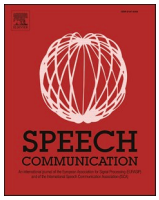
Speech Communication 173 (2025) 103267

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom



The perception of intonational peaks and valleys: The effects of plateaux, declination and experimental task

Hae-Sung Jeon

School of Psychology and Humanities, University of Central Lancashire, Preston PR1 2HE, United Kingdom

ARTICLE INFO

Keywords:
Speech perception
Intonation
Pitch
Prominence
Declination

ABSTRACT

An experiment assessed listeners' judgement of either relative pitch height or prominence between two consecutive fundamental frequency (f_o) peaks or valleys in speech. The f_o contour of the first peak or valley was kept constant, while the second was orthogonally manipulated in its height and plateau duration. Half of the stimuli had a flat baseline from which the peaks and valleys were scaled, while the other half had an overtly declining baseline. The results replicated the previous finding that f_o peaks with a long plateau are salient to listeners, while valleys are hard to process even with a plateau. Furthermore, the effect of declination was dependent on the experimental task. Listeners' responses seemed to be directly affected by the f_o excursion size only for judging relative height between two peaks, while their prominence judgement was strongly affected by the overall impression of the pitch raising or lowering event near the perceptual target. The findings suggest that the global f_o contour, not a single representative f_o value of an intonational event, should be considered in perceptual models of intonation. The findings show an interplay between the signal, listeners' top-down expectations, and speech perception.

1. Introduction

The definition of prominence has been under intense debates in speech prosody literature (see Ladd and Arvaniti, 2022 for a recent survey). In any event, native speakers of West Germanic languages can judge relative intonational prominence between utterance constituents (e.g. Cole et al., 2010; Knight, 2008; Turnbull et al., 2017) and they interpret the prominence in terms of informational salience or importance (e.g. Chen et al., 2007; Krahmer and Swerts, 2001; van Maastricht et al., 2016). For instance, in "I bought her a bottle of whisky, but it turns out that she doesn't LIKE whisky (see Ladd 2008, Chapter 6)", the repeated, second 'whisky' is likely to be spoken with reduced pitch, deaccented, being less informative than the accented word 'LIKE'. However, the relationship between accentuation and information load is not straightforward (see Bolinger, 1972; Ladd, 2008, Section 7.1) and not all intonational pitch accents are equally prominent. For instance, studies in German show that pitch accents with a high pitch, a rise, a late peak, or a long flat stretch of the fundamental frequency (f_o) are judged prominent, and they are associated with new information or focus (e.g. Baumann and Roth, 2014; Baumann and Röhr, 2015; Röhr et al., 2022). On the other hand, pitch accents with a low pitch, a fall, an early peak, or a sharp turn of the f_o contour are associated with low prominence and more accessible or given information (e.g. Chen et al., 2007; Gussenhoven, 2002; Knight, 2008; Pierrehumbert and Hirschberg, 1990 for English and Röhr et al., 2022 for German).

The present study tackles how listeners perceive relative pitch height and prominence for intonational peaks (rise-falls) and valleys (fall-rises). As further discussed below, a large body of research has explored the perception of intonational peaks or high prominence. However, little is known about the perception of valleys or low prominence (see Jeon and Heinrich, 2022a; Barnes et al., 2023), while the complexity of low prominence was pointed out from the early stage of the empirical investigation on intonation (Liberman, 1975). Hereafter, 'fundamental frequency (f_o)' refers to the acoustic parameter and 'pitch' refers to the perceptual sensation. This distinction is made because in this study, the f_o contour was experimentally manipulated to examine its perceptual consequence in pitch. The term 'extrema' is used as a cover term for both peaks and valleys. In the remainder of this paper, we examine listeners' perception as a function of the precise f_o contour shape including the f_o movement direction (peaks vs valleys, see Section 1.1), the precise shape of the f_o turn (i.e. whether it forms a flat stretch of f_o, a plateau of 25 ms, 50 ms or 100 ms and how they are paired (Section 1.2), the presence of

All stimuli are available upon request to the author.

E-mail address: hjeon1@uclan.ac.uk.

https://doi.org/10.1016/j.specom.2025.103267
Received 27 February 2023; Received in revised form 26 April 2024; Accepted 9 June 2025
Available online 10 June 2025

baseline declination (Section 1.3), and the effect of experimental task type (judging 'height' vs 'prominence', Section 1.4). Section 2 discusses the study design and predictions. Section 3 reports a perception experiment and Section 4 presents the experimental results. Section 5 and Section 6 respectively discuss findings and conclusions.

1.1. The perceptual asymmetry between intonational peaks and valleys

The present study focuses on the perceptual asymmetry between f_o peaks and valleys. For instance, in German, pitch accents with f_o rises are perceived as more prominent than pitch accents with low targets or falls (Baumann and Roth, 2014; Baumann and Röhr, 2015; Baumann and Winter, 2018; see Evans, 2015 and Hsu et al., 2015 for discussion on the physiological and psychological bases). The present study builds upon Jeon and Heinrich's (2022a,b) studies which took a psychoacoustic approach (cf. 't Hart et al., 1990). It was shown that listeners are better at discriminating the relative height (Jeon and Heinrich's, 2022a) or prominence (Jeon and Heinrich, 2022b) between two consecutive peaks compared to valleys. In both studies (Jeon and Heinrich, 2022a,b), the experimental materials were resynthesised English question utterances (e.g. 'does Nina know Mona?'). All experimental materials had either two f_o peaks or valleys, and the stressed syllable ('Nína', 'Móna') was associated with an f_o extremum (as in the present experiment; see Fig. 1). Rather than using naturally produced high (H*) or low (L*) pitch accents, the experimental materials were created to have the peaks as mirror images of the valleys. The height and shape of the first extremum remained constant throughout the experiment, but the second extremum varied in its height in five steps and its plateau duration. Native English speakers living in England participated in the experiments.

In two experiments in Jeon and Heinrich (2022a), participants judged which of the two peaks was higher and which of the two valleys was lower. The first experiment examined the effect of the extrema type (peaks, valleys), the second extremum shape (sharp turn, 100 ms plateau), and the item (four different sentences). The stimuli with two peaks and those with two valleys shared the baseline at 200 Hz, i.e. the pitch rose from the baseline to the maximum to form a peak, and it fell from the baseline to reach the minimum to form a valley. The second experiment re-examined the effect of the extrema type (peaks, valleys), using different durational conditions for the second extremum (25 ms, 100 ms plateau) and varying the f_o level of the intonational events (high, 200–302 Hz; low, 132–200 Hz). The results established that listeners' pitch height discrimination was significantly reduced for valleys compared to peaks. A change in one semitone in the second extremum height yielded relatively small changes in the responses for valleys than for peaks. For perceptual equivalence, compared to the peaks, listeners required a larger f_o excursion for the valleys, i.e. when the two valleys were perceived equal in height, the second valley was physically lower than the first. The f_o forming a 100 ms plateau increased the perceived pitch saliency compared to a sharp turn or a 50 ms plateau, making a peak sound higher and a valley lower (see Section 1.2 for further discussion). However, the pitch-saliency-enhancing effect of a long plateau was reduced for valleys, suggesting that this effect was potentially constrained by pitch perceptibility. Furthermore, while listeners showed good discrimination for the peaks regardless of the f_o level, the f_o changes associated with valleys at a low frequency level were not perceptually weighted as much as the same magnitude of f_o changes in semitones occurring in the higher level.

Jeon and Heinrich (2022b) replicated the perceptual asymmetry using a smaller contrast in the second plateau duration (25 ms vs 50 ms).

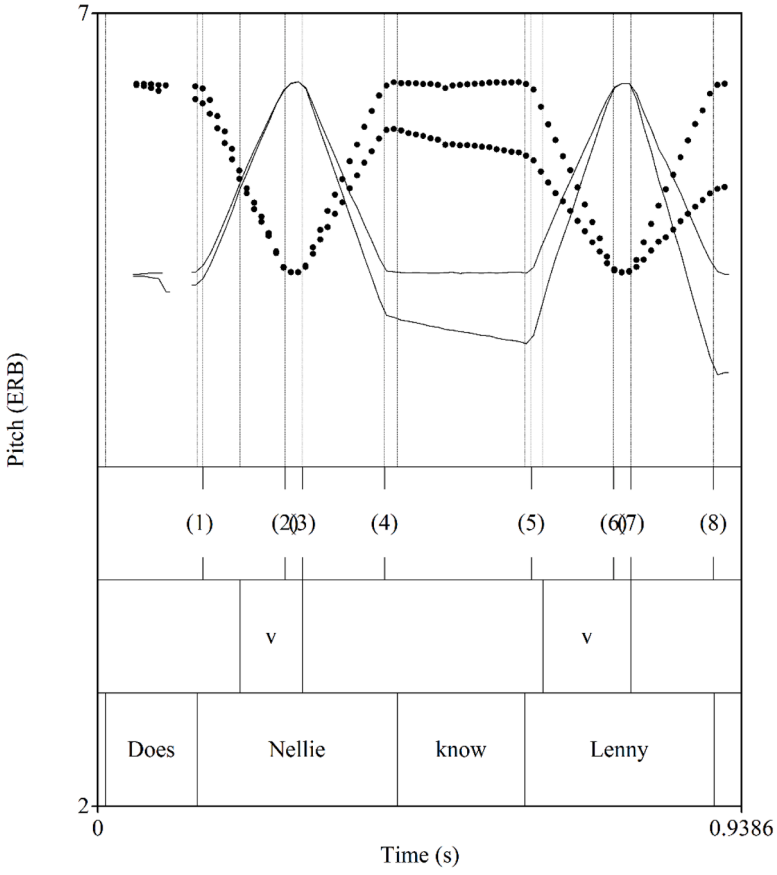


Fig. 1. Sample f_o tracks for two stimulus pairs with either two peaks (line) or two valleys (speckles). All f_o extrema in this figure form a 25 ms plateau aligned at the end of the stressed vowel. In each pair, one stimulus has a flat baseline (no declination) and the other has a declining baseline. The numbers indicate landmarks for f_o manipulation (e.g. 1: beginning of the rise/fall, 2: beginning of the plateau, 3: end of the plateau, 4: end of the rise/fall).

The ERB_N number scale (Glasberg and Moore, 1990; Patterson, 1976) was used to address the concern that the previously observed perceptual asymmetry might have been a by-product of using the logarithmic semitone scale. When the f_o rises and falls were expressed in semitones, the perceived magnitude of the pitch excursion could have been smaller for valleys relative to peaks (Rietveld and Gussenhoven, 1985). The one semitone step was equivalent to 12–14 Hz for the valleys and 14–17 Hz for the peaks when the f_o varied in the range of 200–302 Hz, whereas using the ERB_N number scale introduced similar f_o step sizes in Hertz for the peaks and the valleys (see Nolan, 2003).

1.2. The saliency-enhancing effect of f_o plateaux

A flat stretch of f_o, i.e. a plateau, associated with a peak makes it sound higher and more prominent compared to a sharper peak with the same f_o height (D'Imperio, 2000; Jeon and Heinrich, 2022a,b; Knight, 2008). In Knight's (2008) first experiment, six listeners judged which peak was higher after listening to a pair of 'came with Mánny' utterances which were resynthesised to have the stressed vowel to have a sharp f_o peak, a 50 ms-long plateau or a 100 ms-long plateau. The plateau duration was fully crossed within the utterance pair. The results showed that listeners judged the peak with either a 50 ms or 100 ms-long plateau to be higher than a sharp peak. No significant difference in listeners' responses was found between the 50 ms and 100 ms plateau conditions when each of them was paired with a sharp peak in a trial. When the 50 ms and 100 ms plateaux were paired, however, the 100 ms had a slight advantage compared to the 50 ms plateau. Knight (2008) concluded that the 50 ms plateau was sufficient to trigger the saliency-enhancing effect and the 100 ms plateau did not have a strong additive effect. This finding was interpreted as evidence for listeners' stability-sensitive weighting (Gockel et al., 2001); the pitch perception was more strongly affected by the steady portions in the signal than where the frequency was changing rapidly and the 50 ms plateau was sufficient to trigger phase-locking of the pitch perception mechanisms. Meanwhile, House (1990,1996) proposed that listeners track f_o movements as dynamic events only when enough cognitive resources are available, and when there is a spectrally stable section in the speech signal of at least 100 ms, as typically provided by a vowel and a following sonorant consonant. Similarly, Barnes et al. (2012a,b, 2014) suggested that the plateau coinciding with high sonority segments are judged higher than those which partially overlap with less sonorous consonants.

In Knight (2008), Barnes et al. (2012a,b) and Barnes et al. (2014), the saliency-enhancing effect of the f_o plateau was investigated only for peaks. Meanwhile, Jeon and Heinrich (2022a) showed that the plateau associated with a valley had a reduced effect. When the second peak formed a 100 ms plateau with a higher f_o than the first peak, listeners' responses were at ceiling, judging the second peak to be higher than the first. On the contrary, the 'ceiling effect' was not found for valleys; the valley forming a 100 ms plateau did not achieve a comparable saliency-enhancing effect as what was found for peaks. While this finding suggests that the perceptual weight of a plateau is context-sensitive, it is still not clear how the durational threshold of the plateau triggering the saliency-enhancing effect is determined. For instance, Jeon and Heinrich (2022b) provided evidence that even a small contrast between 25 ms and 50 ms plateaux associated with peaks was perceived by young listeners with normal hearing. Returning to Knight (2008) reporting no significant difference in the perception of 50 ms and 100 ms plateaux, one reason could be that each of the 50 ms and 100 ms plateaux was pitted against a sharp peak. A sharp peak might not allow sufficient time for listeners to detect a stable pitch target, failing to serve as a robust standard for a comparison. Note that 50 ms and 100 ms plateaux were in fact discriminated from each other when they were paired in the same trial. Furthermore, as only six listeners were tested, the results are not conclusive.

1.3. Declination

One limitation of Jeon and Heinrich's (2022a,b) studies was that all speech stimuli had a flat f_o baseline which is unlikely to be found in natural speech. Declination, i.e. the downward trend of f_o in utterances (Cohen and 't Hart, 1965, Cohen and 't Hart, 1967), is observed across languages (Ladd, 2008, Chapter 2, and references therein) and it is often visible in the f_o track. The declination implies that the f_o excursion associated with a peak is larger in the earlier than the latter part of an utterance (e.g., Cohen et al., 1982; Terken, 1991). Its well-known perceptual consequence is that when listeners judge the relative height or prominence between two consecutive peaks, the second peak is usually acoustically lower than the first in f_o when they are perceptually equivalent. This is probably because listeners perceptually compensate for their expectation of declination (e.g. Gussenhoven and Rietveld, 1988; Ladd et al., 1994; Pierrehumbert, 1979; Repp et al., 1993; Terken, 1991).

However, in some studies, the effect of declination on perception was smaller than hypothesised (Gussenhoven et al., 1997; Repp et al., 1993), and the physical presence of or listeners' expectations on declination seems to have a complex effect. For instance, Gussenhoven and Rietveld (1988) reported a positive correlation between the f_o of the first peak and that of the second peak when listeners rated the degree of perceived prominence, referred to as the "Gussenhoven-Rietveld effect" in Ladd et al. (1994). That is, the higher first peak creates a stronger expectation of declination, and consequently, increases the perceived prominence of the second peak, as listeners overestimate the second peak's prominence. However, Ladd et al. (1994) replicated the Gussenhoven-Rietveld effect only when the second peak height did not exceed a certain level (145 Hz for male voice). When the second peak was higher than 145 Hz, surprisingly, a reduction of the first peak led to an increase in the perceived prominence of the second. Ladd et al. (1994) concluded that the reference line for judging the peak height or prominence is not calculated directly from the f_o, but listeners make categorical, not continuously variable, judgement on prominence when the peaks are within a normal, non-emphatic pitch range. That is, listeners glean the overall degree of emphasis from the pitch range and assess relative prominence between the accents in a few categories. While the above discussion suggests that listeners execute some abstraction process for judging prominence, the acoustic properties of the signal shape listener's expectations and responses in some way. For instance, Gussenhoven et al. (1997) showed that the f_o height of the utterance onset affected listeners' prominence judgement of a following peak when the onset portion was longer than 400 ms. That is, listeners may establish their reference for prominence judgement based on the acoustic information when they can.

The present study questions how acoustically implemented overt f_o declination affects the perception of pitch peaks and valleys. In fact, there has been little investigation on declination of low accents and for questions which are used as stimuli in the present study. In questions, declination may be suspended or the f_o contour may uplift (e.g., Thorsen, 1980 in Danish, see Vaissiere, 1983 for a review). In any event, it is reasonable to expect the excursion size in consecutive f_o valleys to decrease as the utterance unfolds in time unless a latter valley is associated with a word under emphasis. As no systematic descriptions of declination for low accents in British English questions seems to be available, an example of declination in Belfast English where the low accents are common is presented in Fig. 2.

1.4. Pitch height vs prominence judgement

In previous studies using a forced-choice identification paradigm, participants judged which peak was more prominent (Gussenhoven et al., 1997; Repp et al., 1993) or higher (Pierrehumbert, 1979; Jeon and Heinrich, 2022a) than the other. Some studies assumed that the results would not be affected by whether listeners drew their attention to the

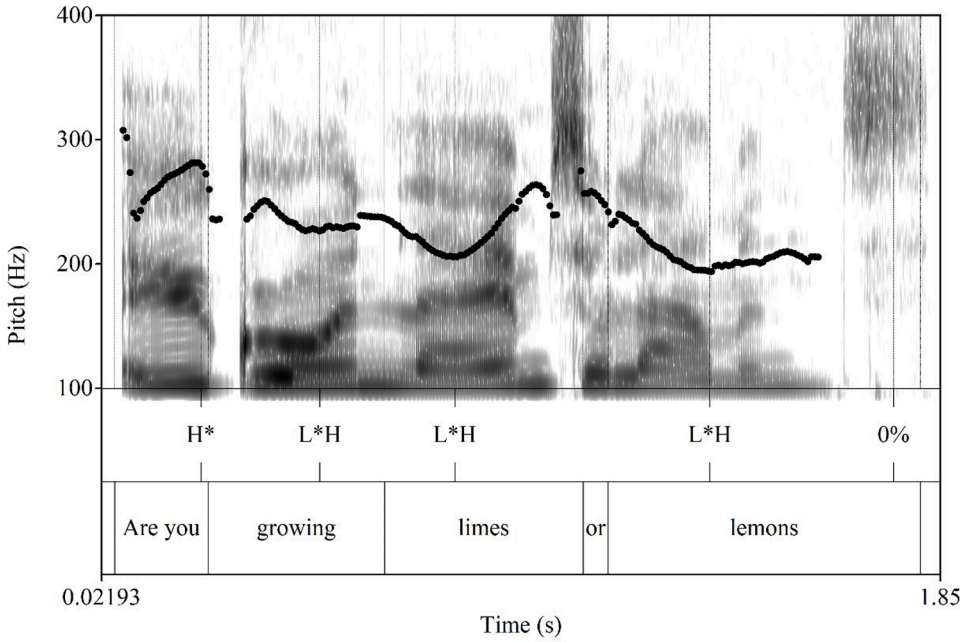


Fig. 2. An utterance with consecutive L*+H accents in Belfast English from the IViE corpus (see Nolan and Post, 2013). The L*+H accent is annotated at the minimum f_o for each accent and there is a gradual lowering. The declination slope between the first and the third accents is −38.58 Hz/sec (= −2.65 ST/sec).

pitch or prominence (Knight, 2008; Pierrehumbert, 1979). For example, when half of the listeners were instructed to identify the more prominent peak and the other half the higher peak in Knight (2008), the results were similar between the two groups. However, Jeon and Heinrich (2022b) showed that pitch height and prominence are two different perceptual dimensions. Listeners in Jeon and Heinrich's (2022b) study were native British English speakers in two age groups, younger (18–30 years old) and older (65+ years old). Half of listeners judged relative pitch height ('height group') and the other half judged prominence ('prominence group'). The task effect was observed for both age groups. First, the 'prominence group' required larger f_o changes to perceive changes in pitch compared to the 'height group'. It seemed that listeners were biased towards perceiving the first peak or valley as prominent, and probably they required a substantial f_o movement of the second extremum or additional cues in duration or loudness to match its prominence to the first. Second, although the listeners in the 'height group' could make a relative judgement on which valley was lower, the 'prominence group' were reluctant to associate an f_o valley with prominence.

Importantly, different instructions change the nature of the task (Sutcliffe, 1972). The pitch height judgement constitutes a psycho-acoustic task probably allowing listeners' cognitive capacity to directly track a single melodic dimension, whilst the prominence judgement would prompt listeners' subjective interpretations, encouraging them to correct their auditory or phonetic decision (Studdert-Kennedy and Hadding, 1973; Terken, 1991). For judging prominence, listeners readily incorporate other acoustic cues, such as duration and intensity, their phonological knowledge, and top-down information such as informativeness and predictability (e.g. Baumann and Winter, 2018; Bishop et al., 2020; Cole et al., 2010; van Heuven and Turk, 2020; Kochanski et al., 2005; Turnbull et al., 2017).

2. Study design and predictions

The experiment investigates how the extrema type (peaks vs valleys), the experimental task (judging height vs prominence), the presence of overt f_o baseline declination (yes vs no), the f_o plateau duration (25 ms, 50 ms, 100 ms), the grouping of the durational difference between the plateaux ('contrast group': 25 ms vs 50 ms, 25 ms vs 100 ms), and the f_o height difference between the two extrema (varied in five steps) affect listeners' relative judgement of two consecutive f_o peaks or valleys. It is expected that listeners' reduced ability to perceive f_o valleys will have an overarching impact on perceiving and interpreting other acoustic information. In all stimuli, the first extremum formed a 25 ms plateau. The second extremum's plateau duration varied between 25 ms and 50 ms for the 'short' contrast condition and it varied between 25 ms and 100 ms for the 'long' contrast condition (Fig. 3). This design will allow us to

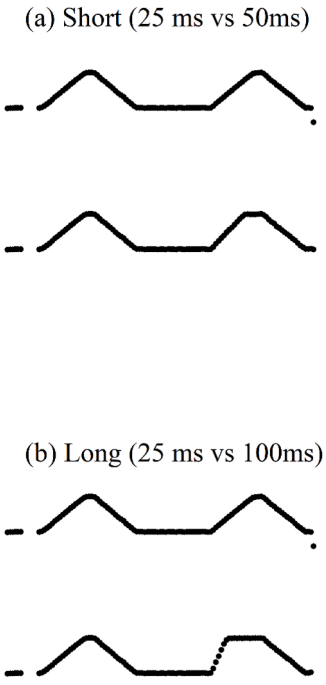


Fig. 3. Example f_o tracks for the Contrast Group conditions. The second plateau duration varied between 25 ms and 50 ms for 'Short' and between 25 ms and 100 ms for 'Long'.

examine whether 50 ms and 100 ms plateaux are perceived differently when presented with a 25 ms plateau respectively, and also conversely, whether a 25 ms plateau is perceived differently depending on whether it is presented with a 50 ms or 100 ms plateau.

For manipulating f_o in the listening stimuli, the ERB_N number scale was used. It is equivalent rectangular bandwidth rates (Patterson, 1976) modified concerning psychoacoustic data (Glasberg and Moore, 1990; also see Moore, 2012, Section 3.3). For denoting the EBR_N number, the unit 'Cam' is used for brevity (Hartmann, 1997). Using the EBR_N number scale, instead of the semitone scale used in Jeon and Heinrich (2022a), can make the perceived degrees of pitch change comparable between peaks and valleys. Declination was implemented as a gradual lowering of the straight baseline. Experimental participants were randomly allocated to two task groups; one group was asked to judge the relative pitch height between the two peaks or valleys, and the other group judged prominence.

Jeon and Heinrich (2022a,b) analysed listeners' responses as a categorical outcome (i.e. whether the first or second extremum was judged more prominent). Jeon and Heinrich (2022b) also calibrated the Point of Subjective Equality (PSE) as a measure of the f_o difference between the two extrema at which they are perceptually equal. However, due to a wide variation in listeners' responses for valleys, the PSE could not be reliably estimated, and it was inappropriate for statistical analyses. Therefore, the present study used only the responses as a categorical variable for statistical modelling. The following predictions are based on the previous findings summarised in Section 1.

(1) We expect to replicate the perceptual asymmetry between f_o peaks and valleys; listeners will show reduced discrimination for valleys compared to peaks.
(2) An interaction between the extrema type and plateau duration effects is expected; while a long f_o plateau will have a saliency-enhancing effect, the effect is expected to be reduced for valleys relative to peaks.
(3) Predicting the 'contrast group' effect is not straightforward, because previous studies did not simultaneously test the effects of the plateau duration and the pairing of different durations. If listeners solely use absolute durational information in judging pitch height or prominence, then the 'contrast group' will not have a significant effect. Alternatively, the way that plateaux of different durations are paired may affect how listeners treat the 25 ms plateau which is present in both 'contrast group' conditions. For instance, when the second plateau duration varies between 25 ms and 100 ms ('Long', Fig. 3), the presence of the long 100 ms plateaux may lead to an overall higher probability of the second extremum being judged as more salient (e.g. by leading listeners choose the 'second' for ambiguous stimuli) compared to when the second plateau duration varies between 25 ms and 50 ms with a small contrast ('Short', Fig. 3). The 'contrast group' effect may also be shown for how listeners treat the 25 ms second plateaux. The presence of the long 100 ms second plateaux may make the 25 ms plateaux in the same position sound relatively less salient, leading to a decrease of the 'second' responses when both the first and the second plateaux in the stimulus are 25 ms long. In this case, when responses only for the 25 ms second plateau conditions are compared, the 'second' response probability for the 'Long' contrast group will be lower than that for the 'Short' contrast group.
(4) There may be a three-way interaction between the extrema type, the experimental task, and declination. If the different tasks (judging pitch height vs prominence) lead listeners to pay attention to different aspects of the f_o movements, there will be an interaction between the task type and the declination effects. In the experimental stimuli, the peak or valley height remains constant while acoustically overt f_o declination is either present or absent. If the 'height' task leads listeners to compare the maxima of the peaks and the minima of the valley, the presence or absence of declination will not affect their judgement. On the other hand, if listeners need a larger acoustic difference in the f_o excursion size for prominence than height perception, the presence of declination may hinder listeners' prominence perception, particularly for valleys; the declination decreases the f_o excursion size for the second valley and may make it sound 'less deep' (see Fig. 1). Therefore, for valleys, the presence of declination will lead to a decrease in the 'second' response probability for the prominence task. Conversely, for peaks, the baseline declination increases the f_o excursion size of the second peak, potentially leading listeners to overestimate its prominence and increasing the 'second' response probability.

3. Experiment

3.1. Participants

Eighty-two participants were recruited on Prolific (www.prolific.co). They were native speakers of British English, living in England. They were under the age of 30 (average = 29.96 years, range 18–30 years) and reported no impairment in hearing or vision. Only those with no professional music training (e.g. without a degree in music) were recruited. The analysis is based on the data from 81 participants (51 female, 29 male, 1 non-binary). Data from one participant who constantly chose the same response throughout one experimental block were removed before the statistical analysis.

3.2. Experimental stimuli

The stimuli were based on the English sentence 'does Néllie know Lénny?', designed to have two disyllabic names with initial stress. The sentence consisted mainly of sonorants to keep f_o perturbations at minimum. The following factors were crossed in the stimulus design: Extrema (Peaks, Valleys) × Declination (Yes, No) × Second Plateau Duration (25, 50, 100 ms) × Height Difference (−0.5, −0.25, 0, 0.25 and 0.5 Cams).

A female native speaker of Standard Southern British English in her 20s read the English sentence several times at a comfortable speaking rate with either peaks or valleys, and also monotonously. A Sennheiser MKH40 cardioid microphone (Wedemark, Germany) and a MixPre-6 digital recorder (Sound Devices, Reedsburg, USA) were used to record the speech at a sampling rate of 44.1 kHz. Recording took place in a sound-attenuated booth in the Phonetics Laboratory at the University of Cambridge. One monotonously spoken utterance (duration 0.89 s) was selected as the base for resynthesis.

Praat ver. 6.1.16 (Boersma and Weenink, 2020) was used for resynthesizing the experimental stimuli. The built-in 'manipulation' function was used for the base sound file as the manipulation object. New pitch tiers were created with the target f_o values (Table 1) and temporal markers as described below. Then the manipulation object's pitch tier was replaced with a newly created pitch tier. The 'publish resynthesis' function using the PSOLA algorithm (Moulines and Charpentier, 1990) was used to generate all experimental stimuli. Half of the stimuli had two f_o peaks and the other half valleys (Fig. 1). The first extremum always formed a 25 ms plateau but the plateau duration of the second extremum varied to be 25 ms, 50 ms or 100 ms long. In the base utterance, time points were identified for f_o stylisation to mark the accented vowels and turning points (see Jeon and Heinrich, 2002a for details). The f_o rise (e.g. the distance between (1) and (2) and that between (5) and (6) in Fig. 1) or fall time (e.g. the distance between (3) and (4) and that between (7) and (8) in Fig. 1) associated with each extremum with a 25 ms plateau was controlled at 120 ms. The f_o plateau was aligned to the end of the stressed vowel.

The flat baseline for Peaks was at 200 Hz, for Valleys at 260 Hz (see Fig. 1). The f_o contours of the valley stimuli were mirror images of the

Table 1
'Difference' refers to the difference between the first and the second extrema. The negative Difference values indicate that the second extremum had a smaller f_o excursion size from the flat baseline than the first. 'Height' for the second extremum was measured from the 200 Hz flat baseline for Peaks and from the 260 Hz flat baseline for Valleys.

Difference (Cams)	Height		
	Cams	Hertz	Semitones
Peaks			
−0.5	6.55	234	14.75
−0.25	6.80	247	15.66
0	7.05	260	16.54
0.25	7.30	273	17.41
0.5	7.55	287	18.25
Valleys			
−0.5	6.34	224	13.94
−0.25	6.09	212	12.98
0	5.84	200	12.00
0.25	5.59	189	10.99
0.5	5.34	178	9.94

peak counterparts. The first extremum was always at the height of 0 Height Difference (260 Hz for Peaks, 200 Hz for Valleys, Table 1). The second extremum varied in five 0.25 Cam steps in its height. The ERB_N number (Cam) was converted from Hertz using the formula in Moore (2012, p. 76, ERB_N number $= 21.4 \times \log_{10}(4.37 \times Hz/1000 + 1)$). When expressed in Hertz, the difference between the consecutive steps was 13–14 Hz for Peaks and 11–12 Hz for Valleys.

The declination slope −1.33 Cams/sec was implemented for the Declination–Yes condition. We referred to our recordings to determine an appropriate slope, but the slope varied widely across utterances between −28 and −88 Hz (−1.07 and −3.20 Cams) per second. In the experimental stimuli, it had to be ensured that the slope of the f_o fall associated with a valley was larger than that of declination; if not, the f_o valleys would not be created. For the Peak stimuli in the Declination–Yes condition, the utterance initial f_o was at 200 Hz and the final f_o at 170 Hz. For Valleys, the utterance initial f_o was at 260 Hz and final *fo* at 226 Hz. For both Extrema, the difference between the utterance initial and final f_o was 0.67 Cams. (Applying the IPO model for the declination slope $(D = (-11)/(t + 1.5$, where D is the slope in semitones and t is time in seconds, ''t Hart et al., 1990) caused a problem for Valleys. The steep declination caused the f_o contour associated with the second valley to move upward from the declining baseline.)

3.3. Experimental procedure

The Gorilla Experiment Builder (www.gorilla.sc) was used to run the experiment (Anwyl-Irvine et al., 2019). Data were collected between 22 July and 25 August 2020. The study was approved by the Business, Arts, Humanities and Social Science Ethics committee of the University of Central Lancashire (BAHSS2 0122).

All participants used a desktop computer and headphones. Before the experiment, they filled in questionnaires on their variety of UK English, gender, age, musical training and experience, and a consent form. Listeners' English variety and musical experience (e.g. how often they listen to music and for how long they have taken music lessons) varied, but none was professionally trained in music. Seven participants indicated that they had absolute pitch.

Participants were asked to wear headphones and adjust the volume to a comfortable level while a 1 kHz pure tone was played at 70 dB for 10 seconds. They took a headphone screening task (Woods et al., 2017) with 12 trials. Participants who were correct for fewer than 10 trials could not proceed to the experiment. Sixteen catch trials were devised to ensure that participants were paying attention to the tasks. They were simple mathematical operations with the correct answer either 1 or 2 (e.g., 4 − 3 = ?) which were visually presented on the screen.

Participants were randomly assigned to one of four groups created as a result of an orthogonal combination between the two tasks (Task–Height and Prominence) and two groups for differentiating the second extremum plateau duration (Contrast Group–Short, 25 ms vs 50 ms and Long, 25 ms vs 100 ms). For each listener, the main experiment consisted of four blocks (2 Extrema × 2 Declination). In each block, there were 10 stimuli (2 Plateau Duration × 5 Height Difference). Each stimulus was presented three times in the block. The presentation of stimuli including four catch trials was randomised for each listener. The order of blocks was counterbalanced in four lists. Two lists (lists A and B) started with Extrema–Peaks, and then the order of Declination was counterbalanced (i.e. the order of conditions for list A: (1) (Extrema–) Peaks-(Declination–)Yes, (2) Peaks–No, (3) Valleys–Yes, and (4) Valleys–No, list B: (1) Peaks–No, (2) Peaks-Yes, (3) Valleys–No, (4) Valleys–Yes). The other two lists (lists C and D) started with Extrema–Valleys.

The written instructions informed the participants that they would hear 'does Nellie know Lenny?' with two high 'peak' accents or two low 'valley' accents in the first vowel in each name, NEllie and LEnny. On the screen, there were two buttons, 'Peaks' and 'Valleys', which a participant could press to listen to a practice stimulus. There was a practice session before the first block and after the second block. The practice session consisted of 8 trials (2 Plateau × 2 Declination × 2 Height Difference [-0.75 Cams, +0.75 Cams]) with the stimulus presentation order randomised for each participant. Participants in the Height task group were instructed to focus their attention on accent height and to identify which accent sounded higher for Peaks or lower for Valleys. Participants in the Prominence group were instructed to identify which accent sounded more prominent, standing out or emphatic. Participants were not informed about the order of blocks in advance.

The stimulus was automatically played 0.5 s after the onset of each trial. Participants could repeat the stimulus presentation up to 20 times. In each trial, a question appeared on the top of the screen. Participants in the Height group were asked: 'which one sounds higher?' for the stimuli with peaks or 'which one sounds lower?' for the stimuli with valleys. Participants in the Prominence group were asked 'which one sounds more prominent?'. Two buttons on the bottom labelled as 'NEllie' and 'LEnny' gave the response options. Participant indicated their choice by clicking the appropriate button with a mouse. The experiment automatically progressed to the next trial when participants pressed a response button. No feedback was provided in the practice session or the main experiment.

3.4. Statistical analysis

Seven participants declared to have an absolute pitch, but their data were included in the analysis. When each participant's response functions were examined, the response functions of one participant with absolute pitch showed the canonical S-shape indicating high accuracy for pitch perception (see Klein, 2001). The other six participants' response functions did not notably deviate from those without absolute pitch. There was one error for the catch trials by one listener, but no one was excluded in the analyses based on the catch trial results.

All statistical analyses were conducted with R Version 4.3.1 (R Core Team, 2023) and R Studio 2022.07.1 + 554 (R Studio Team, 2022). We used the package *tidyverse* Version 2.0.0 (Wickham et al., 2019) for data processing and the package *brms* Version 2.21.0 (Bürkner, 2017) for Bayesian logistic regression modelling. All data and analysis codes are available under the OSF repository (https://osf.io/akwez/, 10.17605/OSF.IO/AKWEZ). The categorical predictors were sum-coded to facilitate the interpretation of the model as indicated in the squared brackets below. The predictors were Extrema (Peaks [1], Valleys [−1]), Contrast Group (Short: 25 ms vs 50 ms [−1], Long: 25 ms vs 100 ms [1]), Task (Height [1], Prominence [−1]), Declination (No [1], Yes [−1]), the second Plateau duration (first contrast: 25 ms [1], 50 ms [0], and 100 ms [−1]; second contrast: 25 ms [0], 50 ms [1], and 100 ms [−1]), and Height Difference (five Cam steps).

The logistic models estimated the maximum likelihood of the 'second (Lenny)' response. The Height Difference effect indicates the change in listeners' 'second' responses along the five steps. While we aimed to construct a model without an overcomplicated and uninterpretable structure, we incorporated the interaction terms involving Extrema to assess its potential overarching effect, Contrast Group × Task × Extrema, Task × Declination × Extrema, and Declination × Extrema × Plateau. Listener was incorporated as a random slope for Extrema because each listener's sensitivity was expected to differ between Peaks and Valleys (Jeon and Heinrich, 2022a).

Default priors were used for the intercept. The weakly informative priors with normal distributions centred at zero ($SD = 0.5$) were used for the coefficients. Hamiltonian Markov Chain Monte Carlo (MCMC) sampling was conducted with four chains and 10,000 iterations (2000 of which were warm-up), resulting in a total of 32,000 posterior samples used for inference. There was no indication of convergence issues (all Rhat values = 1.00).

4. Results

Overall, 51 % of the listeners' responses are the 'second' ('Lenny'); this shows that listeners were not inherently biased towards one response. The 'second' response rates varying across experimental conditions (Table 2) indicate that the experimental manipulation affected listeners' responses. For instance, the lowest 'second' response rate (23 %) is found for the condition combining Contrast Group–Long, Task–Prominence, Extrema–Peaks, Declination–No, Plateau–25 ms. The highest 'second' response rate (71 %) is found for the condition combining Contrast Group–Long, Task–Height, Extrema–Peaks, Declination–Yes, Plateau–100 ms. In Table 2, the 'second' response rates are always at or below 52 % within Task–Prominence, while the rates widely vary for Height.

Table 3 summarises the results of the Bayesian logistic regression modelling. A positive log odd coefficient ($\beta > 0$) indicates that the relevant predictor is associated with an increase in the 'second' responses whilst a negative parameter ($\beta < 0$) indicates a decrease. The 95 % credible interval (CI) not straddling zero is interpreted as indicating a reliable effect of the predictor concerned. The posterior probability indicates the posterior sample distribution of the estimated coefficients. For instance, if the posterior probability of the coefficient being above zero is 1 ($\Pr(\beta > 0) = 1$), this means that not a single posterior sample for this coefficient was below zero, indicating a strong positive effect. Below only the effects with strong evidence assessed based on the credible interval were discussed. Figs. 4–6 show the predicted probability of the 'second' responses as a function of Height Difference and the predictors in interactions. The response functions from the raw data are provided in the OSF repository (https://osf.io/akwez/, 10.17605/OSF.IO/AKWEZ).

The main predictors with a significant effect are discussed first. First, there was strong evidence for the Extrema effect ($\beta = 0.1$, CI [0.04, 0.17], $\Pr(\beta > 0) = 1$). Listeners were more likely to choose the 'second' response for Peaks relative to Valleys. Second, a positive coefficient for the Task ($\beta = 0.24$) was reliable with the 95% credible interval away from zero (CI[0.16, 0.31]) and high posterior probability ($\Pr(\beta > 0) = 1$); the 'second' response probability was higher for Height compared to Prominence. Third, there was strong evidence for the Declination effect

Table 3
Output of the Bayesian regression model (c_1: the first contrast, c_2: the second contrast).

	β	SE	Lower bound	Upper bound	Posterior probability
Intercept	−1.93	0.04	−2	−1.85	1.00
Extrema	0.1	0.03	0.04	0.17	1.00
Contrast Group	−0.08	0.04	−0.16	0.01	0.96
Task	0.24	0.04	0.16	0.31	1.00
Declination	−0.04	0.02	−0.07	0	0.98
Plateau_c1	−0.18	0.02	−0.23	−0.13	1.00
Plateau_c2	−0.08	0.04	−0.16	−0.01	0.99
Height Difference	1.5	0.05	1.4	1.6	1.00
Extrema × Plateau_c1	−0.11	0.02	−0.16	−0.07	1.00
Extrema × Plateau_c2	−0.07	0.04	−0.14	0	0.97
Contrast Group × Task	0.03	0.04	−0.05	0.1	0.75
Extrema × Contrast Group	−0.08	0.04	−0.15	−0.01	0.98
Task × Extrema	−0.04	0.03	−0.11	0.02	0.91
Task × Declination	−0.05	0.02	−0.08	−0.02	1.00
Declination × Extrema	0.03	0.02	−0.01	0.07	0.95
Declination × Plateau_c1	−0.01	0.02	−0.06	0.03	0.69
Declination × Plateau_c2	0.02	0.03	−0.04	0.07	0.73
Contrast Group × Task × Extrema	0.01	0.03	−0.05	0.07	0.63
Task × Declination × Extrema	0.02	0.02	−0.01	0.06	0.92
Declination × Extrema × Plateau_c1	0.02	0.02	−0.03	0.06	0.79
Declination1 × Extrema × Plateau_c2	0.02	0.03	−0.03	0.08	0.81

Table 2
The response frequency (count and percentage for the 'second' ('Lenny') responses. Data were collapsed over the Height Difference conditions.

		Contrast Group–Short			Contrast Group–Long		
Task–Height							
Extrema	Declination	Plateau	Nellie	Lenny (%)	Plateau	Nellie	Lenny (%)
Peaks	No	25 ms	162	138 (46%)	25 ms	181	119 (40%)
		50 ms	141	159 (53%)	100 ms	111	189 (63%)
	Yes	25 ms	161	139 (46%)	25 ms	179	121 (40%)
		50 ms	133	167 (56%)	100 ms	86	214 (71%)
Valleys	No	25 ms	183	117 (39%)	25 ms	176	124 (41%)
		50 ms	174	126 (42%)	100 ms	145	155 (52%)
	Yes	25 ms	135	165 (55%)	25 ms	137	163 (54%)
		50 ms	137	163 (54%)	100 ms	129	171 (57%)
Task–Prominence							
Peaks	No	25 ms	191	124 (39%)	25 ms	231	69 (23%)
		50 ms	180	135 (43%)	100 ms	150	150 (50%)
	Yes	25 ms	202	113 (35%)	25 ms	219	81 (27%)
		50 ms	201	114 (36%)	100 ms	143	157 (52%)
Valleys	No	25 ms	221	94 (30%)	25 ms	208	92 (31%)
		50 ms	213	102 (32%)	100 ms	200	100 (33%)
	Yes	25 ms	220	95 (30%)	25 ms	210	90 (30%)
		50 ms	215	100 (32%)	100 ms	197	103 (34%)

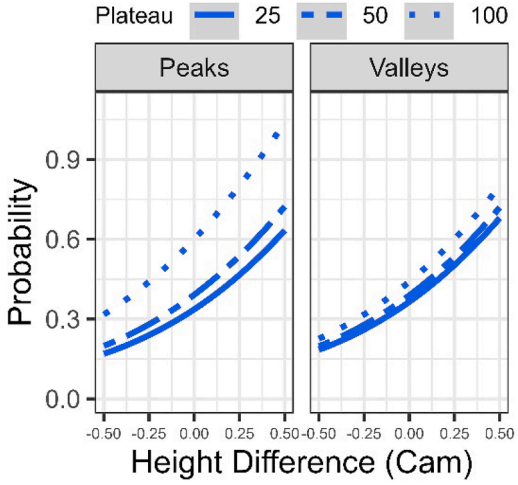


Fig. 4. Estimated 'second' response conditional probability by Extrema and Plateau Duration.

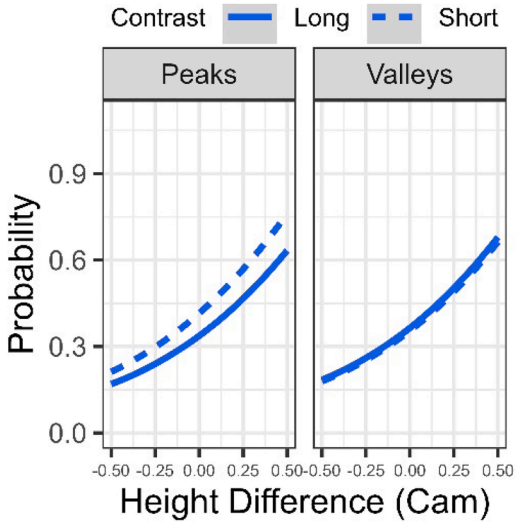


Fig. 5. Estimated 'second' response conditional probability by Extrema and Contrast Group.

($\beta = -0.04$, CI[−0.07, 0], Pr ($\beta < 0$) = 0.98), showing a lower 'second' response probability for No compared to Yes. Fourth, the Plateau effect was reliable; the negative estimate of the first contrast suggested a decrease in the 'second' response probability for the 25 ms condition relative to the 100 ms condition ($\beta = -0.18$, CI[−0.23, −0.13], Pr ($\beta < 0$) = 1); the second contrast also showed a decrease for the 50 ms condition relative to the 100 ms condition ($\beta = -0.08$ CI[−0.16, −0.01], Pr ($\beta < 0$) = 0.99). Last, there was strong evidence for the Height Difference effect; a step increase was associated with an increase of the 'second' response probability (β =1.5, CI [1.4, 1.6], Pr($\beta > 0$) = 1).

None of the three-way interactions in the model was significant (Table 3). However, importantly, all predictors (apart from Height Difference which was not incorporated in interaction terms), Contrast Group, Extrema, Task, Declination, and Plateau Duration, were involved in two-way interactions. First, there was evidence for the Extrema × Plateau interaction (Extrema × Plateau_c1, $\beta = -0.11$, CI[−0.16, −0.07], Pr($\beta < 0$) = 1; Extrema × Plateau_c2, $\beta = -0.07$, CI[−0.14, 0], Pr($\beta < 0$) = 0.97). The negative coefficients suggest that the 'second' response probability showed a larger magnitude of decrease for the 25 ms-long plateaux compared to the 50 ms-long plateaux (c1) and also for the 50 ms-long plateaux compared to the 100 ms-long plateaux (c2) for

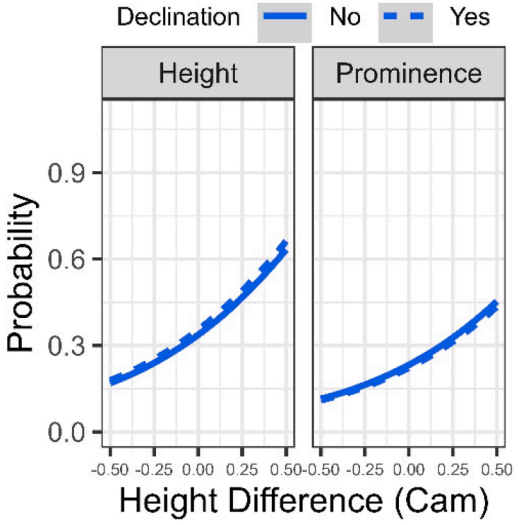


Fig. 6. Estimated 'second' response conditional probability by Task and Declination.

Peaks than for Valleys. Although, for both Peaks and Valleys, the 'second' response probability increases along the Height Difference steps for both Peaks and Valleys in Fig. 4, the slope of the increase is steeper for Peaks compared to Valleys, particularly when the second plateau formed a 100 ms plateau.

Second, the Extrema effect was dependent on the Contrast Group effect (β = -0.08, CI[-0.15, -0.01], Pr($\beta > 0$) = 0.98). The negative coefficient indicates a decrease in the 'second' response probability for the condition combining Extremum–Peaks and Contrast Group–Long. Fig. 5 shows that for Peaks, the 'second' response probability for Contrast Group–Short is higher relative to Long. On the other hand, the Contrast Group effect is not clear for Valleys. The Contrast Group effect also provides insight into how listeners perceived the 25 ms plateau depending on what it was paired with. When the 'second' response probability was examined only for the second 25 ms plateau extremum conditions, for Peaks, the 'second' response probability was higher for Short (mean = 0.42, SE = 0.04) than for Long (mean = 0.34, SE = 0.04), whereas there was no obvious difference for Valleys (Short, mean = 0.35, SE = 0.04; Long, mean = 0.36, SE = 0.04).

Third, the Task predictor interacted with Declination ($\beta = -0.05$, CI [−0.08, −0.02], Pr($\beta < 0$) = 1). The negative coefficient indicates a decrease in the 'second' response probability for the condition combining Task–Height and Declination–No. Fig. 6 shows that for Task–Height, the presence of declination increases the 'second' response probability relative to when there was no declination. For Task–Prominence, listeners' 'second' response probability is lower with a less steep slope for the probability function compared to Height; furthermore, the 'second' response probability is slightly higher without overt declination.

To summarise, the 'second' response probability increased together with an increase in the second plateau duration (Plateau–25 ms, 50 ms, and 100 ms) and also along the increase in the f_o excursion size of the second extremum from the baseline (Height Difference– −0.50, −0.25, 0, +0.25 and +0.50 Cams). The presence of overt baseline declination increased the 'second' response probability, but this effect did not interact with the Extrema (peaks vs valleys) effect as expected. Importantly, first, the Extrema effect interacted with the Plateau Duration effect (Fig. 4). The slope of the 'second' response probability function was steeper for the peaks, indicating that listeners' perception was more strongly influenced by the variation in Plateau Duration conditions (25 ms, 50 ms and 100 ms) relative to the valleys. Second, the Extrema effect was also dependent on the Contrast Group effect (Fig. 5). The Contrast Group effect was obvious only for the peaks where the 'second' response

probability was higher when the second extremum's plateau was shorter at 50 ms. In addition, for the peaks, the 'second' response probability for the 25 ms second plateau condition was higher when it was presented with the 50 ms plateaux compared to the 100 ms plateaux. Finally, the experimental task type interacted with the presence of overt baseline declination (Fig. 6). The 'second' response probability increased more steeply as a function of an increase in the second extremum's excursion size for the height judgement task compared to the prominence judgement task. The 'second' response probability was higher with declination for the height judgement, whereas the declination had an opposite effect for the prominence judgement.

5. Discussion

Based on the previous findings (Section 1), this study aimed to investigate how the perceptual asymmetry, i.e. the relative perceptual advantage of peaks compared to valleys, would affect the role of other factors in perception (see Section 2 for the study design and predictions).

To summarise the findings, first, this study replicated the previously reported perceptual peaks vs valleys asymmetry and the saliency-enhancing f_o plateau effect. One indication of the perceptual asymmetry was that the 'extrema' (peaks vs valleys) effect interacted with the second extremum's plateau duration which was manipulated to be 25 ms, 50 ms or 100 ms. The longer plateau had a more notable saliency-enhancing effect for the f_o peaks compared to the valleys. Second, the 'extrema' effect also interacted with the 'contrast group' effect (see Fig. 3); the 'contrast group' effect was shown only for the peaks, and the presence of the 100 ms-long second plateau decreased the 'second' responses. For the peaks, when the responses for stimuli with the 25 ms-long second plateau were specifically examined, listeners were more likely to choose the 'second' response when 25 ms-long plateaux were presented with 50 ms-long plateaux compared to 100 ms-long plateaux. Third, while the declining f_o baseline had an overall effect of increasing the 'second' response probability for both peaks and valleys, the experimental task type interacted with the presence of declination. The 'second' response probability was higher with declination for the height judgement, whereas the declination had the opposite effect for prominence judgement. These findings are further discussed below.

5.1. The peaks vs valleys asymmetry and durational settings

The slope of the response probability functions for valleys tended to be less steep than those for peaks (Figs. 4 and 5). That is, changes of f_o in 0.25 Cam steps had a stronger influence on listeners' judgement for the peaks compared to the valleys. The interaction effects between (i) the extrema type and plateau duration factors and between (ii) the extrema type and contrast group factors suggest that the perceptual asymmetry between f_o peaks and valleys affected listeners' interpretation of the temporal properties in the stimuli as expected (Section 2).

First, the saliency-enhancing effect of f_o plateaux was more notable for peaks than for valleys. For the peaks, the increase in the plateau duration led to an increase in the 'second' response probability; the response probability functions for the three plateau conditions (25, 50, and 100 ms) were clearly separated (Fig. 4). On the other hand, the saliency-enhancing effect of plateaux was relatively reduced for the valleys. Fig. 4 showed that, for the valleys, listeners were less likely to differentiate a 25 ms plateau from a 50 ms plateau, while the 100 ms plateau did have some saliency-enhancing effect compared to the shorter ones. This result corroborates the potential 'asymmetrical weighting' between f_o rises and falls briefly reported by Barnes et al. (2012b). While investigating the timing of the Tonal Center of Gravity (TCoG) for the L + H* (rise-fall) accent in American English, they observed that any changes occurring in the rising portion seemed to be more heavily weighted in the perception of the timing and scaling of the pitch accent compared to the analogous changes in the falling portion.

Furthermore, the present finding has an implication for the relationship between the f_o contour alignment and perception. Previous studies (e.g. Barnes et al., 2012a,b; House, 1996) suggested that the pitch information should be available over a spectrally stable interval such as a vowel for successful pitch tracking. However, the present results showed that the threshold value of the saliency-enhancing effect of an f_o plateau is not absolute or invariable. The perceptibility is affected by the f_o movement direction, and the alignment between an f_o event and spectral stability is probably a necessary but not sufficient condition for a perceptual 'sweet spot' (Barnes et al., 2012a,b, p. 353). Even when a reasonably long f_o plateau is aligned to a substantial portion of a vowel, listeners would not hear or interpret it as prominence-lending or –cueing if preceded by a falling f_o, although the role of the slope of the fall needs further perceptual validation (cf. Barnes et al., 2010a).

The perceptual asymmetry between f_o peaks and valleys also provides an account for the findings that low accentual valleys with the f_o minimum aligned to the stressed vowel are not as effective as high peaks for signalling high informational load (e.g. Baumann and Roth, 2014; Zahner and Braun, 2018; Zahner et al., 2019). For instance, Baumann and Röhr (2015) suggested a prominence hierarchy in German: $L + H^* > L^* + H > H^* > H + !H^* > H + L^* > L^*$ > no pitch accent (for their schematic representations, see Grice et al., 2019). In this hierarchy, the most prominent accent has a rising f_o in the accented syllable (L + H*). For the second most prominent L*+H, the timing of the rise is later than that of L + H*, followed by H* which has a high f_o in the accented syllable. On the other hand, the low plateau associated with the accented syllable (L*) and the low plateau following a fall (H + L*) are not perceptually as prominent. Baumann and Röhr (2015) concluded that rises and high f_o were perceived more prominent than falls and low f_o and that a steep f_o excursion increased the perceived prominence. However, as Baumann and Röhr (2015) controlled the stimuli in such a way that all pitch accents were followed by a low boundary tone (L %), the prominence hierarchy of the pitch accent types may be manifested differently in different tonal contexts, such as when the accents are followed by a different phrasal tone or high boundary tone. Therefore, it is worth further investigating the relationship between more fine-grained acoustic shapes and their interpretations by listeners as exemplified in different tonal contexts. Another domain for further investigation concerns potential cross-language differences in how the plateau duration, the direction of the f_o movement and its slope are weighted in relation to the prominence hierarchy. As different languages differ in how they code f_o movements in the linguistic structure (e.g. D'Imperio and House, 1997; House et al., 1997; Jongman et al., 2017; Jun 2005; Krishnan et al., 2005), we do not expect the prominence hierarchy to be universal.

Second, the 'contrast group' effect shows that listeners' perception of auditory objects is affected by the stimulus presentation context. In the experiment, the first plateau duration was constant at 25 ms. Listeners were, in general, more likely to choose the 'second' response when the second peak's plateau duration varied between 25 ms and 50 ms compared to when it varied between 25 ms and 100 ms (Fig. 5). For the latter case, probably listeners could make use of the reliable durational information of the 100 ms plateaux to judge the second as higher or more prominent. On the other hand, listeners who listened to the 25 ms vs 50 ms contrast dealt with a smaller difference introducing a higher level of ambiguity. As the durational cue was not strong, probably they were swayed to choose the 'second' even when the first and the second peaks had the same plateau duration at 25 ms. For valleys, the contrast group effect was not shown; the 'second' response probability was overall low, indicating that listeners were biased to choose the 'first'. It seems that the valley stimuli were too hard to perceive to introduce ambiguity which can interfere with the decision process.

The present study cannot determine whether the contextual effect is an outcome of listeners' low-level hearing or high-level decision-making (cf. Davis and Johnsrude, 2007). Nonetheless, the results show that exposure to long plateau-shaped f_o peaks can shift listeners' perceptual threshold or meta-correction strategies. The context effect was not observed when listeners faced perceptual challenges for the valleys and

established a bias.

5.2. The effect of baseline declination for judging pitch height vs prominence

Different experimental tasks seemed to have drawn listeners' attention to different dimensions of the speech signal. For the height judgement task, the 'second' response probability was higher compared to the prominence judgement task, and listeners' discrimination was relatively heightened. On the other hand, listeners showed a stronger bias towards the 'first' response when judging prominence as reported by Jeon and Heinrich (2022b). One possible reason for the bias is that in the auditory stimulus 'does Néllie know Lénny?', 'Néllie' was never deaccented, i.e., it was always realised with a normal f_o excursion size in the speaker's range without any other kinds of phonetic reduction. The first f_o peak or valley always had an excursion size of 60 Hz from the flat baseline. As listeners perceived the first peak or valley as reasonably prominent, they might have required strong acoustic cues associated with the second one to judge it as more prominent to override the already-perceived prominence of the first.

The present finding concerning the interaction effect between declination and the experimental task has broader implications on the question of how to represent intonation. Researchers so far have taken two broadly different approaches, decomposing intonation into either a sequence of static low and high tonal targets or dynamic movements of rises and falls (for a review, see 't Hart et al., 1990; Ladd, 2008, Chapter 3; Nolan, 2022). Whether one takes the target-based or movement-based approach has a consequence in mapping acoustic correlates with the perceived prominence. Taking the target-based approach, for instance, an f_o peak would gain prominence because of its high f_o. On the other hand, based on the movement-based approach, the peak's prominence has to do with the rise. The target-based approach formed a basis for widely used transcription systems such as the Tones and Break Indices (ToBI, cf. Beckman et al., 2005) and for intonational phonology (cf. Gussenhoven, 2004; Ladd, 2008). However, the fact that we have established an analytic framework does not mean that we understand the perceptual process (Barnes et al., 2012a,b, p.340); recent studies emphasise the necessity to incorporate the dynamic f_o movement in modelling intonation on the perceptual grounds (e.g. Barnes et al. 2010a,b, 2012a,b, 2021; D'Imperio, 2000; Niebuhr et al., 2020).

The overtly declining baseline used in this study increased the f_o excursion size for the second peak but decreased it for the second valley (see Fig. 1). Therefore, if listeners relied on the excursion size as a cue, the declination would have given some advantage for the second peak, leading listeners to overestimate the second peak height or prominence. On the other hand, for the valleys, the declination was expected to hinder listeners' judgement by reducing the excursion size, making it sound 'less deep' and increasing their 'first' responses. However, contrary to the expectation, there was no statistical evidence for the declination effect interacting with the extrema type effect.

Nevertheless, there was an interaction effect between the task type and declination; the declining f_o baseline had an effect of increasing the 'second' responses for judging pitch height but decreasing them for judging prominence. That is, the declining f_o baseline made the second peak sound higher but less prominent, and it made the second valley sound lower and less prominent (see Fig. 5).

This result is interpreted that a large falling f_o movement is not a requirement for creating a sensation of 'lowness'. For the 25 ms or 50 ms valley plateaux posing perceptual challenges, it would have been greatly difficult for listeners to track either the f_o movement or the f_o minimum associated with the valley. Then the generally declining f_o contour from the utterance onset probably served as a context for the perceptual target moving downwards in pitch. Consequently, listeners may have relied on the global impression of the f_o movement to carry out the height judgement task for the valleys. On the contrary, when listeners judged relative prominence, the declining baseline decreased the second extremum's perceived prominence. This result is intuitively explicable for the valleys; the declining baseline had an effect of decreasing the f_o excursion size and the mass of the 'area over the curve' for the second valley (see Fig. 1), decreasing its overall acoustic energy. On the other hand, for the peaks, the result suggests that listeners relied neither on the increased excursion size nor the mass of the 'area under the curve' for judging prominence. As discussed above for judging the relative height between the valleys, listeners probably relied on the global impression, perceiving their perceptual target moving downward in the context of the overall f_o downtrend.

What listeners did for judging the relative height between the peaks, tracking the f_o excursion size, might be an unusual strategy which is not always relevant to interpreting speech which is characterised by complex and dynamic intonational variation. For carrying out the height judgement task with the peaks, listeners were provided with the experimental setting facilitating their pitch-tracking with a psychoacoustic task and the trackable stimuli. On the other hand, for carrying out a more linguistic 'prominence' task or for dealing with a perceptual challenge of listening to the valleys, listeners seem to have been strongly affected by the global pitch context surrounding their perceptual target.

The present results highlight the importance of the overall f_o contour over the time-bound f_o target in perceiving intonational prominence. This conclusion may seem to contradict the previous findings arguing for the intonational target's importance (e.g. see Ladd et al., 1994). The discrepancy may be attributable to the differences in the experimental setup. In previous studies, such as Gussenhoven et al. (1997) and Terken (1991), listeners engaged in active tasks such as adjusting the f_o to match the degree of prominence between two peaks or rating prominence using a scale. By contrast, the present study tested naïve listeners' intuitive judgement using forced-choice tasks. The present experimental setup may have discouraged listeners from finely scaling the degree of prominence, while no non-phonetic information such as the informational context was provided (cf. Cole and Shattuck-Hufnagel, 2016). The intuitive judgement task used in the present study is perhaps closer to language users' daily experience of rapidly identifying relative prominence between words in short utterances compared to the tasks requiring listeners' full attention to pitch as a single prosodic dimension or a single intonational peak in an utterance.

6. Conclusions

The experimental results showed the perceptual advantage of f_o peaks compared to valleys. Listeners showed enhanced sensitivity to the f_o change for the peaks and the saliency-enhancing effect of a long f_o plateau was stronger for the peaks. These findings support the association between high rising pitch and high-level attention (e.g. Evans, 2015; Gussenhoven, 2002; Hsu et al., 2015) and that pitch accents with a steep rise and high pitch are perceived more prominent than those with a fall and low pitch (Baumann and Röhr, 2015). Furthermore, the presence of f_o peaks associated with a 100 ms-long plateau seemed to be perceptually salient and affected listeners' overall response rate for other stimuli. The findings suggest that the absolute duration of an f_o plateau associated with an accent would not be a reliable acoustic correlate for its prominence. Importantly, perceived prominence did not seem to be directly derived from a handful of independent acoustic dimensions. Neither f_o nor durational variation, even when it was above the perceptual threshold as shown in the results of the height judgement, seemed to be a reliable cue that listeners use for judging prominence. Unless listeners were carrying out an explicit psychoacoustic task judging the relative height between two f_o peaks, listeners seemed to have based their prominence judgement on the overall impression about pitch, incorporating where the perceptual target was located relative to the preceding target. Therefore, for perceptually modelling speech intonation, we will need to seek a holistic approach than solely analysing local f_o measures such as an accent's f_o height or excursion size.

CRediT authorship contribution statement

Hae-Sung Jeon: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Visualization, 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.

Acknowledgements

This research was supported by the Department for Business, Energy and Industrial Strategy, UK (the British Academy\Leverhulme Small Research Grant, SRG19\190109).

Data availability

Data will be made available on request.

References

Anwyl-Irvine, A.L., Massonnié, J., Flitton, A., Kirkham, N., Evershed, J.K., 2019. Gorilla in our midst: an online behavioral experiment builder. Behav. Res. Methods 52 (1), 388–407. https://doi.org/10.3758/s13428-019-01237-x.

Barnes, J., Brugos, A., Shattuck-Hufnagel, S., Veilleux, N., 2012a. On the nature of perceptual differences between accentual peaks and plateaux. In: Niebuhr, O. (Ed.), Prosodies: Context, Function, Communication. de Gruyter, pp. 93–118.

Barnes, J., Brugos, A., Veilleux, N., Hufnagel, S., 2014. Segmental influences on the perception of pitch accent scaling in English. In: 7th International Conference on Speech Prosody 2014. ISCA, pp. 1125–1129.

Barnes, J., Brugos, A., Veilleux, N., Shattuck-Hufnagel, S., 2021. On (and off) ramps in intonational phonology: rises, falls, and the Tonal Center of gravity. J. Phon. 85, 101020. https://doi.org/10.1016/j.wocn.2020.101020.

Barnes, J., Lee, C.K.C., Brugos, A., Shattuck-Hufnagel, S., Veilleux, N., 2023. Sometimes low really is just the opposite of high: perception of low F0 targets by tone and non-tone language speakers. Proceedings of the 20th ICPhS. Guarant International, Prague, Czech Republic, pp. 1295–1299.

Barnes, J., Veilleux, N., Brugos, A., Shattuck-Hufnagel, S., 2010a. Turning points, tonal targets, and the english l- phrase accent. Lang. Cogn. Process. 25 (7–9), 982–1023. https://doi.org/10.1080/01690961003599954.

Barnes, J., Veilleux, N., Brugos, A., Shattuck-Hufnagel, S., 2010b. The effect of global F0 contour shape on the perception of tonal timing contrasts in American english intonation. Speech Prosody 1–4, 10–14, 2010, 100445ChicagoMay.

Barnes, J., Veilleux, N., Brugos, A., Shattuck-Hufnagel, S., 2012b. Tonal Center of Gravity: a global approach to tonal implementation in a level-based intonational phonology. Lab. Phonol. 3 (2). https://doi.org/10.1515/lp-2012-0017.

Baumann, S., Röhr, C.T., 2015. The perceptual prominence of pitch accent types in German. In: The Proceedings of the ICPhS2015. Glasgow, UK, 0298, pp. 10–14. August.

Baumann, S., Roth, A., 2014. Prominence and coreference on the perceptual relevance of F0 movement, duration and intensity. In: Proc. 7th International Conference on Speech Prosody 2014, pp. 227–231. https://doi.org/10.21437/SpeechProsody.2014-33. Retrieved from.

Baumann, S., Winter, B., 2018. What makes a word prominent? Predicting untrained German listeners' perceptual judgments. J. Phon. 70, 20–38. https://doi.org/10.1016/j.wocn.2018.05.004.

Beckman, M.E., Hirschberg, J., Shattuck-Hufnagel, S., 2005. The original ToBI system and the evolution of the ToBI Framework. In: Jun, S.-A. (Ed.), Prosodic Typology. Oxford University Press, pp. 9–54. https://doi.org/10.1093/acprof:oso/9780199249633.003.0002.

Bishop, J., Kuo, G., Kim, B., 2020. Phonology, phonetics, and signal-extrinsic factors in the perception of prosodic prominence: evidence from Rapid Prosody Transcription. J. Phon. 82, 100977. https://doi.org/10.1016/j.wocn.2020.100977.

Boersma, P. & Weenink, D. (2020). Praat: doing phonetics by computer [Computer program]. available at http://www.praat.org/.

Bolinger, D., 1972. Accent is predictable (if you're a mind-reader). Language 48 (3), 633–644. https://doi.org/10.2307/412039.

Bürkner, P.C., 2017. brms: an R package for bayesian multilevel models using Stan. J. Stat. Softw. 80 (1), 1–28. https://doi.org/10.18637/jss.v080.i01.

Chen, A., den Os, E., de Ruiter, J.P., 2007. Pitch accent type matters for online processing of information status: evidence from natural and synthetic speech. Linguist. Rev. 24 (2–3), 317–344. https://doi.org/10.1515/tlr.2007.012.

Cohen, A., Collier, R., 't Hart, J., 1982. Declination: construct or intrinsic feature of speech pitch? Phonetica 39 (4–5), 254–273. https://doi.org/10.1159/000261666.

Cohen, A., 't Hart, J., 1965. Perceptual analysis of intonational patterns. In: Commins, D. E. (Ed.), Proceedings of the Fifth International Congress on Acoustics, Vol. paper, p. A16.

Cohen, A., 't Hart, J., 1967. On the anatomy of intonation. Lingua 19 (1–2), 177–192. https://doi.org/10.1016/0024-3841(69)90118-1.

Cole, J., Mo, Y., Hasegawa-Johnson, M., 2010. Signal-based and expectation-based factors in the perception of prosodic prominence. Lab. Phonol. 1 (2), 425–452. https://doi.org/10.1515/labphon.2010.022.

Cole, J., Shattuck-Hufnagel, S., 2016. New methods for prosodic transcription: capturing variability as a source of information. Lab. Phonol. 7 (1), 1–29. https://doi.org/10.5334/labphon.29, 8.

Davis, M.H., Johnsrude, I.S., 2007. Hearing speech sounds: top-down influences on the interface between audition and speech perception. Hear. Res. 229 (1–2), 132–147. https://doi.org/10.1016/j.heares.2007.01.014.

D'Imperio, M., 2000. The Role of Perception in Defining Tonal Targets and Their Alignment. The Ohio State University.

D'Imperio, M., House, D., 1997. Perception of questions and statements in neapolitan Italian. EUROSPEECH-1997, pp. 251–254.

Evans, J.P., 2015. High is not just the opposite of low. J. Phon. 51, 1–5. https://doi.org/10.1016/j.wocn.2015.05.001.

Glasberg, B.R., Moore, B.C., 1990. Derivation of auditory filter shapes from notched-noise data. Hear. Res. 47 (1–2), 103–138. https://doi.org/10.1016/0378-5955(90)90170-t.

Gockel, H., Moore, B.C.J., Carlyon, R.P., 2001. Influence of rate of change of frequency on the overall pitch of frequency-modulated tones. J. Acoust. Soc. Am. 109 (2), 701–712. https://doi.org/10.1121/1.1342073.

Grice, M., Baumann, S., Ritter, S. & Röhr, C.T. (2019). GToBI. Übungsmaterialien zurdeutschen Intonation. Retrieved from www.Gtobi.Uni-koeln.De. (last accessed 23 February 2023).

Gussenhoven, C., 2002. Intonation and interpretation: phonetics and phonology. In: Speech Prosody, 2002, pp. 47–57.

Gussenhoven, C., 2004. The Phonology of Tone and Intonation. Cambridge University Press. https://doi.org/10.1017/CBO9780511616983.

Gussenhoven, C., B H Repp, A.R., Rump, W.H., Terken, J., 1997. The perceptual prominence of fundamental frequency peaks. J. Acoust. Soc. Am. 102, 3009–3022. https://doi.org/10.1121/1.420355.

Gussenhoven, C., Rietveld, A., 1988. Fundamental frequency declination in Dutch: testing three hypotheses. J. Phon. 16, 355–369.

Hartmann, W.M., 1997. Signals, Sound, and Sensation. American Inst. of Physics. ISBN: 1563962837.

House, D., 1990. Tonal Perception in Speech. Lund University Press.

House, D., 1996. Differential perception of tonal contours through the syllable. In: Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP 96. IEEE, pp. 2048–2051.

House, D., Hermes, D., Beaugendre, F., 1997. Temporal-alignment categories of accent-lending rises and falls. In: Proceedings of the 5th European Conference on Speech Communication and Technology (Eurospeech 1997), pp. 879–882. https://doi.org/10.21437/Eurospeech.1997–294.

Hsu, C.H., Evans, J.P., Lee, C.Y., 2015. Brain responses to spoken F 0 changes: is H special? J. Phon. 51, 82–92. https://doi.org/10.1016/j.wocn.2015.02.003.

Jeon, H.-S., Heinrich, A., 2022a. Perceptual asymmetry between pitch peaks and valleys. Speech Commun. 140, 109–127. https://doi.org/10.1016/j.specom.2022.04.001.

Jeon, H.-S., Heinrich, A., 2022b. Perception of pitch height and prominence by old and young listeners. Speech Prosody. ISCA, Lisbon, Portugal, pp. 654–658. https://doi.org/10.21437/speechprosody.2022-133.

Jongman, A., Qin, Z., Zhang, J., Sereno, J.A., 2017. Just noticeable differences for pitch direction, height, and slope for Mandarin and English listeners. J. Acoust. Soc. Am. 142 (2), EL163–EL169. https://doi.org/10.1121/1.4995526.

Jun, S.-A., 2005. Prosodic typology. In: Jun, S.A. (Ed.), Prosodic Typology. Oxford University Press, pp. 430–458.

Klein, S.A., 2001. Measuring, estimating, and understanding the psychometric function: a commentary. Percept. Psychophys. 63 (8), 1421–1455. https://doi.org/10.3758/bf03194552.

Knight, R.A., 2008. The shape of nuclear falls and their effect on the perception of pitch and prominence: peaks vs. plateaux. Lang. Speech 51 (3), 223–244.

Kochanski, G., Grabe, E., Coleman, J., Rosner, B., 2005. Loudness predicts prominence: fundamental frequency lends little. J. Acoust. Soc. Am. 118 (2), 1038–1054.

Krahmer, E., Swerts, M., 2001. On the alleged existence of contrastive accents. Speech Commun. 34 (4), 391–405. https://doi.org/10.1016/s0167-6393(00)00058-3.

Krishnan, A., Xu, Y., Gandour, J., Cariani, P., 2005. Encoding of pitch in the human brainstem is sensitive to language experience. Cognitive Brain Res. 25, 161–168. https://doi.org/10.1016/j.cogbrainres.2005.05.004.

Ladd, D., Verhoeven, J., Jacobst, K., 1994. Influence of adjacent pitch accents on each others perceived prominence: two contradictory effects. J. Phon. 22 (1), 87–99. https://doi.org/10.1016/s0095-4470(19)30268-2.

Ladd, D.R., 2008. Intonational Phonology, 2nd Edition. Cambridge University Press. https://doi.org/10.1017/CBO9780511808814.

Ladd, D.R., Arvaniti, A., 2022. Prosodic prominence across languages. Ann. Rev. Linguist. 9 (1). https://doi.org/10.1146/annurev-linguistics-031120-101954.

Liberman, M.Y., 1975. The Intonation Systems of English. Unpublished PhD Thesis. MIT.

Moore, B.C.J., 2012. An Introduction to the Psychology of Hearing. Emerald Group Publishing Limited.

Moulines, E., Charpentier, F., 1990. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. Speech Commun. 9, 453–467. https://doi.org/10.1016/0167-6393(90)90021-Z.

Niebuhr, O., Reetz, H., Barnes, J., Yu, A.C.L., 2020. In: Gussenhoven, C., Chen, A. (Eds.), Fundamental Aspects in the Perception of f0. Oxford University Press, pp. 28–42.
Nolan, F., 2003. Intonational equivalence: an experimental evaluation of pitch scales. In: Solé, M.J., Recasens, D., Romero, J. (Eds.), Proceedings of the 15th International Congress of Phonetic Sciences, pp. 771–774.
Nolan, F., 2022. The rise and fall of the British School of intonation analysis. Prosodic Theory and Practice. The MIT Press, pp. 319–349. https://doi.org/10.7551/mitpress/10413.003.0012.
Nolan, F., Post, B., 2013. The IViE Corpus. In: Durand, J., Gut, U., Kristoffersen, G. (Eds.), The Oxford Handbook of Corpus Phonology. Oxford University Press, pp. 475–485.
Patterson, R.D., 1976. Auditory filter shapes derived with noise stimuli. J. Acoust. Soc. Am. 59 (3), 640–654. https://doi.org/10.1121/1.380914.
Pierrehumbert, J., 1979. The perception of fundamental frequency declination. J. Acoust. Soc. Am. 66 (2), 363–369. https://doi.org/10.1121/1.383670.
Pierrehumbert, J.B., Hirschberg, J., 1990. The meaning of intonational contours in the interpretation of discourse. In: Cohen, P., Morgan, J., Pollack, M. (Eds.), Intentions in Communication. MIT Press, pp. 271–311.
R Core Team, 2023. R: a language and environment for statistical computing. R Foundation for Statistical Computing. Retrieved from. https://www.R-project.org.
R Studio Team, 2022. RStudio: Integrated Development for R. RStudio. PBC, Boston, MA. Retrieved from. http://www.rstudio.com/.
Repp, B.H., Rump, H.H., Terken, J.M.B., 1993. IPO Annual Progress Report. In: IPO Annual Progress Report, 28, pp. 59–62.
Rietveld, A.C., Gussenhoven, C., 1985. On the relation between pitch excursion size and prominence. J. Phon. 13 (3), 299–308.
Röhr, C.T., Baumann, S., Grice, M., 2022. The influence of expectations on tonal cues to prominence. J. Phon. 94, 101174. https://doi.org/10.1016/j.wocn.2022.101174.
Studdert-Kennedy, M., Hadding, K., 1973. Auditory and linguistic processes in the perception of intonation contours. Lang. Speech 16 (4), 293–313. https://doi.org/10.1177/002383097301600401.
Sutcliffe, J.P., 1972. On the role of "instructions to the subject" in psychological experiments. Am. Psychol. 27 (8), 755–758. https://doi.org/10.1037/h0033111.
Terken, J., 1991. Fundamental frequency and perceived prominence of accented syllables. J. Acoust. Soc. Am. 89, 1768–1776.
't Hart, J., Collier, R., Cohen, A., 1990. A Perceptual Study of Intonation: An Experimental-Phonetic Approach to Speech Melody. Cambridge University Press. https://doi.org/10.1017/CBO9780511627743.
Thorsen, N.G., 1980. A study of the perception of sentence intonation: evidence from Danish. J. Acoust. Soc. Am. 67 (3), 1014–1030. https://doi.org/10.1121/1.384069.
Turnbull, R., Royer, A.J., Ito, K., Speer, S.R., 2017. Prominence perception is dependent on phonology, semantics, and awareness of discourse. Lang. Cogn. Neurosci. 32 (8), 1017–1033. https://doi.org/10.1080/23273798.2017.1279341.
Vaissière, J., 1983. Language-independent prosodic features. Springer Series in Language and Communication. Springer Berlin Heidelberg, pp. 53–66. https://doi.org/10.1007/978-3-642-69103-4_5.
van Heuven, V.J., Turk, A., 2020. Phonetic correlates of word and sentence stress. In: Gussenhoven, C., Chen, A. (Eds.), The Oxford Handbook of Language Prosody. Oxford University Press, pp. 150–165.
van Maastricht, L., Krahmer, E., Swerts, M., 2016. Native speaker perceptions of (non-) native prominence patterns: effects of deviance in pitch accent distributions on accentedness, comprehensibility, intelligibility, and nativeness. Speech Commun. 83, 21–33. https://doi.org/10.1016/j.specom.2016.07.008.
Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L.D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T.L., Miller, E., Bache, S.M., Müller, K., Ooms, J., Robinson, D., Seidel, D.P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., Yutani, H., 2019. Welcome to the tidyverse. J. Open Source Softw. 4 (43), 1686. https://doi.org/10.21105/joss.01686.
Woods, K.J.P., Siegel, M.H., Traer, J., McDermott, J.H., 2017. Headphone screening to facilitate web-based auditory experiments. Atten. Percept. Psychophys. 79 (7), 2064–2072. https://doi.org/10.3758/s13414-017-1361-2.
Zahner, K., Braun, B., 2018. F0 peaks are a necessary condition for German infants' perception of stress in metrical segmentation. In: Proceedings of the 17th Speech Science and Technology Conference (SST 2018). Sydney, Australia, pp. 73–76.
Zahner, K., Kutscheid, S., Braun, B., 2019. Alignment of f0 peak in different pitch accent types affects perception of metrical stress. J. Phon. 74, 75–95. https://doi.org/10.1016/j.wocn.2019.02.004.