Your "VOORnaam" is not my "VOORnaam": An acoustic analysis of individual talker differences in word stress in Dutch

Giulio G.A. Severijnen [a]*, Hans Rutger Bosker [a, b], James M. McQueen [a, b]

*Corresponding author: giulio.severijnen@donders.ru.nl, Thomas van Aquinostraat 4, 6525 GD Nijmegen, The Netherlands
[a] Donders Institute for Brain, Cognition, and Behaviour, Radboud University. Thomas van Aquinostraat 4, 6525 GD Nijmegen, The Netherlands
[b] Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands. Wundtlaan 1, 6525 XD Nijmegen, The Netherlands

giulio.severijnen@donders.ru.nl
hansrutger.bosker@donders.ru.nl
james.mcqueen@donders.ru.nl

Accepted for publication in *Journal of Phonetics* on January 16th, 2024

[Authors' accepted manuscript]

Abstract

Different talkers speak differently, even within the same homogeneous group. These differences lead to acoustic variability in speech, causing challenges for correct perception of the intended message. Because previous descriptions of this acoustic variability have focused mostly on segments, talker variability in prosodic structures is not yet well documented. The present study therefore examined acoustic between-talker variability in word stress in Dutch. We recorded 40 native Dutch talkers from a participant sample with minimal dialectal variation and balanced gender, producing segmentally overlapping words (e.g., *VOORnaam* vs. *voorNAAM*; 'first name' vs. 'respectable', capitalization indicates lexical stress), and measured different acoustic cues to stress. Each individual participant's acoustic measurements were analyzed using Linear Discriminant Analyses, which provided coefficients for each cue, reflecting the strength of each cue in a talker's productions. On average, talkers primarily used mean F0, intensity, and duration. Moreover, each participant also employed a unique combination of cues, illustrating large prosodic variability between talkers. In fact, classes of cue-weighting tendencies emerged, differing in which cue was used as the main cue. These results offer the most comprehensive acoustic description, to date, of word stress in Dutch, and illustrate that large prosodic variability is present between individual talkers.

Keywords

Acoustic variability, Lexical stress, Prosody, Individual differences, Cue weighting

Declaration of competing interests: none

Data availability statement

The data and the analysis scripts of this experiment have been made openly available at https://data.donders.ru.nl/login/reviewer-241214775/bbEj3LjTV3m6x6u0JQKatK2RGlMNST-jdqVIzFfHMLc under a CC-BY 4.0 International license. A subset (due to data sharing constraints) of the recordings has been made available in the Corpus of Dutch Lexical Stress (Severijnen et al., 2023). The corpus can be accessed at https://data.donders.ru.nl/login/reviewer-215350326/cbS234GkDzu06Cy8VA2IhW9JhKWYXoIhB83Yq0I1x98 .

Your "VOORnaam" is not my "VOORnaam": An acoustic analysis of individual talker differences in word stress in Dutch

1.0 Introduction

Individual differences between talkers lead to large variability in the speech signal. For example, consider the phrase 'Speech is subject to between-talker variability', produced by two different talkers. Even though the phrase itself is identical, the acoustic realization will be highly different between the two talkers. This variability can have multiple possible causes, such as biological differences between talkers (e.g., fundamental frequency, intensity) and the talker's accent or production strategies. These differences in turn lead to variability in both the segmental structures (e.g., vowels, consonants) and suprasegmental structures (e.g., sentential accentuation, lexical stress) of the speech signal, and can even impede correct perception of the intended message. For example, perceiving the incorrect stress pattern in 'subject' in the example sentence above (i.e., perceiving 'subJECT' instead of 'SUBject', capitalization indicates lexical stress) would result in perception of a different word. An important step in understanding how listeners can deal with such variability is to find out how it is manifested in speech acoustically. In the present study, we take a step in this direction by examining between-talker variability in word stress in Dutch.

1.1 Acoustic variability in segmental information

Between-talker variability affects acoustic properties at the segmental and the suprasegmental level. In the following, we will review how this variability is manifested in speech, and discuss possible sources of the variability. Note that we provide these possible sources (e.g., gender, regional dialects) only as examples where acoustic variability can originate from. The goal of the present study was not to directly test these sources, but rather to examine and describe acoustic variability in word stress in Dutch.

With respect to segmental structures, there is variability in productions of vowels and consonants. For example, in Dutch, women produce vowels with longer durations compared to men (Adank et al., 2004). Furthermore, the first and second formants (F1 and F2) of Dutch vowels are affected by gender and regional dialect (Adank et al., 2004, 2007). Comparable acoustic differences have also been found between men, women, and children in American English (Hillenbrand et al., 1995). Further, acoustic differences between American-English talkers have been found in voice onset times (VOT) for stop consonants (Allen et al., 2003; Theodore et al., 2009), and in centroid frequencies and skewness for fricatives (Newman et al., 2001).

In addition to variability within *single* acoustic cues (e.g., within vowel duration, formant values, or VOT), talkers also vary in how *multiple* acoustic cues are combined. That is, talkers appear to differ in how much they weigh (i.e. how much they prioritize) multiple cues when producing speech sounds. For example, in productions of plosives (e.g., /p/ and /b/), talkers use multiple cues such as voice onset time (VOT), fundamental frequency (F0) of the voiced part after the plosive, and spectral centre of gravity (Lisker, 1986; van Alphen & Smits, 2004) to signal these sounds, and talkers vary in how exactly they combine these cues. That is, there are different cue hierarchies for different languages (Lisker & Abramson, 1964) and regional dialects (Kang, 2013). This phenomenon has been found for different speech sounds and languages (for review, see Schertz & Clare, 2020).

There even appear to be differences in phonetic cue-weighting tendencies for individual talkers. Schertz et al. (2015) recorded native talkers of Korean producing words containing word-initial stops (e.g., /p/ and /b/) in their L1 (Korean) and their L2 (English). Acoustic cues (VOT, F0 and closure duration) were analyzed for each individual participant by running a Linear Discriminant Analysis (LDA) on each participant's data. As Schertz et al. (2015) point out, the LDA can be used to model the extent to which a set of acoustic cues predict category membership (i.e., classified as either a /p/ or a /b/) in each participant's productions, and can be thought of as a metric of how well a given dataset can be separated using an optimized linear combination of a set of dimensions. In other words, the output of the LDA gives a set of

coefficients, reflecting the optimal combination of acoustic cues that predicts whether an observation should be categorized as a /p/ or a /b/, which can thus be interpreted as cue weights. The analyses revealed great variability between talkers in cue weights for VOT, F0, and closure duration, even within the same language. Specifically, they found that each talker used a unique set of cue weights to produce the speech sounds, illustrating individual phonetic cue-weighting tendencies.

In sum, previous research on acoustic properties of segments suggests that acoustic between-talker variability can be attributed to (1) variability within cues and (2) different phonetic cue-weighting tendencies between talkers.

1.2 Acoustic variability in prosody

Prosody in speech refers to information that is conveyed on top of individual vowels and consonants. This is often referred to as the suprasegmental aspect of speech because it defines patterns that are largely independent of the segmental makeup of a given word or phrase. Prosody can take many forms, such as intonation, rhythm, or lexical stress and is signaled by acoustic cues such as F0, intensity and duration (for review, see Cole, 2015). Previous research has shown that prosodic structures, such as intonation, are also affected by between-talker variability on a group level. For instance, Haan & Van Heuven (1999) found that in Dutch, women produce questions using a wider pitch range than men. Furthermore, variation in pause distributions and pitch accents has been established as a function of regional dialect and gender in American English (Arvaniti & Garding, 2007; Clopper & Smiljanic, 2011).

1.2.1 Variability between individual talkers

Similar to variability in segmental information, suprasegmental information also varies between *individual* talkers (i.e., within a relatively homogeneous population). For instance, Niebuhr et al. (2011) found talker-specific production strategies to differentiate between two pitch accent categories in German and in Italian. Specifically, they found that talkers either relied more strongly on F0 peak alignment to the vowel onset, or more strongly on the shape of the pitch accent (a more or less symmetrical shape). Cangemi et al. (2015) further found similar talker-specific differences in production of focus in German. They found that talkers differentiated between broad, narrow, and contrastive focus by varying the number of used acoustic cues and the informativeness (i.e., how strongly a cue distinguishes between categories) of each cue.

More recently, Xie et al. (2021) examined how variability is manifested in the distributional structure of phonetic cues to sentence prosody. They recorded native talkers of American English, producing declarative questions vs. statements (e.g., 'It's raining?' vs. 'It's raining'), and analyzed F0 and duration in the final overlapping syllable (i.e., '-ing'). Results indicated that talkers differed from each other in two ways. First, they differed in the category means and shapes of the F0 and duration distributions, showing that some talkers produced the sentences with different mean values, and a wider range of F0 and duration values than others. Second, talkers differed in how correlated the cues were. That is, while F0 was the primary cue for most talkers, some talkers also used duration to a higher degree. In sum, these studies suggest that the two types of variability found in segmental information (i.e., within-cue variability and differences in phonetic cue weighting) are also present in sentence prosody.

1.2.2 Acoustic correlates of word stress

Another form of prosody is lexical stress, a structural, linguistic property of a word that specifies which syllable in the word is acoustically more prominent than any of the others. Depending on the language one speaks, lexical stress has important consequences for word recognition. For instance, in English, the words 'SUBject' vs. 'subJECT' are segmentally identical, but the stress pattern determines their meaning. It appears that listeners are not deaf to these differences but actually benefit from using lexical stress in online word recognition to correctly perceive spoken words (Cutler, 1986; Cutler & Van Donselaar, 2001; Reinisch et al., 2010; Sulpizio & McQueen, 2012).

Previous research has identified a number of acoustic cues that signal lexical stress. In most languages, including Dutch (Rietveld & Van Heuven, 2009), a stressed syllable is produced using a higher mean F0, higher intensity, and longer duration compared to an unstressed syllable (for review, see Gordon & Roettger, 2017). Importantly, researchers have argued that F0 is not an acoustic correlate of lexical stress but of sentential accentuation (e.g., in English: Beckman & Edwards, 1994; Pierrehumbert, 1980). Moreover, as Roettger & Gordon (2017) point out, since pitch accents often co-occur with stressed syllables, one should be cautious with ascribing findings to word-level stress. In the present study, we do not aim to draw conclusions about whether our findings originate from sentential accentuation or lexical stress, but aim instead to describe the acoustic properties and variability of stress in accented and unaccented words[1]. We consider F0 as a necessary part of this description. An additional reason for doing so comes from the perspective of the listener: Even if F0 were specified in speech production to mark sentential accentuation, pitch accents will typically surface on a stressed syllable and thus still serve as an indirect cue to word stress in speech perception.

Further examining the interplay between word stress and sentential accentuation, Sluijter & Van Heuven (1996) previously found that syllable duration, overall intensity, and spectral tilt were affected by both sentential accentuation and word stress. Specifically, stressed syllables were longer, louder, and had shallower spectral tilt in stressed syllables, but even more so when these appeared in an accented word. However, no such strengthening effects of sentential accentuation stress were found for vowel quality. This was further confirmed by van Bergem (1993), who found that the contribution of sentential accentuation on acoustic vowel reduction was of minor importance compared to word stress.

Other acoustic cues to stress have also been identified, such as spectral tilt (Sluijter & Van Heuven, 1996) and F0 variation (measured as F0 slope in Plag et al., 2011). Finally, while vowel quality is a reliable cue to lexical stress in English, other languages, such as Dutch, do not rely that heavily on vowel quality (Cutler, 1986; Cutler & Pasveer, 2006). Nonetheless, some vowel quality reduction has still been found in Dutch (van Bergem, 1993). As van Bergem (1993) states, this involves acoustic vowel reduction (i.e., a more centralized position in the vowel space) instead of lexical vowel reduction (i.e., replacing a full vowel with a schwa, as is the case in English).

It is important to mention that these cues to word stress are not weighted equally in production. Similar to phonetic cue-weighting tendencies in segmental information (Schertz & Clare, 2020), different cue hierarchies exist in different languages. For instance, in Dutch, F0 is considered to be the strongest cue to word stress, but only when the word appears in an accented position (e.g., the word 'yellow' in 'Did you mean the green circle? No, I meant the yellow circle'). When the word appears in an unaccented position (e.g., 'Did you mean the yellow square? No, I meant the yellow circle'), duration serves as the strongest cue, followed by spectral tilt, intensity and vowel quality (Rietveld & Van Heuven, 2009). In English, there is a much larger role for segmental differences in productions of word stress (Cutler, 1986; Cutler et al., 2007). Specifically, unstressed syllables often contain reduced vowels, and the suprasegmental cues play a smaller part in word stress production, as corroborated by acoustic comparisons of L1 English and Dutch L2 English (Braun et al., 2008; Cooper et al., 2002).

Acoustic realizations of word stress are further affected by demographic groups (e.g., gender), speaking context, and the native language one speaks. Eriksson et al. (2016) measured acoustic correlates of word stress in Italian talkers across different speaking contexts. They recorded male and female participants while producing spontaneous speech, words in word lists, and in isolation. Using these recordings, mean F0, F0 variation (measured as the standard deviation of F0), duration and spectral tilt were measured in stressed and unstressed syllables. Results showed several differences between gender and speaking contexts. First, the difference in mean F0 between stressed and unstressed syllables was larger for women than for men. Second, women produced unstressed syllables with more F0 variation compared to men. Third,

[1] For the rest of the paper, we will use the term 'word stress' as we will measure acoustic properties of single words. We do not make any claims as to whether the results originate from lexical stress or sentential accentuation.

the duration of syllables was longer for women. Moreover, this difference was even larger for stressed syllables. Fourth, men produced stressed syllables with steeper spectral tilt. In addition to these gender differences, the speaking context (word lists, phrases or spontaneous speech) added to the variation in the abovementioned acoustic cues. In English, similar results have been found, except for the difference in mean F0, which was larger in men, and spectral tilt, which showed no difference between genders (Eriksson & Heldner, 2015). Variability caused by different native languages has been observed by Tseng et al. (2013), who found differences in F0-usage between L1 Taiwan Mandarin and L1 Beijing Mandarin talkers of English compared to L1 English talkers, illustrating that even within the same target language, productions of word stress are affected by the talker's L1.

1.2.3 Word stress perception

Despite the presence of variability in word stress, evidence from speech perception experiments shows that listeners are still able to correctly perceive spoken words. With regard to L2 stress perception, for instance, listeners can adapt to non-canonically produced words (i.e., with the incorrect stress pattern), spoken by L2 talkers (Reinisch & Weber, 2012). Further, listeners can also track talker-specific usage of acoustic cues to word stress in L1 talkers. For instance, listeners can learn how a given talker produces word stress and generalize this learning to the perception of new words from this talker (Bosker, 2022). Furthermore, Severijnen et al. (2021) taught participants to map novel non-word minimal stress pairs onto different object referents (e.g., *USklot* meant 'lamp', *usKLOT* meant 'train'). Crucially, the non-words were spoken by two male talkers who used different cues to word stress (e.g., Talker 1 used only F0 while Talker 2 used only intensity). In a subsequent test phase, participants heard semantically constraining sentences containing the non-words (e.g., 'The word for lamp is *USklot'*) produced with either the expected cue (e.g., Talker 1 using F0) or the unexpected cue (e.g., Talker 1 using intensity). Results on a sentence verification task showed that participants were slowed down in a 2AFC task when presented with the unexpected cues, illustrating that participants had learned information about which cue was used by either talker, and used this in perception. Converging evidence has also been found in a similar experiment using existing Dutch words (Severijnen et al., 2023).

In sum, previous research suggests that several factors (e.g., gender, production strategies) affect acoustic realizations of word stress. However, descriptions of these acoustic realizations have previously been limited to group-level differences while there are reasons to believe that individual talker differences in word stress are also present. First, previous research has shown that individual differences in sentence prosody are present in speech (Cangemi et al., 2015; Niebuhr et al., 2011; Xie et al., 2021). Second, perception experiments illustrate that listeners can adapt to how individual talkers produce word stress (Severijnen et al., 2021). If this talker-specific learning mechanism reflects a task that listeners are faced with in real-life speech perception, it suggests that variability between individual talkers is also present in word stress. However, we know surprisingly little about how individual talkers actually differ in their productions of word stress.

1.3 The present study

The present study therefore asks: How do individual Dutch talkers differ in how they produce word stress? To address this, we recorded 40 native speakers of Dutch, from a participant pool in which we minimized dialectal variation and balanced gender, producing Dutch word pairs with segmentally overlapping syllables but differing in lexical stress (e.g., *VOORnaam* vs. *voorNAAM*, 'first name' vs. 'respectable'). To measure the acoustic correlates of word stress in accented and unaccented words separately, as well as in isolation and sentence context, the target words were recorded in three speaking conditions: in isolation, in an accented position in a sentence (e.g., *Toen had Jan het woord <u>voorNAAM</u>* gezegd, [Then had Jan the word <u>respectable</u> said], 'Then, Jan said the word <u>respectable</u>'), and in an unaccented position in a sentence (*Daarna had <u>Koen</u> het woord voorNAAM gezegd,* [Afterwards had <u>Koen</u> the word respectable said], 'Afterwards, <u>Koen</u> said the word respectable'; <u>underlining indicates the</u>

accented word). We then measured, for each individual participant, a set of acoustic cues to stress in Dutch: mean F0, duration, intensity, spectral tilt, vowel quality, and F0 variation.

The present study has two main goals. First, while this was not the main goal of the study, we aim to describe the group-level acoustic correlates of word stress in Dutch. This is certainly not the first study to report on this (cf. Sluijter & Van Heuven, 1996), but we aim to contribute to the literature by testing a larger number of participants, producing a larger number of different words, and measuring a range of cues to word stress. Second, our main goal was to examine how individual talkers produce word stress. In other words, we examine whether individual talkers follow the group-level patterns or whether individual talkers adopt their own production strategies. While previous studies have already established such variability for sentence prosody (Cangemi et al., 2015; Niebuhr et al., 2011; Xie et al., 2021), we aim to extend this to word stress.

Regarding the group-level results, we had the following predictions. In line with previous literature on word stress (for review, see Gordon & Roettger, 2017), we expected stressed syllables to have a longer duration, higher intensity, shallower spectral tilt (i.e., a shallower slope downwards from low to high frequencies; Hayward, 2000, p. 243), and acoustically fuller (i.e., less reduced) vowels. Moreover, we expected stressed syllables in accented words to have a higher mean F0 and a larger F0 variation (measured as F0 range) which would indicate the presence of a pitch accent on the stressed syllable. Following the results in Sluijter & Van Heuven (1996), we further expected an enhancing effect of sentential accentuation on stressed syllables, resulting in an even longer duration, higher overall intensity, and shallower spectral tilt compared to stressed syllables in unaccented words. Regarding gender differences, based on Eriksson et al. (2016), we expected larger differences between stressed and unstressed syllables for women in mean F0 (but see Eriksson & Heldner, 2015), F0 variation and duration. We also expected steeper spectral tilt, for men in stressed syllables. We did not have specific predictions for effects of stress on different syllable positions.

Further, regarding cue weighting, we expected that in accented words, mean F0, duration and intensity are the three strongest cues while in unaccented words, duration and intensity are the strongest cues (Rietveld & Van Heuven, 2009). Spectral tilt has previously been found to be a strong cue to lexical stress in Dutch (Sluijter & Van Heuven, 1996), but recent evidence showed that this was largely due to the vowels that were tested (Severijnen et al., 2022). Therefore, we expected that spectral tilt will be a relatively weak cue in the present study, in which a more representative set of vowels was tested. Moreover, we expected talkers to rely less on vowel quality (cf. Sluijter & Van Heuven, 1996) and F0 variation (cf. Plag et al., 2011). Note that we did not make predictions about the shape of the F0 contour in the syllable. The reason for this is that this acoustic feature is difficult to quantify into one single measure, which is required for the LDA. Therefore, we restricted ourselves to mean F0 and F0 variation, which we believe best capture the F0 dynamics of the entire syllable in two complementary measures.

To examine the variability in word stress, we conducted two types of analyses. First, we looked into variability across talkers *within* each acoustic cue. In line with previous studies, we expected to find large between-talker variability in cue means and distributions. Similar to Xie et al. (2021), we descriptively illustrated each participants' acoustic values and compared them to the group mean for each cue. To further test this statistically, we built linear mixed effects models for each acoustic cue with the critical predictor Stress Status and added by-participant random slopes for Stress Status to the final model (Allen et al., 2003; Clayards, 2018). We expected that the models with by-participant random slopes would improve the model fit to the data compared to the model without random slopes, confirming that taking between-talker variability into account would significantly improve the model.

Second, we examined each participants' cue-weighting tendency (i.e., variability *between* acoustic cues). Following the approach in Schertz et al. (2015), we quantified these weights using LDAs, resulting in a set of cue weights for each individual participant. We expected that each participant would employ a unique set of cue weights when producing word stress, and explored the possible presence of cue-weighting tendencies. Note that both analyses

are mostly descriptive in nature. That is, we did not test specific hypotheses, but rather tried to provide a thorough description of acoustic variability in word stress in Dutch.

2.0 Materials and methods

2.1 Participants

We recruited 47 native talkers of Dutch from the Radboud University participants pool. All participants gave informed consent and received a monetary reward or course credits for their participation. Seven participants were excluded for not following the correct instructions to the experimental task. The remaining 40 participants reported not having any hearing and/or reading problems (20 male, 20 female, age range: 17-33, M_{age} = 22.5, SD_{age} = 4.03). The study was approved by the Ethics committee of the Social Sciences faculty of Radboud University (project code: ECSW2016-1403-391).

To reduce acoustic variability caused by Dutch dialects or accented Dutch, we attempted to recruit only participants speaking the standard Dutch variety by screening participants before inviting them to participate. Since screening regional accents is highly subjective, we adopted the self-report approach offered by Pinget (2015). In an online survey, participants responded to the following two questions: (1) When speaking standard Dutch, do you think you speak accented Dutch? (2) When speaking standard Dutch, do others perceive your speech as having an accent? Only participants who responded "no" to both questions were invited to participate.

By means of a Language Background Questionnaire (see section 2.2.2), we assessed participants' linguistic background. Six participants reported being raised bilingually, and four spoke a Dutch dialect next to standard Dutch. Among the languages the bilingual participants spoke, were Afghan, English, Mandarin, Italian, Russian, and German. Among the dialects spoken by the participants were "Brabants" (spoken in the province of Noord-Brabant), "Arnhems" (spoken in the city Arnhem), and "Sallands" (spoken in the province of Overijssel).

The majority of the participants grew up in the province of Gelderland (21/40), the province where the experiment was carried out. The rest of the participants came from different regions of the Netherlands: Noord-Brabant (5/40), Utrecht (3/40), "the Achterhoek" (3/40) close to the German border, Friesland (1/40), "the Randstad" (3/40), Zeeland (1/40), Limburg (1/40), and Overijssel (2/40). For further details about language exposure and proficiency, see Supplementary Information (section 2.3).

2.2 Materials

2.2.1 Stimuli

The study aim was to measure acoustic cues to word stress. To minimize any influence from segmental information in words, we carefully selected a set of target words with identical consonants and vowels that differed only in lexical stress. This set included 6 Dutch disyllabic minimal stress pairs, differing in stress pattern (e.g., *VOORnaam* vs. *voorNAAM,* 'first name' vs. 'respectable', capitalization indicates lexical stress) and 56 partially overlapping disyllabic pairs (i.e., words with one fully overlapping syllable and a variable number of overlapping segments in the other syllable; *TAlen* /ˈtaːlən/ vs. *taLENT* /taːˈlɛnt/, 'languages' vs. 'talent'). The partially overlapping pairs were included to increase external validity, since the number of Dutch minimal stress pairs is limited. In these partially overlapping words, we measured prosodic cues only in the overlapping syllable (e.g., *ta*). We selected 28 first-syllable partially overlapping pairs and 28 second-syllable pairs which resulted, together with the minimal stress pairs, in a total number of 62 word pairs (for the complete stimulus list, see Supplementary Table S1). Note that 24 of the SW and 12 of the WS partially overlapping words were morphologically complex words. The number of morphologically complex words was thus not perfectly balanced between SW and WS items. We nevertheless used this selection of words because (1) to our knowledge, it remains unclear how morphology would affect lexical stress

production and (2) we were constrained by the limited number of words with the same segmental structure in Dutch. In section 3.1.2, we assessed to what extent differences between items contributed to the observed results.

Next, we created a set of carrier sentences in which the target words would appear, so that we could attempt to separate the acoustic correlates of word stress from those of sentential accentuation (cf. Sluijter & Van Heuven, 1996). Target words appeared once in an accented position and once in an unaccented position. The sentences were created in such a way that they would naturally induce correct sentential accentuation placement, but for clarity, we underlined the accented words in the recording script that we gave to the participants. Moreover, the stressed syllables in the target words were given in capitals. An example of the sentences for the word *voorNAAM* ('respectable') is:

Example 1.

(1) *Eerst had Jan met enthousiasme <u>boot</u> gezegd,*
'First had Jan with enthusiasm <u>boat</u> said,'

(2) Toen had Jan het woord <u>voorNAAM</u> gezegd,
'Then had Jan the word <u>respectable</u> said,'

(2) Daarna had <u>Koen</u> het woord voorNAAM gezegd.
'Afterwards had <u>Koen</u> the word respectable said.'

Finally, to avoid sentence-final prosodic properties (e.g., sentence-final lengthening, amplitude and F0 drop), the target words never appeared at the end of the sentence. The two latter sentences containing the target words remained identical across all trials except for the name of the actor (e.g., *Jan, Koen*) and the target word (*voorNAAM*).

2.2.2 Language background questionnaire

We created a language background questionnaire to obtain a clear image of the linguistic background of each participant. This questionnaire addressed the following three issues (for the complete questionnaire, see Supplementary Information, section 1). First, participants were asked to report which languages they spoke, including second languages, dialects, and whether they were raised bilingually. Second, they reported on the languages they had been exposed to during their childhood. More specifically, they reported the region in which they and their parents grew up as well as which language/dialect their parents spoke to them. Third, participants were asked to rate their proficiency of each language they spoke on a 7-point scale from 1 (not proficient) to 7 (native) for speaking, listening, writing, and reading. Finally, participants gave the age of first exposure to each language. In sum, this questionnaire provided us with a clear impression of the languages and dialects each participant spoke and which languages participants had been exposed to throughout their lives.

2.3 Procedure

The experiment consisted of one single recording session per participant. Participants were seated in a sound-attenuating booth and wore a head-mounted Omnitronic HS-1100 microphone. We further used a Behringer X-Air XR18 mixer and the recordings were digitized at a 44.1 kHz. The data were collected in three different recording labs because of various COVID-19 related restrictions, but we ensured that recording settings were as similar as possible by taking the intensity of a silent recording in all labs as a proxy for how much background noise was present.

The target words and sentences were visually presented in Dutch orthography using SpeechRecorder (Draxler & Jänsch, 2004), and participants were instructed to read them aloud as naturally as possible. Participants were informed to pay attention to the capitalization in the target words, and that this indicated which syllable should be stressed. Each trial consisted of two speaking conditions: the target word was first produced in isolation, followed by the three carrier sentences as in Example 1. If a mispronunciation or a disfluency was present in the recording, as detected by the experimenter during the session, the same trial was repeated. Trials

were presented in a pseudorandomized order, ensuring that the two members of the same word pair were at least 62 trials apart. The stimulus list was repeated twice, each time in a different order, and was preceded by four practice trials with words that did not appear in the experiment. All participants received the same order of presentation, which reduced any between-talker variability caused by possible order effects. After the recording session, participants completed the Language Background Questionnaire and were debriefed on the purpose of the experiment.

2.4 Data analysis

2.4.1 Acoustic measurements

The recordings were automatically forced-aligned using the WebMAUS Basic tool (Kisler et al., 2017), which segmented the target words and their individual segments. The resulting annotated TextGrid files were subsequently manually checked by six phonetically trained research assistants who additionally segmented the syllables in the target words. Reliability analyses confirmed that there was limited variation between the researchers' annotations (for details, see Supplementary Information, section 2.1).

We measured six acoustic cues to word stress. These cues were mean F0, duration, intensity, spectral tilt, vowel quality, and F0 variation. All acoustic cues were measured using Praat (Boersma & Weenink, 2019) and processed in R (R Core Team, 2020). For each minimal stress pair (e.g., *VOORnaam* vs. *voorNAAM*), we measured these acoustic cues in both syllables. For each partially overlapping pair (e.g., *Talen* vs. *taLENT*), we only measured the acoustic cues on the overlapping syllable (*ta*). Prior to performing the measurements, we excluded any trials that contained a disfluency or noise in the recording ($N = 30$ trials; < 0.01%).

The measurements for mean F0, duration, Intensity, and F0 variation were performed across the entire syllable instead of the vowel because the segmental information in the stimuli was identical across the measured stressed and unstressed syllables. Thus, any observed differences could not be due to different consonants in the syllable[2]. If applicable, any additional justification for measuring across the entire syllable for specific cues is provided in the corresponding sections below.

2.4.1.1 Mean fundamental frequency (F0)

Mean F0 in Hertz was measured in the voiced part of each syllable using the 'To Pitch…' function in Praat. The voiced part included the vowel of the syllable, and any voiced part in voiced stops (e.g., /b/). We used different pitch settings for males and female participants (male: 75-250 Hz, female: 100-500 Hz, time step = 10 ms). To reduce the number of measurement errors, we automatically identified syllables containing octave jumps and measurement errors, and re-measured these using participant-specific pitch settings (for details, see Supplementary Information, section 2.2). After this pre-processing procedure, we excluded observations, separately for each participant, that were more extreme than $M_{F0\ talker} \pm 3 * SD_{F0\ talker}$ ($N = 125$ tokens; < 0.01%). Finally, the F0 measurements were converted to semitones relative to 50 Hz. This semitone conversion transformed raw Hz values to a log-scale, and thus accounted for higher dispersion of higher frequencies (cf. Clayards, 2018). The resulting distributions were normally distributed, which is more appropriate for the statistical analyses.

2.4.1.2 Duration

Duration was computed by measuring the total syllable duration and the values were converted to log-scale. This accounted for skewed distributions, again making the resulting distributions more suitable for statistical analyses. We measured total syllable duration (instead of vowel duration) because greater articulatory effort due to the realization of stress has been previously linked to an increase in consonant duration (Slis, 1971). This has been found for some Dutch consonants (Cho & McQueen, 2005; Nooteboom, 1972) but not consistently (Cho

[2] To confirm, we ran the same analyses but with intensity, mean F0, and F0 variation measured on the vowel. These yielded similar results as the analyses on the syllables (See Supplementary Tables S8 – S10).

and McQueen observed prosodic lengthening in stressed syllables for /d, s, z/ and for /t/ closure duration, but shortening of VOT in stressed syllables for /t/). Given these effects, it seemed appropriate to include consonant duration in the analyses and thus used total syllable duration.

2.4.1.3 Intensity

We measured overall intensity in dB in the entire syllable using the 'Get intensity (dB)…' function in Praat. We took intensity as an absolute measure (instead of relative to the previous syllable) because the stimulus list also included partially overlapping pairs. Since the intensity in the non-overlapping syllable is presumably affected by the varying segmental structure, it cannot serve as a controlled reference value. For this reason, we only measured intensity in the overlapping syllable.

2.4.1.4 Spectral tilt

Spectral tilt was measured by computing a linear regression on the spectrum of the vowel in the target syllable. First, we measured the power in frequency bins of 10 Hz from 0-4000 Hz using the 'Tabulate…' function in Praat. The center frequency of each bin was then converted to semitones relative to 50 Hz. Next, we ran a linear regression on the spectrum of each vowel and took the slope as a measure of spectral tilt. We opted for this method to avoid defining a priori frequency bands, which may strongly affect the results (cf. Severijnen et al., 2022).

2.4.1.5 Vowel quality

We followed the procedure in Karlsson & van Doorn (2012) to quantify vowel quality in terms of vowel formant dispersion (VFD). VFD is calculated by taking the Euclidean distance from a centroid location (i.e., a weighted midpoint of all the vowels) to the location of each vowel in the F1/F2 vowel space (all in raw Hz). This makes it possible to compare changes across conditions and talkers (Karlsson & van Doorn, 2012). To obtain this measure, we first measured F1 and F2 values in the middle 1/3 part of each vowel. We only included monophthongs for these measurements (only 8% of the data contained diphthongs). Following Escudero et al. (2009), we calculated participant-specific formant settings for formant estimation. We then calculated the centroid location, which was defined as the mean F1 and a weighted mean F2. The weighted mean F2 is calculated as the mean F2 of only the vowels with an F1 lower than the mean F1, and thus only based on the more closed vowels. As Karlsson & van Doorn (2012) point out, a weighted F2 is used for two reasons. First, it keeps the vowel space center fixed across systems containing three or four vowels, improving comparability across systems. Second, the weighted F2 is more suitable for a system in which front, open vowels are missing. Next, the VFD was calculated for each participant and vowel separately.

2.4.1.6 Fundamental frequency variation

F0 variation was computed by subtracting the minimum F0 value from the maximum F0 value in each syllable. We first measured the maximum and minimum F0 in each syllable, and performed semitone conversions (relative to 50 Hz). The F0 minimum (in semitones) was then subtracted from the F0 maximum (in semitones) to obtain our measure of variation. Positive values thus indicate larger variation.

2.4.2 Linear Mixed Effects Models

The acoustic measures were analyzed at the group level using linear mixed effects models with the *lmerTest* package (Kuznetsova et al., 2017) in R (R Core Team, 2020). A model was built for each acoustic cue separately to examine the effect of stress pattern and other predictors (e.g., sentential accentuation, gender) on the acoustic measures. We then performed pairwise comparisons for any significant interactions using *emmeans* (Length, 2022). Furthermore, we established whether there was significant evidence for between-talker variability in each cue by comparing the final model without a by-participant random slope for word stress to a model with one. For the analyses for vowel quality and spectral tilt, syllables

containing diphthongs were excluded, resulting in 30, 080 observations. The analyses for all other cues included both monophthongs and diphthongs, resulting in 32, 474 observations.

For each acoustic cue, we tested for the following fixed factors: Stress Status (categorical predictor with two levels, deviance coding with unstressed coded as -0.5 and stressed coded as 0.5), Speaking Condition (categorical predictor with three levels, dummy coding with the unaccented condition mapped onto the intercept), Gender (categorical predictor with two levels, deviance coding with male coded as -0.5 and female coded as 0.5), Syllable (categorical predictor with two levels, deviance coded with the first syllable coded as -0.5 and the second syllable coded as 0.5). We also tested for interactions between Stress Status and Speaking Condition, Stress Status and Gender, and Stress Status and Syllable. Each model included random intercepts for Item (individual items, not item pairs) and Participant, and we added by-participant random slopes for the fixed effects using forward modeling. We selected the model that (1) successfully converged and (2) demonstrated the best fit to the data using log-likelihood model comparisons. Addition of by-participant random slopes for Stress Status are assessed in section 3.2.2. For the final model for each cue, see Supplementary Tables S2-S7.

2.4.3 Linear Discriminant Analyses

We ran Linear Discriminant Analyses (LDA) to obtain sets of cue weights for each individual participant. An LDA tries to find the optimal linear combination of a set of predictors (e.g., acoustic cues to word stress) to separate a dataset into different classes (e.g., stressed vs. unstressed syllables). Following the procedure in Schertz et al. (2015), the LDA models were built and tested on the same data, which makes them descriptive rather than predictive. For these analyses, we wanted to exclude overall mean differences between talkers because we were interested only in the magnitude of change for each cue (i.e., how much a cue is increased or decreased). We created these 'talker-normalized' data as follows. For duration and intensity, the talker mean of each corresponding cue was subtracted from the value of each individual. For mean F0 and F0 variation, we converted the observed frequency in Hz to semitones relative to the talker mean. The slope of spectral tilt was measured based on frequency bands relative to the talker mean. Finally, vowel quality (measured as VFD) could already be considered a talker-normalized cue, since each measure is relative to a talker-specific centroid location. We then used these cues as predictors, and attempted to predict the stress/unstressed status in each syllable in each participant's productions. Using the *lda* function from the MASS package in R (Venables & Ripley, 2002), we built a model for each individual participant. The output coefficients from each model represented how much each cue is weighted in producing word stress by a given talker. All cues were converted to z-scores prior to running the model, which means that the coefficients are comparable even when originating from different acoustic dimensions.

3.0 Results

We present the results in different steps. First, we present the group-level results, including results from linear mixed effects models and group means of the LDAs, which will inform us on general tendencies of word stress production in Dutch, and provide a basis for interpreting the individual differences. Third, we will present several analyses and visualizations that address how individual talkers vary. These will include visualizations of cue distributions from individual talkers, comparisons of linear mixed effects models, and results from LDAs on individual participants. We refer the reader to our Data Availability Statement for access to the raw data.

3.1 Group-level correlates of word stress

Means for the raw measurements of each cue are provided in Table 1, averaging across the various conditions (isolation, accented, unaccented). Density plots for the talker-normalized cues are depicted in Figure 1. Moreover, an acoustic vowel diagram, illustrating formant values for stressed and unstressed vowels, is depicted in Figure 2. For all acoustic cues, the difference between stressed and unstressed syllables is in the expected direction: a stressed syllable is produced with a higher mean F0, longer duration, higher intensity, shallower spectral tilt,

greater spectral vowel dispersion, and more F0 variation. In addition, while there is a mean difference between stressed and unstressed syllables for all cues, Figure 1 shows considerable overlap between the cue distributions for stressed and unstressed productions. Finally, while unstressed vowels are overall more acoustically reduced, Figure 2 also shows large variability between different vowels. More specifically, the low (open) vowels appear to be more strongly reduced than the high (closed) vowels.

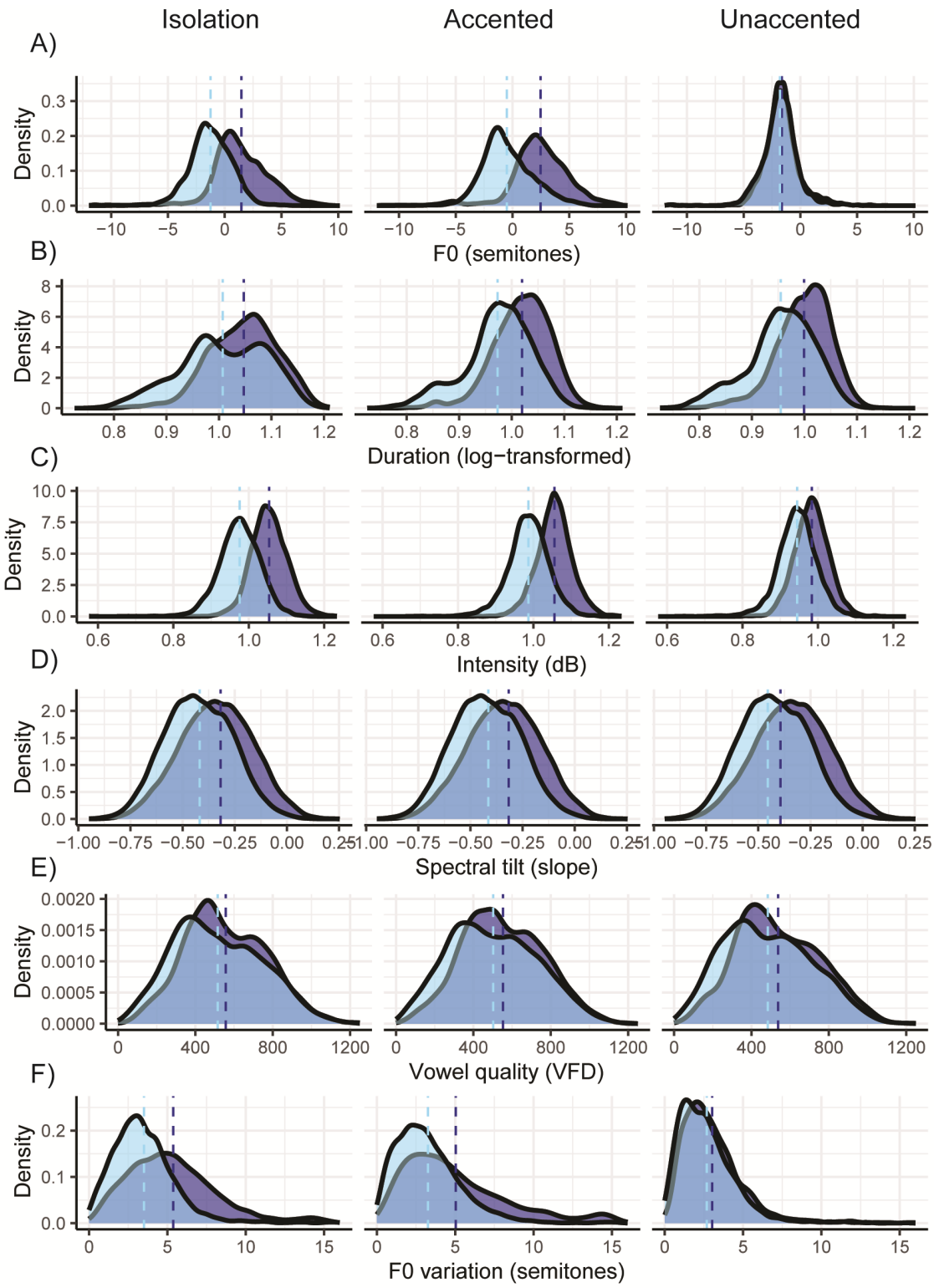

Figure 1.

Density plots of each cue for stressed (in dark blue) and unstressed (in light blue) productions. The plots are depicted separately for the isolation (left), accented (middle), and unaccented (right) condition. The dotted lines represent the means. All plots represent the talker-normalized measures. As the figure shows, stressed syllables generally had higher mean values compared to their unstressed counterparts in all conditions and measures. For mean F0 and F0 variation in the unaccented condition, this difference was reduced.

Table 1
Cue means for stressed and unstressed syllables, averaged across conditions (isolation, accented, and unaccented).

Syllable	**1**		**2**	
Cue	**Stressed**	**Unstressed**	**Stressed**	**Unstressed**
Mean F0 (Hz)	174	158	162	142
Mean F0 (semitones)	20.46	18.95	19.35	17.20
Duration (ms)	232	169	321	278
Intensity (dB)	70	65	68	64
Spectral tilt (slope)	-0.37	-0.46	-0.31	-0.39
Vowel quality (VFD)	549	483	549	521
F0 variation (Hz)	40	26	49	31
F0 variation (semitones)	3.90	2.78	5.07	3.55

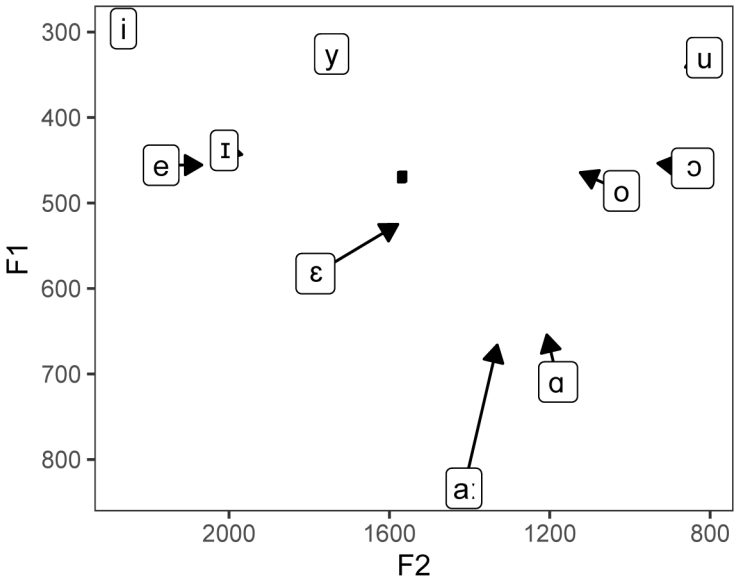

Figure 2.
Acoustic vowel plot (F_1 and F_2 values in Hz) for monophthongs, averaged across conditions (isolation, accented, and unaccented). The arrows indicate the shift from a vowel produced in a stressed syllable to one produced in an unstressed syllable. The dot in the middle represents the centroid location of all vowels.

3.1.1 Linear Mixed Effects models

Linear mixed effects models were performed on each cue separately. Below, we will discuss the effects for Stress Status, Speaking Condition, Gender, and any interactions. The full model output is provided in the Supplementary Tables S2-S7.

3.1.1.1 Mean fundamental frequency (F0)

The model for mean F0 revealed a main effect of Stress Status (β = 0.40, *SE* = 0.14, *t* = 2.79, *p* < .05), indicating a higher mean F0 in stressed syllables compared to unstressed syllables. Moreover, the model revealed a larger mean F0 difference between stressed and unstressed syllables in the accented condition compared to the unaccented condition (β = 2.74, *SE* = 0.05 *t* = 57.39, *p* < .001), and in the isolation condition compared to the unaccented condition (β = 2.50, *SE* = 0.05 *t* = 52.22, *p* < .001). Inspecting the beta estimates, this shows that the effect of word stress on mean F0 was nearly canceled out in the unaccented condition, suggesting that only in accented words, mean F0 is an indirect cue to word stress. That is, mean F0 is potentially a cue to accentedness, which surfaces as a pitch accent on the stressed syllable. This was further confirmed by pairwise comparisons that showed a significant, but smaller pairwise effect in the unaccented condition (Δ = 0.40, *SE* = 0.144, *z* = 2.79, *p* < .01), compared to the isolation (Δ = 2.89, *SE* = 0.144, *z* = 20.06, *p* < .0001) and accented condition (Δ = 3.14, *SE* = 0.144, *z* = 21.68, *p* < .0001). Finally, a main effect of Gender was found (β = 10.75, *SE* = 0.51 *t* = 20.88, *p* < .001), illustrating an overall higher mean F0 for women compared to men. No interaction between Stress Status and Gender was found.

3.1.1.2 Duration

The model for duration revealed a main effect of Stress Status (β = 0.28, *SE* = 0.008, *t* = 34.36, *p* < .001), illustrating longer durations in stressed syllables compared to unstressed syllables. Furthermore, we found significant interactions between Stress Status and the accented condition (β = 0.01, *SE* = 0.005, *t* = 2.67, *p* < .01), and between Stress Status and the isolation condition (β = -0.02, *SE* = 0.005, *t* = -3.68, *p* < .0001). Results from pairwise comparison illustrated that the difference between stressed and unstressed syllables was largest in the accented condition (Δ = 0.29, *SE* = 0.008, *z* = 36.01, *p* < .0001), followed by the unaccented condition (Δ = 0.28, *SE* = 0.008, *z* = 34.37, *p* < .0001), and the isolation condition (Δ = 0.26, *SE* = 0.008, *z* = 32.10, *p* < .0001). This further suggests that duration is a stronger cue to word stress when in sentence context compared to isolation. Finally, we found a main effect of Gender (β = 0.08, *SE* = 0.02, *t* = 3.33, *p* < .05), illustrating that women produced longer syllables than men, irrespective of word stress.

3.1.1.3 Intensity

The model for intensity revealed a main effect for Stress Status (β = 2.84, *SE* = 0.02, *t* = 15.33, *p* < .001), confirming that stressed syllables are produced with a higher intensity. We further found significant interactions between Stress Status and the accented condition (β = 1.96, *SE* = 0.07, *t* = 29.59, *p* < .001), and between Stress Status and the isolation condition (β = 2.61, *SE* = 0.06, *t* = 39.37, *p* < .001). Results from pairwise comparisons revealed that the difference between stressed and unstressed syllables is largest in the isolation condition (Δ = 5.44, *SE* = 0.185, *z* = 29.43, *p* < .0001), followed by the accented condition (Δ = 4.80, *SE* = 0.185, *z* = 25.93, *p* < .0001), and the unaccented condition (Δ = 2.84, *SE* = 0.185, *z* = 15.33, *p* < .0001). These results suggest that intensity is a stronger cue to word stress when produced in accented words (isolation and accented condition) compared to unaccented words.

3.1.1.4 Spectral tilt

The model for spectral tilt revealed a main effect of Stress Status (β = 0.07, *SE* = 0.006, *t* = 12.85, *p* < .001), illustrating that the spectral slope is less negative (i.e., shallower) for stressed vowels. We additionally found an interaction between Stress Status and the accented condition (β = 0.04, *SE* = 0.002, *t* = 16.72, *p* < .001), and between Stress Status and the unaccented condition (β = 0.04, *SE* = 0.002, *t* = 18.77, *p* < .001). Results from pairwise comparisons revealed that the effect of word stress largest in the isolation condition (Δ = 0.11, *SE* = 0.006, *z* = 20.48, *p* < .0001), followed by the accented condition (Δ = 0.11, *SE* = 0.006, *z* = 19.65, *p* < .0001), and the unaccented condition (Δ = 0.07, *SE* = 0.006, *z* = 12.85, *p* < .0001), illustrating that the effect of word stress is larger in accented words. Finally, we found a main

effect of Gender (β = 0.07, *SE* = 0.02, t = 3.69, $p < .001$), illustrating that women overall produce vowels with shallower slopes for spectral tilt.

3.1.1.5 Vowel quality

The model for vowel quality revealed a main effect of Stress Status (β = 56.03, *SE* = 5.18, t = 10.82, $p < .001$), illustrating that the VFD is larger for stressed vowels compared to unstressed vowels. In other words, unstressed vowels are closer to the centroid vowel location (i.e., more reduced; see Figure 2). We did not find any interactions with Speaking Condition, suggesting that the amount of reduction was similar across conditions. Finally, we found a main effect of Gender (β = 105.42, *SE* = 14.88, t = 7.09, $p < .001$), illustrating that women overall produce vowels with formant values that are further away from a centroid location.

3.1.1.6 F0 variation

The model for F0 variation revealed a main effect of Stress Status (β = 0.56, *SE* = 0.17, t = 3.29, $p < .01$). Recall that F0 variation was calculated as the difference between $F0_{max}$ and $F0_{min}$. The main effect thus indicates that there is a larger F0 difference in stressed syllables compared to unstressed syllables. Further, we found a significant interaction between Stress Status and the accented condition (β = 1.26, *SE* = 0.06, t = 21.77, $p < .001$), and between Stress Status and the isolation condition (β = 1.32, *SE* = 0.06, t = 22.83, $p < .001$). Results from pairwise comparisons showed that the difference between stressed and unstressed syllables is smaller in the unaccented condition (Δ = 0.56, *SE* = 0.170, z = 3.30, $p < .001$) compared to the isolation (Δ = 1.88, *SE* = 0.170, z = 11.06, $p < .0001$) and accented (Δ = 1.82, *SE* = 0.170, z = 10.71, $p < .0001$) condition. Similar to mean F0, this illustrates that the effect of Stress Status on F0 variation is strongly reduced in the unaccented condition. Finally, we found a main effect of Gender (β = 0.48, *SE* = 0.19, t = 2.28, $p < .05$), illustrating overall larger F0 variation for women compared to men.

3.1.2 Linear Discriminant Analyses

We ran LDAs on each individual participant to assess their cue-weighting tendencies in producing word stress. Mean LDA coefficients, reflecting group-level cue-weighting strategies, are depicted in Figure 3. These coefficients represent the cue weights for each cue, with higher coefficients indicating a higher cue weight (i.e., it is more important in signaling word stress). Figure 3 shows that, on average, talkers use all cues in the same direction. More specifically, talkers primarily used mean F0, intensity, and duration to produce word stress in Dutch in accented words. While the other three cues (spectral tilt, vowel quality, and F0 variation) are also used, they are much weaker in Dutch. Furthermore, the strength of mean F0 is drastically reduced in the unaccented condition, again confirming that mean F0 is a weak cue to word stress for words in unaccented position.

Next, we examined how different factors in the experiment could possibly affect the LDA coefficients. Specifically, we ran separate LDAs for the minimal pairs and the partially overlapping pairs to assess whether there would be any influence from different items on the cue weights. Also, we ran separate LDAs for the bilingual and the monolingual participants, to assess possible influences from the cue weights in the L2. All analyses showed the same ordering of cue weights in either item category and either bi- or monolinguals (see Supplementary Figure S1-2), suggesting that the observed group-level cue weights are relatively stable.

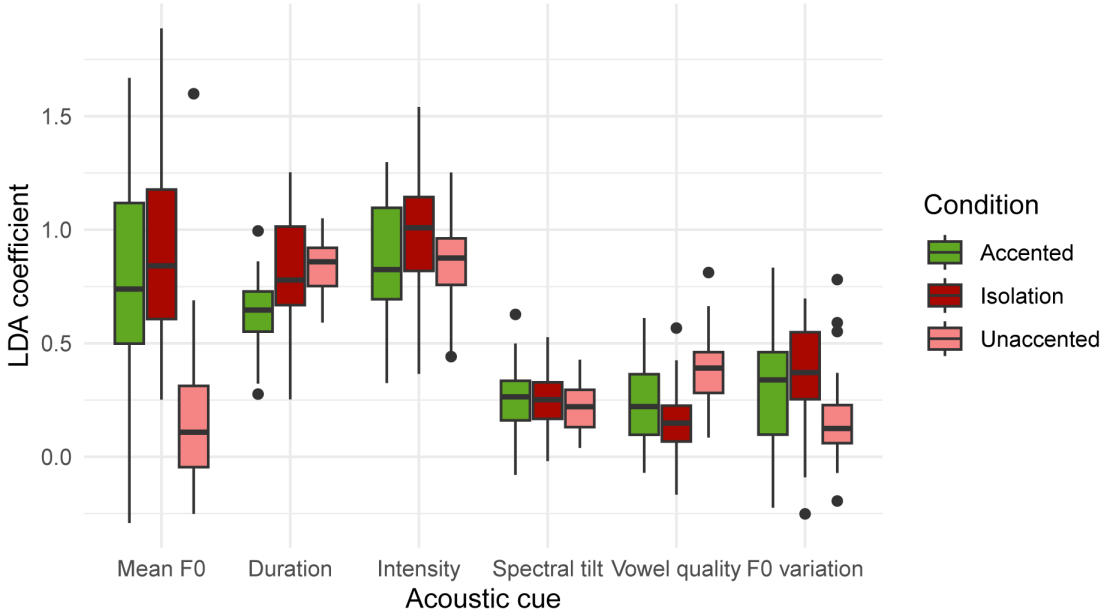

Figure 3.
Boxplots of the mean Linear Discriminant Analyses (LDA) coefficients for different acoustic cues to word stress, when averaging across participants. Cue weights are displayed in three conditions: isolation, accented, and unaccented. Higher coefficients for duration, mean F0, and intensity demonstrate that – on average – these cues are the primary cues to word stress in Dutch.

3.2 Individual differences in production of word stress

In the previous sections we have observed clear differences between stressed and unstressed syllables within acoustic cues (Figure 1). We also described general group-level cue-weighting tendencies in Dutch (Figure 3). In the following sections, we examine how individual talkers vary in how they produce word stress. In other words, do all Dutch talkers follow these mean patterns or do they differ in their usage of the acoustic cues? We will address this question both for within-cue variability and differences in cue-weighting tendencies.

3.2.1 Visualization of cue distributions

To illustrate within-cue variability, Figure 4 visualizes cue distributions of unnormalized duration and intensity values for eight illustrative participants (four per cue; see Supplementary Figures S3-S8 for distributions of all cues and all participants). These plots show a large degree of variability, but also similarities between talkers in their use of intensity and duration. More specifically, some talkers seem to differ from each other, as well as from the group mean (Figure 4 A and B), in their cue means and the shape of the distributions, while others do show more similar distributions. As Figure 4 shows, for duration, talkers 7 and 23 have wider distributions, while talker 32 has much peakier distributions. For talker 15, the distributions for stressed and unstressed syllables even almost completely overlap. The intensity distributions show a similar pattern: talker 1 and 28 have much peakier distributions than talker 18 and 36. This illustrates that the mean distributions, observed on a group-level, are aggregate statistics calculated across the variable behavior of individuals in our participant sample.

3.2.2 Assessing statistical significance of talker variability

Next, we assessed the statistical evidence for talker variability in word stress production. For each cue, we took the linear mixed effects model with the best fit to the data, containing only random intercepts for items and participants, and added by-participant random slopes for word stress to that model. Log-likelihood comparisons between these two models

quantify the statistical evidence, in terms of model fit, for talker variability in word stress production. Results showed that, for all cues, the model with by-participant random slopes for word stress explained significantly more variance in the data (see Table 1), confirming the presence of between-talker variability. While these results provide statistical evidence for the overall presence of talker variability, they do not inform us on how this variability is structured, which was assessed by the following LDA analyses.

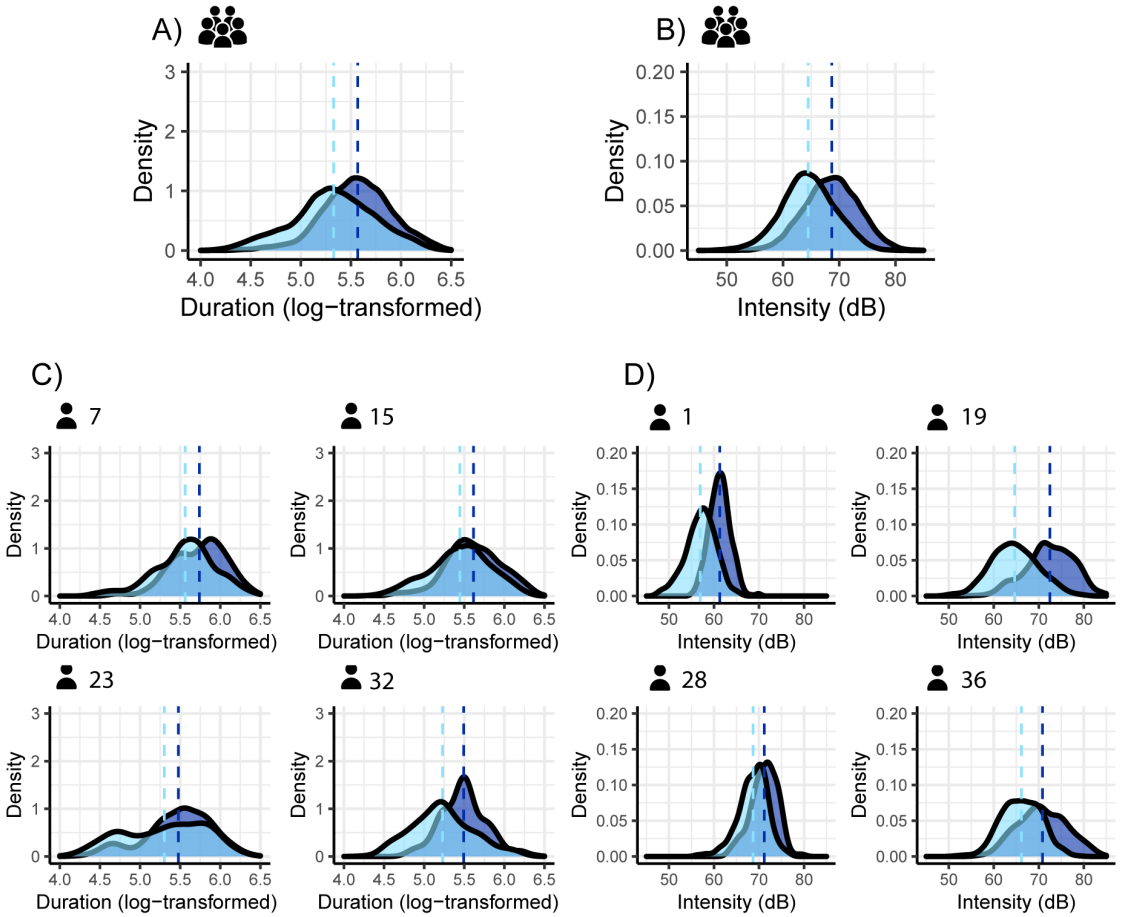

Figure 4.
Distributions of unnormalized duration and intensity values in stressed (dark blue) and unstressed (light blue) syllables. Vertical lines represent means across items. **A.** Mean distributions across all participants for duration. **B.** Mean distributions across all participants for intensity. **C.** Cue distributions from four individual participants (talkers 7, 15, 23, and 32) for duration. **D.** Cue distributions from four individual participants (talkers 1, 19, 28, and 36) for intensity.

Table 1
Results from model comparisons of models with only random intercepts for items and participants, and models with by-participant random slopes. Results are depicted for each cue separately.

Cue	Log-likelihood		Chi-square	*p*
	(1 | Item) + (1 | Participant)	(1 | Item) + (Stress Status | Participant)		
Mean F0	-64 249	-63 538	1 421	< .001
Duration	6 465	6 563	196	< .001
Intensity	-81 147	-81 758	778	< .001
Spectral tilt	20 002	20 169	333	< .001
Vowel quality	-200 409	-200 396	26	< .001
F0 variation	-69 699	-69 023	1 351	< .001

3.2.3 Linear Discriminant Analyses

We now turn to the cue-weighting tendencies. Recall that, on the group level, we observed consistent high use of duration, intensity, and mean F0 in accented words and words in isolation, but low use of mean F0 in unaccented words. In contrast, spectral tilt, vowel quality and F0 variation were weak cues. Next, we examined whether these group-level cue weights, shown in Figure 3 hold for all talkers. For example, we ask whether all talkers consistently use duration and intensity as their strongest cues in unaccented words, or does one take priority in some talkers? Additionally, we examine whether there are similarities between different talkers (i.e., clustering approach) and further examine what kind of cue-weighting relation (i.e., cue-trading or cue-enhancement) exists between cues.

LDA coefficients of individual participants are given in Supplementary Tables S11-S13, split for condition (isolation, accented, and unaccented). Recall that these coefficients represent the cue weights for each cue, with higher coefficients indicating a larger cue weight (i.e., more important in signaling word stress). Supplementary Tables S11-S13 show that mean F0, intensity, and duration had generally the highest LDA coefficients, corroborating the tendency to primarily use these cues to signal word stress (cf. Figure 3). However, more critically, we observed a large amount of between-talker variability (see Supplementary Tables S11-S13): each participant produced word stress with a unique combination of cue weights. In other words, even though the group-level data show some consistency in cue weighting, not all individual talkers follow the same tendency, but show individual preferences. For example, when examining the unaccented condition (Supplementary Table S13), some talkers prioritize duration, while others prioritize intensity.

Next, we assessed whether talkers vary in systematic ways in their cue-weighting tendencies. We classified participants into different groups according to which cue (mean F0, intensity, or duration) had the highest LDA coefficient for each particular talker. Different clusters of cue-weighting tendencies emerged from this classification (see Figure 5). More specifically, in the isolation and the accented condition, we observed a large group of primarily F0 users (in dark blue) and a group of primarily intensity users (in purple). In the unaccented condition, we observed a group of duration users and a group of intensity users.

This clustering approach raises the question how stable these cue-weighting tendencies actually are within a given talker. In other words, are the LDA coefficients a reliable measure of talkers' individual cue-weighting tendencies, or do they result from random variability? To assess this, we ran an additional analysis quantifying the split-half reliability of the LDA coefficients. We divided each participant's data into two subsets, data from even trials vs. data from uneven trials. We then ran the same LDA analyses for each participant, for each condition, and – critically – separately for each subset, resulting in two sets of cue weights for each participant in each of the three speaking conditions. If the LDA coefficients reflect reliable

individual talker cue-weighting tendencies, the two sets of cue weights should be highly correlated. Results showed a high correlation between the two sets of LDA coefficients for the isolation condition (r (238) = .90, p < .001), the accented condition (r (238) = .81, p < .001), and the unaccented condition (r (238) = .85, p < .001). Finally, we checked whether the clustering approach, indicating for each talker which cue was used as the main cue, was comparable in both subsets. Results showed that this was the case for 28 participants (73%) in the isolation condition, 31 (70%) in the accented condition, but only 22 (55%) in the unaccented condition. The lower clustering consistency in the unaccented condition is likely due to overall lower LDA coefficients, particularly in mean F0 and intensity, in this condition (see Fig 5). With overall lower coefficients, the difference between cue weights is reduced, making exactly which cue happens to surface as the strongest more variable. Taking these outcomes together, we find evidence for high split-half reliability of the individual LDA coefficients, although the clustering approach was somewhat less reliable particularly in the unaccented condition.

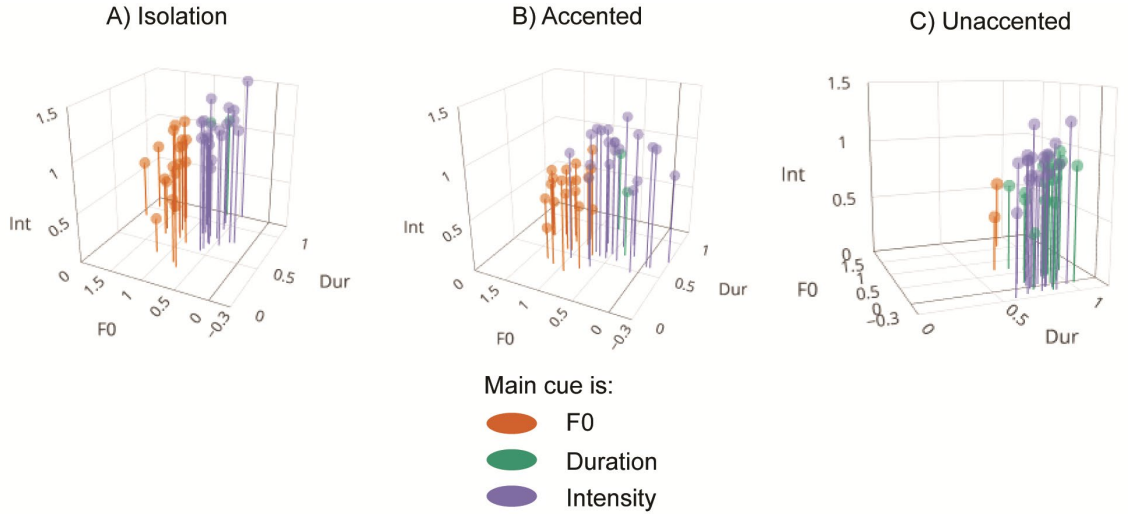

Figure 5.
Scatter plots of Linear Discriminant Analyses (LDA)-coefficients from individual participants. Each data point represents one participant plotted along three acoustic dimensions (mean F0, intensity, and duration), illustrating between-talker variability. The main cue (color coded) is determined as the cue with the highest LDA coefficient of the three dimensions within each participant. The figure suggests that there are groups of primarily F0 or intensity users in the isolation and accented condition, but groups of primarily intensity or duration users in the unaccented condition.

It is important to note that this classification does not imply that talkers use only one cue to produce stressed syllables. To assess the contribution of the other cues, we ran correlation analyses between all the cues (see Figure 6). Here, we will focus on the three strongest cues; mean F0, intensity, and duration. We observed a negative correlation between mean F0 and intensity in the accented (r (38) = -.61, p < .001), unaccented (r (38) = -.33, p < .05), and isolation condition (r (38) = -.64, p < .001). We further found a positive correlation between duration and intensity in the isolation condition (r (38) = .47, p < .01) and a positive trend in the unaccented condition (r (38) = .28, p = .08). These results suggest that there is a clear cue-trading relation between mean F0 and intensity: if a talker primarily used mean F0 to cue stress, that talker typically down-weighted the intensity cue, regardless of what accentual condition the words occurred in. Note that in unaccented words, the group-level data already suggest low LDA coefficients for mean F0, the negative correlation thus could suggest that talkers accommodate for the lack of mean F0 by relying more on intensity. On the other hand, the positive correlation between duration and intensity suggests a cue-enhancement relation between these cues: when talkers rely more on one of these cues, the other cue additionally

receives more weight. Other cue-trading and cue-enhancement relations are further present with the other three cues; spectral tilt, vowel quality, and F0 variation (see Figure 6).

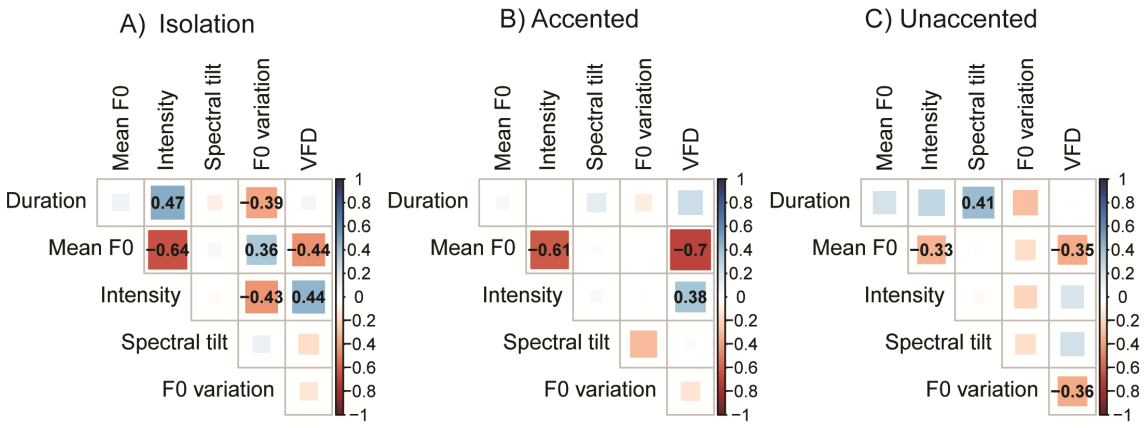

Figure 6.
Correlation matrices of the LDA coefficients for the analyzed cues to word stress. Correlations are depicted separately for the three speaking conditions (isolation, accented, unaccented). Correlations are color coded with blue colors indicating a positive correlation and red indicating a negative correlation. For significant correlations (p <.05), the exact correlation coefficient (r) is given. The figure shows a variety of positive and negative correlations between the different cues, suggesting that different cue-trading and cue-enhancement relations exist between different cues.

4.0 Discussion

The present study examined acoustic correlates of word stress in Dutch. First, we investigated group-level acoustic correlates, providing the most comprehensive description of word stress in Dutch to date. Second, we investigated acoustic differences between individual talkers in word stress production in Dutch. Results showed that − on top of a general tendency to primarily use mean F0, intensity, and duration − individual talkers each reliably used a unique set of cue weights to produce word stress. The latter finding emphasizes that there is large variability between individual talkers. Moreover, classes of cue-weighting tendencies could be identified in the data, with groups of talkers differing in which cue (either mean F0, intensity, or duration) they used as their most important cue.

The first goal of the present study was to describe the acoustic correlates of word stress in Dutch in a larger number of participants, producing a larger number of different words, and measuring a range of cues to stress in Dutch. In this description, we confirm previous findings (Rietveld & Van Heuven, 2009; van Bergem, 1993) showing that, in Dutch, stressed syllables have a longer duration, higher mean intensity, shallower spectral tilt, acoustically less reduced (i.e., fuller) vowels, and, in accented words, higher mean F0 and larger F0 variation. We further found that the effect of word stress was further enhanced in accented words for all cues except for vowel quality (cf. Sluijter & Van Heuven, 1996; van Bergem, 1993). In contrast to Eriksson (2016), we did not find gender differences in how strongly cues were enhanced in stressed syllabled compared to unstressed syllables.

Interestingly, we found that low (open) vowels were more strongly reduced in unstressed syllables compared to high (closed) vowels. While we did not have specific predictions about this, we offer the following explanation. Previously, Lindblom (1963) measured spectral and temporal characteristics of Swedish vowels and concluded that the temporal reduction in unstressed vowels often leads to formant undershoot (i.e., a failure to reach the intended vowel formant values). In the present study, since open vowels are simply

further away from a central vowel location in the vowel space, the formant undershoot more strongly affects low vowels compared to high vowels. In other words, since higher values for F1 have to be reached in open vowels, this is harder to reach in unstressed, temporally reduced vowels.

Further, we found that duration is a stronger cue to word stress when in sentence context compared to words in isolation. We offer two possible interpretations for this finding. First, it could be that, since in isolation syllable durations tend to be longer compared to in sentence context regardless of stress, the difference between stressed and unstressed syllables becomes smaller. Therefore, duration may be a less useful cue to word stress in words in isolation. Second, the reverse is also possible. In sentence context, talkers generally produce more syllables compared to words in isolation. Therefore, talkers provide much more information to the listener about their average syllable duration for stressed and unstressed syllables. In other words, talkers could possibly rely more on duration as a cue to word stress in sentences, making it a more useful cue in sentences.

Regarding cue-weighting tendencies, the group-level results from the LDAs showed that Dutch talkers generally use mean F0, intensity and duration to signal stress in accented words and words in isolation, while only intensity and duration are strong cues in unaccented words. We further found that spectral tilt, vowel quality and F0 variation are generally weak cues in all conditions (isolation, accented, and unaccented). This ordering of mean F0, intensity and duration is not surprising based on previous research (Rietveld & Van Heuven, 2009), but the relatively low LDA coefficients for spectral tilt seem to contradict Sluijter & Van Heuven (1996), who reported spectral tilt as a strong cue to word stress in Dutch.

There are (at least) two important differences between the present study and Sluijter & Van Heuven (1996) that could underly these results. First, Sluijter & Van Heuven (1996) tested mainly the vowel /a:/, while the present study included a more representative sample of vowels. Indeed, Severijnen et al. (2022) recently demonstrated that, when including a more representative sample of vowels and a larger participant sample, the advantage for spectral tilt disappeared – even when using the same spectral tilt metric as S&vH1996. Second, the metric in both Severijnen et al. (2022) and Sluijter & Van Heuven (1996) is confounded with the vowel's characteristics. Specifically, they both measured spectral tilt as the intensity in four contiguous frequency bands (bins B1-B4: 0-0.5, 0.5-1, 1-2, 2-4 kHz). This measure is highly affected by the formant characteristics of a vowel, but could also be affected by a priori decisions on which frequency bands to use (e.g., 0-1 kHz instead of 0-0.5 kHz for B1). In the present study, we avoided these a priori decisions by measuring the slope of a linear regression on the spectrum (cf. Van Heuven, 2018). While this is a first step towards removing the confounds in measuring spectral tilt, other measures that further remove vowel influences, such as H1-A3 (cf. Hanson & Chuang, 1999), or measures obtained through inverse filtering, should provide converging evidence before drawing firm conclusions on the relative contribution of spectral tilt.

The second and main goal of the present study was to examine acoustic differences between talkers, and we found that individual talkers each reliably used a unique set of cue weights to produce word stress. These findings are in line with previous research on segmental phonetic cue-weighting differences between languages (Lisker & Abramson, 1964), dialects (Kang, 2013), and individual talkers (Schertz et al., 2015). These studies showed that the strength of different acoustic cues that signal speech categories largely depend on the native language and dialect one speaks. Moreover, on top of these group-level cue weights, talkers further vary on an individual level, differing in how much they follow the group cue-weighting tendencies. Building on these studies, and on Schertz et al. (2015) in particular, we show for the first time that individual talkers also each use different sets of cue weights to produce word stress. More specifically, the present study illustrated that on top of the group-level tendency to mainly use three cues to signal word stress (mean F0, intensity, and duration), there was a large degree of variability on an individual talker level. Specifically, while some talkers did follow the group-level tendency to use these three cues, others appeared to prioritize one cue (mean F0, intensity, or duration); though note that the other cues still contributed. This illustrates that, with

regard to word stress in Dutch, the group-level tendencies do not generally reflect how individual talkers actually produce word stress, highlighting the importance of taking talker-specific differences into account.

This is also evident in the use of mean F0 in the different speaking conditions. That is, we observed that mean F0 was a stronger cue for accented words (isolation and accented speaking condition) compared to unaccented words, in line with previous studies (Beckman & Edwards, 1994; Pierrehumbert, 1980; Rietveld & Van Heuven, 2009). Specifically, mean F0 being a stronger cue in accented words is indicative of the presence of a pitch accent on the stressed syllable. However, not all talkers used mean F0 as the strongest cue in accented words; some used intensity more strongly, as indicated by our clustering approach as well as the negative within-talker correlation between mean F0 and intensity. This does not imply that for those talkers, mean F0 did not contribute to stress production in accented words. Instead, they used both cues but weighted intensity more heavily, again illustrating the relevance of talker-specific differences.

We further found that talkers clustered into different cue-weighting tendencies, generally prioritizing either mean F0 or intensity (in isolation and accented words), or duration and intensity (unaccented words). This suggests that, while there is a large amount of variability (i.e., each talker had a unique set of cue weights), talkers do seem to cluster together regarding which cue is their main cue. Note that this does not imply that every talker within a cluster is equally similar. While they do share the main cue, there was still considerable variability between talkers within a cluster as to how much the other cues contribute, as also evidenced by the different cue-trading and cue-enhancement relations between cues. Still, the shared main cue possibly helps listeners in dealing with between-talker variability in speech perception, by allowing listeners to generalize between talkers with similar production strategies (cf. Kleinschmidt & Jaeger, 2015). Two interesting questions for future research emerge from these results. First, where does the observed between-talker variability come from? More specifically, is the observed variability completely random, or is it in some way affected by the previous language experience of the talkers? Second, how is cue weighting of word stress in production and perception linked? Specifically, do talkers prioritize the same cues to word stress in production and perception?

Concerning the former question, the present study took a first step into describing how exactly talkers vary in word stress production in Dutch. The next step would be to directly examine the possible sources of this variability. As already mentioned, acoustic differences have been observed between gender (Adank et al., 2004; Haan & Van Heuven, 1999), regional dialects (Adank et al., 2007; Clopper & Smiljanic, 2011), and individual production strategies of talkers (Schertz et al., 2015; Xie et al., 2021). The observed variability in the present study could be driven by any (or a combination) of these sources. For example, there could be differences between dialects, but on top of that, talkers who speak the same dialect could further vary due to individual production strategies. The present study could not properly address this, since we tried to minimize variability in linguistic backgrounds. Moreover, any differences in participants' background that were still present (e.g., monolingual vs. bilingual participants, the region in which participants grew up), were not strictly controlled for. Examining which of these, and possibly other, sources contribute to the observed variability would require more systematic recruitment such as in Adank et al. (2007). Specifically, systematically recruiting participants from different regions, while balancing gender and minimizing further linguistic variability (e.g., bilinguals, or other regional dialects), would allow experimenters to isolate each possible source and measure its contribution.

The latter question relates to a larger body of research that has examined the link between production and perception. More specifically, some researchers regard the perception and production system to be closely linked (e.g., Liberman & Mattingly, 1985; Pickering & Garrod, 2013), which has further been confirmed by several studies. For example, Harrington (2008) and Kleber et al. (2012) found that talkers who produce speech with less coarticulation also compensated less for coarticulation in perception. Also, Newman (2003) found correlations between perceptual prototypes of speech sounds (e.g., VOT for stop consonants) and average productions of those sounds. Moreover, Pinget et al. (2020) illustrated that participants who

start to participate in a sound change (specifically, devoicing of labiodental fricatives and bilabial stops in Dutch), change their perceptual patterns *before* changing their productions. In other words, not only do perception and production appear to be linked, but changes in perception *precede* changes in production. Thus, if complex cue-weighting patterns, with various types of interactions (e.g., cue-trading, cue-enhancement), are present in production, these might also be present in perception. Evidence in favor of such interactions between cue weights in perception comes from research using Active Learning systems (Einfeldt et al., 2024), illustrating that models with various cue-weighting interactions outperformed models without these interactions.

In the current study, this would then predict that the individual cue-weighting tendencies in production of word stress may be preceded by similar perceptual cue-weighting tendencies. Future research could examine this prediction by incorporating a perceptual task and compare individual cue-weighting tendencies in production and perception. However, note that these perceptual cue weights are also subject to change based on the short-term speech regularities a listener is exposed to (Severijnen, et al., 2021; 2023).

Another interesting question concerns how the present study can be explained by existing models of speech production. More specifically, how do these cue-weighting tendencies emerge in speakers? Are they a result of intrinsic preferences within a talker, or does every talker start with more general tendencies which are then shaped for each talker by particular linguistic experiences? Once the cue-weighting tendencies have been established, how can models of speech production incorporate them? In other words, at which representational level do these tendencies influence production, and how are the tendencies represented/stored? The present study was not designed to test models of speech production, but it does suggest that models, in particular those of speech motor control such as the DIVA model (Tourville & Guenther, 2011), may be extended to incorporate mechanisms capable of generating between-talker variability.

On top of the observed variability in cue weighting, we also observed between-talker variability *within* cues. More specifically, visualizations of the cue distributions illustrated that individual talkers differed from each other with regard to the cue means and the shape of the distributions. These within-cue differences were further confirmed by the LMER model comparisons between a model without and with by-participant random slopes for word stress. These results illustrated that there was a significant increase in model fit when random slopes were added for each cue, confirming the presence of between-talker variability within each cue. These findings are in line with previous literature that also observed within-cue variability between talkers for other speech contrasts (Adank et al., 2004, 2007; Allen et al., 2003; Hillenbrand et al., 1995; Theodore et al., 2009; Xie et al., 2021). While this type of variability is not concerned with cue-weighting tendencies, it still adds variability to the acoustic signal. Specifically, this means that the same absolute cue values can signal a stressed syllable for one talker, while it can signal an unstressed syllable for another talker. Possible sources of this within-cue variability could be biological differences (e.g., differences in overall mean F0 due to gender), differences in production strategies (e.g., differences in syllable durations due to speech rate variation), or just random variability in speech production. Regardless of the exact source, the within-cue differences add more variability to the differences in cue-weighting tendencies. Together, these two sources of variability illustrate the large amount of variability that listeners have to take into account when perceiving word stress in Dutch (cf. Bosker, 2022; Severijnen et al., 2021; 2023).

5.0 Conclusions

The present study illustrated how individual talkers vary in their productions of word stress in Dutch. We found that, on top of the group tendency to use mainly mean F0, intensity, and duration, there is a large amount of between-talker variability that lies underneath these group tendencies, both within and between cues. Moreover, classes of cue-weighting tendencies emerged, which could inform and support listeners in their perception of new talkers. These results exemplify the large-scale acoustic variability in speech and underline the immense challenge listeners face in perceiving various and novel talkers.

Acknowledgements

This research was funded by a DCC Internal Round grant, awarded to J.M. and H.R.B. Funding was also received from an ERC Starting Grant (HearingHands, 101040276) from the European Union, awarded to H.R.B. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them. Part of this work has been reported in Severijnen, G.G.A., Bosker, H.R., & McQueen, J.M. (2022) Acoustic correlates of Dutch lexical stress re-examined: Spectral tilt is not always more reliable than intensity. *Proceedings of the International Conference on Speech Prosody 2022,* Lisbon. We would further like to thank Sanne van Eck, Dennis Joosen, Esther de Kerf, Inge Pasman, Carlijn van Herpt, and Abdellah Elouatiq, who helped to annotate the data. Finally, we would like to thank Gerard van Oijen for his help with setting up the sound recording lab.

Reference List

Adank, P., van Hout, R., & Smits, R. (2004). An acoustic description of the vowels of Northern and Southern Standard Dutch. *The Journal of the Acoustical Society of America*, *116*(3), 1729–1738. https://doi.org/10.1121/1.1779271

Adank, P., van Hout, R., & Van de Velde, H. (2007). An acoustic description of the vowels of northern and southern standard Dutch II: Regional varieties. *The Journal of the Acoustical Society of America*, *121*(2), 1130–1141. https://doi.org/10.1121/1.2409492

Allen, J. S., Miller, J. L., & DeSteno, D. (2003). Individual talker differences in voice-onset-time. *The Journal of the Acoustical Society of America*, *113*(1), 544–552. https://doi.org/10.1121/1.1528172

Arvaniti, A., & Garding, G. (2007). Dialectical variation in the rising accents of American English. In J. Cole & J. H. Hualde, *Papers in laboratory phonology* (Vol. 9, pp. 547–576). Mouton de Gruyter.

Beckman, M. E., & Edwards, J. (1994). Articulatory Evidence for Differentiating Stress Categories. In P. A. Keating (Ed.), *Phonological Structure and Phonetic Form: Papers in Laboratory Phonology III* (pp. 7–33). Cambridge University Press.

Boersma, P., & Weenink, D. (2019). *Praat: Doing phonetics by computer* (6.065) [Computer software]. www.praat.org

Bosker, H. R. (2022). Evidence For Selective Adaptation and Recalibration in the Perception of Lexical Stress. *Language and Speech*, *65*(2), 472–490. https://doi.org/10.1177/00238309211030307

Braun, B., Lemhöfer, K., & Cutler, A. (2008). English word stress as produced by English and Dutch speakers: The role of segmental and suprasegmental differences. *Proceedings of Interspeech 2011*, 1.

Cangemi, F., Krüger, M., & Grice, M. (2015). Listener-specific perception of speaker-specific production in intonation. *Individual Differences in Speech Production and Perception*, 123–145.

Cho, T., & McQueen, J. M. (2005). Prosodic influences on consonant production in Dutch: Effects of prosodic boundaries, phrasal accent and lexical stress. *Journal of Phonetics*, *33*(2), 121–157. https://doi.org/10.1016/j.wocn.2005.01.001

Clayards, M. (2018). Individual Talker and Token Covariation in the Production of Multiple Cues to Stop Voicing. *Phonetica*, *75*(1), 1–23. https://doi.org/10.1159/000448809

Clopper, C. G., & Smiljanic, R. (2011). Effects of gender and regional dialect on prosodic patterns in American English. *Journal of Phonetics*, *39*(2), 237–245. https://doi.org/10.1016/j.wocn.2011.02.006

Cole, J. (2015). Prosody in context: A review. *Language, Cognition and Neuroscience*, *30*(1–2), 1–31. https://doi.org/10.1080/23273798.2014.963130

Cooper, N., Cutler, A., & Wales, R. (2002). Constraints of Lexical Stress on Lexical Access in English: Evidence from Native and Non-native Listeners. *Language and Speech*, *45*(3), 207–228. https://doi.org/10.1177/00238309020450030101

Cutler, A. (1986). Forbear is a Homophone: Lexical Prosody Does Not Constrain Lexical Access. *Language and Speech*, *29*(3), 201–220. https://doi.org/10.1177/002383098602900302

Cutler, A., & Pasveer, D. (2006). Explaining cross-linguistic differences in effects of lexical stress on spoken-word recognition. *3rd International Conference on Speech Prosody*.

Cutler, A., & Van Donselaar, W. (2001). Voornaam is not (really) a Homophone: Lexical Prosody and Lexical Access in Dutch. *Language and Speech*, *44*(2), 171–195. https://doi.org/10.1177/00238309010440020301

Cutler, A., Wales, R., Cooper, N., & Janssen, J. (2007). Dutch listeners' use of suprasegmental cues to English stress. *16th International Congress of Phonetic Sciences (ICPhS 2007)*, 1913–1916.

Draxler, C., & Jänsch, K. (2004). *SpeechRecorder – a Universal Platform Independent Multi-Channel Audio Recording Software* [Computer software].

Einfeldt, M., Sevastjanova, R., Zahner-Ritter, K., Kazak, E., & Braun, B. (2024). The use of Active Learning systems for stimulus selection and response modelling in perception experiments. *Computer Speech & Language*, *83*, 101537. https://doi.org/10.1016/j.csl.2023.101537

Eriksson, A., Bertinetto, P. M., Heldner, M., Nodari, R., & Lenoci, G. (2016). *The Acoustics of Lexical Stress in Italian as a Function of Stress Level and Speaking Style*. 1059–1063. https://doi.org/10.21437/Interspeech.2016-348

Eriksson, A., & Heldner, M. (2015). The Acoustics of Word Stress in English as a Function of Stress Level and Speaking Style. *Proc. Interspeech 2015*, 41–45. https://doi.org/10.21437/Interspeech.2015-9

Escudero, P., Benders, T., & Lipski, S. C. (2009). Native, non-native and L2 perceptual cue weighting for Dutch vowels: The case of Dutch, German, and Spanish listeners. *Journal of Phonetics*, *37*(4), 452–465. https://doi.org/10.1016/j.wocn.2009.07.006

Gordon, M., & Roettger, T. (2017). Acoustic correlates of word stress: A cross-linguistic survey. *Linguistics Vanguard*, *3*(1), 20170007. https://doi.org/10.1515/lingvan-2017-0007

Haan, J., & Van Heuven, V. (1999). Male vs. Female pitch range in Dutch questions. *Proceedings of the 13th International Congress of Phonetic Sciences*, 1581–1584.

Hanson, H. M., & Chuang, E. S. (1999). Glottal characteristics of male speakers: Acoustic correlates and comparison with female data. *The Journal of the Acoustical Society of America*, *106*(2), 1064–1077. https://doi.org/10.1121/1.427116

Harrington, J., Kleber, F., & Reubold, U. (2008). Compensation for coarticulation, /u/-fronting, and sound change in standard southern British: An acoustic and perceptual study. *The Journal of the Acoustical Society of America*, *123*(5), 2825–2835. https://doi.org/10.1121/1.2897042

Hayward, K. (2000). *Experimental Phonetics: An introduction* (1st ed.). Routledge.

Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *The Journal of the Acoustical Society of America*, *97*(5), 3099–3111.

Kang, K.-H. (2013). F0 Perturbation as a Perceptual Cue to Stop Distinction in Busan and Seoul Dialects of Korean. *Phonetics and Speech Sciences*, *5*(4), 137–143. https://doi.org/10.13064/KSSS.2013.5.4.137

Karlsson, F., & van Doorn, J. (2012). Vowel formant dispersion as a measure of articulation proficiency. *The Journal of the Acoustical Society of America*, *132*(4), 2633–2641. https://doi.org/10.1121/1.4746025

Kisler, T., Reichel, U., & Schiel, F. (2017). Multilingual processing of speech via web services. *Computer Speech & Language*, *45*, 326–347. https://doi.org/10.1016/j.csl.2017.01.005

Kleber, F., Harrington, J., & Reubold, U. (2012). The Relationship between the Perception and Production of Coarticulation during a Sound Change in Progress. *Language and Speech*, *55*(3), 383–405. https://doi.org/10.1177/0023830911422194

Kleinschmidt, D. F., & Jaeger, T. F. (2015). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review*, *122*(2), 148. https://doi.org/10.1037/a0038695

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). **lmerTest** Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, *82*(13). https://doi.org/10.18637/jss.v082.i13

Length, R. (2022). *emmeans: Estimated Marginal Means, aka Least-Square.* (R package version 1.8.2) [Computer software]. http://cran.r-project.org/package=emmeans

Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition*, *21*, 1–36. https://doi.org/10.1016/0010-0277(85)90021-6

Lindblom, B. (1963). Spectrographic Study of Vowel Reduction. *The Journal of the Acoustical Society of America*, *35*(11), 1773–1781. https://doi.org/10.1121/1.1918816

Lisker, L. (1986). "Voicing" in English: A Catalogue of Acoustic Features Signaling /b/ Versus /p/ in Trochees. *Language and Speech*, *29*(1), 3–11. https://doi.org/10.1177/002383098602900102

Lisker, L., & Abramson, A. S. (1964). A Cross-Language Study of Voicing in Initial Stops: Acoustical Measurements. *WORD*, *20*(3), 384–422. https://doi.org/10.1080/00437956.1964.11659830

Newman, R. S. (2003). Using links between speech perception and speech production to evaluate different acoustic metrics: A preliminary report. *The Journal of the Acoustical Society of America*, *113*(5), 2850–2860. https://doi.org/10.1121/1.1567280

Newman, R. S., Clouse, S. A., & Burnham, J. L. (2001). The perceptual consequences of within-talker variability in fricative production. *The Journal of the Acoustical Society of America*, *109*(3), 1181–1196. https://doi.org/10.1121/1.1348009

Niebuhr, O., D'Imperio, M., Fivela, B. G., & Cangemi, F. (2011). Are there "shapers" and "aligners"? Individual differences in signaling pitch accent category. *Proceedings of the 17th International Congress of Phonetic Sciences*, 120–123.

Nooteboom, S. G. (1972). *Production and perception of vowel duration: A study of durational properties of vowels in Dutch* [Ph.D. dissertation]. Universiteit Utrecht.

Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral and Brain Sciences*, *36*(4), 329–347. https://doi.org/10.1017/S0140525X12001495

Pierrehumbert, J. (1980). *The Phonology and Phonetics of English Intonation* [Ph.D. dissertation]. MIT.

Pinget, A.-F. (2015). *The actuation of sound change* [Ph.D. dissertation]. Universiteit Utrecht.

Pinget, A.-F., Kager, R., & Van de Velde, H. (2020). Linking Variation in Perception and Production in Sound Change: Evidence from Dutch Obstruent Devoicing. *Language and Speech*, *63*(3), 660–685. https://doi.org/10.1177/0023830919880206

Plag, I., Kunter, G., & Schramm, M. (2011). Acoustic correlates of primary and secondary stress in North American English. *Journal of Phonetics*, *39*(3), 362–374. https://doi.org/10.1016/j.wocn.2011.03.004

R Core Team. (2020). *R: A Language and Environment for Statistical Computing* [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Reinisch, E., Jesse, A., & McQueen, J. M. (2010). Early use of phonetic information in spoken word recognition: Lexical stress drives eye movements immediately. *Quarterly Journal of Experimental Psychology*, *63*(4), 772–783. https://doi.org/10.1080/17470210903104412

Reinisch, E., & Weber, A. (2012). Adapting to suprasegmental lexical stress errors in foreign-accented speech. *The Journal of the Acoustical Society of America*, *132*(2), 1165–1176. https://doi.org/10.1121/1.4730884

Rietveld, A. C. M., & Van Heuven, V. J. (2009). *Algemene fonetiek* (3rd ed.). Coutinho.

Roettger, T., & Gordon, M. (2017). Methodological issues in the study of word stress correlates. *Linguistics Vanguard*, *3*(1), 20170006. https://doi.org/10.1515/lingvan-2017-0006
Schertz, J., Cho, T., Lotto, A., & Warner, N. (2015). Individual differences in phonetic cue use in production and perception of a non-native sound contrast. *Journal of Phonetics*, *52*, 183–204. https://doi.org/10.1016/j.wocn.2015.07.003
Schertz, J., & Clare, E. J. (2020). Phonetic cue weighting in perception and production. *WIREs Cognitive Science*, *11*(2). https://doi.org/10.1002/wcs.1521
Severijnen, G., Bosker, H. R., & McQueen, J. (2022). Acoustic correlates of Dutch lexical stress re-examined: Spectral tilt is not always more reliable than intensity. *Speech Prosody 2022*, 278–282. https://doi.org/10.21437/SpeechProsody.2022-57
Severijnen, G. G. A., Bosker, H. R., & McQueen, J. M. (2023). *Corpus of Dutch Lexical Stress (CooLeSt)* (Version v1) [Donders Repository]. https://doi.org/10.34973/vkkk-yg79
Severijnen, G. G. A., Bosker, H. R., Piai, V., & McQueen, J. M. (2021). Listeners Track Talker-Specific Prosody to Deal With Talker-Variability. *Brain Research*.
Severijnen, G. G. A., Di Dona, G., Bosker, H. R., & McQueen, J. M. (2023). Tracking talker-specific cues to lexical stress: Evidence from perceptual learning. *Journal of Experimental Psychology: Human Perception and Performance*, *49*(4), 549–565. https://doi.org/10.1037/xhp0001105
Slis, I. H. (1971). Articulatory Effort and its Durational and Electromyographic Correlates. *Phonetica*, *23*(3), 171–188. https://doi.org/10.1159/000259338
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*(4), 2471–2485. https://doi.org/10.1121/1.417955
Sulpizio, S., & McQueen, J. M. (2012). Italians use abstract knowledge about lexical stress during spoken-word recognition. *Journal of Memory and Language*, *66*(1), 177–193. https://doi.org/10.1016/j.jml.2011.08.001
Theodore, R. M., Miller, J. L., & DeSteno, D. (2009). Individual talker differences in voice-onset-time: Contextual influences. *The Journal of the Acoustical Society of America*, *125*(6), 3974–3982. https://doi.org/10.1121/1.3106131
Tourville, J. A., & Guenther, F. H. (2011). The DIVA model: A neural theory of speech acquisition and production. *Language and Cognitive Processes*, *26*(7), 952–981. https://doi.org/10.1080/01690960903498424
Tseng, C., Su, C., & Visceglia, T. (2013). Levels of lexical stress contrast in english and their realization by L1 and L2 speakers. *2013 International Conference Oriental COCOSDA Held Jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE)*, 1–5. https://doi.org/10.1109/ICSDA.2013.6709853
van Alphen, P. M., & Smits, R. (2004). Acoustical and perceptual analysis of the voicing distinction in Dutch initial plosives: The role of prevoicing. *Journal of Phonetics*, *32*(4), 455–491. https://doi.org/10.1016/j.wocn.2004.05.001
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 Heuven, V. J. (2018). Acoustic Correlates and Perceptual Cues of Word and Sentence Stress. In R. Goedemans, J. Heinz, & H. van der Hulst (Eds.), *The Study of Word Stress and Accent: Theories, Methods and Data* (1st ed., pp. 15–59). Cambridge University Press. https://doi.org/10.1017/9781316683101
Venables, W. N., & Ripley, B. D. (2002). *Modern Applied Statistics with S* (Fourth). Springer.
Xie, X., Buxó-Lugo, A., & Kurumada, C. (2021). Encoding and decoding of meaning through structured variability in intonational speech prosody. *Cognition*, *211*, 104619. https://doi.org/10.1016/j.cognition.2021.104619