Acoustic correlates of stress in speech perception

Petroula Mousikou a,1,2, Patrycja Strycharczuk b,1, Kathleen Rastle a,*

a Department of Psychology, Royal Holloway, University of London, United Kingdom
b Linguistics and English Language, The University of Manchester, Oxford Rd, Manchester M13 9PL, United Kingdom

A R T I C L E   I N F O

Keywords:
Word stress
Speech perception
Acoustics

A B S T R A C T

Stress is an important property of English spoken words. Research conducted over the past 70 years has sought to determine how acoustic cues, including duration, pitch, and intensity influence stress perception; however, the evidence remains conflicting. In the present study, we used a large dataset of 10 speakers’ productions of disyllabic nonwords to investigate how listeners make use of these cues to perceive stress. Over 100 listeners made stress judgements on nearly one thousand items each, yielding a total of nearly 75,000 analysable responses. Results of average performance showed that stress judgments were influenced by all three cues, both individually and in combination. However, the relative importance of any one cue depended on the value of the other cues, particularly in the frequent situations in which cues offered conflicting stress information. Results of individual performance showed that listeners often use the same acoustic information regarding stress in different ways, but that speakers also sometimes offer different information about stress. Our mega-study approach to investigating word-stress perception eclipses previous studies in terms of its power, and offers new insights into our understanding of how listeners perceive stress.

Introduction

Word stress refers to the emphasis placed on a certain syllable in a word, making that syllable more prominent than the others. In languages characterised by word stress, stress can be either fixed or variable. Hungarian is an example of a language with fixed stress, as words are always stressed on the initial syllable. English is an example of a language with variable stress, as words might belong to the same grammatical category and even have a similar syllabic structure, yet are stressed on different syllables (e.g., CAmel vs. caNAL).

Understanding how listeners perceive word stress has emerged as an important area of research because of the possible role of stress in speech segmentation (e.g., Cutler & Norris, 1988) and spoken word recognition (e.g., Jesse, Poellmann, & Kong, 2017). According to an influential theory of speech segmentation, locating word boundaries in continuous speech is achieved by assuming that every stressed syllable is a word onset (Cutler & Norris, 1988; see Mattys, White, & Melborn, 2005 for discussion). This strategy is fairly successful in English, given that approximately 80% of English words have a strong (i.e., stressed) initial syllable (Cutler & Carter, 1987). Empirical research in this domain also shows that English adult listeners show a processing preference for words with initial stress (Mattys & Samuel, 2000), and that this preference is already evident in infants between 6 and 9 months of age (Jusczyk, Cutler & Redanz, 1993; see also Jusczyk, Houston, & Newsome, 1999, for evidence that English-learning infants use this information to segment speech in the first year of life). Beyond the challenge of finding words in continuous speech, research also suggests that English adult listeners use stress to disambiguate words in online recognition before this is possible using segmental information (e.g., admiral-admiration; Jesse et al., 2017; see Reinisch, Jesse, & McQueen, 2010, for similar evidence in Dutch). These findings all imply that listeners use sublexical (signal-derived) cues to perceive stress. Thus, a critical question has been to determine what these are.

One key distinction in this literature is between the concepts of lexical stress and metrical stress (e.g., Cutler, 1986; Cutler & van Donselaar, 2001; Slowiaczek, Soltano, & Bernstein, 2006). Lexical stress refers to the degree of emphasis in the syllables of a word that is spoken in isolation, with the stressed syllable being the one with greatest emphasis. Languages with lexical stress, such as English, allow segmentally identical words to be distinguished on the basis of the
relative emphasis of their syllables (e.g., TRUSTy vs. trusTEE). Research has shown that the main acoustic cues to lexical stress are duration, pitch, and intensity (see e.g., Cutler, 2005 for a review). Stressed syllables are produced with greater physiological effort than unstressed syllables on the part of the speaker at all stages of the speech production process (i.e., subglottal, glottal, supraglottal). Bigger articulatory movements require more time to be exerted, which results in stressed syllables being longer in duration than unstressed syllables. More effort exerted by the glottis influences the frequency with which the vocal folds vibrate, causing a rise or fall of pitch (note that pitch is the perceptual consequence of fundamental frequency, typically referred to as f0). Likewise, increased effort at the subglottal level pushes more air through the glottis, causing an increase in intensity (for more details about each of these processes, see van Heuven, 2018). In contrast, the concept of metrical stress refers to the vowel quality of a syllable, with strong syllables containing full vowels and weak syllables containing vowels that are reduced, often to schwa (e.g., Cutler, 1986; Cutler & Jesper, 2021).

Our focus in this article is on the perception of lexical stress: specifically, on how listeners use information about duration, pitch, and intensity to judge which syllable has the greater emphasis. Researchers for over 70 years have asked not only whether each of these cues influences the perception of stress, but also whether some cues might be more important than others. However, the studies in this domain are limited in scope and have often led to different conclusions. Seminal work by Fry (1955) showed that both duration and intensity are important predictors of stress judgments, with duration being a more effective predictor than intensity. Subsequent work by Fry (1958; 1965) confirmed the importance of duration, but suggested that pitch may ultimately play a more important role. This conclusion is consistent with findings reported by Morton and Jassem (1965), suggesting that variations in pitch have a greater impact on stress judgments than variations in intensity and duration (see also Beckman, 1986). Turk and Sawusch (1996) confirmed the importance of duration in stress perception and argued that it is a more important form of information than intensity. In contrast, Lieberman (1960) reported that both pitch and intensity are important predictors of stress judgments, and that both of these are more important than duration. Likewise, Chrabaszcz et al. (2014) reported that of these three cues, pitch has the greatest influence on English stress judgments, with intensity and duration having a weaker influence. According to Mattys (2000), however, both pitch and duration (but not intensity) influence stress judgments. Taken together, there is evidence that duration, pitch, and intensity may influence stress judgments, but there is variation in the pattern of results across studies, and there is disagreement on the relative influence of each cue.

The studies conducted in this domain frequently adopt a very similar approach to one another. One stimulus or a small number of stimuli are selected and then synthesized to yield tokens that vary along a particular acoustic dimension. The analyses then probe which dimension has the more substantial influence on stress judgments. The seminal studies of Fry (1955, 1958) used synthesized variations of the stimuli object, subject, digest, contract, and permit. Likewise, several studies in this domain have used synthesized variations of a single stimulus: for example, muma (Turk & Sawusch, 1996), mubu (Chrabaszcz et al., 2014), desert (Zhang & Francis, 2010), object (Lee, 2022). This approach offers very tight experimental control over the acoustic property under investigation; however, the restriction in the number of stimuli being used in these studies severely limits the extent to which conclusions can be generalized to different vowels and phonetic contexts (a fact not often recognized in this literature). Duration, pitch, and intensity are likely to be highly sensitive to context; thus, using a limited set of items leaves those studies vulnerable to item-level and non-generalizable effects.

In addition to the small number of items that these studies have been based on, the way that acoustic properties are systematically varied in these studies does not allow stimuli to reflect conflicts between acoustic cues that may occur in natural speech. Simply knowing that a particular acoustic property has perceptual consequences in stimuli that have been varied along a specific dimension does not allow one to infer how that form of information might be used in a situation in which there are multiple acoustic cues to stress that may be inconsistent with one another. We have known for decades that these types of cue conflicts do arise: for example, Lieberman (1960) presented data based on productions of 24 noun–verb pairs (e.g., digest, contract) suggesting that pitch and intensity cues trade against one another. That is, if fundamental frequency is higher in the unstressed syllable than in the stressed syllable, then this is compensated for in the amplitude difference between the syllables. Indeed, some of the earliest work in this domain recognized that the experimental approach that has come to dominate investigations of the perceptual cues to stress is “a drastic simplification of the conditions in which the [stress] judgment is made” and that “all judgments of stress in natural speech depend on the complicated interaction of a number of cues” (Fry, 1958, p. 151). Yet, to our knowledge, the way that listeners respond to these complicated interactions in natural speech has not been investigated.

This body of work has also been severely limited by low power. We have already highlighted the very small number of items used in studies in this domain. Though the early studies by Fry (1955, 1958, 1965) typically included 100 or more listeners, other studies have paired small numbers of items with small numbers of participants: for example, two listeners (Lieberman, 1960), 15 listeners (Chrabaszcz et al., 2014), 24 listeners (Zhang & Francis, 2010). Recent influential work has shown that substantial numbers of observations are needed to investigate interactions between variables (Brystaert, 2019). Therefore, in all of these studies, it would have been impossible to examine potential interactions due to the very small numbers of participants and items used.

Finally, the earliest research in this domain (Fry, 1955; Lieberman, 1960) provided hints that there may be individual differences both in the way that speakers express stress and in the way that listeners use this information. In particular, Fry (1955, p.765) noted the “considerable variation in the behavior of the speakers with respect to the placing of the accent in different words”, while Lieberman (1960) noted instances in which listeners disagreed in their stress judgments. Both authors, however, simply excluded items that presented any variation; in other words, such variation was left unstudied in previous investigations. To our knowledge, no study has considered individual differences in how listeners use duration, pitch, and intensity in the perception of English stress.

The present study addresses the above-mentioned limitations of previous research by making use of the mega-study approach (e.g., Balota, Cortese, Sergent-Marshall, Spieler, & Yap, 2004; Keuleers, Lacey, Rastle & Brystaert, 2012), in which the influence of particular variables and their interactions is studied using very large datasets comprising hundreds or thousands of items and/or participants. Specifically, we make use of a mega-study of disyllabic nonword reading aloud conducted by Mousikou, Sadat, Lucas, and Rastle (2017). In this study, 41 native English speakers were asked to read aloud 915 disyllabic nonwords (e.g., ejome, vengle) that varied on a number of characteristics. Our approach in this new study was to ask over 100 listeners to make stress judgments on the reading aloud responses from Mousikou et al. (2017) and to investigate how those stress judgments are influenced by the duration, pitch, and intensity of the original productions. Use of this dataset provides a unique opportunity for studying the acoustic cues that underpin stress perception, because the nonwords were read aloud naturally, and thus may contain varying acoustic cues to stress. The nonwords also contained a range of vowels in a variety of phonetic contexts, thus increasing the generalizability of our findings, and minimizing expectancy effects based on lexical knowledge. Moreover, the very large number of datapoints allows us to test for interactions between acoustic cues, something which has not been possible in previous studies, given the limited number of items and participants used. Finally, the power in our design further allows us to explore whether there are individual differences in how listeners use acoustic information in...
judging stress, and to begin to probe why those differences arise.

**Data availability**

Materials, data and analysis code are available on https://osf.io/hngfb/.

**Method**

**Participants**

A total of 175 undergraduate students from Royal Holloway, University of London, signed up for the study. Of these, 111 fulfilled the screening criteria for participating in the main study (see details in the Apparatus and Procedure section). Due to a technical error, the data from one participant were lost, while seven participants provided no response to 10% or more of the trials. The data from these eight participants were discarded, yielding a total of 103 participants to be included in the analyses. The rationale for this sample size is provided in the Design section. Participants were monolingual native speakers of Southern British English and reported no hearing or language difficulties. They provided written consent prior to participating in the main study and received £15 upon its completion.

**Materials**

Stimuli were recordings of reading aloud responses from Mousikou et al.’s (2017) mega-study of disyllabic nonword reading. Specifically, we selected 10 speakers from that study (out of the original 41 participants), all of whom were monolingual native speakers of Southern British English, and all of whom had read aloud 915 disyllabic nonwords (yielding a total of 9150 reading aloud responses). Participants in Mousikou et al. (2017) were instructed to read each nonword as naturally as possible, as if it were a real word, at their own pace and without hesitation. Participants were not given any special instructions as to how to stress each nonword.

The 10 speakers selected from Mousikou et al. (2017) consisted of five males and five females. Speaker selection was based primarily on overall clarity of speech and the quality of the recordings. Items were removed if they were pronounced in an unexpected way, if they were unclear due to noise interference, or if they were produced with hesitations, inter-syllabic pauses, or following self-correction. This cleaning process led to the exclusion of four items from all speakers (alyth, diparge, jemsim, sputsam), and a number of additional items excluded for individual speakers. The number of spoken productions remaining for each speaker were as follows: Speaker 5 (888), Speaker 7 (846), Speaker 11 (883), Speaker 16 (849), Speaker 17 (894), Speaker 20 (893), Speaker 22 (881), Speaker 32 (883), Speaker 36 (892), Speaker 40 (892). For consistency, we refer to the original participant numbers from Mousikou et al. (2017).

Recordings of twenty additional disyllabic nonwords from a study by Ktori, Mousikou, and Rastle (2018) were used in the screening tests and practice session (see Apparatus and Procedure section for details about how these were administered). The additional nonwords were read aloud by a female monolingual native speaker of Southern British English who was naive to the purposes of the study. This speaker read aloud the nonwords with stress clearly placed on the first syllable, and then with stress clearly placed on the second syllable; for example, PRAket and prakET (the full list of items is included in Appendix A and the corresponding recordings can be found on https://osf.io/hngfb/). Recordings took place in a sound-attenuated cabin at the Department of Psychology at Royal Holloway, University of London.

**Design**

Ten participants (i.e., Listeners) were assigned to each Speaker. Due to an oversight, one speaker was assigned 11 listeners and another speaker was assigned 12 listeners, resulting in the final sample of 103 participants. We opted for using this particular design, namely, having each listener exposed to a single speaker (rather than to a combination of different speakers), because the pitch of primary-stressed syllables produced by a certain speaker could be equivalent to the pitch of secondary-stressed syllables produced by another speaker, in which case, the task could have become misleading and particularly difficult for the listeners. The total number of items differed slightly for each speaker due to the exclusions detailed in the Materials section. However, for all speakers, there were a total of six blocks of trials, with each block containing approximately 150 trials. A break was administered between the blocks, yielding five breaks in total. Participants were encouraged to pause for as long as they wanted during the break. Each block lasted approximately 10 min, so it took participants roughly an hour to complete the main task.

**Apparatus and procedure**

Before the study began, participants were given the following spoken instructions: “This study is about how we perceive word stress, the relative emphasis that we put on certain syllables. So for example, if I say ‘reMOVE’ I put the emphasis on ‘move’, so on the second syllable. If I say ‘REflex’ I put the emphasis on ‘re’, so on the first syllable. With real words, this is a relatively easy task to do, because you can think about the words and figure out where the stress is. However, it becomes really difficult if I ask you to do the same thing with made-up words. In this study, you will be listening to made-up words that you have never heard before. All of these made-up words will be stressed either on the first or on the second syllable. Your task is to decide whether the stress is on the first or the second syllable. If you think it is the first, you have to press 1; if you think it is the second you have to press 2.” Subsequently, participants were informed about the screening criteria. In particular, they were told that before doing the main study, they had to do a screening test. Only if they achieved over 85% accuracy on this test (which implied making correct stress judgements for 17 out of 20 items) would they be able to proceed to the main study. If they scored less than 85% correct, they would have to undergo a practice session, and subsequently, they would be asked to do a second screening test. Again, only if they achieved 85% accuracy on this test (which also implied making correct stress judgements for 17 out of 20 items) would they be able to proceed to the main study. In the event they failed the screening test twice, they would not be able to participate in the main study.

Participants were tested individually, seated approximately 60 cm in front of a CRT monitor. They wore headphones, while stimulus presentation was controlled by DMDX software (Forster & Forster, 2003). Each trial started with the visual presentation of a + sign, which remained on the screen for half a second, and was followed by the aural presentation of the target. Participants had five seconds to decide whether the target item was stressed on the first or the second syllable by pressing the corresponding keys on the keyboard. The next trial started as soon as participants responded, or after five seconds if no response was given. The order of trial presentation was randomized across participants. Once the main study started, participants were not given feedback on their performance.

4 It is worth noting that the first eight participants had a timeout threshold of 3 s. Three of these participants yielded over 10% of timeouts and were thus excluded from the analyses. To avoid a similar situation with the rest of the participants the timeout threshold was increased to 5 s.
Screening tests
Each screening test included 20 items. The nonwords included in the first screening test (see Appendix A, List 1) were segmentally identical to those included in the second screening test (see Appendix A, List 2); however, they were stressed on different syllables across the two lists (see Materials section). Each trial started with the visual presentation of a + sign, which remained on the screen for one second, and was followed by the aural presentation of the target. Participants were told that they had five seconds to respond. The next trial started as soon as they responded, or after five seconds if they failed to respond. Participants were informed that “no responses” would count as errors. The order of trial presentation was randomized across participants. At the end of each screening test, the experimenter checked the output. If participants had made three or fewer errors (minimum 85% correct), they proceeded to the main study. In contrast, if they had made more than three errors, they proceeded to a practice session after the first screening test, or they left the study after the second screening test.

Practice session
Before the practice session began, participants were told that a list of nonwords would be played to them, and that they had to listen to them carefully through the headphones. Each nonword would be played twice. The first time the stress would be on the first syllable, while the second time it would be on the second syllable. Participants were also asked to repeat each nonword as they heard it. The list of nonwords included in the practice session corresponded to the nonwords of Lists 1 and 2 combined (see Appendix). Each trial started with the visual presentation of a + sign, which remained on the screen for one second, and was followed by the aural presentation of the target, which participants had to repeat. Participants had two seconds to do that, after which the next trial started. The order of trial presentation was identical for every participant. Participants’ responses during the practice session were not evaluated. Following the practice session, participants were asked to do the second screening test. Out of all participants included in the analyses (N = 103), 65 did the practice session before proceeding to the main study.

Acoustic analyses
The FAVE Forced Aligner (Rosenfelder et al., 2014) was used to align the 8,801 spoken nonwords to the phone level (https://github.com/Jof rhwld/FAVE/wiki/FAVE-align). The first author (P.M.) subsequently hand-corrected the phone-level alignment and manually labeled each of the nonwords and their relevant segments (e.g., the vowels and the consonants of the first and the second syllable) using Praat (Boersma & Weenink, 2020). Alignment decisions are likely to be subjective. Therefore, a trained research assistant, who was naïve to the purposes of the study, carried out the same procedure as P.M. for four of the speakers (40% of the data). Hand-checking of the phone-level alignment was done via visual (and when necessary, auditory) inspection of the acoustic waveform and the spectrogram. To ensure alignment reliability, both P.M. and the research assistant followed the segmentation criteria established in the ANDOSI database (Croot, Fletcher, & Harrington, 1992) and by Rastle, Croot, Harrington, and Coltheart (2005). To evaluate agreement between the two individuals, Kendall rank correlation tests were calculated for vowel durations. This was done for all items, for each of the four speakers separately. Critically, correlations were high and significant for all vowels and for all four speakers (Speaker 20: r = 0.87, p < .001, for both the 1st and the 2nd vowel; Speaker 22: r = 0.90, p < .001, for the 1st vowel, and r = 0.88, p < .001, for the 2nd vowel; Speaker 36: r = 0.83, p < .001, for both the 1st and the 2nd vowel; Speaker 40: r = 0.92, p < .001, for the 1st vowel, and r = 0.84, p < .001, for the 2nd vowel), thus suggesting high fidelity to the established segmentation criteria. We provide labeled segmentations of all items uttered by one example speaker (i.e., Speaker 20) on https://osf.io/hngfb/.

We measured intensity and f0 at the acoustic midpoint of each vowel, using a Praat script (note that we use the term ‘f0’ here rather than ‘pitch’, since we are referring to acoustic measurements). The f0 was not measurable in 480 tokens (4.7% of the data). Praat returned the reading as ‘undefined’ for either the 1st, 2nd, or both vowels in these cases. Furthermore, the distributions of the remaining f0 measurements were characterised by a large number of outliers. Some of the outliers were clear measurement errors, representing values in unrealistic data ranges, such as f0 of 400 Hz or higher. There were also outliers suggesting non-modal phonation, e.g., 80 Hz for a female speaker, suggesting the presence of creaky phonation. A spot-check of the data confirmed that non-modal phonation was quite common in the data, partially due to individual speaker voice characteristics, and partially due to the test items appearing in isolation, and therefore adjacent to a major prosodic boundary.

In order to ensure the measurements were reliable, we removed all f0 measurements that fell outside the inter-quartile f0 range for any individual speaker. This procedure led us to discard another 1112 spoken tokens. While this is a substantial portion of the data (12.63%), the data trimming was deemed necessary to avoid erroneous measurements skewing the results. The final analysis included 7209 spoken tokens. For these tokens, we calculated the difference between V1 and V2 in duration, f0, and intensity. The duration difference was measured in seconds. The f0 was measured in Bark (first converted from Hertz). Bark scale was used for f0 analysis because it is more representative of perception than Hertz and our questions concern perception. The intensity difference was measured in decibel (dB). Because the decibel scale is logarithmic, it is more appropriate to express a contrast between two measurements in terms of a difference rather than a ratio. For consistency and ease of interpretation, we used a difference for the other two measurements as well.

The Praat script that we used to extract the measurements is available online: https://web.mit.edu/zqi/www/uploads/1/4/8/9/14891652/get_measurements.praat.

Statistical analysis
The data were analysed using mixed effect logistic regression, predicting the likelihood of a word being perceived as having initial stress. The number of observations in the model (equivalent to the number of stress judgments for spoken tokens following data cleaning) was 74,404. The fixed predictors included V1-V2 duration difference, V1-V2 intensity difference and V1-V2 f0 difference. In order to test the significance of the main predictors, we fitted a model with the random structure as described above, and a three-way interaction between V1 and V2 duration difference, V1-V2 intensity difference, and V1-V2 f0 difference. We then compared it with a series of simpler models, starting with a model that contained two-way interactions between the individual fixed predictors, then dropping the two-way interactions one-by-one, and finishing with a model that only contained main effects of V1-V2 duration difference, V1-V2 intensity difference, and V1-V2 f0 difference.

The model also included random intercepts for speaker, listener, and item, as well as uncorrelated random slopes for V1-V2 duration difference, V1-V2 intensity difference, and V1-V2 f0 difference by listener. The by-listener random slopes were included to account for the possibility that listeners respond to the same cue differently. This was the most complex random structure we could include before encountering convergence issues. We used the random part of the model to reduce autocorrelation and to gauge individual differences between listeners in how they responded to different stress cues.

Results

Descriptive results

Of the 74,404 listener responses analysed in the study, 43,683
(58.71%) indicated perception of initial stress, whereas 30,721 (41.29%) responses indicated perception of final stress. The frequencies and relative percentages of initial and final responses, depending on the speaker who produced the stimuli, are summarised in Table 1.

Overall, we can see some preference for initial stress over final stress, with further variation depending on the speaker.

Results of average performance

Below, we present the results of the model comparison procedure, and the generalizations that emerge from the selected best model.

The three-way interaction between V1 and V2 duration difference, V1–V2 intensity difference, and V1–V2 f0 difference was significant. This was established through the model comparison procedure described previously. The model with a three-way interaction had a lower AIC, a lower BIC, and it also had a significantly higher log likelihood compared to the model with three two-way interactions (chi.sq = 27.43, df = 1, p < .001). Table 2 provides a summary of the model comparisons.

Multicollinearitity was assessed for the model by analysis of the VIF values, all of which were within acceptable range. The VIF values are reported in Appendix B, along with a summary of correlation values between the independent variables (both aggregated across participants and for individual participants). Table 3 summarises the influences of the fixed effects included in the regression model on the likelihood of perceiving initial stress.

The results of the model suggest that duration difference, intensity difference and f0 difference all affect stress perception. Moreover, the presence of a significant three-way interaction indicates that the contribution of each of these cues to the perception of initial stress depends on the value of the remaining two cues.

In order to understand the effect of the individual cues and the nature of their interaction, let us first consider the isolated effects of each V1–V2 difference in the absence of other differences. The top left panel of Fig. 1 illustrates the effect of duration difference between V1 and V2, when there is no difference between the two vowels in terms of f0 or intensity. In this scenario, when the duration difference equals 0, stress is ambiguous: it is equally likely to be judged as initial or as final.

Positive duration differences are associated with initial stress, and the greater the positive durational difference, the greater the likelihood of initial stress. Negative duration differences are associated with final stress: the likelihood of initial stress becomes smaller as the durational difference between V1 and V2 decreases. In other words, when the two vowels in a word are identical in terms of intensity and f0, the longer of the two vowels tends to be perceived as stressed. Similarly, when the duration difference and f0 difference between V1 and V2 equals 0 (top right panel of Fig. 1), positive intensity differences are associated with initial stress, whereas negative intensity differences are associated with final stress. When the duration difference and intensity difference equal 0 (bottom left panel of Fig. 1), the likelihood of initial stress increases for relatively positive V1–V2 f0 differences, and it decreases for negative V1–V2 f0 differences.

The effects in Fig. 1 illustrate a situation in which only duration differs between the two vowels, but not intensity or f0, or when only intensity varies between the two vowels, but not duration or f0, etc. However, let us now turn to situations in which V1 and V2 differ according to multiple cues. Fig. 2 shows the interaction between f0 difference and intensity difference when there is a relatively large negative V1–V2 difference in duration, specifically when V1 is 161 ms shorter than V2. The y-axis corresponds to the likelihood of initial stress. The value of −161 ms difference corresponds to the 10th percentile of the distribution of the V1–V2 duration difference. V1–V2 intensity difference is represented using a colour scale. Five reference values are shown in the plot, corresponding to five equally spaced points on the distribution of this difference, between 10th and 90th percentile. When a negative duration difference of −161 ms is combined with no V1–V2 difference in f0 or intensity, stress is likely to be perceived as final (likelihood of initial stress in such a case is only 0.10). The likelihood of initial stress is even smaller when the −161 ms duration difference is accompanied by a negative f0 difference, or a negative intensity difference, or both. However, the effect of duration can be offset by a very large positive f0 difference. For example, if the −161 ms duration difference is combined with an f0 difference of 0.61 Bark or more and no difference in intensity, stress is somewhat more likely to be perceived as

<table>
<thead>
<tr>
<th>Speaker</th>
<th>Number of stress-initial responses</th>
<th>Number of stress-final responses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speaker11</td>
<td>3757 (49.89%)</td>
<td>3773 (50.11%)</td>
</tr>
<tr>
<td>Speaker16</td>
<td>4997 (52.31%)</td>
<td>4555 (47.69%)</td>
</tr>
<tr>
<td>Speaker17</td>
<td>4738 (68.27%)</td>
<td>2202 (31.73%)</td>
</tr>
<tr>
<td>Speaker20</td>
<td>3456 (46.39%)</td>
<td>3994 (53.61%)</td>
</tr>
<tr>
<td>Speaker22</td>
<td>3656 (56.51%)</td>
<td>2814 (43.49%)</td>
</tr>
<tr>
<td>Speaker32</td>
<td>4383 (65.52%)</td>
<td>2307 (34.48%)</td>
</tr>
<tr>
<td>Speaker36</td>
<td>3935 (62.07%)</td>
<td>2405 (37.93%)</td>
</tr>
<tr>
<td>Speaker40</td>
<td>4585 (54.91%)</td>
<td>3765 (45.09%)</td>
</tr>
<tr>
<td>Speaker5</td>
<td>5245 (64.04%)</td>
<td>2697 (33.96%)</td>
</tr>
<tr>
<td>Speaker7</td>
<td>4901 (69.06%)</td>
<td>2209 (30.94%)</td>
</tr>
</tbody>
</table>

Table 2
Summary of model comparison, depending on different combinations of fixed predictors.

| Duration diff + Intensity diff + f0 diff | 65,010 | 65,102 | −32,495 |
| Duration diff: f0 diff + Intensity diff: f0 diff | 65,012 | 65,123 | −32,494 |
| Duration diff: Intensity diff + Intensity diff: f0 diff | 65,011 | 65,121 | −32,493 |
| Duration diff: Intensity diff + f0 diff | 65,010 | 65,120 | −32,493 |
| Duration diff: f0 diff + Intensity diff: f0 diff | 65,011 | 65,131 | −32,493 |
| Duration diff: Intensity diff + Intensity diff: f0 diff | 65,011 | 65,115 | −32,479 |

Table 3
Influences of durational differences, intensity differences, and f0 differences between syllables on the likelihood of initial stress.

<table>
<thead>
<tr>
<th>Estimate</th>
<th>Std Error</th>
<th>Z</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>0.06</td>
<td>0.13</td>
<td>0.45</td>
</tr>
<tr>
<td>Duration difference</td>
<td>13.73</td>
<td>0.95</td>
<td>14.40</td>
</tr>
<tr>
<td>Intensity difference</td>
<td>0.15</td>
<td>0.01</td>
<td>11.89</td>
</tr>
<tr>
<td>f0 difference</td>
<td>4.17</td>
<td>0.37</td>
<td>11.28</td>
</tr>
<tr>
<td>Duration difference: Intensity difference</td>
<td>0.36</td>
<td>0.07</td>
<td>4.94</td>
</tr>
<tr>
<td>Duration difference: f0 difference</td>
<td>3.87</td>
<td>1.78</td>
<td>2.18</td>
</tr>
<tr>
<td>Intensity difference: f0 difference</td>
<td>−0.11</td>
<td>0.03</td>
<td>−3.92</td>
</tr>
<tr>
<td>Duration difference: Intensity difference: f0 difference</td>
<td>−1.41</td>
<td>0.28</td>
<td>−5.04</td>
</tr>
</tbody>
</table>
initial than final (note that this particular scenario is not directly plotted in Fig. 2). The likelihood of initial stress then increases for greater magnitudes of V1–V2 f0 difference, or if the positive f0 difference is combined with a positive intensity difference. In contrast, a large positive intensity difference alone is not sufficient to offset a large negative duration difference, where ‘large’ is defined as 90th percentile of intensity difference (9.08 dB) and 10th percentile of the duration difference (−161 ms).

The median V1–V2 duration difference value in the data is −49 ms. Fig. 3 illustrates the effect of f0 difference and intensity difference on the likelihood of initial stress when the V1–V2 duration difference is equal to −49 ms. When the f0 difference and intensity difference both equal 0, the −49 ms duration difference corresponds to 0.35 likelihood of initial stress. This means that initial stress is less likely than final, but the likelihood of initial stress is nevertheless higher than in the case of −161 ms duration difference, considered above. If the −49 ms duration difference is accompanied by a positive f0 difference (ca. 0.2 Bark or more), stress perception is likely to be initial, even in the absence of an intensity difference. If the −49 ms duration difference is accompanied by a high positive intensity difference (5 dB or more), stress perception is likely to be initial, even in the absence of an f0 difference. In the presence of a large positive intensity difference, or a large positive f0 difference, the likelihood of initial stress increases.

Fig. 4 presents the interaction between V1 and V2 f0 difference and V1–V2 intensity difference when the V1–V2 duration difference is equal to 35 ms. While this positive duration difference might seem numerically small compared to the negative difference we have seen so far, it represents the 90th percentile of the V1–V2 duration difference distribution, and is therefore close to the upper limit of what we find in the data. In the presence of a 35 ms positive duration difference, stress is generally more likely to be initial, unless it is accompanied by both a negative f0 difference and a negative intensity difference. If the 35 ms duration difference is enhanced by a positive f0 difference of 0.7 Bark or more, the likelihood of initial stress approaches 1. Similarly, the presence of a positive intensity difference increases the likelihood of initial stress.

In sum, the results of the logistic regression modelling confirm that of the two syllables in a word, the one with longer duration, greater f0 or greater intensity tends to be perceived as stressed. However, the picture becomes more complex whenever a conflict arises between these cues; for example, when V1 has longer duration but lower f0 than V2. Cases of such cue conflict are common in our data: in fact, they are the norm. Only 15.99% of the stimuli used in our study are characterized by a positive duration difference, a positive f0 difference and a positive
intensity difference. Instances of all three differences being negative are even rarer, at 2.36%.

As shown in Fig. 5 below, the V1–V2 duration difference for the stimuli used in our study is typically negative, whereas the f0 and intensity differences are typically positive. Thus, V1 is shorter than V2 but higher in f0 and intensity most of the time. These tendencies are likely due to prosodic boundary effects influencing duration, f0, and intensity. The tendency for V1 to have a higher intensity and f0 than V2 is potentially a strategy to mark the beginning of the word. The tendency for V2 to be longer than V1 is likely a reflection of final lengthening (Klatt, 1975), especially prominent if the V2 is in an open syllable.

It would appear that boundary-related prosodic pressures create an inherent conflict between duration, f0, and intensity as stress cues. The presence of such a conflict considerably complicates the task of the listener in making stress judgements, requiring them to normalise the relative values of duration, f0, and intensity. Put differently, the task of the listener in making a stress judgement is not simply a matter of judging whether the first or second vowel in the word is longer (or higher in f0 or intensity), but rather, whether the same vowel in the same position would have been longer (or higher in f0 or intensity) under a different stress condition.

To a large extent, listeners are able to apply such normalization. This is evident for instance from the fact that the effect of the V1–V2 duration difference is asymmetrical. As discussed above, a V1–V2 duration difference of −49 ms can lead listeners to either a positive or a negative stress judgement, depending on the values of the remaining cues. We
could attribute this effect to the listener’s expectation that V2 is generally longer than V1, requiring them to use other cues to disambiguate the stress. In contrast, a V1–V2 duration difference of +35 ms, is strongly indicative of initial stress, because listeners do not generally expect V1 to be longer than V2.

**Individual variation**

While listeners are clearly capable of normalization in stress perception, the complexity of the task creates scope for ambiguity. In order to quantify the extent of this ambiguity, we calculated inter-listener agreement rates, summarised in Table 4 below. Kappa-Fleiss values were calculated separately for each speaker, because different

---

**Fig. 4.** Interaction between f0 difference (Bark) and intensity difference (dB) when the V1–V2 duration difference equals 35 ms.

**Fig. 5.** Distributions of the V1–V2 differences in duration (s), f0 (Bark), and intensity (dB) in the stimuli used in the study.
speakers were rated by different subsets of listeners. The agreement was significant overall; however, there were considerable differences in the level of agreement, depending on the speaker. Based on the classification proposed by McHugh (2012), there was no agreement for speaker 36, minimal agreement for speakers 7, 5, 17, 11 and 16, weak agreement for speakers 32, 22 and 40 and moderate agreement for speaker 20. Note that there were relatively more listeners assigned to speakers 5 and 16, which might have negatively affected the Kappa values in these two cases.

The results of the consistency analysis show that listeners’ agreement levels vary considerably, depending on the speaker. This suggests that some speakers tend to produce stress patterns that are highly ambiguous, whereas others pronounce the nonwords in a way that cues the stress for the listeners with less ambiguity. The perception results suggest that this ambiguity could arise in two ways. One potential source of ambiguity is producing V1–V2 differences that are close to 0, since we have seen that a zero difference between V1 and V2 in terms of duration, f0 or intensity leads to ambiguous stress perception. Alternatively, ambiguity could arise from producing strongly conflicting cues; for example, producing a large negative V1–V2 duration difference along with a large positive V1–V2 f0 difference.

We conducted an initial analysis to quantify the magnitude of phonetic contrasts produced by each speaker, and to assess the impact of this variable on listeners’ agreement of the stress. We first calculated the standard deviation for each phonetic cue for each speaker. A large standard deviation would indicate that a speaker tends to produce large phonetic contrasts, whereas a small standard deviation would arise if a speaker does not use the whole available range of a particular phonetic dimension. We then calculated the Pearson’s correlation coefficient between the standard deviation for each speaker and the speaker’s Kappa value. The correlations are illustrated in Fig. 6. We observed a moderate positive correlation between the standard deviation of duration difference and the Kappa value (r = 0.66, p < .05). This would suggest that overall, listeners show greater agreement in their stress judgements when the speakers produce large duration differences

<table>
<thead>
<tr>
<th>Speaker</th>
<th>Kappa</th>
<th>N (listeners)</th>
<th>Z</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speaker 36</td>
<td>0.2</td>
<td>10</td>
<td>39.91</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Speaker 7</td>
<td>0.25</td>
<td>10</td>
<td>48.02</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Speaker 5</td>
<td>0.31</td>
<td>11</td>
<td>67.48</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Speaker 17</td>
<td>0.34</td>
<td>10</td>
<td>67.33</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Speaker 11</td>
<td>0.37</td>
<td>10</td>
<td>73.57</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Speaker 16</td>
<td>0.38</td>
<td>12</td>
<td>90.41</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Speaker 32</td>
<td>0.46</td>
<td>10</td>
<td>92.24</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Speaker 22</td>
<td>0.47</td>
<td>10</td>
<td>93.11</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Speaker 40</td>
<td>0.5</td>
<td>10</td>
<td>100.35</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Speaker 20</td>
<td>0.63</td>
<td>10</td>
<td>125.83</td>
<td>&lt;.001</td>
</tr>
</tbody>
</table>

Fig. 6. Correlation plots between the standard-deviation of the V1–V2 differences measures for each speaker, and the by-speaker Kappa values.
between V1 and V2. Speaker 20 appears to be an exception, because this speaker receives most consistent stress judgements of all, while their use of duration contrast is comparable to three other speakers, for whom agreement levels are lower. The correlation coefficient of 0.48 for the standard deviation of intensity difference suggests a low positive correlation with the Kappa value, but the correlation was not significant (perhaps given that there were only 10 speakers). For the standard deviation of f0 difference, we did not find any correlation with the Kappa value.

The results of the consistency analysis begin to suggest that there may be individual differences in the way that speakers communicate stress that have consequences for listeners.

In order to explore this possibility further, we conducted an analysis of the random by-listener effects captured in the model presented in Table 3 and Figs. 1-4. The fixed part of the model, as discussed in the previous section, represents an averaged effect across all speakers and listeners. However, the average effect is not necessarily representative of each and every individual. This is confirmed by the fact that the by-listener random slopes were all highly significant.

We tested the significance of by-listener slopes by removing the within-listener effects in a stepwise fashion, and comparing the model fit. Removing the within-listener effect for duration difference significantly decreased the log likelihood ratio (chi-square = 2153.5, df = 1, p < .001). Similarly, removing the within-listener effect for f0 difference negatively affected the model fit (chi-square = 936.91, df = 1, p < .001), and the same was true for removing the within-listener effect for intensity difference (chi-square = 712.2, df = 1, p < .001). From this, we conclude that the predictive value of any individual cue for stress assignment differs significantly by listener, i.e., individual listeners may respond to the same cues differently.

Model predictions for individual listeners are shown in Fig. 7. The figure illustrates the likelihood of initial stress for each individual listener, depending on the value of the relevant stress cue. In each case, the intercept is set to 0, and so, for instance, the prediction for the V1–V2 duration difference assumes no difference in f0 or intensity. The prediction for V1–V2 f0 difference assumes no difference in duration or intensity, etc. This visualisation parallels Fig. 1, except that each individual listener is visualised in this case. From visual inspection, it is clear that listeners vary considerably in how they use the individual cues. For each cue, there is a group of listeners who show categorical perception, which corresponds to an S-shaped curve with a steep slope. However, the same is not true of all listeners.

There is no established approach to defining a curve as representing categorical perception, but we can generalise that categorical perception involves the presence of a category boundary, and thus a steep slope for the curve predicting the probability of the two response categories, depending on the value of the predictor. For each by-listener curve, we calculated the slope, using Discrete Cosine Transformation (Harrington, 2010). We then performed a clustering analysis on the slope values, using k-means clustering, and assuming two clusters. This approach allows us to assign the individual by-listener curves into two groups, depending on the steepness of the slope, in a more systematic way than by simply eyeballing the data. As we can see in Fig. 8, this approach identifies two distinct groups of listeners. Those with relatively steeper slopes are marked in red. These listeners show categorical perception for a specific cue. In contrast, the listeners marked in blue do not show categorical perception for a given cue; they may use that cue gradiently, or not at all.

Using this diagnostic for categorical perception, we investigated how frequently listeners show categorical perception of stress for specific cues, depending on the speaker. Fig. 9 compares two sets of listeners in this respect, i.e., the listeners who heard the stimuli produced by speaker 20 (top panel), and those who heard the stimuli produced by speaker 36. Those two speakers were chosen because they represent the two...
extremes in terms of inter-listener agreement (Table 4). Speaker 20 had the highest Kappa score, meaning listeners were most consistent in responding to the stimuli, whereas speaker 36 had the lowest Kappa score, corresponding to the lowest level of inter-listener consistency.

The data provided in Fig. 9 suggest that listeners tended to respond categorically to cues produced by Speaker 20. Out of 30 possible listener-cue combinations, 25 (83%) were categorical. In contrast, for Speaker 36, 14 out of 30 (46.6%) of possible speaker-cue combinations were categorical. These data suggest that Speaker 20 offers information about stress that listeners are generally able to interpret categorically, while the information being offered by Speaker 36 does not allow listeners to make a categorical distinction regarding stress as easily.

However, it is important to appreciate that variation in the way that listeners use acoustic cues to stress is not totally determined by the information being offered by the speaker. Fig. 9 shows that listeners vary in how they use the information provided by a single speaker, not only in terms of whether they interpret the information categorically, but also in the slope of the gradient and in the position of the category boundary (i.e., the magnitude of the V1–V2 difference corresponding to a 50% likelihood of initial stress). This variation in how listeners use information provided by individual speakers is shown for all listeners and speakers in our experiment in Appendix C.

We tested whether the higher inter-listener agreement rates described in Table 4 correspond to increased use of categorical perception. For each speaker, we calculated the relative frequency with which a listener used a cue categorically. The total frequency corresponds to the number of listeners exposed to specific speaker’s stimuli (typically 10) multiplied by the three cues: duration difference, f0 difference, and intensity difference. We then investigated the correlation between the relative frequency of categorically used cues per speaker and the Kappa value, as in Table 4. The correlation was significant at $r = 0.74, p = 0.015$ (see Fig. 10), suggesting that listeners tend to be more in
agreement about their stress judgements when relatively more of the available cues are used categorically.

We further considered how individual listeners integrate various stress cues, in terms of which cues they use categorically. Fig. 11 shows which cues were used categorically by individual listeners, comparing data from two subsets of listeners. Specifically, these were listeners who heard stimuli produced by Speakers 20 and 36, as previously in Fig. 9. In the case of Speaker 20, most listeners tended to use all three cues categorically. One listener did not use any of the cues in a categorical way, though Fig. 9 would suggest they had a gradient use of duration difference and intensity difference. Two speakers used duration difference and intensity difference categorically, but not f0 difference. Listeners who heard the stimuli produced by Speaker 36 present more variation, consistent with them also having lower agreement rates. Here, in a sample of 10 listeners, we find listeners using three, two, one, or none of the cues categorically. Three listeners used all three cues categorically, two used two of the three cues categorically; in each case, the cues were duration difference and intensity difference. One listener only used duration difference categorically, but not f0 difference or intensity difference, and four listeners did not use any of the cues in a categorical way.

This pattern of variation is interesting, because it suggests that all three cues are, in principle, available to be used categorically, but only a small number of listeners use all of them in this way. Furthermore, listeners seem to show some asymmetries with respect to which specific cues they use in a categorical way: duration difference was most likely to be used categorically, whereas f0 difference was the least likely. We know that Speaker 36 tended to produce quantitatively small differences between V1 and V2, especially for the f0 difference (Fig. 6). We therefore speculate that phonetically small contrasts may make it harder for listeners to use a cue categorically to determine stress.

Data like those presented in Fig. 11 are available for all speakers and listeners in Appendix D. We used the data in Appendix D to investigate the frequency with which listeners combine duration, f0, and intensity cues to stress (see Table 5). This descriptive analysis suggests that over half of listeners (56 out of 103 listeners) used all three cues categorically. The numbers of listeners using two, one, or zero cues were comparable, at 16, 18 and 13 respectively. The most likely of these cue combinations were duration difference and intensity difference (9 listeners), or duration difference alone. Only one listener used f0 difference + intensity difference categorically, but not duration difference. Across all possible cue combinations, 80 listeners used duration difference.
categorically, 72 used intensity difference categorically, and 66 used f0 difference categorically. The results of the exploratory analysis of by-listener variation provide important context for the interpretation of the average results, presented in Table 3 and Figs. 1-4. The averaged data suggest a level of gradience in the likelihood of stress which is not typical of most individual listeners, since a typical listener shows categorical perception for at least some of the stress cues we have analysed. The gradience in the averaged data arises from the fact that individual listeners vary. They may vary in whether they use a specific cue categorically, or according to where the category boundary lies. To an extent, this variation appears to be down to the individual listener, since listeners who heard the same stimuli may use different cues to make stress judgements. However, there also seems to be a level of variation conditioned by the speaker, in respect of the acoustic information that different speakers offer regarding stress.

![Fig. 10.](image1.png)

**Fig. 10.** Correlation between by-speaker Kappa values and the frequency of individual perception cues being used categorically across all listeners.

![Fig. 11.](image2.png)

**Fig. 11.** Categorical vs. non-categorical use of individual perception cues, depending on the listeners. Left panel shows listeners who heard stimuli produced by Speaker 20. Right panel shows listeners who heard stimuli produced by Speaker 36. The row labels represent individual listener codes.

<table>
<thead>
<tr>
<th>Number of cues used categorically</th>
<th>Cues</th>
<th>No of listeners</th>
</tr>
</thead>
<tbody>
<tr>
<td>three cues</td>
<td>duration difference</td>
<td>6 16</td>
</tr>
<tr>
<td></td>
<td>+ f0 difference</td>
<td>9 16</td>
</tr>
<tr>
<td></td>
<td>+ intensity difference</td>
<td>1 16</td>
</tr>
<tr>
<td>two cues</td>
<td>duration difference</td>
<td>6 16</td>
</tr>
<tr>
<td></td>
<td>+ intensity difference</td>
<td>9 16</td>
</tr>
<tr>
<td></td>
<td>+ f0 difference + intensity difference</td>
<td>1 16</td>
</tr>
<tr>
<td>one cue</td>
<td>duration difference</td>
<td>9 18</td>
</tr>
<tr>
<td></td>
<td>f0 difference</td>
<td>3 18</td>
</tr>
<tr>
<td></td>
<td>intensity difference</td>
<td>6 18</td>
</tr>
<tr>
<td>zero cues</td>
<td></td>
<td>13</td>
</tr>
</tbody>
</table>

**Table 5**

Frequency of categorical use of specific stress cues by individual listeners.
Discussion

Lexical stress is an important source of information in the process of mapping the spoken signal to meaning. Research has shown that listeners use information about lexical stress to segment continuous speech into words (e.g., Mattys & Samuel, 2000) and in the process of recognising spoken words (e.g., Cooper, Cutler & Wales, 2002; Jesse et al., 2017; Reinsch et al., 2010; Soto-Faraco, Sebastian-Galles, & Cutler, 2001; van Donselaar, Koster, & Cutler, 2005). Thus, understanding the sources of information that contribute to the perception of lexical stress has been an important research question for nearly seventy years (e.g., Fry, 1955).

Our research sought to advance understanding of how different forms of acoustic information influence the perception of lexical stress. Research has consistently identified duration, pitch, and intensity as important perceptual cues (see e.g., Cutler, 2005, for review). However, while there is evidence that all of these forms of acoustic information may influence stress judgments, there has been disagreement around their relative importance. Our review of the literature indicated that one problem in this area has been the use of a limited number of items that are varied systematically on a particular property (e.g., intensity). This design choice severely limits scope to generalize from these studies; for example, finding that a particular acoustic property influences the perception of stress in the single stimulus ‘maba’ (Chrabaszcz et al., 2014) does not allow researchers to generalize the importance of that acoustic property in judging stress in other phonetic contexts. Likewise, the approach of systematically varying stimuli on a particular acoustic property while holding the others constant by setting specific parameter values does not allow researchers to investigate natural speech situations in which the acoustic cues to stress are in conflict (or indeed, even to estimate how frequently this occurs). Finally, the approach taken in previous studies has led to severe power limitations that limit opportunity to investigate how interactions between acoustic properties might influence stress judgments and that restrict analyses to average performance.

Our study sought to address these different limitations by developing a new approach based on the mega-study methodology (e.g., Balota et al., 2004; Keuleers et al., 2012). In the present study, over 100 listeners made stress judgements on nearly one thousand spoken productions of disyllabic nonwords each, yielding nearly 75,000 analyzable responses following data cleaning. These spoken productions were taken from a previous mega-study, in which adult participants were asked to read aloud disyllabic nonwords (Mousikou et al., 2017). This approach gives us unprecedented power to detect the influences of individual acoustic parameters and their interactions on stress judgments across a wide range of phonetic contexts, and it also allows us to begin to investigate the extent to which individuals use acoustic information in the same way in the perception of stress.

Results of the analysis of averaged data showed significant effects of duration, pitch, and intensity on stress judgments, replicating previous findings but in a much wider range of phonetic contexts. More importantly, our data revealed how these acoustic cues interact in the perception of lexical stress. Listeners tend to perceive stress when there are large differences between V1 and V2 in duration, pitch, or intensity, and when more than one cue signals a difference in the same direction between V1 and V2. These data suggest that the previous focus on the relative importance of these cues may be misguided. Our findings suggest that the relative importance of any one cue depends on the value of the other cues. For example, if the duration and pitch difference both suggest initial stress, then intensity has only a very small impact on the likelihood of judging initial stress (see the left-hand side of Fig. 2). However, if the duration and pitch difference conflict, then intensity has a transformative impact on the likelihood of judging initial stress (see the left-hand side of Fig. 4). Indeed, the interactions described in Figs. 2 through 4 reveal that large differences between V1 and V2 on any one cue can be overturned by differences in the other direction for the other cues. We believe that the nature of the designs in this research area specifically, the tendency to systematically vary a single acoustic property in a stimulus or handful of stimuli and measure its influence on stress judgments - has led researchers to think about duration, pitch, and intensity in a stable hierarchy of importance (see e.g., Guo, Chen & Guo, 2017, who summarise the different hierarchies that have been proposed). Our findings show that listeners weigh different forms of acoustic information in judging stress in a more nuanced way that takes account of those forms of information in combination in different spoken tokens.

The insight that listeners weigh different forms of acoustic information in combination is important, because our study also revealed that duration, pitch, and intensity were in conflict in over 80% of the naturally-produced tokens used. This fundamental observation that cue conflict is the norm has not been appreciated in previous studies, because those studies have used synthesized speech with carefully-controlled acoustic properties (e.g., studies where one property was varied while the others were held constant). The conflict between different forms of acoustic information substantially complicates the task of the listener in judging stress. Listeners need to go beyond judging the direction of the difference between V1 and V2 on any single parameter. Instead, they need to assess the magnitude of any difference between V1 and V2 on one parameter, often in relation to the magnitude of any difference between V1 and V2 on the other parameters. Moreover, these judgments need to be made in relation to any baseline differences that might be expected for the different parameters; for example, the expectation that V2 will usually be longer in duration than V1 (see Fig. 5). Further research is needed to understand how listeners weigh different (conflicting) sources of information in combination, and how they determine baseline expectations for differences across syllables in relevant acoustic parameters. For example, do these baseline expectations reflect a listener’s overall language experience, or do listeners adjust to individual speakers’ ranges of duration, pitch, and intensity? This question cannot be addressed in our design but would make an interesting follow-up study.

In the analysis of the averaged data, the interactions offer new understanding of how listeners use information about duration, pitch, and intensity in combination to judge stress. However, the mega-study approach developed in this study allowed us to go beyond averaged data to ask whether listeners vary in how they use these sources of information. Our results showed that most listeners (55%) used all three acoustic cues categorically in judging stress. However, a significant percentage of listeners used two cues (16%), one cue (17%) or zero cues (12%) categorically. In these cases, listeners were most likely to use duration categorically to inform their stress judgments, but it must be noted that 22% of listeners in our sample did not use duration in this manner. This individual variation appears to arise both from variation in the information that different speakers provide about stress and from variation in how listeners use this information. It appears that different speakers may offer higher- or lower-quality information about stress; for example, 80% of participants listening to Speaker 32 used all cues categorically while just 30% of participants listening to Speaker 36 used all cues categorically. The analyses reported in Fig. 6 suggest that Speaker 32 uses a wider range of phonetic contrasts than Speaker 36 on all three of the dimensions studied, and therefore provide converging evidence that Speaker 32 offers higher-quality information about stress than Speaker 36. However, our data also show that listeners who responded to the same spoken tokens used the acoustic information in different ways, depending on the specific tokens and the context in which they were presented.

5 We speculate that this situation arises, because the phonetic cues used to signal stress are also used for other prosodic functions – for example, to signal whether a syllable is initial or final – and this creates an inherent conflict between duration, pitch, and intensity. We speculate that the prosodic effects associated with the position of the syllable within the word persevere even though our task involved reading tokens in isolation.
those tokens differently. Results showed that the extent to which listeners can use acoustic cues categorically is strongly associated with the consistency with which listeners judge stress for a particular speaker (Fig. 10).

These initial findings regarding individual differences are important, because they further undermine the notion that acoustic cues to lexical stress exist in some type of hierarchy of importance. We have already established from the analysis of average data that the relative importance of any one acoustic cue depends on the value of the other acoustic cues. These findings from the analysis of individual data further suggest that the importance of any one acoustic cue may depend on the nature of information offered by the speaker and on perceptual ‘settings’ of the listener. It will be important for future research to determine how different speakers vary with regard to the stress information they offer, and to determine why different listeners appear to prioritize different sources of information (even while listening to the same speaker).

Our findings from the analysis of individual data are also important because they modulate the interpretation of the results from the averaged data (and, indeed, the interpretation of all previous research drawing inferences on the basis of averaged data). Specifically, results from our analyses of averaged data suggest that listeners use duration, pitch, and intensity in a gradient manner. However, the appearance of gradience emerges because of the averaging of individual profiles; most listeners use all three cues categorically. Likewise, previous research based on averaged data has frequently investigated the slope of the function relating some acoustic cue to the percentage of participants making a particular stress judgment; these studies then infer that the cue with the steepest slope is most important (e.g., Fry, 1955; Zhang & Francis, 2010). Our findings suggest that these types of analyses may reflect different numbers of participants using particular cues categorically rather than a genuine difference in the gradience of the functions.

One surprising outcome of our study was the large number of participants who failed the screening tests and were unable to participate in the study. The fact that nearly 37% of the original sample could not judge stress well enough to participate in the study was unexpected given that (a) the screening test used disyllabic nonwords with clearly articulated stress, and (b) the procedure session used minimal nonword pairs that differed only in stress, and thus should have drawn participants’ attention to the acoustic information signalling stress. Substantial research has investigated stress ‘deafness’ in second-language learners (e.g., Dupoux, Pallier, Sebastian, & Mehler, 1997), and whether this arises as a result of linguistic or perceptual factors (e.g., Dupoux, Sebastian-Galles, Navarrete, & Peperkamp, 2008; Zhang & Francis, 2010). However, we are unaware of research suggesting that listeners have difficulty in judging where word stress falls in their native language. If some native speakers find it difficult to perceive word stress, then it raises the question of the extent to which stress can be used as a distributional cue to word boundaries in speech segmentation (Cutler & Norris, 1988). However, it is also possible that these participants perceived stress normally, but were non-compliant, or had difficulty in understanding the concept of stress (i.e. the instructions) or in making explicit stress judgements. Likewise, it is important to remember that the criterion for participation was relatively high at a minimum of 85% correct on the second screening test; this means that participants with scores as high as 80% were considered to have failed the screening tests. It will be for future research to investigate whether some listeners have particular difficulty in perceiving stress in their native language.

This point brings us to consider the ecological validity of our study. We have already undertaken research on how vowel reduction influences stress judgments, and the studies that have been undertaken typically use a single stimulus (e.g., maba; Cribaszczy et al., 2014) synthesized on a full-vowel-to-reduced-vowel continuum. Quantifying the degree of vowel reduction of the vowels in our dataset would not have been possible, as this would have required full-vowel realizations of each vowel for each speaker against which the produced vowels could be measured. Furthermore, as our stimuli were nonwords, it would not have been possible to assume what the full vowel realization would have been. Understanding how the analysis of vowel reduction may influence stress judgments in naturalistic speech, and understanding how any analysis of this metrical stress information may combine with the analysis of duration, pitch, and intensity, as described in this article, is an important matter for future research.

To conclude, the use of a mega-study approach (Balota et al., 2004; Mousikou et al., 2017) has allowed us to make a substantial advance in our understanding of how English-speaking adults use different forms of acoustic information in the perception of lexical stress. Our work confirms that listeners use duration, pitch, and intensity in judging stress, but adds a series of important new insights to this literature. These include the observations that (a) these acoustic cues to stress often conflict with one another in natural speech; (b) the importance of any one of these cues depends on the value of the other cues in a particular token; (c) speakers differ in the quality of information that they offer regarding stress; and (d) listeners differ in how they use that information. Thus, despite decades of experimental research, our study shows that there is much still to learn about the nature of stress perception, and this may require posing different types of research questions. In particular, we may no longer need to ask questions about the relative importance of different sources of acoustic information, but about how different listeners combine these often-conflicting sources of information in different phonetic, environmental, and speaker contexts. Likewise, while our paradigm has allowed us to magnify the influence of acoustic information in judging lexical stress, it will be important to move toward a theory of how listeners use acoustic information in combination with segmental, lexical, and contextual cues available in spoken language (see e.g., Mattys et al., 2005, for a similar problem in relation to speech segmentation).

CRediT authorship contribution statement

Petroula Mousikou: Writing – review & editing, Writing – original draft, Visualization, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization.
Patrycja Strycharczuk: Writing – review & editing, Writing – original draft, Visualization, Methodology, Investigation, Formal analysis, Data curation. Kathleen Rastle: Conceptualization, Funding acquisition, Methodology, Project administration, Supervision, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Our data and analysis code are available on the Open Science...
Framework (https://osf.io/hngfb/).

Acknowledgments

This research was supported by a research grant from the British Academy (SG151288) awarded to Kathleen Rastle and Petroula Mousikou. The authors thank Luke Burston, Becky Crowley, Cheng-Yu Hsieh, Clare Lally, and Svetlana Rodinskaya for research assistance. The authors have no competing interests to declare.

Appendix A

Nonword items included in the screening test and in the practice session.

**Screening test**


*List 2.* doPOM /dəˈpɒm/, praDUS /prəˈdəs/, mesFOD /mesfəd/, ROJAT /rəʊdʒət/, doRUS /dərəs/, prALel /prələl/, doZAT /dʊzət/, prASOF /prəˈsɒf/, prAPEM /prəpəm/, MESSut /mezˈdʌt/, DOrof /dərəf/, PRAkET /prəkɛt/, doTep /dətɛp/, rONEN /rəʊnɛn/, PRAzan /prəˈzæn/, PRAtoB /prətəb/, ROvON /rəʊˈvʌn/, PRAvun /prəˈvʌn/, ROSom /rəʊˈsʌm/.

**Practice session**


Appendix B

Summary of the VIF values for the predictors in the logistic regression model reported in Table 2.

**Duration difference** 1.17

Intensity difference 1.44

f0 difference 1.24

Duration difference: Intensity difference 2.56

Duration difference: f0 difference 2.07

Intensity difference: f0 difference 2.33

Duration difference: Intensity difference: f0 difference 3.78

The table below summarises the correlations between pairs of independent variables in the same model. Pearson’s correlation coefficients are reported in each case.

<table>
<thead>
<tr>
<th></th>
<th>Duration difference: Intensity difference</th>
<th>Duration difference: f0 difference</th>
<th>Intensity difference: f0 difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>All speakers combined</td>
<td>0.35</td>
<td>0.15</td>
<td>0.28</td>
</tr>
<tr>
<td>Speaker 11</td>
<td>0.43</td>
<td>0.42</td>
<td>0.62</td>
</tr>
<tr>
<td>Speaker 16</td>
<td>0.39</td>
<td>0.24</td>
<td>0.48</td>
</tr>
<tr>
<td>Speaker 17</td>
<td>0.35</td>
<td>0.07</td>
<td>0.19</td>
</tr>
<tr>
<td>Speaker 20</td>
<td>0.25</td>
<td>0.19</td>
<td>0.13</td>
</tr>
<tr>
<td>Speaker 22</td>
<td>0.43</td>
<td>0.20</td>
<td>0.31</td>
</tr>
<tr>
<td>Speaker 32</td>
<td>0.15</td>
<td>-0.02</td>
<td>0.15</td>
</tr>
<tr>
<td>Speaker 36</td>
<td>0.28</td>
<td>-0.05</td>
<td>-0.11</td>
</tr>
<tr>
<td>Speaker 40</td>
<td>0.61</td>
<td>0.36</td>
<td>0.47</td>
</tr>
<tr>
<td>Speaker 5</td>
<td>0.18</td>
<td>0.17</td>
<td>0.28</td>
</tr>
<tr>
<td>Speaker 7</td>
<td>0.22</td>
<td>0.05</td>
<td>0.28</td>
</tr>
</tbody>
</table>

Appendix C

By-listener effects of the different cues on the perception of stress, grouped by speaker and cue. Each line represents the predicted likelihood of perceiving initial stress by an individual listener. The colour represents whether the cue was used categorically by a listener or not.
Appendix D

Categorical vs. non-categorical use of individual perception cues, for each combination of speaker and listener.