generated by a wall source parallel to the airflow. No obstacle and wall source are required in the production of bilabial fricatives where frication noise is generated by turbulent airflow. Additionally, secondary articulations such as palatalization or aspiration can complexify the articulatory and acoustic structure observed in fricatives. Though typologically rare, phonologically aspirated voiceless fricatives involve for instance the production of both frication and aspiration noise leading to further challenges in their characterization (Rabha et al., 2019).

The production of voiced fricatives differs from the production of voiceless fricatives. Voiced fricatives involve glottal vibration over at least a period of time (Stevens et al., 1992). As in voiceless fricatives, the production of voiced fricatives implies the combination of two simultaneous aeroacoustic sources. In contrast to voiceless fricatives, these two sources are of a different nature. The first source is represented by the vibration of vocal folds generating a periodic signal as in vowel production. The second source is identical to the noise source in voiceless fricatives and frication noise is generated in the cavities (Jesus and Jackson, 2008; Pincas and Jackson, 2006). Thus, it is assumed that these two sources are not just simply overlapping in voiced fricatives. They undergo a complex interaction, which also represents an aerodynamic challenge, as the production of turbulence is complicated by the lower airstream velocity produced by vocal vibration. As a consequence, some voiced fricatives lack frication during articulation, and the realizations of these sounds are more approximant-like (Jesus and Jackson, 2008). The combination of voice and frication sources co-occurs not only in voiced fricatives but also in the vowel-fricative transition regions of voiceless fricatives. The interaction between these two sources is determined by their relative timings, the on and offset, the fundamental frequency (f0) and the levels of voicing and frication (Jesus and Jackson, 2008).

2.3 Linguistic and speaker information in fricatives

Speech production is influenced by both anatomical predispositions and movements of the vocal tract, which are controlled by neuromuscular programming (Dellwo et al., 2007). Any change in place and length of the constriction causes a change in the size and shape of the cavities behind and in front of the constriction. This, in turn, can result in a change of the acoustic characteristics in the produced speech signals. Therefore, all speech signals carry both linguistic information, such as place of articulation or voicing, and certain speaker-specific characteristics.

Human listeners are capable of perceiving and understanding speech even in difficult listening conditions. Regardless of, for instance, the volume of the speaker, background noise, or in whispered speech, a part of the linguistic information seems to be preserved in the speech signal. One explanation is provided by the *invariant theory*, which predicts unique and distinctive temporal, spectral and/or amplitude characteristics in the acoustic signals. These acoustic properties can be extracted from speech signals and serve as crucial perceptual cues (Blumstein and Stevens, 1981).

While such an approach was successful in finding, for example, voice onset time and formants as stable acoustic and perceptual characteristics for stop consonants and vowels, when it comes to fricatives, such acoustic invariant properties are highly debated. Several studies even argued that there is no set of properties that characterizes all fricatives and that only a distinction between the sibilants and non-sibilants can be made (Ladefoged and Maddieson, 1996).

Nevertheless, there is abundant literature that tried to identify measurements allowing the description and classification of fricatives. Most work has concerned the English voiceless fricatives, and the contrasts in places of articulation (Behrens and Blumstein, 1988; Jassem, 1965, 1995; Jongman et al., 2000; Maniwa et al., 2009; McMurray and Jongman, 2011; Shadle, 1986, 1990; Shadle and Mair, 1996; Strevens, 1960). Fricative inventories of other languages are less studied. Single analysis exist on for instance Spanish (de Manrique and Massone, 1981), Polish (Jassem, 1995; Żygis and Padgett, 2010), Japanese (Funatsu and Kiritani, 1998), Dutch (Kissine et al., 2003) Greek (Lilley et al., 2021; Nirgianaki, 2014), Romanian (Spinu et al., 2012), Lebanese Arabic (Al-Tamimi and Khattab, 2015) and Azerbaijani (Ghaffarvand Mokari and Mahdinezhad Sardhaei, 2020).

In addition to their linguistic meanings, speech sounds also convey information about the speaker. Research on idiosyncrasy assumes that motor control in speech is highly individual like other modes of human movements such as human gait (Matovski et al., 2010). Individual characteristics are expected to be reflected in the physical properties of speech sounds (He and Dellwo, 2014). Therefore, idiosyncrasies in speech such as gender, accent, language, emotions, or health status can be exploited not just by listeners, but also by technological application through the extraction of acoustic cues (Dellwo et al., 2007). One of the best-described speaker characteristics is gender and the acoustic contrast between female and male speakers is argued to be well understood and explained by physiological and sociophonetic differences (e.g. Jongman et al., 2000; Munson et al., 2006). Perception experiments (Schwartz, 1968) and acoustic cue-based recognition tasks (Ghaffarvand Mokari and Mahdinezhad Sardhaei, 2020; Spinu et al., 2018; Spinu and Lilley, 2016) provide evidence that gender information can be obtained from fricative sounds.

To which extent fricatives contain speaker-specific information, besides gender, is still debated. Some research demonstrated, for example, that models built to identify and distinguish speakers in vowels exhibit a decrease in performance for fricatives and nasals. The authors concluded that less speaker information is contained in these sound categories (Gendrot et al., 2020). Significant differences between voiced and voiceless sounds in general, and between fricatives, in particular, were reported for phoneme base speaker identification in Arabic consonants (Alsulaiman et al., 2017). Voiced Arabic fricatives [ʕ, h], and [z] showed high classification rates and voiceless fricatives a very poor performance (Alsulaiman et al., 2017). Moderate performance in fricatives was also identified in forensic voice comparison of six sound categorise (Ajili et al., 2017). Opposite findings were reported for English fricatives where a high recognition rate was achieved with vowels and also fricatives (Antal, 2008). Additionally, numerous studies

on idiosyncratic information in fricatives found substantial speaker variation and significant potential for speaker discrimination (Gordon et al., 2002; Hughes and Halle, 1956; Kavanagh, 2011; Narayanan et al., 1995; Newman et al., 2001; Silbert and de Jong, 2008; Smorenburg and Heeren, 2020), suggesting their further exploration (Kavanagh, 2012; Schindler and Draxler, 2013).

2.4 Acoustic features measured in fricatives and automatic classification

Different vocal tract configurations during the production of fricatives affect various acoustic measures. It is generally agreed that the size and shape of the vocal tract determine the spectrum of a fricative (Stevens, 1998), which is argued to be well described by the acoustic features of the *Spectral Peak Location* and the first four *Spectral Moments* (spectral *mean*, *spread*, *skewness* and *kurtosis*) (Hoelterhoff and Reetz, 2007; Jesus and Shadle, 2002; Jesus and Jackson, 2008; McMurray and Jongman, 2011; Shadle and Mair, 1996).

The spectral peak location and the four spectral moments are the most common acoustic features measured to investigate linguistically stable properties as well as speaker-specific characteristics in fricative sounds. The spectral peak location (*peak*) measures the frequency of the highest amplitude, which is connected to tongue movements during the production of fricatives at different places of articulation (Hughes and Halle, 1956). The first spectral moment describes the mean distribution of spectral energy or the centre of gravity (*cog*) of the fricative (Forrest et al., 1988). The second spectral moment *sdev* refers to the spectral spread or variance of the energy around the mean. Skewness (*skew*) gives insights into the spectral tilt and measures the overall asymmetry of the energy distribution. A skewness of 0 means a symmetrical distribution around the mean. A positive skewness suggests a negative tilt with a concentration of energy in the lower frequencies. A negative skewness infers a positive tilt and a predominance of energy in the higher frequencies (Newell and Hancock, 1984; Peeters, 2003). Finally, kurtosis (*kurt*) specifies the "peakedness" or flatness of the distribution: a spectral kurtosis equal to 3 indicates a normal distribution, while a smaller value than 3 suggests a flat distribution, and a higher value stands for a narrow distribution (Newell and Hancock, 1984; Peeters, 2003).

Acoustic-phonetic research focusing on the linguistic properties of fricatives is characterised by controversial results. As an example, it is still debated whether the spectral peak and spectral moments are sufficient to identify and categorise all fricatives. Controversially, other investigations argued that while the spectral moments carry important information about fricatives, they cannot reliably distinguish places of articulation of these fricatives (Shadle and Mair, 1996)

Nevertheless, the spectral domain is excessively studied in fricative research. Several researchers suggested that the frequency of the spectral peak is connected to the tongue movements during the production of fricatives at different places of articulation, and this value decreases from high to low frequencies as the tongue moves from front to back (Hughes and Halle, 1956; Jongman et al., 2000). However, this could not be confirmed for Greek fricatives (Nirgianaki, 2014). The spectral peak was found to distinguish between sibilants and non-sibilants, and, within the former, between the alveolars and post-alveolars (Behrens and Blumstein, 1988; Heinz and Stevens, 1961; Jassem, 1965; Shadle, 1990; Strevens, 1960). On the other hand, these and other analyses of spectral cues have identified a main effect of the

speaker, gender (Hughes and Halle, 1956; Jongman et al., 2000; Nirgianaki, 2014), and of the vowel context, which influences the tongue body during the production of the fricative (Mann and Repp, 1980; Nirgianaki, 2014; Soli, 1981; Stevens, 1998). As an example, the impact of the following vowel was stronger for [f] than for [s], and even less for [ʃ] (Stevens, 1998).

The first spectral moment is often considered in the analysis of frication noise. Several studies agree that the centre of gravity can distinguish between non-sibilants and sibilants, and within sibilants (Jongman et al., 2000; Kochetov, 2017; Nittrouer et al., 1989). Across the sibilant fricatives *cog* is higher in [s] than in [ʃ]/ (Funatsu and Kiritani, 1998; Jongman et al., 2000; Nittrouer et al., 1989; Padgett and Zygis, 2007; Zsiga, 2000). In Russian, the centre of gravity was reported to be gender and speaker dependent, with higher values in word-initial than in word-medial positions (Kochetov, 2017).

The second spectral moment or spectral variance is less reported in the literature. The spectral variance was found to be lower for sibilants and higher for non-sibilants (Jongman et al., 2000), with the post-alveolar fricative [ʃ] having the lowest variance (Shadle and Mair, 1996).

More findings can be reviewed for the third and the fourth spectral moments, skewness and kurtosis. Several studies indicate that skewness and kurtosis may distinguish between [s] and [ʃ] (McFarland et al., 1996; Nittrouer et al., 1989). A negative skewness was reported for [s] and a positive one for [ʃ] (Jongman et al., 2000; McFarland et al., 1996; Nittrouer et al., 1989). For kurtosis, large positive values were measured for [s] and smaller positive or a negative ones for [ʃ] (Jongman et al., 2000; McFarland et al., 1996; Nittrouer et al., 1989).

The temporal properties of fricatives were far less investigated, with most studies agreeing that duration is not a distinct cue in fricatives at all (Jongman et al., 2000; Kochetov, 2017), or that duration can only contrast non-sibilants and sibilants (Behrens and Blumstein, 1988).

Spectral acoustic properties also vary within and across speakers. Several studies argued that speaker variation is place of articulation dependent and a greater variation was reported in anterior fricatives (Gordon et al., 2002; Kochetov, 2017). With regard to gender variation, a cross-linguistic study showed, for instance, that in some languages female speakers articulate front fricatives differently than males, resulting in acoustic gender variation (Gordon et al., 2002). Furthermore, gender variation in the spectral acoustic properties of fricative sounds was found in several studies. A very early perception experiment on English fricatives determined that human listeners can identify speakers' gender in isolated voiceless sibilant fricatives, by relying on higher spectral energy in female productions. This effect was not present for the nonsibilants (Schwartz, 1968). Several follow-up papers also revealed gender differences in the spectral domain. They often reported higher values for female speakers in the centre of gravity, and peak frequency (Flipsen et al., 1999; Gordon et al., 2002; Jongman et al., 2000; Kochetov, 2017; Ludger et al., 2021; Newman et al., 2001). Furthermore, gender variation was identified in the spectral skewness (Flipsen et al., 1999; Ludger et al., 2021; Munson et al., 2006; Stuart-Smith, 2007).

Research on idiosyncratic characteristics in fricatives revealed that besides gender information further speaker properties are encoded in the spectral shape of fricative sounds. In an early study, it was shown that the spectral peak frequency in voiceless fricatives is highly variable between speakers and one speaker's alveolar peak can appear as the post-alveolar peak frequencies of another speaker (Hughes and Halle, 1956). The spectral moments are considered to serve as reliable acoustic cues for speaker discrimination in [f] and [s] sounds (Schindler and Draxler, 2013), while the strongest inter-speaker variability was reported in the spectral shape

of the alveolar [s] (Gordon et al., 2002; Kavanagh, 2011, 2012). Additionally, speaker variation was found in the temporal domain. A laryngographic analysis of voiced obstruents showed that vocal fold vibration varies between speakers, resulting in different frication and voicing duration as well as in different patterns of devoicing (Barry, 1995).

Several studies argue that gender and speaker-specific properties are not solely reflected in the spectral and temporal domains but also in further acoustic characteristics. For instance, it was claimed that females produce stronger acoustic distinctions and articulate contrasting vowels and consonants more clearly. The productions of vowels (Diehl et al., 1996; Weirich and Simpson, 2014) and fricatives (Weirich and Simpson, 2015) of female speakers tend to occupy a larger phonetic space than male speakers. Furthermore, duration and spectral analysis are not equally informative for all speakers when it comes to discriminating between them. While some speakers can be identified by these measures, others cannot. Acoustic measures can provide information on speakers, but more for individuals at the extremes rather than in the middle of the distribution (Kavanagh, 2012). Consequently, not all speakers can be identified with the same performance rate by acoustic cues. In a study comparing spectrograms and phonetic features extracted from vowels, differences between speakers were also reported. The authors concluded that there are some *good speakers* showing the best results in the identification task and *poor speakers* showing poor results (Gendrot et al., 2019). On the other hand, some articulation and acoustic studies claim, that intra-speaker variability in obstruents is contrast- and/or cue-specific rather than a general talker characteristic (Harper, 2021; Romeo et al., 2013).

In distinction to acoustic speech features (ASFs) such as the spectral peak and the spectral moments, a spectrum of speech sound can also be described through the more mainstream approach based on Mel-frequency coefficients (MFCCs). The advantage of ASFs is that they can be related to certain articulation mechanisms and they can contribute to a better understanding of perceptual crucial properties in speech sounds. MFCCs, on the other hand, are often used in speaker recognition tasks and ASR technologies as they encode most of the information found in speech signals. However, MFCCs represent an abstract set of cues which are difficult to interpret and relate to articulation and perception mechanisms.

In terms of the predictive power reported in the literature, temporal and spectral measures achieved quite a low accuracy of about 77% (Jongman et al., 2000) and between about 79% and 85% (McMurray and Jongman, 2011) for predicting the place of articulation of English fricatives. The accuracy was about 61 % for Greek fricatives (Nirgianaki, 2014). In terms of MFCCs, several recent studies have focused on the extraction of cepstral coefficients (CCs) on the Mel (Kong et al., 2014) or the Bark scales, to describe and distinguish fricative place, voicing and palatalization contrasts (Ghaffarvand Mokari and Mahdinezhad Sardhaei, 2020; Jesus and Jackson, 2008; Lilley et al., 2021; Spinu et al., 2018; Spinu and Lilley, 2016), achieving a much better predictive power of around 90%.

Some of these studies also compared the classification of spectral measures and cepstral coefficients in predicting gender. The results on Azerbaijani (Ghaffarvand Mokari and Mahdinezhad Sardhaei, 2020), Romanian (Spinu and Lilley, 2016) and a subset of Russian fricatives (Spinu et al., 2018) showed that cepstral coefficients clearly outperform common spectral, temporal and amplitude measures. The difference in accuracy was very similar with classification rates, with around 60% for ASFs and around 80% and higher for CCs.

2.5 Challenges in fricative research

The noisy and aperiodic nature of fricative sounds causes a number of factors that complicate the understanding of the acoustics of fricatives. The following outline should provide some insight into the current methodological challenges of fricative research.

One of the most common problems in linguistic and speaker research on fricative sounds is the low pass filtering of frication noise. Fricative sounds consist either of only aperiodic energy as in the case of voiceless fricatives or of the interaction of periodic and aperiodic components as in voiced fricatives. By comparing the frequency ranges of fricative sounds with other consonants, it was discovered that aperiodic spectral energy is presented in much higher frequency ranges than in other sounds (Strevens, 1960). Nevertheless, studies investigating the spectral shape of fricatives mostly considered only the information coded in bandwidths up to around 10kHz (Flipsen et al., 1999) and, more generally, the relevance of high-frequency has been overlooked, as underlined by Monson and colleagues (2014). In the past, this filter was motivated by the technological limitation that only a particular bandwidth could be analysed (Strevens, 1960). Even though speech processing technologies developed in the last decades (e.g. through bandwidth extension (Jax and Vary, 2003)), studies on linguistic and speaker information in fricatives continued investigating a frequency threshold up to 8 to 12 kHz (Forrest et al., 1988; Gordon et al., 2002; Jongman et al., 2000; Kavanagh, 2011; Kochetov, 2017). Studies of speech production and perception in patients with cochlear implants and studies of hearing loss in the elderly have both argued that frequencies above the filtered ranges also matter in speech perception. In the systems of cochlear implants, for instance, not all the acoustic information about the spectral shape in the high-frequency range is sufficiently provided to the user (Moore, 2003). This in turn can cause developmental difficulties in children perceiving and articulating fricative sounds correctly (Grandon and Vilain, 2020). As a result of the low pass filtering of fricative sounds, the understanding of how much linguistic and speaker information and variation is coded in the higher frequency ranges remains unclear.

Further aspects contributing to the complexity of identifying stable and invariant cues in fricative sounds are articulatory and acoustic language-specific characteristics. Language-specific and cross-linguistic studies on linguistic features as well as speaker variation show strong differences in the articulation and acoustics of fricatives among languages, suggesting the existence of different acoustic features of the same sound (Catford, 1988, 1997; Gordon et al., 2002; Hayward, 2000; Jongman et al., 2000; Ladefoged and Wu, 1984; McMurray and Jongman, 2011; Nirgianaki, 2014; Reidy, 2016). Speakers of different languages can apply various strategies to produce the same phoneme, resulting in varying acoustic characteristics, as reported by acoustic (Gordon et al., 2002; Hayward, 2000; Ladefoged and Maddieson, 1996) and articulatory studies (Narayanan et al., 1995). For instance, speaker and gender variation in the duration and the spectral shape is prominent only for some languages, but not for others (Gordon et al., 2002).

Apart from language-dependent factors, variation of fricatives is also identified between speakers and within speaker (Catford, 1988; Gordon et al., 2002; Hayward, 2000; Ladefoged and Wu, 1984; Newman et al., 2001; Reidy, 2016). Speakers of the same language can apply different articulation strategies and show diverse acoustic characteristics. Speaker-specific acoustic properties complicate the understanding of stable linguistic cues, but on the other hand, it shows that there is speaker variation.

Moreover, fricatives as continuous and complex aperiodic sounds with diffused energy, could

so far not be convincingly described by unique and distinct acoustic properties, because most measured features, such as the spectral peak location or the four spectral moments are found to be vowel context dependent (Jongman et al., 2000; McMurray and Jongman, 2011; Nirgianaki, 2014; Reidy, 2016). Other studies also report coarticulation effects in fricative sounds and that the articulatory movements have salient acoustic consequences on the spectral energy distribution. As a consequence, the spectro-temporal trajectory has also been successfully exploited to study fine-grained differences among voiceless fricatives (Reidy, 2016).

Studies on the acoustics of fricatives comparing different speaking styles report considerable acoustic variation in regard to linguistic and speaker-specific features. For instance, a study comparing the acoustic properties of fricatives from productions of clear speech where speakers emphasised the contrast of minimal pairs, with fricatives which were produced in a non-contrastive context, found considerable variation. The study demonstrates that there are systematic acoustic-phonetic modifications in the production of clear fricatives (Maniwa et al., 2009). Another study dealing with the comparison of speaker-specific features in nasals and fricatives between read and spontaneous speech showed slightly better performance in a speaker identification task for spontaneous speech. The authors hypothesised that the speaker's individual characteristics are more represented in spontaneous than in read speech (Schindler and Draxler, 2013).

Another constraint in the investigation of fricative sounds represents the limited datasets employed in these studies. The analyses often considered voiceless fricatives with a focus on the English language and the alveolar [s]. Few studies took into account voiced and palatal fricatives. Furthermore, previous analyses often included only a limited number of speakers. They were rarely productions of more than ten speakers of the same gender (e.g. Jongman et al., 2000; Kochetov, 2017; Spinu et al., 2018).

2.6 Methods in fricative research

Various methods have been employed for describing and characterising linguistic information or speaker-specific properties in fricatives. For the statistical analysis of variance, the most common methods are ANOVA and logistic regression models (Ghaffarvand Mokari and Mahdinezhad Sardhaei, 2020; McMurray and Jongman, 2011; Spinu and Lilley, 2016).

Additionally, machine learning techniques received much attention in acoustic and speech processing research and found a wide range of applications (Bianco et al., 2019; Michalopoulou et al., 2021). Further fields which are advanced by machine learning techniques include music perception, bioacoustics, hearing and hearing aids, and emotion recognition (Michalopoulou et al., 2021). Diverse machine learning techniques were developed and applied in phoneme recognition tasks (Chorowski et al., 2015) and in phoneme-based speaker recognition for different contexts and languages (Alsulaiman et al., 2017; Antal, 2008). Machine learning is also frequently combined with deep learning for visual phoneme recognition (Algabri et al., 2020) and for speaker recognition (Gendrot et al., 2019, 2020) by using broadband spectrograms of speech sounds.

The advantage of machine learning and deep learning methods in phonetic-acoustic research is that they can enhance the extraction of statistical-based information from acoustic data. They can help to clarify and describe complicated acoustic phenomena by showing how features interact and can be used for recognizing patterns in different sounds. However, even though machine learning is capable of advancing phonetic and phonological research, for instance by

automatically identifying speech sounds by their acoustic properties, they are rarely utilized for research purposes. As an example, in fricative research, only a few studies approached the identification of fricatives using automatized methods, and they mostly used deep learning methods (Anjos et al., 2020; Nagamine et al., 2015).

2.7 Russian phoneme inventory and research on fricatives

The articulatory and acoustic features of the Russian sound system are well described by a number of studies (Bolla, 1981; Fant, 1960; Halle, 2011; Jones and Ward, 1969; Shupljakov et al., 1968; Timberlake, 2004; Zsiga, 2000). However, it is to be noted that these grammars show some controversy in the description of the production and classification of vowels and consonants. Considering consonants, there are different counts of places of articulation reported in these grammars. In addition, several consonants are assigned to different places of articulation. To name one example, the [s] and [z] fricatives are in one grammar defined as a dental place of articulation (Timberlake, 2004), while in another study the same phoneme refers to an alveolar place of articulation (Jones and Ward, 1969). Such discrepancies can be found in many language descriptions, which reflect on the one hand certain speaker-specific characteristics of the participants in the production of speech sounds and on the other hand the complexity of grasping the phoneme inventory of a language. An overview of the generally agreed upon the consonant inventory of Russian is shown in Table 2.2.

Table 2.2: Phoneme inventory of Russian consonants (This table is adapted from (Jones and Ward, 1969)).

OBSTRUENTS	abial	labiodental	dental	alveolar	post-alveolar/palatal	velar
voiceless stop	p p^j		t t^j		k k^j	
voiced stop	b b^j		d dj		g g^j	
affricate			ts		tɕ	
voiceless fricative		f f^j		s s^j	ʃ ɕ	x x^j
voiced fricative		v v^j		z z^j	ʒ $ʒ^j$	
SONORANTS						
glide					j	
nasal stop	m m^j	n n^j				
lateral			l l^j			
trill			r r^j			

Despite variations between different grammars, the major characteristic of the Russian phonological system is stress in vowels and palatalization in consonants. Russian vowels can be stressed or unstressed. Therefore, stress operates on different levels. Phonetically, stressed vowels are longer than unstressed vowels and it is claimed that they show acoustically more distinct features (Timberlake, 2004). Furthermore, the stress in Russian vowels is relevant to the lexicon and morphology. Unlike other Slavonic languages, the stress in Russian does not follow any rules and can occur on any syllable of a word. However, words can only contain one stressed syllable, and stress can differentiate the meaning between words. In terms of prosody,

stress is an important factor to define the intonation contours which are on or around the stressed syllable (Timberlake, 2004).

Consonants are generally divided into two groups: Obstruents and Sonorants. Most consonants in Russian exist as non-palatalised and palatalised phonemes. Additionally, most obstruents can be voiced and voiceless. Thus, palatalization and voicing are significant factors to distinguish between words. While voicing contrast exists in many languages of the world, palatalization contrast is very rare. For instance, only 10 out of 806 languages in the LAPSyD database distinguish between non-palatal [s] and palatal [sʲ], as the Russian language does (Maddieson et al., 2013).

The palatalization contrast in Russian differs between sounds articulated at different places. There are also some sounds without a palatalized or non-palatalized counterpart. Those include for example the palatal affricate [tɕ] and the non-palatal affricate [ts] and the glide [j] (Timberlake, 2004). The non-palatal post-alveolar fricatives [ʃ] and [ʒ] and the palatal [ɕ] and [ʒʲ]are regarded not to be paired in this sense, because [ʃ] and [ʒ] do not follow the same rules as other consonants do (become palatalized at the end of a noun in the locative singular or in the conjugation of verbs) (Timberlake, 2004). Nevertheless, in phonetic acoustic studies these sounds are often treated as palatal and non-palatal pairs (Kochetov, 2017; Spinu et al., 2018).

In Russian, both palatalized and non-palatalized variants occur before vowels, after vowels in word-final position, and in consonant clusters. In CV syllables, the presence/absence of palatalisation affects the production of the following vowel. In the word-final position, the palatalisation contrast is intrinsic to the consonant (Bolla, 1981). The phonotactic rules in consonant clusters define that if the second consonant is palatal, the first consonant will also be palatalized. If the second consonant is non-palatalized, the first one can be both palatal or non-palatal (Bolla, 1981). Other distribution and phonotactic rules apply to voiced and voiceless consonants. Voiceless obstruents occur in both CV and VC syllables. At the word-final position, the voiced consonants become devoiced. In consonant clusters, the second consonant determines the voicing of the preceding sound: if a voiced consonant is followed by a voiceless consonant, it will also be devoiced. If the second consonant is voiced the first one will also be produced as a voiced (Bolla, 1981).

To summarise, the Russian consonant inventory offers an interesting research field, since it exhibits a rich and rare variety. The fricative inventory in particular represents a large set including voiceless, voiced, palatal and non-palatal fricatives. Such a variation offers a wide range of possibilities for the investigation of the articulation and acoustics of fricative sounds.

There is a number of studies investigating the articulatory properties of Russian fricative sounds (e.g. Bolla, 1981; Fant, 1960; Kedrova et al., 2008; Litvin, 2014), but only a handful dealt with acoustic analyses. Furthermore, existing surveys on the acoustics of Russian fricatives are limited and do not take into account all fricative consonants, or only consider a small set of tokens, vowel contexts, word positions, and/or speakers (Derkach et al., 1970; Kochetov, 2017; Padgett and Zygis, 2007; Spinu et al., 2018, 2012). In linguistic research, this leads to a lack of systematic documentation of topologically contrasting fricatives (Kochetov, 2017). Research on gender or speaker variation in Russian fricatives is also rare. These studies are usually based on the productions of a few speakers and do not report much on speaker-specific characteristics. Speaker and gender variation are reported so far for sibilant palatal vs. non-palatal fricatives (Kochetov, 2017; Spinu et al., 2018) and for the variation of vocal fold vibration in voiced fricatives from eight speakers (Barry, 1995). One of the main aims of this thesis is to fill this gap of data by providing an open-access acoustic database of fricatives in Russian.

Chapter 3

The database

This section is represented by the paper: Ulrich N. (2022), Database description: Russian fricatives. The description is submitted to the Journal of the Acoustical Society of America as a letter to Editors on the 14th of October 2022.

The database description provides detailed reports on the data retrieval process, the experimental design, the participants, and the recording procedure. The paper also gives an overview of the pre-processing steps of the data, the resulting data files, and possible reuse options. The database is published and available for scientific research. The usage instructions and accessibility of the database are also described.

Data base description: Russian fricatives

Natalja Ulrich [1a]
[1] *Lab Dynamics of Language UMR 5596, CNRS and University Lyon 2, France*

This paper presents a speech database primarily designed to investigate linguistic and speaker information in fricative sounds in Russian. Acoustic recordings were obtained from 59 native Russian speakers. The resulting dataset consists of 198 sentences for each speaker. In these sentences, real words containing one of the fricatives [f], [s], [ʃ], [x], [v], [z], [ʒ], [sʲ], [ɕ], [vʲ], [zʲ] were embedded. The total amount of fricative tokens is 22,561. The number of observations per sound differs across categories, because of their natural distribution. The dataset is made available as a collection of audio files in *wav* format along with companion Praat *TextGrid* files for each sentence. Target fricatives are furthermore available as individual *wav* files. The database can be accessed with the DOI https://doi.org/10.48656/4q9c-gz16.
Additionally, the experimental design allows the investigation of other sound categories. The number of speakers recorded gives further possibilities for phonetic-oriented speaker identification studies.
[https://doi.org(DOI number)]

[XYZ] Pages: 1–4

I. BACKGROUND

The construction of the database was primarily motivated by the intention to investigate linguistic and speaker-specific information of complex sounds such as fricatives.

The Russian fricative inventory offers an interesting research field for examining complex acoustic phenomena since it exhibits a rich and rare variety. For instance, only 10 out of 806 languages in the LAPSyD database distinguish between non-palatal [s] and palatal [sʲ], as the Russian language does (Maddieson *et al.*, 2014).

Furthermore, the Russian phonetic inventory contains at least 12 fricatives, at four (other descriptions of Russian fricative inventory state that there are five places of articulation (Bolla, 1981)) places of articulation [f, s, ʃ, x], with voicing [v, z, ʒ] and palatalization [fʲ, vʲ, sʲ, ɕ:, zʲ] contrasts (Timberlake, 2004).

Even though fricative sounds were extensively studied in the past, accessible databases designed for investigations of fricative sounds in a format suitable for reuse are rare or inaccessible. In terms of data size, most studies on fricatives were based on data from a few speakers. In general, the data and research available are very limited, especially for understudied languages such as Russian.

It is to note that there exists a large open-source Russian language data set – OpenSTT, available online at `https://github.com/snakers4/open_stt`. However, this Russian corpus is oriented towards deep learning and not phonetics studies.

To fill this gap, the experimental design and the number of speakers included in the current database give the possibility for further investigation of fricative sounds and other sound categories, along with intra- and inter-speaker variation or speaker identification tasks.

II. EXPERIMENTAL DESIGN AND DATA RETRIEVAL

A. The participants

The participants were 59 students (30 female) between 18 and 30 years old, studying at different departments of St. Petersburg University in Russia. All participants were born or lived in St. Petersburg since early childhood. No participants reported any speech or hearing impairment.

B. The stimuli

To obtain recordings of the target fricatives, a list of 94 real words containing minimal pairs (words that vary by only a single sound contrasted by place of articulation, voicing, and palatalisation) of the target fricatives was collected. Each word contains one of the target fricatives in a) word-initial position as a CV syllable (e.g. ***f****ara*), b) word-medial position in inter-vocal context (e.g. *ma****S****a*), and in c) word-final position in a VC syllable (e.g. *ve****s***).

The stimuli consist of a list of 198 Russian sentences including 293 target fricatives. All sentences are listed in Supplementary Materials 1 (SuppPub1_sentence_list). Each word appears three times in the stimuli list and was embedded in two different sentence structures: i) carrier sentence with the structure of “She said ”X” and not “Y””. Minimal pairs of real words containing one of the 11 tested fricatives were placed in both “X” and “Y” positions, ii) natural language sentence including each of the lexemes. Target fricatives in carrier sentences always occur in inter-vocal positions, with the exception of fricatives at the end of the word or at the end of the sentence. In natural sentences, the vowel context was not

[a] ulrichnatalja@gmail.com

TABLE I. Token count by sound. Each speaker produced the same amount (N=293) of token for each fricative category. The number of fricatives is not the same across all fricatives due to restrictions of occurrences in different surrounding contexts in Russian.

sound	[f]	[s]	[ʃ]	[x]	[v]	[z]	[ʒ]	[sʲ]	[ɕ]	[vʲ]	[zʲ]
frequency by speaker	36	67	55	11	29	27	24	15	15	11	3
total automatically and manually aligned	2301	3953	3245	649	1711	1593	1416	885	885	649	177
total automatically aligned	648	1206	990	198	522	486	432	270	270	198	54

controlled. Therefore, stimuli of type a) can appear with a preceding consonant. It is to note that the amount of tokens per sound differs strongly between the categories due to the naturally uneven distribution found in Russian as shown in Table I.

As an illustration, the fricative [x] is only involved in a few minimal pairs contrasted by place of articulation. This fricative was therefore infrequently recorded. Fricatives [zʲ] and [vʲ] also appear very rarely in words, being involved in a few minimal pairs contrasted by place of articulation or palatalization. Due to these reasons, these three sounds contain a very small number of examples per speaker. Furthermore, the preceding and following vowels vary between the fricatives.

C. Procedure

The recording sessions were conducted in an audiometric booth at the phonetic laboratory of the Phonetic Institute in St. Petersburg. Participants were briefly introduced to the purpose of the experiment, the expected duration, and the procedure. To avoid a distortion of the productions only after the recordings it was announced that it is a research on Russian fricatives. Participants who were invited to a second recording session were informed after the second recording of the purpose of the experiment. They were told that they have the right to withdraw at any time during (and after) the experiment and they were provided with the contact details of a person that can answer all their questions concerning the research and their rights. The participants were compensated for their participation.

Demographic data, such as sex and age, were registered before the experiment started. 19 Participants agreed to a second recording session.

The recording program `Speech-Recorder` version 3.28.0 https://www.bas.uni-muenchen.de/Bas/software/speechrecorder/ was used at a sample rate of 44.1 kHz (16-bit encoding). A clip-on microphone (Sennheiser MKE 2-P) was placed at a distance of 15cm from the speakers' mouth. The microphone was connected through an audio interface (Zoom U-22) to a laptop computer.

The participants were then instructed to read the 198 sentences aloud from a computer screen. The sentences were presented one by one in random order. The participants could repeat a sentence in the case of a production error.

III. DATASET DESCRIPTION

The dataset includes 198 audio files from 77 recording sessions. For each of the 198 sentences, a *wav* and a *TexGrid* are provided. The fricative data set consists of 293 tokens for each of the 77 recording sessions.

All audio files were automatically pre-processed using the Munich Automatic Segmentation System, `MAUS` (Kisler *et al.*, 2017; Schiel, 1999) available online at https://www.bas.uni-muenchen.de/Bas/BasMAUS.html. During this process *TextGrid* annotation files were generated. The TextGrid output contains three Tiers (Figure 1). The first tier shows the sentence segmented by word in Russian. In the second tier, the sentence was transcribed in Sampa (Gibbon *et al.*, 2017). The third tier provides time-aligned word segmentation at a phonemic level. In this tier, the boundaries of the target fricatives were manually corrected using `Praat` version **3.9** (Boersma and Weenink, 2022). Before prepossessing the raw data, the audio files were filtered below 80 and above 20050Hz with a smoothing of 80Hz to get rid of parasite noise. It is to be noted that only the first recordings were manually corrected, and the 18 second recordings were only automatically reprocessed. In total there are 22561 fricatives in the database, of which 17287 were extracted and are available as individual *wav* files.

In order to define the onset and offset of the full consonant, the broadband spectrogram was considered more relevant than the start of an aperiodic waveform with rising zero crossing rates. In intervocalic fricatives, the presence of formant columns is defined as the onset and offset of the fricative (following (Skarnitzl and Machač, 2011)). Some speakers ended their voiceless fricatives with a somehow long post-aspiration in intervocalic positions and/or when the fricative appeared at the end of the word and sentence. In these cases, the fricatives were segmented according to the changes in high-energy events and the post-aspiration part was not considered. The voiced fricatives represented also a segmentation challenge, because the waveform and spectrogram may be insufficiently informative to define the onset and offset. The boundaries of these sounds were identified according to perceptual judgments. In general, it should be noted

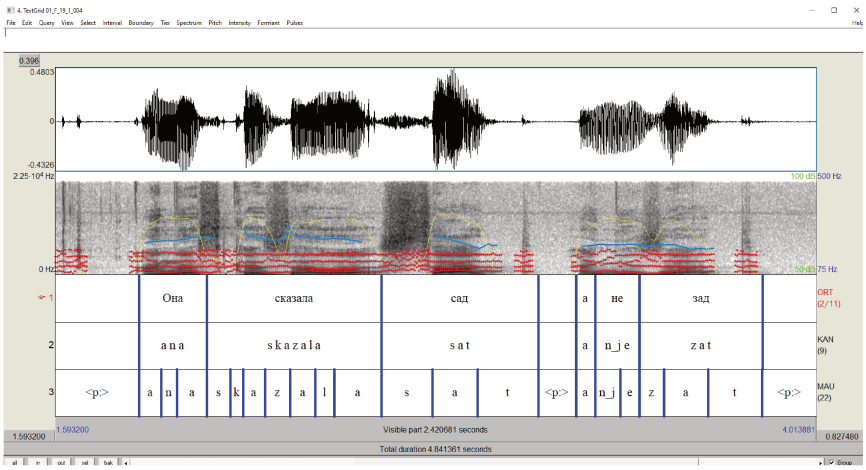

FIG. 1. The Praat window shows an example of a carrier sentence with the three tiers.

that it is hardly possible to standardise segmentation criteria across all fricatives and speakers. In case one defines the on and offset of the fricative boundaries differently, the manual correction process can be redone.

IV. USAGE NOTES

The data is available at *www.swissubase.ch* through the DOI *https://doi.org/10.48656/4q9c-gz16*. To access the data, a SWITCH edu-ID at *https://eduid.ch* needs to be created. This ID allows the user to log in to *SWISSUbase* and to download the data along with the Supplementary materials. The data has a closed license and is only available for teaching and research purposes.

The citation of the database should include both the current description and the data set. The dataset should be cited as follows: Natalja Ulrich: Russian Fricatives [Dataset]. Université Lumière Lyon 2. Distributed by SWISSUbase, Lausanne, 2022. https://doi.org/10.48656/4q9c-gz16.

The data is organized in three folders: *Recording_1*, *all_sounds_session_1*, *Recording_2*. Additionally two *csv* spread sheets are provided: a sentence list *sentence_list* (SuppPub1_sentence_list) and a *metadata* table (SuppPub2_metadata).

Folder *recording_session_1* stores all automatically and manually processed files. It is organized by speaker and consists of one folder for each speaker (N= 59). Each speaker's folder contains all sentences as *wav* and *TextGrid* files (N=396). This folder also includes the folder *all_sounds_session_1* with all the extracted fricatives from the first recording session (N= 17287). In a third folder *recording_session_2*, there are 18 recordings of the second session, which were only automatically processed.

To avoid showing the speakers' identities, the recordings were anonymized. The names of folders e.g *01_F_22_1* are coded as follows: The first two digits refer to speaker ID (01-59). The characters *F* and *M* refer to the speaker's gender. The following two digits indicate the age (18-30) of the participant at the time of the recording (2018). The last number indicates the recording session (1,2). The filenames of the sentences e.g. *01_F_22_1_001* represent the same information as the folder names and additionally the sentence number in the last three digits. The single sound files e.g. *01_F_1_001_11_v* contain information of speaker ID, gender, recording session, sentence no, interval number of the third tier and the fricative label in Sampa.

The *sentence_list* contains all recorded sentences. The first column *sentenceNo* corresponds to the sentence number in the file name. The second column presents the sentence in Russian and the third gives a transcription in IPA. A fourth column lists separately the target words from which the fricatives were extracted in IPA. The last column *comments* contains comments on the omitted sentences from the fricative extraction. More specifically, some of the sentences turned out to be miss-constructed and were excluded from the fricative database. A few sentences were exchanged in the middle of the experiment, and were therefore different between the participants (N=3). These sentences are marked in the sentence list and were excluded from the investigations so far (see Section V). Some of the natural sentences turned out to be useless for the defined research purposes and were also excluded from the data base (N=11). To fill the gaps of missing fricative examples from words embedded in natural sentences, these fricatives were extracted from other words. Moreover, in the process of automatically generating the *TextGrid* tiers and the transcription, some Russian vowels were wrongly transcribed. Only some vowels in words containing a target fricative have been corrected.

The *metadata* table gives information on the target fricatives. The columns *SentenceNo* and *intervalNo* indicate the sentence and interval number by which the fricatives can be identified. *SentenceTyp* refers to one of the two types of sentences: *CS* stands for carrier sentence and *NC* for none-carrier sentence. The column *Position* specifies the location of the target fricative in a carrier sentence. The annotation *1* stands for *"X"* position and *2* for *"Y"* position. The column *FricativePosition* shows the word-position of the fricative. The annotation *B* stands for word-initial position, *M* for word-middle, and *E* for word-final. The columns *Sampa* and *IPA* indicate the transcription format. The columns *PrecedingSound* and *FollowingSound* contain, as the column names indicate, the preceding and following sound of the target fricatives. The columns *Voicing* and *Palatalization* show whether the target fricative is voiced or palatalized. The sentence number, interval number and fricative label of each fricative filename correspond to the columns *SentenceNo*, *IntervalNo* and *Sampa* in the metadata table.

V. STUDIES ON THE DATA

Different samples were so far used in publications. The first study dealt with the identification of stable acoustic cues in the fricatives /f/, /s/ and /S/ (Ulrich *et al.*, 2021). In this study a sample of 40 speakers and the three fricatives was used. Several acoustic measures (peak frequency and peak amplitude, 4 spectral moments, duration and zcr) were compared with MFCCs. A second study focusing on intra- and inter-speaker variation in eight of the 11 fricatives is in revision (13.10.2022). For this study the full data set of the first recordings session was employed. The peak frequency, spectral moments, duration and harmonic to noise ratio were measured and compared with MFCCs.

VI. CODE AVAILABILITY

The Praat codes used to filter the sound files (SuppPub3_1_filter_the_noise) and to extract the target fricatives (SuppPub4_2_fricative_extraction) are available in Supplementary Materials.

ACKNOWLEDGMENTS

I wish to thank our participants, the Phonetic Lab in St. Petersburg (and the sound engineer Tatiana Chukaeva in particular), the University of Zürich for financial support, technical support, and help with the design of the experiment (to Volker Dellwo in particular). NU was partly supported by a grant from the Doctoral Program of Linguistics of the Faculty of Arts and Social Sciences, University of Zürich, Switzerland, NU, was funded by the IDEXLyon Fellowship Grant 16-IDEX-0005 (2018-2021), and indirectly by the Labex ASLAN (ANR-10-LABX-0081) of the University of Lyon within the program Investissements d'Avenir (ANR-11-IDEX-0007) of the French National Research Agency (ANR).

Boersma, P., and Weenink, D. (**2021**). "Praat: doing Phonetics by Computer" https://www.fon.hum.uva.nl/praat/.

Bolla, K. (**1981**). "A conspectus of Russian speech sounds" Publisher: Publ. House of the Hungarian Academy of Sciences.

Gibbon, D., Moore, R., and Winski, R (**1997**). "Handbook of standards and resources for spoken language systems," Walter de Gruyter.

Kisler, T., Reichel, U., and Schiel, F. (**2017**). "Multilingual processing of speech via web services," Computer Speech & Language **45**, 326–347, https://linkinghub.elsevier.com/retrieve/pii/S0885230816302418, doi: 10.1016/j.csl.2017.01.005.

Maddieson, I., Flavier, S., Marsico, E., and Pellegrino, F. (**2013**). "2020 LAPSyd: lyon-albuquerque phonological systems databases, version 1.0,".

Schiel, F. (**1999**). "Automatic Phonetic Transcription of Non-Prompted Speech,".

Skarnitzl, R., and Machač, P. (**2011**). "Principles of Phonetic Segmentation," Phonetica **68**, 198–9, doi: 10.1159/000331902.

Timberlake, A. (**2004**). "A Reference Grammar of Russian," Cambridge University Press.

Ulrich, N., Allassonnière-Tang, M., Pellegrino, F., and Dediu, D. (**2021**). "Identifying the Russian voiceless non-palatalized fricatives /f/, /s/, and /S/ from acoustic cues using machine learning," The Journal of the Acoustical Society of America **150**(3), 1806–1820, https://asa.scitation.org/doi/10.1121/10.0005950, doi: 10.1121/10.0005950.

Ulrich, N. (**2022**). "Russian Fricatives [Dataset]," Université Lumière Lyon 2. Distributed by SWISSUbase, Lausanne, doi: https://doi.org/10.48656/4q9c-gz16

Chapter 4

Defining linguistic information in fricatives sounds

In this section, the segmentation and the analysis of three voiceless fricatives are themed. The following study is a published paper: Ulrich, N., Allassonnière-Tang, M., Pellegrino, F., Dediu, D. (2021). Identifying the Russian voiceless non-palatalized fricatives [f], [s], and [ʃ] from acoustic cues using machine learning. The Journal of the Acoustical Society of America, 150(3), 1806-1820.

JASA ARTICLE

Identifying the Russian voiceless non-palatalized fricatives /f/, /s/, and /ʃ/ from acoustic cues using machine learning[a)]

Natalja Ulrich,[b)] Marc Allassonnière-Tang,[c)] François Pellegrino, and Dan Dediu
Laboratoire Dynamique Du Langage (DDL) UMR 5596, CNRS/Université Lyon 2, Lyon, France

ABSTRACT:
This paper shows that machine learning techniques are very successful at classifying the Russian voiceless non-palatalized fricatives [f], [s], and [ʃ] using a small set of acoustic cues. From a data sample of 6320 tokens of read sentences produced by 40 participants, temporal and spectral measurements are extracted from the full sound, the noise duration, and the middle 30 ms windows. Furthermore, 13 mel-frequency cepstral coefficients (MFCCs) are computed from the middle 30 ms window. Classifiers based on single decision trees, random forests, support vector machines, and neural networks are trained and tested to distinguish between these three fricatives. The results demonstrate that, first, the three acoustic cue extraction techniques are similar in terms of classification accuracy (93% and 99%) but that the spectral measurements extracted from the full frication noise duration result in slightly better accuracy. Second, the center of gravity and the spectral spread are sufficient for the classification of [f], [s], and [ʃ] irrespective of contextual and speaker variation. Third, MFCCs show a marginally higher predictive power over spectral cues (<2%). This suggests that both sets of measures provide sufficient information for the classification of these fricatives and their choice depends on the particular research question or application.
 https://doi.org/10.1121/10.0005950

(Received 1 February 2021; revised 4 August 2021; accepted 5 August 2021; published online 13 September 2021)
[Editor: Bozena Kostek] Pages: 1806–1820

I. INTRODUCTION

Building efficient techniques for the (semi)automatic identification of different speech sounds from their acoustic properties is very important not only for practical applications in speech processing, but also for advancing fundamental research in phonetics and phonology. While certain sound categories, such as vowels and stop consonants, are relatively well understood, more complex ones, such as fricatives, still represent a challenge, as it is currently unclear how they can be efficiently identified and classified using acoustic cues. Fricatives, as continuous and complex aperiodic sounds with diffused energy, have so far not been convincingly described by unique and distinct acoustic properties, because most measured features, such as, for instance, the spectral peak location or the four spectral moments, show considerable speaker variation, vowel context dependencies, and language-specific properties (Jongman *et al.*, 2000; McMurray and Jongman, 2011; Nirgianaki, 2014; Reidy, 2016).

In this paper, a machine learning-based approach is proposed to tackle this question by showing that computational classifiers are successful at correctly identifying the Russian fricatives [f], [s], and [ʃ] from a set of spectral and temporal acoustic cues. This process identifies a subset of acoustic cues that carry most of the information about these fricatives, helping advance the theoretical understanding of the perception and processing of fricatives in speech. The predictive power of these parameters is also compared with that of the more mainstream approach based on mel-frequency cepstral coefficients (MFCCs). Moreover, by making the computer code available in the spirit of open science, this study should contribute to the emergence of a standardised computational toolkit in phonetic science.

The paper is structured as follows: Sec. II surveys the literature concerning the most commonly measured acoustic cues for fricatives, discussing their applicability, limitations, and remaining gaps. Section III then introduces the dataset composed of 6320 tokens containing productions of the voiceless non-palatal fricatives [f], [s], and [ʃ] by 40 young native speakers of Russian from St. Petersburg. Please note that the sample analyzed here is only one part of a larger-scale investigation of Russian fricatives. The full dataset contains 22 854 tokens, including voiced and voiceless non-palatal and palatal fricatives, from 78 recording sessions with 59 (29 females) native speakers of Russian, of whom 19 (nine females) participated in a second recording session. The manual and automatic segmentation steps as well as the acoustic measurement procedure are also described. Moreover, an original classifier based on changes in zero crossing rate to identify the noise part of a fricative sound is introduced.

Section IV compares four different classifiers (decision trees, random forests, support vector machines, and feed-forward neural networks with backpropagation) on a large set of acoustic cues derived from different approaches and on 13 MFCCs to predict the fricative sounds. It shows, first,

[a)]This paper is part of a special issue on Machine Learning in Acoustics.
[b)]Electronic mail: natalja.ulrich@univ-lyon2.fr
[c)]ORCID: 0000-0002-9057-642X.

https://doi.org/10.1121/10.0005950

that all classifiers and both types of measurements have high predictive power and, second, that traditional measurements do so while using only a small subset of acoustic cues.

The paper ends with a discussion of the advantages and limitations of the methods and of the implications of the findings for understanding fricatives in general and Russian fricatives in particular.

II. AN OVERVIEW OF FRICATIVES

Even though fricatives have been extensively studied, neither the relationship between the articulators and their acoustic output, on the one hand, nor the perception mechanisms involved, on the other, are currently fully understood.

Despite this, the basic mechanisms involved in the production of voiceless fricatives are relatively well described: they are produced by a turbulent airflow in the pharynx and the oral cavities. The most significant parameters for acoustics are the length of the front cavities, the flow rate, and the presence of an obstacle.

During the production of voiceless fricatives, frication noise can in general be generated by two mechanisms: the first source of frication noise is a "channel turbulence" resulting from the air flow passing through a narrow constriction of the vocal tract, producing random fluctuations of the air-stream (Catford, 1977; Stevens, 1998). Depending on the fricative place of articulation, frication noise can also be generated by a second source, due to the airflow encountering a wall or an obstacle (e.g., the teeth), generating energy in the high frequency range of the noise spectrum (Catford, 1977; Shadle, 1990). Additionally, secondary articulations such as palatalization or aspiration can complexify the articulatory and acoustic structure observed in fricatives. Though typologically rare, phonologically aspirated voiceless fricatives involve, for instance, the production of both frication and aspiration noise, leading to further challenges in their characterization (Rabha *et al.*, 2019).

Based on the *invariant theory*, which predicts that unique and distinctive temporal, spectral, and/or amplitudinal characteristics of acoustic signals serve as crucial perceptual cues (Blumstein and Stevens, 1981), many studies have tried to find reliable and distinct acoustic cues of fricatives. While such an approach was successful in finding, for example, voice onset time and formants as stable acoustic and perceptual characteristics for stop consonants and vowels, when it comes to fricatives, such acoustic invariant properties are highly debated. On the other hand, several studies argue that there is no single property that characterizes all fricatives and that in grouping them, only a distinction between the sibilants and non-sibilants can be made (Ladefoged and Maddieson, 1996). Recent attempts to automatically classify the fricative manner of articulation (vs stop or affricate manners) confirmed both that a high level of accuracy can be reached and that performance significantly differs between sibilant and non-sibilant segments (Patil and Rao, 2008; Vydana and Vuppala, 2016). Moreover, cross-linguistic studies show strong differences in the articulation and acoustics of fricatives among languages and speakers, suggesting the existence of different acoustic features of the same sound (Catford, 1988; Gordon *et al.*, 2002; Hayward, 2000; Ladefoged and Wu, 1984; Reidy, 2016).

Nevertheless, there is an abundant literature that tries to identify measurements allowing the description and classification of fricatives. Most work has concerned the English voiceless fricatives and the contrasts in places of articulation (Behrens and Blumstein, 1988; Jassem, 1965, 1995; Jongman *et al.*, 2000; Maniwa *et al.*, 2009; McMurray and Jongman, 2011; Shadle, 1986, 1990; Shadle and Mair, 1996; Strevens, 1960), while the fricative inventories of other languages, such as Spanish (de Manrique and Massone, 1981), Polish (Jassem, 1995; Żygis and Padgett, 2010), Japanese (Funatsu and Kiritani, 1998), Dutch (Kissine *et al.*, 2003), and Greek (Nirgianaki, 2014), are much less studied. The research on the Russian sound system in general, and in particular on fricatives, is also strongly unrepresented, which results in a lack of systematic documentation of topologically contrasting fricatives (Kochetov, 2017). The Russian phonetic inventory is particularly interesting due to its complex phonetics and rich fricative inventory: there are at least 12 fricatives, at four places of articulation [f, s, ʃ, x], with voicing [v, z, ʒ] and palatalization [fʲ, vʲ, sʲ, ɕ:, zʲ] contrasts (Timberlake, 2004), offering thus a wide range of possibilities for the investigation of fricatives. However, only a handful of studies provide a description of the Russian phoneme inventory (Bolla, 1981; Shupljakov *et al.*, 1968; Timberlake, 2004), and most surveys of Russian fricatives (Derkach *et al.*, 1970; Kochetov, 2017; Padgett and Żygis, 2007) either do not take into account all its fricative consonants or only consider a small set of tokens, vowel contexts, word positions, and/or speakers.

Concerning the effects of different vocal tract configurations during the production of fricatives on various acoustic measures, it is in general agreed that the size and shape of the vocal tract determines the spectrum of a fricative (Stevens, 1998), and it is argued to be well described by the acoustic features of the *spectral peak location* and the first four *spectral moments* (spectral *mean*, *spread*, *skewness*, and *kurtosis*) (Hoelterhoff and Reetz, 2007; Jesus and Shadle, 2002; Jesus and Jackson, 2008; McMurray and Jongman, 2011; Shadle and Mair, 1996). Moreover, fricatives are not immune to co-articulation, and the articulator movements have salient acoustic consequences for the spectral energy distribution. As a consequence, the spectro-temporal trajectory has also been successfully exploited to study fine-grained differences among voiceless fricatives (Reidy, 2016). The spectral peak location is probably the most studied acoustic cue and is defined as the frequency with the highest amplitude. It has been argued that the frequency of the spectral peak is connected to the tongue movements during the production of fricatives at different places of articulation: this value supposedly decreases from high to low frequencies as the tongue moves from front to back (Hughes and Halle, 1956; Jongman *et al.*, 2000), but this could not be confirmed for Greek fricatives

(Nirgianaki, 2014). Moreover, spectral peak may serve to distinguish between sibilants and non-sibilants and, within the former, between the alveolars and palato-alveolars (Behrens and Blumstein, 1988; Heinz and Stevens, 1961; Jassem, 1965; Shadle, 1990; Strevens, 1960). Controversially, a number of studies have found a main effect of speaker and gender (Hughes and Halle, 1956; Jongman *et al.*, 2000; Nirgianaki, 2014) and of the vowel context, which influences the tongue body during the production of the fricative (Mann and Repp, 1980; Nirgianaki, 2014; Soli, 1981; Stevens, 1998). Indeed, the impact of the following vowel is stronger for [f] than for [s] and even less for [ʃ] (Stevens, 1998).

The first spectral moment is also often used and refers to the mean of the distribution of spectral energy or to the *center of gravity* of the fricative (Forrest *et al.*, 1988). Several studies show that center of gravity can distinguish between non-sibilants and sibilants and even within sibilants (Jongman *et al.*, 2000; Kochetov, 2017; Nittrouer *et al.*, 1989): higher values were found for sibilants than for non-sibilants (Tomiak, 1991) and for [s] than for [ʃ] (Funatsu and Kiritani, 1998; Jongman *et al.*, 2000; Nittrouer *et al.*, 1989; Padgett and Żygis, 2007; Zsiga, 2000). In Russian, the center of gravity was reported to be gender- and speaker-dependent, with higher values in word-initial than in word-medial positions (Kochetov, 2017).

An acoustic cue less considered in the literature is the second spectral moment, which refers to the spectral spread or variance of the energy around the mean. Spectral variance was found to be lower for sibilants and higher for non-sibilants (Jongman *et al.*, 2000; Tomiak, 1991), with the post-alveolar fricative [ʃ] having the lowest variance (Shadle and Mair, 1996).

More findings are reported for the third and the fourth spectral moments, skewness and kurtosis. Skewness describes the spectral tilt and measures the overall asymmetry of the energy distribution. A skewness of zero indicates a symmetrical distribution around the mean. A positive skewness suggests a negative tilt with a concentration of energy in the lower frequencies, and a negative skewness infers a positive tilt and a predominance of energy in the higher frequencies (Newell and Hancock, 1984; Peeters, 2004). Kurtosis refers to the "peakedness" or flatness of the distribution: spectral kurtosis equal to 3 indicates a normal distribution, while a value smaller than 3 suggests a flat distribution and a higher value stands for a "peaker" distribution (Newell and Hancock, 1984; Peeters, 2004). Several studies suggest that skewness and kurtosis may distinguish between [s] and [ʃ] (McFarland *et al.*, 1996; Nittrouer *et al.*, 1989; Tomiak, 1991). A negative skewness was found for [s] and a positive one for [ʃ] (Jongman *et al.*, 2000; McFarland *et al.*, 1996; Nittrouer *et al.*, 1989), but others report a greater positive skewness for [s] than for [ʃ] (Tomiak, 1991). For kurtosis, a large positive value was measured for [s] and a small positive or a negative one for [ʃ] (Jongman *et al.*, 2000; McFarland *et al.*, 1996; Nittrouer *et al.*, 1989; Tomiak, 1991).

Thus, multiple studies show that the spectral moments may be able to distinguish fricatives (Forrest *et al.*, 1988; Jongman *et al.*, 2000; Tomiak, 1991), but others argue that while they carry important information about fricatives, they cannot reliably distinguish their places of articulation (Shadle and Mair, 1996). On the other hand, the temporal properties of fricatives were so far much less investigated, with most studies agreeing that duration is not a distinct cue in fricatives at all (Jongman *et al.*, 2000; Kochetov, 2017) or can only contrast non-sibilants and sibilants (Behrens and Blumstein, 1988).

In terms of the predictive power found in the literature, temporal and spectral measures achieve quite a low accuracy of about 77% (Jongman *et al.*, 2000) and between about 79% and 85% (McMurray and Jongman, 2011) for English fricative place of articulation and of only about 61% for Greek fricatives (Nirgianaki, 2014). In contrast, several recent studies have focused on the extraction of cepstral coefficients on the mel scale (Kong *et al.*, 2014) or the Bark scale to describe and distinguish fricative place, voicing, and palatalization contrasts (Ghaffarvand Mokari and Mahdinezhad Sardhaei, 2020; Jesus and Jackson, 2008; Spinu *et al.*, 2018; Spinu and Lilley, 2016), achieving a much better predictive power of around 90% and higher than the traditional measures. Even fewer studies approached the identification of fricatives using machine learning, and they mostly used deep learning methods (Anjos *et al.*, 2020; Nagamine *et al.*, 2015). However, while very interesting, it is generally harder, when using such methods, to understand how the acoustic cues participate in the classification process.

III. PRIMARY DATA AND ACOUSTIC CUES

The following R packages are used for the quantitative analysis: data.table (Dowle and Srinivasan, 2019), e1071 (Meyer *et al.*, 2019), ggfortify (Tang and Horikoshi, 2016), neuralnet (Fritsch *et al.*, 2019), nnet (Venables *et al.*, 2002), recipes (Kuhn and Vaughan, 2019), randomForest (Liaw and Wiener, 2002), randomForestExplainer (Paluszynska and Biecek, 2017), recipes (Kuhn and Wickham, 2019), rpart (Therneau and Atkinson, 2019), rpart.plot (Milborrow, 2019), rsample (Kuhn *et al.*, 2019), scales (Wickham and Seidel, 2020), and tidyverse (Wickham, 2017).

A. Participants and primary data collection

The participants were 40 students (20 female) between 18 and 30 years old, studying in different departments of St. Petersburg University in Russia. These participants were born or had lived since their early childhood in St. Petersburg. No participants reported any speech or hearing impairment, and only one had to be excluded as he was a professional musician. All participants were first introduced to the purpose of the experiment, the expected duration, and the procedure. They were told that they had the right to withdraw at any time during the experiment, and they were provided with the contact details of a person who could answer all their questions concerning the research and their rights. The participants were compensated for their

participation. Demographic data, such as sex and age, were recorded before the experiment started. The recording sessions were conducted at the phonetic laboratory of the Phonetic Institute in St. Petersburg, in an audiometric booth using the recording program SpeechRecorder (Draxler and Jänsch, 2018) at a sample rate of 44.1 kHz (16-bit encoding). For the recordings, a clip-on microphone [Sennheiser (Wedemark, Germany) MKE 2-P] was placed at a distance of 15 cm from the speakers' mouth and connected through an audio interface [Zoom (San Jose, CA) U-22] to a laptop computer.

The participants were instructed to read 198 sentences from a computer screen. The stimuli were presented one by one in a pseudo-random order by the experimenter, and the participants could repeat a sentence in the case of a production error. Ninety-four real words containing one of the 12 Russian fricatives at four places of articulation, voicing contrast, and palatalization were embedded either in sentences where the fricatives occurred without contrast ($N = 94$) or as minimal pairs in carrier sentences in which the fricatives were in contrast ($N = 104$). Sentences not containing a contrasting fricative were natural-sounding language sentences, such as "his name is Sasha [salʲ]" and "I like your [ʃalʲ]" (scarf),[1] while the contrasting ones were more constrained: for example, for the minimal pair [salʲ] and [ʃalʲ], the carrier sentences were "She said [salʲ] and not [ʃalʲ]" and "She said [ʃalʲ] and not [salʲ]."[2] Some target words have two minimal pairs (for instance, the word [salʲ] is embedded in two different carrier sentences, once contrasting with [ʃalʲ] and a second time with [ʒalʲ]), explaining the higher number ($N = 104$) of carrier sentences.

B. The fricatives

The current study focuses on the differences in the place of articulation between three Russian fricatives: the labiodental [f], the dental [s], and the hard alveolar-palatal [ʃ]. The velar [x] and other voiced and palatalized fricatives were excluded for several reasons. First, while the contrast in places of articulation in Russian fricatives has been studied previously, a gap still exists in the literature (Kochetov, 2017). Studies of Russian fricatives have mostly concerned pairwise comparisons of places of articulation, such as the contrast between [s] and [ʃ], while [f] generally has not been considered so far. In terms of acoustic cues, most studies have measured noise intensity, F1, F2, F3 onset/offset, and consonant duration (Kochetov, 2017). Noise spectra have not been much considered, except in studies that involved the production from a single speaker (Bolla, 1981) or only measured the center of gravity (Kochetov, 2017). Since the documentation of Russian voiceless fricatives is rather limited, it is preferable to start with a smaller sample and go deeper in the analysis to achieve a better understanding of how different acoustic cues interact with each other in the identification of these fricatives.

Second, the velar fricative [x] was excluded, since its realisations are often very short and show strong co-articulatory effects, meaning that no or only a very short noise portion could be detected by the manual and automatic methods. Therefore, the acoustic cues could only be obtained from the raw sounds, and even there we saw a very high variation in the estimated values, suggesting that further research is needed to determine how to measure the velar [x] in a comparable way to the other fricatives. Furthermore, the occurrence of [x] is much less frequent than of the other fricatives in Russian, which makes its sample size too small to be investigated in the current controlled study.

Third, palatalized and voiced fricatives are not included to avoid interference between voicing, palatalization, and place of articulations. That is to say, by only considering voiceless non-palatalized fricatives, the current study allows a clear view of how acoustic cues interact with each other to distinguish fricatives with different places of articulation. Arguably, this strength can also be construed as a weakness, since the results shown in the current study are restricted to a certain subset of Russian fricatives, but since the current state-of-the-art is relatively limited when it comes to Russian fricatives and to machine learning, this more focused approach may be preferable (this is further developed in Sec. IV).

The final data consist of 6320 sounds: 1440 (22.7%) [f], 2680 (42.4%) [s], and 2200 (34.8%) [ʃ], each equally distributed among tokens recorded by male and female speakers (e.g., there are 720 [f] sounds recorded by males and 720 recorded by females). Due to the structure of the Russian lexicon, there are fewer [f] sounds than [s] and [ʃ].

C. Automatic and manual segmentation

The audio files were filtered below 80 and above 20 050 Hz with a smoothing of 80 Hz and were first preprocessed online automatically using the Munich Automatic Segmentation System (MAUS) (Kisler *et al.*, 2017; Schiel, 1999). Its output is a TextGrid containing, among other things, a tier with the phonetic boundaries, which was used for further manual boundary corrections, followed by the extraction of the fricatives with Praat (Boersma and Weenink, 2021). To define the onset and offset of the full consonant, the broadband spectrogram was considered as more important than the start of an aperiodic waveform with rising zero crossing rates, and in intervocalic fricatives, the presence of formant columns is defined as the onset and offset of the fricative [following Skarnitzl and Machač (2011)].

Applying this segmentation strategy means that the full segment of a fricative in an intervocalic positions will also contain part of the transition zone, with co-articulatory effects of the preceding and following sounds, as can be seen in Fig. 1. Fricatives preceded by consonants, or in the last word and sentence position, were segmented according to the presence of high energy in the spectrogram.[3]

A third segmentation step was performed to better separate the full consonant into temporal components and to

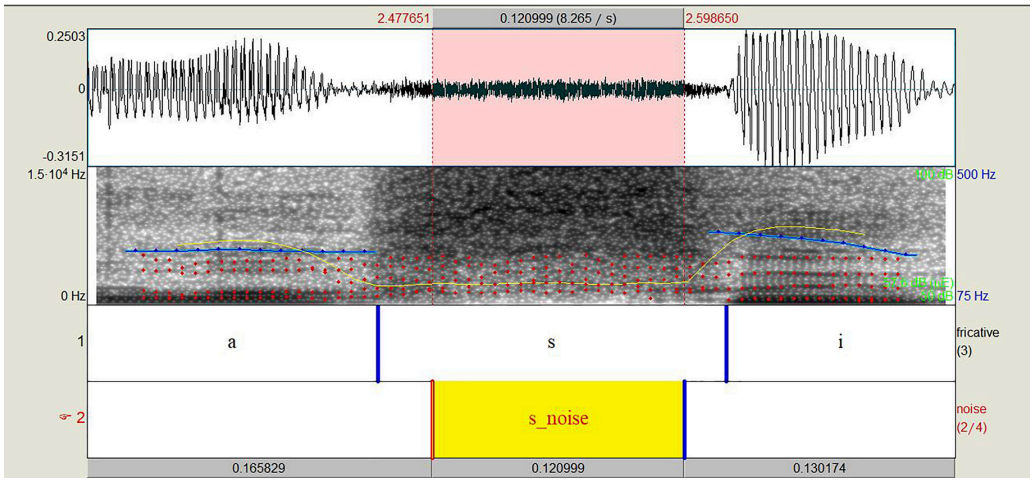

FIG. 1. (Color online) An example of a fricative sound. The first tier of this screenshot from Praat shows the full duration of the fricative, while the second shows only the noise part, excluding the effects of any potential co-articulation.

extract the relevant frication noise portion of the sound. As shown in Fig. 1, the oscillogram of the full duration of the consonant is not equal to the pure noise part of the fricative. Noise is in general defined as an aperiodic signal with high frequencies and therefore a high number of zero crossings in a given time, i.e., a high zero crossing rate (*zcr*). This is known to detect the voiced and unvoiced parts in speech, and we used it here to detect the frication noise part in fricatives. To visualize the number of zero crossings in Praat, a PointProcess object[4] was generated, as shown in Fig. 2.

The blue bars represent the points where the waveform passes through zero, and the noise parts of the fricative are characterized by the high density of the blue bar (appearing almost as a solid blue rectangle), while the gaps between the blue bars at the beginning and the end of the sound indicate fewer zero crossings, which can arise from co-articulatory effects. Our data show that, in connected speech, the distribution of zero crossings along the sound duration depends to some degree on linguistic and non-linguistic factors, such as co-articulation, stress, or speaker-specific production characteristics. Furthermore, many sounds did not show a clear

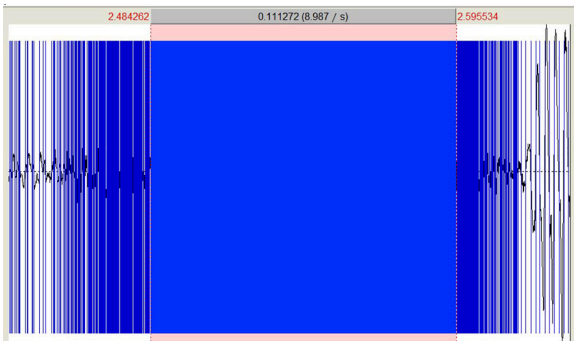

FIG. 2. (Color online) Visualizing the zero crossings in Praat. The increase in the spatial density of the blue bars shows a rapid increase in *zcr*.

middle noise portion without any interruption, in which case no all-encompassing rule could be applied and, to detect the relevant region, each token had to be considered individually, explaining why the segmentation of the noise part is very time-consuming and resists full automatisation and standardisation.

To overcome these difficulties and allow the full automatisation of the extraction of the noise part, we introduce here a new method based on training a tree-based computational classifier, built on the assumption that the zero crossing rate provides sufficient information to divide a speech signal into a purely aperiodic portion and portions containing periodics. With this model, each sound is separated into different *windows* based on a certain amount of zero crossing points. The *zcr* within each window is then measured and compared with the zero crossing rate of the preceding window (if any). The difference of zero crossing rate between the two windows (*diff*) is then computed and used as a cue to identify the beginning and the end of the noise part of a sound. Typically, we expect that a rise of zero crossing rate across two windows indicates the beginning of the noise, while a drop of zero crossing rate across two windows represents the end of the noise. To have a better understanding of which settings are optimal for the model, we tested different window lengths (here, 64, 128, 256, or 512 points) with different levels of overlap (0%, 30%, 50%, or 80%); please note that the window lengths are considered in terms of number of zero crossings and do not represent the window's absolute duration in terms of wall-clock time, as the same number of zero crossings may cover different absolute durations for different sounds.

A "gold standard" subset of 560 fricative sounds, which had their noise duration identified manually, was used to annotate each window with noise = TRUE or noise = FALSE depending on its occurrence within or outside the noise part identified manually. For the sake of argument, let us consider

https://doi.org/10.1121/10.0005950

a recording of a certain length, within which there is only one manually annotated noise part that starts at t_s seconds and ends at t_e seconds. Each possible window is annotated with a unique time mark, t_i, representing the moment at which the window starts; if, for a particular window i, this time mark falls between the starting time and the ending time of the manually annotated noise of the sound ($t_s \leq t_i \leq t_e$), the window is marked as noise = TRUE, but if the time mark is found before the starting time ($t_i < t_s$) or after the ending time of the noise ($t_e < t_i$), the window is marked as noise = FALSE. This procedure ensures that each window within each of the sounds is annotated as noise = TRUE or noise = FALSE, annotations that are used for training a tree-based computational classifier (Breiman *et al.*, 1984) to identify the TRUE or FALSE value of each window based on the gap of zero crossing rates between two consecutive windows.

The classifier was trained on a randomly chosen 70% of the data (the "training subset") and evaluated on the remaining 30% of the data (the "test subset"). The random splitting of the "gold data" into the "training" and "test" samples was repeated 100 times. For each of these 100 training/test samples (replications), we evaluated all the possible combinations of window length and overlap so as to identify which of them generate the highest accuracy at identifying the noise parts of the sounds. We thus estimated a total of 4 lengths × 4 overlap values = 16 possible combinations of parameters, which were replicated 100 times each, resulting in a total of 1600 replications. An example decision tree for window length 512 and 50% overlap is shown in Fig. 3.

The overall performance of the classifier is measured by its *accuracy*, which is equal to the percentage of the correctly classified windows out of the full set of windows (e.g., if a sound is segmented into ten windows and the model classifies correctly seven of them, the accuracy of the model is $7/10 = 70\%$). A summary of the accuracy of each of the 16 possible combinations of window lengths and overlaps is shown in Fig. 4, where each boxplot represents the distribution of the accuracies of the 100 replications of the corresponding combination of parameters.

We see that all combinations of parameters result in accuracies between 78% and 83%, with the best accuracy being found for a large window length (512 zero crossings) and a standard overlap (50%), with mean = median = 80.8% across the 100 replications.[5] It is important to note that these models are much more accurate than the "majority baseline," which is equal to what would be obtained by conducting a deterministic allocation of all the data points into the majority category (please see below for more details). For our best parameters (window length = 512 and overlap = 50%), the majority baseline is equal to the share of the TRUE sound segments in the data, i.e., 39 842/61 485 = 64.8%, but the accuracy of the model (80.8%) is much higher than this. Thus, the sound segments classified by this model can then be used for the extraction of acoustic cues.

However, in general, 80% is far from excellent performance and can only be considered as good. Therefore, we also conducted a brief analysis of the performance of the classifier for the noise classification task:[3] the closer analysis of the errors generated by the classifier indicates that the predictions of the classifier tend to wrongfully predict windows without noise as having noise, which is to say, the model predicts noise parts that are larger than the actual noises.

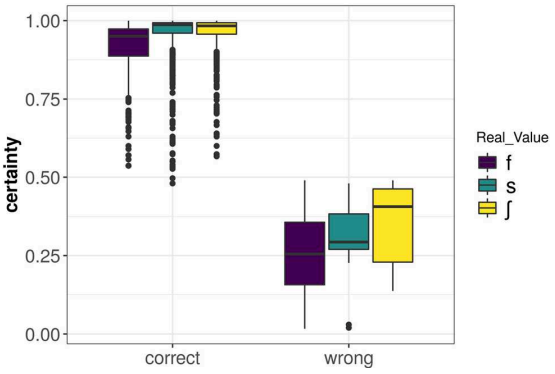

FIG. 3. (Color online) A decision tree generated for window length = 512 points and 50% overlap. *zcr*, zero crossing rate; *diff*, the gap of zero crossing rate between two consecutive windows. A positive *diff* value represents an increase in the zero crossing rate, while a negative value refers to a decrease in the zero crossing rate. The values Sound_TRUE and Sound_FALSE refer to the presence of noise in a window: a window with Sound_TRUE is located within the noise part of the sound, while a window with Sound_FALSE is not. Such a tree is interpreted as follows: the color of the rounded rectangles ("buckets") at the bottom of the tree represents the ratio of correctly predicted TRUE/FALSE value of noise, with the numbers within showing the number of tokens classified as such (the denominator) and, of those, which were correctly identified (the numerator). The prediction for a given token starts from the top node and ends in a bucket at the bottom of the tree. For instance, starting from the top node 1, if $zcr < 0.14$, the segment is interpreted as noise = FALSE; this path classifies 2525 tokens as noise = FALSE, among which 2042 are correctly identified as noise = FALSE, resulting in an accuracy of 2042/2525 = 80.9% for this prediction. As another example, if the $zcr \geq 0.14$ and if the gap of zero crossing rate with the previous sound ranges between –0.036 (node 3) and 0.05 (node 7), the sound segment is interpreted as noise = TRUE. This path classifies 6858 tokens as noise = TRUE, of which 5688 are classified correctly, resulting in an accuracy of 5688/6858 = 82.9%. The same logic applies for the other branches of the tree. The variables that are shown in the decision tree are the variables considered to have statistically significant explanatory power given the data, while the variables not shown are considered to not help in identifying the TRUE/FALSE value of the windows; here, both *zcr* and *diff* are relevant.

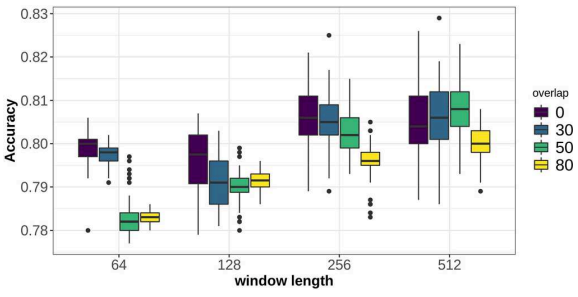

FIG. 4. (Color online) The accuracy of the classifiers trained with different parameters of window length and percentage of overlap (add percentage in graph). Each combination of parameters is trained and tested for 100 replications with different training and testing data.

These errors are equally frequent at the beginning and at the end of the noise parts of a sound. Furthermore, the windows from the [f] sounds seem harder to classify, as the accuracies for the three sounds are [f] = 75%, [s] = 85%, and [ʃ] = 80%, which is not surprising given that [f] typically has a shorter noise duration. Additional tuning of the parameters (such as window length and overlap) may help to further improve the performance of the classifier, but this goes beyond the aims of the current study, whose main goal is to investigate whether machine learning may improve distinguishing fricatives. Further discussions can be found in Sec. V.

D. Acoustic cue definition and extraction

To extract acoustic cues, most studies use single spectral slices from the middle and sometimes the beginning and end of the fricative or of the frication noise, with window sizes between 25 ms (Kochetov, 2017) and 40 ms (Jongman *et al.*, 2000). The extraction of the acoustic cues for fricatives is generally not conducted on the full duration of the consonant. Because here we want to both follow the examples of previous studies and develop new machine learning methods, we combined two dimensions for pre-processing the sound files to subsequently extract the acoustic cues.

For the acoustic analysis, two data sets were used. The first data set includes the whole corpus of 6320 sounds (denoted in the following as "A" = all sounds): there are 1440 [f] sounds (22.7%), 2680 [s] sounds (42.4%), and 2200 [ʃ] sounds (34.8%). The second data set is the subset of 6068 sounds for which a frication noise window of minimum 30 ms could be detected by applying the above mentioned automatic noise detection strategy (denoted as "N" = noise sounds); thus, 252 sounds were discarded [[f] = 171 (2.7%), [s] = 5 (0.08%), [ʃ] = 76 (1.2%)]. To extract the acoustic cues, four regions of the fricative are considered: (a) the full consonant duration derived from the manual segmentation (denoted as "C" = consonant), (b) the identified frication noise duration from the automatic segmentation ("F" = frication), (c) the 30 ms window placed in the middle of the consonant ("W" = window), and (d) the 30 ms window place in the middle of the frication noise ("M" = middle). Combining these two dimensions results in six *acoustic cue extraction techniques* (ACETs): first, extracting the acoustic measures from the whole corpus ("A"; 6320 tokens), using (i) the full consonant duration ("AC") or (ii) the middle 30 ms ("AW") and, second, extracting the acoustic measures from the "N" subset (6068 tokens), using (iii) the full duration of the consonant ("NC"), (iv) the frication noise ("NF"), (v) the 30 ms window placed in the middle of the sound ("NW"), or (vi) the 30 ms window placed in the middle of the frication noise ("NM") (Table I).

Table II shows the acoustic cues extracted for this study. All measures were extracted using Praat (Boersma and Weenink, 2021) and standard settings. The spectral measurements *central peak location* (*peak*) and the four *spectral moments* (*cog*, *sdev*, *kurt*, *skew*) are the most commonly used cues for fricatives and are discussed above.[6] In the temporal domain, we measured the *zcr* and the *duration of the entire consonant* (*dur*). Furthermore, 13 MFCCs from the middle 30 ms of the sound were extracted.

TABLE I. The six theoretically possible acoustic cue extraction techniques (ACETs). The abbreviations shown in each cell are used to refer to each ACET within the following text. The first letter (A/N) of the abbreviation refers to the data sample used for the extraction of acoustic measures (all sounds/noise sounds), and the second letter (C/F/W/M) indicates the considered region of each sound (full consonant duration/frication noise duration/middle 30 ms of duration/middle 30 ms of noise).

	All sounds (A)	Noise sounds (N)
Consonant duration (C)	AC	NC
Middle 30 ms of duration (W)	AW	NW
Frication duration (F)		NF
Middle 30 ms of frication (M)		NM

Figure 5 compares the main acoustic cues computed using the three ACETs.[3] It can be seen that the acoustic cues behave differently across ACETs, with, for example, [f] showing more variation for *cog* and *skew* than the other sounds. Likewise, there is variation in the acoustic cues between the sounds, the most variable being *cog*, *peak*, *sdev*, and *zcr*.[3]

We also conducted a principal component analysis (PCA) to visualize the relationships between the acoustic cues. PCA is a technique used for unsupervised dimension reduction (Jolliffe, 2002). Because multidimensional data often include variables that are correlated, it is preferable to transform them before applying other types of analysis. PCA transforms the correlated input variables into a set of uncorrelated principal components (PCs) derived from them and explaining the same variation. The PCs are ordered decreasingly in terms of the amount of variation in the data they explain (thus, PC1 explains most of the variance, PC2 explains most of the remaining variance, and so on). Figure 6 shows the data projected on the PC1 (x axis) and PC2 (y axis), which explain together 96.52% of the variance. 77.66% of the variance is explained by PC1, which is mostly driven by *zcr*, *cog*, and *peak*, and 18.86% is

TABLE II. Summary of the acoustic cues included in the present study.

Cue	Variable	Description
Fricative duration	*dur*	Duration of the entire sound obtained from manual segmentation
Zero crossing rate	*zcr*	Number of times the wave crosses 0, computed for each time frame of the signal
Peak frequency	*peak*	Frequency of the highest amplitude
Peak amplitude	*peak_a*	Amplitude of the highest frequency
Spectral mean	*cog*	Mean distribution of spectral energy (center of gravity)
Spectral variance	*sdev*	Spectral spread or variance of the energy around the mean
Spectral skewness	*skew*	Spectral tilt, overall asymmetry of the energy distribution
Spectral kurtosis	*kurt*	Spectral flatness of the distribution

JASA
https://doi.org/10.1121/10.0005950

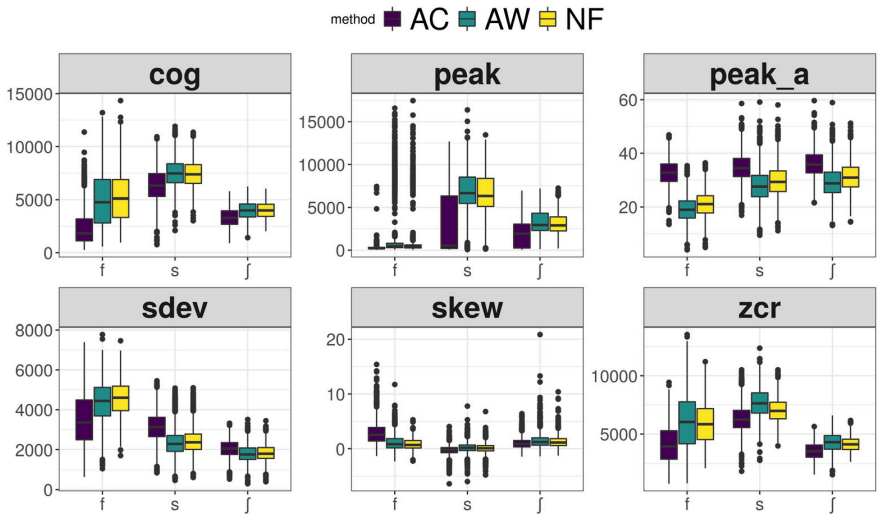

FIG. 5. (Color online) The comparison of acoustic cues based on the three main ACETs reported in the experiments. The names of the ACETs refer to the acoustic cue extraction techniques listed in Table I.

explained by PC2, which is driven mostly by *peak*, *zcr*, *cog*, and *sdev*.

The clusters of [s] and [ʃ] sounds generally stand out from each other, which implies that the classifiers will probably not have difficulty in differentiating those two sounds based on their acoustic cues. On the other hand, the tokens of [f] are a bit blurred with the [s] and [ʃ] sounds. This shows that [f] sounds may represent some difficulty for the classifiers.

IV. PREDICTING FRICATIVES FROM ACOUSTIC CUES

Four computational classifiers were used to predict fricatives from acoustic cues. The information about the sex of the speakers as well as their unique (anonymous) identifiers was also provided to the classifiers to assess their potential relevance to the classification of fricatives.[7] The first two are based on binary recursive partitioning (Breiman *et al.*, 1984): the first classifier generates a single *decision tree* based on the data and helps visualize the interactions between the variables (incidentally, we also used such a classifier above for sound filtering).

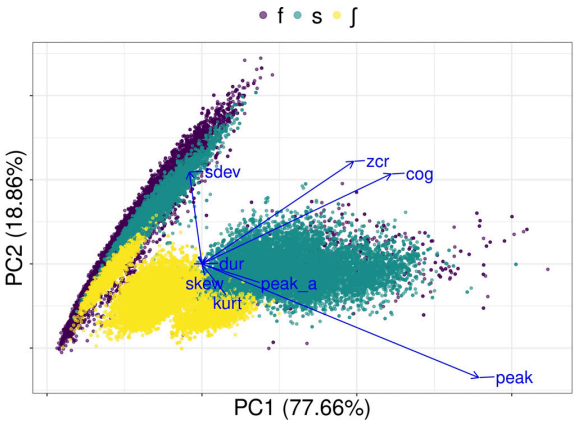

FIG. 6. (Color online) The PCA visualization of the acoustic cues for each sound. The length of the arrows relates to how much information is contributed by the acoustic cues to the PCs. *cog*, *peak*, *sdev*, and *zcr* are the most relevant.

The second, called a "random forest" (Breiman, 2001), generates a series of 300 decision trees[8] that are analyzed as a whole and used to assess the importance of each variable with regard to correctly predicting the fricatives. For each tree, it uses a bootstrap sub-sample of observations and a random subset of the variables from the entire dataset. This process of random sampling is also the main strength of random forests, as it allows the analysis of small-scale data and consideration of the possible auto-correlation of variables (Tagliamonte and Baayen, 2012).

The third classifier is called "support vector machines" (SVMs), which are able to separate subsets of the data even when the separation boundary is not linear.

The fourth classifier uses a neural network architecture (Haykin, 1998; Parks *et al.*, 1998), which searches for non-linear boundaries between the data points. Here, we use a feed-forward neural network that consists of an input layer, a hidden layer, and an output layer, each layer having a specific number of neurons that are connected to the neurons of the next layer. The input layer has one neuron for each variable (predictor) in the classification task, while the output layer has one neuron for each type of predicted sound. The hidden layer is set to ten neurons in the current experiment.

We chose these four classifiers for the following reasons. The first classifier generates an explicit decision tree that captures the hierarchical interactions of the variables within the dataset. The second classifier provides information about the relative importance of the predictors. The third and the fourth classifiers are among the best at dealing

with complex non-linear problems, at the cost of an easy understanding of the decision process. The interest of comparing these four classifiers is in trying to find the best trade-off in terms of transparency and performance for the classification of fricatives based on acoustic cues.

All classifiers were trained on 70% of the data (the training subset), and their accuracy was evaluated on the other, non-overlapping, 30% of the data (the test subset). Importantly, both the training and testing subsets have the same frequency of the predicted sounds as the full dataset (e.g., as [f] appears 1440 times in the data, that is, 1440/6320 = 22.8% of the time, the subsets each contain about 22% [f] sounds). To be able to generalize the results, we ran ten replicates, each with the data randomly partitioned into such training and testing subsets.[9]

The performance of the computational classifiers was captured using three measures: *accuracy*, *precision*, and *recall*. Accuracy provides an overview of the performance on the entire dataset, and it is the proportion of all correctly classified sounds. Its value should be compared with an appropriate baseline. One such baseline would be the accuracy of a model that makes completely random guesses; here, this random baseline would be equal to the square of the proportion of each sound in the data, i.e., $(1440/6320)^2 + (2680/6320)^2 + (2200/6320)^2 = 35\%$, and if our model surpasses this baseline, it would be considered as performing better than chance. However, the random baseline is easily affected by the different sizes of each category in the data, prompting us to use the *majority baseline* as our threshold. This baseline deterministically allocates all sounds to the biggest category in the dataset: since [s] appears in the most tokens in our data (42%, 2680/6320), such a classifier would reach a precision of 42% just by guessing that all the sounds are [s], so that the accuracy of our classifiers should be greater than 42%. The majority baseline is by default at least as good as the random baseline, making it harder to beat and more reliable for evaluating the accuracy of classifiers.

However, accuracy gives only a general idea of the performance of the model, and to have a more precise idea as to how the classifier performs for each sound, we also considered *precision* and *recall* (Ting, 2010). Precision quantifies how many of the sounds classified in each category are correctly classified (e.g., how many of the sounds classified as [f] are actually [f] sounds). Recall quantifies how many of the sounds actually belonging to each category are correctly classified (e.g., how many [f] sounds are correctly classified as [f] sounds by the classifier). Precision and recall are computed for each of the three fricatives, resulting in three estimates of precision and three of recall in total.

We now analyze the results of each of the four classifiers in turn.

A. Single decision tree

The mean output of the 10 replications is shown in Table III. The accuracy does not vary much between the ACETs, as the maximum is 94.6% and the minimum is 93.0%, but the accuracy of NF is consistently the highest.[3] The precision and recall are generally high for all sounds across the ACETs, without much systematic variation.

Focusing on NF, the accuracy is similar across the replications, and we show in Fig. 7 the decision tree generated on the first replication. This tree is to be interpreted in the same way as in Fig. 3 and shows that *cog* and *sdev* are sufficient for the classifier to distinguish between [f], [s], and [ʃ]. For instance, if *cog* is high ($\geq$5486, node 1 to node 2) and *sdev* is also high ($\geq$4002, node 2 to node 4), the classifier predicts an [f] sound, while if *cog* is low (<5486, node 1 to node 3) and *sdev* is also low (<2803, node 3 to node 7), the classifier predicts an [ʃ].

TABLE III. The performance of the classifiers across ten replications ranked according to their mean accuracy. The names of the ACETs refer to the acoustic cue extraction techniques listed in Table I. The baseline indicates the majority baseline. Acc., accuracy; upper, upper confidence interval; lower, lower confidence interval; Pr., precision; Rc., recall. Please note that the slight variation in the accuracy of the majority baseline is due to variations in the dataset size (NF has fewer tokens than AC since the former is only considering the sounds that were detected with noise parts). The values in bold indicate the parameters with the highest accuracy for each classifier.

Classifier	ACET	Baseline (%)	Mean Acc. (95% CI) (%)	Pr. [f] (%)	Rc. [f] (%)	Pr. [s] (%)	Rc. [s] (%)	Pr. [ʃ] (%)	Rc. [ʃ] (%)
Single tree	MFCC	42.4	93.5 (93.1–93.9)	90.4	89.0	92.3	93.1	97.9	96.9
Single tree	AW	42.4	94.6 (94.3–94.9)	91.0	93.4	96.7	93.4	94.7	96.9
Single tree	AC	42.4	93.0 (92.8–93.3)	85.9	94.5	94.6	91.8	96.4	93.4
Single tree	NF	44.1	**94.9 (94.6–95.1)**	92.6	91.9	96.1	94.3	94.7	97.3
Random forest	MFCC	42.4	**98.5 (98.3–98.7)**	97.6	96.8	98.2	98.6	99.5	99.6
Random forest	AW	42.4	97.4 (97.2–97.6)	96.2	96.9	97.1	97.1	98.1	97.7
Random forest	AC	42.4	97.3 (97.0–97.5)	96.2	97.1	97.0	96.9	98.2	97.8
Random forest	NF	44.1	97.7 (97.4–97.9)	97.1	96.4	97.8	97.3	97.9	99.0
SVM	MFCC	42.4	**99.6 (99.5–99.7)**	99.2	99.2	99.6	99.6	1.00	1.00
SVM	AW	42.4	98.0 (97.8–98.2)	96.8	96.6	97.9	97.7	98.9	99.3
SVM	AC	42.4	98.2 (98.0–98.3)	97.6	97.7	98.2	97.6	98.5	99.2
SVM	NF	44.1	98.5 (98.3–98.7)	98.1	97.0	98.4	98.4	98.8	99.6
Neural net	MFCC	42.4	**99.5 (99.4–99.6)**	99.1	99.0	99.4	99.4	99.8	99.9
Neural net	AW	42.4	97.7 (97.4–98.1)	96.8	96.7	97.4	97.4	98.1	98.3
Neural net	AC	42.4	97.8 (97.5–98.1)	97.9	96.9	97.4	97.7	98.2	98.4
Neural net	NF	44.1	98.1 (97.8–98.4)	96.9	97.0	98.3	98.0	98.5	98.8

https://doi.org/10.1121/10.0005950

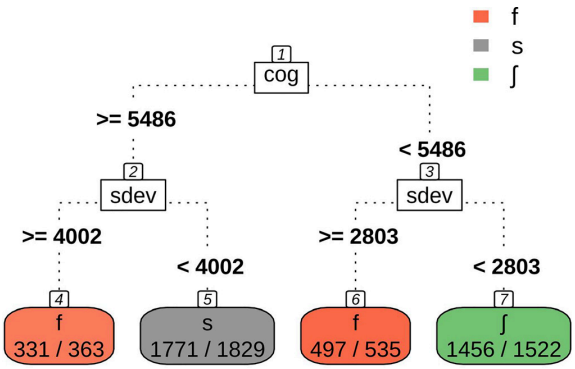

FIG. 7. (Color online) A decision tree generated with the acoustic cues from the ACET of NF. The rules for its interpretation are similar to those of the tree in Fig. 3 except for the color of the "buckets," which now represent the sound.

Interestingly, only *cog* and *sdev* matter, while the other variables (such as *zcr*, *dur*, and even speaker information) are considered as not relevant by the model. This suggests that the information captured by *cog* and *sdev* does not vary much across speakers (see also Sec. V).

Finally, the confusion matrix generated by this decision tree on the testing subset is shown in Table IV. It can be seen that, for example, the testing set includes $358 + 15 + 7 = 380$ [f] sounds and that the classifier predicted $358 + 36 + 6 = 400$ sounds as [f] sounds, correctly predicting 358 [f] sounds, while 36 were in fact [s] and six were in fact [ʃ]. Of the actual [f] sounds, 358 were predicted correctly, while 15 [f] sounds were misjudged as being [s] sounds and seven [f] sounds were misinterpreted as [ʃ] sounds.

To sum up, the single tree classifier performs generally well on the data and reaches similar performances across the three ACETs, but NF consistently ranks first in terms of accuracy.[3] Focusing on one such tree shows that *cog* and *sdev* are the most relevant variables for identifying fricatives, a finding supported by the other trees, which all converge in that *cog* is always at the root, and the two following branches depend on *sdev*.

B. Random forest

The accuracy of the random forest classifiers is shown in Table III, and we can see that, in general, the accuracy is better when compared to the single decision trees across all ACETs, all performing comparably well (accuracy between 97.7 and 97.3). NF has a better accuracy than the other ACETs. However, its accuracy is lower than MFCC-based extraction.

TABLE IV. The confusion matrix generated from the decision tree in Fig. 7. The columns indicate the actual values, and the rows refer to the predictions of the classifier. The values in the matrix are from the test set used to evaluate the accuracy of the classifier, which represents approximately 30% of the data.

	[f]	[s]	[ʃ]
[f]	358 (19.7%)	36 (1.9%)	6 (0.3%)
[s]	15 (0.8%)	741 (40.7%)	10 (0.5%)
[ʃ]	7 (0.4%)	25 (1.4%)	621 (34.1%)

TABLE V. The acoustic cues ranked on their importance as estimated by minimal depth, mean decrease in accuracy, and purity. These numbers are based on acoustic cues from the NF data.

Ranking	Minimal depth		Accuracy		Purity	
1	*cog*	2.3	*sdev*	56.9	*cog*	625.5
2	*peak*	2.4	*cog*	38.8	*sdev*	516.0
3	*sdev*	2.4	*peak_a*	31.5	*zcr*	483.5
4	*zcr*	2.6	*skew*	27.3	*peak*	437.8
5	*peak_a*	2.8	*zcr*	26.3	*kurt*	167.3
6	*skew*	2.9	*peak*	24.6	*peak_a*	164.1
7	*kurt*	2.9	*kurt*	23.2	*skew*	120.0
8	*dur*	3.4	*dur*	14.6	*dur*	38.8

Random forests allow the estimation of the importance of each predictor. Here, we used three measures: *minimal depth*, the decrease in *accuracy*, and *node purity*. The *minimal depth* of a variable indicates how far from the root node is the first node where that specific variable matters (for example, in Fig. 7, *cog* appears at the root node, having thus a minimal depth of zero). A variable frequently close to the root node (thus, with a low minimal depth) is considered to have a high importance. Table V shows the ranked importance of the acoustic cues in terms of minimal depth, of the mean decrease in the accuracy of the model when excluding a variable (a high decrease means that the variable has predictive power), and of the mean decrease in the *purity* (the Gini coefficient), indicating how the variable contributes to the homogeneity of the nodes at the bottom of the tree (a high drop in the purity when removing the variable suggests strong predictive power). While different measures result in slightly different rankings, there is a high degree of consistency, with *cog* and *sdev* being ranked in the top three most important variables.

Figure 8 shows how "consistent" the model is when making decisions, estimated as the probability of the votes

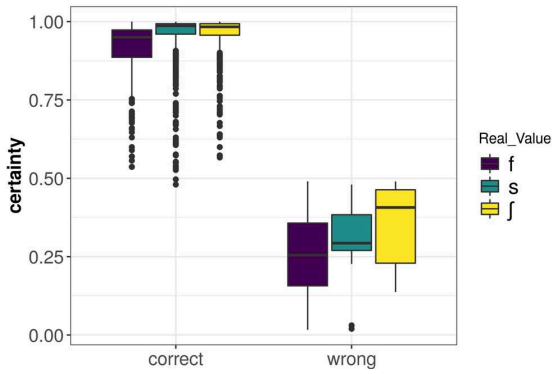

FIG. 8. (Color online) The confidence of the random forest classifier for correct and wrong decisions across [f], [s], and [ʃ] for NF.

across all the trees considered (e.g., if 270 of the 300 trees assign a token to [f], then the confidence of the decision is 270/300 = 90%). We can see that the model generally has a confidence level >85% for correct decisions and ≈35% for the wrong ones, indicating that the model is "confident" about decisions that turn out to be correct but that it also "knows" that a decision is likely to be wrong when it actually is wrong.

In sum, the results of the decision trees and of the random forests with 300 trees converge in identifying a set of variables considered important for predicting fricatives in Russian. The accuracy of these two classifiers exceeds by far the majority baseline.

C. Support vector machine

Again, we perform ten replications using randomly selected training and testing subsets. Their mean output is shown in Table III, and, as for the tree-based classifiers, the accuracy is quite similar across the ACETs (between 98.5% and 98%). NF also has the highest accuracy.

The accuracy of the SVMs is higher, on average, by ≈1% compared to the random forests, showing that the tree-based classifier already captures most of the information encoded in the acoustic cues.

D. Neural networks

Once more, we use ten replications, and their mean output is shown in Table III. As for the tree-based classifiers and the SVM, the accuracy of the ACETs does not vary much (between 98.1% and 97.7%), NF again has the highest accuracy, but the differences between ACETs are extremely small (0.4%).

Thus, the different classifiers have very comparable performances, reaching extremely high accuracies across the ACETs, showing that there is enough information in the acoustic cues to correctly classify the fricative sounds [f], [s], and [ʃ]. Interestingly, NF seems to (very slightly) outperform the other ACETs, suggesting that focusing on the extracted noise may provide the best information for classifying fricatives.

V. DISCUSSION AND CONCLUSIONS

This paper has four (five) inter-related main aims, two substantive and two methodological. Substantively, we wanted (i) to check whether using the entire sound, only a fixed-duration window in the middle of the sound, or only the noise part makes any difference to the amount of useful information contained in the extracted acoustic cues and (ii) to investigate whether conventional acoustic cues do, in fact, contain enough information to correctly classify fricatives, despite previous claims to the contrary. Methodologically, we tested whether four different computational classifiers (decision trees, random forests, support vector machines, and feed-forward neural networks with backpropagation) are capable of (iii) identifying the noise part of a fricative sound using only basic acoustic information and (iv) correctly classifying the Russian fricatives [f], [s], and [ʃ] using acoustic cues. Finally, (v) we compare the predictive power of the acoustic measures with that of the MFCCs.

Starting with aims (i) and (iii), we defined three ACETs using either the full consonant duration (AC), its middle 30 ms (AW), or only the noise part of each sound file (NF) (the noise detection used our classifier-based method).

We found that the accuracy of classifying the fricatives from acoustic cues does not vary much among these ACETs or among the four classifiers, but differences do exist and are informative scientifically and methodologically. All four classifiers perform far above the majority baseline of 44% accuracy (reaching between about 93% and 98% across ACETs). The accuracy of the decision trees is generally lower than of the other three classifiers (as expected, given that this has the simplest architecture), but, importantly, random forests perform almost at ceiling; this result is potentially very important as there is a high interpretability of the decision rules used.

In particular, extracting acoustic measurements from the full noise duration seems better than from a 30 ms window (e.g., for *cog*, *sdev*, and *peak*) for all three fricatives and especially for [f]. That is to say, the most invariant parameters are the ones estimated from the largest section that does not show strong co-articulatory effect. Therefore, we suggest that, depending on the main aim of the investigation, future work should extract acoustic measurements from the full noise duration instead of from a small spectral slice, more so if non-sibilants are the focus of the study. Similarly, the method we propose can also be useful for studying fricatives with secondary features such as palatalization (as in Russian) or aspiration (as in Korean). In both cases, clearly identifying the frication noise section can be crucial for identifying the phoneme (Rabha *et al.*, 2019).

The ACET NF does not include the speech sounds where the noise portion was absent or too short to be detected by the automatic segmentation, resulting in only 6068 tokens being retained (of the 6230 in total), allowing us to test the potential impact of such errors on the detection of the fricatives. Most such errors were found in the realization of [f], but it is unclear whether this can be generalized to other datasets. This prompts us to suggest that production errors should be carefully checked and probably excluded from the analysis; if the higher error rate for [f] is a general feature, then this might be particularly relevant for studies of contrasting front non-sibilant fricatives as is, for example, the case for English. Furthermore, while our study is relatively well powered in terms of number of tokens per speaker and the set of speakers, it might be the case that smaller samples, as typically used in previous studies, do not have the power to extract the useful information from the noise.

Focusing now on (ii) and (iv), we think that our study clearly shows that acoustic cues do contain enough information for the correct classification of the Russian fricatives [f], [s], and [ʃ], in particular, and gives hope that this may

https://doi.org/10.1121/10.0005950

be the case for other fricative sounds in other languages. A few acoustic cues seem to be necessary and sufficient, including *cog*, *sdev*, and possibly *zcr* and *peak*. The importance of *sdev* echoes previous studies emphasizing the importance of dynamical features and spectro-temporal variations in identifying fricatives (Patil and Rao, 2008; Reidy, 2016). Interestingly, the vowel context does not seem to matter, as is also the case for the speaker's sex and identity, suggesting that we may have identified *context-independent characteristics* of the fricative sounds themselves beyond and above the effects of phonetic context (Mann and Repp, 1980; Nirgianaki, 2014; Soli, 1981; Stevens, 1998) and of sex and other individual-specific factors (Hughes and Halle, 1956; Jongman *et al.*, 2000; Kochetov, 2017; Nirgianaki, 2014).

Concerning (v), as shown in Table III, our results did not find a large difference in predictive power between the acoustic measures and the MFCCs, strikingly smaller than that reported in the literature. In fact, while the MFCCs perform better than the acoustic measures (formally, statistically significantly so), this difference is very small in terms of effect size (less than 2% accuracy), both performing effectively at ceiling (above 97% for random forests, SVMs, and neural nets), and this difference is smaller when the full frication noise is used. (The fact that such small real-world differences are statistically significant here is due to the very small variation between replications.) Thus, both methods are very good and comparably so at classifying the sounds [f], [s], and [ʃ], showing that the information necessary for correctly classifying these three fricatives can be extracted in several manners. We also considered the performance of models trained with both acoustic cues and MFCCs.[3] While the results indicate that merging acoustic cues and MFCCs does not result in a better performance than the MFCCs, the ranking of the variables represents a mix between acoustic cues and MFCCs, suggesting that further studies should investigate how such acoustic cues are captured by the MFCCs. More precisely, it is not possible at this point to determine whether the absence of improvement observed when both acoustic cues and MFCCs are considered is due to the simplistic merging approach or to a ceiling effect related to the somehow limited variability offered by our corpus. The choice of which manner to use should therefore depend on the particular research question or practical application at hand, each having its advantages and disadvantages: the MFCCs are probably more appropriate in an engineering context, while the acoustic measures give more insight into the articulatory and perceptual mechanisms relevant for fundamental research.

It is perhaps important to note that our approach here is to use the acoustic cues to classify the fricative sounds, identifying, in the process, those cues that matter the most, in contrast to, for example, McMurray and Jongman (2011), which, within a regression framework, tries to find statistically significant differences for a cue given the type of fricative sound. We replicated and extended the methodology in McMurray and Jongman (2011) using a maximum-likelihood mixed effects regression approach where the value of given cue is predicted from the *method* (the ACETs), the sound *classification* ([f], [s], or [ʃ]), and their interaction as the predictors of interest, controlling for sentence *type* (carrier or normal sentence), fricative *position* (beginning, middle, or end), the sounds *preceding* and *following* the fricative (several classes), and *sex* (F/M) as fixed effects and for *sentence* and *speaker* as random effects (sentence embedded within speaker). In a nutshell, our findings[3] suggest that, as expected, there is a high similarity within speakers and sentences for all cues (high intra-class correlations) and that there are significant differences between sounds for all cues, with varying influences of sentence type, fricative position, and context but, again, not of sex. While they are concordant with our machine learning results and confirm that, indeed, acoustic cues differ between fricatives, these results cannot be directly used to *classify* fricatives *from* acoustic measures as our classifiers do, which, arguably, is the relevant question both scientifically and practically.

Comparing our results of spectral and temporal cues with the previous findings, we find both overlaps and differences. *Spectral peak location* is probably one of the most promising cues in the literature, but our classifiers did not find it as crucial for distinguishing fricatives. As for Greek fricatives (Nirgianaki, 2014), we do not find a clear decrease in frequency as the place of articulation moves from front to back, in opposition to other previous research (Hughes and Halle, 1956; Jongman *et al.*, 2000). In our data, *cog* is the most important cue for distinguishing [f], [s], and [ʃ]. Higher values are reported for sibilants than for non-sibilants (Tomiak, 1991) and for [s] than for [ʃ] (Funatsu and Kiritani, 1998; Jongman *et al.*, 2000; Nittrouer *et al.*, 1989; Padgett and Żygis, 2007; Zsiga, 2000), which our data confirm, to a certain extent: [f] has the lowest values around 4000 Hz (but reaching up even above 7000 Hz), while the energy of [s] is centered around 7500 Hz and that of [ʃ] is centered around 4500 Hz.

Despite the *spectral spread* being much less considered in the literature, we found that this is one of the most important cues in our data: the lowest spread was found for [ʃ] and the highest for [f] (Jongman *et al.*, 2000; Shadle and Mair, 1996; Tomiak, 1991).

For the other two spectral moments, *skewness* and *kurtosis*, our results did not match with previous findings suggesting that these two cues are stable characteristics of fricatives (McFarland *et al.*, 1996; Nittrouer *et al.*, 1989; Tomiak, 1991). Not only there are no significant differences across the methods, but both measures are plagued by many outliers.

Temporal measures, such as the full consonant duration and the frication noise duration, are not distinct cues in our data. Only the zero crossing rate seems to contain relevant information, but it is not an important cue for distinguishing [f], [s], and [ʃ].

Our study has several limitations, probably the most important being that we are focusing here only on a subset of the Russian fricative inventory of read speech.

Nevertheless, we believe our study is a potentially important contribution to several current debates in phonetics and linguistic typology and to the application of machine learning techniques to acoustic studies. First, it found that there may be a set of acoustic cues (*cog* and *sdev*) that can reliably distinguish the Russian fricatives [f], [s], and [ʃ]. This supports the invariant theory and suggests that stable and descriptive acoustic characteristics can be found (Blumstein and Stevens, 1981). Second, the results also support the view that the configuration of the vocal tract during the production of fricatives shapes their spectrum, with the relevant spectral cues not residing primarily in the frequency of the highest amplitude but in the spectral mean and spread, but more research is needed in this direction. Finally, this paper shows that acoustic and phonetics studies can be helped by machine learning (and, more generally, data science) approaches: on the one hand, they can help to identify the voiced and unvoiced parts of a fricative and extract the frication noise and, on the other, to find patterns in the acoustic correlates extracted from speech sounds.

ACKNOWLEDGMENTS

We wish to thank our participants, the Phonetic Lab in St. Petersburg (and sound engineer Tatiana Chukaeva in particular), and the University of Zürich for financial support, technical support, and help with the design of the experiment (Volker Dellwo in particular). N.U. was partly supported by a grant from the Doctoral Program of Linguistics of the Faculty of Arts and Social Sciences, University of Zürich, Switzerland; N.U., M.A.T., and D.D. were funded by IDEXLyon Fellowship Grant No. 16-IDEX-0005 (2018–2021) and indirectly by the Labex ASLAN (Grant No. ANR-10-LABX-0081) of the University of Lyon within the program Investissements d'Avenir (Grant No. ANR-11-IDEX-0007) of the French National Research Agency (ANR).

[1]RU: [evo zavut saʃa [salʲ]], mnʲ e nravitsa tvoja [ʃalʲ]].
[2]RU: [ana skazala [salʲ]], a nʲ e [ʃalʲ]].
[3]See supplementary material at https://www.scitation.org/doi/suppl/10.1121/10.0005950 for a comparison of the waveform and spectrogram between the three fricatives (SuppPub5.pdf); (SuppPub1.pdf) for other measures, such as precision and recall, further explained in Sec. IV; (SuppPub2.pdf) for all cues and all ACETs; (SuppPub4.pdf) for more details about the acoustic cues across fricatives and ACETs; (SuppPub2.pdf) for the full list of the outputs from each replication; (SuppPub2.pdf) for similar results when six possible ACETs are considered; (SuppPub2.pdf) for the detailed output of the performance of models trained with both acoustic cues and MFCCs; (SuppPub4.pdf) for the detailed output of the regression analysis.
[4]A PointProcess object represents a sequence of points, t_i, ordered in time, defined on a domain $[t_{min}, t_{max}]$, with the index i between 1 and the number of points (Boersma and Weenink, 2021).
[5]Due to the very similar accuracy between window length 512 with overlap 50% and window length 256 with overlap 0%, we also analyzed the latter, as can be seen in the supplementary materials (SuppPub3.html). The results did not vary much between the two settings, but considering the continuous and overlapping nature of speech sounds and the best accuracy of the former settings, we only report the results from the former setting in the main text of the paper.
[6]Additionally, we measured the *central peak amplitude* and computed *peak*, *cog*, and *sdev* on the Bark scale. We also measured the *duration of the frication noise* (*ndur*). Neither the Bark scale nor the frication noise resulted in a divergence of performance; this information is thus included in the raw data but not reported in the current study.
[7]The information of the preceding and following context is also available in the raw data. Our testing shows that including this information does not result in a different performance of the models; the information is thus not included in the reported results but is available for readers.
[8]The number 300 was chosen based on the stabilisation point of the predictions. Further details are available in the supplementary material (SuppPub2.html).
[9]We also tested 100 replications with very similar results, but more computationally expensive.

Anjos, I., Eskenazi, M., Marques, N., Grilo, M., Guimarães, I., Magalhães, J., and Cavaco, S. (**2020**). "Detection of voicing and place of articulation of fricatives with deep learning in a virtual speech and language therapy tutor," in *Proceedings of Interspeech 2020*, October 25–29, Shanghai, China, pp. 3156–3160.
Behrens, S. J., and Blumstein, S. E. (**1988**). "Acoustic characteristics of English voiceless fricatives: A descriptive analysis," J. Phon. **16**(3), 295–298.
Blumstein, S. E., and Stevens, K. N. (**1981**). "Phonetic features and acoustic invariance in speech," Cognition **10**(1), 25–32.
Boersma, P., and Weenink, D. (**2021**). "Praat: Doing phonetics by computer (version 3.9) [computer program]," https://www.fon.hum.uva.nl/praat/ (Last viewed 4/7/2021).
Bolla, K. (**1981**). *A conspectus of Russian speech sounds* (Böhlau Verlag, Vienna, Austria).
Breiman, L. (**2001**). "Random forests," Mach. Learn. **45**(1), 5–32.
Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. (**1984**). *Classification and Regression Trees* (Taylor & Francis, New York).
Catford, J. C. (**1977**). *Fundamental Problems in Phonetics* (Indiana University, London).
Catford, J. C. (**1988**). *A Practical Introduction to Phonetics* (Oxford University, London).
de Manrique, A. M. B., and Massone, M. I. (**1981**). "Acoustic analysis and perception of Spanish fricative consonants," J. Acoust. Soc. Am. **69**(4), 1145–1153.
Derkach, M., Fant, G., and de Serpa-Leitao, A. (**1970**). "Phoneme coarticulation in Russian hard and soft VCV-utterances with voiceless fricatives," STLQPSR **11**(2–3), 1–7.
Dowle, M., and Srinivasan, A. (**2019**). "data.table: Extension of data.frame," R package version 1.12.2, https://CRAN.R-project.org/package=data.table (Last viewed 8/4/2021).
Draxler, C., and Jänsch, K. (**2018**). "SpeechRecorder (version 3.28.0) [computer program]," https://www.bas.uni-muenchen.de/Bas/software/speechrecorder/ (Last viewed 12/28/2020).
Forrest, K., Weismer, G., Milenkovic, P., and Dougall, R. N. (**1988**). "Statistical analysis of word-initial voiceless obstruents: Preliminary data," J. Acoust. Soc. Am. **84**(1), 115–123.
Fritsch, S., Guenther, F., and Wright, M. N. (**2019**). "neuralnet: Training of neural networks," R package version 1.44.2, https://CRAN.R-project.org/package=neuralnet (Last viewed 8/4/2021).
Funatsu, S., and Kiritani, S. (**1998**). "Perceptual properties of Russians with Japanese fricatives," in *Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP 98)*, November 30-December 4, Sydney, Australia.
Ghaffarvand Mokari, P., and Mahdinezhad Sardhaei, N. (**2020**). "Predictive power of cepstral coefficients and spectral moments in the classification of Azerbaijani fricatives," J. Acoust. Soc. Am. **147**(3), EL228–EL234.
Gordon, M., Barthmaier, P., and Sands, K. (**2002**). "A cross-linguistic acoustic study of voiceless fricatives," J. Int. Phon. Assoc. **32**(2), 141–174.
Haykin, S. (**1998**). *Neural Networks: A Comprehensive Foundation* (Prentice Hall, Englewood Cliffs, NJ).
Hayward, K. (**2000**). *Longman Linguistics Library Experimental Phonetics*, 2nd ed. (Longman, New York).
Heinz, J. M., and Stevens, K. N. (**1961**). "On the properties of voiceless fricative consonants," J. Acoust. Soc. Am. **33**(5), 589–596.

Hoelterhoff, J., and Reetz, H. (**2007**). "Acoustic cues discriminating German obstruents in place and manner of articulation," J. Acoust. Soc. Am. **121**(2), 1142–1156.
Hughes, G. W., and Halle, M. (**1956**). "Spectral properties of fricative consonants," J. Acoust. Soc. Am. **28**(2), 303–310.
Jassem, W. (**1965**). "The formants of fricative consonants," Lang. Speech **8**(1), 1–16.
Jassem, W. (**1995**). "The acoustic parameters of Polish voiceless fricatives: An analysis of variance," Phonetica **52**(3), 251–258.
Jesus, L. M. T., and Jackson, P. J. B. (**2008**). "Frication and voicing classification," in *Computational Processing of the Portuguese Language*, edited by A. Teixeira, V. L. S. de Lima, L. C. de Oliveira, and P. Quaresma (Springer, Berlin), pp. 11–20.
Jesus, L. M. T., and Shadle, C. H. (**2002**). "A parametric study of the spectral characteristics of European Portuguese fricatives," J. Phon. **30**(3), 437–464.
Jolliffe, I. (**2002**). *Principal Component Analysis* (Springer, New York).
Jongman, A., Wayland, R., and Wong, S. (**2000**). "Acoustic characteristics of English fricatives," J. Acoust. Soc. Am. **108**(3), 1252–1263.
Kisler, T., Reichel, U., and Schiel, F. (**2017**). "Multilingual processing of speech via web services," Comput. Speech Lang. **45**, 326–347.
Kissine, M., Van de Velde, H., and van Hout, R. (**2003**). "An acoustic study of standard Dutch /v/, /f/, /z/ and /s/," Linguist. Netherlands **20**, 93–104.
Kochetov, A. (**2017**). "Acoustics of Russian voiceless sibilant fricatives," J. Int. Phon. Assoc. **47**(3), 321–348.
Kong, Y.-Y., Mullangi, A., and Kokkinakis, K. (**2014**). "Classification of fricative consonants for speech enhancement in hearing devices," PLoS One **9**(4), e95001.
Kuhn, M., Chow, F., and Wickham, H. (**2019**). "rsample: General resampling infrastructure," R package version 0.0.5, https://CRAN.R-project.org/package=rsample (Last viewed 8/4/2021).
Kuhn, M., and Vaughan, D. (**2019**). "parsnip: A common API to modeling and analysis functions," R package version 0.0.3.1, https://CRAN.R-project.org/package=parsnip (Last viewed 8/4/2021).
Kuhn, M., and Wickham, H. (**2019**). "recipes: Preprocessing tools to create design matrices," R package version 0.1.6, https://CRAN.R-project.org/package=recipes (Last viewed 8/4/2021).
Ladefoged, P., and Maddieson, I. (**1996**). *The Sounds of the World's Languages* (Blackwell, Oxford, UK).
Ladefoged, P., and Wu, Z. (**1984**). "Places of articulation: An investigation of Pekingese fricatives and affricates," J. Phon. **12**(3), 267–278.
Liaw, A., and Wiener, M. (**2002**). "Classification and regression by randomForest," R News **2**(3), 18–22.
Maniwa, K., Jongman, A., and Wade, T. (**2009**). "Acoustic characteristics of clearly spoken English fricatives," J. Acoust. Soc. Am. **125**(6), 3962–3973.
Mann, V. A., and Repp, B. H. (**1980**). "Influence of vocalic context on perception of the [ʃ]-[s] distinction," Percept. Psychophys. **28**(3), 213–228.
McFarland, D. H., Baum, S. R., and Chabot, C. (**1996**). "Speech compensation to structural modifications of the oral cavity," J. Acoust. Soc. Am. **100**(2), 1093–1104.
McMurray, B., and Jongman, A. (**2011**). "What information is necessary for speech categorization? Harnessing variability in the speech signal by integrating cues computed relative to expectations," Psychol. Rev. **118**(2), 219–246.
Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., and Leisch, F. (**2019**). "e1071: Misc functions of the Department of Statistics, Probability Theory Group (formerly: E1071)," R package version 1.7-2, https://CRAN.R-project.org/package=e1071 (Last viewed 8/4/2021).
Milborrow, S. (**2019**). "rpart.plot: Plot rpart models: An enhanced version of plot.rpart," R package version 3.0.8, https://CRAN.R-project.org/package=rpart.plot (Last viewed 8/4/2021).
Nagamine, T., Seltzer, M., and Mesgarani, N. (**2015**). "Exploring how deep neural networks form phonemic categories," in *Proceedings of Interspeech 2015*, September 6–10, Dresden, Germany, pp. 1912–1916.
Newell, K. M., and Hancock, P. A. (**1984**). "Forgotten moments," J. Mot. Behav. **16**(3), 320–335.
Nirgianaki, E. (**2014**). "Acoustic characteristics of Greek fricatives," J. Acoust. Soc. Am. **135**(5), 2964–2976.
Nittrouer, S., Studdert-Kennedy, M., and McGowan, R. S. (**1989**). "The emergence of phonetic segments," J. Speech Lang. Hear. Res. **32**(1), 120–132.
Padgett, J., and Żygis, M. (**2007**). "The evolution of sibilants in Polish and Russian," J. Slavic Linguist. **15**(2), 291–324.
Paluszynska, A., and Biecek, P. (**2017**). "randomForestExplainer: Explaining and visualizing random forests in terms of variable importance," R package version 0.9, https://CRAN.R-project.org/package=randomForestExplainer (Last viewed 8/4/2021).
Parks, R. W., Levine, D. S., and Long, D. L. (**1998**). *Computational Neuroscience Fundamentals of Neural Network Modeling: Neuropsychology and Cognitive Neuroscience* (MIT, Cambridge, MA).
Patil, V., and Rao, P. (**2008**). "Acoustic cues to manner of articulation of obstruents in Marathi," in *Proceedings of Frontiers of Research on Speech and Music (FRSM)*, edited by A. Okrent and J. Boyle, February 20–21, Kolkata, India.
Peeters, G. (**2004**). "A large set of audio features for sound description (similarity and classification) in the CUIDADO project," CUIDADO First Project Report (IRCAM, Paris, France).
Rabha, S., Sarmah, P., and Prasanna, S. R. M. (**2019**). "Aspiration in fricative and nasal consonants: Properties and detection," J. Acoust. Soc. Am. **146**(1), 614–625.
Reidy, P. F. (**2016**). "Spectral dynamics of sibilant fricatives are contrastive and language specific," J. Acoust. Soc. Am. **140**(4), 2518–2529.
Schiel, F. (**1999**). "Automatic Phonetic Transcription of Non-Prompted Speech," in *Proceedings of the 14th International Congress of Phonetic Sciences*, August 1–7, San Francisco, CA.
Shadle, C. H. (**1985**). "The acoustics of fricative consonants," Doctoral dissertation, Massachusetts Institute of Technology.
Shadle, C. H. (**1990**). "Articulatory-acoustic relationships in fricative consonants," in *Speech Production and Speech Modelling* (Springer, New York), pp. 187–209.
Shadle, C. H., and Mair, S. (**1996**). "Quantifying spectral characteristics of fricatives," in *Proceedings of the Fourth International Conference on Spoken Language Processing, ICSLP '96*, October 3–6, Philadelphia, PA, Vol. 3, pp. 1521–1524.
Shupljakov, V., Fant, G., and de Serpa-Leitao, A. (**1968**). "Acoustical features of hard and soft Russian consonants in connected speech: A spectrographic study," STL-QPSR **9**(4), 1–6.
Skarnitzl, R., and Machač, P. (**2011**). "Principles of phonetic segmentation," Phonetica **68**, 198–199.
Soli, S. D. (**1981**). "Second formants in fricatives: Acoustic consequences of fricative-vowel coarticulation," J. Acoust. Soc. Am. **70**(4), 976–984.
Spinu, L., Kochetov, A., and Lilley, J. (**2018**). "Acoustic classification of Russian plain and palatalized sibilant fricatives: Spectral vs. cepstral measures," Speech Commun. **100**, 41–45.
Spinu, L., and Lilley, J. (**2016**). "A comparison of cepstral coefficients and spectral moments in the classification of Romanian fricatives," J. Phon. **57**, 40–58.
Stevens, K. N. (**1998**). *Acoustic Phonetics* (MIT, Cambridge, MA).
Strevens, P. (**1960**). "Spectra of fricative noise in human speech," Lang. Speech **3**(1), 32–49.
Tagliamonte, S. A., and Baayen, H. (**2012**). "Models, forests, and trees of York English: Was/were variation as a case study for statistical practice," Lang. Var. Change **24**, 135–178.
Tang, Y., and Horikoshi, M. (**2016**). "ggfortify: Unified interface to visualize statistical result of popular R packages," R J. **8**(2), 474–489.
Therneau, T., and Atkinson, B. (**2019**). "rpart: Recursive partitioning and regression trees," R package version 4.1-15, https://CRAN.R-project.org/package=rpart (Last viewed 8/4/2021).
Timberlake, A. (**2004**). *A Reference Grammar of Russian* (Cambridge University, Cambridge, UK).
Ting, K. M. (**2010**). "Precision and recall," in *Encyclopedia of Machine Learning*, edited by C. Sammut and G. I. Webb (Springer, Boston, MA).
Tomiak, G. R. (**1991**). "An acoustic and perceptual analysis of the spectral moments invariant with voiceless fricative obstruents," Doctoral dissertation.
Venables, W. N., Ripley, B. D., and Venables, W. N. (**2002**). *Statistics and Computing Modern Applied Statistics with S*, 4th ed. (Springer, New York).
Vydana, H. K., and Vuppala, A. K. (**2016**). "Detection of fricatives using S -transform," J. Acoust. Soc. Am. **140**(5), 3896–3907.

Wickham, H. (**2017**). "tidyverse: Easily install and load the Tidyverse," R package version 1.2.1, https://CRAN.R-project.org/package=tidyverse (Last viewed 8/4/2021).
Wickham, H., and Seidel, D. (**2020**). "scales: Scale functions for visualization," R package version 1.1.1, https://CRAN.R-project.org/package=scales (Last viewed 8/4/2021).
Zsiga, E. C. (**2000**). "Phonetic alignment constraints: Consonant overlap and palatalization in English and Russian," J. Phon. **28**(1), 69–102.
Żygis, M., and Padgett, J. (**2010**). "A perceptual study of Polish fricatives, and its implications for historical sound change," J. Phon. **38**(2), 207–226.

Chapter 5

Defining speaker-specific information in fricatives

This section deals with the speaker information encoded in fricative sounds. The section is based on the paper: Ulrich, N., Allassonnière-Tang, M., Pellegrino, F., Inter- and intra- speaker variation in eight Russian fricatives. The paper was accepted under the consideration of a Revision due to the 6th of December 2022.

The aim is to understand where and how do speaker exhibit individual differences in the speech signal. First, gender-specific characteristics are investigated. Second, the analysis is zoomed into the individual variation of speakers.

THIS MANUSCRIPT IS UNDER REVISION

Intra and inter-speaker variation in eight Russian Fricatives

Natalja Ulrich [1],[a] François Pellegrino [1], and Marc Allassonnière-Tang [2]
[1]*Lab Dynamics of Language UMR 5596, University Lyon 2, France,* [2]*Lab Ecological-Anthropology UMR 7206, National Museum of Natural History, France*

This paper is part of a special issue on Perception and Production of Sounds in the High-Frequency Range of Human Speech.

The current study shows that voiceless, voiced and palatal fricative sounds contain information specific to gender and individual speakers. The data consist of 15812 tokens from eight Russian fricatives and 59 speakers. Two types of acoustic cues are selected. First, 11 acoustic speech features (ASFs) including spectral cues, duration and HNR measures and second, 13 Mel Frequency Cepstral Coefficients (MFCCs) are extracted. Classifiers based on single decision trees and random forests were trained and tested to predict speakers' gender and ID from the two types of acoustic cues. Additional quantitative methods were utilized to understand the distribution of gender and speaker information across different fricatives and acoustic cues. The results show gender can be predicted with a good performance by both ASFs and MFCCs, whereby MFCCs clearly outperform ASFs. The individual speakers can only be predicted by MFCCs. ASFs encode speakers' idiosyncrasies in fricative sounds in a highly individual manner, and no set of cues can predict those idiosyncrasies. The study concludes that commonly extracted measures in phonetic research are insufficient to understand the complexity of speakers' individuality coding in complex speech sounds.

[https://doi.org(DOI number)]

[XYZ] Pages: 1–16

I. INTRODUCTION

Speech production is influenced by both anatomical predispositions and movements of the vocal tract, which are controlled by neuromuscular programming (Dellwo *et al.*, 2007). Any change in place and length of the constriction causes a change in the size and shape of the cavities behind and in front of the constriction. This, in turn, can result in a modification in the acoustic characteristics of the produced speech signals related to these cavities (Stuart-Smith, 2007).

Research on idiosyncrasy assumes that motor control in speech is highly individual like other modes of human movements similar to human gait (Matovski *et al.*, 2010). Those individual features are also reflected in the physical properties of speech sounds (He and Dellwo, 2014). Consequently, all produced speech signals carry both the linguistic meaning, for instance, the place of articulation or voicing, and certain speaker-specific characteristics. Research focusing on the idiosyncrasies in speech shows that a speaker's gender, accent, language, emotions, or health status can be exploited not merely by listeners, but also be determined by technological application and the extraction of acoustic cues (Dellwo *et al.*, 2007).

The investigation of inter- and intra- speaker variation and discrimination potential in each speech sound category represents a different challenge. An influential aspect which determines our understanding of the acoustic nature of sounds is the distribution of periodic and aperiodic energy. It has been demonstrated that speech sounds with predominantly periodic energy as vowels are better understood than speech sounds with a high degree of aperiodic energy, such as fricatives. Most papers focusing on idiosyncratic information considered vowel formants (McDougall and Nolan, 2007; Rose, 2007), and nasals (Enzinger and Balazs, 2011; Kavanagh, 2012).

Fricative sounds on the contrary are more challenging. They consist of either only aperiodic energy as in the case of voiceless fricatives or of the interaction of periodic and aperiodic components as in voiced fricatives. Comparing the frequency ranges of fricative sounds with other consonants, it was early noted that aperiodic spectral energy is presented in much higher frequency ranges than in other sounds (Strevens, 1960). Nevertheless, analysis investigating the spectral shape of fricatives primarily considered the bandwidths up to around 10kHz (Flipsen *et al.*, 1999) and more generally, the relevance of high-frequency sounds has been overlooked, as underlined by Monson and colleagues, 2014 (Monson *et al.*, 2014). In the past, the main motivation for this filter

[a]ulrichnatalja@gmail.com

was the technological limitation that only a particular bandwidth could be analysed (Strevens, 1960). Even though speech processing technologies developed in the last decades (e.g. through bandwidth extension (Jax and Vary, 2003)), description of linguistic and speaker aspects in fricatives continued investigating a frequency threshold up to 8 to 12 kHz (Forrest *et al.*, 1988; Gordon *et al.*, 2002; Jongman *et al.*, 2000; Kavanagh, 2011; Kochetov, 2017). Considering this frequency range, existing studies on idiosyncrasies in fricative sounds found substantial speaker variation and significant potential for speaker discrimination in fricatives (Kavanagh, 2011; Narayanan *et al.*, 1995; Newman *et al.*, 2001; Silbert and de Jong, 2008), suggesting their further exploration (Kavanagh, 2012; Schindler and Draxler, 2013). These analyses typically involved voiceless fricatives with a focus on the alveolar [s], whereby almost no observations exist on voiced and palatal fricatives. Research on the acoustics of Russian fricatives, in particular, is generally rare. These investigations were often based on the productions of a few speakers and speaker-specific attributes were not reported. Speaker and gender variation were considered so far for sibilant palatal vs. non-palatal fricatives (Kochetov, 2017; Spinu *et al.*, 2018) and for the variation of vocal fold vibration in voiced fricatives from eight speakers (Barry, 1995).

As a result of the low pass filtering of fricative sounds, the understanding of how much linguistic and speaker information and individual variability are coded in the higher frequency ranges remains unclear. Articles on speech production and perception in patients with cochlear implants and studies of hearing loss in the elderly have both observed that higher frequencies matter in speech perception. In the systems of the cochlear implants, for instance, not all the acoustic information about the spectral shape in the high-frequency ranges is sufficiently provided to the user (Moore, 2003). This, in turn, can cause developmental difficulties in children perceiving and articulating fricative sounds correctly (Grandon and Vilain, 2020).

The present study aims to fill these gaps by investigating how the speaker's gender and identity are encoded in the acoustic features of voiceless, voiced, and palatal fricatives in Russian. The article starts with an overview of the existing literature on gender and speaker variation in fricative sounds in Section II. Section III describes the process of data collection, acoustic cue extraction and acoustic analysis methods. For the current investigation, we used the same database and similar methods as in Ulrich et al (2021). However, in the present study, the recordings of 59 Russian native speakers were considered. From a dataset of 15812 tokens including the voiceless [f] [s] [ʃ], voiced [v] [z] [ʒ] and palatal [sʲ] [ɕ] fricatives, 11 Acoustic Speech Features (ASF) (spectral cues, duration and HNR measures) and 13 Mel Frequency Cepstral Coefficients (MFCCs) were extracted. To capture speaker-specific characteristics across all bands, the sounds were filtered only below 80 and above 20050Hz. For data analysis, two machine learning classifiers as well as several statistical methods were applied. Section IV explores the acoustic differences between gender categories and individual speakers. The first objective was to challenge the predictive power of acoustic measures to classify speakers' gender and to identify gender-specific traits in fricative sounds. Both sets of acoustic cues (ASFs, MFCCs) were tested to determine if they provide sufficient information to predict a gender by acoustic cues using a machine learning approach. Several statistical methods were employed to investigate gender-specific acoustic features and variation described in previous research. The second objective was to zoom into the individual level of a speaker. The application of the same machine learning techniques should show whether speakers' ID can also be predicted by the extracted cues. Then, a Principle Component Analysis (PCA) was compared with the ratio of standard deviation, which gave insights into the distribution of intra- and inter-speaker variation in the investigated fricatives. The paper finishes with a discussion on gender and inter- and intra-speaker variation in section V

II. PREVIOUS RESEARCH ON SPEAKER INFORMATION IN FRICATIVE SOUNDS

For the investigation of speaker-specific properties and variation in speech sounds, fricatives are particularly interesting because they consist mainly of turbulent noise and represent one of the most complex sound categories in terms of articulation, acoustics and perception.

From a typological view, fricatives represent the second largest group of obstruents (after stop consonants) across the world's languages (Maddieson and Disner, 1984). Fricatives exist at various places and voice settings and can undergo several secondary articulation processes such as palatalisation or aspiration (Ladefoged and Maddieson, 1996; Maddieson *et al.*, 2013). Fricative inventories can vary widely across the languages of the world, as reported in the LAPSYD database (Maddieson *et al.*, 2013). There are languages that arguably lack fricatives, as in Australian languages (Butcher, 2003; Maddieson and Disner, 1984; Maddieson *et al.*, 2013), to languages like Russian, with 12 phonological fricatives of five places of articulation, voicing and palatalization contrast (Bolla, 1981).

Furthermore, in the articulation and acoustics of fricative sounds, language-specific triads were identified. Speakers of different languages can apply various strategies to produce the same phoneme, resulting in varying acoustic properties, as observed by acoustic (Gordon *et al.*, 2002; Hayward, 2000; Ladefoged and Maddieson, 1996) and articulatory studies (Narayanan *et al.*, 1995). For instance, speaker and gender variation in the duration and the spectral shape was found to be prominent only for some languages, but not for others (Gordon *et al.*, 2002).

Even though a wide range of factors influences the acoustics of fricatives, several attempts were made to explore how the properties of the speech signal may

vary consistently between female and male speakers and across speakers in general and whether speaker-specific attributes can be exploited by human listeners or be predicted by technological applications using a set of acoustic cues. The literature suggests that regularities exist within those variations (Gordon *et al.*, 2002; Hughes and Halle, 1956; Jongman *et al.*, 2000; Kavanagh, 2011; Narayanan *et al.*, 1995; Newman *et al.*, 2001; Silbert and de Jong, 2008; Smorenburg and Heeren, 2020), which might help to discriminate between- speakers (Schindler and Draxler, 2013). Different fields, including acoustic phonetics, ASR, and forensic speaker comparison, are interested in investigating gender and inter- and intra-speaker variation. Therefore, a number of methods and measures were developed and applied in connection with the purpose of the research.

One of the best-described speaker characteristics is gender and the acoustic contrast between female and male speakers is argued to be well understood and explained by physiological and sociophonetic differences (e.g. Jongman *et al.*, 2000; Ludger *et al.*, 2021; Munson *et al.*, 2006). Perception experiments (Schwartz, 1968) and acoustic cue-based recognition tasks (Ghaffarvand Mokari and Mahdinezhad Sardhaei, 2020; Spinu *et al.*, 2018; Spinu and Lilley, 2016) provided evidence that gender information can be obtained from fricative sounds. Several studies argued that speaker variation is dependent on the place of articulation and greater gender variation was identified in anterior fricatives (Gordon *et al.*, 2002; Kochetov, 2017). A cross-linguistic study showed, for instance, that in some languages female speakers articulate front fricatives differently than males, resulting in acoustic gender variation (Gordon *et al.*, 2002).

To evaluate acoustic differences between female and male speakers, a number of acoustic speech features including spectral, temporal and amplitude cues were measured and analysed. These studies have concluded that the spectral domain provides crucial information on speakers' gender. A very early perception experiment on English fricatives concluded that human listeners can identify speakers' gender in isolated voiceless sibilant fricatives, relying on higher spectral energy in female productions. This effect was not present for the nonsibilants (Schwartz, 1968). Several follow-up studies also reported higher values for female speakers in the centre of gravity, and peak frequency (Flipsen *et al.*, 1999; Gordon *et al.*, 2002; Jongman *et al.*, 2000; Kochetov, 2017; Ludger *et al.*, 2021; Newman *et al.*, 2001). Gender variation is further found in spectral skewness (Flipsen *et al.*, 1999; Ludger *et al.*, 2021; Munson *et al.*, 2006; Stuart-Smith, 2007). Differences between female and male speakers were also observed in the duration of single-tone and germinate fricatives (Al-Tamimi and Khattab, 2015).

To predict gender from acoustic cues, several studies compared the performance between spectral measures and cepstral coefficients (Ghaffarvand Mokari and Mahdinezhad Sardhaei, 2020; Jesus and Jackson, 2008; Spinu *et al.*, 2018; Spinu and Lilley, 2016). The findings on Azerbaijani (Ghaffarvand Mokari and Mahdinezhad Sardhaei, 2020), Romanian (Spinu and Lilley, 2016) and a subset of Russian fricatives (Spinu *et al.*, 2018) showed that cepstral coefficients clearly outperform common spectral measures. Thereby, very similar accuracy rates with around 60% for ASFs and around 80% and higher for CCs were obtained. These results indicate that gender variation is best captured by the spectral envelope information measured by CCs.

A literature review on idiosyncrasies in fricatives reveals controversial results regarding whether individual speakers can also be predicted from fricative sounds. Some research demonstrated, for example, that models built for visual speaker recognition in vowels exhibit a decrease in performance for fricatives and nasals. The authors concluded that less speaker information is contained in these sound categories (Gendrot *et al.*, 2020). Significant differences in speaker discrimination potential were observed between voiced and voiceless sounds in general and between fricatives in particular for phoneme-based speaker identification in Arabic consonants (Alsulaiman *et al.*, 2017). Moderate performance in fricatives was noted in forensic voice comparison of six sound categories in French (Ajili *et al.*, 2017). Opposite findings were reported for English fricatives where a high recognition rate was achieved with vowels and also fricatives (Antal, 2008). Idiosyncratic information in fricatives and significant potential for speaker discrimination was further detected in ASFs (Gordon *et al.*, 2002; Hughes and Halle, 1956; Kavanagh, 2011; Narayanan *et al.*, 1995; Newman *et al.*, 2001; Silbert and de Jong, 2008; Smorenburg and Heeren, 2020), suggesting their further exploration (Kavanagh, 2012; Schindler and Draxler, 2013). For instance, the spectral peak frequency in voiceless fricatives was found to be highly variable between- speakers, and one speaker's alveolar peaks can appear as the post-alveolar peak frequencies of another speaker (Hughes and Halle, 1956). Also, the spectral moments were considered to serve as reliable acoustic cues for speaker discrimination in [f] and [s] sounds (Schindler and Draxler, 2013). The most substantial inter-speaker variability was identified in the spectral shape of the alveolar [s] (Gordon *et al.*, 2002; Kavanagh, 2011, 2012). As another example, laryngographic analysis of voiced obstruents showed that vocal fold vibration varies between-speakers, resulting in different frication and voicing duration as well as in different patterns of devoicing (Barry, 1995).

Several researchers argued that gender and speaker-specific properties are not solely reflected in the spectral and temporal domains but also in further acoustic characteristics. For instance, it was claimed that females produce stronger acoustic distinctions and articulate contrasting vowels and consonants more clearly. The productions of vowels (Diehl *et al.*, 1996; Weirich and Simpson, 2014) and fricatives (Weirich and Simpson, 2015) of female speakers tend to occupy a larger phonetic space than male speakers. Furthermore, duration and spectral

analysis were discovered to be unequally informative for different speakers. In discrimination tasks solely some speakers were identified by these measures. Acoustic properties were found to provide the best information for individuals at the extremes rather than in the middle of the distribution (Kavanagh, 2012). In a study comparing spectrograms and phonetic features extracted from vowels, it was confirmed that significant differences between-speakers exist. The investigation concluded that there are some *good speakers* giving the best results in the identification task and *poor speakers* showing poor results (Gendrot *et al.*, 2019). On the other hand, some articulation and acoustic studies claim, that intra-speaker variability in obstruents is contrast- and/or cue-specific rather than a general talker characteristic (Harper, 2021; Romeo *et al.*, 2013).

Regarding the Russian fricative inventory, solely a few studies were conducted to investigate gender variation. Most works on Russian fricatives examined the place of articulation or palatalisation contrast. These analyses were usually based on the productions of a few speakers and did not include speaker-specific descriptions as summarized in (Kochetov, 2017). Speaker and gender variation were reported so far for sibilant palatal and non-palatal fricatives produced by ten speakers (Kochetov, 2017; Spinu *et al.*, 2018) and the variation of vocal fold vibration in voiced fricatives from eight speakers (Barry, 1995).

To summarize, the research on idiosyncratic information in fricatives shows that speaker information is contained in fricative sounds. However, the literature review does not provide a clear overview of which acoustic cues are the most crucial for speaker recognition. In addition, so far, the study of inter- and intra-speaker variation mostly focused on spectral cues and conclusions have often been made based on a limited number of speakers or the acoustics of a small set of fricatives. The present study aims to provide a deeper understanding of idiosyncratic information in noise sounds by looking at Russian voiceless, voiced, and palatal fricatives in a large data sample in terms of the number of speakers and tokens.

III. PRIMARY DATA AND ACOUSTIC ANALYSIS

A. Participants, Data Collection and Segmentation

The participants were 59 students (30 females and 29 males) between 18 and 30 years old, studying at different departments of the St. Petersburg University in Russia. They were born or lived since their early childhood in St. Petersburg. No participants reported any speech or hearing impairment. All participants were first introduced to the purpose of the experiment, the expected duration and the procedure. They were told that they have the right to withdraw at any time during the experiment. They were provided with the contact details of a person that can answer all their questions concerning the research and their rights. The participants were compensated for their participation.

The recording sessions were conducted at the phonetic laboratory of the Phonetic Institute in St. Petersburg, in an audiometric booth using the recording program `Speech-Recorder` version 3.28.0 `https://www.bas.uni-muenchen.de/Bas/software/speechrecorder/` at a sample rate of 44.1 kHz (16-bit encoding). For the recordings, a clip-on microphone (Sennheiser MKE 2-P) was placed at a distance of 15cm from the speakers' mouth and connected through an audio interface (Zoom U-22) to a laptop computer.

Demographic data, such as gender and age, were recorded before the experiment started. The participants were instructed to read 198 sentences in random order from a computer screen. Two sentence structures were designed to obtain each real-word lexeme produced in three different contexts. One type of sentence is a so-called carrier sentence with the structure of "She said "X" and not "Y"" (RU: [ana skazala **salʲ**, a nʲe **ʃalʲ**]). Minimal pairs of real words for instance [salʲ] and [ʃalʲ] containing one of the 11 tested fricatives were placed in both "X" and "Y" positions. The second type of the predesigned sentence is a natural language sentence including each of the lexemes for instance "his name is Sasha [salʲ]" and "I like your [ʃalʲ]"(scarf) (RU: [evo zavut saʃa **[salʲ]**, mnʲe nravitsa tvoja **[ʃalʲ]**]). The distribution of voiceless, voiced and palatal fricatives depend on several phonotactic rules (e.g. Bolla, 1981; Timberlake, 2004). For example, voiceless fricatives can appear at the initial, medial, and word-final positions, while voiced fricatives undergo devoicing at the word-final position. Furthermore, not for all contrastive fricatives, minimal pairs exist. Consequently, a different number of tokens for each fricative could be recorded. The raw audio files were first automatically pre-processed by applying the online tool Munich Automatic Segmentation system, `MAUS` (Kisler *et al.*, 2017; Schiel, 1999) available at `https://www.bas.uni-muenchen.de/Bas/BasMAUS.html`. Then, the files were filtered below 80 and above 20050Hz with a smoothing of 80Hz, and the boundaries were manually corrected using `Praat` (Boersma and Weenink, 2022). In order to determine the onset and offset of the full consonant, the broadband spectrogram was considered more important than the start of an aperiodic waveform with rising zero crossing rates, and in intervocalic fricatives, the presence of formant columns is defined as the onset and offset of the fricative (following (Skarnitzl and Machač, 2011)).[1] The number of sounds per speaker employed in the current study is summarized in Table I. [2]

TABLE I. Token count by sound. Each speaker produced the same amount of token for each fricative category.

sound	[f]	[s]	[ʃ]	[v]	[z]	[ʒ]	[sʲ]	[ɕ]
freq	36	67	55	29	27	24	15	15

TABLE II. Summary of the acoustic cues included in the present study.

Cue	Variable	Description
Fricative Duration	*dur*	Duration of the entire sound obtained from manual segmentation
Peak Frequency	*peak*	Frequency of the highest amplitude
Spectral Mean	*cog*	Mean distribution of spectral energy (center of gravity)
Spectral Variance	*sdev*	Spectral spread or variance of the energy around the mean
Spectral Skewness	*skew*	Spectral tilt, overall asymmetry of the energy distribution
Spectral Kurtosis	*kurt*	Spectral flatness of the distribution
HNR mean	*hmean*	The mean of Harmonics to Noise Ration (HNR)
HNR sd	*hsd*	Standard deviation of HNR
HNR max	*hmax*	Maximum of HNR
HNR tmax	*htmax*	Time to the maximum HNR
tilt	*tilt*	Spectral tilt. Computed by H1-H2

B. Acoustic cue definition and extraction

The identification of speaker idiosyncratic information in speech sounds is the aim of diverse research fields with various purposes. It shows, therefore, significant theoretical and methodological diversity in the selection of extracted features as well as in the application of analytical techniques. One of the main aims of phonetic and acoustic research is to enhance the fundamental understanding of the relationship between articulation and acoustic properties by measuring and comparing acoustic features across phoneme classes and speakers. The most extracted and best understood acoustic cues are the peak frequency and the spectral moments, which are known to be correlated with articulatory and anatomical properties of a speaker (Newman *et al.*, 2001; Schindler and Draxler, 2013; Smorenburg and Heeren, 2020). Research focusing on automatic speaker recognition aims less to understand the interaction of articulation and acoustics, but to enhance speaker recognition systems by obtaining more abstract features such as for instance Mel-frequency coefficients (MFCCs) (Ganchev *et al.*, 2005).

For the current acoustic analysis, two sets of measurements were extracted from a data sample of 15812 tokens including the fricatives [f], [s], [ʃ], [v], [z], [ʒ], [sʲ], [ɕ] with `Praat` (Boersma and Weenink, 2022) and standard settings. The first set of features, to which we will refer in the following as Acoustic Speech Features (ASF), contains 11 measurements and is summarized in Table II.

Besides spectral measures which describe the distribution of the aperiodic energy across the frequency bands, harmonic-to-noise ratio (HNR) measures were extracted. The HNR cues give insights into the distribution of periodic and aperiodic energy. HNR mean and maximum values around zero indicate equal energy in harmonics and noise. A value of 20 indicates 99% of Harmonics and 1% of noise in the signal (Boersma and Weenink, 2022). The second set of acoustic features is represented by the extraction of the 13 Mel Frequency Cepstral Coefficients (MFCCs).

The extraction of these measures was conducted based on the entire fricative duration. The spectral analysis was performed on 10ms non-overlapping windows and averaged over the entire sound. The examination of all measurements was performed on the frequency bands from 80 to 20050Hz.

C. Methods for acoustic analysis

To investigate the inter- and intra- acoustic variation in the target fricatives, first machine learning techniques were tested to predict speakers' gender and ID. For further exploration of acoustic gender and speaker variation, a number of statistical analyses were performed. The significance of gender differences was visualized and assessed with Wilcoxon tests' using the Bonferroni multiple testing correction. The significance of the variables was then computed with machine learning methods such as *random forest* (RF). The extracted values were normalized considering their z-score (*(x-mean(x))/sd(x)*). If a value is exactly equal to the mean of all the values of the feature, it will be normalized to 0. If it is below the mean, it will be a negative number, and if it is above the mean it will be a positive number.

In order to estimate the potential of gender and speaker prediction by acoustic cues, we utilized Machine Learning techniques. The performance of the ASFs and MFCCs was tested and compared. Two classifiers based on binary recursive partitioning (Breiman *et al.*, 1984) were employed: the first classifier generates a single *decision tree* (DT) based on the data and helps to visualize the interactions between the variables. The output of this classifier is an explicit decision tree that captures the hierarchical interactions of the variables within the data set. The second is a 'random forest' (RF) (Breiman, 2001). It generates a series of 200 decision trees[3] that are

analyzed as a whole and used to assess the importance of each variable with regard to correctly predicting the fricatives. For each tree, the algorithm uses a bootstrap sub-sample of observations and a random subset of the variables from the entire data set. We adopted this approach, encouraged by the limited size of the data set because it is adequate to identify the features that exhibit an interesting identification potential. Additionally, we were not focused on the absolute classification performances, and the risk of overestimating them was not crucial in this context. The two classifiers were trained on 70% of the data (the training subset) and their accuracy was evaluated on the other, non-overlapping, 30% of the data (the test subset). Additional details on the parameters are available in Supplementary Materials 1 at [URL will be inserted by AIP].

The accuracy is the proportion of all correctly classified sounds. It provides an overview of the performance of the entire (balanced) dataset. However, accuracy gives only a general idea of the performance of the model. To have a more precise idea as to how the classifier performs for each sound we also reported *precision* and *recall* (Ting, 2010). Precision quantifies how often each gender or speaker was correctly classified (e.g., how many of the sounds classified as produced by for instance female speakers were actually classified as produced by female speakers). Recall quantifies how many of the sounds actually belonging to each category are correctly classified (e.g., how many sounds produced by female speakers are correctly classified as produced by female speakers). Precision and recall are computed for both genders, resulting in two estimates of precision and two of recall. Due to the lack of balance in terms of the token number between the produced fricatives, the *kappa* was also used to assess the overall accuracy. The kappa metric compares the observed accuracy of a classifier with the expected accuracy under random classification. This metric is generally used to compare the performance of different classifiers on different sets of data, as the comparison with the accuracy under random classification allows a cross-model and data comparison. It is calculated with the following formula: (observed accuracy - accuracy based on random classification)/(1 - accuracy based on random classification). For example, if the accuracy of the classifier is 0.7 and the accuracy based on random classification is 0.5, the kappa is equal to 0.4. A kappa higher than 0.75 means excellent performance. A kappa between 0.40 and 0.75 indicates a fair performance, while a kappa lower than 0.40 shows a poor performance.

In previous research gender and speaker variation was often accessed by various statistical methods. Frequently computed and compared were the statistical means of spectral and temporal cues across speakers and sounds (Gordon *et al.*, 2002; Kavanagh, 2011; Newman *et al.*, 2001; Silbert and de Jong, 2008). To capture the produced variance within each category, the range (Kavanagh, 2011; Silbert and de Jong, 2008), standard-deviation (Newman *et al.*, 2001), and the Interquartile Range (IQR) (Ferragne and Pellegrino, 2010) were measured.

For the current analysis, first, the significance of gender variation was tested across all raw values for each of the eight fricatives and the 11 ASFs. To visualise gender differences and compare the acoustic properties of Russian fricatives to previous analyses the statistical mean was obtained. Furthermore, the IQR was measured to determine the variation across gender categories. For each sound and cue, the IQR average over gender categories was estimated and the difference between female and male speakers was tested for significance.

Apart from that, several investigations found that male and female speakers organize their fricative contrasts differently, computing pairwise distances between the fricatives produced by male and female speakers (Weirich and Simpson, 2015). As an additional way of assessing the pairwise distances within the current data set, two t-SNE (t-distributed stochastic neighbour embedding) representations of the sound tokens are generated (Van der Maaten and Hinton, 2008). The t-SNE method is selected to represent the high-dimensional data of ASFs and MFCCs on two-dimensional plans. For each representation, the Euclidean distance was computed between all the tokens of two contrastive sounds in the t-SNE representations. For the comparison of female and male speakers, the measured distances were compared by gender and sound. More precisely, the distance is computed for sound pairs contrasted by places of articulation [f]- [s], [s]- [ʃ], [v] -[z], [z]- [ʒ], [sʲ]- [ɕ], voicing [f]- [v], [s]- [z], [ʃ]- [ʒ] and palatalisation [s]- [sʲ], [ʃ]- [ɕ].

To explore speaker inter- and intra- speaker variation, the same methods as used for gender don't lead to results that can be interpreted. We decided to explore speaker variation by Principal Component Analysis (PCA) based on the z-scored ASFs. First, a PCA was performed on each sound with the goal, to detect acoustic cues that explain most of the variation within each fricative. The PCA showed the overall variation, without making statements on the origin of variation, such as whether the variation is caused by a high degree of inter or intra-speaker variation. To evaluate the source of variation, we added a measure that we will refer to as the SD-ratio in the following text. For each fricative, we divided the overall standard deviation of a certain sound and cue by the speakers' standard deviation of the same sound and cue. This measure is expected to capture similar information used in previous studies (Schindler and Draxler, 2013) as a measure to visualize the ratio of inter- and intra-speaker variability. The same two methods were further applied to explore the acoustic properties of individual speakers in more detail. Therefore, a PCA was performed and an SD-ratio was computed for each sound and speaker.

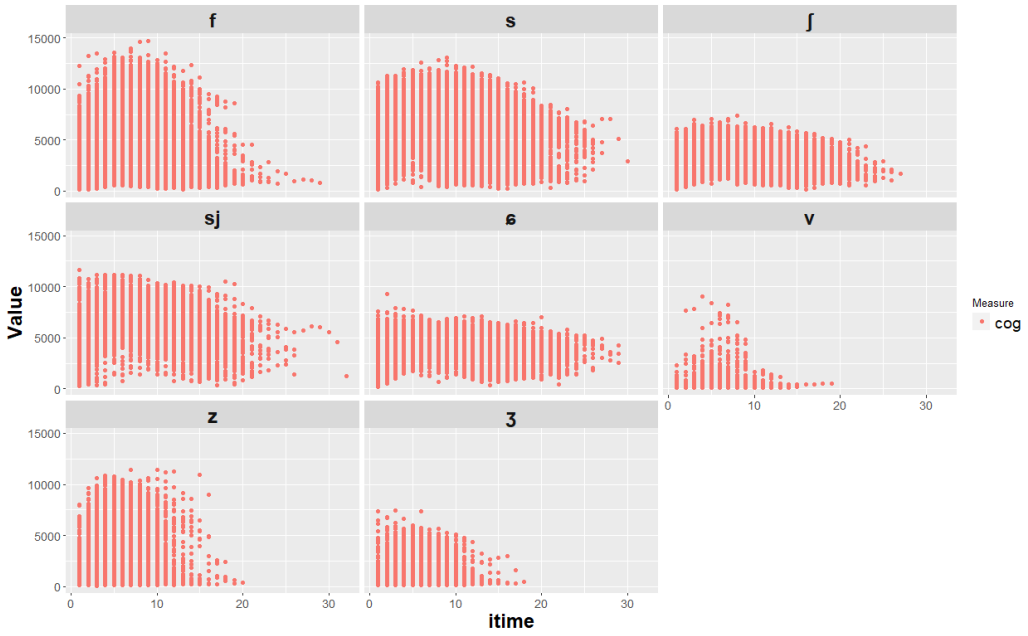

FIG. 1. *Cog* across sounds for all speakers. The x-axis represents the time windows and each line is a window of 10ms. The Y axis in frequency in Hz. The energy extends 10Hz clearly in [f], [s] and [s^{j}]

IV. PREDICTIVE POWER OF SPEAKERS' GENDER AND ID AND INTER AN INTRA-SPEAKER VARIATION

Fricative sounds contain spectral energy in higher frequency ranges than other sounds (Strevens, 1960). Before discussing gender and speaker-specific characteristics of fricatives, Figure 1 provides a sense of how spectral energy is distributed across different bands and time windows. It shows that spectral energy frequently extends above 10kHz, particularly in [f] and [s] and [s^{j}].

For the investigation of idiosyncratic information, two main objectives were defined. The first objective aimed to test the predictive power of the speaker's gender and to explore gender-specific traits in fricative sounds. The second objective focused on predicting individual speakers and describing inter- and intra-speaker variation. In the following outline, the main results of the analysis are reported. Additional details and code can be found in Supplementary Materials 1 at [URL will be inserted by AIP].

A. Predictive power of speaker's gender and acoustic gender variation

Gender prediction was carried out on the entire data sample including all eight fricatives. The performance was tested and compared between the ASFs and the MFCCs. The results from both data sets (ASF and MFCC) and both classifiers are summarized in Table III.

TABLE III. The performance of the two classifiers across ten replications was tested to predict the speaker's gender. The abbreviations are interpreted as follows: DT = single decision tree, RF = random forest, Acc = accuracy, Prec_F = precision female, Rec_F = recall female and Prec_M = precision male, Rec_M = recall male. The majority baseline is always 0.5008, due to an extra speaker in the female sample.

Classifier	Set	Kappa	Acc	Prec_F	Rec_F	Prec_M	Rec_M
DT	ASF	0.28	64%	0.64	0.68	0.65	0.6
RF	ASF	0.45	72%	0.72	0.76	0.74	0.7
DT	MFCC	0.31	66%	0.64	0.73	0.68	0.57
RF	MFCC	0.76	88%	0.88	0.88	0.88	0.87

When comparing the performance of the ASF and MFCCs, we observed accuracies of 64% for ASFs and 66% for MFCCs for the decision tree classifier (DT). With the second classifier [random forest] (RF), the predictive power increases for both sets of data. For both classifiers, MFFCs clearly outperform the ASFs with 88% in comparison to 72%. These results are in line with previous research, which observed classification rates of around 60% for ASFs and around 80% and more for CCs (Ghaffarvand Mokari and Mahdinezhad Sardhaei, 2020; Spinu *et al.*, 2018; Spinu and Lilley, 2016). Thus, it can be con-

TABLE IV. Mean values by gender for each ASF. To estimate the significance of the differences between female and male speakers all tokens were used. The majority of ASFs show significant gender variation and are marked in bold.

gender	sound	peak	cog	sdev	skew	kurt	dur	hmean	hsd	htmax	hmax	tilt
[f]	F	**1465**	**3499**	**3491**	**2.82**	**21.4**	**0.134**	**1.84**	**2.67**	0.040	**8.88**	-3.35
	M	**1662**	**3730**	3545	**2.40**	**16.4**	**0.132**	**1.21**	**2.46**	0.039	**7.74**	-4.64
[s]	F	**5904**	**6837**	**2715**	**0.17**	**3.18**	**0.155**	**0.42**	**1.68**	0.066	**9.84**	**-23.7**
	M	**5249**	**6078**	**2484**	**0.45**	**2.92**	**0.154**	**-0.16**	**1.54**	0.066	**9.32**	**-33.2**
[ʃ]	F	**2790**	**3439**	**1821**	**1.80**	**10.1**	**0.153**	**-0.21**	**1.85**	**0.071**	**10.4**	**-33.9**
	M	**2739**	**3312**	**1783**	**1.57**	**8.41**	**0.148**	**-0.56**	**1.77**	**0.066**	**8.59**	**-27.6**
[sʲ]	F	**5246**	**6252**	2520	**0.52**	3.36	0.160	**-0.19**	1.60	0.063	**11.0**	-44.8
	M	**4643**	**5644**	2459	**0.75**	3.32	0.158	**-0.85**	1.56	0.061	**9.81**	-44.4
[ɕ]	F	**3282**	**4095**	**1994**	**1.29**	**5.43**	0.182	**-0.49**	1.60	0.08	14.0	-59.6
	M	**2928**	**3666**	**1836**	**1.38**	**5.80**	0.18	**-0.26**	1.62	0.082	13	-62.4
[v]	F	**226**	**339**	452	**27**	**1622**	**0.076**	**17.2**	**2.72**	**0.041**	**20.7**	-0.60
	M	**170**	**277**	442	**20**	**1059**	**0.083**	**14.2**	**2.97**	**0.046**	**18.1**	-0.89
[z]	F	**1154**	**2439**	**2268**	**5.44**	**110**	**0.088**	10.74	**3.39**	**0.024**	**15.7**	**1.45**
	M	**631**	**1560**	**1774**	**6.16**	**110**	**0.097**	10.39	**2.73**	**0.034**	**14.5**	**-0.4**
[ʒ]	F	**668**	**1300**	**1326**	**4.27**	**51**	**0.087**	**8.31**	**2.72**	**0.026**	**12.6**	**2.16**
	M	**399**	**944**	**1125**	**5.22**	**72**	**0.095**	**8.63**	**2.23**	**0.041**	**12.05**	**0.53**

cluded that the detailed information encoded by MFCCs on a non-linear spectral scale outperforms the simpler linear features. However, the MFCCs are not ideal to interpret how gender information is coded in fricative sounds. The accuracy based on ASF reaches 72% and has a kappa of 0.45, which means a fair performance. This suggests that gender can be predicted from spectral and HNR cues. The most important cues across the eight fricatives are *peak*, *cog*, *skew* and *hmean*. The precision and recall are generally high for both gender across the two data sets, meaning that gender was generally classified correctly.

To explore further the variation between gender categories, a number of tests were performed. As outlined in the literature review, acoustic characteristics specific to female and male speakers were reported merely for a limited set of fricatives and acoustic cues. However, very narrow information is obtainable about gender differences in, for instance, voiced and palatal fricatives. To address these gaps, the following objective was to investigate gender variation in voiceless, voiced and palatal fricatives of different places of articulation. A statistical analysis of gender-specific properties and variation was performed. For this description, the raw values, the statistical mean and the interquartile range (IQR) of all ASFs from the eight fricatives were considered. In the current data acoustics differences between gender categories were detected in almost all eight fricatives and ASFs. The mean values averaged by gender, sound and cues are summarised in Table IV.

Significant differences between the female and male distributions are denoted by values written in bold. To define the significance of gender variation all realisation of the speakers were considered and not just the statistical mean. Most p values are smaller than 0.001, they are thus reported in Supplementary Materials 1 at [URL will be inserted by AIP] but not listed in this Table.

In line with previous observations (Flipsen *et al.*, 1999; Jongman *et al.*, 2000; Kochetov, 2017; Newman *et al.*, 2001; Schwartz, 1968), female speakers produced higher spectral energy than male speakers in the voiceless sibilant fricatives. Additionally, the same relation accounts for palatal and voiced fricatives. Interestingly, the opposite was found to be the case for the *peak* and *cog* in [f] and females produced lower values than male speakers. Contrary to previous studies which noted a greater gender variation in anterior fricatives (Kochetov, 2017), the current results suggest for almost all cues significant variation in all three places of articulation. The second spectral moment *sdev* was less explored in the literature. We measured significant gender differences and higher values for female speakers in [s] and [ʃ], indicating that female speakers produced more spectral spread than male speakers. On the contrary, the spectral variance in [f] was smaller for females. This is also theoretically expected as the spectral spread is correlated with

the centre of gravity, i.e., a higher *cog* leads to a higher *sdev*. More findings were reported in the literature for the spectral skewness in [s], with a tendency for negative skewness in female speakers and positive values or values centred near to zero for male speakers (Flipsen *et al.*, 1999; Ludger *et al.*, 2021). The present data shows a similar trend with values centred around zero for females and 0.5 for males around. Higher values are found for [f] and [ʃ] and female speakers exhibit higher positive values than males. Kurtosis was so far not reported for gender variation. Across the tested voiceless fricatives *kurt* is the lowest in [s] and the highest in [f]. Female speakers produced thereby in all three fricatives higher values than males.

Gender-specific properties in palatal fricatives were less explored in previous research. The observed patterns are convergent with previous studies for the non-palatal fricatives, with female speakers producing higher spectral energy (Kochetov, 2017) indicated by the *peak* and *cog*. Furthermore, in both sounds, the female speakers showed higher positive *skew*.

Gender variation in voiced Russian fricatives is also understudied. As in the voiceless fricatives, we found gender variation in almost all spectral cues. In *peak* and *cog*, female speakers exhibit higher values for all three voiced fricatives and additionally a higher spectral spread *sdev* in sibilants. The highest skewness *skew* was measured in the bilabial fricative, with higher values in [v] and lower values for [ʒ] for female than for male speakers. The results of kurtosis in voiced fricatives suggest significant gender variation. However, the measured kurtosis is challenging to interpret because the range of positive kurtosis above 3 exhibits a large variation across the sounds, with values over 1000 in [v]. Such high values observed in [v] suggest a very compact spectral distribution, which also can be viewed in Figure 1.

Gender variation was further detected in duration in the non-palatal fricatives. Female speakers produced longer voiceless fricatives and male speakers the voiced fricatives.

While gender differences were widely investigated in the spectral domain of voiceless fricatives, very little is known about further noise features like the distribution of periodic and aperiodic energy. One way to look at it is through the analyses of the harmonic-to-noise ratio (HNR). For almost all cues and sounds, significant gender variation was identified in the HNR measures. The distribution of HNR in the voiceless fricatives differs between the places of articulation. Table IV shows that the fricative [f], contains some harmonic energy, and the sibilant fricatives have higher ratios of noise energy.

Concerning gender variation, female speakers produced relatively higher harmonics proportions (*hmean hmax*) in their articulations of the voiceless fricatives, except for the fricative [ɕ]. The same pattern can be observed in the voiced bilabial and alveolar fricatives, but not for [ʒ]. Furthermore, female speakers had a higher variance (as indicated by higher values in *hsd*) in all fricatives, except for [v]. This suggests that the distribution of periodic and aperiodic energy generally differs between female and male speakers. This outline indicates, that there are gender-specific properties in the acoustics of fricatives and these differences manifest in the most measured cues and sounds.

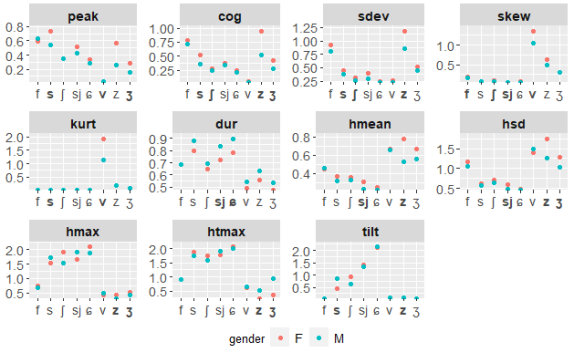

FIG. 2. The interquartile range is based on z-scored values for each ASF and each fricative. Significant gender variation is marked in bold. Female and male speakers show significant variation for several cues and all fricatives.

To explore additional gender-specific acoustic properties, the interquartile range (IQR) was analyzed (Figure 2). The IQR was first computed for each speaker and then averaged over gender categories. The IQR should shed light on whether the produced variance within gender categories differs between female and male speakers and to what extent it is a gender-specific property. When comparing all sounds and cues together, we observed that female speaker produced in general more variant cues ($p = 0.0068$). However, these conclusions cannot be generalized, since the values differ with sounds and cues. Broadly speaking, the IQR values for *peak* were significantly higher in [s] and [ʒ] for female speakers and in [v] for male speakers. A higher IQR in female acoustics was also found for the *cog* in [s], [ʃ], [z], and [ʒ]. Interestingly, male speakers produced significantly more variance in the duration of palatal fricatives. Female speakers showed significantly higher variation in the IQR of *hmean* in [s^{j}] and [z] and *hmax* in [z] and [ʒ].

To sum up the findings so far, it is noticeable that the ranges occupied by female and male productions are to a certain degree dependent on the place and manner of articulation. Furthermore, variation is not systematic across gender categories.

To test further the hypothesis that female and male speakers organize their fricative contrasts differently, the distance was measured for places of articulation between the fricatives pairs [f]- [s], [s]- [ʃ], [v] -[z], [z]- [ʒ], [s^{j}]- [ɕ]. Furthermore, the distance was compared between pairs contrasted by voicing [f]- [v], [s]- [z], [ʃ]- [ʒ] and palatalisation [s]- [s^{j}], [ʃ]- [ɕ]. All the results and Figures are reported in Supplementary Materials 1 at [URL will be inserted by AIP].

Comparing the effect on distances produced by both genders between contrastive categories across the two sets of measures, it is noticeable that they show different patterns. This has mainly to do with the fact that ASFs and MFFCs capture different properties of speech sounds.

For a better overview and for an easier interpretation, only the findings from ASFs were reported in the following outline. Figure 3 presents the results of the place of articulation contrast in the voiceless, voiced and palatal fricatives.

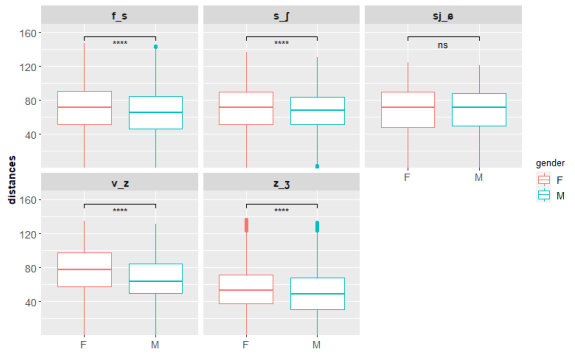

FIG. 3. Distance measures of the five fricative pairs contrasted by place of articulation. Female speakers produce a greater phonetic distance between the tested non-palatal sound pairs. The p-values for fricatives pairs where significant gender variation is found are smaller than 0.0001. The p-value for the palatal pair is 0.79

The observations made for voiceless sibilants by previous research (Weirich and Simpson, 2015) can be extended to other fricative pairs contrasted by place of articulation. Figure 3 shows that for both non-palatal voiceless [f]- [s], [s]- [ʃ], and voiced pairs [v] -[z], [z]- [ʒ], contrasted by place of articulation, female speakers produce in general a larger distance between the tested contrastive fricative pairs. However, in the palatal pair [sʲ]-[ɕ] no such gender variation was observed. The output for fricative pairs contrasted by voicing and palatalisation can be viewed in Supplementary Materials 1 (Section 3) at [URL will be inserted by AIP]. In fricative pairs contrasted by voicing, different patterns can be noticed. Female speakers produced less distance between the bilabial and post alveolar pairs [f]-[v] and [ʃ]-[ʒ]. However, they produced more distance in the alveolar pairs [s]-[z] than male speakers. Female speakers showed also a larger distance between the two sibilant fricative pairs [s]-[sʲ], [ʃ]-[ɕ] contrasted by palatalization.

The analysis of distances between contrastive sounds, confirms place and manner of articulation trends and gender variation seems to be less systematic.

B. Predictive power of individual speaker's and inter- and intra- speaker variation

In the next step, we asked whether more information is coded in the extracted ASFs and MFCCs by zooming into how speakers differ on the individual level. First, we investigated whether individual speakers can be identified with machine learning methods based on ASFs and MFCCs, as it was conducted with the prediction of gender. Second, we explored the variety of acoustic cues between- and within- speakers.

TABLE V. The performance of the two classifiers across ten replications to predict individual speakers. The abbreviations are interpreted as follows: DT = single decision tree, RF = random forest, Acc = accuracy. The majority baseline is 0.02.

Classifier	Set	Kappa	Acc
DT	ASF	0.00	0.014
RF	ASF	0.21	0.22
DT	MFCC	0.04	0.05
RF	MFCC	0.64	0.64

Table VII shows that in terms of predicting speakers, the two decision tree-based algorithms were unable to identify speakers by ASFs. The accuracy of each model is extremely low, and the kappa indicates poor performance. For instance, the kappa for RF based on ASFs is only 0.22. Testing MFCCs, the accuracy and kappa were much higher with around 64%, suggesting a moderate performance in predicting speaker.

Evidence that fricative sounds do contain certain speaker information in the spectral moments was found by previous research (Kavanagh, 2011; Newman *et al.*, 2001; Schindler and Draxler, 2013). One explanation about why these findings were not confirmed and the classifiers failed to predict speakers by ASFs in Russian fricatives is that the data set is unsuitable for applications of machine learning techniques used in the current analyses. The employed data has a high number of speakers (59) to be predicted and a relatively low number of tokens per sound and speaker. There is also the possibility that speaker variation is encoded in a more complex way in ASFs measured in fricative sounds. To explore this complexity of speaker information and individual differences, we provide a detailed description of speaker specificity in fricative sounds.

In order to determine the cues and sounds that explain most of the variation in the data, a Principal Component Analysis (PCA) was performed based on the z-scored ASFs for each sound. The results of the first principal component (PC1) are summarized in Table VI.

The data indicate that the variation in the investigated sounds is characterized by different sets of cues. Nevertheless, some patterns can be detected between sibilant and non-sibilant and voiceless and voiced frica-

TABLE VI. Summary of the PC1 of all eight fricatives. The values indicate how much variation in each sound can be explained by a certain cue averaged over the speaker. The PC variance indicates how much variance of the total variance can be explained by PC1.

cue	[f]	[s]	[ʃ]	[sʲ]	[ɕ]	[v]	[z]	[ʒ]
PC1 variance	0.32	0.41	0.45	0.46	0.46	0.57	0.57	0.43
peak	0.19	0.03	0.01	0.05	0.01	0	0.09	0.06
cog	0.23	0.01	0	0.03	0.01	0.01	0.15	0.1
sdev	0.21	0.03	0.02	0.03	0.01	0.06	0.18	0.13
skew	0.06	0	0.01	0	0	0.32	0.1	0.08
kurt	0.01	0	0	0	0	0.49	0.03	0.02
dur	0.02	0.15	0.13	0.15	0.12	0.01	0.02	0.02
hmean	0.05	0.01	0.02	0	0	0.07	0.12	0.15
hsd	0.17	0.02	0.05	0.02	0.01	0.01	0.23	0.31
hmax	0.06	0.28	0.3	0.24	0.26	0.04	0.04	0.04
htmax	0	0.23	0.18	0.18	0.17	0	0.04	0.07
tilt	0.01	0.23	0.29	0.29	0.39	0	0	0.01

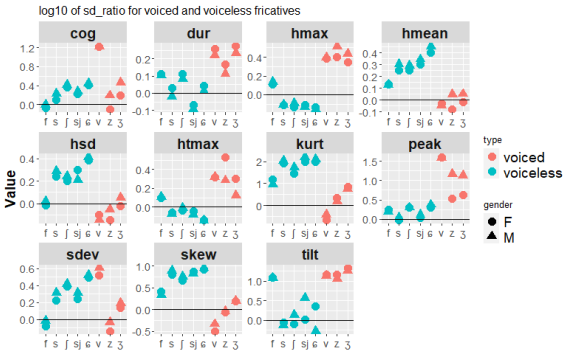

FIG. 4. The SD-ratio averaged over all speakers by sound and cues. For better visualisation, the log10 of an SD-ratio is used. For instance, 3 is now equal to 0.48, and an SD-ratio of 1 equals 0. The values below 0 in this Figure mean that the within-speaker variation is higher than the between-speaker variation. Values above 0 indicate higher between-speaker variation. The larger the SD-ratio, the higher the between-speaker variation and the lower the within-speaker variation, which indicates high speaker-discriminating potential.

tives. The *peak* frequency and the spectral moments were the most variant cues in [f], [v] and [z]. In [v], PC1 mostly consists of *skew* and *kurt*, while the variance in the rest of the fricatives is distributed across several ASFs. Thus, in most fricatives, the spectral domain seems to be less meaningful in explaining variation. In addition, duration also explains some of the variation in voiceless sibilants. Contrary to expectations, the harmonic-to-noise ratio cues in sibilant fricatives were more variable than the spectral cues. Different patterns were observed for voiceless and voiced fricatives. In voiceless sibilants, the most variant cues were *hmax*, *htmax* and *tilt*. In voiced sibilants, the most variant cues were *hmean* and *hsd*. These results suggest that the greatest variation was detected in the distribution of periodic and aperiodic energy in the sibilant fricatives. The PC1 variance indicates furthermore, how much variance of the total variance is explained by the first component. For example, the PC1 variance in [f] is only 32% and it is 57% in [v]. In sibilant fricatives, the lowest PC1 variance value is in [s] and the highest in [z].

PCA reveals the most variant ASFs across fricatives. However, it does not provide information on whether the measured variance is caused by a high variation between or within speakers. To understand the distribution of intra- and inter-speaker variation, we computed the SD-ratio. Figure 4 displays the SD-ratios of the eight fricatives and the eleven ASFs.

In acoustic cues, the SD-ratio identifies speakers' discrimination potential (SDP). It is calculated by dividing the overall standard deviation of a sound and cue by the speaker's standard deviation. The higher the value, the greater the acoustic difference between speakers and the higher the SDP. Figure 4 displays that in some of the cues identified by PCA as highly variant, the variation is caused by high between-speaker variation, while in others it is caused by high within-speaker variation.

Common cues with a high between-speaker variation in [f] and [v] were *peak* and *tilt*, suggesting a high speaker discrimination potential in these cues. The SD-ratios of *cog*, *sdev*, and *hsd* in [f] are below 1 (below 0 in Figure 4), which indicates that the within- speaker variation is higher than the between- speaker variation. This in turn means that these three measures provide little speaker information in the fricative [f], while in [v], *cog* and *sdev* have very high values. In the sibilant voiceless fricatives, a higher between-speaker variation than within-speaker was found in the spectral moments, *hmean* and *hsd*, while *hmax* and *htmax* indicated higher within- speaker variation. In the voiced fricatives, the opposite patterns for the same cues were observed.

Finally, for a better overview of the relations between the PC1 and SD-ratios, the values were compared and the correlation was computed. Figure 5 shows the distribution of the PC1 and SD-ratio values across the eight fricatives and the ASFs.

In most cases, values with a high PC1 display a low SD-ratio. Consequently, the majority of cues identified to explain a large part of the variation were produced by speakers with a high degree of within-speaker variation. ASFs with a low PC1 on the other hand showed in a set of cues a high SD-ratio. Additionally, a correlation analysis of PC1 and SD-ration found a negative correlation for almost all cues and sounds.

These findings suggest that no cue effectively explains variation in general within a sound and has a high

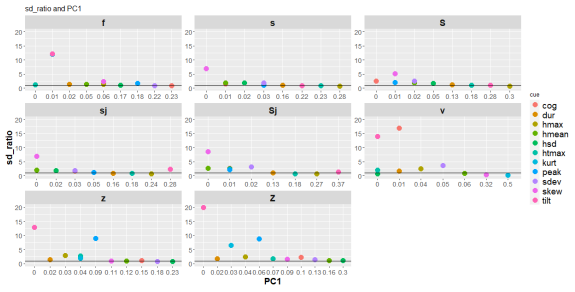

FIG. 5. Comparison of the PC1 and SD-ratio by cue and sound. A higher PC1 value shows that more variation within a sound can be explained by this cue. And as higher the SD-ratio the higher the between-speaker variation.

degree of between-speaker variation. Therefore, it is challenging to conclude which cues could potentially serve to encode speaker information in fricative sounds. To understand better the distribution of within and between-speaker variation, a closer look was taken by analysing individual speakers. Therefore, we provide a more focused example of the interaction between PC1 and SD-ratio across three speakers and two sounds in order to give an indication of how the variation is structured at an individual level.

Three female speakers showing quite different patterns were chosen based on a visual inspection. A PCA was performed on each speaker and the SD- value was computed. The results of the PC1 and SD-ratios across the two fricatives [f] and [ʃ] and the three speakers are summarized in Table VII.

TABLE VII. PC1 and SD-ratio for three speakers and two sounds. The three speakers show remarkable differences in the distribution of the ratio between PC1 and SD-ratio across the two fricatives [f] and [ʃ].

sound	[f]		[f]		[f]		[ʃ]		[ʃ]		[ʃ]	
speaker	1	1	7	7	16	16	1	1	7	7	16	16
measure	SD	PC1	SD	PC1	SD	PC1	SD	PC1	SD	PC1	SD	PC1
peak	1.22	0.1	2.89	0.02	0.25	0.39	1.37	0.01	1.02	0.02	5.52	0.01
cog	1.17	0.13	1.75	0.1	0.48	0.23	1.54	0.01	1.24	0.01	3.05	0
sdev	0.91	0.15	1.27	0.19	1	0.02	1.81	0.01	1.35	0.01	1.43	0.02
skew	1.35	0.02	2.55	0.06	3.74	0.02	1.34	0	2.54	0	0.98	0
kurt	1.62	0	4.4	0.02	15.54	0	1.72	0	2.98	0	0.87	0
dur	1.15	0.08	1.54	0.02	1.18	0.02	1.15	0.12	1.77	0.06	1.79	0.05
hmean	1.06	0.01	1.54	0.1	2.96	0.03	1.36	0	1.56	0.03	1.06	0.04
hsd	0.85	0.08	1.18	0.33	1.42	0.1	0.93	0.03	1.05	0.09	0.94	0.1
hmax	0.89	0.18	0.53	0.11	0.68	0.1	1.06	0.17	1.09	0.4	1.05	0.26
htmax	0.92	0.1	0.82	0.04	1.04	0.04	0.99	0.22	1.07	0.36	1.33	0.03
tilt	0.66	0.15	1.97	0	0.08	0.05	0.95	0.43	1.96	0.03	0.93	0.49
PC1 variace		0.38		0.56		0.56		0.46		0.52		0.55

Extracting the first PC and SD-ratio by speaker gives an idea of which cues are the most variant and the most stable within the speakers. A high PC value indicates a high variation within a speaker and a low value indicates a small variation within a speaker. In order to identify the most stable cues within a speaker and the most variant between- speakers, the PC value needs to be low and the SD-ratio high.

The data in Table VII shows that some ASFs indeed fulfil these constraints. This suggests that these cues were produced by speakers with a low degree of within-speaker variation and their acoustic characteristics differ from other speakers. This also means that these cues could potentially provide speaker-specific information.

It is striking that in both fricatives the three speakers differed greatly in the set of cues in which they produced the most and the least variation. The acoustics of speaker 1 is characterised by a lower idiosyncrasy in fricatives in comparison to the other two speakers. The SD-ratios are between 1 and 2 at the highest and the PC1 variance is 38% in [f] and 46% in [ʃ]. Speaker 7 exhibits some idiosyncrasy in both fricatives, indicated by SD-ratios between 2 and 3, and higher PC1 variance values for both fricatives. Speaker 16 has the greatest degree of individual information in fricatives. The SD-ratios reach up to 15 and the PC1 variance is almost equal in both sounds with 56% and 55%.

V. DISCUSSION

The first objective of the current study was to predict the speaker's gender and identify gender-specific traits in fricative sounds. The second objective was to predict individual speakers and investigate the distribution of intra- and inter-speaker variation in the eight fricatives. To address the defined aims and to understand how speaker information is encoded in Russian fricative sounds, various methods and techniques were applied.

The discrimination task of speakers' gender was based on machine learning models. We compared the performance of MFCCs and ASFs using two ML classifiers, *decision tree* (DT) and *random forest* (RF). The results indicate that MFCCs clearly outperform ASFs with an accuracy of 88% over an accuracy of 72%. Thus, gender can be predicted by acoustic cues and speakers' gender information is best captured by the fine-graded spectral envelope information measured by MFCCs. With moderate accuracy, gender can also be predicted from ASFs and the most important cues are *peak*, *cog*, *skew* and *hmean*.

To explore acoustic gender variation, the ASFs were compared between female and male speakers. The findings suggest that female and male speakers differ significantly in the acoustics of the eight fricatives. Gender variation is found in the min and max values of the ASFs. Also, the threshold of the range varies between male and female speakers, which indicates how large the variance is within and between the gender categories.

A second finding on gender variation shows that female and male speakers differ in how much contrast they produce between contrastive fricative categories (place of articulation, voicing, palatalisation). The analysis

demonstrated that gender variation in contrasting different fricatives is place and manner of articulation dependent. In previous studies, distance measurements were obtained for vowels (Diehl *et al.*, 1996) and for sibilant voiceless fricatives (Weirich and Simpson, 2015), suggesting that females produce more distance between two contrastive sounds and concluding that females produce more distinct categories. Our data confirm only partly these findings and give an alternative explanation taking into account the IQR analysis. It is assumed that a high IQR indicates considerable variance in the produced cue. This, in turn, means that with larger values of the IQR, we find a larger distance between the ASFs. This difference could explain why the measured distance between contrastive fricative pairs is for some pairs higher for females and others higher for male speakers. Having a wider distance between the categories does not imply that they are more distinct from one another.

Taken together, the findings suggest that the overall patterns of gender variation are less systematic across female and male speakers, but more specific to sound and acoustic cues. Gender variation was often evaluated by previous studies measuring spectral moments in [s]. The current analysis showed (IV) a large variation between fricatives of different places of articulation and voicing quality. Patterns found in [s] are not necessarily transferable to for instance [z], [sʲ] or [ʃ]. It is therefore suggested in future studies to extend gender-variation research to other fricatives.

These findings could also explain why machine learning classifiers performed only moderately when predicting gender by ASFs. Most ASFs show significant differences, so they probably all contribute to a certain extent to the distinction between males and females. Nevertheless, to test the importance of the individual ASFs, it may be necessary to compare separately, for example, the performance on spectral cues and HNR measures. Furthermore, future analyses should probably include the comparison of voiced and voiceless fricatives.

The second aim of the study was to test the predictive power of individual speakers and analyze inter- and intra- speaker variation across the eight fricatives and measured acoustic cues. To predict the speaker, we followed the same methodological approach used to predict gender from acoustic cues using machine learning techniques. The two decision tree-based algorithms were unable to identify speakers based on ASFs. In RF, however, the speaker ID could be predicted with a kappa and accuracy of 0.64 using the data set of MFCCs. These observations imply that fricative sounds do contain speaker information which can be determined by technological applications using MFCCs. This observation is not surprising, since MFCCs are successfully used in ASR.

To further explore why the ML classifiers were unable to predict speakers using ASFs, the inter- and intra-speaker variation was investigated in more detail. A PCA by sound was performed to identify the cues that explain the most variation. For the analysis, only the results of the first principal component were considered. The findings show that spectral cues explain the variation only in [f] across voiceless fricatives and to a certain degree the variation in voiced fricatives. Duration and the HNR ratio on the other hand seem to play a role in almost all fricatives with a different distribution of meaningful cues. Additionally, the SD-ratio was computed to give information on whether the difference shown by PC1 is caused by between or within-speaker variation. Consequently, it is concluded that intra- and inter-speaker variation cannot be defined by a set of cues from the measured ASFs across the eight fricatives. Most cues found by the PCA to be variant within a sound are characterized by an SD-ratio below 1 meaning that the within-speaker variation is higher than the inter-speaker variation.

A more detailed analysis of three speakers provided further insights into the distribution of variant and stable cues across speakers and sounds (Table VII). The results showed clearly how largely speakers differ in the cues they produce with a high within-speaker variation and a high constancy. Taking for instance the *peak* in [f], the data shows that while speaker 7 has an SD-ratio of 2.89 and 0.02 in PC1, speaker 16 has the opposite pattern with an SD-ratio of 0.25 and PC1 of 0.39. This in turn means that speaker 7 produces a very stable *peak*, and speaker 16 has high variation within her *peak* frequencies. The analysis demonstrates that speakers can potentially code their individual information in different cues for the same sound. Also, it can be noted that no cue seems to be consistently employed by speakers to code individual information. Variation is higher than expected, and the process is more complex than just detecting the most stable cues within and between- speakers.

From these analyses, it can be concluded that not only do feature distributions exhibit a variation that may not be consistent from one feature to the other, but the level of individuality encoded in fricatives is highly speaker-dependent, which can in turn explain why the general performances aforementioned for individual recognition were so poor using ASFs.

These conclusions contradict previous findings that suggested speaker discrimination potential in the spectral moments. One explanation is that the sample size (both the number of speakers and the number of fricative categories) employed for speaker recognition has a significant effect on performance. To clarify these questions further tests would be needed to explore speaker discrimination performance on the same data set but different data samples. On the other hand, taking into account the conclusions of the identification of place of articulation in the same data set (Ulrich *et al.*, 2021), which found that centre of gravity and spectral spread provide sufficient information to distinguish [f, s, ʃ] strong speaker effects were not expected to be found in the spectral domain.

To summarize, from this analysis it can be concluded that, feature distributions exhibit large variation across sounds and individuality encoded in fricatives is highly speaker-dependent. Furthermore, variation within the speaker across different fricatives depends on the place and manner of articulation. Intra- and inter-speaker vari-

ation is highly complex and no set of cues seems to explain acoustic variability and stability for all fricatives and speakers. Speakers can potentially code their individual information in different cues for the same sound. Whether patterns between- speakers exist and whether some speakers can be grouped together according to a similar distribution of variant and stable cues needs to be further investigated. Furthermore, it is questionable to what extent a speaker's individual variation in one sound can predict the variation in another sound. From the analysis of the two sounds, no such pattern can be obtained.

The current study has implications for phonetic research as well as for ASR applications. In phonetic research, it helps us to understand that individual speaker information is distributed in fricative sounds across all ASFs. Which underlying mechanism define speaker specif patterns and what influences the degree of freedom where a speaker can code information on their individuality needs to be further explored. We found that MFCCs contain more detailed speaker information in fricative sounds than the information that can be obtained from regular spectral, temporal and HNR measures. These findings suggest that the spectral domain contains such fine-graded information on speakers' idiosyncrasies but the spectral measures used in phonetic research do not capture this information sufficiently. The current study indicates that further measurements must be developed to capture more detailed information, similar to MFCCs, but interpretative for phonetic research. The moderate performance on predicting speakers by the MFCCs suggests that also in noisy speech sounds such as fricatives both the periodic and aperiodic parts contain potentially speaker-specific information allowing to discriminate speakers from each other.

VI. ACKNOWLEDGMENTS

We wish to thank our participants, the Phonetic Lab in St. Petersburg (and the sound engineer Tatiana Chukaeva in particular), the University of Zürich for financial support, technical support and help with the design of the experiment (to Volker Dellwo in particular). NU was partly supported by a grant from the Doctoral Program of Linguistics of the Faculty of Arts and Social Sciences, University of Zürich, Switzerland, NU, MAT was funded by the IDEXLyon Fellowship Grant 16-IDEX-0005 (2018-2021), and indirectly by the Labex ASLAN (ANR-10-LABX-0081) of the University of Lyon within the program Investissements d'Avenir (ANR-11-IDEX-0007) of the French National Research Agency (ANR). MAT is also thankful for the support of the Junior researcher grant from the French National Research Agency (ANR-20-CE27-0021).

[1]Some speakers end their voiceless fricatives with a somehow long and unexpected post-aspiration in intervocalic positions, but also when the fricative appeared at the end of the word and sentence. In these cases, the fricatives were segmented according to the changes in high-energy events and the post-aspiration part was not considered. The voiced fricatives represented also a segmentation challenge, because the waveform and spectrogram may be insufficiently informative to define the onset and offset. The boundaries for these sounds were identified according to perceptional judgments. In general, it should be noted that it is hardly possible to standardise segmentation criteria across all fricatives and speakers. Such an investigation would also require taking sociophonetics factors into account and is beyond the scope of the present study.

[2]The fricatives [x], [v^j] and [z^j] are not included in the analysis due to their low count in the data.

[3]The number 200 was chosen based on the stabilisation point of the predictions.

Ajili, M., Bonastre, J.-F., Ben Kheder, W., Rossato, S., and Kahn, J. (**2017**). "Phonological content impact on wrongful convictions in Forensic Voice Comparison context," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, New Orleans, LA, pp. 2147–2151, http://ieeexplore.ieee.org/document/7952536/, doi: 10.1109/ICASSP.2017.7952536.

Al-Tamimi, J., and Khattab, G. (**2015**). "Acoustic cue weighting in the singleton vs geminate contrast in Lebanese Arabic: The case of fricative consonants," The Journal of the Acoustical Society of America **138**(1), 344–360, http://asa.scitation.org/doi/10.1121/1.4922514, doi: 10.1121/1.4922514.

Alsulaiman, M., Mahmood, A., and Muhammad, G. (**2017**). "Speaker recognition based on Arabic phonemes," Speech Communication **86**, 42–51, https://linkinghub.elsevier.com/retrieve/pii/S0167639315300649, doi: 10.1016/j.specom.2016.11.004.

Antal, M. (**2008**). "Phonetic speaker recognition," Proc. of the 7th International Conference COMMUNICATIONS 67–72.

Barry, S. M. E. (**1995**). "Variation in vocal fold vibration during voiced obstruents in Russian," International Journal of Language & Communication Disorders **30**(2), 124–131, doi: https://doi.org/10.3109/13682829509082523.

Boersma, P., and Weenink, D. (**2022**). "Praat: doing phonetics by computer [Computer program]. Version 6.2.14" http://www.praat.org/.

Bolla, K. (**1981**). *A conspectus of Russian Speech Sounds.*, 32 ed. (Köln : Böhlau Verlag).

Breiman, L. (**2001**). "Random Forests," Machine Learning **45**, 5–32, doi: https://doi.org/10.1023/A:1010933404324.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. (**1984**). "Classification and regression trees. Wadsworth & Brooks," Cole Statistics/Probability Series .

Butcher, A. (**2003**). "Australian Aboriginal Languages Consonant-Salient Phonologies and the "Place-of-Articulation Imperative"," Australian Speech Science and Technology Association .

Dellwo, V., Huckvale, M., and Ashby, M. (**2007**). "How Is Individuality Expressed in Voice? An Introduction to Speech Production and Description for Speaker Classification," Speaker Classification I **4343**, 1–20, http://link.springer.com/10.1007/978-3-540-74200-5_1, doi: 10.1007/978-3-540-74200-5_1 iSSN: 0302-9743, 1611-3349 Series Title: Lecture Notes in Computer Science.

Diehl, R. L., Lindblom, B., Hoemeke, K. A., and Fahey, R. P. (**1996**). "On explaining certain male-female differences in the phonetic realization of vowel categories," Journal of Phonetics **24**(2), 187–208, https://linkinghub.elsevier.com/retrieve/pii/S009544709690011X, doi: 10.1006/jpho.1996.0011.

Enzinger, E., and Balazs, P. (**2011**). "Speaker Verification using Pole/Zero Estimates of Nasals," Analele Universitatii "Eftimie" **18**, 33–44.

Ferragne, E., and Pellegrino, F. (**2010**). "Formant frequencies of vowels in 13 accents of the British Isles," Journal of the International Phonetic Association **40**(1), 1–34, https://www.cambridge.org/core/product/identifier/S0025100309990247/type/journal_article, doi: 10.1017/S0025100309990247.

Flipsen, P., Shriberg, L., Weismer, G., Karlsson, H., and McSweeny, J. (**1999**). "Acoustic Characteristics of /s/ in Adolescents," Journal of Speech, Language, and Hearing Re-

search **42**(3), 663–677, http://pubs.asha.org/doi/10.1044/jslhr.4203.663, doi: 10.1044/jslhr.4203.663.
Forrest, K., Weismer, G., Milenkovic, P., and Dougall, R. N. (**1988**). "Statistical analysis of word-initial voiceless obstruents: Preliminary data," The Journal of the Acoustical Society of America **84**(1), 115–123, http://asa.scitation.org/doi/10.1121/1.396977, doi: 10.1121/1.396977 number: 1.
Ganchev, T., Fakotakis, N., and Kokkinakis, G. (**2005**). "Comparative Evaluation of Various MFCC Implementations on the Speaker Verification Task," Proceedings of the SPECOM. 191–194.
Gendrot, C., Ferragne, E., and Pellegrini, T. (**2019**). "Deep learning and voice comparison: phonetically-motivated vs. automatically-learned features," ICPhS .
Gendrot, C., Ferragne, E., and Pellegrini, T. (**2020**). "Informations segmentales pour la caractérisation phonétique du locuteur: variabilité inter-et intra-locuteurs," 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 1: Journées d'Études sur la Parole **1**.
Ghaffarvand Mokari, P., and Mahdinezhad Sardhaei, N. (**2020**). "Predictive power of cepstral coefficients and spectral moments in the classification of Azerbaijani fricatives," The Journal of the Acoustical Society of America **147**(3), EL228–EL234, http://asa.scitation.org/doi/10.1121/10.0000830, doi: 10.1121/10.0000830.
Gordon, M., Barthmaier, P., and Sands, K. (**2002**). "A cross-linguistic acoustic study of voiceless fricatives," Journal of the International Phonetic Association **32**(2), 141–174, http://www.journals.cambridge.org/abstract_S0025100302001020, doi: 10.1017/S0025100302001020.
Grandon, B., and Vilain, A. (**2020**). "Development of fricative production in French-speaking school-aged children using cochlear implants and children with normal hearing," Journal of Communication Disorders **86**, 105996, https://linkinghub.elsevier.com/retrieve/pii/S0021992420300642, doi: 10.1016/j.jcomdis.2020.105996.
Harper, S. K. (**2021**). "Individual differences in phonetic variability and phonological representation.," Ph.D. thesis, Diss. University of Southern California.
Hayward, K. (**2000**). Longman linguistics library *Experimental phonetics*, 2nd ed ed. (Longman, Harlow).
He, L., and Dellwo, V. (**2014**). "Speaker idiosyncratic variability of intensity across syllables," in *Interspeech 2014*, ISCA, pp. 233–237, https://www.isca-speech.org/archive/interspeech_2014/he14_interspeech.html, doi: 10.21437/Interspeech.2014-59.
Hughes, G. W., and Halle, M. (**1956**). "Spectral Properties of Fricative Consonants," The Journal of the Acoustical Society of America **28**(2), 303–310, http://asa.scitation.org/doi/10.1121/1.1908271, doi: 10.1121/1.1908271.
Jax, P., and Vary, P. (**2003**). "On artificial bandwidth extension of telephone speech," Signal Processing **83**(8), 1707–1719, https://linkinghub.elsevier.com/retrieve/pii/S0165168403000823, doi: 10.1016/S0165-1684(03)00082-3.
Jesus, L. M. T., and Jackson, P. J. B. (**2008**). "Frication and Voicing Classification," Computational Processing of the Portuguese Language **5190**, 11–20, http://link.springer.com/10.1007/978-3-540-85980-2_2, doi: 10.1007/978-3-540-85980-2_2.
Jongman, A., Wayland, R., and Wong, S. (**2000**). "Acoustic characteristics of English fricatives," The Journal of the Acoustical Society of America **108**(3), 1252, http://scitation.aip.org/content/asa/journal/jasa/108/3/10.1121/1.1288413, doi: 10.1121/1.1288413 number: 3.
Kavanagh, C. (**2011**). "Intra- and inter-speaker variability in acoustic properties of English /s/," International Association for Forensic Phonetics and Acoustics doi: https://doi.org/10.1121/1.3655046.
Kavanagh, C. M. (**2012**). "New consonantal acoustic parameters for forensic speaker comparison," Ph.D. thesis, University of York.
Kisler, T., Reichel, U., and Schiel, F. (**2017**). "Multilingual processing of speech via web services," Computer Speech & Language **45**, 326–347, https://linkinghub.elsevier.com/retrieve/pii/S0885230816302418, doi: 10.1016/j.csl.2017.01.005.
Kochetov, A. (**2017**). "Acoustics of Russian voiceless sibilant fricatives," Journal of the International Phonetic Association **47**(3), 321–348, doi: https://doi.org/10.1017/S0025100317000019.
Ladefoged, P., and Maddieson, I. (**1996**). *The Sounds of the World's Languages*. (Blackwell, Cambridge).
Ludger, P., Fuchs, S., and Seifert, F. (**2021**). "Differences between male and female speakers in the production of /s/: A cross-linguistic study.," 17. Phonetik und Phonologie im deutschsprachigen Raum (P&P) .
Maddieson, I., and Disner, S. F. (**1984**). *Patterns of sounds* (Cambridge University Press, Cambridge [Cambridgeshire] ; New York).
Maddieson, I., Flavier, S., Marsico, E., Coupé, C., and Pellegrino, F. (**2013**). "LAPSyd: lyon-albuquerque phonological systems database," in *Interspeech 2013*, ISCA, pp. 3022–3026, https://www.isca-speech.org/archive/interspeech_2013/maddieson13b_interspeech.html, doi: 10.21437/Interspeech.2013-660.
Matovski, D. S., Nixon, M. S., Mahmoodi, S., and Carter, J. N. (**2010**). "The effect of time on the performance of gait biometrics," 2010 Fourth IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS) 1–6, http://ieeexplore.ieee.org/document/5634547/, doi: 10.1109/BTAS.2010.5634547.
McDougall, K., and Nolan, F. (**2007**). "Discrimination of speaker using the formant dynamics of /u/ in British English.," Proceedings of the International Congress of Phonetic Sciences 1825–1828, http://icphs2007.de/conference/Papers/1567/1567.pdf.
Monson, B. B., Hunter, E. J., Lotto, A. J., and Story, B. H. (**2014**). "The perceptual significance of high-frequency energy in the human voice," Frontiers in Psychology **5**, http://journal.frontiersin.org/article/10.3389/fpsyg.2014.00587/abstract, doi: 10.3389/fpsyg.2014.00587.
Moore, B. C. J. (**2003**). "Coding of Sounds in the Auditory System and Its Relevance to Signal Processing and Coding in Cochlear Implants:," Otology & Neurotology **24**(2), 243–254, http://journals.lww.com/00129492-200303000-00019, doi: 10.1097/00129492-200303000-00019.
Munson, B., McDonald, E. C., DeBoe, N. L., and White, A. R. (**2006**). "The acoustic and perceptual bases of judgments of women and men's sexual orientation from read speech," Journal of Phonetics **34**(2), 202–240, https://linkinghub.elsevier.com/retrieve/pii/S0095447005000379, doi: 10.1016/j.wocn.2005.05.003.
Narayanan, S. S., Alwan, A. A., and Haker, K. (**1995**). "An articulatory study of fricative consonants using magnetic resonance imaging," The Journal of the Acoustical Society of America **98**(3), 1325–1347, http://asa.scitation.org/doi/10.1121/1.413469, doi: 10.1121/1.413469.
Newman, R. S., Clouse, S. A., and Burnham, J. L. (**2001**). "The perceptual consequences of within-talker variability in fricative production," The Journal of the Acoustical Society of America **109**(3), 1181–1196, http://scitation.aip.org/content/asa/journal/jasa/109/3/10.1121/1.1348009, doi: 10.1121/1.1348009.
Romeo, R., Hazan, V., and Pettinato, M. (**2013**). "Developmental and gender-related trends of intra-talker variability in consonant production," The Journal of the Acoustical Society of America **134**(5), 3781–3792, http://asa.scitation.org/doi/10.1121/1.4824160, doi: 10.1121/1.4824160.
Rose, P. (**2007**). "Forensic speaker discrimination with Australian English vowel acoustics," ICPhS XVI **6**(10).
Schiel, F. (**1999**). "Automatic Phonetic Transcription of Non-Prompted Speech," .
Schindler, C., and Draxler, C. (**2013**). "Using spectral moments as a speaker specific feature in nasals and fricatives," in *Interspeech 2013*, ISCA, pp. 2793–2796, https://www.isca-speech.org/archive/interspeech_2013/schindler13_interspeech.html, doi: 10.21437/Interspeech.2013-639.
Schwartz, M. F. (**1968**). "Identification of Speaker Sex from Isolated, Voiceless Fricatives," The Journal of the Acoustical Society

of America **43**(5), 1178–1179, http://asa.scitation.org/doi/10.1121/1.1910954, doi: 10.1121/1.1910954.
Silbert, N., and de Jong, K. (**2008**). "Focus, prosodic context, and phonological feature specification: Patterns of variation in fricative production," The Journal of the Acoustical Society of America **123**(5), 2769–2779, http://asa.scitation.org/doi/10.1121/1.2890736, doi: 10.1121/1.2890736.
Skarnitzl, R., and Machač, P. (**2011**). "Principles of Phonetic Segmentation," Phonetica **68**(3), 198–199, https://www.degruyter.com/document/doi/10.1159/000331902/html, doi: 10.1159/000331902.
Smorenburg, L., and 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), 949–960, https://asa.scitation.org/doi/full/10.1121/10.0000674, doi: 10.1121/10.0000674 publisher: Acoustical Society of America.
Spinu, L., Kochetov, A., and Lilley, J. (**2018**). "Acoustic classification of Russian plain and palatalized sibilant fricatives: Spectral vs. cepstral measures," Speech Communication **100**, 41–45, https://linkinghub.elsevier.com/retrieve/pii/S0167639317303680, doi: 10.1016/j.specom.2018.04.010.
Spinu, L., and Lilley, J. (**2016**). "A comparison of cepstral coefficients and spectral moments in the classification of Romanian fricatives," Journal of Phonetics **57**, 40–58, https://linkinghub.elsevier.com/retrieve/pii/S0095447016300109, doi: 10.1016/j.wocn.2016.05.002.
Strevens, P. (**1960**). "Spectra of Fricative Noise in Human Speech," Language and Speech **3**(1), 32–49, http://journals.sagepub.com/doi/10.1177/002383096000300105, doi: 10.1177/002383096000300105 number: 1.
Stuart-Smith, J. (**2007**). "Empirical evidence for gendered speech production: /s/ in Glaswegian," Mouton de Gruyter .
Timberlake, A. (**2004**). *A Reference Grammar of Russian* (Cambridge University Press).
Ting, K. M. (**2010**). "Precision and Recall.," Encyclopedia of machine learning **781**.
Ulrich, N., Allassonnière-Tang, M., Pellegrino, F., and Dediu, D. (**2021**). "Identifying the Russian voiceless non-palatalized fricatives /f/, /s/, and // from acoustic cues using machine learning," The Journal of the Acoustical Society of America **150**(3), 1806–1820, https://asa.scitation.org/doi/10.1121/10.0005950, doi: 10.1121/10.0005950.
Van der Maaten, L., and Hinton, G. (**2008**). "Visualizing data using t-SNE.," Journal of machine learning research **9**(11).
Weirich, M., and Simpson, A. P. (**2014**). "Differences in acoustic vowel space and the perception of speech tempo," Journal of Phonetics **43**, 1–10, https://linkinghub.elsevier.com/retrieve/pii/S0095447014000023, doi: 10.1016/j.wocn.2014.01.001.
Weirich, M., and Simpson, A. P. (**2015**). "Gender-specific differences in sibilant contrast realizations in English and German," ICPhS .

Chapter 6

Discussion

In the following section, the results and the conclusions from both studies are summarized and discussed within the context of the asked research questions. Then, the scientific contributions are described. Finally, the investigation ends with an outline of the limitations of the current project and an outlook on further research in fricatives.

6.1 How can machine learning techniques contribute to automatizing and standardising the segmentation of noise duration in voiceless fricatives?

The first question was motivated by the attempt to extract acoustic cues from the entire consonant duration and from the entire noise duration of a fricative. This improves from previous studies, which mostly used a predefined window (e.g. Jongman et al., 2000; Kochetov, 2017).

As a first step towards investigating the acoustic characteristics of speech sounds, the target phonemes need to be aligned and segmented out of the speech sequence. This process is highly time-consuming. More precisely, the time to transcribe speech and then to time-align phonemes is reported to be approximately 800 times longer than the processed speech segment (Schiel et al., 2012). Over the past decade, a number of significant computational methods have been developed to advance transcription, forced alignment, and phoneme segmentation (Gonzalez et al., 2020). Common forced aligners are for instance the Munich Automatic Segmentation System (MAUS) (Schiel, 1999), the Forced Alignment & Vowel Extraction suite (FAVE) (Rosenfelder et al., 2014), the Language, Brain and Behaviour Corpus Analysis Tool (LaBB-CAT) (Fromont and Hay, 2012), and the Montreal Forced Aligner (MFA) (McAuliffe et al., 2017)). In one study, the performance across these four programs was compared. Furthermore, the accuracy was evaluated between the aligners and humans. The results demonstrate differences in performance across the programs. MFA and LaBB-CAT show the highest alignment quality, followed by FAVE and MAUS. Furthermore, humans show more constancy in a segmentation task and manual correction improves alignment accuracy (Gonzalez et al., 2020). In the current investigation, all raw audio files were first pre-processed by applying the Munich Automatic Segmentation System (MAUS) (Schiel, 1999).

The performance was influenced by several factors, including the phonological context, as

found in previous studies (DiCanio et al., 2012). Moreover, differences were observed between speakers. This might be caused, among other factors, by the speaking rate, which was reported earlier to affect the automatic segmentation performance (MacKenzie and Turton, 2020). In general, a boundary correction was needed for all target tokens to define the duration of the fricatives, which was later used for the extraction of acoustic cues. The predefined settings for the on and offset of the fricatives included thereby some parts of the transition zones. Furthermore, obtaining acoustic measures including only frication noise, implied a repeated manual adjustment of the boundaries. In this step, transition zones and therefore co-articulation effects were excluded.

To avoid a second manual segmentation of all target tokens, a method based on training a tree-based computational classifier was introduced. The model is built on the assumption that the zero crossing rate provides sufficient information to divide a speech signal into purely aperiodic and periodic portions. Noise is in general defined as an aperiodic signal with high frequencies and therefore a high amount of zero crossings in a given time, i.e., a high zero crossing rate (or *zcr*). This is known to detect the voiced and unvoiced parts of speech. In the current study, it was used to detect the frication noise part in fricatives. A "gold standard" subset of 560 fricative sounds, which had their noise duration identified manually, was generated to annotate each window with `noise = TRUE` or `noise = FALSE` depending on its occurrence within or outside the noise part identified manually. This data served to train the model.

With the model applied in the current research, each sound is separated into *windows* based on a certain amount of zero crossing points. The zero crossing rate (*zcr*) within each window is then measured and compared with the zero crossing rate of the preceding window (if any). The difference of zero crossing rate between the two windows (*diff*) is then computed and used as a cue to identify the beginning and the end of the noise part of a sound. To have a better understanding of which settings are optimal for the model, we tested various window lengths (here, 64, 128, 256 or 512 points) with different levels of overlap (0%, 30%, 50% or 80%). The model shows that all combinations of parameters result in an accuracy between 78% and 83%, with the best accuracy being found for a large window length (512 zero crossings) and a standard overlap (50%), with mean = median = 80.8% across the 100 replications.

Applying this model to the manually pre-segmented fricatives allowed for the extraction of the full noise duration with a minimum of co-articulation effects. The extracted noise part was then used to extract acoustic cues and to compare the results with acoustic extraction techniques applied in previous research. The output of that comparison is analysed in the second research question.

6.2 What is the effect of window length in extracting acoustic cues from voiceless fricatives?

To extract acoustic cues, most studies used single spectral slices from the middle and sometimes the beginning and end of the fricative or of the frication noise, with window sizes between 25ms (Kochetov, 2017) and 40ms (Jongman et al., 2000). In general, acoustic cues for fricatives were not extracted from the full duration of the consonant or from its noise part.

In the current investigation, the goal was both to follow the examples of previous studies, as well as to develop a new approach. Therefore, three acoustic cue extraction techniques (ACETs) were applied, using either the full consonant duration (AC), its middle 30ms (AW),

or only the noise part of each sound file (NF) defined by the step before. Figure 6.1 shows the values and ranges of acoustic cues extracted with different ACETs.

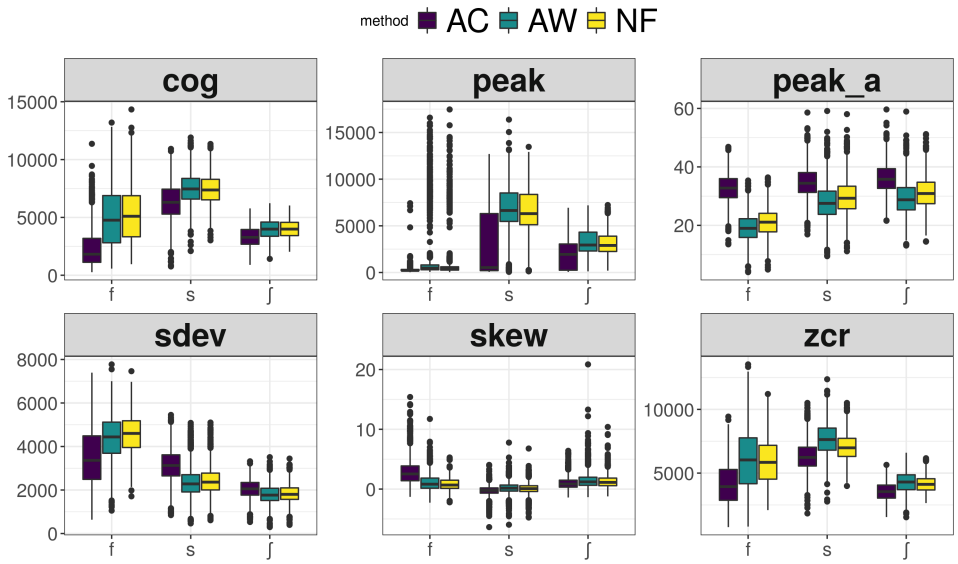

Figure 6.1: The comparison of acoustic cues based on the three main ACETs reported in the experiments. The names of the ACETs refer to the acoustic cue extraction techniques.

The three sets of cues were used by four machine learning classifiers to predict the place of articulation in the three voiceless fricatives [f], [s], and [ʃ]. The accuracy of classifying the fricatives from acoustic cues does not vary much among these ACETs, nor among the four classifiers. However, differences do exist and are informative scientifically and methodologically. All four classifiers perform far above the majority baseline of 44% accuracy, reaching about 93% and 98% across ACETs. The accuracy based on a single decision tree is generally lower than the other three classifiers (as expected, given that this has the simplest architecture), but, importantly, random forests perform almost at ceiling. This result is very meaningful as there is a high interpretability of the decision rules used. In terms of ACETs, extracting acoustic measurements from the full noise duration works better than from a 30ms window (e.g., for *cog*, *sdev* and *peak*) for all three fricatives, and especially for [f]. That is to say, the most invariant parameters are the ones estimated from the largest section that doesn't show a strong co-articulatory effect.

6.3 Can the Russian fricatives [f], [s] and [ʃ] be correctly classified by a set of acoustic cues? And how does the performance of the models differ between using ASFs or/and MFCCs?

Apart from the slightly different results of the ACETs, there is clear evidence that acoustic cues do contain enough information for the correct classification of the Russian fricatives [f], [s] and [ʃ]. A few acoustic cues seem to be necessary and sufficient, including *cog*, *sdev* and possibly

zcr and *peak*. The importance of *sdev* echoes previous studies emphasizing the importance of dynamical features and spectro-temporal variations in identifying fricatives (Reidy, 2016).

Comparing the results of spectral and temporal cues with the previous findings, there are both overlaps and differences. *Spectral peak location* is probably one of the most promising cues in the literature, but our classifiers did not find it as crucial for distinguishing fricatives. As for Greek fricatives (Nirgianaki, 2014), no clear decrease in frequency as the place of articulation moves from front to back was observed. In all ACETs, *peak* is a distinct measure for most of the [s] and [ʃ] sounds, but not for [f], where there is no clear peak. In our data, *cog* is the most important cue for distinguishing [f], [s], and [ʃ]. Higher values were reported for sibilants than for non-sibilants, and for [s] than for [ʃ] (Funatsu and Kiritani, 1998; Jongman et al., 2000; Nittrouer et al., 1989; Padgett and Zygis, 2007; Zsiga, 2000), which our data confirm, to a certain extend: [f] has the lowest values around 4000Hz (but reaching up even above 7000Hz) while the energy of [s] is centred around 7500Hz and of [ʃ] around 4500Hz.

Despite the *spectral spread* being much less considered in the literature, we found that this is one of the most important cues in our data: the lowest spread was measured for [ʃ] and the highest for [f] (Jongman et al., 2000; Shadle and Mair, 1996).

For the other two spectral moments, *skewness* and *kurtosis*, our results did not match with previous findings suggesting that these two cues are stable characteristics of fricatives (McFarland et al., 1996; Nittrouer et al., 1989). Not only there are no significant differences across the methods, but both measures are plagued by many outliers. In general, we see a slightly lower (or negative) skew for [s] than for [f] and [ʃ], supporting earlier findings of a negative skewness for [s] and a positive one for [ʃ] (Jongman et al., 2000; McFarland et al., 1996; Nittrouer et al., 1989). For kurtosis, the differences are even less relevant and difficult to interpret.

Temporal measures, such as the full consonant duration and the frication noise duration, are not distinct cues in the current data. Only the zero crossing rate seems to contain relevant information but is not an important cue for distinguishing [f], [s] and [ʃ].

The second part of the third research question deals with the comparison of the predictive power between the ASFs and MFCCS. The results do not show a large difference in performance between the acoustic measures and the MFCCs. It is strikingly smaller than that reported in the literature. In fact, while the MFCCs perform better than the acoustic measures (formally, statistically significantly so), this difference is very small in terms of effect size (less than 2% accuracy), with both performing effectively at ceiling (above 97% for random forests, SVMs and neural nets). This difference is even smaller when the full frication noise is used (The fact such small real-world differences are statistically significant here is due to the very small variation between replications). Thus, both methods are very good at classifying the sounds [f], [s] and [ʃ], showing that the information necessary for correctly classifying these three fricatives can be extracted in several manners.

Furthermore, the performance of models trained with both acoustic cues and MFCCs was also considered.[1] While the results indicate that merging acoustic cues and MFCCs does not result in a better performance than the MFCCs, the ranking of the variables represents a mix between acoustic cues and MFCCs, suggesting that further studies should investigate how such acoustic cues are captured by the MFCCs. More precisely, it is not possible at this point to determine whether the absence of improvement observed when both acoustic cues and MFCCs are considered is due to the simplistic merging approach or to a ceiling effect related to the

[1]See supplementary material 2 at [URL will be inserted by AIP] for the detailed output.

somehow limited variability offered by our corpus. The choice of which manner to use should therefore depend on the particular research question or practical application at hand, each having its advantages and disadvantages: the MFCCs are probably more appropriate in an engineering context, while the acoustic measures give more insight into the articulatory and perceptual mechanisms relevant for fundamental research.

It is also important to note that our approach here is to use the acoustic cues to classify the fricative sounds and to identify which cues matter the most, in contrast to, for example, McMurray and Jongman (2011), which, within a regression framework, tries to find statistically significant differences for a cue given the type of fricative sound. We replicated and extended the methodology in McMurray and Jongman (2011) using a maximum-likelihood mixed effects regression approach where the value of given cue is predicted from the *method* (the ACETs), the sound *classification* ([f], [s] or [ʃ]) and their interaction as the predictors of interest, controlling for sentence *type* (carrier or normal sentence), fricative *position* (beginning, middle or end), the sounds *preceding* and *following* the fricative (several classes) and *sex* (F/M) s fixed effects, and for *sentence* and *speaker* as random effects (sentence embedded within speaker). In a nutshell, our findings[2] suggest that, as expected, there is a high similarity within speakers and sentences for all cues (high intra-class correlations), and that there are significant differences between sounds for all cues, with varying influences of sentence type, fricative position and context, but, again, not of sex. While being concordant with our machine learning results and confirming that indeed, acoustic cues differ between fricatives, these results cannot be directly used to *classify* fricatives *from* acoustic measures as our classifiers do and which, arguably, is the relevant question both scientifically and practically.

Interestingly, in predicting fricatives from acoustic cues, the vowel context does not seem to matter, as is also the case for the speaker's sex and identity, suggesting that we may have identified *context-independent characteristics* of the fricative sounds themselves beyond and above the effects of phonetic context (Mann and Repp, 1980; Nirgianaki, 2014; Soli, 1981; Stevens, 1998), and of sex and other individual-specific factors (Hughes and Halle, 1956; Jongman et al., 2000; Kochetov, 2017; Nirgianaki, 2014).

These results partly contradict those of previous studies, but also bring new information for current research on the identification of speaker-specific properties, which found considerable speaker variation in the spectral cues of fricative sounds. How much speaker information these cues provide is approached by the following research questions.

6.4 Can speakers' gender be predicted by acoustic cues? And how does the performance of the models differ between using ASFs or/and MFCCs?

Clarifying to which extent common ASFs provide speaker information motivated the investigation of speaker-specific properties in the spectral domain and of further acoustic correlates. At first, it was asked whether a set of acoustic cues extracted from eight fricatives can be identified to predict speakers' gender, using two machine learning classifiers, *Decision Tree* (DT) and *Random Forest* RF.

[2]See supplementary material 4 at [URL will be inserted by AIP] for the detailed output of the regression analysis.

The results suggest that gender can be predicted by ASFs with moderate accuracy and the most relevant cues are *peak*, *cog*, *skew* and *hmean*. Additionally, the performance between ASFs and MFCCs was compared. The output indicates that MFCCs clearly outperform ASFs with an accuracy of 88% over an accuracy of 72%. Thus, gender can be predicted by acoustic cues and speakers' gender information is best captured by the fine-graded spectral envelope information measured by MFCCs. These results are in line with previous findings on a subset of Russian (Spinu et al., 2018), Azerbaijani (Ghaffarvand Mokari and Mahdinezhad Sardhaei, 2020), and Romanian (Spinu and Lilley, 2016) fricatives. These studies also reported that cepstral coefficients clearly outperform common spectral measures. The difference in accuracy was very similar with classification rates around 60% for ASFs and around 80% and higher for CCs.

For further exploration of gender variation in the eight sounds, the ASFs were compared between female and male speakers. The results show that contrary to previous studies which identified a greater gender variation in anterior fricatives (Kochetov, 2017), significant gender variation was detected in all three places of articulation and for almost all measured cues.

In previous studies, higher spectral energy was measured in female than in male speakers in voiceless sibilant fricatives (Flipsen et al., 1999; Jongman et al., 2000; Kochetov, 2017; Ludger et al., 2021; Newman et al., 2001; Schwartz, 1968). In the current data, the same is true for the voiceless, palatal and voiced sibilants and for the non-sibilant [v]. The spectral energy in the voiceless non-sibilant [f], on the other hand, is lower in female speakers, as found in lower values in *peak* and *cog*. In the two palatal sibilant fricatives, the observed patterns are convergent with previous studies, reporting higher spectral energy in female productions (Kochetov, 2017). To mention another general observation, previous research stated that the production of Russian palatal fricatives involves lower spectral energy than non-palatal fricatives (Kochetov, 2017). In the current data, the same pattern accounts only for the alveolar fricatives [s] and [sʲ]. The post-alveolar pair [ʃ] and [ɕ] follow the opposite trend and the values are higher for the palatal [ɕ].

The second spectral moment indicating spectral spread is less explored in the literature. In the current data, significant gender differences and higher values were measured for female speakers in all sibilants except [sʲ]. This is also theoretically expected as the spectral spread is correlated with the centre of gravity, i.e., a higher *cog* leads to a higher *sdev*.

More findings are reported concerning spectral skewness. In [s], a tendency for negative skewness in female speakers and positive values or values centred near zero for male speakers was observed (Flipsen et al., 1999; Ludger et al., 2021).The female speakers in the present analysis show a more symmetrical distribution of energy in [s], as reflected by a skewness around zero. Male speakers generate significantly more energy at lower frequencies, with mean values around 0.5. In the representations of [f] and [ʃ], both genders exhibit an asymmetrical distribution of energy, with higher positive values for female speakers and therefore a predominance of energy at lower frequencies. The highest skewness was measured in the voiced bilabial fricative, and female speakers also produce more energy in lower frequency bands than male speakers in [v]. In the realisations of [z] and [ʒ] male speakers show higher *skew* than females.

Kurtosis describes the peakedness of the energy distribution and has not yet been reported for gender variation. In the present data, only the spectral energy in the alveolar [s] and [sʲ] is normally distributed, as specified by a kurtosis of around 3. All other sounds show different degrees of peaked distribution, while the values decrease in both females and males with the place of articulation moving backwards. Kurtosis of voiced fricatives displays a huge variation,

with values over 1000 in [v]. Such high values observed in [v] suggest a very compact spectral distribution. Thus, the analysis indicates significant gender variation for most sounds. However, the results of kurtosis are challenging to interpret because the range of positive kurtosis above 3 exhibits a large variation across the sounds.

Gender variation is further present in duration in the non-palatal fricatives. Female speakers produce longer duration in voiceless fricatives and male speakers in voiced fricatives.

While gender differences were widely investigated in the spectral domain of voiceless fricatives, very limited information is known about further noise features such as the distribution of periodic and aperiodic energy. One way to look at it is through analyses of the harmonic-to-noise ratio (HNR). The ASFs *hmean* and *hmax* measure the distribution of harmonic and noise energy. Values around zero imply equal energy in harmonics and noise. The overall trend for non-palatal fricatives is a decrease in harmonic energy as the place of articulation moves backwards. Significant gender variation was identified in some of the HNR measures. The data shows that the distribution of periodic and aperiodic energy generally differs between female and male speakers. For instance, the *hmean* in female speakers contains higher harmonic proportions in voiceless fricatives than in male speakers, except for the fricative [ɕ]. In the voiced bilabial [v] and alveolar [z] fricatives, female speakers follow the same pattern and produce more harmonic energy than male speakers. And the opposite is the case for the post-alveolar [ʒ]. Furthermore, female speakers have higher values in *hsd* in all fricatives, except for [v]. This finding was expected since the mean of HNR is correlated with the standard deviation of the HNR mean.

For the exploration of additional aspects of acoustic gender variation, the interquartile range (IQR) was measured. The results show that gender variation exists in the min and max values of the ASFs. The threshold of the range also differs between female and male speakers, which indicates how large the produced variance is within and between gender categories. Generally speaking, female speakers exhibit higher variance in acoustic cues than male speakers. However, this does not account for all ASFs. The findings suggest further that the overall patterns of gender variation in IQR are less systematic across gender, but more specific to sound and cue.

In addition to the comparison of mean and IQR values between female and male speakers, distance measures were obtained across the eight fricatives testing the contrast of place of articulation, voicing and palatalisation. In previous studies, the distance was measured between vowels (Diehl et al., 1996) and sibilant voiceless fricatives (Weirich and Simpson, 2015) contrasted by place of articulation. Both studies conclude that females produce more distance between two contrastive sounds and therefore more distinct sound categories. The observations made for voiceless sibilants by previous research (Weirich and Simpson, 2015) can be extended to other fricative pairs contrasted by place of articulation. For both non-palatal voiceless [f]-[s], [s]- [ʃ], and voiced [v] -[z], [z]- [ʒ] pairs, contrasted by place of articulation, female speakers produce in general a larger distance between the tested contrastive fricative pairs. However, in the palatal fricatives [sʲ]- [ɕ] no such gender variation is observed. Analysis of fricative pairs contrasted by voicing ([f-v], [s-z], [ʃ- ʒ]) suggests trends based on places of articulation. Females produce less contrast between the bilabial [f]-[v] and post-alveolar [ʃ]-[ʒ] pairs and more contrast in the alveolar pair than male speakers. Female speakers also show a larger distance between the two sibilant fricative pairs contrasted by palatalization ([s-sʲ], [ʃ- ɕ]).

The current data provide an alternative explanation for the observed results taking into account the IQR analysis, than the previous conclusions drawn in (Diehl et al., 1996; Weirich and Simpson, 2015). It is assumed that a high IQR indicates high variance in the measures of

a cue. This, in turn, means that with increasing values of the IQR, a greater distance between the ASFs can be detected. This difference could explain why the measured distance between contrastive fricative pairs is for some pairs larger for females and others larger for male speakers. This does not imply that categories with a greater distance of values in the same cue and sound are more distinct from each other, but only that these sounds were produced with a high degree of variation. However, the results suggest that the IQR and the distance produced between contrastive sounds seem to be irrelevant gender-specific properties. Therefore, it is questioned whether the variation and distance produced between contrastive sounds represent a perceptual cue for gender recognition.

Taken together, the findings suggest that female and male speakers differ significantly in their acoustic characteristics, but the overall patterns of gender variation are less systematic across female and male speakers, and more specific to sound and acoustic cues. Gender variation was often evaluated by previous studies measuring spectral moments in [s]. The current analysis shows a large variation between fricatives of different places of articulation and voicing quality. Patterns found in [s] are not necessarily transferable to for instance [z], [s^{j}] or [ʃ]. These findings could also explain why machine learning classifiers performed only moderately when predicting gender by ASFs. Most ASFs show significant differences, so they probably all contribute to a certain extent to the distinction between males and females.

6.5 Can speakers' ID be predicted by acoustic cues? And how does the performance of the models differ between using ASFs or/and MFCCs?

In the next step, the same questions as for gender variation were asked to investigate inter- and intra- speaker variation. First, the two corresponding machine learning classifiers were applied to predict speakers' IDs from ASFs and MFCCs. The results demonstrate that the two decision tree-based algorithms are unable to identify speakers based on ASFs. However, using the dataset of MFCCs, speakers could be predicted with a kappa and accuracy of 0.64 in RF. This evidence gives reasons to believe that fricative sounds do contain speaker information which can be determined by technological applications. This is not surprising, since MFCCs are successfully used in ASR.

The finding that the classifiers fail to predict speakers' ID by ASFs could be attributed to the fact that the dataset is not suitable for machine learning applications of this type, since there is a high number of speakers (59) to predict and a relatively low number of tokens per sound and speaker. Another explanation is that common ASFs measured in fricative sounds may not contain as much precise speaker information, or that speaker variation in general is more complex to evaluate. Evidence that fricative sounds do contain certain speaker information in the spectral moments was found by previous research (Kavanagh, 2011; Newman et al., 2001; Schindler and Draxler, 2013). To explore this complexity of speaker information encoding and individual differences in fricative sounds, the inter- and intra- speaker variation in the ASFs was investigated in more detail.

The application of identical methods employed to identify gender variation within the measures and across the eight fricatives did not lead to an interpretable output. Therefore, a Principal Component Analysis by sound was performed, to define cues that explain the most variation within each sound category. For a better overview, only the results of the first prin-

cipal component were reported and discussed. The PC1 explains thereby between 32% (in [f]) and 57% ([v,z]) of the variation in all realisations of each fricative.

Interestingly, spectral cues seem to explain the variation only in [f] across the voiceless fricatives with the most variable cues of *peak, cog, sdev* and *hsd*. All voiceless sibilants show similar patterns with the most relevant features *dur, hmax, htmax, tilt*. Among the three voiced fricatives, the PCA identified in [v] *skew* and *kurt* as the most variable cues. Furthermore, similar patterns were observed across the two voiced sibilants. For both sounds, some variation is contained by *cog* and *sdev*, but more by *hmean* and *hsd*.

In general, the PCA provide a good overview of variant cues across the fricatives, but it does not show the source of the variety. In particular, it is unclear whether the discovered variation results from a high degree of between or within speaker variation. To clarify this question the SD-ratio was computed for each cue and sound. A high SD-ratio indicates that a given cue is produced by speakers with a high between-speaker variation and a low within-speaker variation. Values below one mean that the inter-speaker variation is higher than the intra-speaker variation.

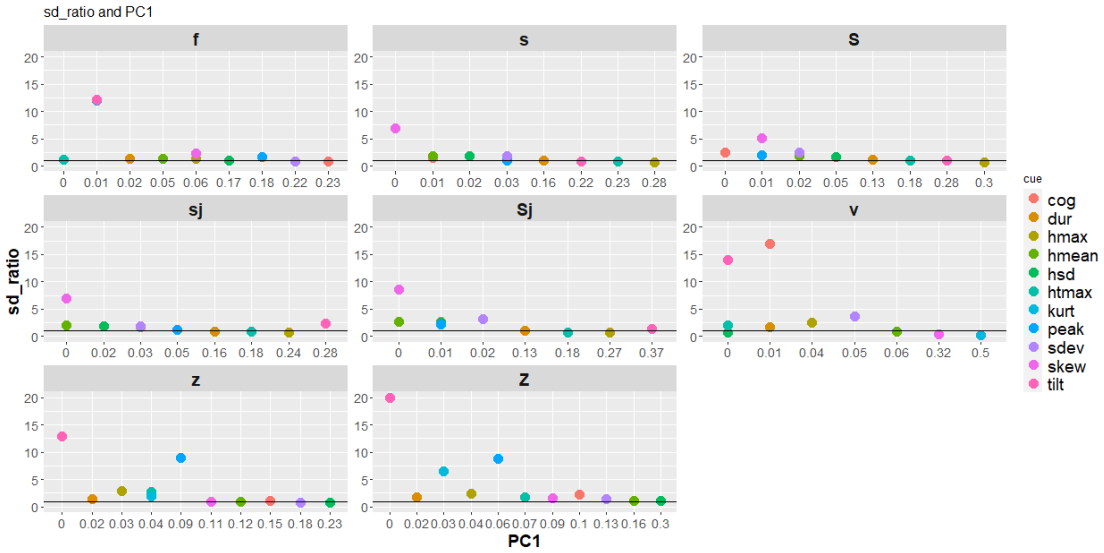

Figure 6.2: Comparison of the PC1 on the x-axis and SD-ratio on the y-axis by cue and sound. The PC1 value shows the variation within a sound which can be explained by this cue. And as higher the SD-ratio the higher the between-speaker variation. The SD-ratio below 1 means that the within-speaker variation is higher than the between-speaker variation.

The comparison in Figure 6.2 shows the distribution of inter- and intra- speaker variation across the tested fricatives and acoustic cues. Combining these two methods, it can be expected that ASFs with high values in PC1 explain a certain amount of variation and a high SD-ratio indicate that a given cue is produced by speakers with a high between speaker variation and a low within speaker variation. Consequently, these cues could potentially contain speaker-specific information. The visualisation in Figure 6.2 suggests that intra- and inter-speaker variation cannot be defined by the measured set of cues across the eight fricatives. Most acoustic features found by the PCA to be variable within a sound are characterized by an SD-ratio below 1 meaning that the within-speaker variation is higher than between speakers. Low values in PC1 on the other hand show for a set of cues a high SD-ratio. This means that most cues detected by the PCA explaining the most variation are produced by the speakers

with a high degree of within-speaker variation. Therefore, it is challenging to determine which cues could potentially serve to encode speaker information in fricative sounds.

In regard to previous research which identified speaker variation and a high speaker discrimination potential in the spectrum of voiceless fricatives, these findings are somewhat surprising. On the other hand, taking into account the conclusions of the first study, in which *cog* and *sdev* were identified as reliable cues for the place of articulation independent of vowel context, gender and speaker, it was not expected to find these cues to be most prominent to distinguishes between speakers.

6.6 How do speakers differ in their acoustic characteristics on the individual level?

To understand better the reported findings for inter- and intra- speaker variation across the measured ASFs and sounds, the acoustic characteristics of three speakers were compared in more detail. Thereby, the interaction between PC1 and SD-ratio was assessed across the three speakers for the two sounds [f] and [ʃ] and is summarised in Table 6.1.

Table 6.1: PC1 and SD-ratio for three speakers and two sounds. The three speakers show remarkable differences in the distribution of the ratio between PC1 and SD-ratio across the two fricatives [f] and [ʃ].

sound	[f]		[f]		[f]		[ʃ]		[ʃ]		[ʃ]	
speaker	1	1	7	7	16	16	1	1	7	7	16	16
Measure	SD	PC1	SD	PC1	SD	PC1	SD	PC1	SD	PC1	SD	PC1
peak	1.22	0.1	2.89	0.02	0.25	0.39	1.37	0.01	1.02	0.02	5.52	0.01
cog	1.17	0.13	1.75	0.1	0.48	0.23	1.54	0.01	1.24	0.01	3.05	0
sdev	0.91	0.15	1.27	0.19	1	0.02	1.81	0.01	1.35	0.01	1.43	0.02
skew	1.35	0.02	2.55	0.06	3.74	0.02	1.34	0	2.54	0	0.98	0
kurt	1.62	0	4.4	0.02	15.54	0	1.72	0	2.98	0	0.87	0
dur	1.15	0.08	1.54	0.02	1.18	0.02	1.15	0.12	1.77	0.06	1.79	0.05
hmean	1.06	0.01	1.54	0.1	2.96	0.03	1.36	0	1.56	0.03	1.06	0.04
hsd	0.85	0.08	1.18	0.33	1.42	0.1	0.93	0.03	1.05	0.09	0.94	0.1
hmax	0.89	0.18	0.53	0.11	0.68	0.1	1.06	0.17	1.09	0.4	1.05	0.26
htmax	0.92	0.1	0.82	0.04	1.04	0.04	0.99	0.22	1.07	0.36	1.33	0.03
tilt	0.66	0.15	1.97	0	0.08	0.05	0.95	0.43	1.96	0.03	0.93	0.49
PC1 variace		0.38		0.56		0.56		0.46		0.52		0.55

Table 6.1 shows clearly how large speakers can differ in the degree of variation and constancy across the measured cues. Taking for instance the *peak* in [f], the data shows that while speaker 7 has an SD-ratio of 2.89 and 0.02 in PC1, speaker 16 has the opposite pattern with an SD-ratio of 0.25 and PC1 of 0.39. This in turn means that speaker 7 produces a very stable *peak*, and speaker 16 has high variation within *peak* frequencies. Similar findings can be observed for most of the ASFs in all eight fricatives.

From this analysis, it can be concluded that the intra- and inter-speaker variation is highly complex and no set of cues seems to explain acoustic variability and stability for all fricatives and speakers. Speakers can potentially code their individual information in different cues for the

same sound. Feature distributions exhibit a large variation across sounds and the individuality encoded in fricatives is highly speaker-dependent. Furthermore, the within-speaker variation across different fricatives depends on place and manner of articulation. Whether patterns between speakers exist and whether some speakers can be grouped together according to a similar distribution of variant and stable cues needs to be further investigated. Furthermore, it is questionable to what extent a speaker's individual variation in one sound can predict the variation in another sound.

6.7 Contributions to linguistic and speaker research in fricative sounds

The first contribution of the present work is the generation of a large database on Russian fricatives. The database provides additional materials to the already existing databases, such as the open-source Russian language dataset – OpenSTT, (available online at https://github.com/snakers4/open_stt) for the study of underrepresented languages in account-phonetic research and beyond.

The two follow-up studies on linguistic and speaker characteristics have a significant impact on several current debates in phonetics, linguistic typology, forensic speaker comparison and ASR. Furthermore, the analysis gives examples of how machine-learning techniques can be applied to acoustic studies.

The analyses provide evidence that there may be a set of acoustic cues (*cog* and *sdev*) that can reliably distinguish the Russian fricatives [f], [s], and [ʃ]. This supports the invariant theory and suggests that stable and descriptive acoustic characteristics can be extracted from speech signals (Blumstein and Stevens, 1981). The performance of identifying the three fricatives by the ASFs is thereby very high and similar to the performance of MFCCs. Both sets of measures seem to capture equally accurate information on the distinct features of fricative sounds. The results also support the view that the configuration of the vocal tract during the production of fricatives shapes their spectrum, with the relevant spectral cues not residing primarily in the frequency of the highest amplitude, but in the spectral mean and spread. However, more research is needed in this direction.

In research focusing on speaker characteristics in complex sounds, the dissertation contributes to a better understanding of how individual speaker information is distributed in sounds like fricatives across common ASFs and MFCCs. The output display that speakers' gender and speakers' ID can be best predicted by MFCCs. ASFs seem to provide less idiosyncratic information. While gender could be predicted by these cues, the classifiers fail to identify speakers' IDs. Furthermore, the study explored gender variation and inter- and intra-speaker variation in more detail and gives possible explanations for the decrease in discrimination tasks in performance for ASFs. The investigation discovered idiosyncratic information in almost all measured acoustic features. For most ASFs, significant gender variation was detected across eight Russian fricatives. Concerning individual differences, the observations suggest that speaker information encoding in fricative sounds is highly complex and can concern diverse acoustic cues. The inter- and intra- speaker variation show for most analysis place and manner of articulation trends. Finally, this investigation demonstrates that acoustic and phonetics studies can be advanced by machine learning (and, more generally, Data Science) approaches: they can help to identify the voiced and unvoiced parts of a fricative and extract the frication noise. They can also be

useful for finding patterns in the acoustic correlates extracted from speech sounds.

6.8 Limitations of the project and an outlook on further research

The current dissertation has several limitations. Probably the most important is that the database represents a subset of the Russian fricative inventory of read speech.

Concerning the application of machine learning methods, in the present study, only basic deep learning methods (feed-forward neural network) without additional tuning were used. Fricative place of articulation, gender and speaker were predicted on pre-extracted and selected measures. Machine learning techniques give also the possibility for pattern and interaction recognition according to an acoustic or visual representation of speech sounds without predefined extracted acoustic correlates. Several studies have shown that for instance speaker discrimination tasks can also be performed on spectrograms, and machine learning techniques can recognize patterns among different sound categories and speakers.

Furthermore, the identification of a set of acoustic cues to predict the place of articulation in fricatives was performed only on three voiceless fricatives. Additional research is needed to identify crucial and distinct cues for the place of articulation contrast in voiced fricatives, voicing contrast of the same place of articulation, and palatalisation contrast.

Moreover, it is worth testing how language-specific are the findings and whether the application of similar acoustic cue extraction techniques, as well as identification methods for the analysis of distinct and contrastive cues, would lead to comparable results in languages other than Russian.

Concerning idiosyncratic information in fricatives, the analysis shows that most of the measured acoustic features differ between female and male speakers. As it was outlined, it is known that females produce higher spectral energy and that these properties help listeners to distinguish speakers' gender. The analysis of gender variation demonstrates that further research is needed to determine the extent to which other acoustic cues such as the harmonics-to-noise ratio serve as crucial perceptual cues. This information could be accessed through perceptual experiments and the manipulation of, for example, harmonic and noise proportions. It is therefore suggested in future studies to extend gender-variation research to other fricatives. Nevertheless, to test the importance of individual ASFs in gender prediction, it may be necessary to compare separately, for example, the performance on spectral cues and HNR measures. Furthermore, future analyses should probably include the comparison of voiced and voiceless fricatives

A further limitation of the study is that the acoustics of only three speakers could be analysed in more detail. Considering the findings of how individual speakers differ in their variant and constant acoustic characteristics, the results argue for a qualitative approach and supplementary investigations of individual speakers. Furthermore, it remains unclear whether other speakers employ comparable strategies and that exhibit similar patterns as the three introduced speakers. Preliminary analyses of the data suggest that if grouping speakers, the production of one sound should be considered. For most speakers, it seems that the detected patterns of idiosyncratic information encoding in one sound are not directly transferable to other sounds. However, this aspect needs also more attention in further phonetic-acoustic research.

Appendix

Table 6.2: Summary of the acoustic cues included in the present study. Some cues were used in both studies, and some others only in one. The column *Paper* refers to in which study the cue was used.

Cue	Variable	Description	Paper
Fricative Duration	*dur*	Duration of the entire sound obtained from manual segmentation	2,3
Zero Crossing Rate	*zcr*	Number of times the wave crosses 0, computed for each time frame of the signal	2
Peak Frequency	*peak*	Frequency of the highest amplitude	2,3
Peak Amplitude	*peak_a*	Amplitude of the highest frequency	2
Spectral Mean	*cog*	Mean distribution of spectral energy (center of gravity)	2, 3
Spectral Variance	*sdev*	Spectral spread or variance of the energy around the mean	2,3
Spectral Skewness	*skew*	Spectral tilt, overall asymmetry of the energy distribution	2,3
Spectral Kurtosis	*kurt*	Spectral flatness of the distribution	2,3
HNR mean	*hmean*	The mean of Harmonics to Noise Ration (HNR)	3
HNR sd	*hsd*	Standard deviation of HNR	3
HNR max	*hmax*	Maximum of HNR	3
HNR tmax	*htmax*	Time to the maximum HNR	3
tilt	*tilt*	Spectral tilt. Computed by H1-H2	3

Bibliography

Ajili, M., Bonastre, J.-F., Ben Kheder, W., Rossato, S., and Kahn, J. (2017). Phonological content impact on wrongful convictions in Forensic Voice Comparison context. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2147–2151, New Orleans, LA. IEEE.

Al-Tamimi, J. and Khattab, G. (2015). Acoustic cue weighting in the singleton vs geminate contrast in Lebanese Arabic: The case of fricative consonants. *The Journal of the Acoustical Society of America*, 138(1):344–360.

Algabri, M., Mathkour, H., Bencherif, M. A., Alsulaiman, M., and Mekhtiche, M. A. (2020). Towards Deep Object Detection Techniques for Phoneme Recognition. *IEEE Access*, 8:54663–54680.

Alsulaiman, M., Mahmood, A., and Muhammad, G. (2017). Speaker recognition based on Arabic phonemes. *Speech Communication*, 86:42–51.

Anjos, I., Eskenazi, M., Marques, N., Grilo, M., Guimarães, I., Magalhães, J., and Cavaco, S. (2020). Detection of Voicing and Place of Articulation of Fricatives with Deep Learning in a Virtual Speech and Language Therapy Tutor. In *Interspeech 2020*, pages 3156–3160. ISCA.

Antal, M. (2008). Phonetic speaker recognition. *Proc. of the 7th International Conference COMMUNICATIONS*, pages 67–72.

Barry, S. M. E. (1995). Variation in vocal fold vibration during voiced obstruents in Russian. *International Journal of Language & Communication Disorders*, 30(2):124–131.

Behrens, S. J. and Blumstein, S. E. (1988). Acoustic characteristics of English voiceless fricatives: a descriptive analysis. *Journal of Phonetics*, 16(3):295–298. Number: 3.

Bianco, M. J., Gerstoft, P., Traer, J., Ozanich, E., Roch, M. A., Gannot, S., and Deledalle, C.-A. (2019). Machine learning in acoustics: Theory and applications. *The Journal of the Acoustical Society of America*, 146(5):3590–3628.

Blumstein, S. E. and Stevens, K. N. (1981). Phonetic features and acoustic invariance in speech. *Cognition*, 10(1-3):25–32.

Bolla, K. (1981). *A conspectus of Russian Speech Sounds.* Köln : Böhlau Verlag, 32 edition.

Bonastre, J.-F., Kahn, J., Rossato, S., and Ajili, M. (2015). Forensic speaker recognition: Mirages and reality. *S. Fuchs/D*, 255.

Butcher, A. (2003). Australian Aboriginal Languages Consonant-Salient Phonologies and the "Place-of-Articulation Imperative". *Australian Speech Science and Technology Association*.

Catford, J. C. (1988). *A Practical Introduction to Phonetics*. Oxford University Press.

Catford, J. C. (1997). *Fundamental Problems in Phonetics.* Indiana University Press.

Chorowski, J. K., Bahdanau, D., Serdyuk, D., Cho, K., and Bengio, Y. (2015). Attention-Based Models for Speech Recognition. *Advances in neural information processing systems*, 28.

de Manrique, A. M. B. and Massone, M. I. (1981). Acoustic analysis and perception of Spanish fricative consonants. *The Journal of the Acoustical Society of America*, 69(4):1145–1153. Number: 4.

Dellwo, V., Huckvale, M., and Ashby, M. (2007). How Is Individuality Expressed in Voice? An Introduction to Speech Production and Description for Speaker Classification. *Speaker Classification I*, 4343:1–20. ISSN: 0302-9743, 1611-3349 Series Title: Lecture Notes in Computer Science.

Derkach, M., Fant, G., and de Serpa-Leitao, A. (1970). Phoneme coarticulation in Russian hard and soft VCV-utterances with voiceless fricatives. *STL-QPSR 11.2-3*, 11(2-3):1–7.

DiCanio, C. T., Nam, H., Whalen, D. H., Bunnell, H. T., Amith, J. D., and Castillo Garcia, R. (2012). Assessing agreement level between forced alignment models with data from endangered language documentation corpora. In *Interspeech 2012*, pages 130–133. ISCA.

Diehl, R. L., Lindblom, B., Hoemeke, K. A., and Fahey, R. P. (1996). On explaining certain male-female differences in the phonetic realization of vowel categories. *Journal of Phonetics*, 24(2):187–208.

Enzinger, E. and Balazs, P. (2011). Speaker Verification using Pole/Zero Estimates of Nasals. *Analele Universitatii "Eftimie"*, 18:33–44.

Fant, G. (1960). *Acoustic Theory of Speech Production.* Walter de Gruyter, 2 edition.

Flipsen, P., Shriberg, L., Weismer, G., Karlsson, H., and McSweeny, J. (1999). Acoustic Characteristics of /s/ in Adolescents. *Journal of Speech, Language, and Hearing Research*, 42(3):663–677.

Forrest, K., Weismer, G., Milenkovic, P., and Dougall, R. N. (1988). Statistical analysis of word-initial voiceless obstruents: Preliminary data. *The Journal of the Acoustical Society of America*, 84(1):115–123. Number: 1.

Fromont, R. and Hay, J. (2012). LaBB-CAT: an Annotation Store.

Funatsu, S. and Kiritani, S. (1998). Perceptual Properties of Russians with Japanese Fricatives. *Fifth International Conference on Spoken Language Processing*.

Gendrot, C., Ferragne, E., and Pellegrini, T. (2019). Deep learning and voice comparison: phonetically-motivated vs. automatically-learned features. *ICPhS*.

Gendrot, C., Ferragne, E., and Pellegrini, T. (2020). Informations segmentales pour la caractérisation phonétique du locuteur: variabilité inter-et intra-locuteurs. *6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 1: Journées d'Études sur la Parole*, 1.

Ghaffarvand Mokari, P. and Mahdinezhad Sardhaei, N. (2020). Predictive power of cepstral coefficients and spectral moments in the classification of Azerbaijani fricatives. *The Journal of the Acoustical Society of America*, 147(3):EL228–EL234.

Gonzalez, S., Grama, J., and Travis, C. E. (2020). Comparing the performance of forced aligners used in sociophonetic research. *Linguistics Vanguard*, 6(1):20190058.

Gordon, M., Barthmaier, P., and Sands, K. (2002). A cross-linguistic acoustic study of voiceless fricatives. *Journal of the International Phonetic Association*, 32(2):141–174.

Grandon, B. and Vilain, A. (2020). Development of fricative production in French-speaking school-aged children using cochlear implants and children with normal hearing. *Journal of Communication Disorders*, 86:105996.

Halle, M. (2011). *Sound Pattern of Russian. A Linguistic and Acoustical Investigation.* De Gruyter Mouton.

Harper, S. K. (2021). *Individual differences in phonetic variability and phonological representation.* PhD thesis, Diss. University of Southern California.

Hayward, K. (2000). *Experimental phonetics*. Longman linguistics library. Longman, Harlow, 2nd ed edition.

He, L. and Dellwo, V. (2014). Speaker idiosyncratic variability of intensity across syllables. In *Interspeech 2014*, pages 233–237. ISCA.

Heinz, J. M. and Stevens, K. N. (1961). On the Properties of Voiceless Fricative Consonants. *The Journal of the Acoustical Society of America*, 33(5):589–596. Number: 5.

Hoelterhoff, J. and Reetz, H. (2007). Acoustic cues discriminating German obstruents in place and manner of articulation. *The Journal of the Acoustical Society of America*, 121(2):1142–1156. Number: 2.

Hughes, G. W. and Halle, M. (1956). Spectral Properties of Fricative Consonants. *The Journal of the Acoustical Society of America*, 28(2):303–310.

Jassem, W. (1965). The Formants of Fricative Consonants. *Language and Speech*, 8(1):1–16. Number: 1.

Jassem, W. (1995). The Acoustic Parameters of Polish Voiceless Fricatives: An Analysis of Variance. *Phonetica*, 52(3):251–258. Number: 3.

Jax, P. and Vary, P. (2003). On artificial bandwidth extension of telephone speech. *Signal Processing*, 83(8):1707–1719.

Jesus, L. M. and Shadle, C. H. (2002). A parametric study of the spectral characteristics of European Portuguese fricatives. *Journal of Phonetics*, 30(3):437–464. Number: 3.

Jesus, L. M. T. and Jackson, P. J. B. (2008). Frication and Voicing Classification. *Computational Processing of the Portuguese Language*, 5190:11–20.

Jones, D. and Ward, D. (1969). *The phonetics of Russian*. Cambridge University Press.

Jongman, A., Wayland, R., and Wong, S. (2000). Acoustic characteristics of English fricatives. *The Journal of the Acoustical Society of America*, 108(3):1252. Number: 3.

Kavanagh, C. (2011). Intra- and inter-speaker variability in acoustic properties of English /s/. *International Association for Forensic Phonetics and Acoustics*.

Kavanagh, C. M. (2012). *New consonantal acoustic parameters for forensic speaker comparison*. PhD thesis, University of York.

Kedrova, G. Y., Anisimov, N. V., Zaharov, L. M., and Pirogov, Y. A. (2008). Magnetic Resonance investigation of palatalized stop consonants and spirants in Russian. *The Journal of the Acoustical Society of America*, 123(5):3325–3325.

Kissine, M., Van de Velde, H., and van Hout, R. (2003). An acoustic study of standard Dutch /v/, /f/, /z/ and /s/. *Linguistics in the Netherlands*, 20:93–104.

Kochetov, A. (2017). Acoustics of Russian voiceless sibilant fricatives. *Journal of the International Phonetic Association*, 47(3):321–348.

Kong, Y.-Y., Mullangi, A., and Kokkinakis, K. (2014). Classification of Fricative Consonants for Speech Enhancement in Hearing Devices. *PLoS ONE*, 9(4):e95001.

Ladefoged, P. and Maddieson, I. (1996). *The Sounds of the World's Languages*. Blackwell, Cambridge.

Ladefoged, P. and Wu, Z. (1984). Places of articulation: an investigation of Pekingese fricatives and affricates. *Journal of Phonetics*, 12(3):267–278.

Lilley, J., Spinu, L., and Athanasopoulou, A. (2021). Exploring the front fricative contrast in Greek: A study of acoustic variability based on cepstral coefficients. *Journal of the International Phonetic Association*, 51(3):393–424.

Litvin, N. (2014). *An Ultrasound Investigation of Secondary Velarization in Russian*. PhD thesis, University of Victoria, University of Victoria.

Ludger, P., Fuchs, S., and Seifert, F. (2021). Differences between male and female speakers in the production of /s/: A cross-linguistic study. *17. Phonetik und Phonologie im deutschsprachigen Raum (P&P)*.

MacKenzie, L. and Turton, D. (2020). Assessing the accuracy of existing forced alignment software on varieties of British English. *Linguistics Vanguard*, 6(s1):20180061.

Maddieson, I. and Disner, S. F. (1984). *Patterns of sounds*. Cambridge University Press, Cambridge [Cambridgeshire] ; New York.

Maddieson, I., Flavier, S., Marsico, E., Coupé, C., and Pellegrino, F. (2013). LAPSyd: lyon-albuquerque phonological systems database. In *Interspeech 2013*, pages 3022–3026. ISCA.

Maniwa, K., Jongman, A., and Wade, T. (2009). Acoustic characteristics of clearly spoken English fricatives. *The Journal of the Acoustical Society of America*, 125(6):3962–3973. Number: 6.

Mann, V. A. and Repp, B. H. (1980). Influence of vocalic context on perception of the [J]-[s] distinction. *Perception & Psychophysics*, 28(3):213–228.

Matovski, D. S., Nixon, M. S., Mahmoodi, S., and Carter, J. N. (2010). The effect of time on the performance of gait biometrics. *2010 Fourth IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, pages 1–6.

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., and Sonderegger, M. (2017). Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. In *Interspeech 2017*, pages 498–502. ISCA.

McDougall, K. and Nolan, F. (2007). Discrimination of speaker using the formant dynamics of /u/ in British English. *Proceedings of the International Congress of Phonetic Sciences*, pages 1825–1828.

McFarland, D. H., Baum, S. R., and Chabot, C. (1996). Speech compensation to structural modifications of the oral cavity. *The Journal of the Acoustical Society of America*, 100(2):1093–1104.

McMurray, B. and Jongman, A. (2011). What information is necessary for speech categorization? Harnessing variability in the speech signal by integrating cues computed relative to expectations. *Psychological Review*, 118(2):219–246. Number: 2.

Michalopoulou, Z.-H., Gerstoft, P., Kostek, B., and Roch, M. A. (2021). Introduction to the special issue on machine learning in acoustics. *The Journal of the Acoustical Society of America*, 150(4):3204–3210.

Monson, B. B., Hunter, E. J., Lotto, A. J., and Story, B. H. (2014). The perceptual significance of high-frequency energy in the human voice. *Frontiers in Psychology*, 5.

Moore, B. C. J. (2003). Coding of Sounds in the Auditory System and Its Relevance to Signal Processing and Coding in Cochlear Implants:. *Otology & Neurotology*, 24(2):243–254.

Munson, B., McDonald, E. C., DeBoe, N. L., and White, A. R. (2006). The acoustic and perceptual bases of judgments of women and men's sexual orientation from read speech. *Journal of Phonetics*, 34(2):202–240.

Nagamine, T., Seltzer, M. L., and Mesgarani, N. (2015). Exploring How Deep Neural Networks Form Phonemic Categories. *Sixteenth Annual Conference of the International Speech Communication Association*.

Narayanan, S. S., Alwan, A. A., and Haker, K. (1995). An articulatory study of fricative consonants using magnetic resonance imaging. *The Journal of the Acoustical Society of America*, 98(3):1325–1347.

Newell, K. M. and Hancock, P. (1984). Forgotten moments: a note on skewness and kurtosis as influential factors in inferences extrapolated from response distributions. *Journal of motor behavior*, 16(3):320–335.

Newman, R. S., Clouse, S. A., and Burnham, J. L. (2001). The perceptual consequences of within-talker variability in fricative production. *The Journal of the Acoustical Society of America*, 109(3):1181–1196.

Nirgianaki, E. (2014). Acoustic characteristics of Greek fricatives. *The Journal of the Acoustical Society of America*, 135(5):2964–2976. Number: 5.

Nittrouer, S., Studdert-Kennedy, M., and McGOWAN, R. S. (1989). The emergence of phonetic segments: Evidence from the spectral structure of fricative-vowel syllables spoken by children and adults. *Journal of Speech, Language, and Hearing Research*, 32(1):120–132.

Padgett, J. and Zygis, M. (2007). The Evolution of Sibilants in Polish and Russian. *Journal of Slavic linguistics*, pages 291–324.

Peeters, G. (2003). A large set of audio features for sound description. *Icram, Paris, France*.

Pincas, J. and Jackson, P. J. B. (2006). Amplitude modulation of turbulence noise by voicing in fricatives. *The Journal of the Acoustical Society of America*, 120(6):3966–3977.

Rabha, S., Sarmah, P., and Prasanna, S. R. M. (2019). Aspiration in fricative and nasal consonants: Properties and detection. *The Journal of the Acoustical Society of America*, 146(1):614–625.

Reidy, P. F. (2016). Spectral dynamics of sibilant fricatives are contrastive and language specific. *The Journal of the Acoustical Society of America*, 140(4):2518–2529.

Romeo, R., Hazan, V., and Pettinato, M. (2013). Developmental and gender-related trends of intra-talker variability in consonant production. *The Journal of the Acoustical Society of America*, 134(5):3781–3792.

Rose, P. (2007). Forensic speaker discrimination with Australian English vowel acoustics. *ICPhS XVI*, 6(10).

Rosenfelder, I., Fruehwald, J., Evanini, Keelan, Seyfarth, Scott, Gorman, Kyle, Prichard, Hilary, and Yuan, Jiahong (2014). FAVE (forced alignment and vowel extraction) suite version 1.1. 3.

Schiel, F. (1999). Automatic Phonetic Transcription of Non-Prompted Speech.

Schiel, F., Draxler, C., Baumann, A., Ellbogen, T., and Steffen, A. (2012). The Production of Speech Corpora.

Schindler, C. and Draxler, C. (2013). Using spectral moments as a speaker specific feature in nasals and fricatives. In *Interspeech 2013*, pages 2793–2796. ISCA.

Schwartz, M. F. (1968). Identification of Speaker Sex from Isolated, Voiceless Fricatives. *The Journal of the Acoustical Society of America*, 43(5):1178–1179.

Shadle, C. H. (1986). The acoustics of fricative consonants. *The Journal of the Acoustical Society of America*, 79(2):574–574. Number: 2.

Shadle, C. H. (1990). Articulatory-acoustic relationships in fricative consonants. In *Speech production and speech modelling*, pages 187–209. Springer, Dordrecht.

Shadle, C. H. and Mair, S. (1996). Quantifying spectral characteristics of fricatives. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96*, volume 3, pages 1521–1524, Philadelphia, PA, USA. IEEE.

Shupljakov, V., Fant, G., and de Serpa-Leitao, A. (1968). Acoustical features of hard and soft Russian consonants in connected speech: A spectrographic study. *STL-QPSR*, 9(4):1–6.

Silbert, N. and de Jong, K. (2008). Focus, prosodic context, and phonological feature specification: Patterns of variation in fricative production. *The Journal of the Acoustical Society of America*, 123(5):2769–2779.

Smorenburg, L. and 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):949–960. Publisher: Acoustical Society of America.

Soli, S. D. (1981). Second formants in fricatives: Acoustic consequences of fricative-vowel coarticulation. *The Journal of the Acoustical Society of America*, 70(4):976–984.

Spinu, L., Kochetov, A., and Lilley, J. (2018). Acoustic classification of Russian plain and palatalized sibilant fricatives: Spectral vs. cepstral measures. *Speech Communication*, 100:41–45.

Spinu, L. and Lilley, J. (2016). A comparison of cepstral coefficients and spectral moments in the classification of Romanian fricatives. *Journal of Phonetics*, 57:40–58.

Spinu, L., Vogel, I., and Timothy Bunnell, H. (2012). Palatalization in Romanian—Acoustic properties and perception. *Journal of Phonetics*, 40(1):54–66.

Stevens, K. N. (1998). *Acoustic Phonetics*. MIT Press, 30 edition.

Stevens, K. N., Blumstein, S. E., Glicksman, L., Burton, M., and Kurowski, K. (1992). Acoustic and perceptual characteristics of voicing in fricatives and fricative clusters. *The Journal of the Acoustical Society of America*, 91(5):2979–3000. Number: 5.

Strevens, P. (1960). Spectra of Fricative Noise in Human Speech. *Language and Speech*, 3(1):32–49. Number: 1.

Stuart-Smith, J. (2007). Empirical evidence for gendered speech production: /s/ in Glaswegian. *Mouton de Gruyter*.

Timberlake, A. (2004). *A Reference Grammar of Russian*. Cambridge University Press.

Weirich, M. and Simpson, A. P. (2014). Differences in acoustic vowel space and the perception of speech tempo. *Journal of Phonetics*, 43:1–10.

Weirich, M. and Simpson, A. P. (2015). Gender-specific differences in sibilant contrast realizations in English and German. *ICPhS*.

Zsiga, E. C. (2000). Phonetic alignment constraints: consonant overlap and palatalization in English and Russian. *Journal of Phonetics*, 28(1):69–102. Number: 1.

Żygis, M. and Padgett, J. (2010). A perceptual study of Polish fricatives, and its implications for historical sound change. *Journal of Phonetics*, 38(2):207–226. Number: 2.