A comparison of cepstral coefficients and spectral moments in the classification of Romanian fricatives

Laura Spinu¹, Jason Lilley¹

^a *University of Western Ontario
London, Ontario, Canada, N6A 3K7*

^b *Nemours Biomedical Research
Wilmington, Delaware 19803*

Abstract

In this paper we explore two methods for the classification of fricatives. First, for the coding of the speech, we compared two sets of acoustic measures obtained from a corpus of Romanian fricatives: (a) spectral moments, and (b) cepstral coefficients. Second, we compared two methods of determining the regions of the segments from which the measures would be extracted. In the first method, the phonetic segments were divided into three regions of approximately equal duration. In the second method, Hidden Markov Models (HMMs) were used to divide each segment into three regions such that the variances of the measures within each region were minimized. The corpus we analyzed consists of 3,674 plain and palatalized word-final fricatives from four places of articulation, produced by 31 native speakers of Romanian (20 females). We used logistic regression to classify fricatives by place, voicing, palatalization status, and gender. We found that cepstral coefficients reliably outperformed spectral moments in all classification tasks, and that using regions determined by HMM yielded slightly higher correct classification rates than using regions of equal duration.

Keywords: fricatives, cepstral coefficients, spectral moments, place of articulation, secondary palatalization, classification, Romanian

*Corresponding author

1. Introduction

The description and classification of fricatives based on spectral measurements has been a popular topic in the phonetics literature in recent decades (Shadle and Mair 1996, Tabain and Watson 1996, Jesus and Shadle 2002, Hoelterhoff and Reetz 2007, Jesus and Jackson 2008, McMurray and Jongman 2011, Koenig et al. 2013, among others), as well as in the signal processing and automatic speech recognition literature (Halberstadt and Glass 1997, Childers et al. 1998, Farooq and Datta 2001) and in the clinical phonetics literature (Todd et al. 2011, Kong et al. 2014). Part of the increased focus on this topic has to do with the relative difficulty in finding reliable, invariant measures distinguishing fricatives from one another, especially as far as place of articulation is concerned (Forrest et al. 1988, Tomiak 1990, Jongman et al. 2000, McMurray and Jongman 2011). Some of the acoustic parameters previously reported for fricatives include spectral peak location, spectral slope, spectral moments, noise duration, F2 onset frequency, static and dynamic amplitude measurements, and locus equations. These measures were typically based on discrete Fourier transforms. More recently, multitaper spectra were recommended as better suited for stochastic parts of speech, as they reduce the bias of spectral estimates when calculated over short intervals (Lousada et al. 2012, Zygis et al. 2012, Koenig et al. 2013). Spectral moment analysis has been used successfully for classifying stops (Forrest et al. 1988, 1990, 1994) and, while not as successful with fricatives in certain studies (Shadle and Mair 1996, Zygis et al. 2012), in others it provided reliable measures associated with categories such as fricative place of articulation, and proved to be more invariant for speaker and vowel context than other measures (Flipsen et al. 1999, Jongman et al. 2000, McMurray and Jongman 2011). McMurray and Jongman (2011) employed a wide combination of measures with the goal of modeling fricative perception and representation. They concluded that there is no invariance in the acoustic signal, and as a result speech categorization by listeners requires massive cue-integration as well as compensatory mechanisms able to handle contextual influences. While their

31 results are robust and were reached based on a large-scale study using sophisti-
32 cated statistical tools to compare the acoustic properties of fricatives to human
33 categorization, it remains to be seen (1) whether there might be other kinds of
34 acoustic features that are better suited for the classification of fricatives, and
35 (2) whether fricatives in other languages might behave differently.

36 In the current study, we use cepstral coefficients (Deller et al. 1993) and
37 directly compare their performance to that of spectral moments. Other differ-
38 ences include dividing the fricatives into regions based on internal variance using
39 a Hidden Markov Model approach (also used for different purposes in Jesus and
40 Jackson 2008) in order to determine whether this may present advantages over
41 the more traditional method of extracting measurements from predefined loca-
42 tions or evenly spaced intervals inside segments (Jongman et al. 2000, Nissen
43 and Fox 2005, McMurray and Jongman 2011, Koenig et al. 2013).

44 We also expand on the research on the acoustic properties of fricatives by
45 examining sounds in a different language, Romanian, some of which are artic-
46 ulated at different places of articulation than those of English. An aspect of
47 Romanian fricatives not yet explored acoustically is the behavior of the voicing
48 contrast. A second under-studied aspect is represented by secondary palatal-
49 ization, to date only addressed in one study (Spinu 2010, Spinu et al. 2012).

50 While a direct comparison with previous studies on fricatives is not pos-
51 sible, given differences in the fricative inventory, the experimental set-up, the
52 scope of the work and the measures utilized, our study adds to the body of
53 research on the relationships between acoustic properties and categories such as
54 place of articulation, voicing, gender, and palatalization by providing a detailed
55 analysis of these relationships for Romanian fricatives.

56 2. Background

57 The literature on the acoustic properties of fricatives is extensive and in-
58 cludes work in acoustic phonetics, signal processing, automatic speech recog-
59 nition and clinical phonetics. Although most early work focused exclusively

60 on English (Gordon et al. 2002), more recent studies have examined fricative
61 acoustics in other languages, including German (Hamann and Sennema 2005,
62 Kuzla et al. 2007), Polish (Nowak 2006, Nawrocki 2008), Portuguese, Italian
63 and German (Jesus and Shadle 2002, 2005; Jesus and Jackson 2008; Pape and
64 Jesus 2014, 2015), and Xitsonga (Lee-Kim et al. 2014). Despite abundant
65 interest in the acoustic properties of fricatives, studies focusing explicitly on
66 classification methods for dimensions such as place of articulation or voicing
67 are more scarce. In what follows, we provide a brief historical perspective by
68 reviewing six studies (see Table 1). Our review is far from comprehensive, but
69 the reasons we selected these particular studies include:

70 **Focus on classification.** These studies are designed to identify acoustic pa-
71 rameters that reliably differentiate or discriminate productions of fricative
72 consonants in terms of place of articulation and voicing, two categories es-
73 pecially relevant for speech perception and recognition.

74 **Prominence in the literature.** Except for the more recent studies, our selec-
75 tions are well-known in the fricative literature. A cursory internet search
76 reveals that each has almost 400 citations or more.

77 **Acoustic parameter sets.** We selected these studies because they all used
78 spectral moments or cepstral coefficients, either on their own or in combi-
79 nation with other measures, and our study makes a direct comparison of
80 these two types of measures.

81 **Diversity in scope.** Three of these studies are interested in acoustic proper-
82 ties only (Forrest et al. 1988, Tomiak 1990, Jongman et al. 2000), while
83 Nissen and Fox (2005) focus on developmental aspects. McMurray and
84 Jongman (2011) are concerned with human perception, and Kong et al.’s
85 (2014) work is in clinical phonetics and engineering. While these perspec-
86 tives shape the selection of acoustic parameter sets in different ways, they
87 converge for the most part in the results obtained.

88 Table 1 summarizes the studies selected. All of them classified corpora of En-
89 glish fricatives at either three or four of the following places of articulation:
90 labiodental, dental, alveolar, postalveolar, glottal. Some included both voiced
91 and voiceless segments, while others only used voiceless ones.

92 [INSERT TABLE 1 HERE]

93 Forrest et al. (1988) employed two sets of spectral moments, linear and
94 Bark-scaled. The data came from 10 speakers (5 females). The fricatives in-
95 cluded were all voiceless: [f, θ, s, ʃ]. The moments were extracted over several
96 overlapping 20 ms analysis windows spaced at 10 ms intervals. The most im-
97 portant contributor to the classification of fricatives was spectral skewness, and
98 slightly higher overall correct classification¹ was achieved with Bark moments
99 (77.7%) than with linear moments (74.5%). Most errors involved the distinction
100 between labial [f] and interdental [θ].

101 Tomiak (1990) also employed linear and Bark-scaled spectral moments.
102 The corpus included recordings from 6 speakers (3 females), and the same voice-
103 less fricatives plus /h/ were examined. The values were extracted from consecu-
104 tive overlapping 15 ms analysis windows spaced 10 ms apart. Unlike in Forrest
105 et al.’s (1988) study, slightly higher correct classification was reported for linear
106 vs. Bark moments (78% vs. 74%), but this difference was not significant.

107 Jongman et al. (2000) used a corpus of recordings from 20 speakers (10
108 females). Eight fricatives were included: [f, v, θ, ð, s, z, ʃ, ʒ]. Six types of
109 acoustic parameters were extracted: spectral moments at four locations, spectral
110 peak location, normalized and relative amplitude, duration, and F2 onset. The
111 spectral moments were obtained from single power FFT spectra derived from
112 40-ms Hamming windows at 4 different locations (onset, mid-point, and end
113 of frication, as well as centered over the fricative-vowel transition). A stepwise
114 linear discriminant analysis with 21 predictor variables yielded an overall correct

¹“Correct classification” here and henceforth refers to the percent of tokens correctly clas-
sified out of the total tokens for each category.

115 classification by place of 77%. Classification of all four places of articulation was
 116 significantly above chance (88% for the sibilants and 66% for the non-sibilants).
 117 As was the case in the earlier studies, most errors involved [f] and [θ]. The
 118 most useful variables in classification were spectral peak location, normalized
 119 and relative amplitude, and the first spectral moment (mean) at the onset and
 120 midpoint of frication. Despite the complexity of the model (acoustic parameters
 121 and locations), the accuracy rate remained unchanged from that yielded by
 122 earlier studies based on spectral moment measurements alone.

123 Nissen and Fox (2005) examined the acoustic structure of voiceless frica-
 124 tives [f, θ, s, ʃ] produced by adults and typically developing 3- to 5-year-old
 125 children in terms of spectral moments, duration, normalized amplitude, and
 126 spectral slope. They used a series of three 40-ms Hamming windows located
 127 at the beginning, middle, and end of each segment. They found that spectral
 128 slope and variance were important acoustic parameters in the differentiation
 129 and classification of the voiceless fricatives, with spectral variance (the second
 130 spectral moment) being the only measure able to separate all four places of ar-
 131 ticulation. A linear discriminant analysis based on a cross-validation jack-knife
 132 procedure yielded 69.2% correct classification for place of articulation overall.
 133 The classification rate for the adult speakers' productions was 65%, while the
 134 rates for the fricatives of the 3-, 4-, and 5-year-old speakers were 70.0%, 70.0%,
 135 and 77.5% respectively.

136 McMurray and Jongman (2011) analyzed data from the same corpus as
 137 Jongman et al. (2000). The 24 predictors for place of articulation used in this
 138 study included the following ten types of acoustic measures: total amplitude (for
 139 frication and vocalic portions separately), RMS amplitude (for the frication and
 140 vowel portions), Narrow Band amplitude (separate amplitudes for frication and
 141 vocalic portions in a specific frequency band near F3 or F5), duration (fricative,
 142 vowel and word duration), pitch, formants 1 through 5, F2 onset frequency
 143 (F2 measured at vowel onset), low frequency energy (mean RMS amplitude
 144 below 500 Hz), spectral peak (frequency of the spectral peak with the highest

145 amplitude), and spectral moments 1-4 (extracted from 40-ms windows at four
 146 locations: onset, middle, and end of frication, as well as centered over the
 147 voicing offset). This combination of predictors yielded 92.9% overall correct
 148 classification. It should be noted that, because the main goal in this study
 149 was to evaluate the performance of several models of phoneme categorization,
 150 the logistic regression analysis yielding these results was one in which the raw
 151 acoustic parameters were processed relative to long term knowledge about how
 152 they behave in different contexts (e.g. different talkers or vowel contexts) and
 153 in that sense, it was not a purely acoustic classification scheme. An additional
 154 categorization model using the raw measures only yielded 79.2%-85.0% correct
 155 classification.

156 The last study we review is Kong et al. (2014). A support vector ma-
 157 chine (SVM) algorithm was used to classify fricatives extracted from a subset
 158 of 168 speakers from the TIMIT database. Spectral analyses used a 128-point
 159 FFT in the frequency range from 1kHz to 8 kHz. Information was extracted
 160 from a randomly chosen 8-ms segment using a Hamming window. Spectral
 161 features including four spectral moments, peak, slope, Mel-frequency cepstral
 162 coefficients (MFCC), Gammatone filter outputs, and magnitudes of fast Fourier
 163 Transform (FFT) spectra were used for the classification. With 13 MFCC coef-
 164 ficients and 14 or 24 Gammatone filter outputs, classification performance was
 165 approximately 85%. It is important to note that while eight English fricatives
 166 were included, the classification was made for only three places of articulation,
 167 as the labial and interdental fricatives were collapsed into one category.

168 To summarize, in the studies described above, the overall correct classi-
 169 fication rate for fricative place of articulation based on raw acoustic measures
 170 ranged from 77% to 85%. The early and mid studies were comparable in their
 171 results, staying below 80% in all cases. The more recent studies showed a 7%
 172 increase, yielding accuracy rates of 85%. The best results were obtained based
 173 on more complex combinations of acoustic parameters, as well as larger corpora
 174 of speakers (which provide more model training data). For all models, the main

challenge was distinguishing the labial and interdental fricatives.

3. Current study

In the current study, we add to previous research in a number of ways. In addition to spectral moments, we also performed a cepstral analysis which captures features of overall spectrum shape. Routinely used in speech recognition (Rabiner and Schafer 2010, Deller et al. 1993), perceptually-weighted cepstral features form the basis for acoustic feature sets, providing useful descriptions for vowels and obstruents. Cepstral coefficients have also been used in clinical phonetics studies, outperforming spectral moments in the classification of stop release bursts (Bunnell et al. 2004). In the acoustic phonetics literature, Mel-frequency cepstral coefficients (MFCCs) were used successfully in a study on the classification of voicing in fricatives in British English and European Portuguese (Jesus and Jackson 2008). Furthermore, Ferragne and Pellegrino (2010) recommended cepstral coefficients as a means to compute distances between vowels in the MFCC space. This method yielded a very good estimate of the acoustic distance between 13 different accents of the British Isles, and led the authors to conclude that “the argument that MFCCs cannot be wrong (while formants can) provides strong support for the use of MFCCs in phonetic studies, if only for practical reasons” (p. 536). Using a set of 13 MFCCs, Kong et al. (2014) obtained 85% correct classification for three places of articulation in fricatives, suggesting that the advantages of cepstral coefficients carry over to segments other than vowels. Our study may further validate this conclusion.

An additional research question we addressed was whether classification rates might benefit from a less arbitrary method of choosing measurement locations than the widely used practice of positioning analysis windows at predefined positions inside a segment (Jongman et al. 2000, Jesus and Shadle 2002, McMurray and Jongman 2011). Through the use of HMMs, we aim to divide each fricative into three regions, such that the variances of the measurements within each region are minimized. In doing so, we attempted to capture basic facts

204 of articulation, as continuants are realized as a succession of transition, steady-
205 state, and transition/release phases (each composed of one or more overlapping
206 gestures), which are not evenly distributed (Browman and Goldstein 1992, Zsiga
207 1997, Gafos 2002, Borroff 2007). Our goal is to determine whether extracting
208 acoustic features from such regions results in more informative data than may
209 be obtained using the more traditional predefined positions.

210 Lastly, we analyze fricatives in Romanian, a language for which phonetic
211 studies remain relatively scarce. No acoustic studies to date have addressed
212 specifically the properties of Romanian fricatives. Our fricative corpus was orig-
213 inally collected by the first author of this study for her dissertation research,
214 which addresses the topic of secondary palatalization more generally (Spinu
215 2010). Due to the focus on secondary palatalization and the characteristics of
216 its distribution in Romanian, there are important differences between our stim-
217 uli and those used in the majority of previous studies, such as the VC syllabic
218 structure, as compared to CV (with exceptions such as Jesus and Shadle 2002,
219 Pape and Jesus 2014, 2015, and Maniwa et al. 2009, who also examined frica-
220 tives in a VCV context) and the inclusion of plain and palatalized fricative pairs,
221 allowing us to compare these two types of segments directly. This is particularly
222 noteworthy because, to our knowledge, no large-scale studies have attempted
223 the classification of plain versus palatalized segments based on their acoustic
224 properties. Because of these differences, our study enables us to consider the
225 general properties of fricatives in novel environments and thus expands upon
226 previous work.

227 Based on the make-up of our corpus and the type of analysis chosen, we
228 can begin to answer questions about the specific properties of Romanian frica-
229 tives, with regard to (a) their voicing behavior, for instance whether Roma-
230 nian behaves more like the majority of Romance languages, such as Italian or
231 Spanish, in exhibiting a strong voicing contrast between phonologically voiced
232 and voiceless fricatives, or more like European Portuguese (as well as many
233 Germanic or Slavic languages) in exhibiting heavy devoicing (Pape and Jesus

2015); (b) the realization of the plain-palatalized contrast across different fricatives and places of articulation; and (c) the potential similarities between our acoustics-based classification and the perceptual behavior of human subjects regarding the plain-palatalized contrast. Spinu (2010) reported very asymmetrical perceptual results, with subjects being most sensitive to the secondary palatalization contrast in dorsals and labials, and least sensitive to this contrast in postalveolars. Additionally, (d) we take a closer look at the properties of the dorsal fricative, which has previously been described as either glottal or velar (Mallinson 1986), without having been systematically examined in an acoustic study. Most recently, Sarlin (2014) lists [x] and [ç] as allophonic realizations of /h/ in Romanian.

Similarly to previous studies, we also pursue the question of which of these measurements are most important for the classifications. Specifically, we aim to identify the most predictive coefficients and moments as well as the most informative regions inside the segments.

To summarize, as a result of conducting this study, we expect to determine:

1. whether cepstral coefficients or spectral moments are better suited to classify Romanian fricatives;
2. whether HMM-defined regions or equal-length regions are better suited to classify Romanian fricatives;
3. which combinations of coefficients, moments, and regions are most informative towards the classification of Romanian fricatives, and
4. some of the properties of Romanian fricatives (including aspects of their voicing and secondary articulation, and the place of articulation of the dorsal fricative) and the ways in which they correlate with previous findings.

260 4. Method

261 4.1. Secondary palatalization and the fricative inventory in Romanian

262 Before describing the experimental set-up, a brief overview of Romanian
 263 fricatives and the mechanism behind secondary palatalization in this language
 264 is in order. In Romanian the plain-palatalized contrast arises as the result of
 265 a rule operating in certain phonological and morphological contexts (Chitoran
 266 2002). Surface palatalized consonants are usually associated with the presence
 267 of one of the two homophonous inflectional suffixes consisting of /-i/: the plural
 268 of nouns and adjectives, and the second person singular present indicative of
 269 verbs.² After consonants the /-i/ itself is not pronounced, but is manifested as
 270 palatalization, characterized by a secondary palatal gesture of the tongue body
 271 accompanying the primary articulation of the respective consonant.³ Given that
 272 secondary palatalization in Romanian depends on the presence of /-i/ after a
 273 consonant at the end of a word, we observe this phenomenon only in codas.

274 Table 2 displays the underlying fricative phonemes of Romanian, as well as
 275 their surface word-final realizations in plain and palatalized forms. Note that
 276 /s/ does not become [sʲ], but rather undergoes full palatalization to [ʃʲ] in the
 277 context of the /-i/ suffix, e.g. [pas] *step* → [paʃʲ] *steps*. The phone [sʲ] does not
 278 exist in Romanian.

279 [INSERT TABLE 2]

280 4.2. Stimuli

281 The stimuli consisted of pairs of words ending in a plain and a palatalized
 282 version of the following five fricatives: /f, v, z, ʃ, h/. Thus, four places of

²There is also a restricted number of monomorphemic items that exhibit final secondary palatalization in the absence of the usual morphemes, as in [ungʲ] *angle* (pl. [ungʲurʲ]). In these cases, the lexical item is taken to end in an underlying /-i/.

³See Spinu et al. (2012) for sample spectrograms and analysis. The possibility that the palatalized segment is actually a sequence of consonant plus vowel is unlikely given two lines of evidence: (1) the palatalized form is not significantly longer; and (2) there is no evidence for an extra syllable.

283 articulation are considered: labial, dental, postalveolar, and dorsal. As the
 284 exact nature of the dorsal fricative in Romanian is yet to be determined in an
 285 acoustic study, our study aims to clarify this issue.

286 Since the corpus collected was originally geared toward the investigation of
 287 secondary palatalization (Spinu, 2010) the segments included do not represent
 288 an optimal, balanced sample for the testing of place of articulation and voicing.
 289 Words ending in /s/ were excluded from the corpus because of the lack of a
 290 palatalized allophone distinct from [ʃ], and not enough words ending in [ʒ]
 291 met the word shape criteria imposed on the corpus. Furthermore, Romanian
 292 does not have a (phonemic) voiced dorsal fricative. Thus, while ideally our
 293 corpus would have contained voiced and voiceless segments from all four places
 294 of articulation, it was only feasible to include those listed above, and exclude [s]
 295 and [ʒ].

296 The list of stimuli is provided in Table 3. There were four pairs (8 target
 297 items) for each fricative. Half of the target items ended in a plain fricative
 298 (morphologically singular forms) and the other half in a palatalized fricative
 299 (corresponding plural forms). The stimuli were presented in a context-neutral
 300 carrier sentence, as in (1), where the targets are shown in phonetic transcription
 301 inside square brackets. All segments of interest were in word-final position in
 302 real words of Romanian, which were all disyllabic and bearing final stress.

303 (1) Am să aleg cuvântul [pantof/pantofʲ] când voi fi gata.

304 'I'll choose the word [pantof/pantofʲ] ('shoe/shoes') when I am ready.'

305 The vowels preceding the target fricatives were not strictly controlled for, as the
 306 other requirements (final stress, real words, final segment) had already imposed
 307 limitations on word selection. However, high front vowels were excluded because,
 308 as typical triggers of secondary palatalization in other languages, their presence
 309 might have resulted in coarticulatory effects on our targets. Thus, the preceding
 310 vowels were restricted to [e, a, o, u]. In all cases, the segment following the target
 311 items was the voiceless stop [k] (the first sound in *când*, the Romanian word for
 312 "when").

[INSERT TABLE 3]

In addition to the target sentences, twice as many fillers were included. The fillers were also paired, but showed inflections other than singular or plural, so as to distract the subjects from the target pairs. The set of 120 items (40 targets + 80 fillers) was presented three times to each subject, with the items automatically randomized for each block. Thus, a full set of recordings contained 120 target items per subject: 5 consonants $\times$ 4 items (words) per consonant $\times$ 2 forms per item (plain and palatalized) $\times$ 3 repetitions.

4.3. *Participants*

Thirty-one subjects (11 males, 20 females) with no history of speech or hearing disorders participated in the study. They were all native speakers of standard Romanian and ranged in age from 19 to 30, with an average age of 21.6 years. Some of the subjects reported speaking foreign languages (most often English and/or French) but had typically learned them in school, after the age of 10, and had not had long-term exposure to native speakers of these languages. None of the subjects reported being bilingual. They were all recruited and tested in Bucharest, Romania, and received compensation for their time.

4.4. *Procedure*

The experiment took place in a quiet room in Bucharest, Romania. Each subject was tested individually. All three blocks were recorded in a single experimental session, with short breaks in between. The subjects were recorded reading the target items and fillers on a Sony Vaio VGN-SZ440 notebook, with a Sennheiser PC 136 USB boomset. No amplifiers were used. The positioning of the microphone was level with the participant's mouth in the vertical dimension. Horizontally, it was placed close to the corner of the participant's mouth, out of the way of the puffs of air emitted by the mouth and nose while speaking. The microphone was positioned as close to the speakers' lips as possible without letting their lips touch it while speaking. Care was taken to ensure that the microphone was placed similarly for all participants. The headset's headphones

342 (which were not used) were placed over the subjects' ears, but did not block
343 them from hearing their own speech.

344 The stimuli were presented and recorded using a prototype of the Mod-
345 elTalker Voice Recorder (Yarrington et al. 2008). Each sentence was displayed
346 on the computer screen in Romanian orthography and a recording was made
347 as the subject read the sentence. The acoustic speech signal was recorded with
348 16-bit quantization and a sampling frequency of 16 kHz. No pre-emphasis was
349 applied. Acceptable ranges for f0 and amplitude were set based on microphone
350 calibration. The f0 range was set as $\pm 20\%$ of the average f0 of practice sentences
351 recorded during calibration, and the amplitude range was -12 to -1 dB. Silence
352 was defined as an amplitude lower than -45 dB at either end of a recording. If
353 a recording was produced outside of the acceptable ranges for f0 or amplitude,
354 or if silence could not be detected at either end, then the recording software
355 prompted the speaker to repeat the sentence. Silence in excess of 200 ms was
356 then trimmed, but no further manipulation of the recordings took place. When
357 a sentence was produced appropriately, the software automatically moved on
358 to the following sentence. Before beginning the actual experiment, the sub-
359 jects had a practice session with 20 items to familiarize themselves with the
360 procedure.

361 All of the subjects except one produced the full set of three repetitions; the
362 remaining subject only produced two repetitions. Some items were discarded
363 due to disfluencies, leaving a total of 3,674 fricatives for acoustic analysis.

364 After each sentence was recorded, the software automatically aligned a
365 phonetic transcription to the acoustic signal to estimate the start and end of
366 each fricative. Each alignment was then inspected manually and all errors were
367 corrected by hand. The segmentation was determined based on both waveform
368 and wideband spectrogram, following Jongman et al. (2000) and Nissen and
369 Fox (2005). The onset of a fricative was the point at which high-frequency
370 energy first appeared on the spectrogram, or the point at which the number
371 of zero crossings increased rapidly. The end of each fricative was the intensity

372 minimum immediately preceding the silence associated with the closure of the
 373 following stop, which was always [k]. These methods do not fundamentally
 374 deviate from other segmentation descriptions reported in the literature (Jesus
 375 and Shadle 2002, 2005, Hoelterhoff and Reetz 2007, Pape and Jesus 2015). As
 376 observed by Jesus and Shadle (2002), “during the vowel–fricative transition,
 377 there is a decrease in amplitude, voicing ceases (for unvoiced fricatives) and
 378 frication noise starts. [...] These events do not occur simultaneously or always
 379 in the same order, making the segmentation a somewhat subjective process” (p.
 380 442).

381 4.5. *Analysis*

382 4.5.1. *Dorsal fricative*

383 To clarify the issue of the exact phonetic realization of the Romanian dorsal
 384 fricative in coda position, all waveforms, spectrograms, and spectra (obtained
 385 using Praat version 5.4.01, Boersma and Weenink 2015) of plain and palatalized
 386 /h/ tokens (a total of 733 items) were visually inspected and assigned to a
 387 phonetic category. Following Nawrocki (2008), a velar segment [x] was identi-
 388 fied when the spectrum displayed 1-3 isolated bands, located at approximately
 389 2, 4-5 and 6-7 kHz. A glottal segment was identified when the spectrum and
 390 spectrogram displayed an arrangement of relatively robust spectral peaks imi-
 391 tating the formant structure of the preceding vowel (as the vowels differed in
 392 quality, these spectral peaks were also differently configured across the glottal
 393 tokens identified). Finally, during the analysis it became apparent that a third
 394 type of segment was present in the data, with almost no noise below 3 kHz and
 395 a friction range stretching from 3-6 kHz, sometimes with two distinguishable
 396 peaks around 3 kHz and 5-6 kHz. This realization closely corresponds to the
 397 voiceless palatal fricative [ç] (Nawrocki 2008), so tokens exhibiting this structure
 398 were categorized as such.

399 4.5.2. Acoustic feature sets

400 We compared two sets of acoustic parameter features. The first set, hence-
 401 forth referred to as the Spectral Moment set (SM), consists of the first four
 402 spectral moments (Bark-scaled), the RMS amplitude, and the total energy of
 403 the spectrum below 500 Hz, henceforth called the low-frequency energy or LFE
 404 (this is similar to McMurray and Jongman’s (2011) Low Frequency Energy mea-
 405 sure). The latter was included in order to compensate for the fact that none
 406 of the first four moments captures much information about the content of the
 407 spectrum at the low end of the frequency scale. The reasoning behind this was
 408 to give this feature set a higher chance of detecting voicing, to the extent that
 409 it manifests as low-frequency energy. The second set, henceforth the Cepstral
 410 Coefficient set (CC), consists of the first 6 cepstral coefficients (c0-c5), also
 411 Bark-scaled. The choice of Bark scaling was motivated by the fact that, while
 412 a pilot study with the same data revealed no significant differences between the
 413 performance of Bark and linear spectral moments, the Bark moments slightly
 414 outperformed the linear ones. Forrest et al. (1988) and Bunnell et al. (2004),
 415 also reported an advantage of Bark-scaled moments or coefficients over linear
 416 ones.

417 4.5.3. Feature extraction and division into regions

418 The acoustic feature sets (either SM or CC) were extracted from each
 419 fricative of interest as a series of analysis windows (also referred to as *feature*
 420 *vectors*) spaced 10 ms apart, using a Hamming window of 20 ms (Forrest et
 421 al. 1988, Halberstadt and Glass 1997).⁴ The set of feature vectors for each

⁴Jongman et al. (2000) opt for a 40-ms window at the expense of resolution in the tem-
 poral domain, but Koenig et al. (2013) discuss the differential distribution of acoustic cues
 over time and urge that future work “implement a more complex algorithm for finding the
 mid-frequency peak, taking into account expected variation as a function of speaker, time
 point, and/or phonetic context” (p.1187). Other studies of frication and voicing classifica-
 tion (Jesus and Jackson 2008, Pape and Jesus 2015, Pape et al. 2015) have found significant
 changes throughout the course of the fricative. As the acoustic characteristics of Romanian

422 phonetic segment was then divided into three regions using one of two methods,
423 as described below.

424 **a. Division into equal regions.** In the first method, we simply divided
425 the vectors into three regions of approximately equal duration. If the number
426 of vectors was not exactly divisible by three, then we partitioned them such
427 that the first and third regions had the same number of vectors, and the middle
428 region had either one more or one less. This method ensured that the middle
429 region stayed centered on the midpoint of the phone.

430 **b. Division by HMMs.** In the second method, we used HMMs to divide
431 the segments into regions of internally minimized variance. Each HMM consists
432 of three *states* arranged in a linear series. Each state models one region of a
433 phoneme, and the state’s parameters consist of the means and variances of the
434 feature vectors within that region. To train the model, the training segments
435 are first divided into regions of approximately equal duration, and the initial six
436 means and six variances of those regions are used as the initial parameters of the
437 model states. However, a feature vector on the edge of a region may lie closer
438 to the means of the adjacent state. Thus, in the next step, the boundaries
439 between the regions are recalculated so that the total likelihood of the data
440 is maximized with reference to the current model parameters. The means and
441 variances of the new regions are then recalculated. This process is repeated until
442 it converges (that is, no feature vectors are reassigned). The effect is to maximize
443 the differences between adjacent states while minimizing the differences among
444 feature vectors within the same state.

445 A cross-validation protocol was used to train 31 sets of models: each set
446 was trained on the data from 30 of the speakers, then used to segment the
447 withheld data from the remaining speaker. For each model set, separate HMMs
448 were trained to model plain and palatalized versions of each of the five target

fricatives, particularly with respect to voicing, are not known, and half of our corpus consists of palatalized segments, we opt here for better temporal resolution.

449 fricatives, for a total of 10 models per set. Following common practice in speech
 450 recognition, the feature vectors were speaker-normalized prior to training (that
 451 is, the speaker’s means were subtracted from the features for that speaker).
 452 Maximum-likelihood HMM training was performed to convergence once using
 453 the Viterbi algorithm (Viterbi 1967), and then again using the Baum-Welch
 454 algorithm (Baum et al. 1970). Finally, the Viterbi algorithm was used again to
 455 align the trained HMMs at the state level to the target segments, dividing each
 456 into three regions corresponding to the three states.

457 Note that we are using HMMs in this study only for segment division,
 458 not for the actual classification. Although the HMMs could be used for the
 459 classification, we chose to use logistic regression analysis for classification in our
 460 study, because regression enables us to examine more carefully the contribution
 461 of each coefficient to the success of the classification (Baayen 2008).

462 To summarize, the four combinations of acoustic feature sets and region
 463 division methods we compare are presented in Table 4.

464 [INSERT TABLE 4]

465 4.5.4. *Statistical analyses*

466 After dividing each target segment into three regions, the means of the
 467 features over all of the vectors in each region were calculated and used as in-
 468 put to the statistical analyses. This resulted in 18 different measures for each
 469 parameter set: 6 features (either cepstral coefficients or spectral moments) $\times$ 3
 470 regions inside each segment. Following McMurray and Jongman (2011), we con-
 471 ducted multinomial logistic regression analyses with consonant identity (place),
 472 palatalization, voicing, or gender as the dependent variable and the 18 measures
 473 as continuous explanatory variables. Logistic regression has been claimed to be
 474 best fitted for categorical response data, as theoretical problems arise with the
 475 application of discriminant analyses to them (Morrison and Kondaurova 2009,
 476 p. 2160). We used Matlab R2013a (MATLAB 2013) to determine the regres-
 477 sion coefficients. The main goal was to determine the combination of acoustic

478 measures and regions that would yield the most successful classifications.

479 The first analysis was run on the entire data, including all 5 fricative
480 phonemes. Because of the confound between place of articulation and voicing
481 in our corpus, we also ran analyses on two subsets: just the voiceless fricatives
482 [f, ʃ, h] and just the voiced fricatives [v, z]. We ran a fourth analysis on a subset
483 composed of only [f] and [v] in order to examine the properties of voicing in
484 Romanian labial fricatives and our method’s classification capabilities in regard
485 to voicing alone, as these two segments are assumed to differ only in laryn-
486 geal state (though Proctor et al. (2010) report that voiced fricatives are also
487 generally produced with a larger pharynx compared to voiceless fricatives, sug-
488 gesting that midsagittal analysis of the vocal tract alone is not sufficient for the
489 characterization of fricative production and voicing). Finally, the performance of
490 various combinations of acoustic measures (e.g. the top predictors from the set
491 of 18 measures) and out-of-sample classifications were examined in additional
492 analyses.

493 5. Results

494 In this section, we first provide descriptive results for the dorsal fricative
495 and visual displays of fricative spectra and HMM state distributions. This is
496 followed by the results of the various logistic regression analyses.

497 5.1. The dorsal fricative

498 The plain dorsal fricative was realized as velar [x] 86.7% of the time and as
499 glottal [h] 13.3%, based on the analysis of 368 items. There was a slight tendency
500 for males to produce more glottal realizations (males 15.9%, females 11.8%),
501 but this difference did not reach statistical significance. As for the palatalized
502 forms (365 items), all but one (99.7%) were realized as the palatal fricative [ç];
503 the exception was realized as a glottal by a female subject. Even though the
504 dorsal phoneme was overall realized at three different places of articulation, we
505 treat these as a single “place” for classification purposes and distinguish only

506 plain and palatalized realizations (as we did with the other fricatives). The
 507 reasons behind this decision include the fact that the palatal realization [ç]
 508 corresponded to the palatalized form 100% of the time, and 0% of the time to
 509 the plain one. This type of allophony may also be present in the other fricatives
 510 produced mainly with a tongue tip/tongue body gesture, such as [z] or [ʃ]. If
 511 there is sufficient acoustic similarity between the plain and palatalized forms,
 512 we expect them to be classified as part of the same place category. As for the
 513 glottal realizations, they are too few to warrant the introduction of a different
 514 place category. We understand that this introduced a source of variability for
 515 the classifier, but ultimately this level of free variation may well exist with
 516 other places tested, for instance labiodental fricatives, which “exhibited the most
 517 variability across speakers” in an articulatory study using magnetic resonance
 518 imaging (Narayanan et al. 1995, p. 1344). We leave it to future work to
 519 determine whether our classification method can reliably distinguish the glottal,
 520 velar, and palatal places from each other.

521 5.2. *Fricative spectra*

522 Cepstral coefficients are calculated by applying the Discrete Cosine Trans-
 523 form, which is completely reversible, to the output of a spectrum passed through
 524 a bank of 16 bandpass filters. Hence, the transform’s inverse can be used on
 525 the parameters of the trained HMM states, which consist simply of coefficient
 526 means, to back-calculate the mean filter outputs, and hence a rough “spectrum”
 527 for each HMM state representing the feature vectors associated with that state.
 528 This procedure was used with the trained CC HMMs to calculate a mean spec-
 529 trum over all speakers and utterances for each state of each fricative, as shown
 530 in Figure 1.

531 [INSERT FIGURE 1]

532 A few characteristics of these spectra are worth pointing out. First, there
 533 are distinct spectral shapes for each phoneme evident in the first region, but
 534 these differences are much less noticeable in the middle and final regions. In

particular, the low-frequency “voice bars” of the voiced consonants /v/ and /z/ are visible in the first region, but completely disappear in the other two regions. On the other hand, the distinctions between plain and palatal allophones are clearly noticeable only in the third regions – except for the dorsal, where there are obvious differences in all three regions. This palatalization, which manifests as high amplitudes in the mid-frequency range, is apparent for the non-sibilants but much less so for the sibilants /j/ and /z/, which have energy in that frequency range in their plain forms as well.

5.3. HMM state distribution

Figure 2 displays the mean distribution of frames assigned to each region of the fricative by the HMMs, using the cepstral coefficient parameter set (left) and the spectral moment set (right). For the most part, the segments were divided as expected, with shorter first and third regions (corresponding to the transition phases) and a longer, steady-state portion in the middle. The most notable difference between sets is in their treatment of the palatalized dorsal fricatives: whereas the CC HMM assigns them a long initial region (mean 106 ms), the SM HMM instead assigns them a long final region (mean 127 ms).

[INSERT FIGURE 2]

5.4. Logistic regression: all measures

5.4.1. All consonants together.

Figure 3 shows the overall classification results for all parameter sets and for each classification task, when all 18 measures are used as predictor variables. Both CC parameter sets performed better than the SM sets in most cases, and the HMM-divided regions generally yielded better classification results within sets.

[INSERT FIGURE 3]

Figure 4 shows detailed results of the consonant identity classification task, broken down by consonant. Overall and for almost every individual consonant,

563 the CC HMM parameter set performed best, with an overall classification rate
 564 of 95.29%, followed by the CC EQ set (90.58%), the SM HMM set (88.49%), and
 565 the SM EQ set (84.35%). On average, then, the CC sets outperform the SM sets
 566 by 6.5%, and region selection by HMM improves classification by another 4.4%.
 567 Particularly large differences between the CC HMM and other sets are noted
 568 with the classification of the labials /f,v/, which may be partly due to the large
 569 differences in region division for palatal /f,v/ between the CC HMM and SM
 570 HMM sets, as shown in Figure 2. For /v/, large differences are also seen between
 571 CC EQ and SM EQ, indicating that feature differences and region differences
 572 each independently account for a portion of the large differences in classification
 573 accuracy. For /f/, however, only the CC HMM set has a substantially different
 574 accuracy, indicating that it is the *combination* of feature sets and region divisions
 575 that produces the difference. Note however that a large difference in region
 576 divisions will not necessarily lead to a large change in classification accuracy;
 577 for example, the most striking difference in Figure 2 between the two feature
 578 sets is in their divisions of palatalized /h/, yet the classification difference for
 579 /h/ is small relative to the difference for /v/. How the regions are divided will
 580 have little effect if, for example, the variance of the features throughout the
 581 consonant is small to begin with.

582 [INSERT FIGURE 4]

583 Table 5 shows the individual consonant confusions for the CC-HMM set.
 584 Although the correct classification is very high overall, /f/ and /h/ yielded
 585 slightly lower correct classification rates as compared to the other three seg-
 586 ments: about 5% of each was classified as the other. We also find about 3%
 587 each of /f/ and /v/ classified as the other. These were also the two most com-
 588 mon confusions for the other three parameter sets. However, the third most
 589 commonly confused pair for the other sets was /v/ and /h/, with error rates
 590 between 3% and 6%. Despite the fact that one is voiced and the other voice-
 591 less, these two segments have similar spectra, as seen in Figure 1, each with two
 592 peaks of similar amplitude, one at low frequency and one at mid-frequency. The

593 CC HMM set, however, is very successful at distinguishing the two, as shown
594 in Table 5, with a mean confusion rate of only 0.34%.

595 [INSERT TABLE 5]

596 For the other classification tasks (by palatalization, by voicing, and by
597 gender), the same ordering of results generally held: accuracies were higher
598 with cepstral coefficients than with spectral moments, and usually higher with
599 HMM-determined regions than with equal-length regions (see Figure 3). For
600 the CC HMM set, the classification rates were 88.2% for palatalization, 97.3%
601 for voicing, and 79.7% for gender.

602 Figure 5 shows a breakdown of the palatalization results for the CC HMM
603 set. As noted in Section 5.2, the spectral differences between plain and palatal-
604 ized segments are least clear with the sibilants. Indeed, most palatalization
605 classification errors were encountered with [ʃ], as about half of the plain tokens
606 were miscategorized as palatalized – a pattern also encountered in human per-
607 ception in an experiment based on this corpus (Spinu et al. 2012). The other
608 coronal fricative, [z], was the segment with the next highest rate of errors –
609 another result found in the perception experiment.

610 [INSERT FIGURE 5]

611 5.4.2. Consonant subsets.

612 For the consonant subsets (voiceless, voiced, and labial), the results of
613 classification by consonant identity are shown in Figure 6 for the four parameter
614 sets. As before, classification rates for the CC parameter sets were higher than
615 for the corresponding SM sets overall, indicating that the higher classification
616 rates for cepstral coefficients obtained previously were not because of the data
617 confounds between place and voice. Although the size of this effect varied from
618 one subset to another, the number of classification errors in the SM-HMM set
619 was reduced 45-83% in the corresponding CC HMM set. As for the advantage of
620 using HMM-defined regions, the effect is very small to nonexistent in both the

voiced and voiceless sets, but quite large in the labial subset. This may indicate that the use of HMM-defined regions is particularly useful for distinguishing between voiced and voiceless fricatives. In fact, it is with plain and palatal [f] that some of the largest deviations from equal length can be observed in the lengths of the regions determined by the HMMs (see Figure 2).

[INSERT FIGURE 6]

Classifications by palatalization and gender were also performed on the consonant subsets (Figure 7). While cepstral coefficients continued to outperform spectral moments in all cases, the advantage of using HMM-divided regions was often small or non-existent – particularly in the classification of gender, where the SM EQUAL set obtained higher classification rates than the SM HMM set for every consonant subset. This may indicate that acoustic cues for gender tend to be evenly spread throughout the fricative. The highest classification rates for gender were obtained with the voiced subset for all parameter sets, suggesting that gender cues are more available in voiced consonants than in voiceless consonants. Much higher classification rates by palatalization were obtained with the labial subset (95-97%) than with either the voiceless or voiced subsets (83-93%). This is likely because the latter two subsets contain [ʃ] and [z], respectively, for which it is particularly difficult to identify palatalization, as already discussed.

[INSERT FIGURE 7]

5.5. Contributions of fricative regions

Figure 8 presents the classification rates obtained from just the acoustic measures extracted from the individual regions of the fricative, for the CC HMM and SM HMM analyses, for classifications by consonant, by palatalization, and by gender. The rates using all three regions are also included for reference.

[INSERT FIGURE 8]

For classification by consonant identity, drops in accuracy from 10 to 33% were recorded when restricting the acoustic measures to a single region, but a

rate as high as 84.6% can still be reached with cepstral coefficients when using only the first region of the fricative. For spectral moments, a slightly higher accuracy was reached with the middle region than the first (78.8% vs. 75.1%). Much lower classification rates were observed with the third region for both parameter sets, where the spectra tend to lose their distinct shapes (see Figure 1). This suggests that the perceptual cues for consonant identity are mostly located in the beginning and middle of the fricative.

In contrast, the highest classification rates by palatalization were obtained with the last region for both parameter sets, with drops of only 2% relative to using all 3 regions. This result is consistent with the observation of distinct spectra in the third region as shown in Figure 1. The largest drops in accuracy – 20% for both parameters sets – were obtained with the first region.

With gender classification, different patterns are seen between the two parameter sets. With the CC HMM set, accuracy decreases by 6% using the first region, and by 10% using either of the other two. With the SM HMM set, however, the rate drops are much smaller: from 2% with the last region, to 4% with the middle region.

Finally, results for classification by voicing are shown in Figure 9, for all fricatives (left) and for the labial subset (right). With all consonants, voicing classification rates decrease less than 3% when using parameters from just the first region of the fricative, but as much as 26% using just the last region. This decrease is even higher – up to 39% – with the labial subset. This suggests that the voice distinction is mostly realized at the beginning of the fricative, and is consistent with the possibility that Romanian fricatives tend to devoice towards the end of the segment. This hypothesis is corroborated by the reconstructed spectra (Figure 1) that indicate that the voice bar for /v, z/ is almost entirely absent in the final two regions.

[INSERT FIGURE 9]

To sum up, high classification rates can still be obtained when using only

single regions of the fricative, although the region required depends on the classification goal and sometimes the parameter set used. While similar drops in accuracy were observed between the two parameter sets when using only one region, cepstral coefficients still generally perform better than moments when using one region, just like they do when using all 3 regions.

5.6. Top measures contributing to classification

In order to investigate more closely which of the parameters in each set were most essential to the classification of the fricatives, we examined the coefficients of the regression models. Each parameter is associated in the model with a number of beta coefficients (equal to the number of categories minus one), and each beta coefficient is associated with a t-statistic (or *t-stat*), a measure that the beta's true value does not equal zero. Roughly speaking, the greater the t-stat, the larger the parameter's contribution to the categorization. Table 6 lists the three most significant parameters for each classification, as determined by the order of the maximum of the absolute value of the t-stats associated with each parameter.⁵

[Insert Table 6]

5.6.1. Consonant identity

Considering this measure, the most significant contributors toward the classification of consonant identity using the CC HMM parameter set are, in rank order, the mean c2 (the second cepstral coefficient) and mean c1 extracted from the consonant's initial region, and mean c1 extracted from the middle region. The regions chosen by this method agree with the analysis of regions in the previous section. As for the coefficients, c1 roughly corresponds with spectral

⁵In most cases, using the mean absolute t-stat instead of the maximum resulted in the same top three parameters. In the few cases where they differed, regressions using the top 3 parameters by maximum t-stat produced higher correct classifications than using the mean.

703 tilt, increasing with low frequencies and decreasing with high ones, while a large
704 c2 is associated with low energy around 650-3500 Hz and high energy at lower
705 or higher frequencies.

706 Table 6 also shows the most significant contributors for the SM HMM
707 parameter set. Again, the first and second regions are selected as most useful.
708 The best performing parameter is m2 (the second spectral moment), which
709 measures the variance of the energy distribution, extracted from the middle
710 region. This is followed by m3, the spectral tilt or skewness, and the low-
711 frequency energy – both extracted from the initial region.

712 New logistic regression classifications were performed using only the top 3
713 parameters for each set. The results, as seen in Table 6, show that reasonably
714 high classification rates are maintained even with only 3 parameters. There is a
715 larger drop in accuracy with the spectral moments (14%) than with the cepstral
716 coefficients (10%), however.

717 Figure 10 displays three-dimensional plots of the top three factors from
718 CC HMM and from SM HMM. The separation of the tokens into relatively
719 distinct areas is apparent, though more so for the cepstral coefficients than
720 for the moments. The fact that the dorsal appears to fall into two distinct
721 regions may reflect the separation of two different phonetic realizations, velar
722 and palatal. One of the regions formed by /h/ overlaps with the postalveolar
723 region, suggesting this might be the case. Some of the outlying dorsal fricatives
724 may be glottal realizations as well.

725 [INSERT FIGURE 10]
726

727 5.6.2. Palatalization

728 For palatalization, the cepstral coefficients c4 and c3 extracted from the
729 third region of consonant segments are by far the strongest contributors, while
730 c2 plays a smaller role. As seen previously, the final region of the consonant is
731 very important for the detection of palatalization. The coefficient c4, which is

732 associated with low energy around 4000 Hz, appears to have played the most
733 important part towards classification. As palatalized segments tend to have high
734 energy around 4000 Hz (see Figure 1), they tend to have a negative value for
735 c4. The coefficient c3 is more associated with high energy around 3000 Hz, and
736 so palatalized segments usually have a positive c3 value. As for the SM HMM
737 set, the top parameters are the first 3 moments, all extracted from the final
738 region. The third spectral moment (skew or spectral tilt) has a much higher
739 t-stat, suggesting a far greater importance for this parameter.

740 As shown in Table 6, logistic regressions using only the top 3 factors for
741 each parameter set maintain a high level of classification, dropping only 3-4% in
742 accuracy relative to the full analyses. Figure 11 plots the data according to the
743 top 3 parameters for each parameter set. Again, a cleaner separation of groups
744 is seen with the cepstral coefficients.

745 [INSERT FIGURE 11]
746

747 5.6.3. Voicing

748 Table 6 also provides the 3 most important contributors to the classification
749 of voicing. As was seen before, the first region of the fricative appears to be
750 particularly important for voicing identification. For the CC HMM set, the
751 large t-stats suggest that c1 and c2 are particularly useful for classification,
752 with c5 (which correlates with energies around 1250 and 4000 Hz) making a
753 smaller contribution. This is a highly informative triplet of measures: a logistic
754 regression using only these three cues maintains a classification rate of 95.8%,
755 a drop less than 1.5% from the regression using all 18 cues.

756 For spectral moments, the top contributors were m3 and m2 from the first
757 region, and m2 from the third region as well. But these three cues are not nearly
758 as informative as the full set: classification drops 10% when using only these 3
759 cues. Note that the Low-Frequency Energy is not one of the top 3 parameters.
760 While this appears surprising, it should be noted that this classification was run

on the set of all five consonants, and larger differences might be found for other parameters between various pairs of voiced and voiceless consonants. Note in Figure 1 that the difference in amplitude in the low-frequency range between the voiced and voiceless is segments is not actually very large even in the first region. Even for the labials, a larger difference in amplitude may be observed in the mid-frequency range.

Figure 12 plots the data by the top 3 cues for each parameter set. A much cleaner separation between voiced and voiceless consonants is observed with the CC set than the SM set.

[INSERT FIGURE 12]

5.7. Out-of-sample analyses.

To test whether the high classification rates we obtained for consonant identity carry over out-of-sample, we conducted a new series of analyses in which we trained CC HMM and SM HMM logistic regression models on randomly-selected sets of subjects, and tested on the out-of-sample subjects. We ran 15 cases, in which the number of held-out subjects varied from 1 to 15. For each case, we ran 100 Monte Carlo scenarios, in which that number of subjects was randomly selected and left out. The remaining data were used to train the logistic regression model. Finally, the beta coefficients from the training set regression were applied to the out-of-sample data to predict the categories of those data.

The standard deviations of the classification rates vary for the CC-HMM set from 3.21% (for one subject withheld) to 0.87% (for 15 subjects withheld); similarly the standard deviations for SM-HMM vary from 4.82% to 1.08%. The results show that classification rates are very robust even when almost half of the data is held out. The difference in mean classification rates between the in-sample and out-of-sample data is always less than 2.5%.

789 6. Discussion

790 6.1. Cepstral coefficients vs. spectral moments

791 While previous work suggests that CCs may present advantages over other
792 types of measures (Ferragne and Pellegrino, 2012), this claim has been only
793 weakly corroborated by recent results such as Kong et al. (2014). That study
794 used a set of 13 MFCCs and only classified three places of articulation for an
795 overall correct classification rate of 85%. Our study shows that higher correct
796 classification rates can be obtained with a more parsimonious set of CCs and an
797 additional place of articulation. The rates we obtained for the classification of
798 consonant identity exceeded 95% both when all four places of articulation were
799 included, and when the corpus was restricted to voiceless consonants (three
800 places of articulation). While the CC HMM set yielded the highest classifi-
801 cation rates for all classifications attempted, particular advantages (over 5%
802 higher accuracy as compared to the SM sets) were noted in the overall classi-
803 fication of consonant identity, the classification of labials in the full consonant
804 set, the overall classification of voicing, and the classification of gender when all
805 consonants were considered as well as in the voiced subset.

806 Furthermore, the robustness of the CC method was apparent when we
807 considered the restriction of the acoustic parameter sets to the top three con-
808 tributors. Using only three predictors, overall classification of the CC HMM set
809 exceeded that of the SM HMM set by approximately 11% for consonant identity,
810 5% for palatalization, and over 14% for voicing.

811 6.2. HMM-determined regions vs. equal-length regions

812 The advantage of HMM-defined over equal regions was not as clearcut as
813 that of CCs over SMs, but it is still the case that for most types of classification
814 a small advantage was noted. In some classifications, the advantage of HMM
815 regions was more apparent with SMs (e.g. the classification of /z/ and /h/ in
816 the full consonant set), while in others it was so with CCs, with a difference
817 exceeding 5% in all cases (e.g. the classification of /f/ and /v/ in the full

consonant set; the overall classification of voicing). SMs stand out in that the equal-region set slightly outperformed HMMs for the overall gender classification (as well as in every subset) and in the classification of palatalization in the labial subset, a pattern never encountered with CCs. Thus it may be concluded that the advantage of HMM-defined regions is most notable when the acoustic measures are based on CCs, but may not play a substantial part when SMs are employed.

While the HMM- versus equal-region comparison is not fully conclusive, it is important to note that using regions as opposed to averaging over an entire segment has noticeable advantages in the classification of palatalization. Spinu et al. (2012) obtained an overall classification rate of 78% based on the same fricative corpus, using averages over the entire segments for Bark CCs 0-5 plus segment duration as predictors. In the current study, 88.2% correct classification for palatalization was obtained with HMM regions and 87.3% with equal regions, even without using segment duration.

6.3. *Top contributors*

Some of the previous studies reported the most important SM contributors towards the classification of place of articulation, but there does not appear to be a high degree of agreement in which these were. Spectral skewness (m3) was the highest contributor in Forrest et al. (1988), spectral mean (m1) was most informative in Jongman et al. (2000), and skewness and variance (m2) were most useful towards classification in Nissen and Fox (2005). Providing a more fine-grained perspective, McMurray and Jongman (2011) determined m1, m2, and m3 obtained from frication portions to be informative for place in sibilants, and m3 and m4 obtained from the vowel-fricative transition to be important to the classification of place in non-sibilants. They also found that m2 is an important contributor in the detection of sibilance, regardless of place.

The results of our study also suggest that fine-grained distinctions can be made depending on the classification goal (consonant identity, palatalization, or

voicing). Thus, for the SM HMM set, the main contributors for the classification of consonant identity were m2, m3, and LFE. For palatalization, m3, m2, and m1 were most useful, in that order. The main contributors in the classification of voicing were m3 and m2. Classification rates obtained using only these measures as predictors ranged from 74.5% to 81.3%, compared to 84.49%-91.43% for the full set of 18 measures. For the CC HMM set, we found c2 and c1 to be the highest contributors to consonant identity (place). For the classification of palatalization, c4, c3, and c2 were most informative and for the classification of voicing c1, c2 and c5 were the top 3 contributors. Correct classification rates based on these coefficients alone were between 85.17% and 95.81%, compared to 88.16%-97.25% based on the full set with 18 coefficients.

As far as the contribution of different regions is concerned, our results suggest that the first and second regions inside fricatives are most useful towards place and voicing classification whereas the last region is best suited for the classification of palatalization. The classification of gender was relatively stable across all regions.

6.4. Classification rates compared to prior studies

Our classification rates for place of articulation (95.3% for the full consonant group and 95.6% for voiceless consonants only) were higher than any of those reported in the studies presented in Section 2. Those studies, however, examined pre-vocalic fricatives rather than post-vocalic ones, so caution must be taken in comparing them. Moreover, our corpus, did not pose the added challenge of distinguishing between two front places of articulation, labiodental and interdental, so a direct comparison is only warranted with Kong et al.'s (2014) study. Even though labiodental and interdental fricatives were included in Kong et al.'s study, they were collapsed into one place and thus only three places of articulation were classified, yielding a correct classification rate of 85%.

It is, however, noteworthy that our high correct classification rates with cepstral coefficients were obtained based solely on one type of measurement,

precluding the need for additional measurements of many different types, such as those employed in Jongman et al. (2000) and McMurray and Jongman (2011). We must also consider the fact that the syllabic position of our fricatives (in the coda) was not ideal, as codas have been claimed to be a weaker position for the realization of acoustic cues for consonants in general (Fujimura et al. 1978, Ohala 1990, Wright 2004). In addition, we have not used any measures that straddle the vowel-fricative transition, unlike most previous studies, and in doing so we may be missing information important to classification, as this region has proven particularly informative in previous research. All of these suggest that our classification method is very robust.

6.5. *The voicing contrast in Romanian*

The production of voiced fricatives involves “the complex interaction of articulatory constraints from three separate goals: the formation of the appropriate oral constriction, the control of airflow through the constriction so as to achieve frication, and the maintenance of glottal oscillation by attending to transglottal pressure” (Proctor et al. 2010, p. 1507). The extent to which this articulatory complexity plays a part in fricative devoicing varies from language to language. Studies of European Portuguese (Jesus and Shadle 2002, Pape and Jesus 2014, 2015) indicate that heavy devoicing occurs in the VC position (up to 76.5% of the time full devoicing for [z] and 48.4% for [v], with additional partially devoiced segments). Our corpus enabled a closer look into this mechanism for Romanian, as the two labials included formed a phonologically voiced and voiceless pair. While we found that /f/ can be distinguished from /v/ over 95% of the time when all regions of a segment are used together, the picture is very different when the regions are considered separately. Only the first region yields correct classification rates above 90%, and a marked decrease is noted for the other two regions, to approximately 57% in the third region. Given that only two consonants are classified, this is only slightly higher than chance. Thus, we may conclude that partial devoicing – possibly to a very large extent – is present in word-final Romanian fricatives. It is not clear whether this process

906 is independent of the context or may best be explained by coarticulation with
907 the following segment, the voiceless obstruent [k]. Future studies may clarify
908 this issue and thus determine whether Romanian behaves more like European
909 Portuguese (and Germanic and Slavic languages) or other Romance languages
910 such as Italian or Spanish.

911 6.6. *Palatalization in Romanian*

912 The classification rates for palatalization status (displayed for the CC
913 HMM set in Figure 4) are strongly reminiscent of the performance of human
914 subjects in a series of perceptual experiments based on stimuli from the same
915 corpus (Spinu 2010, Spinu et al. 2012). Similarly to humans, our feature vec-
916 tors yielded the highest correct classification rates with the dorsal fricative, and
917 the lowest with the postalveolar fricative, with a strong bias to categorize plain
918 postalveolars as palatalized. The increased error rate with the alveolar fricative
919 [z] as compared to the labials was also encountered with human perception.
920 Overall, our classification closely replicated human performance. This suggests
921 that the information in these acoustic measures may do a good job of capturing
922 the similarity relations underlying listeners' categorization.

923 From the perspective of acoustic classification, we have found the third
924 region to be by far the most informative with respect to the presence or absence
925 of secondary palatalization. Given that the HMM-defined third regions were
926 relatively short in some cases (below 30 ms), a tentative conclusion may be
927 that, while primary and secondary gestures can be realized simultaneously, the
928 secondary palatalization gesture in Romanian may to at least some degree follow
929 the primary articulation, such that the acoustic cues associated with it are
930 realized very strongly at the end of the segment. Sequential realization is also
931 partially supported by the results for segment duration in Spinu et al. (2012),
932 with longer duration observed for all palatalized segments compared to their
933 plain counterparts. Significant differences in duration, however, were only found
934 between plain and palatalized /v/ and /h/.

935 6.7. *The dorsal fricative in Romanian*

936 Our findings indicate that plain and palatalized /h/ are realized at distinct
937 places of articulation, thus confirming the validity of the allophones proposed
938 by Sarlin (2014). Furthermore, we have found that plain /h/ is mostly realized
939 as velar [x], with a small number of speakers showing a preference for glottal [h],
940 and other speakers using the glottal realization sporadically. This finding is con-
941 sistent with Mallinson’s (1986) observation that before liquids and in word-final
942 position, /h/ can be realized as a velar. Six percent of the palatalized [çʲ] tokens
943 were also classified as plain (100% of the plain tokens were classified correctly),
944 but the overall correct classification rates collapsing plain and palatalized dor-
945 sals (e.g. CC HMM consonant identity classifications for all consonants or for
946 the voiceless subset) were above 95%. These miscategorizations did not have
947 a strong effect on the overall success of the classifier. While [h], [x], and [çʲ]
948 are distinct in spectrographic displays and human perception, the acoustic dif-
949 ferences between them may be small enough to mimic the variability present
950 with other fricatives even when this variability does not consist of differences in
951 primary place of articulation. In any case, the success of our classifier with the
952 problematic Romanian ‘dorsals’ suggests that it can handle variability well.

953 6.8. *Limitations*

954 For all acoustic parameter sets, a number of methodological issues made
955 the classification tasks relatively challenging. To begin with, the distribution of
956 place of articulation in the corpus was asymmetric, with twice as many labials
957 as each of the other places. The uneven distribution of vowels preceding the
958 fricatives also introduced a confound that could not be controlled for. While
959 the presence of that confound and the voicing confound may have facilitated
960 classification by consonant identity in some cases, the allophonic variation for
961 the dorsal fricative (and possibly other places as well) introduced a source of
962 variability. The segment following the target items, a voiceless obstruent, may
963 have induced devoicing of the phonologically voiced segments and thus interfered

964 with the classification of voicing. The absence of any vocalic information may
 965 have further affected classification rates, as previous studies have shown that
 966 it plays an important role towards distinguishing different fricatives, both on
 967 the basis of their acoustic properties and in human perception. The word-final
 968 environment is generally a weaker position for the realization of acoustic cues.
 969 Finally, the quality of the recording equipment (the Sennheiser PC 276 136
 970 USB boomset has a very limited linear response and high internal noise) and
 971 the restricted sampling frequency range (0-16 kHz) may have further added to
 972 the challenges. The robustness and consistency of our results, however, as well
 973 as their close match with human perceptual responses, suggest that we were
 974 able to collect good research material despite these limitations. For even better
 975 results, in future studies we intend to adhere to the recommendations proposed
 976 in Shadle (2006, 2010).

977 7. Conclusion

978 To conclude, we recommend the combination of cepstral coefficients with
 979 HMM-divided regions as a reliable tool for fricative classification. Our top three
 980 predictors alone were able to yield considerable separation of the data into rel-
 981 evant categories. Using additional predictors resulted in very high classification
 982 rates for categories such as consonant identity, voicing, palatalization, and gen-
 983 der, and also held out-of-sample verification. While our findings provide new
 984 information regarding Romanian fricatives, the limitations of our corpus do not
 985 enable us to compare our results with data from other languages. Fruitful fu-
 986 ture research directions would include addressing the relationship between the
 987 fricative spectra and the most important contributors for both the cepstral co-
 988 efficient and spectral moment set, and using our classification method with En-
 989 glish fricatives to determine whether the advantages of the cepstral coefficients
 990 and HMM-divided windows carry over to other languages. Related to this, we
 991 would like to see if our method fares better than previously tried methods at
 992 distinguishing between the front fricatives [f] and [θ]. Extending our analysis to

993 other languages will have the added benefit of making it easier to compare these
994 techniques and results to those from the past.

995

996 **Acknowledgements**

997 This research was supported in part by doctoral dissertation research im-
998 provement grant #0720231 awarded to the first author by the National Science
999 Foundation. We are particularly indebted to Florin Spinu for his generous help
1000 with the statistical analyses. The authors also thank Irene Vogel and Timothy
1001 Bunnell for their assistance. Portions of this research were presented at the 84th
1002 meeting of the Linguistic Society of America in 2010 and at the Very-Large-Scale
1003 Phonetics Research Workshop (VLSP 2011) in Philadelphia, PA.

Year	Study	Speakers	Fricatives	Acoustic Measures	Classification Method	Classification Accuracy
1988	Forrest et al.	10	f, θ, s, ʃ	SM	discriminant analysis	77.7%
1990	Tomiak	6	f, θ, s, ʃ, h	SM	linear function analysis	78%
2000	Jongman et al.	20	f, v, θ, ð, s, z, ʃ, ʒ	SM + 5 other types	discriminant analysis	77%
2005	Nissen and Fox	10*	f, θ, s, ʃ	SM + 3 other types	discriminant analysis	65%
2011	McMurray and Jongman	20	f, v, θ, ð, s, z, ʃ, ʒ	SM + 9 other types	logistic regression	raw: 79.2%-85% compensated: 87%-92.9%
2014	Kong et al.	168**	f, v, θ, ð, s, z, ʃ, ʒ	MFCC	SVM	approx. 85%***

Table 1: A selection of fricative classification studies over the past two decades. All of the classification schemes are based on acoustic parameters that include either spectral moments (SM) or cepstral coefficients (CC). MFCC = Mel-frequency cepstral coefficients. SVM = Support Vector Machine. ***10 adults.** The study also included different age groups for children, with the adults being the control group. ****168.** A subset of the speakers in the TIMIT database. *****No labial-interdental distinction.** The front fricatives were collapsed into one place category.

	LABIAL	CORONAL	DORSAL
Phonemic Inventory	/f/ /v/	/s/ /z/ /ʃ/ /ʒ/	/h/
Phonetic Realization: Plain	[f] [v]	[s] [z] [ʃ] [ʒ]	[x] [h]
Phonetic Realization: Palatalized	[fʲ] [vʲ]	[ʃʲ] [ʒʲ] [ʃʲ] [ʒʲ]	[çʲ]

Table 2: Romanian fricatives - phonemic and word-final phonetic realizations. The specific places are: labiodental: f, v; dental: s, z; postalveolar: ʃ, ʒ; palatal: ç; velar: x; glottal: h. The items used in the current study are shown in **boldface**.

Final C	Preceding V	Singular	Plural	Translation
[f]/[fʲ]	o	<pantof>	<pantofi>	shoe/s
	a	<vătaf>	<vătafi>	bailiff/s
	o	<cartof>	<cartofi>	potato/es
	u	<zuluf>	<zulufi>	curl/s
[v]/[vʲ]	a	<bolnav>	<bolnavi>	sick/pl
	a	<grozav>	<grozavi>	great/pl
	a	<firav>	<firavi>	feeble/pl
	a	<zugrav>	<zugravi>	house painter/s
[z]/[zʲ]	e	<obez>	<obezi>	obese/pl
	u	<ursuz>	<ursuzi>	morose/pl
	u	<mofluz>	<mofluzi>	grumpy/pl
	e	<chinez>	<chinezi>	Chinese/pl
[ʃ]/[ʃʲ]	a	<codaş>	<codaşi>	slacker/s
	e	<cireş>	<cireşi>	cherry tree/s
	o	<cocoş>	<cocoşi>	rooster/s
	u	<ghiduş>	<ghiduşi>	playful/pl
[x]-[h]/[çʲ]	a	<cazah>	<cazahi>	Cossack/s
	a	<valah>	<valahi>	Wallachian/s
	a	<monah>	<monahi>	monk/s
	a	<paroh>	<parohi>	vicar/s

Table 3: List of target words. The word-final phonetic realizations of each consonant are shown in the first column, first for the singular form and then for the plural form. Orthographic transcriptions are enclosed in angle brackets, < ... >.

Set	Acoustic measures	Segment division
SM HMM	Spectral Moments 1-4, RMSA, LFE	by HMM state
SM EQUAL	Spectral Moments 1-4, RMSA, LFE	equal-duration regions
CC HMM	Cepstral Coefficients 0-5	by HMM state
CC EQUAL	Cepstral Coefficients 0-5	equal-duration regions

Table 4: Acoustic parameter sets employed in our study. RMSA: Root mean square amplitude. LFE: low-frequency energy (total energy at frequencies less than 500 Hz). All acoustic measures were Bark-scaled.

CC HMM	f	v	z	S	h
f	.905	.034	.003	.004	.054
v	.030	.954	.011	.000	.005
z	.003	.007	.988	.000	.003
S	.004	.000	.007	.980	.010
h	.048	.001	.003	.010	.939

Table 5: Breakdown by consonant for the CC HMM classification.

	CON		PAL		VOICE	
	CC	SM	CC	SM	CC	SM
Contributor #1	CC2.R1 (-15.78)	SM2.R2 (-13.98)	CC4.R3 (-19.19)	SM3.R3 (-17.02)	CC1.R1 (16.78)	SM3.R1 (9.33)
Contributor #2	CC1.R1 (-12.71)	SM3.R1 (-13.58)	CC3.R3 (15.18)	SM2.R3 (9.16)	CC2.R1 (13.93)	SM2.R1 (-7.56)
Contributor #3	CC1.R2 (12.64)	LFE.R1 (-11.7)	CC2.R3 (8.44)	SM1.R3 (9.14)	CC5.R1 (-6.94)	SM2.R3 (6.80)
% Class. (All 18)	95.29	88.49	88.16	84.49	97.25	91.43
% Class. (Top 3)	85.71	74.50	85.17	80.16	95.81	81.30

Table 6: The top 3 contributors to the classification by consonant identity, palatalization, and voicing out of the 18 acoustic parameters from the CC and SM HMM sets, in descending order by t-stat (in parentheses). For each variable, the t-stat reported is the largest (in absolute value) of the several t-stats associated with it across the categories. The last two lines present correct classification rates using all 18 parameters, and using just the top 3 parameters.

1004

Figure Captions

1005 Figure 1. Mean plain (left) and palatalized (right) fricative spectra averaged
 1006 over all speakers and repetitions, as estimated from the parameters of the trained
 1007 HMM states. The letter *s* indicates the state (region).

1008 Figure 2. Average number of frames (feature vectors) assigned to each region
 1009 inside the fricatives, as determined using the CC HMM acoustic parameter set
 1010 (left) and the SM HMM acoustic parameter set (right). Plain tokens are in the
 1011 top row and palatalized ones in the bottom row.

1012 Figure 3. Logistic regression classification results for all parameter sets and
 1013 all classification tasks (consonant identity, palatalization, voicing, and gender),
 1014 averaged over all 5 consonants.

1015 Figure 4. Classification results for consonant identity for all parameter sets,
 1016 broken down by consonant.

1017 Figure 5. Results for classification by palatalization, broken down by consonant,
 1018 for the CC HMM set. Labels at the bottom indicate the consonant’s true
 1019 identity as plain or palatalized; colored columns indicate predicted categories.

1020 Figure 6. Consonant identity classification results for all parameter sets for
 1021 models trained and tested on consonant subsets: voiceless consonants (middle
 1022 left), voiced consonants (middle right), and labial consonants (right). The ALL
 1023 column (left) averages over all consonants to provide the mean correct classifi-
 1024 cation for each model.

1025 Figure 7. Palatalization and gender classification results for all parameter sets
 1026 for models trained and tested on consonant subsets: voiceless consonants (mid-
 1027 dle left), voiced consonants (middle right), and labial consonants (right). The
 1028 ALL column (left) averages over all consonants to provide the mean correct
 1029 classification for each model.

1030 Figure 8. Consonant, palatalization, and gender classification results for the
 1031 CC HMM and SM HMM parameter sets for logistic regression models trained
 1032 either on all parameters (black column), or just parameters extracted from a

single region (gray columns).

Figure 9. Voicing classification results for the CC HMM and SM HMM parameter sets for logistic regression models trained either on all parameters (black column), or just parameters extracted from a single region (gray columns). Left: Models trained on all 5 consonants. Right: Models trained on the labial subset.

Figure 10. (color online) Three-dimensional scatterplots of the three most informative acoustic measures for CC HMM (left) and SM HMM (right) for classification of consonant identity. Each point represents a single fricative token in the dataset. The first part of each axis label (e.g. CC1.R1) stands for the coefficient or moment (e.g. c1 or m3), and the last part stands for the region from which it was extracted (1, 2, or 3).

Figure 11. (color online) Three-dimensional scatterplots of the three most informative acoustic measures for CC HMM (left) and SM HMM (right) for classification of palatalization. Each point represents a single fricative token in the dataset. The first part of each axis label (e.g. CC1.R1) stands for the coefficient or moment (e.g. c1 or m3), and the last part stands for the region from which it was extracted (1, 2, or 3).

Figure 12. (color online) Three-dimensional scatterplots of the three most informative acoustic measures for CC HMM (left) and SM HMM (right) for classification of voicing. Each point represents a single fricative token in the dataset. The first part of each axis label (e.g. CC1.R1) stands for the coefficient or moment (e.g. c1 or m3), and the last part stands for the region from which it was extracted (1, 2, or 3).

Figure 13. The first six cosine waves corresponding to the first six cepstral coefficients (0 through 5). Points represent the center points of the 16 triangular bandpass filters used for the study.

References

- Baayen, H. R. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press.

- 1062 Baum, L. E., Petrie, T., Soules, G., & Weiss, N. (1970). A maximization
1063 technique occurring in the statistical analysis of probabilistic functions of
1064 Markov chains. *Annals of Mathematical Statistics*, 41(1), 164-171.
- 1065 Boersma, P. & Weenink, D. (2015). Praat: Doing phonetics by computer
1066 [Computer program]. Version 5.4.01 retrieved from <http://www.praat.org/>
- 1067 Borroff, M. (2007). *A landmark underspecification account of the patterning*
1068 *of glottal stop* (Doctoral dissertation). State University of New York at
1069 Stony Brook.
- 1070 Browman, C. P. & Goldstein, L. (1992). Articulatory phonology: An
1071 overview. *Phonetica*, 49, 155-180.
- 1072 Bunnell, H. T., Polikoff, J. B., and McNicholas, J. E. (2004). Spectral
1073 moment vs. bark cepstral analysis of children's word-initial voiceless stops.
1074 In *Proceedings of the 8th International Conference on Spoken Language*
1075 *Processing*. Jeju, Korea.
- 1076 Childers, D. G., Wu, K., Bae, K. S., & Hicks, D. M. (1998). *Automatic*
1077 *recognition of gender by voice*. Paper presented at the IEEE International
1078 Conference on Acoustics, Speech, and Signal Processing (ICASSP-88).
- 1079 Chitoran, I. (2002). *The phonology of Romanian: A constraint-based ap-*
1080 *proach*. Berlin, New York: Mouton de Gruyter.
- 1081 Deller, J. R. Jr., Proakis, J. G., & Hansen, J. H. L. (1993). *Discrete-time*
1082 *processing of speech signals*. Englewood Cliffs, NJ: Macmillan.
- 1083 Farooq, O. & Datta, S. (2001). Mel filter-like admissible wavelet packet
1084 structure for speech recognition. *IEEE Signal Processing Letters*, 8(7), 196-
1085 198.
- 1086 Ferragne, E. & Pellegrino, F. (2010). Vowel systems and accent similar-
1087 ity in the British Isles: Exploiting multidimensional acoustic distances in
1088 phonetics. *Journal of Phonetics*, 38(4), 526-539.

- 1089 Flipsen, P., Shriberg, L., Weismer, G., Karlsson, H., & McSweeny, J.
1090 (1999). Acoustic characteristics of /s/ in adolescents. *Journal of Speech,*
1091 *Language, and Hearing Research, 42*, 663-677.
- 1092 Forrest, K., Weismer, G., Elbert, M., & Dinnsen, D. A. (1994). Spectral-
1093 analysis of target-appropriate T and K produced by phonologically disor-
1094 dered and normally articulating children. *Clinical Linguistics and Phonet-*
1095 *ics, 8*, 267-281.
- 1096 Forrest, K., Weismer, G., Hodge, M., Dinnsen, D. A., & Elbert, M. (1990).
1097 Statistical-analysis of word-initial K and T produced by normal and phono-
1098 logically disordered children. *Clinical Linguistics and Phonetics, 4*, 327-340.
- 1099 Forrest, K., Weismer, G., Milenkovic, P., & Dougall, R. (1988). Statistical
1100 analysis of word-initial voiceless obstruents: Preliminary data. *Journal of*
1101 *the Acoustical Society of America, 84*, 115-124.
- 1102 Fujimura, O., Macchi, M., & Streeter, L. A. (1978). Perception of stop
1103 consonants with conflicting transitional cues: A cross-linguistic study. *Lan-*
1104 *guage and Speech, 21*, 337-346.
- 1105 Gafos, A. I. (2002). A grammar of gestural coordination. *Natural Language*
1106 *and Linguistic Theory, 20*(2), 269-337.
- 1107 Gordon, M., Barthmaier, P., & Sands, K. (2002). A cross-linguistic acoustic
1108 study of voiceless fricatives. *Journal of the International Phonetic Associ-*
1109 *ation, 32*, 141-174.
- 1110 Halberstadt, A. K. & Glass, J. R. (1997). *Heterogeneous acoustic measure-*
1111 *ments for phonetic classification 1*. EUROSPEECH 1997.
- 1112 Hamann, S. & Sennema, A. (2005). Acoustic differences between German
1113 and Dutch labiodentals. *ZAS Papers in Linguistics, 42*, 33-41.
- 1114 Hedrick, M. S. & Ohde, R. N. (1993). Effect of relative amplitude of frica-
1115 tion on perception of place of articulation. *Journal of the Acoustical Society*
1116 *of America, 94*, 2005-2027.

- 1117 Hoelterhoff, J. & Reetz, H. (2007). Acoustic cues discriminating German
1118 obstruents in place and manner of articulation. *Journal of the Acoustical*
1119 *Society of America*, 121, 1142-1156.
- 1120 Jesus, L. M. (2001). *Acoustic phonetics of European Portuguese fricative*
1121 *consonants* (Doctoral dissertation). University of Southampton, England.
- 1122 Jesus, L. M. & Jackson, P. J. (2008). Frication and voicing classification.
1123 In A. Teixeira, V. L. S. de Lima, L. C. de Oliveira, & P. Quaresma (Eds.),
1124 *Computational processing of the Portuguese language: The 8th Interna-*
1125 *tional Conference on Computational Processing of Portuguese (PROPOR)*
1126 (pp.11-20). Berlin/Heidelberg: Springer Verlag.
- 1127 Jesus, L. M. & Shadle, C. H. (2002). A parametric study of the spectral
1128 characteristics of European Portuguese fricatives. *Journal of Phonetics*,
1129 30(3), 437-464.
- 1130 Jesus, L. M. & Shadle, C. H. (2003). Devoicing measures of European
1131 Portuguese fricatives. In N. Mamede, J. Baptista, I. Trancoso, & M. Nunes
1132 (Eds.), *Computational processing of the Portuguese language* (pp. 1-8). Hei-
1133 delberg: Springer.
- 1134 Jesus, L. M. T. & Shadle, C. H. (2005). Acoustic analysis of European
1135 Portuguese uvular [χ, ʁ] and voiceless tapped alveolar [ɾ̥] fricatives. *Journal*
1136 *of the International Phonetic Association*, 35(1), 1-18.
- 1137 Jongman, A., Wayland, R., & Wong, S. (2000). Acoustic characteristics of
1138 English fricatives. *Journal of the Acoustical Society of America* 10/2000;
1139 108 (3 Pt 1), 1252-1263.
- 1140 Kochetov, A. (2002). *Production, perception, and emergent phonotactic pat-*
1141 *terns: A case of contrastive palatalization*. New York: Routledge.
- 1142 Koenig, L. L., Shadle, C. H., Preston, J. L., & Mooshammer, C. R. (2013).
1143 Toward improved spectral measures of /s/: Results from adolescents. *Jour-*
1144 *nal of Speech, Language, and Hearing Research*, 56, 1175-1189.

1145 Kong, Y.-Y., Mullangi, A., & Kokkinakis, K. (2014). Classification of frica-
1146 tive consonants for speech enhancement in hearing devices. *PLoS ONE*
1147 9(4): e95001. doi:10.1371/journal.pone.0095001

1148 Kuzla, C., Cho, T., & Ernestus, M. (2007). Prosodic strengthening of Ger-
1149 man fricatives in duration and assimilatory devoicing. *Journal of Phonetics*,
1150 35(3), 301-320.

1151 Lee-Kim, S., Kawahara, S., & Lee, S. (2014). The 'whistled' fricative in
1152 Xitsonga: Its articulation and acoustics. *Phonetica*, 71, 50-81.

1153 Lousada, M., Jesus, L., & Pape, D. (2012). Estimation of stops' spectral
1154 place cues using multitaper techniques. *DELTA*, 28(1), 1-26.

1155 Mallinson, G. (1986). Rumanian. In M. Harris & V. Nigel (Eds.), *The*
1156 *Romance languages* (pp.391-419). New York: Oxford University Press.

1157 Maniwa, K., Jongman, A., & Wade, T. (2009). Acoustic characteristics
1158 of clearly spoken English fricatives. *Journal of the Acoustical Society of*
1159 *America*, 125, 3962-3973.

1160 MATLAB and Statistics Toolbox Release 2013a. Natick, Massachusetts:
1161 The MathWorks, Inc.

1162 McMurray, B. & Jongman, A. (2011). What information is necessary for
1163 speech categorization? Harnessing variability in the speech signal by inte-
1164 grating cues computed relative to expectations. *Psychological Review*, 118,
1165 219-246.

1166 Morrison, G. S. & Kondaurova, M. V. (2009). Analysis of categorical re-
1167 sponse data: Use logistic regression rather than endpoint-difference scores
1168 or discriminant analysis. *Journal of the Acoustical Society of America*, 126,
1169 2159-2162.

1170 Narayanan, S., Alwan, A. A., & Haker, K. (1995). An articulatory study
1171 of fricative consonants using magnetic resonance imaging. *Journal of the*
1172 *Acoustical Society of America*, 98, 1325-1347.

- 1173 Nawrocki, G. (2008). Laryngeal articulations of /x/ in Southern Polish.
1174 *ZAS Papers in Linguistics*, 49, 145-165.
- 1175 Nissen, S. & Fox, R. (2005). Acoustic and spectral characteristics of young
1176 children's fricative productions: A developmental perspective. *Journal of*
1177 *the Acoustical Society of America*, 118, 2570-2578.
- 1178 Nowak, P. M. (2006). The role of vowel transitions and frication noise in
1179 the perception of Polish sibilants. *Journal of Phonetics*, 34(2), 139-152.
- 1180 Ohala, J. J. (1990). The phonetics and phonology of aspects of assimilation.
1181 In J. Kingston & M. Beckman (Eds.), *Papers in laboratory phonology I:*
1182 *Between the grammar and the physics of speech* (pp. 258-275). Cambridge:
1183 Cambridge University Press.
- 1184 Pape, D. & Jesus, L. (2014). Production and perception of velar stop
1185 (de)voicing in European Portuguese and Italian. *EURASIP Journal on Au-*
1186 *dio, Speech, and Music Processing*, 6, 1-10.
- 1187 Pape, D. & Jesus, L. (2015). Stop and fricative devoicing in European
1188 Portuguese, Italian and German. *Language and Speech*, 58(2), 224-246.
- 1189 Pape, D., Jesus, L., & Birkholz, P. (2015). Intervocalic fricative perception
1190 in European Portuguese: An articulatory synthesis study. *Speech Commu-*
1191 *nication*, 74(C), 93-103. doi:10.1016/j.specom.2015.09.001
- 1192 Proctor, M. I., Shadle, C. H., & Iskarous, K. (2010). Pharyngeal articu-
1193 lation in the production of voiced and voiceless fricatives. *Journal of the*
1194 *Acoustical Society of America*, 127, 1507-1518.
- 1195 Rabiner, L.R. and Schafer, R. W. (2011). *Theory and Applications of Dig-*
1196 *ital Speech Processing*. Pearson, ISBN 9780137050857.
- 1197 Sarlin, M. (2014). *Romanian Grammar*. Books on Demand, ISBN
1198 9789522868985.

- 1199 Shadle, C. (2006). Acoustic phonetics. In K. Brown (Ed.), *Encyclopedia of*
1200 *language and linguistics* (2nd ed., pp. 442-460). Oxford: Elsevier.
- 1201 Shadle, C. (2010). The aerodynamics of speech. In W. J. Hardcastle & J.
1202 Laver (Eds.), *Handbook of phonetic sciences* (2nd ed., pp. 39-80). Black-
1203 well's.
- 1204 Shadle, C. H. & Mair, S. J. (1996). Quantifying spectral characteristics of
1205 fricatives. In *Proceedings of the International Conference on Spoken Lan-*
1206 *guage Processing (ICSLP 96)* (pp. 1517-1520). Philadelphia.
- 1207 Spinu, L. (2010). *Palatalization in Romanian: Experimental and theoret-*
1208 *ical approaches* (Doctoral dissertation). University of Delaware, Newark,
1209 Delaware.
- 1210 Spinu, L., Vogel, I., & Bunnell, T. (2012). Palatalization in Romanian:
1211 Acoustic properties and perception. *Journal of Phonetics*, 40(1), 54-66.
- 1212 Tabain, M. & Watson, C. (1996). Classification of fricatives. *6th Australian*
1213 *International Conference on Speech Science and Technology* (pp. 623-628).
1214 Adelaide, Australia.
- 1215 Todd, A., Edwards, J., & Litovsky, R. (2011). Production of contrast be-
1216 tween sibilant fricatives by children with cochlear implants. *Journal of the*
1217 *Acoustical Society of America*, 130, 3969-3979.
- 1218 Tomiak, G. (1990). *An acoustic and perceptual analysis of the spectral mo-*
1219 *ments invariant with voiceless fricative obstruents* (Doctoral dissertation).
1220 State University of New York, Buffalo.
- 1221 Viterbi, A. J. (1967). Error bounds for convolutional codes and an asymp-
1222 totically optimum decoding algorithm. *IEEE Transactions on Information*
1223 *Theory*, 13(2), 260-269.
- 1224 Wright, R. (2004). A review of perceptual cues and cue robustness. In B.
1225 Hayes, R. Kirchner, & D. Steriade (Eds.), *Phonetically-based phonology*
1226 (pp. 34-57). Cambridge, UK: Cambridge University Press.

- 1227 Yarrington, D., Gray, J., Pennington, C., Bunnell, H. T., Cornaglia, A.,
1228 Lilley, J., Nagao, K., & Polikoff, J. B. (2008). ModelTalker Voice Recorder:
1229 An interface system for recording a corpus of speech for synthesis. *Proceed-*
1230 *ings of the ACL-08: HLT Demo Session* (pp. 28-31). Columbus, OH.
- 1231 Zsiga, E. C. (1997). Features, gestures, and Igbo vowels: An approach to
1232 the phonology/phonetics interface. *Language*, 73(2), 227-274.
- 1233 Zygis, M., Pape, D., & Jesus, L. (2012). (Non)retroflex Slavic affricates
1234 and their motivation: Evidence from Czech and Polish. *Journal of the*
1235 *International Phonetic Association*, 42(3), 281-329.

Figure 1

Figure 2

Figure 3

Figure 4

Figure 5

Figure 6

Figure 7a

Figure 7b

Figure 8

Figure 9

Figure 10a

Figure 10b

Figure 11a

Figure 11b

Figure 12a

Figure 12b

Figure 13