openjournals.nl

Dutch Journal of Applied Linguistics
https://www.dujal.nl | E-ISSN 2211-7253

Speaker discrimination as a function of vowel realization: does focus affect perception?

Willemijn Heeren[1], Cesko Voeten[2,3] and Tessi Marks[1]
[1]*Leiden University* | [2]*Fryske Akademy* | [3]*Utrecht University*

Abstract The acoustic-phonetic characteristics of speech sounds are influenced by their linguistic position in an utterance. Because of acoustic-phonetic differences between different speech sounds, sounds vary in the amount of speaker information they contain. However, do spectral and durational differences between various realizations of the same sound that were sampled from different linguistic positions also impact speaker information? We investigated speaker discrimination in [–focus] versus [+focus] word realizations. Twenty-one Dutch listeners participated in a same-different speaker discrimination task, using stimuli varying in focus, vowel ([a:], [u]), and word context ([ɦ_k], [v_t]), spoken by 11 different speakers. Results show that an effect of focus on speaker-dependent information was present, but limited to words containing [u]. Moreover, performance on [u] words was influenced by (interactions of) word context and trial type (same- vs. different-speaker). Context-dependent changes in a speech sound's acoustics may affect its speaker-dependent information, albeit under specific conditions only.

Keywords speaker discrimination, focus, vowel quality, speech perception, Dutch

Article history
Received: March 2, 2021
Accepted: March 13, 2022
Online: May 23, 2022

Corresponding author
Willemijn Heeren, w.f.l.heeren@hum.leidenuniv.nl

Author contributions
Willemijn Heeren, conceptualization, methodology, formal analysis, writing – original draft, writing – review & editing, funding acquisition; Cesko Voeten, methodology, formal analysis, writing – original draft, writing – review & editing, visualization; Tessi Marks, conceptualization, investigation, writing – original draft, writing – review & editing

Funding information
This research was supported by a VIDI grant (276-75-010) from the Dutch Research Council.

Conflicting interests
The authors have declared that conflicts of interest did not exist.

Supporting information
None

Appendices
Appendix A: Statement-question pairs used to elicit [±focus] versions of the target words; Appendix B: Visualization of the two significant clusters for the by-trial random slope for the interaction

1 Introduction

Different speech sounds vary in the amount of acoustic speaker information they carry (Van den Heuvel, 1996; Kavanagh, 2012); vowels tend to contain more acoustic information on the speaker than consonants do, and within the class of consonants, nasals

How to cite Heeren, W., Voeten, C. and Marks, T. (2022). 'Speaker discrimination as a function of vowel realization: does focus affect perception?', *Dutch Journal of Applied Linguistics* 11: DOI 10.51751/dujal9420

(e.g. [n, m]) may carry more speaker information than fricatives (e.g. [f, z]) or stops (e.g. [t, b]). Also in speech perception, the speech sounds that make up an utterance influence how well a listener can discriminate speakers (Andics et al., 2007) or identify them (Amino & Arai, 2007). Much acoustic-phonetic research, however, has shown that the realization of one particular speech sound in one particular phonetic context varies as a function of linguistic position. Examples from Dutch are how the realization of a speech sound varies with the presence/absence of stress (Sluijter & Van Heuven, 1996), or with word class (Van Bergem, 1993). For instance, in a stressed versus unstressed position, the syllable 'kom' in 'ˈkom·ma' (*comma*) as opposed to that in 'kom·ˈpas' (*compass*), has a longer duration and less vowel reduction (Van Bergem, 1995, p. 28). Such variation in a speech sound's acoustics might in turn affect the speaker information carried by the sound. Whereas earlier work has compared the speaker information carried by different speech sounds, the main research question in the current investigation is if speaker discrimination performance on the *same* speech sound depends on whether tokens of the speech sound are sampled from focused versus non-focused contexts. Focus may be used to indicate new or contrastive information, and this is accompanied by acoustic changes (e.g., Hanssen et al., 2008; Sluijter & Van Heuven, 1996).

In the rest of this introductory section, speaker information in the acoustics of speech sounds (henceforth also referred to as *segments*) will be explained further in 1.1. This is followed by a discussion of prior research on how speaker discrimination may depend on stimulus composition in 1.2. The research questions and hypotheses are formulated in section 1.3.

1.1 Acoustic-phonetic speaker-dependent information in segments

In theory, an acoustic-phonetic speech feature (such as duration, fundamental frequency or formant frequency) would be highly speaker-specific when individual speakers show little variation in that feature when producing different tokens, i.e. there is little within-speaker variation, whereas different speakers would produce tokens that are very different with respect to the feature, i.e. there is large between-speaker variation. In practice, the speaker-specificity or speaker-discriminatory potential of individual speech features tends to be low to moderate. But crucially, both within-speaker and between-speaker variation of particular features differ in magnitudes across linguistic contexts (e.g., McDougall, 2006; Smorenburg & Heeren, 2020): in some situations speakers show more variation whereas in others they show less. This creates the option of particular linguistic contexts yielding higher speaker-specificity than others.

Most research on the speaker-specificity of acoustic-phonetic information has investigated how much speaker information is carried by individual speech sounds, and by vowels in particular. An early comparative investigation into differences in speaker-dependent acoustics for various speech sounds in Dutch was done by Van den Heuvel (1996). The segments included in that study were the vowels [i, a, u] and consonants

[p, t, k, d, s, m, n, r]. Analyses of productions gathered under controlled circumstances, i.e. read pseudo-words, showed that the vowel [a] was the most speaker-dependent segment and that plosives, such as [p, d] contained the least speaker information. The speaker-discriminating potential of a large set of German phonemes was studied by Schindler and Draxler (2013), who found that some consonants, that is [s, n, m, f], contained more speaker-dependent information than most of the vowels. As in Dutch (see Van den Heuvel, 1996), the vowel [a:] was found to contain more speaker-dependent information than [i, u] (Schindler & Draxler, 2013). In Czech, the vowel [i:] was found to pattern more with [a:] than [u] (Fejlová et al., 2013). Finally, various diphthongs were also found to differ in the speaker-specific information they carry (e.g. Morrison, 2009); the reason may be that, in addition to differences in vowel quality, diphthongs vary in the amount and direction of inherent spectral change. Taken together, the results on vowels suggest that in addition to vowel quality, inherent spectral change and vowel duration may contribute speaker-dependent acoustics. For consonants, the longer segments in which specific spectral characteristics (resonances) can be found seem to carry most speaker information.

There is some evidence that the linguistic context from which a speech sound is sampled influences the amount of within- and between-speaker variation it contains. He and Dellwo (2017) found more between-speaker variability in mouth-closing than mouth-opening gestures in read speech. This was explained as the second half of a syllable having lower articulatory demands relative to the first half, thus allowing for more articulatory freedom in individual speakers. In a follow-up investigation, He, Zhang and Dellwo (2019) demonstrated that also the first formant, which reflects the degree of mouth opening during vowel articulation, shows more between-speaker variability over the second half of a syllable than the first. In a study investigating speaker-dependent information in Dutch fricatives [s] and [x], Smorenburg and Heeren (2020) found that coda fricatives showed more between-speaker variation than onset fricatives, whereas the within-speaker variation showed a change in the opposite direction. Speaker classification scores were slightly better in codas than onsets for Dutch [x], whereas for [s] no positional difference in accuracy was found. A later study on speaker-dependent information in Dutch nasals [n] and [m] showed higher speaker classification scores for [n] in codas than onsets, but an effect in the opposite direction for [m] (Smorenburg & Heeren, 2021). This difference in classification was explained by effects of the phonetic context on speech sound acoustics; context effects were larger in onset [m]s and coda [n]s. Hence, in addition to an effect of articulatory freedom depending on syllable position, the speech sound's susceptibility to co-articulatory influences may also affect how much speaker information can be carried by a speech sound.

As was mentioned in section 1 (under Introduction), the acoustics of a particular speech sound are also influenced by higher-level linguistic factors, such as the presence/absence of stress on the syllable containing the speech sound (Sluijter & Van Heuven, 1996) or the word class that its carrier belongs to (Van Bergem, 1993; Heeren,

2020). Another construct that is known to affect the realization of utterances is focus, which for instance expresses which part of a sentence is new (narrow focus) or contrastive (contrastive focus). In Information Structure theory (Chafe, 1976), focus has been defined as indicating 'the presence of alternatives that are relevant for the interpretation of linguistic expressions' (Krifka, 2007, p. 18). Focus may be marked in different ways; several experiments on Dutch word realization demonstrated that segmental and syllabic acoustic measures may vary with focus condition. For instance, in comparison with broad focus, narrow and contrastive focus were found to affect segmental duration and fundamental frequency contours (e.g. Hanssen et al., 2008, see also Chen, 2009).

In sum, a speech sound's spectral and durational characteristics may change with its position in an utterance. It has been demonstrated that between-sound differences in e.g. spectral composition and duration affect speaker discrimination (see section 1.2. for more details). The question then is if the spectral and durational changes found *within* the same sound as a function of its linguistic context would impact speaker discrimination.

1.2 Speaker discrimination by linguistic content and context

In listeners, short words of different segmental composition elicit differential speaker discrimination performance. In Andics et al. (2007), native Dutch listeners heard CVC words and gave same-different speaker decisions for subsequent tokens that were auditorily presented. The CVCs were made up of onset [m] or [l], nucleus [ɛ] or [ɔ], and coda [s] or [t], thus yielding eight different Dutch words. Results showed that [m], [ɛ] and [s] yielded better discrimination performance than their positional counterparts [l], [ɔ] and [t]. In a trained-to-familiar speaker recognition experiment, Drozdova et al. (2017) found that listener performance was positively affected by the presence of vowels and nasals. These findings roughly correspond to findings from speech acoustics on which speech sounds contain most speaker-dependent information (Van den Heuvel, 1996; Kavanagh, 2012; Schindler & Draxler, 2013, and discussion in section 1.1).

Consistent with Fant's source-filter model of speech production (Fant, 1960), Baumann and Belin (2010) found that the two principal components explaining unfamiliar voice discrimination on sustained vowel data were based on the vocal source on the one hand, and the vocal filter on the other. This suggests that voiced speech sounds have an advantage over unvoiced speech sounds in talker perception, because the former benefit from both information sources. This is supported by results from Orchard and Yarmey (1995), who showed that speaker identification in a whispered voice line-up after having heard a whispering 'perpetrator' was worse than identification in an all-normal voice condition. They furthermore found that samples of longer duration resulted in better performance than shorter samples (see also Cook & Wilding, 1997). Earlier, Bricker and Pruzansky (1966) showed that not only duration *per se*, but also the segmental content within the speech fragment affected listener performance; better identification was

obtained with more phonemes in a same-duration sample. The importance of measures of acoustic variation was recently also reported by Lee, Keating and Kreiman (2019), when investigating the acoustic dimensions that explain within-speaker and between-speaker variation.

The task of speaker discrimination not only entails perceiving differences between speakers, but also perceiving that certain differences belong to the same speaker (assuming that non-identical stimulus pairs are presented). It has been shown that such within-speaker variability is judged differently for familiar than unfamiliar voices (Lavan et al., 2019); given 30 speech samples produced by two voices (15 fragments each) listeners who were familiar with the voices clustered the samples into fewer speaker identities than listeners who were unfamiliar with the voices. Therefore, both within-speaker and between-speaker variation of various acoustic parameters seemingly contribute to the processing of speaker information.

In addition to effects of segmental differences and acoustic variation on the processing of speaker information, earlier work has shown effects of higher levels of linguistic information. At the semantic level, for instance, Van Berkum et al. (2008) demonstrated that indexical information in a voice rapidly influences the semantic processing of what the voice is saying: the mentioning of an alcoholic beverage in a child's voice elicited a different ERP response than in an adult voice. Also, the linguistic relationship between words influences speaker discrimination performance: using a same-different speaker discrimination task, Narayan, Mak and Bialystok (2017) showed that listeners had a tendency to assign linguistically-related word combinations to the same speaker (e.g. words with the same phonological rhyme 'bay-day' as opposed to unrelated words 'day-bee').

Earlier perception studies have thus demonstrated a connection between speaker discrimination and higher-level linguistic information. The current study investigated how the presence versus absence of focus, which would alter the precise phonetic-acoustic content of the same speech sound, affects speaker discrimination. This investigation of sub-segmental differences thus adds to the literature on voice perception, which has so far shown that segments vary in speaker-dependent information, and that higher levels of linguistics may influence voice perception.

1.3 Research questions and hypotheses

This study in the first place investigated if speaker discrimination performance is affected by the presence versus absence of focus on a word. This was done using a same-different perception task, in which listeners judged whether subsequent words were produced by the same or by different speakers, and word stimuli were sampled from focused and non-focused sentence positions (further details in 2.2). In order to include some variation in word stimuli, two vowel nuclei were chosen from the three Dutch corner vowels included in Van den Heuvel (1996), namely those vowels that differed most in acoustic speaker information. This study thus also investigated if the difference in speaker-specific infor-

mation between the Dutch corner vowels [aː] and [u] that is found in acoustics is also observed in perception.

We hypothesized that speaker discrimination is more accurate on [aː] than [u], following the acoustic literature. Differences in both spectral and durational information between the vowels may contribute to this effect. As for the effect of focus, the main topic of this study, the prediction is less straightforward. In the comparison of unfamiliar voices there is an important role for low-level acoustic information (cf. Stevenage, 2018). The comparison of acoustic information between two speech samples leads to a same- or different-voice decision when the listener considers if the perceived variation between two non-identical tokens falls under within-speaker or between-speaker variation. From the literature it is known that listeners tend to underestimate within-speaker variation in unfamiliar speakers (Lavan et al., 2019). Therefore, the closer two realizations are acoustically, the more likely they are to be judged as 'same-speaker'. On the one hand, it can be argued that the relatively precise articulation in a [+focus] word may be more comparable from one token to the next than in a [–focus] condition. Moreover, [+focus] words are expected to be longer, thus giving them a perceptual advantage. On the other hand, with an expected lower occurrence frequency of these canonical forms relative to unfocused realizations in everyday speech, it may also be the case that less articulatory routine results in more within-speaker variation in [+focus] than in [–focus] forms. Moreover, between-speaker variation has been found to be higher in locations where articulatory demands are lower, such as in coda consonants and closing syllable gestures (e.g., He & Dellwo, 2017; Smorenburg & Heeren, 2020). In the current investigation the [–focus] words are therefore predicted to show larger between-speaker variation. Note that only in combination with smaller within-speaker variation, higher speaker-specificity is expected.

We assessed if effects of vowel quality and focus would hold across word contexts by including two carrier words. The word contexts [ɦ_k] and [v_t] were selected so that the speaker information contained by the different onset and offset consonants was estimated to be low and roughly comparable. Schindler and Draxler (2013) showed that the ratio of between-to-within speaker variation was quite low for [v] and [h] sampled from spontaneous German speech, with perhaps a small advantage for [v]. Moreover, according to Van den Heuvel (1996) Dutch stop consonants are expected to contribute relatively low speaker information. Moreover, these word contexts gave target word frequencies falling in the mid-frequency range (*haak*/*hoek*, *vaat*/*voet*); word frequency is known to influence articulation (Bell et al., 2009) as high-frequency words are more subject to reduction than low-frequency words.

The speaker discrimination results are accompanied by an acoustic analysis, including within- and between-speaker variation, of how speech sounds differ between conditions. Underlying this investigation is the assumption that speech sounds' acoustics vary with focus condition (Eefting, 1991; Van Heuven, 1997), but to ensure that this was also the case for the stimuli presented to our listeners, the acoustic analysis was carried out.

2 Method

Following Andics et al. (2007), a same-different forced-choice one-back task was used for the speaker discrimination perception task. This means that listeners were presented with a series of stimuli and decided if the speaker of the current stimulus is the same as or different from the speaker of the previous stimulus.

2.1 Participants

Twenty-one Dutch listeners without self-reported hearing problems volunteered to take part in this perception study (12 females, 9 males). Their mean age was 23 years (SD = 1.5 years). All participants beforehand gave their informed consent for taking part in the study, and afterwards received a modest thank-you gift.

2.2 Stimuli

Two Dutch minimal word pairs were used as stimulus contexts for the vowels [a:] and [u]: *haak*/*hoek* ([ɦa:k]/[ɦuk], 'hook'/'corner') and *vaat*/*voet* ([va:t]/[vut], 'dishes'/'foot'). These words, with word frequencies in the mid-frequency range, were selected using the SUBTLEX-NL corpus (Keuleers et al., 2010): *haak* (12.6/million), *hoek* (49.4/million), *vaat* (1.4/million), *voet* (50.8/million).

These four words were each recorded in a sentence context to evoke realizations with and without focus. The target sentences were answers to statement-question pairs that speakers saw on a computer screen, and speakers were instructed to produce answers in the form of full sentences. Two examples of the statement-question pairs and their intended answers are given here (see the appendix for a full list), where the first pair was intended to elicit [–focus] tokens of the target word, and the second pair to elicit [+focus] tokens:

Prompt 1:	Hij zet zijn vaat in de wasbak. Waar zet hij zijn vaat?
	He puts his dishes in the sink. Where does he put his dishes?
Answer 1:	Hij zet zijn vaat in de **wasbak**.
	He puts his dishes in the sink.
Prompt 2:	Hij zet zijn vaat in de wasbak. Wat zet hij in de wasbak?
	He puts his dishes in the sink. What does he put in the sink?
Answer 2:	Hij zet zijn **vaat** in de wasbak.
	He puts his dishes in the sink.

Eleven male speakers of Standard Dutch, aged 20 to 26 years, were recruited for the recordings. The stimuli were recorded in a sound-attenuated booth at the Leiden University Centre for Linguistics, using Praat software (Boersma & Weenink, 2018), a

Sennheiser MKH 416T microphone, and a FocusRite Scarlet 2i4 sound card. Recordings were saved as mono wave files (22,050 Hz, 16 bits). In an information sheet, speakers were instructed to produce the answer to a question about a short statement (see example above). Both the statement and question were shown orthographically on a computer screen in the recording booth. Before participation, speakers gave their informed consent.

Each speaker produced each target word-focus combination six times, so that a sufficient number of tokens would be available for the perception experiment in which five repetitions of each would be needed. Tokens with the highest signal intensity relative to the background noise were kept for further processing. Target words were cut from the carrier sentences, resulting in 440 stimuli (11 speakers × 5 repetitions × 2 focus conditions × 2 vowel nuclei ([aː], [u]) × 2 word contexts ([v_t], [ɦ_k]). Stimulus intensities were all set to 65 dB SPL.

2.3 Stimulus acoustics and their statistical assessment

As a first step, we evaluated statistically whether stimulus acoustics varied by focus condition and by vowel, as they are commonly assumed to do. The results are presented in section 3.1. In addition to stimulus duration, the mean fundamental frequency (F0) was taken as a measure of vocal source information, and as vocal filter parameters the first and second formants (F1 and F2) were extracted. Using Praat (Boersma & Weenink, 2018), F0 was measured over the full duration of a stimulus using an autocorrelation method, and formants were measured using the Burg method at the point in time where vowel intensity was maximal (window size = 25 ms).

Acoustic effects of the fixed factors Vowel, Focus and Word Context were evaluated in linear mixed-effects models (with α Bonferroni-corrected to .05/4 = .0125, given four acoustic measures). Factor levels [aː], [–focus] and [ɦ_k] were used default levels, with predictions for acoustic differences between factor levels being directional, e.g. duration is expected to shorten from [aː] to [u], and to lengthen from [–focus] to [+focus]. Using function buildmer from R package buildmer (Voeten, 2020), the maximally-converging models were obtained following a stepwise forward procedure. Duration was $\log_{10}$-transformed and F0, F1 and F2 were transformed to the Bark frequency scale before modelling.

Both within-speaker and between-speaker variances were determined by vowel, by focus condition and by word context. The ratio of between- to within-speaker variances is called the 'speaker-specificity index' (SSI, Van den Heuvel, 1996, p. 53), which was computed per acoustic parameter; the larger the variation between speakers relative to that within speakers, the higher the SSI, and the better speakers can presumably be separated. Because this is a descriptive view on the data, the measurements in Hertz and milliseconds were used for the sake of interpretability.

2.4 Design and procedure of the perception task

For each of the four words included (*haak, hoek, vaat, voet*), a pseudo-random presentation list was made, in which each different-speaker pair occurred once (11×10/2) and each same-speaker pair occurred five times (11×5). In this way the same number (i.e. 55) of different- and same-speaker trials were included per word. The latter trial type made use of the different recordings from the same speaker, so that no identical recordings were used in a comparison. During the full experiment, each individual token was used twice, and different tokens by the same speaker were presented not more than three times in a row.

Stimulus lists were distributed over 24 presentation blocks, containing one of eight vowel+word context+focus combinations each (3 blocks/combination). Per word context, there were 110 pairs for comparison, divided into blocks of 37, 37 and 36 trials each. A block lasted for about 1.5 minutes and the order of the blocks was randomized per listener. Across listeners 18,480 responses were collected. In this one-back discrimination task, not all speaker pairs occurred in both orders equally frequently. In the statistical analysis this is controlled for by the inclusion of random effects in the modelling.

To each token's onset and offset a 5-ms fade-in or fade-out was applied in order to prevent clicks at trial onset or offset. Between subsequent stimuli, i.e. during the listener's response time, pink noise at an intensity of 50 dB was played. The next stimulus started 2,400 ms after the onset of the previous one.

The perception experiment was run in a sound-attenuated booth at the phonetics laboratory of the Leiden University Centre for Linguistics, using E-Prime (Psychology Software Tools, 2012). Stimuli were presented at a standardized, comfortable listening level, over Beyerdynamic DT 770 PRO headphones.

Listeners were instructed to carefully listen to the subsequent tokens, and to respond, after every token (but the first), whether the speaker of the latter token was the same as that of the former token or not. Responses were given by pressing one of two buttons on a QWERTY keyboard, 'X' or 'N', one for 'same' and the other for 'different' speaker. The response buttons were counterbalanced across listeners.

Before the actual experiment started, listeners completed a short practice round including tokens of the word *vis* ('fish') to get used to the task. Including instruction, practice and breaks, the experiment lasted for about 45 minutes.

2.5 Statistical analysis of the perception data

Initial examination of the mean correct responses per speaker showed that none of the eleven voices were especially hard or easy for the listeners. Hence, all speakers were included in the analysis.

In earlier work on speaker discrimination by speech sound (Andics et al., 2007), responses from same-speaker trials were analyzed separately from responses to different-

speaker trials. We chose to analyze all data together, but added trial type (same-speaker, different-speaker) as a factor into the design. The other fixed factors in the design were focus (+, –), vowel nucleus ([a:], [u]), and word context ([v_t], [fi_k]), with full interactions. All factors were coded using deviation coding. Random intercepts by participants and by trials (that is, speaker1–speaker2 combinations) were included in the design, as were analogous random slopes for all factor combinations included as fixed effects. A diagonal random-effects covariance matrix was assumed. The dependent variable was the correctness of the response given by the listener, coded as 0 for incorrect and 1 for correct. The data were analyzed using a mixed-effects logistic-regression tree (Fokkema et al., 2018; see Tagliamonte & Baayen, 2012 for an accessible introduction to tree-based models in linguistics). Function `buildmertree` from R package `buildmer` (Voeten, 2020) was used to find the maximal random-effects structure that still converged non-singularly, based on the random effects' contributions to the AIC (Akaike, 1971) of the model. The results of the perception experiment are presented in section 3.2.

3 Results

3.1 Stimulus acoustics: the effect of focus condition

Before a perceptual effect of [±focus] on speaker discrimination was evaluated, the effect of [±focus] realizations on stimulus acoustics was assessed. Per acoustic measure, the model output on effects of Focus, Vowel and Word Context is given in Table 1. Significant terms are printed in bold. Three out of four acoustic measurements showed a main effect of or interaction with Focus; these mainly reflected a higher F0 and longer stimulus durations in focused realizations. The effect was not equally strong in all stimulus conditions. As for F1, there was a tendency for more mouth opening in [a:] (given the marginal main effect main effect of Focus, $t = 2.2$), but in [u] this effect was countered under focus, presumably because of increased rounding in its pronunciation. Moreover, stimulus acoustics were influenced by vowel quality, as may be expected by intrinsic differences between the vowels, and to a lesser extent by the word context and interactions between the linguistic factors.

Table 2 shows, in a descriptive manner, the within-speaker variance by condition and by acoustic measure as well as the speaker-specificity index (SSI, the ratio of between-to-within-speaker variances). The F0 has higher within-speaker variation in focused than unfocused words, but also a higher SSI meaning that between-speaker variances also increase with focus. F1 shows less within-speaker variance with focus, whereas SSI tends to increase. F2 shows an increase in within-speaker variance with focus in the [v_t] condition, but a decrease in [fi_k], whereas SSI shows behavior in the opposite direction. For [a:] duration, but not for [u], SSI is higher when focused.

Table 1 Modelling results for the acoustic parameters log-duration, and F0, F1 and F2 (all three in Bark, ***: $p < .001$, **: $p < .01$, *: $p = .01$). Significant terms in bold

	F0 [Bark]		Log(duration)	
	Estimate	***t*-value**	**Estimate**	***t*-value**
Intercept	**0.699**	18.0***	**−0.626**	−28.0***
Focus_[+]	**0.448**	5.8***	**0.079**	4.0***
Vowel_[u]	**0.141**	4.8***	−0.009	−0.7
WordContext_[v_t]	**0.145**	5.0***	**0.093**	5.1***
Focus_[+]: Vowel_[u]	0.011	0.3	**−0.063**	−3.5***
Focus_[+]: WordContext_[v_t]	**−0.211**	−5.2***	−0.019	−1.1
Vowel_[u]: WordContext_[v_t]	−0.058	−1.4	**−0.116**	−6.5***
Focus_[+]: Vowel_[u]: WordContext_[v_t]	0.102	1.8	**0.097**	3.9***
	F1 [Bark]		**F2 [Bark]**	
	Estimate	***t*-value**	**Estimate**	***t*-value**
Intercept	**6.942**	52.6***	**10.391**	265.2***
Focus_[+]	0.401	2.2	0.107	0.5
Vowel_[u]	**−3.803**	−15.7***	**−2.213**	−5.4***
WordContext_[v_t]	−0.277	−1.8	−0.380	−1.7
Focus_[+]: Vowel_[u]	**−0.600**	−2.8**	−0.640	−2.1
Focus_[+]: WordContext_[v_t]	−0.143	−0.7	0.013	0.04
Vowel_[u]: WordContext_[v_t]	0.053	0.3	**1.171**	3.8***
Focus_[+]: Vowel_[u]: WordContext_[v_t]	0.611	2.0	0.354	0.8

Table 2 Mean within-speaker variance and speaker-specificity index (SSI) per acoustic parameter. Measures are given by vowel, focus condition and word context

		[−focus]		[+focus]	
	Parameter	**[fi_k]**	**[v_t]**	**[fi_k]**	**[v_t]**
[a:]	F0_within	161	163	223	136
	F0_SSI	1.03	1.39	3.09	3.65
	F1_within	1,940	1,419	1,224	902
	F1_SSI	2.12	1.39	2.42	2.80
	F2_within	6,363	1,774	2,849	1,989
	F2_SSI	1.41	3.34	2.02	2.45
	Duration_within	0.001	0.001	0.001	0.001
	Duration_SSI	2.33	2.33	4.60	3.99

Table 2 Mean within-speaker variance and speaker-specificity index (SSI) (*cont.*)

		[–focus]		[+focus]	
	Parameter	**[ɦ_k]**	**[v_t]**	**[ɦ_k]**	**[v_t]**
[u]	F0_within	76	73	136	152
	F0_SSI	2.22	2.60	8.05	3.62
	F1_within	1,753	905	836	712
	F1_SSI	1.16	2.79	1.86	2.39
	F2_within	9,138	4,509	2,276	7,627
	F2_SSI	0.95	3.01	1.70	1.41
	Duration_within	0.005	0.001	0.004	0.002
	Duration_SSI	1.53	2.1	1.55	1.75

Together, these results show that stimulus acoustics as well as the SSI vary by focus condition, supporting the hypothesis of differential acoustic speaker information by factor combination.

3.2 Perception results: effects of focus condition and vowel

The main research questions on speaker discrimination as a function of focus condition and vowel quality were evaluated using a same-different perception task. The results of the statistical analysis are shown in Figure 1, and reflect that listeners were generally quite successful in speaker discrimination, but also that performance varied by factor combination. The first, and hence most important, split in the tree model is on vowel, [aː] versus [u]. On [aː] trials listeners gave 87.5% correct responses, irrespective of focus, word context or trial type. The bar charts show that performance on [u] conditions was somewhat lower.

Within [u] trials, the first split was on the predictor Focus and performance further depended on the combination of word context and trial type. For [–focus] words containing [u], Word Context was the next split, followed by Trial Type on both contexts. Different-speaker trials (75.1%) were better than same-speaker trials (68.2%) containing non-focused tokens of [ɦuk]. For non-focused tokens of [vut] same-speaker trials (84.0%) were better than different-speaker trials (77.3%).

For [+focus] words containing [u], the first split was on Trial Type, followed by a split on Word Context for different-speaker trials only. On same-speaker trials, listeners got 88.5% correct responses. On different-speaker trials, [vut] tokens received more correct responses (80.7%) than [ɦuk] tokens (73.9%).

The model also contained a number of random effects of participant and of trial, including random slopes. To further investigate the random effects and thus look for

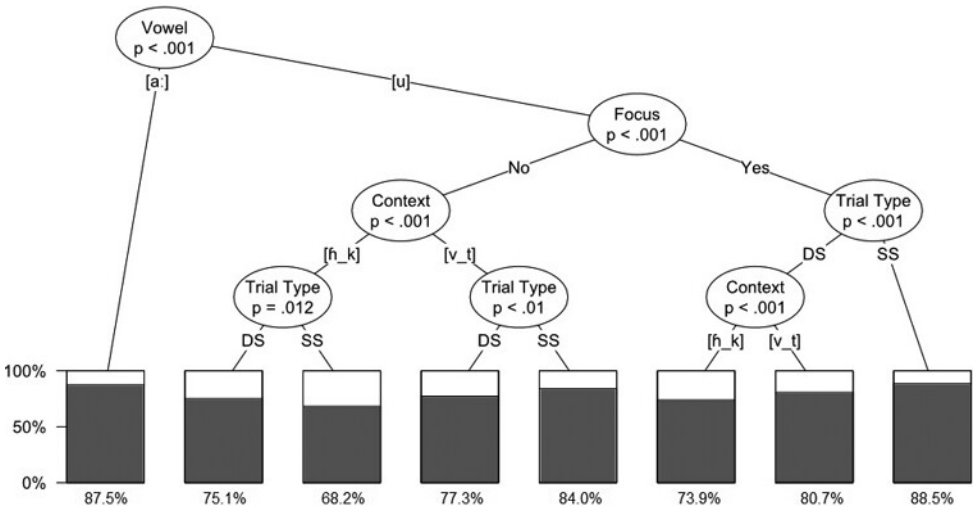


Figure 1 The fitted mixed-effects logistic-regression tree, modelling participants' correct responses as a function of Vowel ([a:]-[u]), Focus ([+/-]), Word Context ([ɦ_k], [v_t]), and Trial Type (DS = different speaker, SS = same speaker)

potential confounds, a cluster analysis was applied, following the method in Voeten (2020). No interpretable[1] structure was found in the by-participants or by-items random-effects structure, which suggests that the experiment was free from confounds due to these random effects.

4 Discussion and conclusion

This study investigated if speaker discrimination performance varies with the presence of focus on target words and if the difference in speaker-specific information between the Dutch corner vowels [a:] and [u] found in acoustic analyses (Van den Heuvel, 1996) is also observed in perception. On average, listeners reached over 80% correct responses for both same-speaker and different-speaker trials. The regression tree further showed that different factor combinations, reflecting different stimulus acoustics, affected the listeners' ability to discriminate speakers.

In line with expectations, listeners gave more correct responses to stimulus pairs containing [a:] words than [u] words. These results complement the perceptual results obtained by Andics et al. (2007) on Dutch vowels [ɛ] and [ɔ], and follow the literature on acoustic speaker information in [a:] versus [u] in Dutch (Van den Heuvel, 1996), and also in other languages (Fejlová et al., 2013; Schindler & Draxler, 2013). Perception of tokens containing the vowel [a:] was not influenced further by linguistic context. For [u], the situation was different: the first split in the model was made by Focus, where

focused realizations reached higher discrimination performance than unfocused ones. Best performance was observed for [+focus] same-speaker trials, where the phonetic-acoustic differences at the surface are apparently small enough for good-quality voice matching. This result shows that the phonetic-acoustic content resulting from the linguistic position in which a word is pronounced may influence the speaker-dependent information available to listeners. Yet, the effect obtained here is restricted to the less informative vowel, [u].

In the literature it has been reported that listeners tend to underestimate within-speaker variation when listening to unfamiliar speakers (Lavan et al., 2019), and are thus more likely to perceive same-speaker samples as coming from different speakers. This would especially lower listeners' accuracy in conditions where within-speaker variance is high. The finding that listeners performed better using focused than unfocused [u], however, suggests that the mere amount of within-speaker variance does not fully explain the discrimination results; as Table 2 shows, within-speaker variance of F0, an important parameter in speaker discrimination (Baumann & Belin, 2010), tended to be larger in focused than unfocused words. At the same time, however, the SSIs of F0 showed that there was also a tendency for between-speaker variance to increase under focus, and also to increase more than within-speaker variance. This suggests that listeners use both within-speaker and between-speaker variance in perception.

Additional evidence that within-speaker variance alone may not explain the speaker discrimination results comes from the comparison of listener performance on [a:] versus [u]. For F1, within-speaker variances are larger for [a:], whereas for F2, within-speaker variances are larger for [u]. SSI, however, is in most cases larger in [a:] than [u], again suggesting a contribution for both sources of variation in perception. The acoustic analysis in this study was limited to one vocal source and two vocal filter parameters, and these are only a small subset of the parameters that have been included in the recent literature (Lee et al., 2019). That work furthermore showed an important contribution of measures capturing phonetic change within stimuli, which was not included here. More detailed analyses of which information listeners use to perform speaker discrimination tasks are left for future research.

The two splits on Word Context, within [u], reflect that speaker discrimination was better on [vut] than [fiuk] contexts, even though word contexts had been selected to contain similar amounts of acoustic speaker information. As plosives have been reported to contain the least speaker-dependent information (Van den Heuvel, 1996), the word-context effect must be attributed mainly to the differential information contained by the onsets. Schindler and Draxler (2013) demonstrated that there was slightly more speaker-dependent information in the spectra of [v] than of [h]. In addition, in the current investigation [v] words in most cases were longer than [fi] words (see Table 1), where additional duration is generally beneficial in perceptual tasks (Bricker & Pruzansky, 1966; Orchard & Yarmey, 1995; Cook & Wilding, 1997). Finally, the two fricatives are expected

to differ in coarticulation with the vowel, with [v] presumably showing larger between-speaker differences in coarticulation than [ɦ]. This may also contribute to explaining the word context difference (Bricker & Pruzansky, 1966; Lee et al., 2019).

The current investigation was limited in its scope, studying the effect of focus using two vowels, in two word contexts each. The higher-order interactions in the results showed that listeners are sensitive to subtle differences conditioned by linguistic structure when processing voice information. However, in the current experiment very short utterances were presented, produced by a relatively homogeneous group of male speakers. In natural communicative settings, more variation between (non-seen) speakers is likely to occur, and is expected to improve speaker discrimination. Those circumstances potentially reduce the effect of relatively subtle linguistic cues; such cues here affected speaker discrimination only in the more-challenging words containing [u], but not those containing [aː]. Moreover, if longer utterances had been used, as would be found in natural interaction, listeners would receive additional speaker information from the utterance. However, if within-speaker variability increases over longer phrases, speaker discriminability may also be compromised.

To conclude, an effect of focus on speaker-dependent information contained by a word was present, but limited. Moreover, additional evidence was found for the claim that different speech sounds differ in speaker-dependent information ([aː] versus [u], and possibly [v] versus [ɦ]). At the same time, under more real-world conditions than the current discrimination task, the detection of speaker changes in speech is unlikely to be affected by focus. When needed, however, listeners seem skilled at exploiting the little information that is available.

Acknowledgements

This research was supported by a VIDI grant (276-75-010) from the Dutch Research Council.

Notes

1 Two significant clusters were identified for the by-trial random slope for the interaction Focus by Vowel, but the resulting two clusters (see Appendix B) seem to be spurious. We note that Appendix B's Cluster 2 seems to be relatively tightly concentrated around zero, while Cluster 1 seems to stay farther away, but we do not have any meaningful interpretation for these differences.

References

Akaike, H. (1971). Information theory and an extension of the maximum likelihood principle. In B.N. Petrov, & F. Csáki (Eds.), *2nd International Symposium on Information Theory, Tsahkadsor, Armenia, USSR*, September 2–8, 1971 (pp. 267–281). Budapest, Akadémiai Kiadó.

Amino, K., & Arai, T. (2007). Contribution of consonants and vowels to the perception of speaker identity. In *Japan-China Joint Conference on Acoustics*. Sendai, Japan.

Andics, A., McQueen, J.M., & Van Turennout, M. (2007). Phonetic content influences voice discriminability. In J. Trouvain, & W.J. Barry (Eds.), *Proceedings of the 16th International Congress of Phonetic Sciences* (pp. 1829–1832). Dudweiler: Pirrot. [https://pure.mpg.de/pubman/faces/ViewItemOverviewPage.jsp?itemId=item_57725]

Baumann, O., & Belin, P. (2010). Perceptual scaling of voice identity: Common dimensions for different vowels and speakers. *Psychological Research* PRPF, 74(1), 110. [https://link.springer.com/article/10.1007/s00426-008-0185-z]

Bell, A., Brenier, J.M., Gregory, M., Girand, C., & Jurafsky, D. (2009). Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language*, 60(1), 92–111. https://doi.org/10.1016/j.jml.2008.06.003

Boersma, P., & Weenink, D. (2018). *Praat. Doing phonetics by computer* (Version 6.0.42) [Computer program].

Bricker, P.D., & Pruzansky, S. (1966). Effects of stimulus content and duration on talker identification. *The Journal of the Acoustical Society of America*, 40, 1441–1449. https://doi.org/10.1121/1.1910246

Chafe, W.L. (1976). Givenness, contrastiveness, definiteness, subjects, topics and point of view. In C.N. Li (Ed.) *Subject and topic* (pp. 27–55). Academic Press.

Chen, A. (2009). The phonetics of sentence-initial topic and focus in adult and child Dutch. In M. Vigário, S. Frota, & M. João Freitas (Eds.), *Phonetics and phonology: interactions and interrelations* (pp. 91–106). John Benjamins Publishing Company.

Cook, S., & Wilding, J. (1997). Earwitness testimony: Never mind the variety, hear the length. *Applied Cognitive Psychology*, 11(2), 95–111. https://doi.org/10.1002/(SICI)1099--0720(199704)11:2<95::AID-ACP429>3.0.CO;2-O

Drozdova, P., Van Hout, R., & Scharenborg, O. (2017). L2 voice recognition: The role of speaker-, listener-, and stimulus-related factors. *The Journal of the Acoustical Society of America*, 142(5), 3058–3068. https://doi.org/10.1121/1.5010169

Eefting, W. (1991). The effect of "information value" and "accentuation" on the duration of Dutch words, syllables, and segments. *The Journal of the Acoustical Society of America*, 89(1), 412–424.

Fant, G. (1960). *Acoustic theory of speech production*. Mouton and Co.

Fejlová, D., Lukeš, D., & Skarnitzl, R. (2013). Formant contours in Czech vowels: Speaker-discriminating potential. *Proceedings of Interspeech 2013*, 25–29 August 2013, Lyon, France (pp. 3182–3186).

Fokkema, M., Smits, N., Zeileis, A., Hothorn, T., & Kelderman, H. (2018). Detecting treatment-subgroup interactions in clustered data with generalized linear mixed-effects model trees.

Behavior Research Methods, 50, 2016–2034. [https://link.springer.com/article/10.3758/s13428-017-0971-x]

Hanssen, J., Peters, J., & Gussenhoven, C. (2008). Prosodic effects of focus in Dutch declaratives. *Proceedings of the 4th international conference on Speech Prosody*. Campinas, Brazil, pp. 609–612.

He, L., & Dellwo, V. (2017). Between-speaker variability in temporal organizations of intensity contours. *The Journal of the Acoustical Society of America*, 141(5), EL488–EL494. https://doi.org/10.1121/1.4983398

He, L., Zhang, Y., & Dellwo, V. (2019). Between-speaker variability and temporal organization of the first formant. *The Journal of the Acoustical Society of America*, 145(3), EL209–EL214. https://doi.org/10.1121/1.5093450

Heeren W.F.L. (2020). The effect of word class on speaker-dependent information in the Standard Dutch vowel /a:/. *The Journal of the Acoustical Society of America*, 148(4), 2028–2039. https://doi.org/10.1121/10.0002173

Kavanagh, C. (2012). *New consonantal acoustic parameters for forensic speaker comparison* Doctoral dissertation. University of York. [https://etheses.whiterose.ac.uk/3980/

Keuleers, E., Brysbaert, M., & New, B. (2010). SUBTLEX-NL: A new frequency measure for Dutch words based on film subtitles. *Behavior Research Methods*, 42(3), 643–650. https://doi.org/10.3758/BRM.42.3.643

Krifka, M. (2007). Basic notions of information structure. In C. Féry, G. Fanselow, & M. Krifka (Eds.), *The notions of information structure* (pp. 13–55). Universitätsverlag Potsdam.

Lavan, N., Burston, L.F.K., & Garrido, L. (2019). How many voices did you hear? Natural variability disrupts identity perception from unfamiliar voices. *British Journal of Psychology*, 110, 576–593. https://doi.org/10.1111/bjop.12348

Lee, Y., Keating, P., & Kreiman, J. (2019). Acoustic voice variation within and between speakers. *The Journal of the Acoustical Society of America*, 146(3), 1568–1579. https://doi.org/10.1121/1.5125134

McDougall, K. (2006). Dynamic features of speech and the characterization of speakers: Towards a new approach using formant frequencies. *International Journal of Speech, Language and the Law*, 13(1), 89–126.

Morrison, G.S. (2009). Likelihood-ratio forensic voice comparison using parametric representations of the formant trajectories of diphthongs. *The Journal of the Acoustical Society of America*, 125(4), 2387–2397. https://doi.org/10.1121/1.3081384

Narayan, C.R., Mak, L., & Bialystok, E. (2017). Words get in the way: Linguistic effects on talker discrimination. *Cognitive Science*, 41(5), 1361–1376. https://doi.org/10.1111/cogs.12396

Orchard, T., & Yarmey, A.D. (1995). The effects of whispers, voice-sample duration, and voice distinctiveness on criminal speaker identification, *Applied Cognitive Psychology*, 9(3), 249–260. https://doi.org/10.1002/acp.2350090306

Psychology Software Tools. (2012). *E-Prime* (*Version 2.0*). https://www.pstnet.com

Schindler, C., & Draxler, C. (2013) Using spectral moments as a speaker specific feature in nasals and fricatives. *Proceedings of Interspeech* (pp. 2793–2796), Lyon, France, 25–29 August 2013.

Sluijter, A.M.C., & Van Heuven, V.J. (1996). Spectral balance as an acoustic correlate of linguistic

stress. *The Journal of the Acoustical Society of America*, 100, 2471–2485. https://doi.org/10.1121/1.417955

Smorenburg, B.J.L., & Heeren, W.F.L. (2020). The distribution of speaker information in Dutch fricatives /s/ and /x/ from telephone dialogues. *Journal of the Acoustical Society of America*, 147(2), 949–960. https://doi.org/10.1121/10.0000674

Smorenburg, B.J.L., & Heeren W.F.L. (2021). Acoustic and speaker variation in Dutch /n/ and /m/ as a function of phonetic context and syllabic position. *The Journal of the Acoustical Society of America*, 150(2), 979–989. https://doi.org/10.1121/10.0005845

Stevenage, S.V. (2018). Drawing a distinction between familiar and unfamiliar voice processing: A review of neuropsychological, clinical and empirical findings. *Neuropsychologia*, 116, 162–178. https://doi.org/10.1016/j.neuropsychologia.2017.07.005

Tagliamonte, S.A., & Baayen, R.H. (2012). Models, forests, and trees of York English: Was/were variation as a case study for statistical practice. *Language Variation and Change*, 24, 135–178. [https://clarinoai.informatik.uni-leipzig.de/fedora/objects/oai:mrr/datastreams/info/content]

Van Bergem, D.R. (1993). Acoustic vowel reduction as a function of sentence accent, word stress, and word class. *Speech Communication*, 12(1), 1–23. https://doi.org/10.1016/0167-6393(93)90015-D

Van Bergem, D.R. (1995). Acoustic and lexical vowel reduction. PhD dissertation, University of Amsterdam. [https://dare.uva.nl/search?identifier=6ba47af3-8bf4-4b46-81cb-2adb65dbc955]

Van Berkum, J.J., Van den Brink, D., Tesink, C.M., Kos, M., & Hagoort, P. (2008). The neural integration of speaker and message. *Journal of Cognitive Neuroscience*, 20(4), 580–591. https://doi.org/10.1162/jocn.2008.20054

Van den Heuvel, H. (1996). *Speaker variability in acoustic properties of Dutch phoneme realisations*. Doctoral dissertation, Katholieke Universiteit Nijmegen. [https://repository.ubn.ru.nl/bitstream/handle/2066/76416/76416.pdf

Van Heuven, V.J. (1997). Effects of focus distribution and accentuation on the temporal and melodic organisation of word groups in Dutch. In S. Barbiers, J. Rooryck, & J. van de Weijer (Eds.), *Small words in the big picture. Squibs for Hans Bennis*. HIL Occasional Papers no. 2 Leiden: Holland Institute of Generative Linguistics. 37–42.

Voeten, C.C. (2020). buildmer: Stepwise elimination and term reordering for mixed-effects regression. *R* package version 1.5. https://CRAN.R-project.org/package=buildmer

Appendix A: Statement-question pairs used to elicit [±focus] versions of the target words

Wat stoot hij aan de tafel? Hij stoot zijn **voet** aan de tafel
'How did he hit the table? He bumped his foot into the table'

Waar stoot hij zijn voet aan? Hij stoot zijn voet aan de **tafel**
'How did he hit his foot? He bumped his foot into the table'

Wat zet hij in de wasbak? Hij zet zijn **vaat** in de wasbak
'What does he place in the sink? He puts his dishes in the sink'

Waar zet hij zijn vaat? Hij zet zijn vaat in de **wasbak**
'Where does he put his dishes? He puts his dishes in the sink'

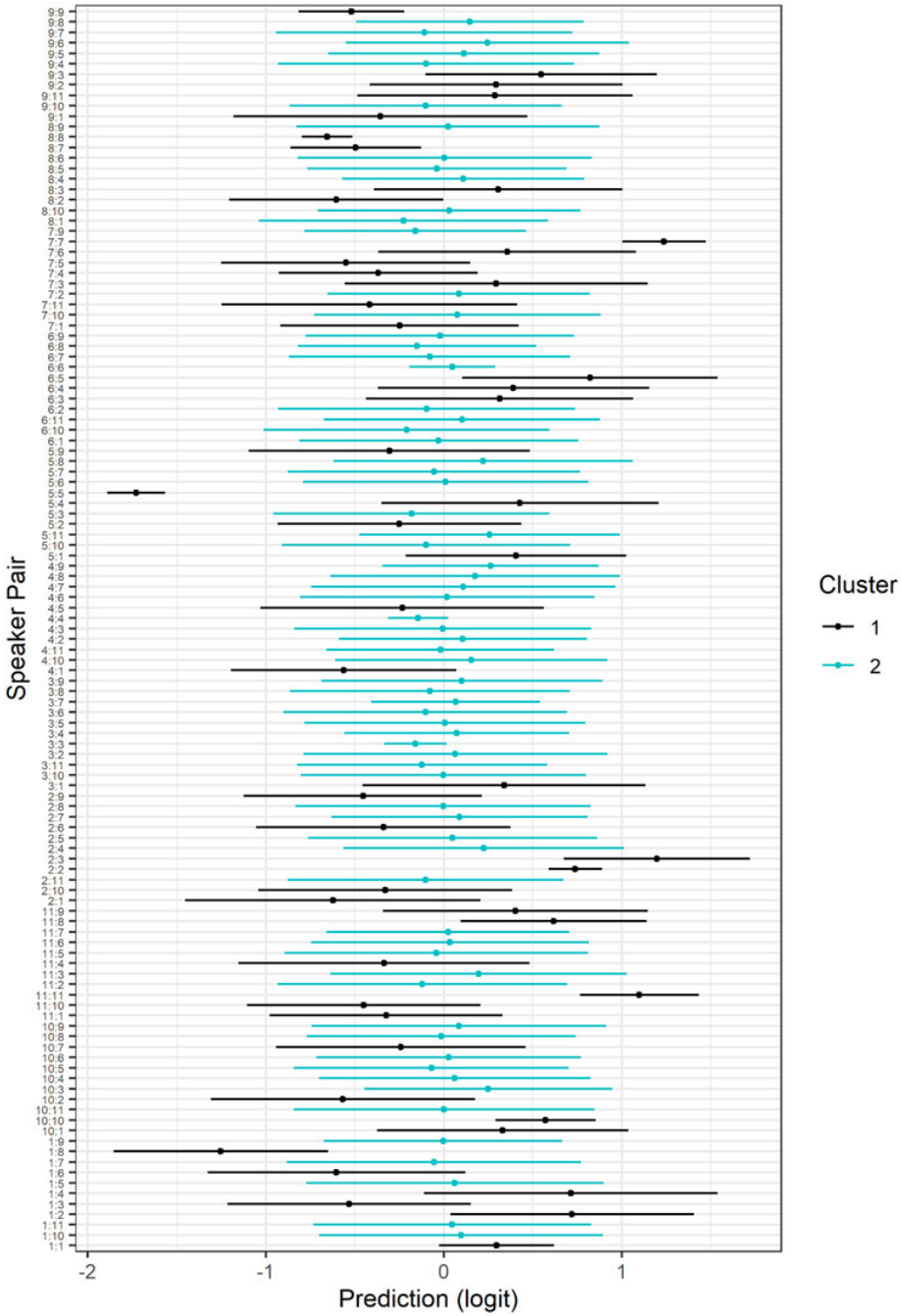
Waar staat ze te wachten? Ze staat op de **hoek** te wachten
'Where is she waiting? She's waiting at the corner'

Wat staat ze te doen op de hoek? Ze staat op de hoek te **wachten**
'What is she doing at the corner? She's waiting at the corner'

Waar is de vis aan geslagen? De vis is aan de **haak** geslagen
'With what was the fish caught? The fish was caught on the hook'

Wat is er aan de haak geslagen? De **vis** is aan de haak geslagen
'What was caught on the hook? The fish was caught on the hook'

Appendix B: Visualization of the two significant clusters for the by-trial random slope for the interaction Focus by Vowel



 DOI 10.51751/dujal9420