The structure of Polish nasalized vowels: results based on spatial energy distribution and formant frequency analysis

Anita Lorenc¹,², Katarzyna Klessa³, Daniel Król⁴

¹Department of Speech Therapy and Applied Linguistics, Maria Curie-Skłodowska University, Lublin, Poland
²Institute of Applied Polish Studies, Department of Speech and Language Therapy and Voice Production, Warsaw University, Poland
³Department of Phonetics, Institute of Linguistics, Adam Mickiewicz University in Poznań, Poland
⁴Department of Technology, Higher State Vocational School, Tarnów, Poland

anita.lorenc@uw.edu.pl, klessa@amu.edu.pl, dankrol@gmail.com

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 recent results obtained by Lorenc (2016) 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 in their internal structures also with regard to the number of stages which is not only a new contribution to the state of knowledge about their features but may also have important implications for speech technology.

Keywords: speech analysis, acoustic camera, formant frequency, acoustic field energy distribution, nasalized vowels

1. Introduction

The Polish vowel system is sometimes referred to as relatively simple in terms of description and usage, as it consists of (only) six oral vowels [a, e, i, o, u, ı]. 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., Puppel et al., 1977).

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., Dukiewicz, 1967; Rochwalski, 2010; Wierczowska, 1971). 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 Wierczowska, 1971), they were usually described as oral or oral-nasal ones (with an increasing presence of nasality but always accompanied by the oral resonance).

The two-segment approach has been reflected in the publications dedicated to applications and technical solutions, as well as phonetic transcription. For example, Steffen-Batogowa (1975) supported the idea of transcribing the sounds with [ę] and [ı] respectively in her work on automatic grapheme-to-phoneme conversion rules for Polish. In the Illustration of the IPA: Polish by W. Jassem (2003), 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)¹. However, the standard version of SAMPA (Wells, 1997) for Polish includes [e~], [o~] transcription labels for the two sounds. An extended version of the SAMPA alphabet was used in several studies on speech synthesis and in corpus annotation. Among others, it includes the [ew~] [ow~] labels instead of [e~] [o~], as well as separate labels for the nasalized approximants [w~] [j~]. In some works, it was postulated to treat the realizations of ę and ą as sequences of subsequent phones and subsequently to identify segment boundaries between their oral and nasal parts (e.g., Demenko et al., 2010). However, this task appeared to be very difficult in practice (the “boundary” was actually a continuous transient and any segmentation applied for the needs of unit selection resulted in glitches and disfluencies in synthesized speech). Klessa et al. (2007) reported that better speech synthesis results could be obtained by avoiding the segmentation into stages and treating the oral and nasal components as a whole that represents inherently diversified vowel segments.

Quite surprisingly, the normative approaches to the problem indicate a lack of consensus, even in terms of

¹ In this paper, we will follow the two-segment approach and transcribe the front vowel ę with [ew] and the back vowel ą with [ow]. 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.
For the purpose of the present experiment, a 60 x 60 mm fragment of the acoustic camera image was chosen that 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.

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 to +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

Formant frequencies were measured in Praat using the Formant listing option. Further processing was performed with Annotation Pro (Klessa et al., 2013). 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 (Section 2.2). This way, information 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. 2 (formant values are displayed as dotted lines, the software enables also numerical display).

![Formant measurement and data processing](image)

Fig. 2: Annotation Pro: an example view of time-aligned segmentation data, formant measurement results and three realization stages for the vowel [ɔw].

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 (see also: Król et al., 2015; Lorenc et al., in progress). The method of generating the information is illustrated in Fig. 1.

![Vowel stages and stage boundary positions](image)

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

2.1. Speakers and speech material

Speech recordings were 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 (2016).

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 [ɛw] and 85 realizations of [ɔw]). 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 [vɛzeɻ] (En. 'knot') or pączek [pɔɾˈɕɛk] (En. 'blushes'). The word recordings were manually segmented into phones using Praat (Boersma, 2002).

2. Data, methods and tools

2.1. Speakers and speech material

Speech recordings were 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 (2016).

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 [ɛw] and 85 realizations of [ɔw]). 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 [vɛzeɻ] (En. 'knot') or pączek [pɔɾˈɕɛk] (En. 'blushes'). The word recordings were manually segmented into phones using Praat (Boersma, 2002).

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 (see also: Król et al., 2015; Lorenc et al., in progress). The method of generating the information is illustrated in Fig. 1.

![Vowel stages and stage boundary positions](image)

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

For the purpose of the present experiment, a 60 x 60 mm fragment of the acoustic camera image was chosen that 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.

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 to +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

Formant frequencies were measured in Praat using the Formant listing option. Further processing was performed with Annotation Pro (Klessa et al., 2013). 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 (Section 2.2). This way, information 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. 2 (formant values are displayed as dotted lines, the software enables also numerical display).

![Formant measurement and data processing](image)

Fig. 2: Annotation Pro: an example view of time-aligned segmentation data, formant measurement results and three realization stages for the vowel [ɔw].

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 (see also: Król et al., 2015; Lorenc et al., in progress). The method of generating the information is illustrated in Fig. 1.

![Vowel stages and stage boundary positions](image)

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

For the purpose of the present experiment, a 60 x 60 mm fragment of the acoustic camera image was chosen that 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.

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 to +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

Formant frequencies were measured in Praat using the Formant listing option. Further processing was performed with Annotation Pro (Klessa et al., 2013). 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 (Section 2.2). This way, information 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. 2 (formant values are displayed as dotted lines, the software enables also numerical display).

![Formant measurement and data processing](image)

Fig. 2: Annotation Pro: an example view of time-aligned segmentation data, formant measurement results and three realization stages for the vowel [ɔw].
The Praat formant listing files provided information about the frequencies of the formants F1-F4 for the whole utterances (container words). We used a C# plugin scripts to link the relevant formant frequency values to the respective vowel stages, and also to calculate mean formant values per stage. Further statistical analysis using both the mean values and raw measurement results was carried out with Statistica (Statsoft Inc., 1984-2011).

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 [əw] were determined as 2-stage ones. Altogether, 57 out of 76 instances of [əw] 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 [əw], 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

3.2.1. The nasalized vowel [əw]

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

<table>
<thead>
<tr>
<th>Structure</th>
<th>Resonance type</th>
<th>Speaker gender</th>
<th>Mean F1 [Hz]</th>
<th>Mean F2 [Hz]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-stage</td>
<td>oral</td>
<td>F</td>
<td>635</td>
<td>1821</td>
</tr>
<tr>
<td></td>
<td></td>
<td>M</td>
<td>475</td>
<td>1486</td>
</tr>
<tr>
<td>2-stage</td>
<td>oral</td>
<td>F</td>
<td>672</td>
<td>1768</td>
</tr>
<tr>
<td></td>
<td></td>
<td>M</td>
<td>494</td>
<td>1490</td>
</tr>
<tr>
<td></td>
<td>both</td>
<td>F</td>
<td>435</td>
<td>1673</td>
</tr>
<tr>
<td></td>
<td></td>
<td>M</td>
<td>365</td>
<td>1459</td>
</tr>
</tbody>
</table>

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

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 < 0.001). The frequency values in 1-stage realizations are close to Polish non-nasalized [ə] reported by Jassem (1973).

Fig. 3 displays scatter-plots of the mean formant frequency values in the F2-F1 space for the 2-stage realizations of [əw] (inverted axes were used as in typical vowel charts). Though the values obtained for the subsequent stages overlap to a certain extent, two different areas can still be distinguished.

Fig. 3: Scatter-plots of the mean formant values in the F2-F1 space for 2-stage [əw] (left: female speakers, right: male speakers).

Median values and the respective percentile ranges of F1 and F2 for 1-stage and 2-stage realizations of [əw] are presented in the box-whiskers plots in Fig. 4 and Fig. 5.

Fig. 4: Median values of F1 in 1-stage and 2-stage realizations of [əw] by male and female speakers.

In 2-stage realizations of [əw], 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 Introduction), 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.

![Graph showing F2 values for 1-stage and 2-stage realizations of [ow] by male and female speakers.](image)

Fig. 5: Median values of F2 in 1-stage and 2-stage realizations of [ow] by male and female speakers.

### 3.2.2. The nasalized vowel [ow]

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

<table>
<thead>
<tr>
<th>Structure</th>
<th>Resonance type</th>
<th>Speaker gender</th>
<th>Mean F1 [Hz]</th>
<th>Mean F2 [Hz]</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-stage</td>
<td>oral</td>
<td>F</td>
<td>644</td>
<td>1206</td>
</tr>
<tr>
<td></td>
<td></td>
<td>M</td>
<td>558</td>
<td>1008</td>
</tr>
<tr>
<td></td>
<td>both</td>
<td>F</td>
<td>542</td>
<td>1404</td>
</tr>
<tr>
<td></td>
<td></td>
<td>M</td>
<td>497</td>
<td>1161</td>
</tr>
<tr>
<td></td>
<td>oral</td>
<td>F</td>
<td>717</td>
<td>1129</td>
</tr>
<tr>
<td></td>
<td></td>
<td>M</td>
<td>518</td>
<td>977</td>
</tr>
<tr>
<td></td>
<td>both</td>
<td>F</td>
<td>708</td>
<td>1364</td>
</tr>
<tr>
<td></td>
<td></td>
<td>M</td>
<td>519</td>
<td>1043</td>
</tr>
<tr>
<td></td>
<td>nasal</td>
<td>F</td>
<td>399</td>
<td>1410</td>
</tr>
<tr>
<td></td>
<td></td>
<td>M</td>
<td>398</td>
<td>1193</td>
</tr>
<tr>
<td></td>
<td>both2</td>
<td>F</td>
<td>450</td>
<td>1225</td>
</tr>
<tr>
<td></td>
<td></td>
<td>M</td>
<td>391</td>
<td>1251</td>
</tr>
<tr>
<td></td>
<td>oral2</td>
<td>F</td>
<td>285</td>
<td>1268</td>
</tr>
<tr>
<td></td>
<td></td>
<td>M</td>
<td>315</td>
<td>1304</td>
</tr>
<tr>
<td></td>
<td>oral</td>
<td>F</td>
<td>583</td>
<td>1065</td>
</tr>
<tr>
<td></td>
<td></td>
<td>M</td>
<td>558</td>
<td>1037</td>
</tr>
<tr>
<td></td>
<td>both</td>
<td>F</td>
<td>616</td>
<td>1183</td>
</tr>
<tr>
<td></td>
<td></td>
<td>M</td>
<td>602</td>
<td>1176</td>
</tr>
<tr>
<td></td>
<td>nasal</td>
<td>F</td>
<td>551</td>
<td>1470</td>
</tr>
<tr>
<td></td>
<td></td>
<td>M</td>
<td>453</td>
<td>1343</td>
</tr>
<tr>
<td></td>
<td>both2</td>
<td>F</td>
<td>242</td>
<td>1305</td>
</tr>
<tr>
<td></td>
<td></td>
<td>M</td>
<td>507</td>
<td>1579</td>
</tr>
</tbody>
</table>

Table 2. Mean F1 and F2 values in 2-stage and 3-stage realizations of [ow] (F- female, M-male; both2 and oral2 denote the second occurrence of a stage based on the respective resonance).

The majority of multiple-stage realizations of [ow] began with the oral resonance, however, in certain cases at the beginning, both oral and nasal resonances were active. Table 2 includes data for all the observed stages with regard to resonance types for [ow] 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 and oral2 used in the Table 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 and Fig. 7.

![Box-whiskers plots showing F1 and F2 values for 2-, 3-, and 4-stage realizations of [ow] by male and female speakers.](image)

Fig. 6: Median values of F1 in 2-, 3- and 4-stage realizations of [ow] by male and female speakers.

![Box-whiskers plots showing F2 values for 2-, 3-, and 4-stage realizations of [ow] by male and female speakers.](image)

Fig. 7: Median values of F2 in 2-, 3- and 4-stage realizations of [ow] by male and female speakers.

As might be seen, the case of the nasalized back vowel [ow] appears to be more sophisticated than [ew] 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 [ew] 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.

5. Conclusions & 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.

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 so-far 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. The front vowel [ew] might be seen as a prevalently two-stage, diphthong-like vowel (in agreement with many earlier studies). However, a different situation occurs with regard to the back vowel [ow], 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 values occur. Consequently, the case of [ow] should be seen as much more sophisticated and prone to individual differences than [ew].

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 stages. Another step will be the inspection of other related parameters, e.g., formant bandwidths that have been reported to influence speech intelligibility, especially vowel identification (Cheveigne, 1999; Kuwabara et al, 1987). The same parameters will be studied with regard to higher formant frequency values (F3, F4).

As a follow-up to the present paper, 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.

Acknowledgement

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).

References