Assessing the link between speech perception and production through individual differences

Matthias K. Franken[1,2], James M. McQueen[3,1,2], Peter Hagoort[1,2] & Daniel J. Acheson[1,2]

(1) Radboud University, Donders Institute for Brain, Behaviour and Cognition, (2) Max Planck Institute for Psycholinguistics, (3) Radboud University, Behavioural Science Institute
m.franken@donders.ru.nl

ABSTRACT

This study aims to test a prediction of recent theoretical frameworks in speech motor control: if speech production targets are specified in auditory terms, people with better auditory acuity should have more precise speech targets.

To investigate this, we had participants perform speech perception and production tasks in a counterbalanced order. To assess speech perception acuity, we used an adaptive speech discrimination task. To assess variability in speech production, participants performed a pseudo-word reading task; formant values were measured for each recording. We predicted that speech production variability to correlate inversely with discrimination performance. The results suggest that people do vary in their production and perceptual abilities, and that better discriminators have more distinctive vowel production targets, confirming our prediction. This study highlights the importance of individual differences in the study of speech motor control, and sheds light on speech production-perception interaction.

Keywords: Speech production, Speech perception, Individual variability, Phonemic goals.

1. INTRODUCTION

Several lines of research have shown that speech production and speech perception are not independent processes, but interact in complicated ways. Investigations of these perception-production interactions can largely be placed in two categories. The first type focuses on short-term effects of perception on production. For example, when a speaker's auditory feedback is manipulated or distorted, his or her speech production is affected [3, 6, 15]. These studies have shown that although auditory feedback is not strictly necessary for regular speech production [9], when feedback is distorted, it affects speech production.

The second line of research focuses on longer-term interactions between speech production and perception, usually by studying correlations between the two. Here, the guiding hypothesis is that if production and perception interact on a daily basis, this will lead to co-variation across individuals. One example is a study [12] where the authors investigated correlations between acoustic measures taken on listeners' perceptual prototypes for a given speech category and on their average production of members of that category. The authors found that people whose perceptual prototype had longer voice onset time (VOT) also tended to produce these consonants with longer VOT.

Another example concerns studies that showed a correlation between auditory acuity and vowel production [13, 14]. In these studies, participants carried out two tasks, (1) a discrimination task on a vowel continuum and (2) a reading task. The results showed that participants that were better at the discrimination task produced vowels more consistently (less within-phoneme variability) but spaced further apart in vowel space (larger between-phoneme acoustic distance). These ideas are largely in line with several models of speech motor control [7, 16]. The authors interpret their findings as follows: better auditory acuity would be reflective of more precise speech targets (e.g., smaller goal regions in acoustic space), which in turn would lead to more consistent speech production, as a smaller goal region would reject non-prototypical productions more quickly as 'speech errors'. A related study is reported by Villacorta et al. [17], where they show that people with higher auditory acuity show stronger compensations in an altered auditory feedback paradigm.

In the present study, we addressed whether auditory acuity as measured by a speech discrimination task would be associated with individual variability in vowel productions, in an attempt to replicate a previously reported study [14].

2. METHODS

2.1. Subjects

Forty healthy volunteers (age: M = 20, SD = 2.2; 24 females) participated after providing written informed consent in accordance with the Declaration of Helsinki and the local ethics board committee (CMO region Arnhem / Nijmegen). All participants

had normal hearing, were native speakers of Dutch and had no history of speech and/or language pathology.

2.2 Stimuli

For the discrimination task, 2 speech continua were created based on recordings of the pseudowords *skef* and *skaf*, spoken by a male native Dutch speaker. From these recordings, 2 continua (/skɛf/-/skɪf/ and /skɑf/-/skɔf/) were made by manipulating F1 and F2 values. First, the vowel was excised from the recordings. Using Burg's linear predictive coding framework (LPC), a filter model was obtained by estimating 5 formants between 0 and 5000Hz. A source model was obtained using 8 prediction coefficients. A number of filter models were created by changing F1/F2 values in a stepwise manner, and the endpoints of the continua were based on the average F1 and F2 values for a male Dutch speaker [1]. For the *skaf-skof* continuum, 1001 steps were used, each one having a change of -0.176Hz in F1 and -0.351Hz in F2. For the *skef-skif* continuum, 543 steps were created, so the Euclidian distance between successive steps was similar to the first continuum (F1 change was -0.210Hz, F2 change was 0.332Hz). These filter models were combined with the source model. The vowels were lowpass-filtered at 2000Hz and combined with the band-pass filtered original signal (2000Hz-6000Hz). All vowels were manipulated so their average intensity matched that of the original sounds.
For the reading task, pseudowords were created using a $C_1V_1C_1C_1V_1C_2$ structure, where C_1 could be either one of /k/, /p/ or /t/, V_1 either one of /ɛ/, /ɪ/, /ɑ/ or /ɔ/, and C_2 either one of /p/, /t/, /k/, /f/, /s/ or /x/. Using all possible combinations resulted in 72 pseudowords (e.g., *kekkef, poppos*).

2.3 Procedure

The experiment consisted of two tasks, which were administered in counter-balanced order within a single session.
The discrimination task consisted of a 4-interval 2-alternative forced choice task [4] with a staircase procedure using the weighted up-down procedure [8, 10]. On every trial subjects heard 4 auditory stimuli, among which three standards and one deviant stimulus. The standard stimuli were always one extreme of the continuum (*skef* for the *skef-skif* continuum, *skaf* for the *skaf-skof* continuum), while the deviant stimulus varied on a trial-by-trial basis. The deviant stimulus occurred in position 2 or 3, and the participant was instructed to push the left button when he or she thought the deviant was the second stimulus, and to push the right button if he or she thought it was the third. If the participant responded correctly, the difference between the standard and the deviant in the next trial was decreased, otherwise it was increased.
The discrimination task consisted of 4 blocks, which alternated between continua. Every block started with a fairly large interval (250 continuum steps or Euclidian distance in F1xF2 space of around 98.2ΔHz between standard and deviant stimulus). 'Reversal' trials were trials were subjects gave a correct response after a previous incorrect trial, or vice versa. The block ended after a total of 20 reversal trials. The amount of change in the interval size from trial to trial was initially large (a decrease of 25 steps after a correct trial, an increase of 75 after an incorrect trial), and became smaller after the second reversal trial of a block (a decrease of 10 after a correct trial, an increase of 30 after an incorrect trial). Because the increase in interval size after an incorrect trial was always three times the decrease of the interval size after a correct trial, the interval size should theoretically converge to a threshold interval size where people would give a correct answer on 75% of the cases [8].
The reading task was a simple pseudoword reading task. Subjects were instructed to read aloud the pseudowords that appeared on the screen, while trying to maintain a constant, normal volume and making sure stress is on the second syllable (which was also printed in capitals). Subjects were positioned about 30 cm from the microphone and asked to try to keep this distance throughout. The task consisted of four blocks, each of which presented all 72 pseudowords in randomized order.

2.4 Analysis

2.4.1 Perception

For the results from the discrimination task, we calculated a threshold value per block, by averaging the interval sizes for the last 16 reversal trials. Subsequently, we took the minimal threshold per continuum for each subject. As another measure of discrimination performance, we quantified the consistency between blocks of the same continuum in the following way: we created a linear mixed effects model, with Block and Continuum as fixed effects, Subject as a random effect (with random slopes for Block and Continuum) and the calculated thresholds as dependent variables. The absolute value of the random slopes for Block were taken as a measure of between-block consistency. Finally, we calculated "discrimination score" by multiplying this between-block consistency by the minimal threshold value.

Furthermore, we carried out a correlation analysis between the minimal threshold and between-block consistency measures, in order to characterize the relationship between these two measures.

2.4.2 Production

For all recordings, the beginning and ending of the vowel in the second syllable, which always carried stress, was manually determined. Then the duration and formant values were extracted. Formant values were calculated by averaging over a 400ms time window at the center of the vowel. 5 formants were estimated between 0 and 5kHz (males) or 5.5kHz (females) using a Burg algorithm in Praat [2].
In order to capture subjects' production variability, two different measures were taken. The first was vowel dispersion, or the average Euclidian distance from a vowel token to the centroid for that phoneme. This was calculated per vowel, and the results were averaged across vowels within subjects. The second measure was average vowel spacing (AVS), which was the average Euclidian distance between the centroids of all possible vowel pairs for a given subjects. Both dispersion and AVS were calculated in F1xF2 space as well as in F1xF2xF3 space.
Additionally, we did similar analyses with mel-frequency cepstral coefficients (MFCCs), which are psycho-acoustically motivated and often used to represent speech in automatic speech recognition (ASR) systems [5]. This way, dispersion was quantified as the mean Euclidian distance to the centroid in 12-dimensional space (defined by 12 MFCCs), and AVS as the average pairwise distance between vowel centroids in the 12-dimensional space.

2.4.3 Perception vs. production

In order to assess the correlations between perception and production variability, multiple correlation analyses were carried out between the perception and production measures.

3. RESULTS

3.1 Perception

Test-retest correlations between the (log) thresholds for the discrimination task on each continuum were significant, showing that participants who were better in discrimination on the first block also did better on the second block (/ɪ/-/ɛ/: $r(38) = 0.63$, $p < 0.001$; /ɑ/-/ɔ/: $r(38) = 0.59$, $p < 0.001$). The correlations between minimal threshold and between-block consistency (see methods) were positive, although not significant (/ɪ/-/ɛ/: $r(37) = 0.237$, $p = 0.146$; $r(37) = 0.233$, $p = 0.153$). This suggests that participants who were more consistent in their discrimination score across the blocks, also performed better. In the remainder of the analyses, we therefore used a 'discrimination score' to take into account both consistency and best (lowest) threshold (see Table 1). A lower discrimination score indicates a lower minimal threshold or higher consistency (i.e., better performance).

Table 1: Overview of Discrimination Score, Dispersion and Average Vowel Spacing (both MFCC-based).

	Score (/ɪ/-/ɛ/)	Score (/ɑ/-/ɔ/)	Dispersion	AVS
Mean	4.37	4.88	4.48	5.46
SD	1.25	1.21	0.17	0.12
Min.	1.16	1.76	4.24	4.86
Max.	7.13	7.40	5.36	5.63

3.2 Production

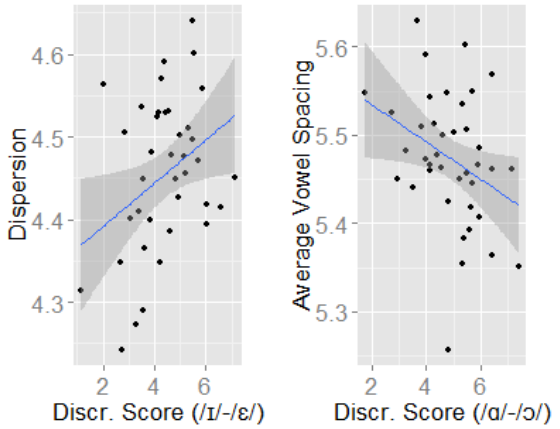
Vowel dispersion and average vowel spacing (AVS) were calculated in F1xF2 space. Dispersion (log-transformed) was found to be marginally correlated with AVS ($r(37) = 0.315$, $p = 0.051$), i.e. participants who produced vowels spaced further apart in vowel space also produced them with more within-phoneme variability. However, the correlation disappeared when the same analysis was done in F1xF2xF3 space ($r(38) = 0.006$, n.s.).

3.3 Perception vs. production

Next, we investigated whether the perception and production measures described above were correlated with each other.
We found small positive, but insignificant, correlation coefficients when correlating discrimination score with (log) dispersion (/ɪ/-/ɛ/: $r(37) = 0.159$, n.s.; /ɑ/-/ɔ/: $r(37) = 0.261$, n.s.), suggesting no association between discrimination score and speech variability. Negative correlation coefficients were found when correlating the discrimination score with AVS (/ɪ/-/ɛ/: $r(37) = -0.211$, n.s.; /ɑ/-/ɔ/: $r(37) = -0.194$, n.s.), although again they were not significant. Note however, that the direction of the trends is in line with our expectations.
One could argue that taking the formant values to represent vowel spectra is a sub-optimal characterization of the speech signal and quite distinct from what speakers actually do [11]. As such, we conducted similar analyses but quantified the production variability using MFCCs to represent the vowel spectral features (Table 1, Figure 1). Here, correlations between dispersion and discrimination score were stronger, and significant for the *skif-skef* continuum (/ɪ/-/ɛ/: $r(37) = 0.354$, $p = 0.027$; /ɑ/-/ɔ/:

$r(37) = 0.251$, $p = 0.124$). When looking at AVS with MFCCs, we also found stronger correlations, this time only significant for the *skaf-skof* continuum (/ɪ/-/ɛ/: $r(37) = -0.238$, $p = 0.145$; /ɑ/-/ɔ/: $r(37) = -0.339$, $p = 0.035$).

Figure 1: MFCC-based dispersion/average vowel spacing as a function of discrimination score for the /ɪ/-/ɛ/, /ɑ/-/ɔ/ continuum.



4. DISCUSSION

In this study, we investigated speaker's variability in vowel discrimination and vowel production. The discrimination test results revealed that although test-retest correlations were significant, they were rather small (around 0.60), indicating that individuals' performance varied between blocks of the same task. This may reflect changes in attention or a learning effect as they get used to the task. We distinguish between two measures of discrimination performance. The first we called minimal threshold, referring to the lowest threshold participants reached (irrespective of in which block). The second measure refers to participants' consistency across blocks, indicating how consistent participants scored across blocks. Although both measures did not correlate significantly, a trend was visible, showing that participants who were more consistent also reached better performance.

Production variability was quantified as dispersion (i.e., within-phoneme variability) or average vowel spacing (AVS). These measures correlated positively with each other, indicating that people tend to 'fill' their vowel spaces: if the produced vowels are spaced further apart, they also show more within-vowel variability.

The main research question of this study, however, asked whether the variability in speech production and speech discrimination are correlated. Although analyses with formant values (similar to [14]) did not reveal significant correlations, the direction of the trends seemed to confirm Perkell et al.'s results [14]: better discriminators produce vowels spaced further apart in vowel space, and with less within-vowel variability. Additional analyses with a more realistic representation of the vowel spectra based on MFCCs showed significant correlations between (1) discrimination score and dispersion for /ɪ/-/ɛ/ and (2) discrimination score and AVS for /ɑ/-/ɔ/, confirming the trend in the formant-based analyses.

These results can be interpreted along the lines of several recent influential models of speech production [7, 16]. Speakers with higher auditory acuity supposedly have more precise auditory representations of speech sounds. Higher precision of auditory representations can be visualized as smaller goal areas in acoustic space. When producing these speech sounds, these speakers will more quickly recognize an outlying production as a speech error and therefore self-repair. Over time, this would lead to less variable productions.

This interpretation should be taken with a grain of salt, given that the formant-based analyses did not reach significance, and the MFCC-based analyses did so only partially. This may be surprising, given the rather strong correlations reported in Perkell et al. [14]. This difference in outcomes may be due to a number of factors: (1) our materials may have led to less variability, given that we had only stressed vowels, in contrast to the materials in [14], which included vowels in both stressed and unstressed positions. Other differences with [14] also exist: we used pseudoword reading in isolation, whereas [14] embedded the pseudowords in a sentence. With respect to the discrimination task, [14] performed discrimination around individually-specified phoneme boundaries, whereas we performed the discrimination task at the phoneme center. Aside from the methodological differences with Perkell et al., we can assume that single-(non)-word production in the laboratory may show reduced variability compared to more natural language use, causing reduced effect sizes.

In conclusion, this study highlights the importance of individual variability when investigating speech perception and production. Individuals show variability in their speech discrimination ability, as well is in their speech production, both in terms of between-phoneme distance as well as within-phoneme variability. Interestingly, some of the results presented here suggest perceptual and production variability are associated: individuals with better auditory acuity seem to produce more distinct phonemes (i.e., spaced further apart and less variable). This is reflective of a long-term intricate interaction between speech production and perception.

5. REFERENCES

[1] Adank, P., van Hout, R., & Smits, R. (2004). An acoustic description of the vowels of Northern and Southern Standard Dutch. *Journal of the Acoustical Society of America*, *116*, 1729–1738.

[2] Boersma, P., & Weenink, D. (2013). Praat: doing phonetics by computer [Computer Program]. Retrieved from http://www.praat.org

[3] Fairbanks, G., & Guttman, N. (1958). Effects of Delayed Auditory-Feedback Upon Articulation. *Journal of Speech and Hearing Research*, *1*, 12–22.

[4] Gerrits, E., & Schouten, M. E. H. (2004). Categorical perception depends on the discrimination task. *Perception & Psychophysics*, *66*, 363–376.

[5] Gold, B., Morgan, N., & Ellis, D. (2011). *Speech and Audio Signal Processing: Processing and Perception of Speech and Music*. 2nd edition. Wiley: Hoboken, NJ.

[6] Houde, J. F., & Jordan, M. I. (1998). Sensorimotor adaptation in speech production. *Science*, *279*, 1213–1216..

[7] Houde, J. F., & Nagarajan, S. S. (2011). Speech production as state feedback control. *Frontiers in Human Neuroscience*, *5*.

[8] Kaernbach, C. (1991). Simple Adaptive Testing with the Weighted up-down Method. *Perception & Psychophysics*, *49*, 227–229.

[9] Lane, H., & Webster, J. W. (1991). Speech Deterioration in Postlingually Deafened Adults. *Journal of the Acoustical Society of America*, *89*, 859–866.

[10] Levitt, H. (1971). Transformed up-down Methods in Psychoacoustics. *Journal of the Acoustical Society of America*, *49*, 467–&.

[11] Moore, B. C. J. (2008). Basic auditory processes involved in the analysis of speech sounds. *Phil. Trans. R. Soc. B*, *363*, 947-963.

[12] Newman, R. S. (2003). Using links between speech perception and speech production to evaluate different acoustic metrics: A preliminary report. *Journal of the Acoustical Society of America*, *113*, 2850–2860.

[13] Perkell, J. S., Guenther, F. H., Lane, H., Matthies, M. L., Stockmann, E., Tiede, M., & Zandipour, M. (2004). The distinctness of speakers' productions of vowel contracts in related to their discrimination of the contrasts. *Journal of the Acoustical Society of America*, *116*, 2338–2344.

[14] Perkell, J. S., Lane, H., Ghosh, S. S., Matthies, M. L., Tiede, M., Guenther, F. H., & Ménard, L. (2008). Mechanisms of Vowel Production: Auditory Goals and Speaker Acuity. In *8th International Seminar on Speech Production* (pp. 29–32). Strasbourg, France.

[15] Purcell, D. W., & Munhall, K. G. (2006). Adaptive control of vowel formant frequency: Evidence from real-time formant manipulation. *Journal of the Acoustical Society of America*, *120*, 966–977.

[16] Tourville, J. A., & Guenther, F. H. (2011). The DIVA model: A neural theory of speech acquisition and production. *Language and Cognitive Processes*, *26*, 952–981.

[17] Villacorta, V. M., Perkell, J. S., & Guenther, F. H. (2007). Sensorimotor adaptation to feedback perturbations of vowel acoustics and its relation to perception. *Journal of the Acoustical Society of America, 122*, 2306-2319.