Identifying the Russian voiceless non-palatalized fricatives /f/, /s/, and / / from acoustic cues using machine learning

Natalja Ulrich, Marc Allassonnière-Tang, François Pellegrino, Dan Dediu

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

Natalja Ulrich, Marc Allassonnière-Tang, François Pellegrino, et al.

Citation: The Journal of the Acoustical Society of America **150**, 1806 (2021); doi: 10.1121/10.0005950
View online: https://doi.org/10.1121/10.0005950
View Table of Contents: https://asa.scitation.org/toc/jas/150/3
Published by the Acoustical Society of America

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 [f^j, v^j, s^j, ɕ:, z^j] 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

https://doi.org/10.1121/10.0005950

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 [sal^{j}]" and "I like your [ʃal^{j}]" (scarf),[1] while the contrasting ones were more constrained: for example, for the minimal pair [sal^{j}] and [ʃal^{j}], the carrier sentences were "She said [sal^{j}] and not [ʃal^{j}]" and "She said [ʃal^{j}] and not [sal^{j}]."[2] Some target words have two minimal pairs (for instance, the word [sal^{j}] is embedded in two different carrier sentences, once contrasting with [ʃal^{j}] and a second time with [ʒal^{j}]), 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

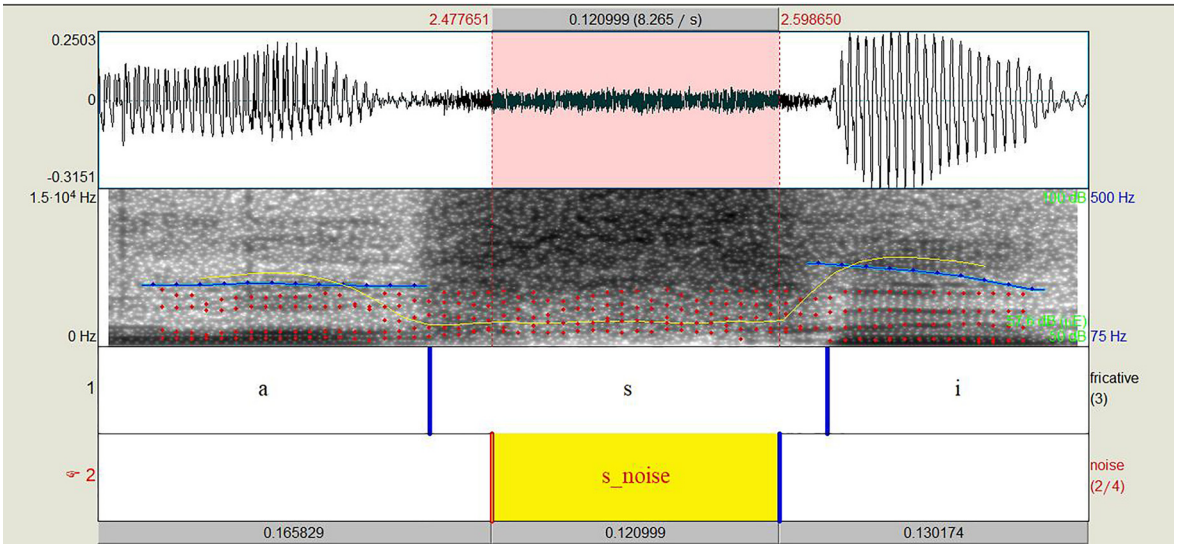

FIG. 1. (Color online) An example of a fricative sound. The first tier of this screenshot from Praat shows the full duration of the fricative, while the second shows only the noise part, excluding the effects of any potential co-articulation.

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

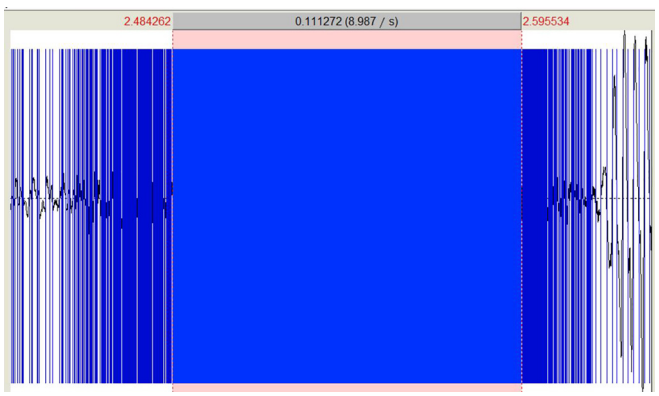

FIG. 2. (Color online) Visualizing the zero crossings in Praat. The increase in the spatial density of the blue bars shows a rapid increase in *zcr*.

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.

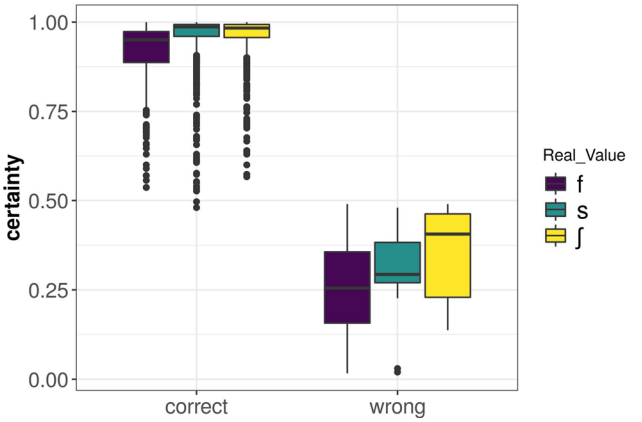

FIG. 3. (Color online) A decision tree generated for window length = 512 points and 50% overlap. *zcr*, zero crossing rate; *diff*, the gap of zero crossing rate between two consecutive windows. A positive *diff* value represents an increase in the zero crossing rate, while a negative value refers to 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.

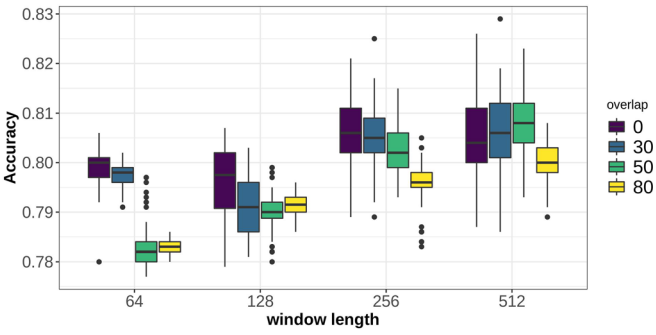

FIG. 4. (Color online) The accuracy of the classifiers trained with different parameters of window length and percentage of overlap (add percentage in graph). Each combination of parameters is trained and tested for 100 replications with different training and testing data.

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

https://doi.org/10.1121/10.0005950

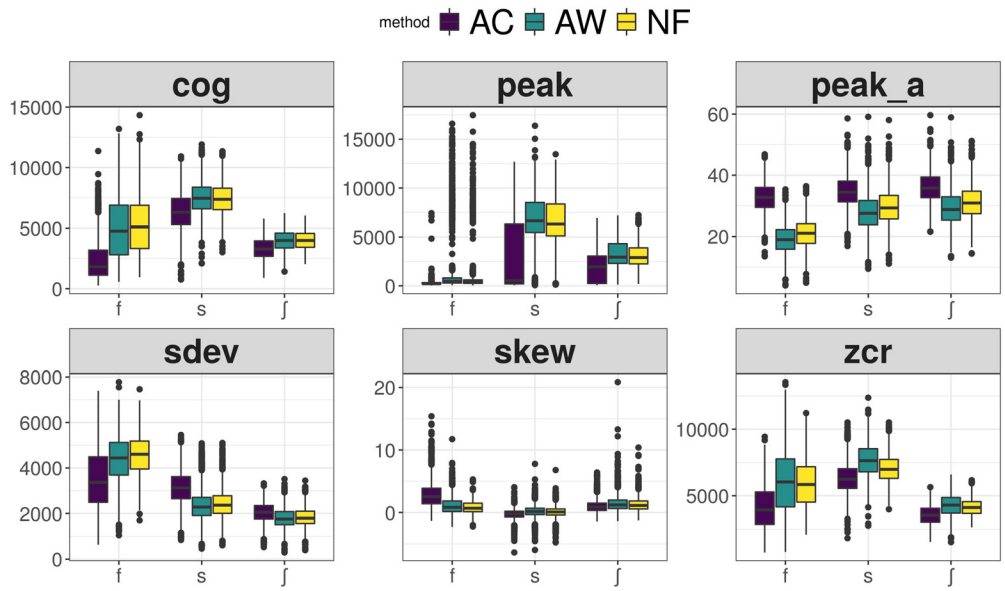

FIG. 5. (Color online) The comparison of acoustic cues based on the three main ACETs reported in the experiments. The names of the ACETs refer to the acoustic cue extraction techniques listed in Table I.

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).

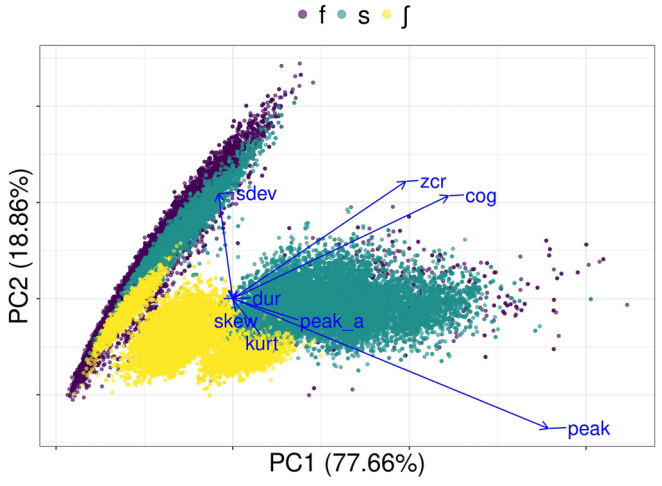

FIG. 6. (Color online) The PCA visualization of the acoustic cues for each sound. The length of the arrows relates to how much information is contributed by the acoustic cues to the PCs. *cog*, *peak*, *sdev*, and *zcr* are the most relevant.

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 nonlinear 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 (≥5486, node 1 to node 2) and *sdev* is also high (≥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

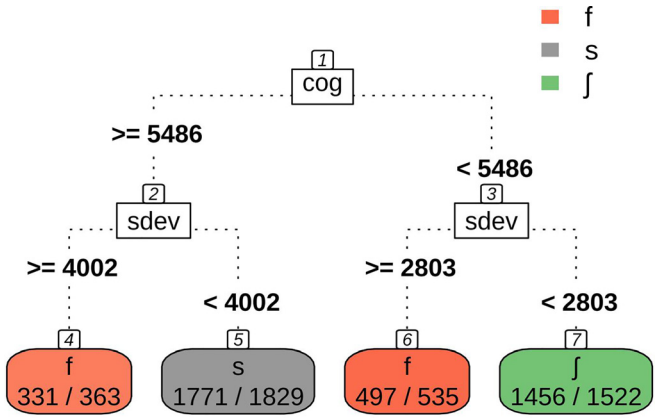

FIG. 7. (Color online) A decision tree generated with the acoustic cues from the ACET of NF. The rules for its interpretation are similar to those of the tree in Fig. 3 except for the color of the "buckets," which now represent the sound.

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

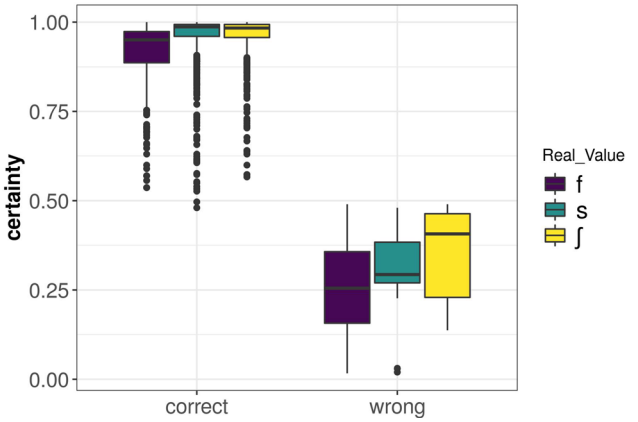

FIG. 8. (Color online) The confidence of the random forest classifier for correct and wrong decisions across [f], [s], and [ʃ] for NF.

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

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.