SPEECH STABILITY ACROSS TIME: EVIDENCE FROM NORWEGIAN VOWELS IN SPONTANEOUS SPEECH PRODUCTION

Peng Li[1], James Emil Flege[2], Clara D. Martin[3,4], Natalia Kartushina[1]

[1] Center for Multilingualism in Society across the Lifespan, Institute for Linguistics and Scandinavian Studies, University of Oslo, Norway
[2] University of Alabama at Birmingham, USA
[3] Basque Center on Cognition, Brain and Language, Spain
[4] Ikerbasque, Basque Foundation for Science, Spain
peng.li@iln.uio.no; jeflege@uab.edu; c.martin@bcbl.eu; natalia.kartushina@iln.uio.no

ABSTRACT

This study investigated the stability of speech sound production across time. Thirty-four monolingual Norwegian speakers participated in a three-session psycho-socio quiz experiment over three weeks, which was designed to collect spontaneous speech data. Participants had to answer questions orally by choosing either one of three plausible answers or one of two sentences for "I don't know". The "I don't know" sentences contained five target vowels /i, æ, ɑ, o, u/. Using the Pillai score, we analyzed the stability of vowel categories and the distinctness of vowel contrasts across tests. We found that, overall, the tested vowel categories and the three vowel contrasts /i-æ/, /ɑ-o/, and /o-u/ were stable over time. However, the backness-based contrasts /æ-ɑ/ and /u-i/ showed instability. Future research needs to examine the role of phonetic environment, repetition, and prosodic structure on sound variability.

Keywords: stability, vowel, Norwegian, spontaneous speech production, formant analysis

1. INTRODUCTION

Individuals produce speech sounds with considerable variability in both native and nonnative speech [1], which may arise from multiple sources such as dialect [2], the number of distinctive sounds in a given language (i.e., phoneme inventory) [3], and speech style (i.e., elicited vs. spontaneous speech) [4], among others. Recent research has established several lines to study the stability of sound production. Yet, it is unclear whether these differences are stable over time, systemic or sound specific, and whether they generalize to spontaneous speech.

First, very little previous research has evaluated segmental production is stable, that is, sought to determine if individuals produce vowels and consonants the same way in multiple test sessions. Heald and Nusbaum [5] asked American English speakers to produce the English vowels in isolation at three different timepoints on three different days. They found that although the mean values of F0 and F1 changed within one day, there was no significant change in the mean values of the acoustic measures between days. This study showed the stability of the production of vowels in *citation* forms between days. Later, Pierce et al. [6] had native English speakers read English sentences and produce /a/ in isolation for seven consecutive days with three timepoints per day. Their study revealed similar results to Heald and Nusbaum's study. It seems that the time of day had a significant influence on speech production, but speech production is stable on different days at the same time point, as demonstrated by the consistent production of vowels and consonants in repeated readings of scripted passages.

Second, individual differences in speech production can be attributed to other factors [7]. For instance, sound realization might reflect the precision of underlying phonemic perceptual representations [8]. Also, speakers' personalities (e.g., introverted vs. extroverted), mental status (e.g., depression and anxiety), and living habits (e.g., alcohol consumption) may affect the acoustic variability of speech [6]. Research with bilingual speakers found that those who scored higher in L1 intelligibility tended to have more intelligible L2 speech [9]. Likewise, precision in L1 vowel production can affect accuracy in L2 vowel production [10]. Although appealing, these results need to be taken with care, as it is not clear whether these L1 differences in precision are stable, subject-intrinsic, systematic, reflecting higher-order representations, or whether these are momentary modulations in speech production.

Most previous studies measured intra-speaker speech variability using scripted speech production tasks, like paragraph-reading, sentence-reading, and even repetitive production of single vowels [5], [6], focusing participants' attention on speech *per se*. Compared to scripted speech, which often involves a deliberate and careful speaking style, spontaneous and more natural speech production may exhibit greater variability, such as vowel reduction, which can affect the acoustic properties of vowels [11]. A study on the Mixtec language confirmed that vowels

in spontaneous speech are shorter, more contextually assimilated, and show a less dispersed vowel space than in scripted speech [4]. Therefore, caution should be exercised when extrapolating results obtained from elicited speech to spontaneous speech.

Furthermore, the existing methods for eliciting speech samples are lab-based, and the subjects are fully aware that their speech will be acoustically scrutinized. These methods may lead to unintended variations in speech sounds, as participants may deliberately alter their pronunciation in response to the experimental conditions [12]. Therefore, a new technique called Characteristic Speech Production (CSP) was developed to elicit data in which participants concentrate on the meaning of their responses to relevant questions rather than on the pronunciation of specific sounds [13].

This study is the first to use the CSP sampling method to investigate the stability of speech production over multiple days. We developed an online psycho-socio quiz that was administered three times over three weeks to collect samples of spontaneous speech in Norwegian, which is the participants' native language. The target speech sounds were five Norwegian vowels /i, æ, ɑ, o, u/. Note that Norwegian has more vowels, but for the current study, we selected the most frequent ones, three of which were corner vowels /i, æ, u/. We assessed the stability of speech production by analyzing the consistency of vowel categories and the distinctiveness of vowel contrasts using formant analyses. Based on previous research [5], [6] and because the target vowels are distinctive phonemic categories in Norwegian [14], we considered the following hypotheses:

- H1: Vowel categories are stable over time: Norwegian speakers would produce each target vowel in a consistent way across time.
- H2: Distinctness of vowel contrasts is stable over time: Norwegian speakers would show stable vowel pair distinctness across time.

2. METHODS

2.1. Participants

We recruited 34 Norwegian native speakers to participate in this study. Nine participants were excluded: Eight completed only one test, and one returned low-quality recordings across all three tests. The remaining 25 participants (aged 18 – 39, female = 13, male = 12) were analyzed. The participants reported no documented speech disorders or hearing impairment. They all signed a consent form approved by the Norwegian Centre for Research Data. Each of them received a small amount of monetary compensation for their time.

2.2. The psycho-socio quiz

We created 150 questions (50 per testing test) with various degrees of difficulty (see Table 1 for an example):

- *Control questions* (n = 9) required common sense knowledge. E.g., Give the result for 5 +10.
- *Easy questions* (n = 36) required some background knowledge. E.g., Which of these foods is NOT toxic to dogs?
- *Average questions* (n = 45) required basic knowledge in a certain field. E.g., Which is the oldest Norwegian political party?
- *Hard questions* (n = 30) required specific knowledge in a certain field. E.g., What is the atomic number of Uranus?
- *No-concrete answer questions* (n = 30) were questions with no objective or reasonable answer. E.g., Who was Norway's best artist?

Table 1: A trial example from the psycho-socio quiz (English translation is provided for the reader but was not included in the quiz).

	Hvem av disse har verdensrekorden for lengst hår? **"Who among these holds the world record for longest hair?"**
1	*Kate Moss*
2	*Xie Qiuping*
3	*Linda Evangelist*
4	*Jeg vet **i**kke sv**a**ret på spørsmålet h**e**r, men jeg tr**o**r n**o**en **a**ndre k**a**n v**i**te det.* "I don't know the answer to the question here, but I think someone else might know it."
5	***I**ngen k**a**n sv**a**re p**å** dette h**e**r, det **e**r j**o** n**o**e s**o**m **e**r umul**i**g **å** s**i*** "No one can answer this here; it is something that is impossible to say."

On each question, participants were presented with five possible answer choices, including three plausible answers and two long sentences for 'I don't know' (options 4-5 in Table 1). The two long sentences were the target sentences, which were identical across all trials and tests, and contained the target vowels (/i, æ, ɑ, o, u/, indicated in bold in Table 1). Each vowel appeared at least four times over the two sentences. Hard and vague questions triggered as many target sentences as possible to ensure an adequate number of target tokens.

Participants did the experiment on an online survey platform Prolific (https://www.prolific.co/), in a quiet place that they were familiar with (e.g., the living room). The productions were recorded three times (T1, T2, T3) at the same hour of the day, on the same day of the week, over a span of three weeks. At each trial, participants first saw the question on the screen, followed by five options. They had to say one

of the five options/prompts to answer each question. On each day, participants answered 50 unique questions in random order.

2.3. Data coding and analyses

2.3.1. Data screening

We obtained 3,750 recordings from the three tests, each containing one sentence. A Norwegian-speaking research assistant identified 1,557 target sentences. The rest of the utterances (e.g., answers like "Kate Moss" to the example question in Table 1) were discarded. We then excluded 65 recordings due to heavy background noise or poor audio quality. Finally, the remaining 1,492 recordings were used for the acoustic analyses.

2.3.2. Acoustic analyses

We segmented our corpus into words and phones with Montreal Forced Aligner [15]. Since the boundary locations may be inaccurate in spontaneous speech due to deletion and non-canonical pronunciation patterns, we manually checked and corrected all the labels in Praat [16]. In total, we obtained 15,635 tokens. Then, we applied a Praat script bundle to extract the mean values (Hertz) of the first four formants (i.e., F1-F4) from the middle third part of each target vowel [17]. The script bundle extracts formants using the dynamic seeding method [18], taking the vowel formant of Norwegian in previous research as the reference [19], which can minimize errors in formant extraction.

2.3.3. Statistical analyses

To minimize the intrinsic differences across speakers, we converted the Hertz values of F1, F2, and F3 to Bark (Z1, Z2, Z3, respectively). Then, the vowel height was represented by Z3-Z1, and the vowel frontness was represented by Z3-Z2. Fig 1 plots the bark normalized F1 and F2 values with the ellipses representing one standard deviation (SD) for each of the five Norwegian vowels across the three tests.

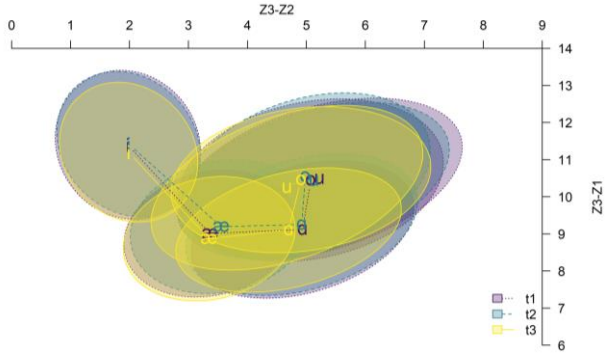

Fig. 1: Five Norwegian vowels plotted by bark normalized vowel height (Z3-Z1) and frontness (Z3-Z2) produced at three tests. The mean is marked by vowel label, and the ellipses show 1 SD distance from the mean.

To evaluate the degree of distinctness between the vowels, we calculated the Pillai score (ranging from 0-1) using MANOVA test [20], taking Z3-Z2 and Z3-Z1 as the dependent variables for each participant.

- *Vowel category stability across tests*: we compared the degree of distinctness of each target vowel within speakers between T1-T2 and between T1-T3 (T1 being the baseline), which yielded 250 Pillai scores (25 participants × 5 vowels × 2 test comparisons).
- *Distinctness of vowel contrasts across tests*: we compared with-in speakers five pairs of vowels (i.e., /i-æ/, /æ-ɑ/, /ɑ-o/, /o-u/, and /u-i/) at each test, yielding 375 Pillai scores (25 participants × 5 vowel pairs × 3 tests).

To address the two hypotheses, we built two generalized linear mixed effects models using the *glmmTMB* package [21] in R [22]. Both models took the Pillai Score as the dependent variable. Model 1 included vowel (/i, æ, ɑ, o, u/), test comparison (T1-T2, and T1-T3), and their interaction as the fixed effects. Model 2 involved test (T1, T2, T3), vowel pair (/i-æ/, /æ-ɑ/, /ɑ-o/, /o-u/, and /u-i/), and their interaction as fixed factors. Participant was added as random intercept in both models. We selected the best random slopes in *buildmer* package [23], which involved a by-participant random slope of test comparison for Model 1 and a by-participant random slope of vowel pair for Model 2. We calculated the significance of the fixed factors with *Anova()* function from the *car* package [24] and tested post-hoc comparisons using *emmeans* package [25] with significance adjusted by false discovery rate for multiple comparisons. All analyses have been preregistered on the OSF (https://osf.io/d3w2k).

3. RESULTS

3.1 The stability of vowel categories over time

Model 1 revealed a significant main effect of vowel, $\chi^2 = 15.93$, $p = .003$. Post-hoc comparisons showed that /u/ had a significantly higher Pillai score than /i/, $t = 3.27$, $p = .011$, /æ/, $t = 2.92$, $p = .032$, /ɑ/, $t = 3.11$, $p = .018$, and /o/, $t = 3.26$, $p = .011$. However, all the vowels showed low Pillai scores (Fig. 2), indicating a considerable within-category overlap over time. More importantly, there was no significant main effect of test comparison, $\chi^2 = 0.21$, $p = .651$, or interaction of Test comparison × Vowel, $\chi^2 = 2.28$, $p = .684$. Hence, the vowels in general were stable over time.

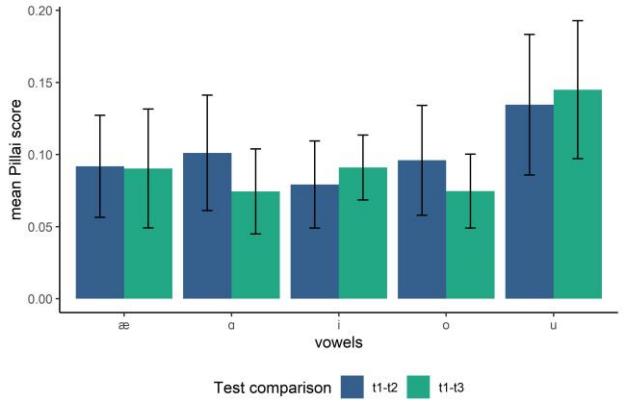

Fig. 2: Mean Pillai score (0 = merge, 1 = separation) of each target vowel compared between T1 and T2, and between T1 and T3. Error bars mark ± 2 *SE*.

3.2 The stability of distinctness in vowel contrasts over time

Model 2 revealed a significant main effect of vowel pair, $\chi^2 = 567.73$, $p < .001$, and a two-way interaction of Vowel pair × Test, $\chi^2 = 21.71$, $p = .005$, suggesting that the degree of distinctness between each of the vowel pairs varied across tests. The Pillai scores of /æ-ɑ/ decreased from T1 to T2, $t = 3.33$, $p = .003$, and from T1 to T3, $t = 3.66$, $p = .001$; the Pillai scores of /u-i/ decreased from T1 to T3, $t = 2.58$, $p = .028$. That is, their distinctness was not stable over time. The other three contrasts showed no effect of test. Hence their distinctness was stable over time. See Fig. 3 for the descriptive data.

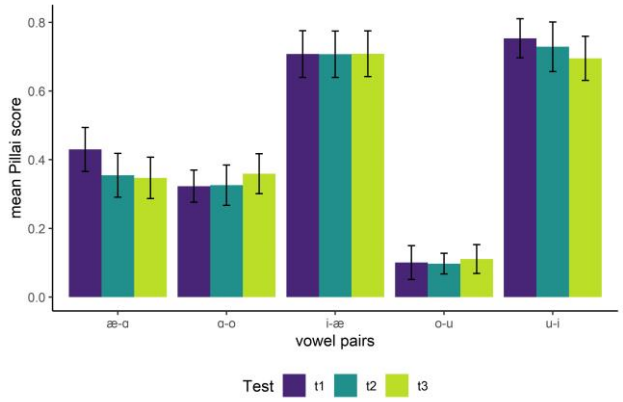

Fig. 3: Mean Pillai score (0 = merge, 1 = separation) compared between each vowel pair across tests. Error bars mark ± 2 *SE*.

4. DISCUSSION AND CONCLUSION

This study investigated the stability in Norwegian vowel production using an innovative speech-eliciting method—Characteristic Speech Production, which instructed the participants to answer meaningful questions in the target language (Norwegian) to elicit the natural production of the target sounds (/i, æ, ɑ, o, u/) without focusing their attention to speech. The test was administered three times over three weeks. The production stability was measured by Pillai score, derived from the formant analyses of the target vowels. We hypothesized that (H1) Norwegian speakers would produce each target vowel in a consistent way over time and (H2) Norwegian speakers would show stable distinctness between the vowel contrasts over time.

Regarding H1, the vowels were stable over time, as revealed by low Pillai scores, suggesting that Norwegian speakers produced the vowels consistently and with overlap across the three tests. Although the five target vowels were produced stably across the three tests, it appeared that, overall, /u/ was more dispersed (with higher Pillai scores) than the other vowels. Because three out of the four words containing /u/ had either a schwa following the target vowel /u/ (i.e., *noe* [nuə], *noen* [nuən]) or a palatal approximant preceding the target vowel (i.e., *jo* [ju]), the phonetic environment may have centralized /u/ and the centralization became more evident through multiple repetitions. In sum, our data supported H1 in terms of stability in individual vowel production.

With respect to H2, the stability of vowel category distinctness was subject to the concrete phonemic contrasts with the two high vowels /u-i/ and the two low vowels /æ-ɑ/ showing closer rapprochement at the follow-up tests compared to the first test.

We interpret our results as follows. First, the repetition across T1 to T3 may have reduced the articulatory efforts for the vowels contrasting mainly in backness, which was reflected by the centralization between the front and back vowel pairs /u-i/ and /æ-ɑ/. Since the two pairs are corner vowels and show clear distinctness, speakers might be less careful in maintaining the distinctness over multiple repetitions. Second, the prosody-segment interaction may have played a role. In spontaneous speech elicitation, the target vowels are inevitably embedded in functional words, which are often unstressed and less prominent on the intonational level. The prosodic structure may have led to vowel reduction, which was a resource of centralization observed in our data. Future research may try to minimize the influence of prosody on speech sound realization.

In conclusion, the current study showed that individual speech sound production was stable in phoneme categories over time. However, some of the category distinctness were reduced. The variability in sound production may be affected by phonetic environment, repetition, and prosodic structure. Future studies in L1 and L2 speech production may consider these factors in the research design.

5. ACKNOWLEDGEMENT

This study was supported by the Research Council of Norway through its Centres of Excellence funding scheme [223265]. CDM was supported by the Spanish Ministry of Economy and Competitiveness [PID2020-113926GB-I00], the Basque Government [PIBA18-29], and funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program [819093]. We thank Silje Robberstad, Arun Singh, Johannes Sletten, and Sander Vassanyi for their assistance in data collection and analysis.

6. REFERENCES

[1] X. Xie and T. F. Jaeger, "Comparing non-native and native speech: Are L2 productions more variable?," *J. Acoust. Soc. Am.*, vol. 147, no. 5, pp. 3322–3347, May 2020, doi: 10.1121/10.0001141.

[2] P. Foulkes and G. Docherty, "The social life of phonetics and phonology," *J. Phon.*, vol. 34, no. 4, pp. 409–438, Oct. 2006, doi: 10.1016/j.wocn.2005.08.002.

[3] S. Y. Manuel, "The role of contrast in limiting vowel-to-vowel coarticulation in different languages," *J. Acoust. Soc. Am.*, vol. 88, no. 3, pp. 1286–1298, Sep. 1990, doi: 10.1121/1.399705.

[4] C. DiCanio, H. Nam, J. D. Amith, R. C. García, and D. H. Whalen, "Vowel variability in elicited versus spontaneous speech: Evidence from Mixtec," *J. Phon.*, vol. 48, pp. 45–59, Jan. 2015, doi: 10.1016/j.wocn.2014.10.003.

[5] S. L. M. Heald and H. C. Nusbaum, "Variability in vowel production within and between days," *PLOS ONE*, vol. 10, no. 9, p. e0136791, Sep. 2015, doi: 10.1371/journal.pone.0136791.

[6] J. Pierce, K. Tanner, R. Merrill, L. Shnowske, and N. Roy, "Acoustic variability in the healthy female voice within and across days: How much and why?," *J. Speech Lang. Hear. Res.*, vol. 64, pp. 1–17, Jul. 2021, doi: 10.1044/2021_JSLHR-21-00018.

[7] C. C. Heffner and E. B. Myers, "Individual Differences in Phonetic Plasticity Across Native and Nonnative Contexts," *J. Speech Lang. Hear. Res.*, vol. 64, no. 10, pp. 3720–3733, Oct. 2021, doi: 10.1044/2021_JSLHR-21-00004.

[8] J. S. Perkell *et al.*, "The distinctness of speakers' productions of vowel contrasts is related to their discrimination of the contrasts," *J. Acoust. Soc. Am.*, vol. 116, no. 4, pp. 2338–2344, Oct. 2004, doi: 10.1121/1.1787524.

[9] A. R. Bradlow, M. Blasingame, and K. Lee, "Language-independent talker-specificity in bilingual speech intelligibility: Individual traits persist across first-language and second-language speech," *Lab. Phonol. J. Assoc. Lab. Phonol.*, vol. 9, no. 1, p. 17, Oct. 2018, doi: 10.5334/labphon.137.

[10] N. Kartushina and U. H. Frauenfelder, "On the effects of L2 perception and of individual differences in L1 production on L2 pronunciation," *Front. Psychol.*, vol. 5, 2014, doi: 10.3389/fpsyg.2014.01246.

[11] C. Meunier and R. Espesser, "Vowel reduction in conversational speech in French: The role of lexical factors," *J. Phon.*, vol. 39, no. 3, pp. 271–278, Jul. 2011, doi: 10.1016/j.wocn.2010.11.008.

[12] J. E. Flege, "New methods for second language (L2) speech research," in *Second language speech learning, Theoretical and empirical progress*, R. Wayland, Ed., Cambridge: Cambridge University Press, 2021, pp. 119–156.

[13] J. E. Flege, "A distributional learning account of L2 speech learning," presented at the 10th International Symposium on the Acquisition of Second Language Speech, Barcelona, Spain, 2022.

[14] G. Kristoffersen, *The Phonology of Norwegian*. Oxford: Oxford University Press, 2000.

[15] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi," in *Proc. Interspeech 2017*, 2017, pp. 498–502. doi: 10.21437/Interspeech.2017-1386.

[16] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [computer program]." 2020. [Online]. Available: http://www.praat.org

[17] M. Zhang, "Dynamic seeding scripts bundle," 2022. https://github.com/ZenMule/dynamic_seeding_scripts

[18] W.-R. Chen, D. H. Whalen, and C. H. Shadle, "F0-induced formant measurement errors result in biased variabilities," *J. Acoust. Soc. Am.*, vol. 145, no. 5, pp. EL360–EL366, May 2019, doi: 10.1121/1.5103195.

[19] A. Rosslund, J. Mayor, G. Óturai, and N. Kartushina, "Parents' hyper-pitch and low vowel category variability in infant-directed speech are associated with 18-month-old toddlers' expressive vocabulary," Dec. 2022, doi: 10.34842/2022.0547.

[20] F. Santiago, "L'accentuation contribue-t-elle à l'acquisition du contraste arrondi vs non-arrondi des voyelles orales en français langue étrangère ?," *Etudes Linguist. Appliquée*, pp. 74–90, 2021.

[21] M. E. Brooks *et al.*, "glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling," *R J.*, vol. 9, no. 2, pp. 378–400, 2017.

[22] R Core Team, "R: A language and environment for statistical computing." R Foundation for Statistical Computing, Vienna, Austria, 2014. [Online]. Available: http://www.r-project.org/

[23] C. C. Voeten, "buildmer: Stepwise Elimination and Term Reordering for Mixed-Effects Regression." 2021. https://cran.r-project.org/package=buildmer

[24] J. Fox and S. Weisberg, *An {R} Companion to Applied Regression*. Thousand Oaks {CA}: Sage, 2019. https://socialsciences.mcmaster.ca/jfox/Books/Companion/

[25] R. Lenth, H. Singmann, J. Love, P. Buerkner, and M. Herve, "Emmeans: Estimated marginal means, Aka Least-Squares means." R package 1.5.1, 2020. https://cran.r-project.org/package=emmeans