2022 ACTA UNIVERSITATIS CAROLINAE PAG. 83–95
PHILOLOGICA 1 / PHONETICA PRAGENSIA

INTRA- AND INTER-SPEAKER VARIABILITY OF VOWEL SPACE USING THREE DIFFERENT FORMANT EXTRACTION METHODS

ALŽBĚTA HOUZAR AND RADEK SKARNITZL

ABSTRACT

Individual speakers' voices display various unique patterns, one of the most prominent of which is vowel articulation. This study focuses on vowel space properties of 15 Czech speakers in read and spontaneous speech, comparing outputs of three formant extraction methods, measuring formants: (1) in the vowels' temporal midpoints, (2) as their mean from the vowels' middle thirds, and (3) in the vowels' articulatory targets. The results show extensive variability across speakers, but also great within-speaker variability between the two speech styles, with spontaneous speech manifesting more centralised vowel pronunciation than read utterances. The first two measurement methods did not yield systematically different results, while formant values extracted from acoustically defined articulatory targets lead to noticeably larger vowel spaces. The results suggest that care should be taken when interpreting formant values obtained by different methods.

Key words: vowel space area, vowel formants, intra-speaker variability, inter-speaker variability, Czech

1. Introduction

Variability is an inherent characteristic of human speech and in speech science, perhaps the best-known example is illustrations of speakers' vowel systems. While traditional depictions of a vowel system will show discrete points corresponding to individual phonological vocalic qualities, nothing could be further from phonetic reality of everyday speech.

This study deals with vowel formants, i.e., the resonance frequencies of the vocal tract. The lowest two resonances – F1 and F2 – and partly also F3 depend on the vowel quality (i.e., the momentary vocal tract setting), while higher formants tend to remain relatively stable (Reetz & Jongman, 2009: 184). F1–F3 values therefore significantly vary within an individual's speech, but their patterns to some extent reflect their idiosyncrasies and they differ across speakers.

https://doi.org/10.14712/24646830.2022.30

1.1 Formant-based parameters

Vowel formants can be parametrized using several methods. Among the most common ones is extracting formant values in individual vowels, with one vowel token characterized by a single value per formant. Extracting formant values from vocalic segments (tokens) of a voice sample allows for observing their variability in the given vowel (type) as well as across vowels. Vowel formants can be examined individually, but it is also possible to observe multiple formants at once, viewing an analyzed vowel segment as a point in a multi-dimensional space defined by the given formants. For example, by plotting individual vowels in a two-dimensional F1~F2 space, we obtain the speaker's vowel space. Such a plot correlates with the speaker's vocal tract physiology, but also their articulation habits: centralized articulation (hypoarticulation) yields a smaller vowel space, while more distinct vocalic articulation (hyperarticulation) results in an expanded space. A parameter based on vowel space that can be measured is vowel space area (VSA), i.e., its area expressed as the formant measurement unit squared. VSA is usually delimited by formant values of the most peripheral (that is, front/back/open/close) vowels, of which at least three are needed for VSA analysis (Fletcher et al., 2015) but it is possible to include other ones as well (as seen for example in Weirich & Simpson, 2013).

Another example of vowel formant parametrization is long-term formant distributions (LTFs). This metric was introduced by Nolan and Grigoras (2005), and it is determined by each formant's distribution throughout a voice sample regardless of individual vowel qualities. Formant values are extracted from equidistant points within vowel or voiced intervals, i.e., multiple formant values are extracted from one segment. LTFs reflect the dimensions of the speaker's vocal tract as well as their articulation habits such as a tendency towards palatalization or lip rounding (Nolan & Grigoras, 2005). Analogously to individual vowels' formants and vowel space, a multi-dimensional representation of LTFs (LTF1 and LTF2) is possible as well, resulting in vowel space density (VSD; Story & Bunton, 2017). F1 and F2 values extracted at multiple time points throughout a voice sample are plotted in a two-dimensional space defined by F1 and F2, while the points' density constitutes the third dimension. Similarly to vowel space and VSA, these metrics also reflect the speaker's vowel articulation patterns.

1.2 Vowel formant and VSA measurement methods

Formant values characterizing individual vowels may be obtained using several different procedures. The most straightforward one appears to be measuring formant values in the temporal midpoint of each vowel; this method appears to be used in the majority of current studies (e.g., Nolan & Grigoras, 2005; Skarnitzl et al., 2015; Pettinato et al., 2016; Cavalcanti et al., 2021).

Some authors argue that the temporal middle of the vowel may not be the optimal point for formant extraction. Jacewicz et al. (2007) measured formant values at 20% and 35% of the vowels' duration – as the authors state, "[t]hese two measurement locations present the most expanded characterization of the working vowel space", while formant values in later points "tend to portray relatively centralized vowels which would tend to reduce

the vowel space area" (Jacewicz et al., 2007: 1466). Fletcher et al. (2015) also presumed the articulatory target can be reached at a different point than in the vowel's temporal midpoint and examined formant values also in the articulatory target, i.e., "at a time where there was minimal movement in formant tracks – for the best approximation of the vowels' steady-state target" (Fletcher et al., 2015: 2134) between 20% and 80% of the vowel's duration. The results of their experiment indeed show that VSAs based on formant values extracted from the temporal midpoints and articulatory targets significantly differ, the latter being larger, which supports their hypothesis. Measuring vowel formants in the (automatically identified) articulatory targets[2] was also performed for example by Fletcher et al. (2017). In the studies mentioned above, F1 and F2 were measured at the same temporal point; however, as Rose (2015) points out, finding a single articulatory target in the vowel that would be universal for every formant can be problematic, as "the putative target lies at different duration points for each formant" (Rose, 2015: 4822); therefore, finding the target position of each formant separately could also be beneficial. To its advantage, this method, unlike the temporal midpoint (see above) or a mean of multiple values in the mid-section of the vowel (see below), is relatively independent of the placement of the vowel start- and end-points which can be inconsistent among labellers; as Fuchs (2017) points out, speech signal is a continuum, where finding distinct points inherently optimal for extraction of the given values can be problematic (Fuchs, 2017: 11). On the other hand, it can be presumed that an algorithm made to extract the formant's most extreme values might tend to cling to outliers.

The downside of extracting formant values from a single timepoint lies in the possible occurrence of erroneous values; an individual point may not be representative of the whole vowel, as it can be affected by a momentary fluctuation. Therefore, it may be more beneficial to use a mean value of several points within the vowel, excluding its edges that can be influenced by segmental environment. For example, Skarnitzl and Volín (2012) calculated F1 and F2 as an arithmetic mean of seven equidistant points in the middle third of a given vowel, while Tykalová et al. (2021) determined formants' values from "30-ms segment close to the middle section of a vowel where F1 and F2 formant patterns were visible and stable" (Tykalová et al., 2021: 931.e25).

Vowel space area measurements are also affected by the vowel selection which is employed. Fletcher et al. (2015, 2017) extracted F1 and F2 in three most extreme vowels – front [iː], open [ɐː], and back [oː] – in New Zealand English. Using three vowels was also opted for by Pettinato et al. (2016), who measured F1 and F2 of [iː], [ɔː] and [æ] in recordings of Southern British English speakers; as the authors say, those vowels were selected for analysis because "they were the most frequent per individual participant recordings" and "they had the best differentiation in terms of front–back and high–low distinctions and therefore covered the largest distances in the F1–F2 space" (Pettinato et al., 2016: 5). Three-vowel VSA was also analysed by Tykalová et al. (2021), using Czech phonologically short corner vowels [a], [ɪ], and [u] (although the long [iː] has a markedly more peripheral, "corner" quality in Czech). Jacewicz et al. (2007) analysed VSA based on four and five vowel qualities in three regional varieties of American English: [i], [æ], [ɑ]

[2] In this study, the "articulatory target" is defined acoustically as a presumed target of the articulatory gesture derived from vowel formant dynamics, analogically to the studies mentioned above.

and [u], subsequently also adding the diphthong [ɔɪ]. Simpson & Ericsdotter (2007) as well as Weirich & Simpson (2013) measured VSA in German using five vowels, namely [iː], [ɛ], [aː], [ɔ], and [uː].

Studies analyzing vowel formants also differ in their position on the scale between automatic formant extraction and manual measurements. Fully automatic formant extraction (applied, for example, by Weirich & Simpson, 2013 or Pettinato et al., 2016) can be considered unbiased and it is considerably faster; it is, however, prone to errors such as merged or missing formants (Tykalová et al., 2021: 931.e25). Potential errors can be avoided by manually correcting the extracted values (see, e.g., Fletcher et al., 2015 or Tykalová et al., 2021); the drawback of this approach lies in it being relatively time-consuming and, to some degree, subjective, potentially introducing the researcher's confirmation bias into the data.

This study analyzes F1 and F2 values in Czech monophthongs and focuses on vowel space. Its goal is to examine the three methods of vowel formant extraction which were described above. It appears that formant extraction from the middle third of vowel segments is most ecologically valid; the articulatory target method seems prone to extreme values, and extracting from a single temporal point in the middle of vowels increases the likelihood of obtaining erroneous values. We decided to apply all three methods to examine differences between their outputs and, by extension, comparability of studies using different formant extraction methods. Based on the formant values obtained, we will compare the variability of vowel space across speakers and speech styles.

2. Method

2.1 Material

Recordings from 15 Czech male speakers were used for the analysis. The speakers were randomly chosen from the Database of Common Czech (Skarnitzl & Vaňková, 2017) – a reference database for forensic purposes, containing voice samples from 100 male speakers aged between 19 and 50 (mean = 25.6 years, SD = 6.7 years), who performed several speaking tasks, representing different speech styles. For this study, recordings of two speech styles were used: (1) reading a phonetically rich text of 150 words (corresponding to reading time around 1 minute) and (2) a one-minute excerpt from a spontaneous interview where the speakers were encouraged to talk on a topic of their own choice. The recordings were obtained in quiet environments in the speakers' home or workplace (subtle acoustic discrepancy among individual speakers' recordings thus cannot be ruled out) in a WAV format with 48-kHz sampling frequency, using a professional portable recorder Edirol HR-09.

The recordings were automatically segmented using the Prague Labeller (Pollák et al., 2007); afterwards, phone boundaries were manually corrected in Praat (Boersma & Weenink, 2015), following segmentation principles described in Machač and Skarnitzl (2009).

The Czech phonemic inventory contains 10 monophthongs and 3 diphthongs: /iː ɪ ɛ ɛː a aː o oː u uː/ and /a͡ʊ̯ o͡ʊ̯ ɛ͡u̯/. In our analysis, only monophthongs were used. Since short

and long vowels' realizations generally do not significantly differ in their quality, the short vowels and their long counterparts were merged into single categories, with the exception of /ɪ/ and /iː/ (see Skarnitzl and Volín, 2012 or Šimáčková et al., 2012 for more details on the Czech vowel inventory); therefore, these two phonemes were treated as separate vowel qualities in the analyses below.

2.2 Extraction of formant values

F1 and F2 values in Hz were automatically extracted from each vowel token using three approaches:

- in a single timepoint in the middle of the vowel duration;
- as the mean value in the middle third of the vowel;
- from articulatory targets that were automatically detected based on the formants' shifts inside the vowel between 20% and 80% of the vowel's duration, the onset and offset 20% being excluded to eliminate potential interference of segmental environment. In front vowels /ɪ iː ɛ ɛː/, the articulatory target was identified as the point of F2 peak, in back vowels /u uː o oː/ as the point of F2 minimum, and in the open vowels /a aː/ as the point of F1 peak (analogously to Fletcher et al., 2015). Both F1 and F2 were measured at the described timepoints (i.e., F1 and F2 values for each vowel were extracted from the same timepoint).

This study's methodology represents a synthesis of formerly used procedures (described above), and its objective is to compare their output. The most prevalent of the examined methods appears to be formant extraction from the vowel midpoint which has been used by a variety of studies (see section 1.2). Extraction of mean formant values from the middle third of a vowel was employed by Skarnitzl and Volín (2012). The last method we analysed, i.e., extraction of formant values from the (acoustic) articulatory target, was based on the methodology described in Fletcher et al., 2015, who explain the procedure like this:

> Articulatory point measurement criteria were designed with the aim of extracting values at a time where there was minimal movement in formant tracks – for the best approximation of the vowels' steady-state target. For the front vowel, [iː], this point was set at peak F2 frequency; for the open [ɐː] vowel the target was extracted when F1 was at its maximum; and for the back [oː] vowel the target point was taken when the lowest value of F2 was reached (Watson and Harrington, 1999; Watson et al., 1998). (*Fletcher et al., 2015: 2134*)

All formant extractions were performed using the Burg LPC algorithm in Praat in three different settings. The default setting contained the detection of 5 formants between 0 and 5,500 Hz (i.e., 500 Hz more than a five-formant default range for an adult male – this expanded range was applied to avoid potential misses of higher formant values in front vowels). The secondary setting, on the other hand, used a reduced frequency range, detecting 5 formants within 0–3,000 Hz, in order to resolve potential errors of the first setting which tended to merge F1 and F2 in back vowels into one detected formant. Lastly, a tertiary setting was also present, extracting 10 formant values in 0–3,000 Hz band, in case neither one of the previous settings yielded accurate results.

All the extracted formant values were manually checked and corrected if necessary; in cases where the values obtained by the default setting prominently diverged from standard formant values for the given vowel quality, the spectrograms were both visually and auditorily inspected and when appropriate, the default values were replaced by those obtained using the secondary or tertiary settings (when those values included detection of random noise as formants due to the reduced frequency range, they were excluded based on the spectrogram visual inspection). Those abnormal values included F1 below 200 Hz or above 800 Hz, F2 below 600 Hz and above 1,500 Hz in back vowels, and F2 below 1,000 Hz and above 2,300 Hz in front vowels. Solely the strongly significant abnormalities in the default formant extraction output were manually corrected in order to avoid introducing confirmation bias into the data.

The output of this procedure was F1 and F2 values from 15 speakers, 2 speech styles, 6 vowel qualities and 3 extraction areas (temporal midpoint, middle third and articulatory target); in total, the final dataset contains F1 and F2 values of 24,646 vowels. The formant values were converted to the Bark frequency scale which is more psychoacoustically relevant compared to Hz.

2.3 Analysis

The vowel space area (VSA) was calculated for each speaker, each speech style, and each formant extraction method as the surface area in a two-dimensional F1~F2 space delimited by formant value medians of the individual vowel qualities, using the formula:

$$VSA = \left| \frac{(x_1 y_2 - y_1 x_2) + (x_2 y_3 - y_2 x_3) \ldots + (x_6 y_1 - y_6 x_1)}{2} \right|$$

with x being median F1, y being median F2 and numbers 1-6 corresponding to individual vowel qualities in the order [iː ɪ ɛ a o u]. VSA will be expressed in Bark squared (Bark^2).

Vowel space illustrations were prepared in R (R Core Team, 2021) and the *ggplot2* package (Wickham, 2016).

3. Results and discussion

The general results are depicted in Figure 1, which shows all the speakers' F1~F2 vowel space area in read and spontaneous speech, as extracted by the three methods described in section 2.2: using the vowels' temporal midpoint, the mean from the middle third of the vowel, and the articulatory target.

First, it is clear that speaking style affects VSA to a great extent: in most speakers, VSA in read speech (shown in circles in Fig. 1) is larger than in spontaneous speech (triangles); the difference is particularly salient in speakers HROK and especially KALT. Five speakers manifest an opposite tendency in at least one extraction method; only speaker NVAT's vowel space area turned out to be larger in spontaneous speech using all three extraction methods.

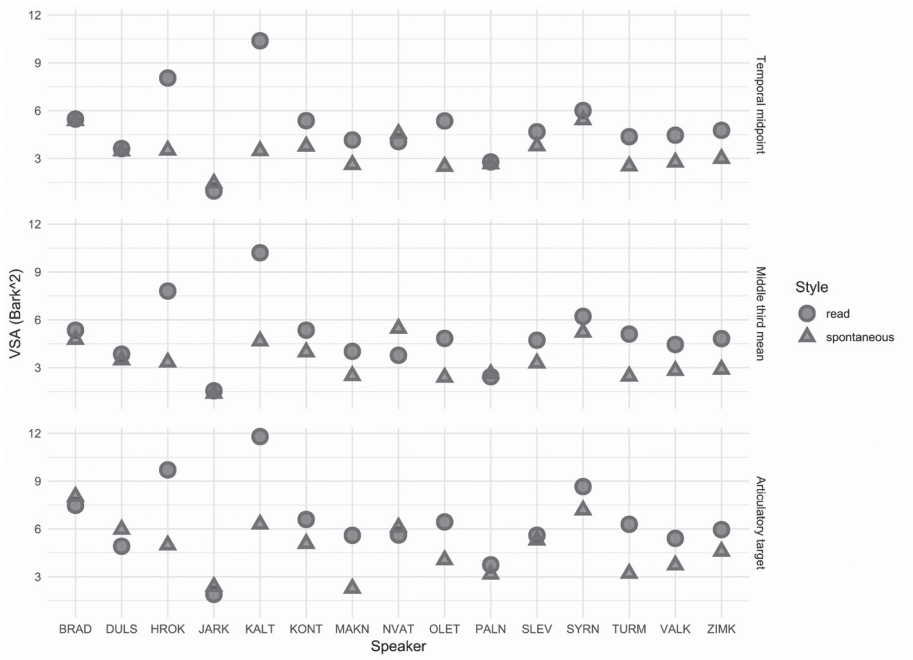

Figure 1. Vowel space area (VSA) of individual speakers in read and spontaneous speech, using three methods of formant extraction.

Second, although most speakers' VSAs fall between approximately 3 and 6 Bark^2, Figure 1 suggests that there are some between-speaker differences. Speaker JARK's vowel space area is strikingly small (see also below), and consistently so across the three measurement methods and across the two speaking styles. Indeed, his speech does sound remarkably centralized.

Third, the extraction methods themselves yield different results (more details follow below). It is not surprising that, almost without exception, the largest VSA is obtained by the articulatory target method, shown in the bottom panel of Figure 1 (*cf.* section 1.2.). The temporal-midpoint and middle-third-mean methods (see the top and middle panel, respectively) tend to yield comparable VSA values.

It is instrumental to examine not only the vowel space area, but to focus in more detail on selected vowel spaces. That will allow us to compare the extraction methods in a better way. Figure 2 shows vowel spaces, as extracted by the three methods, for four speakers who manifested some noteworthy tendencies or who may be regarded as representing more speakers with similar patterns. The plots, along with the VSA values, confirm what has been written above, namely that vowel spaces are considerably larger when formants are extracted from the articulatory targets. When we compare vowel spaces in read and spontaneous speech, it is clear that the shifts which underlie the overall reduction of vowel space in spontaneous speech are not identical in the four depicted speakers. It is only in speaker HROK (and similarly also in speaker OLET, not shown in

the figure) that all vowels except the long [iː] are centralized when compared with read speech. Most speakers realized the Czech back vowels – [u uː o oː] – with a higher F2 value, which may, in articulation terms, correspond to centralization and/or weaker or absent lip rounding. However, the close back vowels [u uː] appear to be pronounced in a more peripheral manner in spontaneous speech by speaker JARK, whose vowel space is otherwise extremely small, and also by speakers BRAD, and KALT and NVAT (not shown in Fig. 2). Most speakers also produce more open [a aː] vowels in read speech (in addition to BRAD and HROK in Fig. 2, this applies to another five speakers).

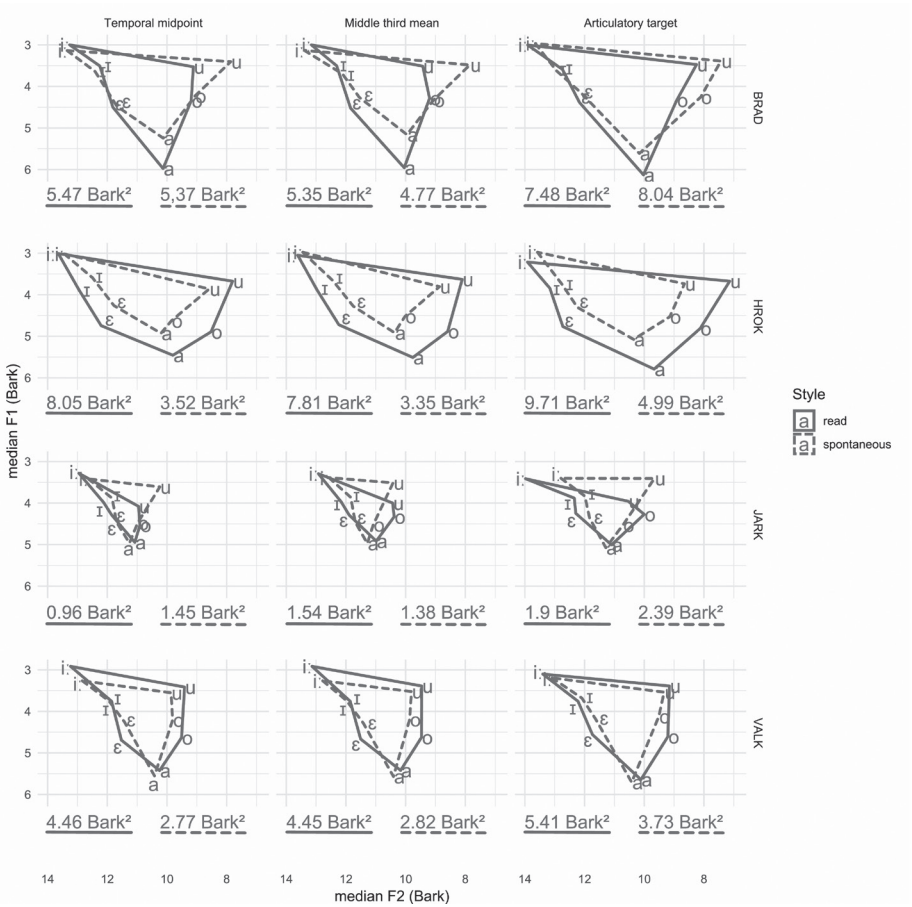

Figure 2. Vowel space of four speakers in read and spontaneous speech, using three methods of formant extraction. The points correspond to medians of F1 and F2. VSA values are provided below each plot.

It is to be expected that median values conceal considerable variability in the data. In Figure 3, we therefore provide another look at the vowel spaces of the four selected speakers in read and spontaneous speech. For the sake of easier comparison, only one extraction method – data based on the middle third mean of each vowel – is shown (as

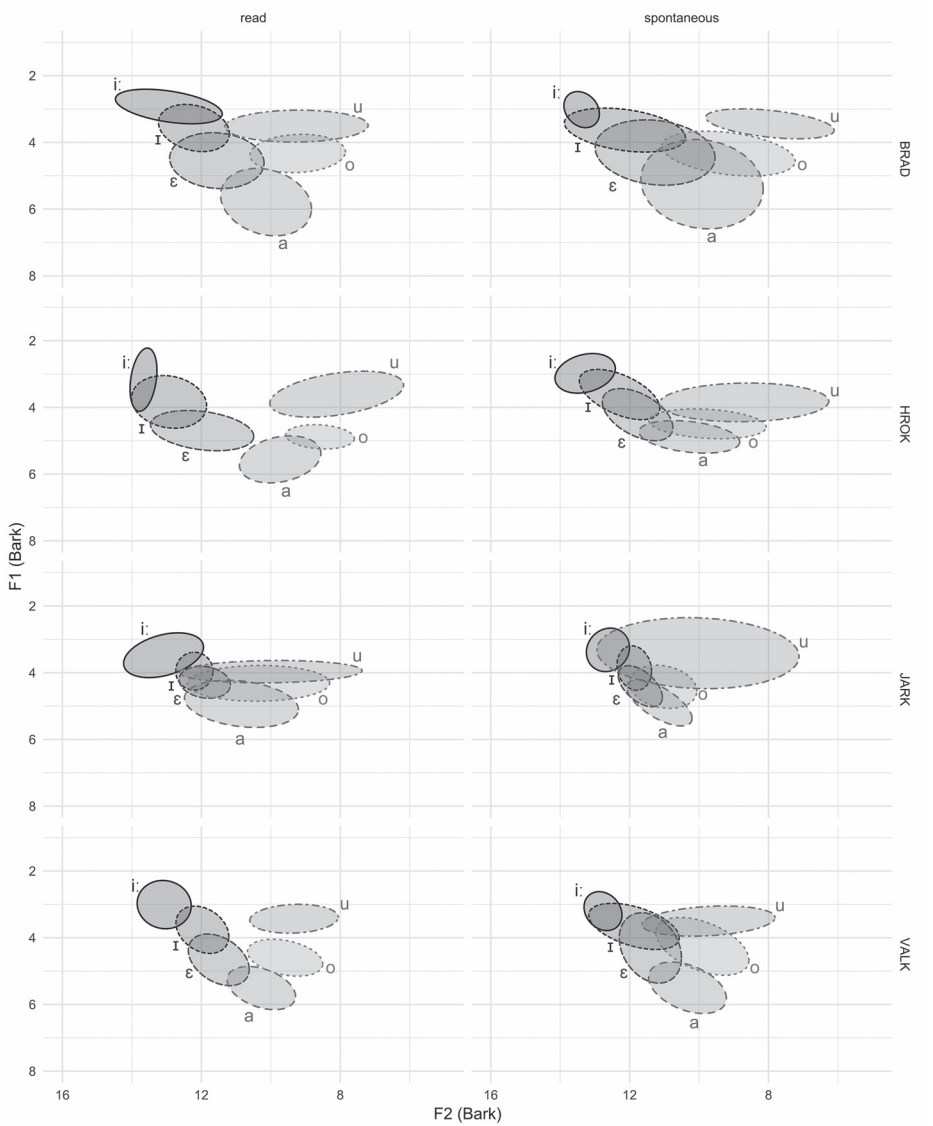

Figure 3. *Vowel spac*e of four speakers in read and spontaneous speech, using the middle third mean method of formant extraction. The ellipses correspond to 68% of the data.

formant extraction from the middle third of a vowel appears to be the most ecologically valid method; see section 1.2).

It is not surprising that there is less overlap between the distributions of formants of individual vowels in read speech. In addition, the overall display of vowel quality distribution in Figure 3 reveals an interesting detail, namely the huge variability of the [u uː] vowels, especially in F2. While the analyses of Skarnitzl and Volín (2012)

indicated a possible change in Common Czech, with the short [u] becoming slightly more centralized than the long [u:], a closer analysis of our data indicates considerable inter- and intra-speaker variability, as shown in Fig. 4. One can see that in most of our speakers the long [u:] is, on average, more peripheral (i.e., has lower F1 and F2 values),

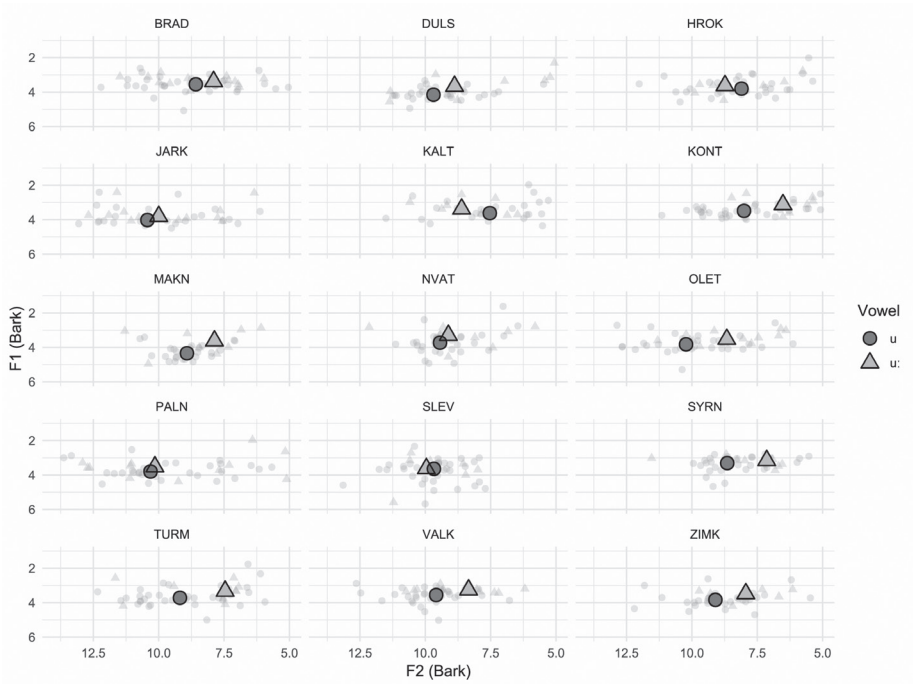

Figure 4. [u] and [u:] F1 and F2 values in individual speakers; the smaller and less opaque points represent single realisations, while the bolder points reflect median values.

but there are several exceptions. Along the horizontal (F2) axis, both vowels manifest considerable variability; indeed, auditory inspection of [u u:] confirmed salient fronting in some tokens.

4. General discussion and conclusion

In our study, we examined variability of vowel space across speakers, in two different speaking styles within a speaker, and using three formant extraction methods. The vowel space characteristics we observed were VSA, i.e., the size of the vowel space area, and distribution of vowel realizations within the acoustic vowel space.

Our results indicate considerable between-speaker differences in vowel space, but also great within-speaker variability between the two speaking styles. We may conclude that vowel space characteristics would not generally be able to differentiate between individual speakers. However, specific speakers may manifest interesting idiosyncratic tendencies which are stable across different conditions and, in comparison with the

comparable population, quite atypical. Speaker JARK in our dataset may serve as an example: as shown in Figure 2, his VSA is markedly smaller compared to other speakers in both speech styles.

Regarding differences between speech styles, speakers' vowel space area tends to be larger in read utterances, reflecting a more distinct, less centralized pronunciation of vowels than in spontaneous speech. Inside the vowel space, there is also an apparent smaller dispersion of values in read speech compared to spontaneous speech, suggesting more consistent articulation of individual vowel qualities (see Figure 3). It must be emphasized that these "divergent" realizations matched perception; in other words, they are not due to faulty formant extraction.

As for the three methods of formant extraction, we calculated formants as the values from the temporal midpoint of a vowel, their mean from the vowel's middle third, and as a single point from the articulatory target defined as the stage where the given formant reached its maximum or minimum. The extraction from the temporal midpoint of vowels represents the most frequent procedure reported in literature (see section 1.2), but it could be argued that a formant value taken from a single time point might frequently correspond to an outlier. This risk can be avoided by taking into account multiple formant values within a vowel, excluding its edges where formants are influenced by the flanking segments, and calculating the mean of those values. However, the two extraction methods did not yield systematically different formant values and VSAs in this study, although the results are certainly not identical (see Figures 1 and 2).

On the other hand, extraction of formants from the vowels' articulatory targets did result in noticeably more extreme values and thus also larger VSAs – at least when considered visually (a quantitative analysis was not an objective of this study). This conclusion is in accordance with Fletcher et al.'s (2015) hypothesis that the articulatory target is not necessarily located in the middle of the vowel's duration. However, it can also be possible that the algorithm set to identify the highest/lowest formant values tends to pick outliers occurring due to faulty formant extraction. Also, the suitability of this extraction method for spontaneous speech can be considered questionable, because – as mentioned above – individuals vowels' pronunciation in reality often corresponds to what appears to be phonetically very different vowel qualities; for example, identifying the articulatory target of an /u/ realization at the F2 minimum can be problematic when the segment's actual pronunciation is closer to a much fronter [y]. This method's validity needs to be further examined; it could be beneficial to identify the placement of identified articulatory targets within vowels and observe whether it is consistent across vowels, and to see if it matches Jacewicz et al.'s (2007) premises (*cf.* section 1.2) or whether its location appears to be random, suggesting the tendency of the algorithm to cling to outliers. In case it shows consistent patterns, it could also be useful to investigate whether the location of the target differs in individual formants, as mentioned by Rose (2015). Moreover, in future research, it could be interesting to focus on LTFs and vowel space density analysis and examine how its outputs correspond to the results of this study.

To conclude, this study has shown that caution should be taken when comparing results of different studies which analyze vowel formants: different extraction methods may provide rather diverging results, and interpretation may thus be less straightforward than it appears.

Acknowledgements

The work was supported by the grant SVV 2020 – 260555 realized at the Charles University, Faculty of Arts, and by the European Regional Development Fund-Project "Creativity and Adaptability as Conditions of the Success of Europe in an Interrelated World" (No. CZ.02.1.01/0.0/0.0/16_019/0000734). We would also like to thank Tomáš Bořil for his assistance with data extraction.

REFERENCES

Boersma, P., & Weenink, D. (2015). *Praat: doing phonetics by computer (Version 6.0)*. Retrieved from http://www.praat.org

Cavalcanti, J. C., Eriksson, A., & Barbosa, P. A. (2021). Acoustic analysis of vowel formant frequencies in genetically-related and non-genetically related speakers with implications for forensic speaker comparison. *Plos One, 16*(2), e0246645.

Fletcher, A. R., McAuliffe, M. J., Lansford, K. L., & Liss, J. M. (2015). The relationship between speech segment duration and vowel centralization in a group of older speakers. *Journal of the Acoustical Society of America, 138*(4), 2132–2139.

Fletcher, A. R., McAuliffe, M. J., Lansford, K. L., & Liss, J. M. (2017). Assessing vowel centralization in dysarthria: A comparison of methods. *Journal of Speech, Language, and Hearing Research, 60*(2), 341–354.

Fuchs, S. (2017). Changes and challenges in explaining speech variation: A brief review. Available at: https://www.researchgate.net/publication/320991961_Changes_and_challenges_in_explaining_speech_variation_A_brief_review.

Jacewicz, E., Fox, R. A., & Salmons, J. (2007). Vowel space areas across dialects and gender. In *Proceedings of the 16th ICPhS*, 1465–1468.

Machač, P., & Skarnitzl, R. (2009). *Principles of phonetic segmentation.* Epocha.

Nolan, F., & Grigoras, C. (2005). A case for formant analysis in forensic speaker identification. *International Journal of Speech, Language and the Law, 12*(2), 143–173.

Pettinato, M., Tuomainen, O., Granlund, S., & Hazan, V. (2016). Vowel space area in later childhood and adolescence: Effects of age, sex and ease of communication. *Journal of Phonetics, 54*, 1–14.

Pollák, P., Volín, J., & Skarnitzl, R. (2007). HMM-Based Phonetic Segmentation in Praat Environment. *Proceedings of SPECOM 2007*, 537–541. MSLU.

R Core Team (2021). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. Available at: https://www.R-project.org/.

Reetz, H., & Jongman, A. (2009). *Phonetics: Transcription, production, acoustics, and perception*. Blackwell.

Rose, P. (2015). Forensic voice comparison with monophthongal formant trajectories – a likelihood ratio-based discrimination of "schwa" vowel acoustics in a close social group of young Australian females. *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4819–4823.

Simpson, A., & Ericsdotter, C. (2007). Sex-specific differences in f0 and vowel space. In *Proceedings of the 16th ICPhS*, 933–936.

Skarnitzl, R., Vaňková, J., & Bořil, T. (2015). Optimizing the extraction of vowel formants. In: Niebuhr, O. & Skarnitzl, R. (Eds.), *Tackling the complexity in speech*, 165–182. Charles University, Faculty of Arts.

Skarnitzl, R., & Vaňková, J. (2017). Fundamental frequency statistics for male speakers of Common Czech. *Acta Universitatis Carolinae – Philologica*, 3, 7–17.

Skarnitzl, R., & Volín, J. (2012). Referenční hodnoty vokalických formantů pro mladé dospělé mluvčí standardní češtiny. *Akustické listy, 18*, 7–11.

Story, B. H., & Bunton, K. (2017). Vowel space density as an indicator of speech performance. *Journal of the Acoustical Society of America, 141*(5), EL458–EL464.

Šimáčková, Š., Podlipský, V. J., & Chládková, K. (2012). Czech spoken in Bohemia and Moravia. *Journal of the International Phonetic Association, 42*(2), 225–232.

Tykalová, T., Škrabal, D., Bořil, T., Čmejla, R., Volín, J., & Rusz, J. (2021). Effect of Ageing on Acoustic Characteristics of Voice Pitch and Formants in Czech Vowels. *Journal of Voice, 35*(6), 931.e21–931.e33.
Weirich, M., & Simpson, A. (2013). Investigating the relationship between average speaker fundamental frequency and acoustic vowel space size. *Journal of the Acoustical Society of America, 134*(4), 2965–2974.
Wickham H (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. Available at: https://ggplot2.tidyverse.org/.

Alžběta Houzar
Institute of Phonetics
Faculty of Arts, Charles University
Prague, Czech Republic
E-mail: alzbeta.houzar@ff.cuni.cz

Radek Skarnitzl
Institute of Phonetics
Faculty of Arts, Charles University
Prague, Czech Republic
E-mail: radek.skarnitzl@ff.cuni.cz