Year: 2022

Idiosyncratic lingual articulation of American English /æ/ and /ɑ/ using network analysis

Lins Machado, Carolina ; Dellwo, Volker ; He, Lei

DOI: https://doi.org/10.21437/interspeech.2022-10397

Originally published at:
Lins Machado, Carolina; Dellwo, Volker; He, Lei (2022). Idiosyncratic lingual articulation of American English /æ/ and /ɑ/ using network analysis. In: Interspeech 2022, Incheon, Korea, 18 September 2022 - 22 September 2022, ISCA.
DOI: https://doi.org/10.21437/interspeech.2022-10397

Interspeech 2022
18-22 September 2022, Incheon, Korea

Idiosyncratic lingual articulation of American English /æ/ and /ɑ/ using network analysis

Carolina Lins Machado[1], *Volker Dellwo*[1], *Lei He*[12]

[1] Dept. Computational Linguistics, University of Zurich, Zurich, Switzerland
[2] Dept. Phoniatrics and Speech Pathology, Clinic for Otorhinolaryngology, Head and Neck Surgery, University Hospital Zurich (USZ), Zurich, Switzerland
cmachado@ifi.uzh.ch, volker.dellwo@uzh.ch, lei.he@uzh.ch

Abstract

Formant dynamics are believed to reflect the characteristic articulatory behavior of a speaker. The present study aims to explore individual articulatory behaviors when producing American English /æ/ and /ɑ/. The two vowels differ in the degree of inherent spectral change, a property believed to carry information about vowel-phoneme identity, which may be reflected in the articulatory movements. We measured first and second formants together with tongue blade and dorsum trajectories from 20 speakers producing 330 words in citation forms. Using the network analysis, the relationships between acoustic and kinematic variables were revealed. In particular, between-speaker articulatory behaviors were most dissimilar in /ɑ/ which requires less inherent spectral change. Moreover, when networks of speakers with similar formant patterns were compared, it was revealed that their articulatory behaviors also shared similarities, although they seemed to be organized in characteristic ways. These findings contribute to our understanding of the complex interaction between articulatory variables and the acoustic outcome.

Index Terms: speaker-specific articulation, formant dynamics, network analysis, electromagnetic articulography

1. Introduction

During speech production the characteristic articulatory behavior of a speaker is expressed in terms of acoustic dynamic features [1]. Formant dynamics is one particular feature considered a rich source of speaker-specific information believed to be related to an individual's articulatory strategy [2–4]. A speaker's vocal tract morphology and the articulatory strategy employed to produce a vowel results in the alteration of the vocal tract resonant frequencies (i.e. formants) in characteristic ways. Yet, the actual articulatory strategy employed by a speaker during vowel production affecting the resulting formant dynamics remains understudied. This paper aims to explore the relationship between idiosyncratic articulatory behaviors of the tongue and the first two formants using network analysis.

Primarily, formant dynamics carry vital information about vowel-phoneme identity in a property denominated "vowel-inherent spectral change" or VISC [5]. VISC refers to vowel-specific changes in spectral properties over the time course of a vowel. However, while for some vowels substantial formant changes pertain to vowel identity, for others a lesser amount of inherent change may carry phonemic information. For the latter, this is believed to be due to vowel-specific saturation effects; i.e., some vowels may require less intrinsic spectral movement because they have relative stable acoustic goals [6]. As a result, such vowels may be produced with more variability in the underlying articulatory strategy allowing the speaker to reduce the need for precision in articulatory movements [7]. Consequently, reduced articulatory control could result in the presence of more speaker-dependent information in the VISC, since the linguistic constraints imposing articulatory control may be less strong in these vowels and allow room for speakers' preferred articulatory strategies.

Previous research demonstrated that individual differences in vowel production were a result of speaker-specific articulatory patterns [8]. Articulatory analyses revealed that even though there are common features in vowel production across individuals, speakers recruited articulators differently to produce the same vowel [8, 9]. In the task-dynamic framework [10] differences in the recruitment of articulators to produce a specific auditory outcome is referred to as "inter-articulator coordination". Speakers can articulate in different ways and still achieve the desired acoustic goal thanks to "motor equivalence", which is the ability to achieve a similar goal employing different strategies.

Typically, inter-articulatory coordination is investigated by examining position data of the articulators involved in the production of a speech sound. Although valuable information is obtained from this type of analysis, there are still important factors which remain unaccounted for. For instance, while the contribution of each articulatory variable can be related to the acoustic outcome, limited information is revealed regarding the relationship between articulators, their spatio-temporal configurations, and the acoustic outcome. Moreover, the visualization of position data does not display concealed variables that may be indirectly affecting the acoustic signal.

The present study is a proof-of-concept that explores dynamic individual articulatory behavior. We examine both acoustic and articulatory productions of the English vowels /æ/, containing substantial vowel-specific formant trajectories, and /ɑ/, exhibiting vowel-specific saturation effects and consequently less precision in articulatory movements. As argued, it is likely that speakers coordinate their articulators for speech in characteristic ways, resulting in idiosyncratic information present in dynamic formant measures. Based on this assumption, we seek to characterize speaker-specific articulatory behavior and how formant dynamics are affected in the process using network analysis.

A network is a collection of nodes representing different variables. The connections between nodes are called edges, the weights of which indicate the connection strengths between nodes or variables [11]. The structure of the data can thus be unveiled. In addition, the intrinsic properties of a system can

10.21437/Interspeech.2022-10397

also be revealed from the network. For instance, the particular patterns of interaction between the variables in a system helps us understand how a variable's action implicitly affects the outcome of another variable. Furthermore, dynamic networks can reveal how a system's behavior changes over time. Thus, analyzing dynamic networks may deepen our understanding of speaker-specific speech behavior by exposing overt and covert levels of articulation.

2. Method

2.1. Material

The data used in this study comes from the EMA-MAE corpus [12]. We selected the L1 group consisting of 10 male and 10 female native speakers of English, with an upper Midwest American English dialect background. Subjects were typical individuals with no history of speech, language, or oral pathology, no history of orofacial surgery or use of medication that may affect motor performance (age range = 18–40).

From three speaking tasks, productions of the vowels /æ/ and /ɑ/ were selected from the "MAE minimal contrast word pairs" task. Vowel tokens produced in the context of rhotic, lateral, nasal, and approximant syllable onset or codas were excluded from the selection, since coarticulatory effects related to these consonants have been shown to affect vowel formants in complex ways [13]. Acoustic and kinematic data were collected using 330 text-prompted words in single-word citation form. Kinematic data (sampling rate = 400Hz) were collected using the NDI Wave electromagnetic articulograph along with time-synchronized acoustic recordings (sampling rate = 22kHz). Position data from the tongue dorsum (TD) and the tongue blade (TB) were analyzed in both anterior-posterior and superior-inferior directions (x and y coordinates respectively). The kinematic data were head-corrected using a 6DOF reference sensor and calibrated using a bite-plate for individual speakers. The tongue kinematic data were not decoupled from jaw movement. The tongue kinematic measurements in our analysis contain tongue-jaw compound movements, and hence correspond to the tract variables of *tongue body constriction location/degree* and *tongue tip constriction location/degree* in Articulatory Phonology [14].

2.2. Data extraction

Prior to data extraction, the audio files were automatically annotated at two levels, word and vowel, in Praat [15] using the phoneme-level transcriptions provided in the corpus. Annotation was manually checked to ensure correct boundary placement before token segmentation.

For each token, measures of F1 and F2 were taken between vowel onset and offset using the Burg method in Praat with formant ceiling adjusted per gender (F = 5500Hz; M = 5000Hz; window length = 25ms; time step = 6.25ms; pre-emphasis from 50Hz). VISC trajectory points were then linearly interpolated at nine equally spaced steps. Finally, as an attempt to minimize the effect of coarticulation from the adjacent consonants, only the five innermost points (hereinafter, analysis points) were kept for further analysis [16]. Kinematic data points (x and y coordinates) of the two tongue sensors were sinc interpolated (depth = 70) at the same time landmarks measured in the acoustic analyses. Prior to extraction, the EMA signal was low pass filtered at 10 Hz (Hann filter; smoothing = 5Hz). Again, only the five innermost points of the kinematic trajectories were kept for further analysis.

Acoustic and kinematic data were subsequently processed in R [17]. After outlier removal, the data were normalized per speaker by vowel type. Next, subsets of the data were created per vowel containing the two acoustic variables F1 and F2, and four kinematic variables: tongue blade and tongue dorsum position values in the x- and y-dimensions: TDx, TDy, TBx, TBy. Overall the datasets consisted of 1220 data points for /æ/ and 970 for /ɑ/.

2.3. Statistical analysis

Static and dynamic networks were estimated in R using the *bootnet* package [18]. All networks were defined as follows: *nodes* comprised two acoustic variables F1 and F2, and four kinematic variables TDx, TDy, TBx, TBy. The weights of node-to-node *edges* are quantified as the partial correlations between all nodes. The Spearman's ρ was used, because the variables to be correlated deviated from the normal distribution. It is important to note that partial correlation networks can indicate potential indirect causal pathways [19].

For each vowel /æ/ and /ɑ/, the networks were constructed across all speakers, capturing the common articulatory strategies. Specifically, two types of networks, static and dynamic, were constructed. Static networks were calculated from all five analysis points for all tokens per vowel. Here, "static" possesses the meanings of "averaged" and "holistic", rather than "reaching an equilibrium or homeostasis". Dynamic networks contained a group of 5 networks calculated at individual analysis points for all tokens per vowel, showing temporal characteristics. Both types of networks for /æ/ and /ɑ/ are referred to as *vowel static networks* and *vowel dynamic networks*.

Similarly, static and dynamic networks were calculated for each speaker for all tokens per vowel. They are referred to as *speaker static networks* and *speaker dynamic networks*. For vowel networks (both static and dynamic) and speaker static networks, edges with significant correlations (alpha = 0.05) were displayed . For speaker dynamic networks at each analysis point, edges with an absolute ρ value of 0.30 were displayed.

To evaluate how speakers' articulatory behaviors deviated from each vowel network and to compare articulatory strategies between speakers, similarity between networks was calculated using the Jaccard similarity index (the ratio between the cardinalities of the intersection and the union of two sets, which, in the context of this paper, are displayed network edges) [20]. An index of 1.0 indicates networks with identical displayed edges (edge weights may vary, nevertheless).

3. Results

Due to the page constraint, we only present the results of six speakers with the most similar VISC contours in each vowel case.

Visual inspection of the vowel networks reveal the common articulatory strategies to produce /æ/ (Figure 1.A) and /ɑ/ (Figure 1.B). In general, the interrelation between the tongue blade and tongue dorsum is observable by the interaction between the kinematic variables, exhibiting a complex behavior in the production of each formant. For example, during the production of vowel /æ/ at points 3 and 4, tongue dorsum height (TDy) seems to affect the outcome of F1 via an indirect path with intermediate nodes TDx and TBy. Interestingly, when contrasting both vowels' networks it is apparent that in /æ/

networks connections between the tongue kinematic variables and both formants are present in all analysis points. However, networks of the vowel /ɑ/ show no connection between the tongue variables and the F1 at two points (3 and 4), suggesting that other articulatory variables may affect this formant at these points or that the articulatory movements measured here were not salient enough to surface a relationship.

A comparison between vowel networks and each speaker's networks (see Supplementary Figures S1 and S2 via https://doi.org/10.17605/OSF.IO/S7H2E) reveals more articulatory variability in the production of /ɑ/ than in /æ/. The Jaccard index (Table 1) indicates that almost all speakers had at least two points which were at least 50% similar to the networks of /æ/. The only exception was speaker 34ENM, who showed the highest similarity score (0.42) at the center analysis point. Conversely, speaker networks were less similar to the networks of /ɑ/, where the majority of indices were around 0.3. Surprisingly, networks of male speakers were more similar to the networks of this vowel than networks of female speakers. This result is reflected in the VISC contours as a greater variability in terms of curve shapes for female speakers.

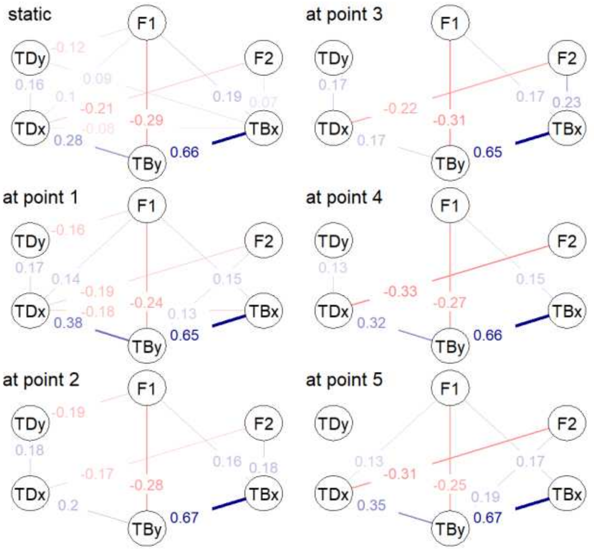
Table 1: *Jaccard similarity index of dynamic network comparison between the vowel and speaker networks.*

Speaker	1	2	3	4	5
/æ/					
07ENF	0.29	0.50	0.50	0.50	0.70
09ENF	0.54	0.55	0.43	0.38	0.40
21ENF	0.57	0.36	0.50	0.30	0.17
15ENM	0.70	0.42	0.55	0.30	0.45
33ENM	0.55	0.50	0.55	0.38	0.55
34ENM	0.36	0.38	0.42	0.31	0.13
/ɑ/					
07ENF	0.31	0.23	0.18	0.36	0.33
09ENF	0.27	0.38	0.27	0.33	0.36
21ENF	0.50	0.18	0.22	0.25	0.42
15ENM	0.30	0.33	0.56	0.50	0.40
33ENM	0.36	0.17	0.15	0.40	0.42
34ENM	0.42	0.42	0.30	0.44	0.33

Turning now to the comparison of speakers with similar VISC patterns, results demonstrated that there is also similarity between their networks. Speakers 09ENF and 21ENF had very similar VISC patterns for the vowel /æ/. This similarity was also manifested in their articulatory patterns where three points (1, 3 and 5) were more than 50% similar. Speaker 07ENF, who also had a similar VISC pattern, showed similar articulatory behavior only at two points: between 0.5 at point 1 and around 0.7 at points 4 and 5. The three male speakers with most similar VISC patterns also showed equivalent articulatory behavior. Speakers 15ENM and 34ENM were more similar in the three innermost points (ranging from 0.5 to 0.9). Speakers 33ENM and 34ENM showed articulatory similarity above 50% at points 2 (0.54) and 4 (0.77).

The results for speakers with similar /ɑ/ VISC patterns were also reproduced in the comparison between their articulatory strategies. Among the female speakers with the most similar patterns, 07ENF and 09ENF showed more than 55% similarity in the three innermost points. In this vowel case, the most dissimilar VISC contour was between speakers 07ENF and 21ENF, which was reflected in a similarity score below 0.42 for almost all points. Surprisingly though, their articulatory behavior at point 5 was 70% similar. Among the male speakers, 33ENM and 34ENM had the most similar VISC contours, while speaker 15ENM had the most idiosyncratic contour. The most similar speakers also showed more articulatory similarities in their dynamic networks, where points 2 and 3 were, respectively, 0.73 and 0.50 similar. Speaker 15ENM shared only one point with similarity around 0.50, the remaining points ranged between 0.46 and 0.25.

A – vowel networks for /æ/



B – vowel networks for /ɑ/

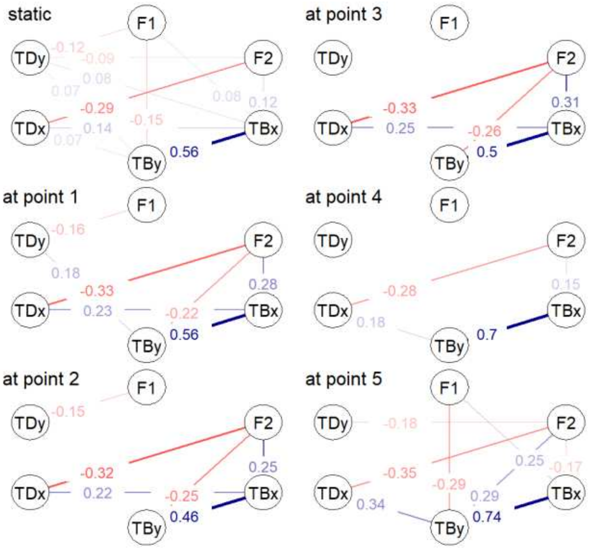


Figure 1: *Static and dynamic vowel networks (respectively labeled "static" and "at points 1 through 5") of the vowels /æ/ (A) and /ɑ/ (B). Nodes represent the acoustic (F1, F2) and kinematic variables (TBx, TBy, TDx, TDy). Red edges represent negative correlations and blue edges represent positive correlations. Only significant edges are displayed (alpha = 0.05).*

4. Discussion

In this study we set out to explore individual articulatory behaviors employing network analysis. We examined acoustic and articulatory data of the vowels /æ/ and /ɑ/ produced by native U.S. English speakers.

Our preliminary results revealed that /æ/, a vowel which relies on VISC to convey its phonemic identity, was produced by the six speakers presented here with less articulatory variability than the vowel /ɑ/, which is believed to rely less on VISC trajectories due to its relatively stable acoustic goal. This finding suggests that vowels requiring less precision in articulatory movements may result in a more idiosyncratic acoustic output, since speakers may be less restricted by linguistic constraints and consequently employ preferred articulatory strategies.

Furthermore, because male speakers showed more similar articulatory strategies than female speakers in vowel /ɑ/ networks, it seems plausible that, in addition to biological differences, sociophonetic discrepancies between male and female speakers could have played a role in the amount of variability in this vowel. However, we did not test this assumption, therefore, these results need to be interpreted with caution. Nevertheless, it has been reported that female speakers tend to vary more than male speakers in their articulatory strategies in the production of open vowels [21].

The differences between speakers are more intricate and offer no straightforward interpretation. Overall, our results suggest that speakers with similar VISC contours seem to have a similar kinematic profile. Inter-speaker similarity reflected the structure of networks, but it did not consider the signed weight of shared node-to-node connections. This result leads us to believe that similarity in formant contours may be partially related to the same node-to-node connections speakers share. For instance, at their most similar dynamic network (at point 1) of the vowel /æ/, speakers 09ENF and 21ENF shared nine out of 13 connections, however, from these nine, five connections had same-sign weight. The fact that important differences related to strength and direction of correlation cannot be factored in the Jaccard index alone, reflects the value of visualizing networks.

It is undeniable that there is no simple 1:1 mapping between the acoustic outcome and the articulatory strategy employed by each speaker in the production of these vowels. The structure of speakers' networks highlighted motor equivalences, the ability to produce a very similar acoustic outcome using different articulatory strategies. In task-dynamics [10], speech entails a succession of tasks to be completed, where a task represents the movement of articulators towards a physical target. In the context of the current study, tongue variables were coordinated to achieve a common goal, which was either the vowel /æ/ or /ɑ/. Theoretically, the variables employed in the production of these sounds should be the same, but the physical details of how the goal is achieved differs per individual. This difference is apparent in the structure of each speaker's network, which revealed that the organization of the same articulatory variables resulted in different strategies for the production of similar formant contours.

Moreover, networks revealed that some articulatory variables may indirectly affect the acoustic variables. Interpreting a network at the behavioral level reveals that each variable's action have implicit consequences for the outcome of every other variable in the network. Our results suggest that some articulatory variables may affect the outcome of formants indirectly. For instance, upon inspecting the static /æ/ networks of all female speakers, it is apparent that, among the most employed articulatory strategies, the height of tongue dorsum (TDy) seems to indirectly affect the outcome of F1. For speakers 07ENF and 21ENF, this variable is connected to F1 via TDx. For speakers 09ENF and 07ENF, this variable is connected to F1 via TDx and subsequently TBy. This suggests that for all three speakers an action by TDy could indirectly affect F1, however, the path between these variables is layered in characteristic ways. Further, even though these connections are shared between speakers, the weight and sign of the edges differ; e.g., for speaker 21ENF an increase in TDy leads to an increase in TDx, which increases F1, while for speaker 07ENF an increase in TDy leads to a decrease in TDx, which in turn increases the value of F1. Taken together, our results seem to indicate that articulatory behavior may also be layered in characteristic ways.

There are two major limitations of this study. First, the Jaccard similarity index does not take into consideration the weight and sign of the edges when computing similarity. However, it served as a starting point to characterize similarity quantitatively. Second, the small sample size in speaker dynamic networks did not allow a significance test to detect statistically significant edges. Future research should address these issues by including a larger sample size to allow significance testing in speaker dynamic networks, and by developing a similarity measure which also incorporates weight and sign in its computation. Notwithstanding these limitations, the present study is the first of its kind to use network analysis to explore idiosyncratic articulatory behaviors in vowel dynamics.

Speech behavior is a key factor in an array of fields from speech synthesis to biometric applications. Thus, examining kinematic data helps clarify the complex interaction between articulatory variables during speech production. Moreover, understanding speaker-specific articulatory behaviors and accounting for individual differences in speech production is fundamental to explain the inherent complexity of motor behavior.

5. Acknowledgements

This work was supported by the Swiss National Science Foundation (Grant #PZ00P1_193328 to LH). The EMA-MAE dataset is based on work supported by the National Science Foundation of the United States under Grant #IIS-1142826 to Marquette University, which support does not constitute an endorsement. We wish to thank Willemijn Heeren for valuable advice and insightful observations.

6. References

[1] T. Kitamura and M. Akagi, "Speaker Individualities in Speech Spectral Envelopes and Fundamental Frequency Contours," in *Speaker Classification II*, vol. 4441, C. Müller, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 157–176. doi: 10.1007/978-3-540-74122-0_14.

[2] X. Yang, J. B. Millar, and I. Macleod, "On the sources of inter- and intra- speaker variability in the acoustic dynamics of speech," *Proceedings of Fourth International Conference on Spoken Language Processing*. ICSLP '96, Oct. 1996, vol. 3, pp. 1792–1795 vol.3. doi: 10.1109/ICSLP.1996.607977.

[3] K. McDougall, "Speaker-specific formant dynamics: An experiment on Australian English /aI/," *International Journal of Speech, Language and the Law*, vol. 11, no. 1, pp. 103–130, Jun. 2004, doi: 10.1558/sll.2004.11.1.103.

[4] K. McDougall, "Dynamic features of speech and the characterization of speakers: towards a new approach using formant frequencies," *International Journal of Speech,*

Language and the Law, vol. 13, no. 1, pp. 89–126, Jun. 2006, doi: 10.1558/sll.2006.13.1.89.

[5] T. M. Nearey and P. F. Assmann, “Modeling the role of inherent spectral change in vowel identification,” *The Journal of the Acoustical Society of America*, vol. 80, no. 5, pp. 1297–1308, Nov. 1986, doi: 10.1121/1.394433.

[6] K. N. Stevens, “On the quantal nature of speech,” *Journal of Phonetics*, vol. 17, no. 1–2, pp. 3–45, Jan. 1989, doi: 10.1016/S0095-4470(19)31520-7.

[7] J. Perkell *et al.*, “Speech motor control: Acoustic goals, saturation effects, auditory feedback and internal models,” *Speech Communication*, vol. 22, no. 2–3, pp. 227–250, Aug. 1997, doi: 10.1016/S0167-6393(97)00026-5.

[8] K. Johnson, P. Ladefoged, and M. Lindau, “Individual differences in vowel production,” *The Journal of the Acoustical Society of America*, vol. 94, no. 2, pp. 701–714, Aug. 1993, doi: 10.1121/1.406887.

[9] Y. Ji *et al.*, “Speech Behavior Analysis by Articulatory Observations,” *Procedia Computer Science*, vol. 111, pp. 463–470, 2017, doi: 10.1016/j.procs.2017.06.048.

[10] E. L. Saltzman and K. G. Munhall, “A Dynamical Approach to Gestural Patterning in Speech Production,” *Ecological Psychology*, vol. 1, no. 4, pp. 333–382, Dec. 1989, doi: 10.1207/s15326969eco0104_2.

[11] M. Newman, *Networks*, vol. 1. Oxford University Press, 2018. doi: 10.1093/oso/9780198805090.001.0001.

[12] A. Ji, J. J. Berry, and M. T. Johnson, “The Electromagnetic Articulography Mandarin Accented English (EMA-MAE) corpus of acoustic and 3D articulatory kinematic data,” in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, May 2014, pp. 7719–7723. doi: 10.1109/ICASSP.2014.6855102.

[13] W. Labov, S. Ash, and C. Boberg, *The atlas of North American English: phonetics, phonology, and sound change: a multimedia reference tool*. Berlin ; New York: Mouton de Gruyter, 2006.

[14] C. P. Browman and L. Goldstein, “Articulatory gestures as phonological units,” *Phonology*, vol. 6, no. 2, pp. 201–251, 1989, doi: 10.1017/S0952675700001019.

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

[16] G. Schwartz, “The phonology of vowel VISC-osity – acoustic evidence and representational implications,” *Glossa: a Journal of General Linguistics*, vol. 6, no. 1, Mar. 2021, doi: 10.5334/gjgl.1182.

[17] R Core Team, *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, 2021. [Online]. Available: https://www.R-project.org/

[18] S. Epskamp, D. Borsboom, and E. I. Fried, “Estimating psychological networks and their accuracy: A tutorial paper,” *Behavior Research Methods*, vol. 50, no. 1, pp. 195–212, Feb. 2018, doi: 10.3758/s13428-017-0862-1.

[19] S. Epskamp and E. I. Fried, “A tutorial on regularized partial correlation networks.,” *Psychol Methods*, vol. 23, no. 4, pp. 617–634, Dec. 2018, doi: 10.1037/met0000167.

[20] P. Jaccard, “The Distribution of the Flora in the Alpine Zone.1,” *New Phytologist*, vol. 11, no. 2, pp. 37–50, 1912, doi: https://doi.org/10.1111/j.1469-8137.1912.tb05611.x.

[21] M. Weirich *et al.*, “Mumbling: Macho or Morphology?,” *Journal of Speech, Language, and Hearing Research*, vol. 59, no. 6, Dec. 2016, doi: 10.1044/2016_JSLHR-S-15-0040.