EXPLORING THE ROLE OF FORMANT FREQUENCIES IN THE CLASSIFICATION OF PHONATION TYPE

Bogdan Ludusan[1], Mattias Heldner[2], Marcin Włodarczak[2]

[1]Bielefeld University, [2]Stockholm University
bogdan.ludusan@uni-bielefeld.de, {heldner,wlodarczak}@ling.su.se

ABSTRACT

Phonation type has been found to interact with other acoustic characteristics of speech, including vowel formants. Here, we investigated the relation between phonation type and formants in three Swedish vowels, produced in three phonation types (neutral, breathy and pressed). Our findings show that the formant values change from neutral to breathy and pressed, respectively, but that this effect is more consistent for pressed than for breathy phonation. Moreover, while some systematic patterns can be identified, there is also substantial speaker variation in the size and in the direction of the change. We then employed the formant values, along with other acoustic features, to automatically classify phonation type. The addition of formant information to the classifier significantly improved the performance of the system.

Keywords: phonation type, formant values, pressed, breathy, classification.

1. INTRODUCTION

Phonation type has been primarily described in terms of the degree of glottal adduction [1], which is, in turn, linked to other aspects of voice production, such as variation in acoustic intensity [2] and fundamental frequency [3]. In addition, given that the degree of adduction is associated with vertical larynx position [4, 5], the glottal configuration also affects formant frequencies by changing the vocal tract length [6].

A review of the effects of phonation type on formant frequencies across a number of languages [1] found that breathy phonation is associated with lower formant frequencies (in particular F1) than creaky phonation (but cf. [7], who found no significant effect of phonation type on F1 in Jalapa Mazatec). In a more controlled experimental setting, with participants producing the syllable /pæː/ in modal and pressed phonations, [8] found that ratings of pressedness by speech-language pathologists were linked to variation in formant frequencies (F1 in females and F2 in males). While all these effects were attributed to the impact of phonatory setting on larynx height, recent developments, such as the Laryngeal Articulator Model [9, 10] have proposed a more comprehensive view, whereby raising the larynx also involves the aryepiglottic constriction and tongue retraction.

Whatever the mechanisms involved, instrumental classification of phonation types has predominantly employed methods explicitly designed to be insensitive to formant effects, whether by using inverse filtering [11], electroglottography [12], miniature accelerometers attached to the neck [13] or acoustic measures in the cepstral domain [14]. However, recent work [15] found that automatic classification of phonation type based on acoustic features derived directly from the speech signal reaches similar performance to those calculated from inverse filtering and neck-surface acceleration.

Given the evidence that supraglottal resonances are affected by phonatory posturing, in this paper we analyse the effect of formant frequencies on the automatic classification of phonation type and their relationship with other acoustic measures of phonation type.

2. MATERIALS

The speech materials consisted of diminuendo sequences of syllables with three different vowels [piː, pæː, poː]; produced in three phonation types (neutral, breathy, pressed); at three self-selected pitch levels (habitual, low, high); by five trained singers (two females: S3 and S5, and three males). Each combination of factors was repeated in at least three diminuendo sequences. The recordings were made in a sound treated room using a head-mounted omnidirectional microphone (Sennheiser HSP2) positioned 6 cm from the mouth and digitized using an Expert Sleepers ES-9 audio interface (48 kHz, 24 bit). The recording included other transducers as well, but only the microphone signals was used in the present paper. This procedure resulted in a total of 2640 syllables (897 pæː, 887

pi: and 856 po:), corresponding to breathy (790), neutral (934), and pressed (916) phonation.

3. METHODS

The syllables were annotated by two of the authors. The position of the burst of the stop consonant and the end of the vowel portions were marked and the middle point between these two landmarks was used in the subsequent analysis.

In a first step, the values of first three formants (F1, F2 and F3) were extracted using the *burg* method of the Praat software [16], employing the default parameters in the case of female speakers and a lower formant ceiling (5 kHz) for male speakers. We then investigated whether differences occur in terms of F1-F3 between the three phonation types, for each speaker. The significance of the differences was determined by means of Wilcoxon rank sum tests, corrected for multiple comparisons with the method in [17]. The statistical analyses were run using the appropriate function of the R software [18]. We also examined the relation between formant values and a number of acoustic features previously employed for the classification of phonation type [15]. These were extracted (using Praat), from the same time instants where the formant values were computed. The considered set includes the following features:

- *alpha* - alpha ratio. It represents the ratio of acoustic energy between the high (1-5 kHz) and the low (0-1 kHz) frequency bands.
- *cpps* - smooth cepstral peak prominence. It is defined as the amplitude of the first cepstrum rahmonic relative to the regression line of the cepstrum of the signal [19].
- *f0* - fundamental frequency. The vibration rate of the vocal folds.
- *h1h2*. It consists of the difference between the level of the first two harmonics.
- *hrf* - harmonic richness factor. Defined as in [20], being the ratio between the summed amplitudes of the 2nd to 10th harmonics and the amplitude of the fundamental.

In a second step, we tested if formant information may be used to improve the automatic detection of phonation type. We used the Random Forest algorithm [21] for the classification experiment (the implementation offered by the *randomForest* R package [22]). Random Forest has previously been employed in phonetic research (e.g., [23]), due to its characteristic of being able to return also the importance of the features involved in the classification task. We considered three cases: neutral-breathy (N-B), neutral-pressed (N-P) and breathy-pressed (B-P). For each case, we trained a Random Forest classifier, consisting of 500 trees, with the set of features employed in [15] and another classifier with the same features plus the values of the first three formants. The out-of-bag (OOB) error was chosen as evaluation measure for the goodness of classification. It is calculated on a subset of the data on which the trees were not fitted and a lower error value represents a better discrimination performance.

We determined the importance of each of the features used in the classification experiments. It was operationalized as the total decrease in node impurities (as given by the Gini index) when performing a split on that feature, averaged across all trees. A more discriminative feature would be represented by a higher importance. We then normalized the importance of each feature by the sum of importance values of all features considered in that condition. After this process, the sum of the importance of all the features in each condition will be equal to 1, allowing for an easier comparison between conditions.

The experiments were run on a per-speaker basis and all vowel instances produced by the speaker in the two phonation types discriminated in the experiment were considered. For each speaker and phonation type pair, a total of 100 runs were performed and the mean OOB error and mean importance of the features over the 100 runs reported. To test the statistical significance of the differences between the discrimination of various conditions, as well as between the two feature sets (without/with formant values), we used bootstrapping (by means of the functions of the R package *boot* [24]). It is a non-parametric method which performs sampling with replacement from the set containing the OOB error rates given by the random forest classifier. We repeated the process for 10,000 times, computing at each iteration the mean error and the 95% confidence intervals. If the obtained confidence intervals do not overlap, it means the differences are statistically significant.

4. RESULTS

4.1. Acoustic analysis

The variation of the formant values with respect to the three investigated phonation types is illustrated in Figure 1. For most speaker/vowels combinations, we can see that changing the phonation type from neutral to breathy or pressed results in a change in the mean formant values of the vowel, and

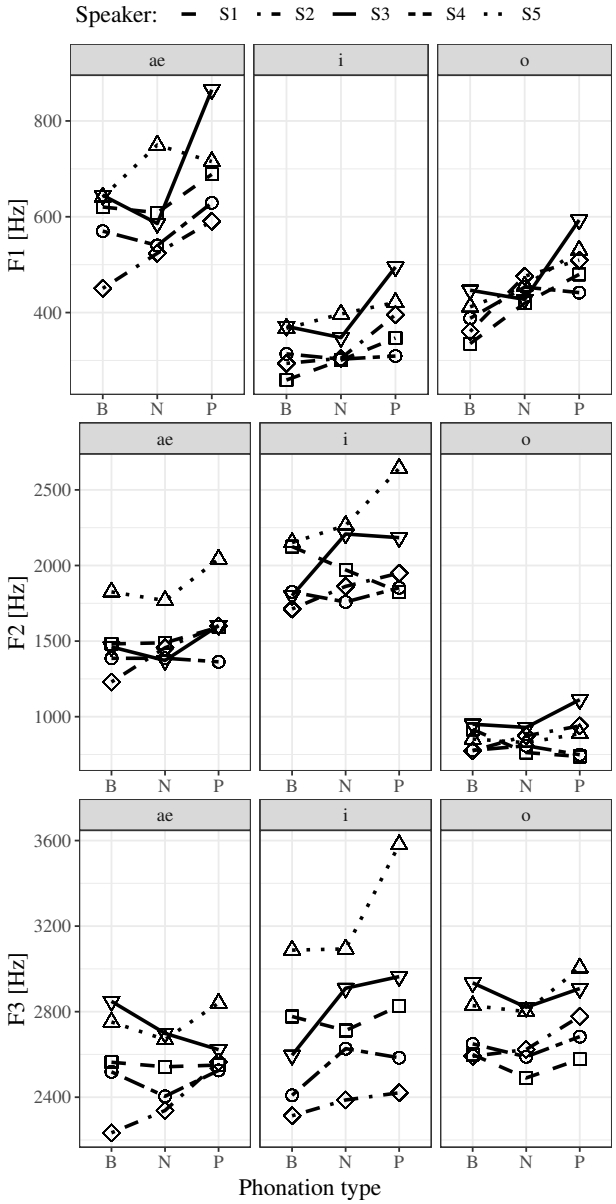

Figure 1: Mean F1, F2 and F3 values across speakers, for each of the investigated vowels in the three phonation type conditions (B - breathy, N - neutral, P - pressed).

formant values generally increased from breathy to pressed. However, one may note a variety of patterns involving the neutral condition, which implies the use of different strategies by the speakers.

We tested the differences between formant values for all pairs of phonation types, within each speaker, by means of Wilcoxon rank sum tests. The results are presented in Table 1. It shows that all speakers modify their formants in one way or another, across phonation types, but also that important individual variation exists. The number of cases (phonation type/formant) where significant changes occurred varied, from the maximal number of cases (9), for

Spkr.	N-B			N-P			B-P		
	F1	F2	F3	F1	F2	F3	F1	F2	F3
S1	×	−	×	×	−	×	×	−	−
S2	×	×	×	×	×	×	×	×	×
S3	−	−	−	×	×	−	×	×	−
S4	−	−	−	×	−	×	×	−	×
S5	−	×	×	×	×	×	×	×	×
All	×	×	−	×	×	×	×	×	×

Table 1: The significance of the change in formant (F1-F3) values between phonation types, for each speaker and for all speakers. $\times/-$ denote statistical significance/lack of it ($\alpha = .05$).

S2, to about half the cases (4), for S3 and S4. While the majority of speakers produced some sort of change in all three investigated phonation type contrasts, speakers S3 and S4 produced no change between neutral and breathy phonation. Overall (last row of Table 1), we see that all but F3 in the N-B case exhibit significant differences (although three of the five speakers changed their F3 between neutral and breathy, the direction differed between the speakers, resulting in a null overall effect).

Finally, we checked the relation between the values of the formants and those of acoustic features previously employed for automatic phonation type classification. The obtained Spearman ρ is illustrated in Table 2. Three of the five features (alpha, h1h2 and hrf) exhibit medium correlations with the first formant value and one of them (alpha) also with the value of the second formant.

4.2. Automatic phonation type detection

The results of the two experiments (without/with the values of the first three formants) are presented in Figure 2. In both experiments, the highest error was obtained in the N-B condition, followed by N-P and the lowest one in the B-P case. A reduction in error rate was seen for all three conditions when the formants were added to the feature set (N-B: 0.9%, N-P: 2.2%, B-P: 1.3%, which corresponds to an additional 25, 57 and 34 correctly classified vowels).

feature	F1	F2	F3
alpha	.49	.55	-.08
cpps	.20	.07	-.14
f0	.31	.12	.29
h1h2	-.45	.19	-.04
hrf	.56	-.23	-.03

Table 2: Spearman rho for the correlation between formant values and acoustic features, across all phonation types and vowels. All but hrf-F3 are significant ($\alpha = .05$).

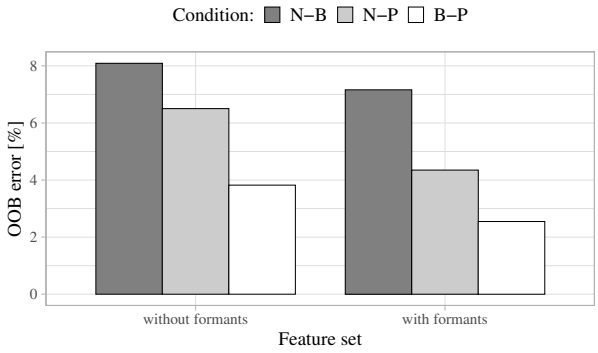

Figure 2: Classification results (mean OOB error) for the two feature sets: without and with formant values, and for the following the phonation type pairs: neutral-breathy (N-B), neutral-pressed (N-P) and breathy-pressed (B-P).

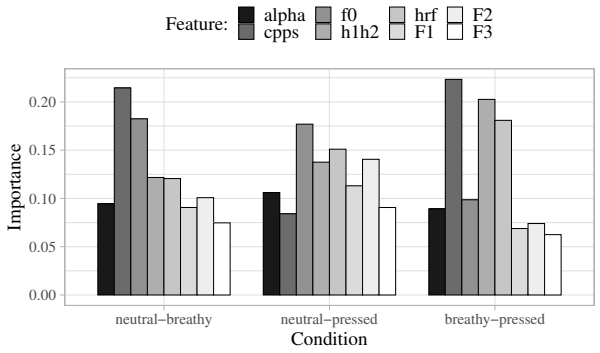

Figure 3: The importance of the features used in the classification experiments considering the formant values, in each of the three conditions.

We tested the differences between conditions within each of the two experiments, as well as between the experiments. There was no overlap between the 95% confidence intervals of any of the pairs compared, which means that all pairwise differences (within and between experiments) are significant. Looking at the per-speaker results in more detail, we observed a decrease in error in 14 of the 15 cases (3 conditions $\times$ 5 speakers): the classification of S2 data did not improve in the N-B condition.

The importance of all the features used in the second experiment (including the formant values) is illustrated in Figure 3. It shows that formant values helped most the N-P classification, all formants being ranked higher than at least one of the other five features (F2 - 3rd, F1 - 5th and F3 - 7th). In the N-B case only F2 was more important than the least important of the other features (alpha), while for the B-P classification the formants had the lowest importance. However, the importance of the formants varied with the speaker, the classification algorithm relying more on formants for some speakers than for others.

5. CONCLUSIONS

We have conducted an investigation of the effects of phonation type on formant values, by systematically analyzing the first three formants of three Swedish vowels, across five speakers. It revealed that all three formants are affected by phonation type, with the changes in formant values for pressed phonation being more systematic than for breathy phonation (compared to neutral phonation production). A higher number of cases showing significant differences in formant values in the comparisons involving pressed speech (as seen from Table 1) suggest that the most marked changes to supraglottal resonances, and implicitly in the position of the formants, occurs in the pressed condition. The observed individual variation might be due to the ability of the speakers to control their larynx height (speakers S1, S3 and S4, who are more experienced singers, show less of a difference between phonation types). We also analysed the relation between formants and other acoustic measures of phonation type. The formant values were mostly correlated to acoustic features that characterize spectral slope (alpha), but also partly the strength of the fundamental (h1h2, hrf). By contrast, cpps, an acoustic measure in the cepstral domain designed to be robust to formant variation, was the least correlated with F1-F3.

We then used the formant information for automatic classification of phonation type, by comparing the performance of one set of features previously used successfully for this task with that of the same set enriched with the formant values. Despite their correlation with some of the other acoustic features and their relatively low mean importance for the classification task (especially in the B-P classification), the use of formant values helped decrease the OOB error rate in all conditions. This shows that formant information is beneficial to automatic detection of phonation type, in a speaker-dependent manner. Future work may explore whether this type of information can be used also in a speaker-independent fashion.

ACKNOWLEDGEMENTS

This work was funded by Swedish Research Council project *Prosodic functions of voice quality dynamics* (VR 2019-02932) to MW. The authors would like to thank Johan Sundberg for the useful feedback.

6. REFERENCES

[1] M. Gordon and P. Ladefoged, "Phonation types: a cross-linguistic overview," *Journal of Phonetics*, vol. 29, no. 4, pp. 383–406, 2001.

[2] J. Sundberg and M. Nordenberg, "Effects of vocal loudness variation on spectrum balance as reflected by the alpha measure of long-term-average spectra of speech," *Journal of the Acoustical Society of America*, vol. 120, no. 1, pp. 453–457, 2006.

[3] J.-M. Hombert, J. J. Ohala, and W. G. Ewan, "Phonetic explanations for the development of tones," *Langauge*, vol. 55, no. 1, pp. 37–58, 1979.

[4] J. Sundberg and A. Askenfelt, "Larynx height and voice source. A relationship?" *STL-QPSR*, vol. 22, no. 2–3, pp. 23–36, 1981.

[5] M. Saldias, M. Guzman, G. Miranda, and A.-M. Laukkanen, "A computerized tomography study of vocal tract setting in hyperfunctional dysphonia and in belting," *Journal of Voice*, vol. 33, no. 4, pp. 412–419, 2019.

[6] J. Sundberg and P.-E. Nordström, "Raised and lowered larynx – the effect on vowel formant frequencies," *STL-QPSR*, vol. 17, no. 2–3, pp. 35–39, 1976.

[7] M. Garellek and P. Keating, "The acoustic consequences of phonation and tone interactions in jalapa mazatec," *Journal of the International Phonetic Association*, vol. 41, no. 2, pp. 185–205, 2011.

[8] M. Millgård, T. Fors, and J. Sundberg, "Flow glottogram characteristics and perceived degree of phonatory pressedness," *Journal of Voice*, vol. 30, no. 3, pp. 287–292, 2016.

[9] J. H. Esling, S. R. Moisik, A. Benner, and L. Crevier-Buchman, *Voice Quality: The Laryngeal Articulator Model*. Cambridge University Press, 2019.

[10] J. A. Edmondson and J. H. Esling, "The valves of the throat and their functioning in tone, vocal register and stress: Laryngoscopic case studies," *Phonology*, vol. 23, no. 2, pp. 157–191, 2006.

[11] J. Sundberg, "Objective characterization of phonation type using amplitude of flow glottogram pulse and of voice source fundamental," *Journal of Voice*, vol. 36, no. 1, pp. 4–14, 2020.

[12] M. Borsky, D. D. Mehta, J. H. Van Stan, and J. Gudnason, "Modal and nonmodal voice quality classification using acoustic and electroglottographic features," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 12, pp. 2281–2291, 2017.

[13] M. Borsky, M. Cocude, D. D. Mehta, M. Zañartu, and J. Gudnason, "Classification of voice modes using neck-surface accelerometer data," in *Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, 2017, pp. 5060–5064.

[14] J. Hillenbrand, R. A. Cleveland, and R. L. Erickson, "Acoustic correlates of breathy vocal quality," *Journal of Speech Language and Hearing Research*, vol. 37, no. 4, 1994.

[15] M. Włodarczak, B. Ludusan, J. Sundberg, and M. Heldner, "Classification of voice quality using neck-surface acceleration: Comparison with glottal flow and radiated sound," *Journal of Voice*, 2022.

[16] P. Boersma and D. Weenink, "Praat: doing phonetics by computer," Computer program, 2021, http://www.praat.org/.

[17] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal statistical society: series B (Methodological)*, vol. 57, no. 1, pp. 289–300, 1995.

[18] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2020. [Online]. Available: https://www.R-project.org/

[19] J. Hillenbrand and R. A. Houde, "Acoustic correlates of breathy vocal quality: Dysphonic voices and continuous speech," *Journal of Speech, Language, and Hearing Research*, vol. 39, no. 2, pp. 311–321, 1996.

[20] L. Eskenazi, D. G. Childers, and D. M. Hicks, "Acoustic correlates of vocal quality," *Journal of Speech, Language, and Hearing Research*, vol. 33, no. 2, pp. 298–306, 1990.

[21] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[22] A. Liaw and M. Wiener, "Classification and regression by randomForest," *R News*, vol. 2, no. 3, pp. 18–22, 2002. [Online]. Available: https://CRAN.R-project.org/doc/Rnews/

[23] B. Ludusan, P. Wagner, and M. Włodarczak, "Cue interaction in the perception of prosodic prominence: The role of voice quality," in *Proc. of Interspeech*, 2021, pp. 1006–1010.

[24] A. Canty, "Resampling methods in R: the boot package," *The Newsletter of the R Project Volume*, vol. 2, no. 3, pp. 2–7, 2002.