Analysis of Polish Nasalized Vowels Based on Spatial Energy Distribution and Formant Frequency Measurement

Anita Lorenc[1], Katarzyna Klessa[2(✉)], Daniel Król[3], and Łukasz Mik[3]

[1] Institute of Applied Polish Studies, University of Warsaw, Krakowskie Przedmieście 26/28, 00-927 Warszawa, Poland
anita.lorenc@uw.edu.pl
[2] Institute of Applied Linguistics, Adam Mickiewicz University in Poznań, Niepodleg łości 4, 61-874 Poznań, Poland
klessa@amu.edu.pl
[3] University of Applied Sciences in Tarnów, Mickiewicza 8, 33-100 Tarnów, Poland
dankrol@gmail.com, l_mik@pwsztar.edu.pl

Abstract. In this paper, we discuss the results of the analysis of F1 and F2 frequency measurements in Polish nasalized vowels represented in writing by the graphemes *ę* and *ą* (realized before voiceless fricatives). The speech material included recordings of isolated word items provided by 20 adult native speakers of Polish (10 females and 10 males). According to the claims often presented in phonetic studies, the two vowels are phonetically realized as diphthongs composed of two subsequent stages of realization: an oral and a nasal stage. In our investigation, we refer to the results obtained by Lorenc et al. (cf. [13] and [14]) based on the analyses of spatial distribution of the acoustic field which indicate that the structure might be even more complex in certain cases and include three or even more stages. We measure formant frequencies within these stages using the stage timestamps obtained with a novel infrastructure composed of a multi-channel recorder with a circular microphone array. Among others, the results indicate that the two vowels differ significantly with regard to their internal structures as expressed by the number and types of the stages as well as frequency formant characteristics of those stages.

Keywords: Speech analysis · Acoustic camera · Formant frequency · Acoustic field energy distribution · Nasalized vowels

1 Introduction

The Polish vowel system is often referred to as relatively simple in terms of description and usage, as it consists of (only) six oral vowels [a, e, i, o, u, ɨ]

Research described in this paper was supported by grant no. 2012/05/E/HS2/03770 financed by The Polish National Science Centre (decision no. DEC-2012/05/E/HS2/03770).

© Springer Nature Switzerland AG 2020
Z. Vetulani et al. (Eds.): LTC 2017, LNAI 12598, pp. 186–196, 2020.
https://doi.org/10.1007/978-3-030-66527-2_14

[7]. However, some difficulties remain, and one of them is certainly the widely discussed problem of the status of two so-called nasalized (or nasal) vowels, denoted in writing by the graphemes *ę* and *ą* and phonotactically constrained to positions before fricative consonants (e.g., [15]). The questions are posed not only from the point of view of fundamental research, but also in the context of speech and language technology. One of the questions concerns the internal structure of the sounds resulting from their specific manner of articulation. According to many empirical studies, the nasalized vowels are produced asynchronously with (at least) two subsequent stages of realization (e.g., [3,16,20]). The first stage is often referred to as (prevalently) an oral one, while for the second one, an important influence of a nasal resonance is observed. Although a possibility of 3-stage realizations was mentioned earlier (e.g., by Wierzchowska [20]), they were usually described as oral or oral-nasal ones. The oral-nasal realizations were described as ones where the presence of nasality was increasing throughout the vowel, however, it was always preceded by the initial oral resonance.

The two-segment approach has been reflected in the publications dedicated to the description of the Polish phonemic inventory, as well as some of the works on automatizing experimental procedures in phonetics and technical solutions. For example, Steffen-Batogowa [18] supported the idea of transcribing the sounds with [ẽũ] and [õũ] respectively in her work on automation of grapheme-to-phoneme conversion rules for Polish. In the *Illustration of the IPA: Polish* by W. Jassem [7], the realizations of the sounds represented by the ę and ą graphemes are treated as sequences of two distinct phones, i.e. an oral vowel [e] or [o] followed by a nasalized component (an approximant or a nasal consonant depending on the context within the utterance).

On the other hand, the standard version of SAMPA (computer-readable alphabet, [19]) for Polish includes [e∼], [o∼] transcription labels for the two sounds, representing the the graphemes *ę* and *ą*, respectively. Therefore, in SAMPA the Polish vowel inventory includes eight vowels as not only does it include the oral vowels [a, e, i, o, u, ɨ] mentioned above but also the two nasalized sounds.

Following the postulates by Steffen-Batogowa [18], extended variants of the SAMPA alphabet were proposed and tested in studies in the context of speech technology or in corpus annotation (e.g., [10]). They included e.g., the two-symbol [ew∼], [ow∼] representations instead of [e∼], [o∼], as well as separate labels for the nasalized approximants [w∼] [j∼]. The realizations of *ę* and *ą* were thus treated as sequences of subsequent phones. From the practical point of view, however, it appears (e.g., to be very difficu)lt to define the position of boundaries between the oral and nasal segment (the "boundary" is actually a continuous transient and any segmentation applied for the needs of unit selection resulted in glitches and disfluencies in synthesized speech). Consequently, better speech synthesis results could be obtained by avoiding the segmentation into stages and treating the oral and nasalized sounds as the components of one inherently diversified vowel segment (corresponding to the graphemic representation). Similar, single-segment approach was assumed for speech recognition tasks (e.g., [21]).

The normative approaches to the problem indicate a lack of consensus, even in terms of defining the basic pronunciation standards or guidelines. For example, regardless of the results reported in the domains of acoustic or instrumental phonetic research, the dictionary of standard Polish pronunciation [8], recommends a synchronic and monophthong pronunciation of the vowels in front of fricatives and in the prepausal position. Two-segment, diphthong-like realizations are referred to as surprising and not common enough to be treated as normative.

In the present work, the realizations of the nasalized vowels are investigated with the use of data obtained from an innovative acoustic camera infrastructure developed by Król et al. [11] and by means of F1-F2 formant frequency analysis. While the results confirm that the two-stage realization is the most typical (although not exclusive) for the front vowel [ewˉ], the back vowel [owˉ] appears to be more diversified in terms of both articulatory features and formant frequency values. As far as the notation is concerned, we follow the two-segment approach and transcribe the front vowel *ę* with [ewˉ] and the back vowel *ą* with [owˉ]. It should be noted, however, that in the light of our findings these transcription labels might also become a matter of discussion especially with regard to the back vowel.

The structure of the paper is as follows: Sect. 2 of the paper provides information about the study material as well as the applied methods and tools. In Sect. 3, the results of the F1, F2 measurements are discussed. Section 4 includes conclusions and a concise outline of the future work.

2 Data, Methods and Tools

2.1 Speakers and Speech Material

Speech material was provided by 20 adult native speakers of Polish (10 females and 10 males) aged from 22 to 46. The speakers were selected from a larger group of candidates and at the preliminary stage, their pronunciation was carefully evaluated by a team of experts (phoneticians and speech therapists). All the speakers used contemporary standard Polish, declared having university education, and represented nine (out of sixteen) Polish voivodeships. The recordings are one of the outcomes of a larger project described by Lorenc [13].

The material selected for the present study consists of 161 wave files (16 bit PCM, 96 kHz) including recordings of isolated two-syllable words, treated as containers for the target nasalized vowels (76 realizations of [ew] and 85 realizations of [ow]). The target vowels were always located in the initial stressed syllables of the container words, before a voiceless fricative consonant [s]. The preceding context was either the voiceless plosive [p] or voiced fricative [v] consonant. The words were meaningful Polish words, such as for instance: węzeł [vɛw̃zɛw] (En. 'knot') or *pąsy* [pɔw̃sɨ] (En. 'blushes'). The word recordings were manually segmented into phones using Praat [1].

2.2 Vowel Stages and Stage Boundary Positions

Vowel stage boundary positions were established based on spatial acoustic field distribution in the function of time obtained from acoustic camera (for details see: [11,14]). The method of generating the information is illustrated in Fig. 1.

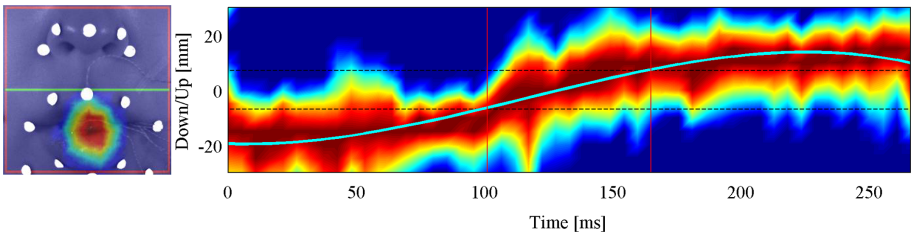

Fig. 1. The method of generating time-aligned spatial distribution of acoustic field.

A 60 × 60 mm fragment of the acoustic camera image was chosen for the needs of the present experiment. The fragment shows the area of the speaker's mouth and nose. The division point was set to a sensor located directly above the upper lip (point 0 at the vertical Down/Up axis in Fig. 1). The information about the sensor position enables constant adjustments to the movements of the speaker's head and consequently, stabilization of the selected area. We applied a 3rd polynomial approximation of maximum pressure, which made it possible to eliminate minor signal fluctuations (specific mainly to the stages where both oral and nasal resonances were active), and this way to obtain a clear image of the pressure changes over time (Fig. 2).

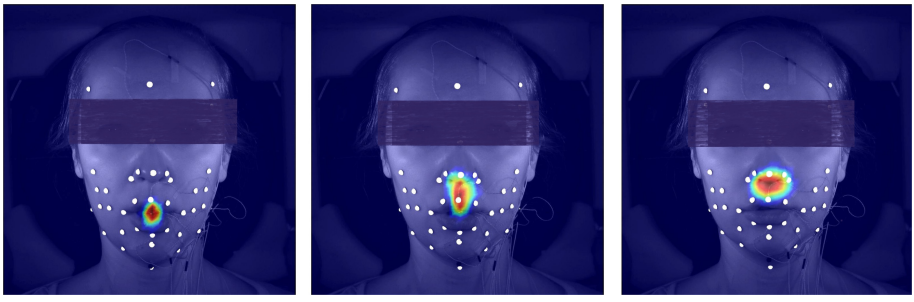

Fig. 2. Generating time-aligned spatial distribution of acoustic field. Example illustrations of three kinds of resonances: oral resonance (left), oronasal (middle), and nasal resonance (right).

For the purpose of the source of emission, a 3 dB acoustic pressure drop threshold was assumed. Three areas of acoustic field distribution were defined:

oral (range: −30 mm to −15 mm), oral-nasal (15 mm to +15 mm) and nasal (+15 mm do +30 mm). The areas correspond to three possible stages of vowel realization with either oral, both oral and nasal (oronasal), or nasal resonance, respectively. In Fig. 1, the areas are indicated by the black dotted lines. The information about the timestamps and durations of particular stages (red vertical markers in Fig. 1) was generated based on the cross points of the approximation polynomial line with the lines denoting spatial boundaries of particular phases.

2.3 Formant Measurement and Data Processing

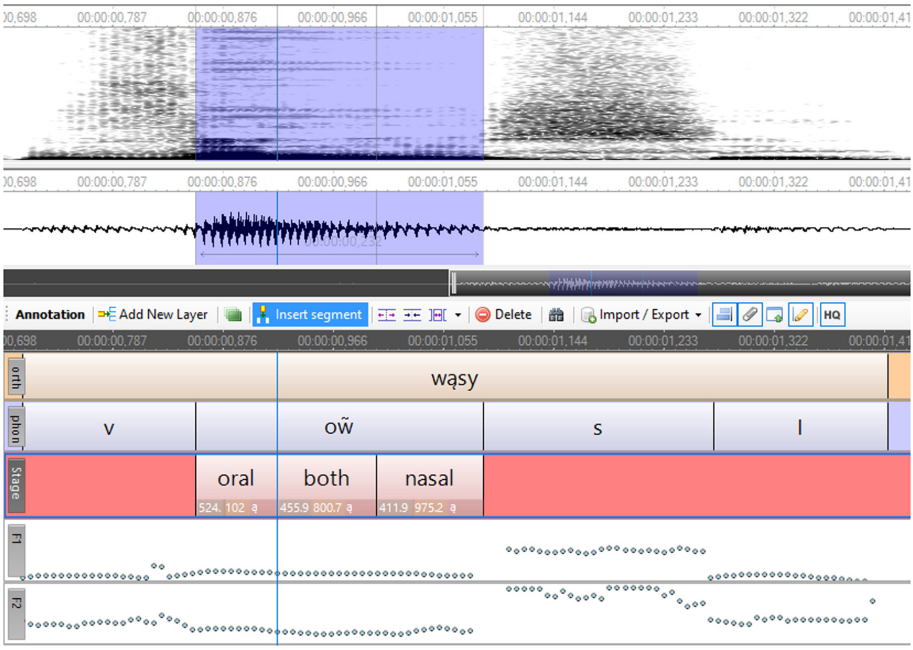

Fig. 3. Annotation Pro: an example view of time-aligned segmentation data, formant measurement results and three realization stages (oral, both, nasal) for the vowel [ow̃] in the word *wąsy*.

Formant frequencies were measured in Praat using the Formant listing option. Further processing was performed with Annotation Pro [9]. The formant listing files were automatically imported to annotation files (.ANT format, Annotation Pro) and synchronized with the time-aligned transcriptions and the vowel stage boundary timestamps (Sect. 2.2). This way, the data originally coming from the acoustic field energy distribution analysis and formant values were saved within the same workspace. An example view of the imported data is shown in Fig. 3 (formant values are represented by the dotted lines, the software enables also numerical display).

The Praat formant listing files provided information about the frequencies of the formants F1, F2, F3, F4 for the whole utterances (container words). We then used a C# plugin scripts for Annotation Pro [9] to link the obtained formant frequency values to the respective vowel stages, and also to calculate mean formant values per stage. Subsequent statistical analysis using both the measurement results and the mean values was carried out with Statistica software package [17]. In this contribution, we focus on the analysis of F1 and F2 frequency values within the vowel realization stages.

3 Results

3.1 The Number of Vowel Stages

According to the results generated based on the spatial acoustic field distribution in the function of time, the majority of the realizations of the vowel [ew̃] were determined as 2-stage ones. Altogether, 57 out of 76 instances of [ew̃] were produced with two subsequent stages: an oral stage and a stage where both oral and nasal resonances were detected. 14 instances were produced with the use of just one, oral resonance. The remaining 5 vowels were identified as: 3-stage (four occurrences), 4-stage (one occurrence) or 5-stage (one occurrence) realizations.

In the case of the vowel [ow̃], out of the total 85 instances, only 22 were produced as 2-stage realizations (again, in the 2-stage variant, one of the stages was always produced with an oral resonance, and the second one with both oral and nasal resonances active). 25 realizations were identified as 3-stage (the third stage was based on a nasal resonance without any oral component). Further 27 vowels were realized using a 4-stage structure (based on alternately activating the three types of resonance, usually either the oral or oronasal stage occurred for the second time). In 9 cases five stages were detected, and only two realizations were produced with a single type of resonance, i.e. with a 1-stage structure.

3.2 Formant Analysis

The Nasalized Vowel [ew̃]. The mean values of the first two formant frequencies for the nasalized front vowel [ew̃] are presented in Table 1.

Only the data for the 1-stage and 2-stage realizations have been included in the Table because these two types of structure represent the vast majority of cases in the present material. The formant frequency values observed for these two stages confirm the distinction in case of 2-stage realizations. The differences between means are statistically significant according to ANOVA ($p \leq 0.001$).

The frequency values in 1-stage realizations are close to Polish non-nasalized [e] reported by Jassem [6].

Figure 4 displays scatter-plots of the mean formant frequency values in the F2-F1 space for the 2-stage realizations of [ew̃] (inverted axes were used as in typical vowel charts). Although the values obtained for the subsequent stages overlap to a certain extent, two different areas can still be distinguished.

Table 1. Mean F1 and F2 values in 1-stage and 2-stage realizations of the vowel [ew̃] (F- female, M-male).

Structure	Resonance type	Speaker sex	Mean F1 [Hz]	Mean F2 [Hz]
1-stage	Oral	F	635	1821
		M	475	1486
2-stage	Oral	F	672	1768
		M	494	1490
	Both	F	435	1673
		M	365	1459

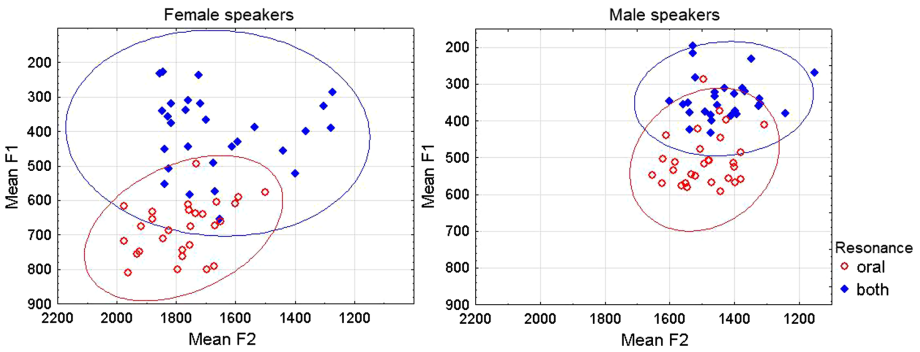

Fig. 4. Scatter-plots of the mean formant values in the F2-F1 space for 2-stage [ew̃] (left: female speakers, right: male speakers).

In 2-stage realizations of [ew̃], the stage based on both oral and nasal resonances was characterized by lower F1 and F2 median (and mean) values when compared to the stage based on exclusively oral resonance, which is in line with the results of earlier studies (see Fig. 5), assuming a decrease in the openness and frontness of the vowel in the course of its realization.

As expected, formant values for men are lower on average than those obtained for women. When looking at the differences between formants for male and female voices, similar tendencies can be observed with regard to the differences between stages, as well as slightly smaller dispersion around the middle value for male voices.

The Nasalized Vowel [ow̃]. Table 2 provides mean formant frequency values obtained for 2-, 3-, and 4-stage realizations of [ow̃] (i.e. the most frequent types of structures detected for this vowel) (Fig. 6).

The majority of multiple-stage realizations of [ow̃] began with the oral resonance, however, in certain cases at the beginning, both oral and nasal cavities were active. Table 2 includes data for all the observed stages with regard to resonance types for [ow̃] but it does not fully account for the order of appearance of these stages inside the vowel as it varied across speakers. The labels 'both2'

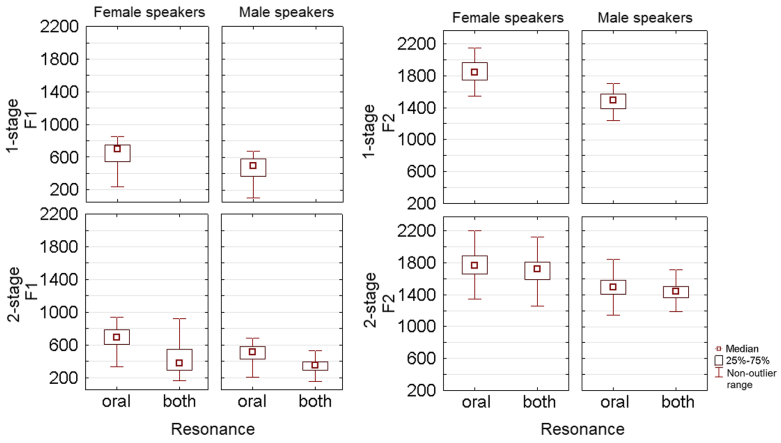

Fig. 5. Median values of F1 (left) and F2 (right) in 1-stage and 2-stage realizations of [ewˆ] by male and female speakers.

Table 2. Mean F1 and F2 values in 2-stage, 3-stage and 4-stage realizations of the vowel [owˆ] (F- female, M-male).

Structure	Resonance type	Speaker sex	Mean F1 [Hz]	Mean F2 [Hz]
2-stage	Oral	F	644	1206
		M	558	1008
	Both	F	542	1404
		M	497	1161
3-stage	Oral	F	717	1129
		M	519	977
	Both	F	708	1364
		M	365	1043
	Nasal	F	399	1410
		M	398	1193
	Both2	F	450	1225
		M	391	1251
	Oral2	F	285	1268
		M	315	1304
4-stage	Oral	F	635	1821
		M	475	1486
	Both	F	616	1183
		M	602	1176
	Nasal	F	551	1470
		M	453	1343
	Both2	F	242	1305
		M	507	1579

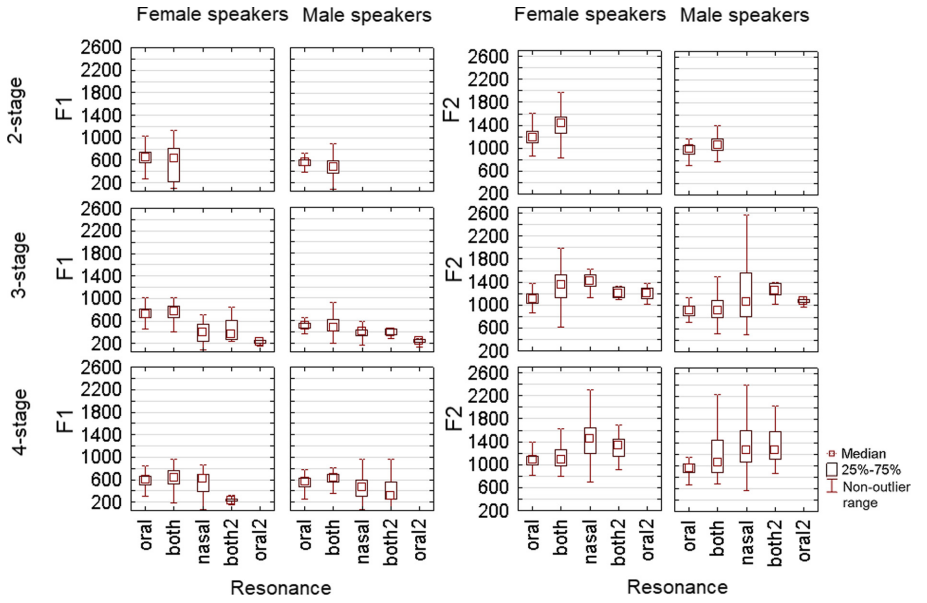

Fig. 6. Median values of F1 (left) and F2 (right) in 2-stage, 3-stage and 4-stage realizations of [ow̃] by male and female speakers.

and 'oral2' used in Table 2 denote the second occurrence of a stage based on the respective resonance (both or oral) in the course of the vowel realization. The median values of F1 and F2 are shown in the box-whiskers plots in Fig. 6.

As it might be seen, the case of the nasalized back vowel [ow] appears to be more sophisticated than [ew̃] in terms of the number of realization stages as well as by the results of formant frequency values measurements.

The average and median values of F1 were the highest for the oral stage in 2-stage realizations, however, for female voices, the dispersion around the middle value was significant. Furthermore, in 3-stage and 4-stage structures, F1 was similar or even higher for the oronasal stage (realized with both resonators) than for the oral stage, i.e. unlike for [ew̃] where F1 was systematically higher in the oral stage.

The values of F2 in 2-stage realizations were higher for stages produced with both resonances than for the oral ones. In 3-stage and 4-stage realizations, the frequency values appeared to be even higher but also more dispersed around the median.

4 Conclusions and Future Work

In this contribution, we presented preliminary results of the analysis of Polish nasalized vowels structure using a combined methodology. Formant frequencies were measured and analyzed for the subsequent stages of vowel realization. The

boundaries of the stages were established based on time-aligned spatial energy distribution data obtained from acoustic camera [11,14].

We provide new empirical input with regard to the structure of the vowels, which may be useful both for basic research and application purposes, cf. the difficulties reported for handling the two sounds in speech technology (such as speech segmentation tasks or acoustic modeling), and the on-going discussion in the subject literature.

Based on the findings, it may be concluded that the two nasalized vowels should not be treated in the same way when considering their internal structure, even if they are realized in the same preceding and following context within the utterance. The front vowel denoted in orthography by the grapheme ę might be seen as a prevalently two-stage, diphthong-like vowel (in agreement with many earlier studies). Notably, many of the remaining realizations of ę were produced as single-segment vowels. However, a different situation occurs with regard to the back vowel orthographically spelt with ą, where 2-, 3-, or even 4-stage realizations are equally common. The mean formant values do differ between these stages, but much overlapping and dispersion of the frequency values occur. That is consistent with the results obtained in our earlier production studies [13]. Consequently, the case of ą should be seen as much more sophisticated and prone to individual differences than ę which might have implications for both fundamental studies and practical applications in speech therapy or speech technology, e.g., acoustic modelling tasks.

Future work will include more detailed investigation of formant frequency variability, such as identification of F1 and F2 values at potential steady states in the course of particular vowel production stages or the differences in minimum vs. maximum values of the formant frequencies. As observed by Goldstein, speaker-identifying features based on formant tracks relate to the F1 and F2 variability [5]. Another step will be the inspection of other related parameters, e.g., formant frequency bandwidths that have been reported to influence speech intelligibility, and to enhance vowel identification processes [2,12]. The same parameters will be studied with regard to higher formant frequency values (F3, F4) that in turn might be associated with individual or more fine-grained differentiation of speech sounds. The differentiation might follow varying patterns depending on the speech sound category,cf. the differences in the importance of the higher formants in two vowel groups for Japanese (/o/ and /a/ as compared to /u/ and /e/) found by Fujisaki and Kawashima [4].

As a follow-up to the present contribution, we will report in more detail on the levels and potential differences between the formant frequencies within the stages produced with the same active resonances but at different positions in the course of the vowel, with a view to closer investigate the role of the order of appearance of particular resonances as well as the actual variability of formant trajectories.

References

1. Boersma, P., Weenink, D.: Praat: Doin (2014). http://www.praat.org. Accessed 20 Dec 2017
2. De Cheveigné, A.: Formant bandwidth affects the identification of competing vowels. ICPhS: International Congress of Phonetic Sciences, 2093 2096 (1999)
3. Dukiewicz, L.: Polskie głoski nosowe: analiza akustyczna. PWN, Warszawa (1967)
4. Fujisaki, H., Kawashima, T.: The roles of pitch and higher formants in the perception of vowels. IEEE Trans. Audio Electroacoust. **16**(1), 73–77 (1968)
5. Goldstein, U.G.: Speaker-identifying features based on formant tracks. J. Acoust. Soc. Am. **59**(1), 176–182 (1976)
6. Jassem, W.: Podstawy fonetyki akustycznej. Państwowe Wydawnictwo Naukowe (1973)
7. Jassem, W.: Illustrations of the IPA: Polish. J. Int. Phonetic Assoc. **33**(1), 103–107 (2003)
8. Karaś, M., Madejowa, M. (eds.): Słownik wymowy polskiej PWN. PWN, Warszawa (1977)
9. Klessa, K., Karpiński, M., Wagner, A.: Annotation Pro-a new software tool for annotation of linguistic and paralinguistic features. In: Proceedings of the Tools and Resources for the Analysis of Speech Prosody (TRASP) Workshop, Aix en Provence, pp. 51–54 (2013)
10. Klessa, K., Szymanski, M., Breuer, S., Demenko, G.: Optimization of Polish segmental duration prediction with cart. In: SSW, pp. 77–80 (2007)
11. Król, D., Lorenc, A., Święciński, R.: Detecting laterality and nasality in speech with the use of a multi-channel recorder. In: 40th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5147–5151. IEEE, Brisbane (2015)
12. Kuwabara, H., Ohgushis, K.: Role of formant frequencies and bandwidths in speaker perception. Electron. Commun. Jpn. **70**(9), 11–21 (1987)
13. Lorenc, A.: Wymowa normatywna polskich samogłosek nosowych i spółgłoski bocznej (En. Normative pronunciation of Polish nasalized vowels and the lateral consonant). Dom Wydawniczy Elipsa, Warszawa (2016)
14. Lorenc, A., Król, D., Klessa, K.: An acoustic camera approach to studying nasality in speech: the case of Polish nasalized vowels. J. Acoust. Soc. Ame **144**(6), 3603–3617 (2018)
15. Puppel, S., Nawrocka-Fisiak, J., Krassowska, H.: A Hand-Book of Polish Pronunciation for English Learners. PWN, Warszawa (1977)
16. Rocławski, B.: Podstawy wiedzy o języku polskim dla glottodydaktyków, pedagogów, psychologów i logopedów. Glottispol, Gdańsk (2010)
17. StatSoft, I.: Statistica (data analysis software system), version 6. Tulsa, USA 150 (2001)
18. Steffen-Batogowa, M.: Automatyzacja transkrypcji fonematycznej tekstów polskich. PWN, Warszawa (1975)
19. Wells, J.C.: SAMPA computer readable phonetic alphabet. In: Gibbon, D., Moore, R., Winski, R. (eds.) Handbook of Standards and Resources for Spoken Language Systems, Part IV, Section B. Mouton de Gruyter, Berlin and New York (1997)
20. Wierzchowska, B.: Wymowa polska. PZWS, Warszawa (1971)
21. Ziółko, B.: Speech Recognition of Highly Inflective Languages. Ph.D. thesis, Department for Computer Science, University of York (2009)