Title

Perceptual Asymmetry between Pitch Peaks and Valleys

Author names and affiliations

Hae-Sung Jeon

School of Humanities, Language and Global Studies

University of Central Lancashire

Preston, PR1 2HE

United Kingdom

Email: hjeon1@uclan.ac.uk

Tel: +44 (0)1772 893143

Antje Heinrich

Manchester Centre for Audiology and Deafness

School of Health Sciences

University of Manchester

Manchester M13 9PL

United Kingdom

Corresponding Author

Hae-Sung Jeon

Highlights

- Listeners' capacity to perceive dynamic f0 change is reduced for 'valleys' compared to 'peaks'.
- A long f0 plateau at a 'peak' or 'valley' makes a 'peak' sound higher and a 'valley' lower.
- The magnitude of the pitch saliency-enhancing effect of an f0 plateau is reduced for 'valleys' compared to 'peaks'.
- Listeners' insensitivity to f0 'valleys' is exacerbated at a low frequency.

Accepted for publication April 2022, Speech Communication

Abstract

Perceiving pitch in spoken utterances is an important part of the speech comprehension process. Previous studies on pitch perception concentrated on 'peaks', neglecting 'valleys', thus making it difficult to know how generalisable the findings are. In two experiments, we took a factorial approach and contrasted the pitch movement direction associated with a stressed syllable ('peaks' vs. 'valleys'), the turning point shape (sharp turn, 25 ms, 100 ms plateau), the accent height (varied in one-semitone steps), and pitch levels by shifting the whole contour up or down. We employed different sentences and also non-speech stimuli. Results showed that listeners were better at discriminating the pitch height between 'peaks' compared to 'valleys'. In particular, 'valleys' in the low level posed challenges. A long pitch plateau forming a flat turn at a 'peak' or 'valley' made a 'peak' sound higher and a 'valley' lower compared to a sharp turn or short plateau of the same height. Similar results were observed across speech and non-speech stimuli. The perceptual asymmetry may lead listeners to allocate more attention to 'peaks' than 'valleys' in speech processing, while the effect of listeners' native language deserves further investigation.

Keywords

pitch, intonation, perception, English, accents, plateau

Perceptual Asymmetry between Pitch Peaks and Valleys

1. Introduction

Pitch variation in speech fulfils important roles, such as indicating whether an utterance is a statement or question, and signalling word stress, emphasis, speaker identity, and emotion. Therefore, accurate tracking of the dynamic pitch is important in speech comprehension and in improving communication in challenging situations, such as in noisy backgrounds (e.g., Binns & Culling, 2007; Miller, Schlauch, & Watson, 2010; Shen & Souza, 2017, 2018, 2019).

Although there is an extensive literature on dynamic pitch perception in speech, large gaps in our knowledge remain. For instance, the vast majority of previous studies investigated high pitch or pitch 'peaks' associated with stressed syllables in West Germanic languages (e.g., Fry, 1958; Kohler, 2008; Niebuhr & Winkler, 2017). The emphasis on peaks was probably because a natural relationship is assumed between prosodic prominence and high pitch (cf. Gussenhoven, 2004, p. 85). However, it is not just local pitch 'peaks' that appear in high-information sites in an utterance; 'valleys' may also occur. In various languages, prosodic prominence can be acoustically realised by low fundamental frequency associated with the stressed syllable, particularly in intonational phrase-final position (e.g., Italian, Grice, 1995; Dutch, Gussenhoven, 2005; German, Grice, Baumann, & Benzmüller, 2005; and Estonian, Asu & Nolan, 2007). Although it is known that a low pitch associated with a stressed syllable may be perceptually less salient than a high pitch (e.g., Baumann, 2014; Baumann & Röhr, 2015; Zahner, Kutscheid, & Braun, 2019), little attention has been paid to how the 'valley' contours are perceived. Investigating the acoustic realisation and perception of both 'peaks' and 'valleys' is crucial to improving our understanding of the auditory mechanisms of speech intonation.

Although the terms ‘fundamental frequency (f0)’ and ‘pitch’ are often used interchangeably in the literature, we differentiate ‘f0’ as the physical parameter from ‘pitch’, which refers to the perceptual sensation. We use the term ‘accent’ to refer to either the local maximum point in a ‘peak’ f0 contour or the minimum point in a ‘valley’ f0 contour. Note that this use of the term ‘accent’ differs from that in intonational phonology, where it entails abstract prominence cued by multiple acoustic parameters, such as duration, pitch, and amplitude (Pierrehumbert, 1980; see Ladd, 2008, Chapters 3 and 5). Our choice of the term ‘accent’ is solely for the sake of simplicity and we by no means suggest that f0 is the only correlate of word or sentence stress. Nonetheless, the acoustically defined f0 ‘accent’ is one of the cues to phonological prominence (see Zahner, et al., 2019 for a review), and therefore, the present study of f0 ‘accent’ has implications for the perception of phonological prominence.

One important step in the empirical investigation of the relationship between acoustic parameters, perceived pitch, and phonological prominence is to describe the f0 contour in quantitative terms. In the autosegmental-metrical (AM) theory of intonational phonology (e.g., Pierrehumbert, 1980), intonation is analysed as a sequence of low (L) and high (H) targets and it is assumed that the pitch is interpolated between the successive targets. Theoretically, the tonal targets are abstract, but in the analysis of the f0 tracks as their phonetic correlates, the target is commonly represented by two parameters, its scaling, i.e., the f0 height, and its alignment, i.e., the temporal coordination of the f0 turning point within the segmental context (see Ladd, 2008, Chapters 2 and 5).

However, the precise f0 turning points can be challenging to identify. The f0 turning point is often in the shape of a plateau, i.e., a flat stretch, rather than forming a sharp turn (see House, et al., 1999; Knight & Nolan, 2006; Knight, 2008 for a review, and Astruc et al., 2012 for plateaux in children’s speech). In addition, researchers experience discrepancy between the perceived pitch height and the f0 as estimated by the algorithms implemented in the speech

analysis tools; perceived pitch of the f0 turning point is affected by factors other than f0, such as f0 peak location within the utterance, excursion size, f0 at the utterance onset, the shape of the f0 peak, the presence of an intonational phrase boundary, listeners' expectations, and speaker pitch range (e.g., Rietveld & Gussenhoven, 1985; Terken, 1991, 1994; Gussenhoven et al.,1997; Gussenhoven & Rietveld, 1998; Knight, 2008). The focus of our investigation is directly comparing the accents which differ in the direction of f0 movement associated with them, rise-fall vs. fall-rise.

We contrast Λ-shaped f0 'peaks' and V-shaped 'valleys' which were a mirror image of each other. Although both a 'peak' and a 'valley' are composed of a rise and a fall, their ordering determines the relative height of the turning point. In a 'peak', listeners hear the rise first, leading to a local maximum f0, but for a 'valley', the f0 falls towards the local minimum. Such differences may lead to an asymmetry in listeners' perception of 'peaks' and 'valleys'. A rise over a vowel (Evans, 2015; Hsu et al., 2015) or a short tone and a nonsense word (Turner et al., 2019) seems to be perceived more saliently and accurately than a fall. Listeners are relatively insensitive to variations in the f0 of minima compared to maxima in utterances (Sluijter, 1991, cited in Gussenhoven et al., 1997), and discrimination of f0 excursion size is easier in rises than in falls in an utterance ('t Hart, 1991). Similarly, listeners have more difficulties in identifying f0 falls than rises with tone glides (e.g., Collins & Cullen, 1978; Gordon & Poeppel, 2002; Kishon-Rabin et al., 2004) and also with nonsense words (Turner et al., 2019). However, to our knowledge, no study has directly compared listeners' perception of f0 'peaks' and 'valleys' with their f0 excursion manipulated in a symmetrical manner in utterance-length stimuli.

Although a high pitch accent (H*) and a low pitch accent (L*) in the AM analysis technically refer to a local peak and a valley, respectively, their acoustic forms in speech are usually not symmetrical. For instance, in American English, the low pitch accent (L*) is often

realised with a long stretch of f0 in the low range of the speaker's voice, whereas the local peak for the high accent (H*) may be clearly definable (see Ladd, 2008, Section 3. 1. 3). Accordingly, speech perception or processing studies have used asymmetrical acoustic shapes between high and low accents for naturalness. For instance, Kutscheid, et al. (2021) showed that German listeners are more likely to associate prominence at the word or sentence level with high pitch accents (e.g., L+H*) than with low pitch accents (e.g., L*+H, L*, H+L*). Similarly, Zahner et al. (2019) demonstrated the processing advantage of the high pitch accents (L+H*) compared to the low pitch accents (H+L*). In these studies, the high accents tended to have a clearly defined local f0 peak, whereas the low accents had a gradual fall with a less salient f0 movement compared to the high accents. Then, the advantages of the high accents in perception and processing might have been due to the lack of sufficient acoustic contrasts for the low accents. Therefore, we created acoustically symmetrical 'peaks' and 'valleys' to clarify the source of their perceptual asymmetry.

Second, we manipulated the f0 turn shape in the stressed syllable associated with an accent. Plateau-shaped f0 peaks are perceived higher compared to sharp peaks with the same maximum f0 ('t Hart, 1991; Knight, 2008; see Barnes, Veilleux, Brugos, & Shattuck-Hufnagel, 2012 for a survey). However, it is not yet clear how the pitch saliency-enhancing effect of plateaux interacts with different f0 dimensions. For instance, the effect may be diminished for 'valleys' compared to 'peaks' because of the interaction between f0 and duration. Level tones higher in f0 tend to sound longer than lower tones (Lehiste, 1976; Rosen, 1977; Jeon & Fricke, 1997), and the same trend was found in perceiving vowels by listeners whose native language was Dutch or Chinese (Gussenhoven et al., 2013) or English (Yu, 2010). On the other hand, the experimental results on the duration perception of a tone with a f0 rise or fall are mixed and studies show a potential influence of listeners' linguistic experience. For instance, while native speakers of Finnish perceived a tone with an f0 rise longer than that with a fall of the same

duration in Dawson, et al. (2017), the opposite was true in Šimko et al. (2015). Listeners' duration perception was unaffected by whether a vowel had a rise or a fall in Yu (2010) and Gussenhoven & Zhou (2013). Therefore, while it is unclear how the f0 movement direction interacts with the perceived duration in an utterance, a plateau of the same duration may be perceived as longer when occurring with 'peaks' than with 'valleys'.

Third, we examined the generalisability of the effect of f0 manipulation across different sentences. The conventional approach in perception studies is to resynthesise one utterance to systematically vary the f0 contours to keep other variables under control. However, listeners' pitch perception is affected by segmental properties, such as the intrinsic f0 of vowel (e.g., Lehiste, 1970; Whalen & Levitt, 1995), the distribution of acoustic energy (Cangemi, Albert, & Grice, 2019), and the intensity or sonority of utterance constituents (Barnes et al., 2011; Barnes et al., 2014). The interaction between the acoustic properties of utterance constituents and f0 poses a broad question which deserves thorough investigation (cf. Barnes et al., 2021). In our experiment, different sentences were used as experimental materials, although they mainly consisted of vowels and sonorants.

Finally, we examined the effect of the whole f0 contour shifting up or down, operationalised as the f0 level. For instance, an adult female with a voice f0 median of 200 Hz can raise her voice to be at 250 Hz, while keeping the f0 excursion associated with an accent constant. The variation in voice pitch level or range across speakers or within a speaker does not seem to be a problem in real life, because listeners linguistically interpret a particular pitch contour, for instance, a rise, as signalling a stressed syllable or a question (see Ladd, 2008, Section 5. 2). What needs further investigation is the detailed process of how listeners derive the linguistic interpretation from the acoustic form. Much research has been carried out to model between- or within-speaker variation in f0 range, for instance, in terms of speaker characteristics (Patterson & Ladd, 1999) and scaling intonational targets (Shriberg et al., 1996;

Gussenhoven & Rietveld, 2000). However, the perception of low or falling f0 in the utterance context in varying f0 levels is yet to be examined. It is possible that height discrimination is more difficult in the relatively low than high level of human voice. For pure tones, an inverse relationship exists between the thresholds for sound pressure level (dB SPL) and f0 in the typical adult speech range (70–400 Hz, cf. Fletcher & Munson, 1933; ISO, 2003), and the same may be true for speech. That is, at a low f0, listeners may need a higher intensity level to achieve sensitivity comparable to that for the higher f0 level. In addition, if high or rising f0, but not low or falling f0, captures listeners' attention (e.g., Hsu et al., 2015), then the perceptual ease of 'peaks' would be less likely to be affected by suboptimal contexts such as in soft sound or at low f0. On the other hand, the low sensitivity to 'valleys' might lead to a reduced pitch perceptibility at particularly low levels compared with higher levels. We aimed to explore the perceptual consequences of the same magnitude of f0 changes in semitones in 'peaks' and 'valleys' in different levels.

Below we present two experiments investigating the effects of accent type ('peaks' vs. 'valleys'), plateau duration (0 ms, 25 ms, 100 ms) and f0 level (high at 200–302 Hz vs. low at 132–200 Hz). In addition to the English speech stimuli, Experiment 1 used complex tones and reiterated speech, and Experiment 2 used complex tones. In the interests of keeping the paper concise, we only present the results of the English speech trials. The results across stimulus types (English speech, reiterated speech, and tones) were comparable (see Appendices A and B).

2. Research questions and hypotheses

The experiments address the following questions: (1) whether there is a perceptual asymmetry in discriminating pitch height between f0 'peaks' and 'valleys' in utterance contexts other things being equal; (2) how the pitch saliency-enhancing effect of f0 plateaux is

manifested across ‘peaks’ and ‘valleys’; and (3) whether listeners’ pitch height discrimination is comparable across f0 levels. We measured listeners’ ability to identify the relative pitch height between two ‘accents’ (cf. Rietveld & Gussenhoven, 1985; Gussenhoven & Rietveld, 1998; Terken, 1991), rather than their prominence to direct listeners’ attention to speech melody. In all auditory stimuli, the first accent’s height was kept constant throughout the experiment while the second accent’s height was orthogonally varied in five one-semitone steps with the plateau duration. We tested the following hypotheses.

Hypothesis 1: The effect of f0 height manipulation on pitch height perception is less pronounced for ‘valleys’ than for ‘peaks’.

Hypothesis 2: Listeners’ reduced capacity in perceiving ‘valleys’ interacts with the f0 contour shape associated with an accent. A relatively diminished saliency-enhancing effect of f0 plateaux for ‘valleys’ compared to ‘peaks’ is expected.

Hypothesis 3: The sentence ‘item’ factor has a significant effect on listeners’ responses.

Hypothesis 4: Changes in the f0 level has a significant effect only for ‘valleys’.

3. Experiment 1

3.1 Method

3.1.1 Participants

Twenty native speakers of British English (8 male, 12 female, age Mean = 24.6, SD = 3.44, range 18–30 years) with self-reported normal hearing and corrected-to-normal vision took part in the experiment. Participants were recruited with posters and by online advertisements at the University of Manchester. All participants provided written informed consent prior to the study. They received a small monetary compensation for participation after the experiment. All response data were anonymised for analysis. The study was approved by

the local ethics committees of the University of Manchester (2018-4595-6682) and the University of Central Lancashire (STEMH 922).

3.1.2 Experimental design and stimulus types

The experimental design was a 2 Accent Type (Peaks, Valleys) × 2 Second Accent Shape (Sharp Turn, 100 ms Plateau) × 5 levels in Accent Height Difference (-2, -1, 0, +1, +2 ST) × 4 Items (Lemmy, Nellie, Mona, Nina). All auditory stimuli had two accents, which were either two high f0 peaks or two low valleys. The stimuli were based on four English sentences: 'is Lemmy near Nellie?' ('Lemmy' Item), 'is Nellie near Lemmy?' ('Nellie' Item), 'does Mona know Nina?' ('Mona' Item), and 'does Nina know Mona?' ('Nina' Item). These sentences were designed to have words with initial stress (e.g., Lémmy and Néllie) for the accents to be aligned to and to contain sonorants to keep f0 perturbations minimal. All sentences were six syllables long and were designed to have two accented syllables on the same location on the second and the fifth syllables. The same name was placed in different positions in a sentence pair to counterbalance a potential effect of intrinsic pitch related to segmental composition.

3.1.3 Recording and resynthesis of experimental stimuli

A female native speaker of Standard Southern British English in her 30s read the English sentences at a comfortable speaking rate neutrally with 'peak' and 'valley' accents several times (see Figs. 1 and 2). The speech was recorded at a sampling rate of 51.2 kHz using a Samurai 2.2.6 sound level meter (Sinus Messtechnik GmbH, Leipzig, Germany) and recording software onto a Toshiba Satellite Pro laptop using a calibrated ½-inch condenser microphone (Type 4134 Brüel and Kjær, Nærum, Denmark), preamplifier (Type 26AM G.R.A.S. Sound and Vibration, Holte, Denmark) and acoustic analyser (Apollo-Box, Sinus). Recording took place in a custom-built anechoic chamber at the University of Nottingham and

the microphone was positioned 20 cm from the speaker's mouth with the diaphragm at a 0 degree orientation.

All stimuli were resynthesised using Praat ver. 6.0.29 (Boersma & Weenink, 2017). Half of the stimuli had an f0 contour with two 'peak' accents and the other half had two 'valley' accents. The first accent always formed a sharp turn and the second one formed either a sharp turn or a 100 ms plateau. The 100 ms plateau duration was chosen following Knight (2008).[1] By choosing 0 and 100 ms we intended to maximise the contrast in the plateau duration between the two accents to establish a benchmark for future studies investigating the effect of varying plateau duration.

The auditory stimuli were resynthesised from one utterance chosen for each of the four English sentences with accent peaks as a base utterance. The utterances with accent valleys were not ideal for resynthesis due to creaks which appeared near the floor of the speaker's f0 voice range.[2] In the process of selecting the base utterance, the f0 and durational properties of each utterance were examined and the token which had a similar duration between the two areas under the two accents [(3)–(1) and (8)–(4) in Figure 1] was chosen. To prepare for resynthesis, in each utterance, the interval between the beginning of the f0 rise and the end of the fall around each accent was measured; the duration of the shorter interval between the two was chosen as the target duration to be implemented in the resynthesised stimuli. In each utterance, the first author who is trained in prosodic analysis identified eight time points for f0

[1] The shape of the f0 accent peak varied across previous studies. For instance, Pierrehumbert (1979) used resynthesised utterances from natural production without specifying the peak shape, Gussenhoven and Rietveld (1998) used sharp f0 peaks, while the f0 contours formed a 30 ms plateau in Terken (1991).

[2] An anonymous reviewer questioned whether the original utterance spoken with peak accents could have sounded unnatural and synthetic when resynthesised to create the 'valley' stimuli. Indeed, some change in the timbre might have been noticed by careful listeners when high pitch was resynthesised to be low. However, such changes were unavoidable, because the resynthesis can only reduplicate pulses from the original, while the pulse shapes vary depending on the pitch in natural speech production. The trends in the results in Experiment 1 were replicated when a neutrally spoken utterance was resynthesised into both 'peaks' and 'valleys' in Experiment 2. Therefore, it is unlikely that any artefact in the resynthesis process significantly affected the results.

stylisation (see Figure 1): (1) the beginning of the first f0 rise, (2) the mid-point of the first vowel in the first name, (3) the end of the f0 fall, (4) the beginning of the f0 rise for the second accent, (5) a time point 50 ms before the vowel mid-point, (6) the mid-point of the first vowel in the second name, (7) a time point 50 ms later than the vowel mid-point, and (8) the end of the final f0 fall. Points (2), (5), (6), and (7) were identified first. Then the f0 'elbows' were semi-automatically identified by using the minimum pitch detection function in Praat for deciding the candidate location of (1), (3), (4), and (8). Then the points were slightly moved from the semi-automatically detected 'elbows' to an appropriate place if necessary to ensure that the duration of the interval under the first f0 curve [(3)–(1)] and that under the second f0 curve [(8)–(4)] could be the same in each utterance (Mean = 354.75 ms, SD = 28.65). Consequently, they were not strictly associated with the precise f0 turning point or the segmental boundary in the original utterance. In real speech, the f0 maximum or minimum in an accented syllable is not precisely aligned to the vowel mid-point, and the alignment may vary depending on the sentential position, being earlier in a nuclear accent than in a pre-nuclear accent (e.g., Ladd et al., 2009). The simplistic approach aligning the f0 turn to the vowel mid-point was taken here to avoid confounds related to the alignment. The mean duration of the resynthesised English utterances was 1.18 s (SD = 0.03). The present design did not allow a control of the f0 movement slope towards the maximum of a 'peak' or the minimum of a 'valley'. Implementing the 100 ms plateau resulted in a decreased slope of f0 compared to a sharp turn.

Fig. 1. The spectrogram and f0 track of the recorded utterance ‘does Nina know Mona?’ with two accent peaks. The word boundaries, the location of the eight points for f0 stylisation and the segmental boundaries are annotated. In the last tier, segments are annotated as c (consonant) and v (vowel). The points (1), (3), (4), and (8) were decided to the same durations of the interval under the first f0 curve [(3)–(1)] and that under the second f0 curve [(8)–(4)].

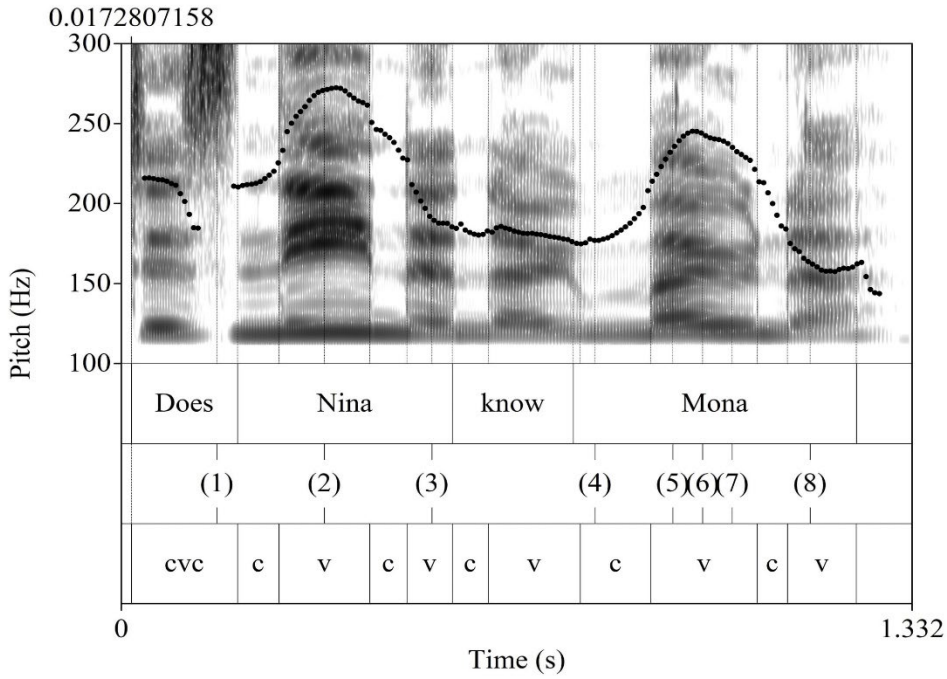
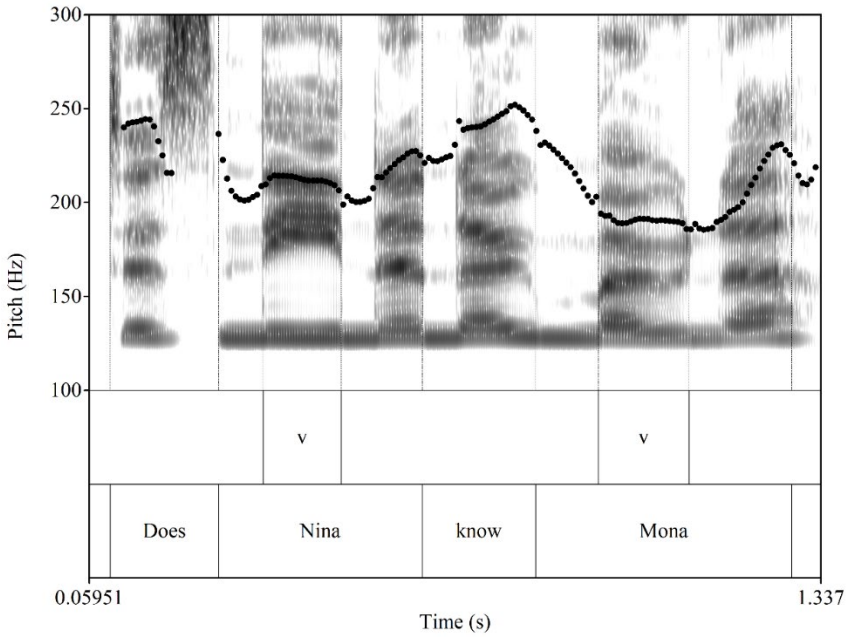


Fig. 2. A sample spectrogram and f0 track of the recorded utterance ‘does Nina know Mona?’ with two accent valleys. The word boundaries and the stressed vowels are annotated.



The f0 baseline was set at 200 Hz, which was close to the speaker's value near the beginning of the utterance (approximately 205 Hz). The height differences between the two accents were expressed in the musical semitone (ST) scale, which seemed most appropriate for comparing pitch events occurring at different levels (cf. Graddol, 1986; Nolan, 2003). For resynthesis of the Peaks stimuli, the first f0 accent maximum was set at 273 Hz (5.4 semitones higher than the 200 Hz baseline, Figure 3) in all stimuli. The second accent f0 maximum varied in five one-semitone steps (Table 1). The maximum difference between the baseline and the second accent, 7.4 ST, was chosen based on Knight (2008). The f0 contours of the Valleys stimuli were mirror images of the Peaks counterparts. The accents occurred between 200 Hz and 307 Hz for Peaks and between 130 Hz and 200 Hz for Valleys. The root-mean-square amplitude of all base utterances was scaled at 70 dB before further resynthesis. Sample stimuli are available as Supplementary Materials.[3]

3.1.4 Experimental procedures

Sound output was calibrated to 70 dB SPL using a B&K 2250 sound level meter and a B&K artificial ear type 4153. Sounds were delivered via a DELL 17-7000 (i7-4510 @2 gHz) with Windows 8 Pro, using a Realtek ALC3253CG soundcard with Waves Maxx Audio Pro and Sennheiser HD280 headphones. All trials were presented with Praat ver. 6.0.26 (Boersma and Weenink, 2017). Testing was carried out in a double-wall sound-attenuating booth (Industrial Acoustics Company, Winchester, UK) at the University of Manchester.

[3] In the sample sound files (expt1_English_peaks_sharp.wav, expt1_English_valleys_sharp.wav, expt1_English_peaks_plateau.wav., expt1_English_valleys_plateau.wav), the two accents have the same f0 height at the stressed vowel mid-point.

Fig. 3. f0 tracks of the resynthesised stimuli 'is Lemmy near Nellie?' with the second accent in Plateau (100 ms) in Peaks and Valleys. In the Sharp Turn stimuli (not shown in the figure), the f0 maxima in Peaks and minima in Valleys were aligned to (6), the mid-point of the vowel. The constant f0 at the beginning, between the two accents, and at the end is referred to as the baseline at 200 Hz (Experiment 1).

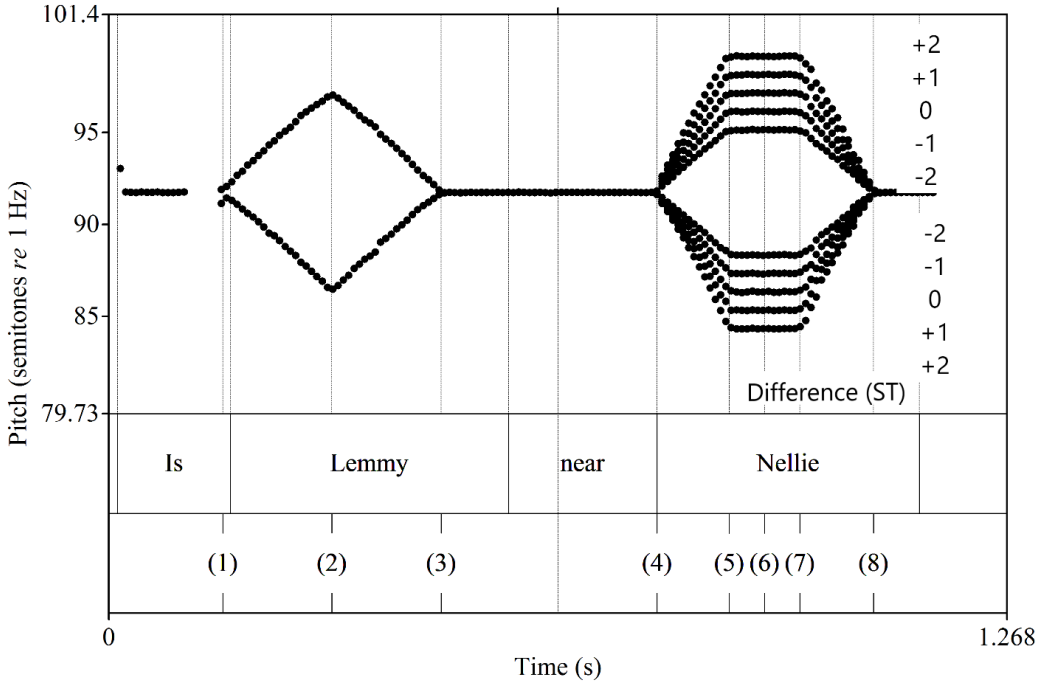


Table 1. The difference between the first accent height and the second accent height measured from the 200 Hz baseline. The first accent height always corresponded to the height for 0 ST Accent Height Difference. The negative Accent Height Difference values indicate that the second accent had a smaller f0 excursion from the baseline than the first (See Fig. 3., Experiment 1).

	Accent Height Difference	Difference from baseline	f0 (Hz)
Peaks	-2 ST	3.4 ST	243.40
	-1 ST	4.4 ST	257.87
	0 ST	5.4 ST	273.21
	+1 ST	6.4 ST	289.45
	+2 ST	7.4 ST	306.67
Valleys	-2 ST	-3.4 ST	164.33
	-1 ST	-4.4 ST	155.11
	0 ST	-5.4 ST	146.41
	+1 ST	-6.4 ST	138.19
	+2 ST	-7.4 ST	130.44

The Experiment consisted of three Stimulus Type blocks (Complex Tone, Reiterated Speech, and English Speech). The details of the stimuli, experimental procedure and results of

the Complex Tone and Reiterated Speech trials are shown in Appendix A. For all stimuli, the presentation order of the blocks was counterbalanced, and within each block, the stimulus presentation order was randomised for each participant. Half of the participants listened to the 'Lemmy' and 'Nellie' stimulus pairs and the other half listened to the 'Mona' and 'Nina' pairs. Each participant was randomly assigned to the stimulus pair group.

When given instructions, participants were shown schematic representations of two 'peaks' or 'valleys' in line drawings and informed that they would hear auditory stimuli with either two 'peaks' or 'valleys' in the melody. They were instructed to listen to English utterances 'is Lemmy near Nellie?', 'is Nellie near Lemmy?', 'does Mona know Nina?' and 'does Nina know Mona?' as appropriate for their stimulus group. Then they were instructed to identify which melodic 'peak' sounded higher or which 'valley' sounded lower respectively by listening to the height of the peaks or valleys. There was a practice session with eight stimuli [2 Accent Type (Peaks, Valleys) × 2 Second Accent Shape (Sharp Turn, Plateau) × 2 Accent Height Difference (-2, +2)] before the main experiment.

Each stimulus was played 0.5 seconds after the onset of each trial. In each trial, the question 'which one sounds higher?' for the Peaks trials or 'which one sounds lower?' for the Valleys trials appeared at the top of the screen as relevant to Accent Type. Two buttons labelled 'first' and 'second' appeared on the screen. Participants could repeat the stimulus presentation as often as they liked. Participants indicated their choice by clicking the appropriate button with a mouse. They were allowed to change their response after repeated presentations within a trial. Only the final response, before clicking 'ok', was recorded. No feedback was given. The experiment was self-paced, and all stimuli were played three times throughout the experiment. The experiment lasted approximately 30 minutes.

3.1.5 Analysis

There were 2,400 data points in total [2 Accent Type (Peaks, Valleys) × 2 Second Accent Shape (Sharp Turn, 100 ms Plateau) × 5 Accent Height Difference (-2, -1, 0, +1, +2 ST) × 2 Items per participant × 3 repetitions × 20 participants]. The averaged response frequency in percentages for all experimental conditions is shown in Figure 4. We explored the effects of experimental factors and their interaction terms by comparing mixed-effect logistic models fitted to the response data (the *glmer* function in the *lme4* package, Bates, et al., 2015) using R ver. 4.0.3 (R Core Team, 2020). Our data analysis focused on the model comparisons rather than fitting the maximal model (Barr, et al., 2014) because of the complexity of the experimental design. Our logistic models involved multi-level factors, and a minor decision in the model-building process, such as the reference level or the ordering of factors, affects the process and results (cf. Clopper, 2013).

The logistic models estimated the maximum likelihood of the positively coded 'second accent' response. Initially, a model was fitted with all fixed factors, Accent Type, Second Accent Shape, Accent Height Difference, and Item, with Listener as a random intercept. Item was incorporated as a fixed factor rather than a random factor so that we could examine its interaction with other fixed factors, Accent Type and Accent Shape, in particular. We then tested the effect of each fixed factor and their interactions by examining whether the model containing the effect of interest (as listed in Table 3) significantly improved the model fit compared to a lower-order model built without it. We interpreted that a lower value of the Akaike information criterion (AIC, Akaike, 1974) indicated a better fit and used the log-likelihood test with the *anova* function ($\alpha = 0.05$). We tested the two- and three-way interactions in the same manner. For instance, for assessing a three-way interaction, we constructed a model with all fixed factors, the three-way interaction and all possible two-way interactions between the fixed factors involved in the three-way interaction. Then we compared it with a lower-order model, which was built with all fixed factors and the two-way

interactions. We did not build models with four-way interactions because they are not interpretable, and also because we did not observe any significant three-way interactions.

3.2 Results

The model comparison results are summarised in Table 3. To interpret interactions, we present the 'second accent' response frequency (%) for all experimental conditions together with the frequency averaged for each fixed factor in Table 2. Figure 4 shows the 'second accent' response functions, and the horizontal reference line is at 50% 'second accent' response frequency. Although none of the three-way interactions was statistically significant (Table 3), Figure 4 presents the results by Accent Type, Item, and Second Accent Shape, because they were part of two-way interactions. The point where the response function crossed the reference line was regarded as the Point of Subjective Equality (PSE), i.e., the Accent Height Difference at the chance-level 'second accent' response (50%), indicating perceptual equivalence between the height of the two accents. Above the PSE in semitones, listeners perceived the second accent as more salient in pitch, i.e., higher for 'peaks' and lower for 'valleys', than the first accent.

Accent Type was a significant main effect ($p < 0.001$). Overall, listeners had a stronger bias to perceive the second accent as more salient in pitch than the first for Peaks (Mean = 61.92, SD = 39.52) compared with Valleys (Mean = 49.58, SD = 37.16).

The main effect of Second Accent Shape ($p < 0.001$) showed that listeners were more likely to perceive the second accent as more salient than the first for Plateau (Mean = 63.33, SD = 38.27) than for Sharp Turn (Mean = 48.17, SD = 37.94). Further, Second Accent Shape did not interact with any other fixed factors; Plateau had a consistent effect of increasing the pitch saliency across Second Accent Shapes and Items. In Figure 4, the response functions for

Plateau were relatively shifted up compared to Sharp Turn in each Item pair. In addition, in all Plateau–Sharp Turn pairs, the PSE was lower for Plateau; i.e., a smaller f0 difference between the two accents, when the second accent was lower than the first for Peaks and higher for Valleys, led to perceived equivalence in pitch for Plateau more than for Sharp Turn.

Item did not have a significant main effect, and listeners' 'second accent' response frequency was 50-60 % across Items.

The Accent Height Difference effect shows the significant change in listeners' responses evoked by a one-semitone f0 change in the second accent height. In Figure 4, listeners' heightened capacity to perceive the second accent height change resulted in a steep positive slope of the response function, while a flat function indicates that listeners reacted little. Accent Height Difference had a significant main effect ($p < 0.001$); in all experimental conditions, the response function showed a positive slope (Fig. 4).

The absence of a significant Accent Type × Second Accent Shape interaction effect could be interpreted as indicating that the Sharp Turn vs. Plateau contrast had a comparable effect across Peaks and Valleys. On the other hand, there was a significant Accent Type × Item interaction effect ($p < 0.001$). Listeners' stronger bias towards the second accent saliency for Peaks than for Valleys was shown for two out of the four Items, Lemmy (Peaks, Mean = 63.33, SD = 47.71; Valleys, Mean = 49.67, SD = 34.97) and Mona (Peaks, Mean = 69, SD = 33.92; Valleys, Mean = 34.33, SD = 32.64). For these Items, the response functions for Peaks and Valleys (Figure 4) were clearly separated. In addition, the difference in the PSEs for Peaks and Valleys was notable, showing that listeners needed a larger f0 excursion from the baseline for Valleys than for Peaks to achieve pitch equivalence between the two accents. In particular, Figure 4 shows markedly low 'second accent' response frequency for Mona, Valleys. On the other hand, the Peaks vs. Valleys difference was relatively reduced and the response functions

were clustered for Nellie (Peaks, Mean = 58.33, SD = 40.86; Valleys, Mean = 60.67, SD = 36.81) and Nina (Peaks, Mean = 57, SD = 40.56; Valleys, Mean = 53.67, SD = 39.33). The PSE differences between Peaks and Valleys were minimal for Sharp Turn for these Items. The main source of the Accent Type × Item interaction is likely to be Nellie, which was exceptional in that overall listeners' 'second accent' response frequency was slightly higher for Valleys (Mean = 60.67, SD = 36.81) than for Peaks (Mean = 58.33, SD = 40.86). However, the Peaks vs. Valleys asymmetry was still present for Nellie. Despite the marginally higher overall 'second accent' response frequency for Valleys than for Peaks, the response function for Valley for Nellie (Fig. 4) still has a less steep slope than for Peaks.

Accent Height Difference was observed in significant two-way interactions. First, although the effect of Second Accent Shape × Accent Height Difference ($p = 0.58$) did not reach the statistical significance threshold, Figure 4 shows that Plateau led the 'second accent' responses to ceiling at +1 and +2 ST Accent Height Differences for Peaks. Second, the Accent Type × Accent Height Difference interaction ($p < 0.001$) was shown by the steeper response functions for Peaks than for Valleys (Fig. 4), i.e., listeners' discrimination was heightened for Peaks. Third, the Item × Accent Height Difference interaction ($p < 0.001$) indicated that the response function slope was significantly different across Items.

To summarise the results, an asymmetry in perception of different accent types, f0 'peaks' and 'valleys', was observed. Listeners showed heightened discrimination for 'peaks' than for 'valleys', and they were more prone to perceive the second accent as salient in pitch than the first for 'peaks' than for 'valleys'. The 'accent type' had a significant interaction effect with 'items', i.e., the four utterances used in the experiment ('Lemmy', 'Nellie', 'Mona' and 'Nina'), but the 'peaks' vs. 'valleys' asymmetry shown in the response functions was consistent across the 'items'. The plateau-shaped accent seemed to have a constant pitch-saliency-enhancing effect for both 'peaks' and 'valleys'. Listeners were more likely to perceive

an accent with a 100 ms plateau as salient compared to an accent with a sharp turn other things being equal. In addition, for the 'plateau' stimuli, listeners were highly likely to perceive the second accent as more salient than the first when the second accent was lower in f0 than the first by 1–2 semitones, particularly for 'peaks'.

Table 2. Frequency (%) means and standard deviations of listeners' 'second accent' responses for each Item (Experiment 1).

		Lemmy		Mona		Nellie		Nina		Total	
Shape	Type	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Plateau	Peaks	74.67	37.23	74.00	31.07	64.67	41.19	64.67	39.50	69.50	37.48
	Valleys	56.67	37.04	43.33	34.50	69.33	36.79	59.33	40.57	57.17	38.16
	Total	65.67	38.04	58.67	36.12	67.00	38.92	62.00	39.93	63.33	38.27
Sharp	Peaks	52.00	43.20	64.00	36.17	52.00	39.93	49.33	40.54	54.33	40.14
	Valleys	42.67	31.62	25.33	28.22	52.00	35.10	48.00	37.59	42.00	34.61
	Total	47.33	37.96	44.67	37.67	52.00	37.40	48.67	38.90	48.17	37.94
Total	Peaks	63.33	41.71	69.00	33.92	58.33	40.86	57.00	40.56	61.92	39.52
	Valleys	49.67	34.97	34.33	32.64	60.67	36.81	53.67	39.33	49.58	37.16
	Total	56.50	39.00	51.67	37.47	59.50	38.81	55.33	39.88	55.75	38.83

Fig. 4. The averaged frequency of 'second accent' responses (%) across all participants by Item, Second Accent Shape, and Accent Type in Experiment 1.

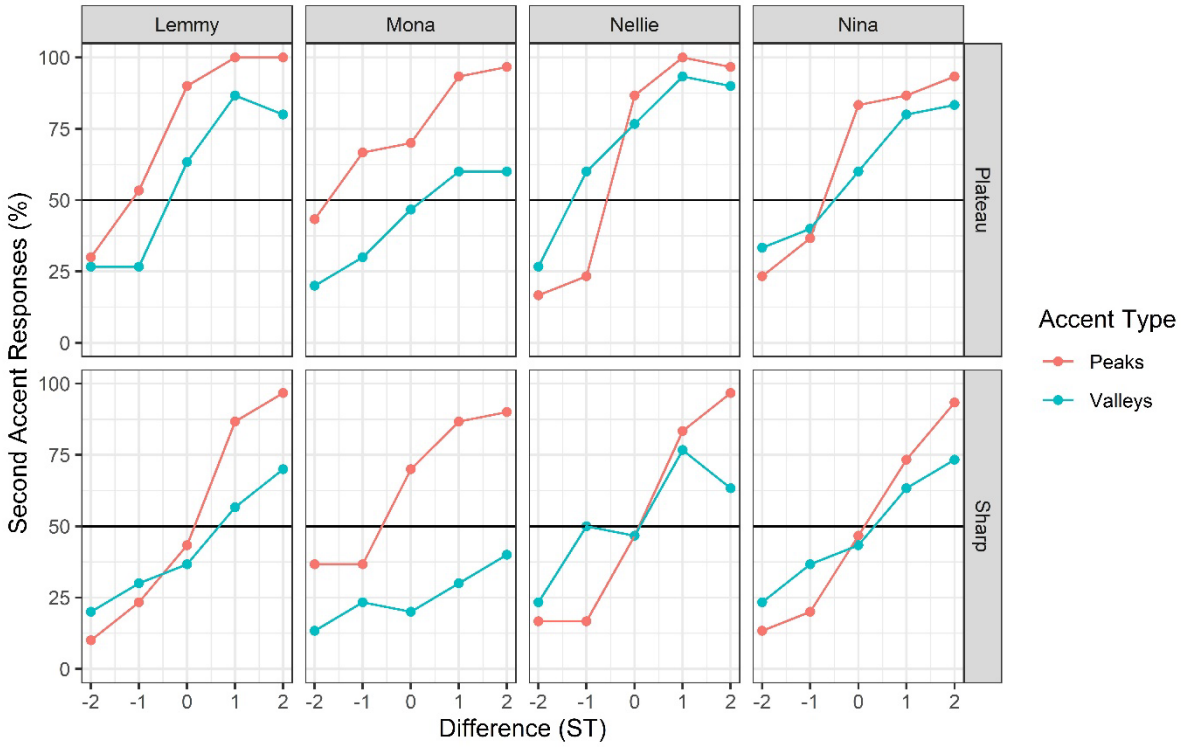


Table 3. Results of the model comparisons ($\alpha = 0.05$). The reference level for all models was Accent Type-Peaks, Second Accent Shape-Sharp Turn and Item-Lemmy (Experiment 1).

	χ^2	df	p
Accent Type	50.34	1	< 0.001 ***
Shape	75.54	1	< 0.001 ***
Item	5.31	3	0.15
Difference	582.01	1	< 0.001 ***
Accent Type × Shape	0.038	1	0.85
Accent Type × Item	67.26	3	< 0.001***
Accent Type × Difference	57.63	1	< 0.001***
Shape × Item	1.26	1	0.73
Shape × Difference	3.6	1	0.58
Item × Difference	25.33	3	< 0.001***
Accent Type × Shape × Item	4.17	3	0.24
Accent Type × Shape × Difference	7	7	0.43
Shape × Item × Difference	4.91	3	0.18

3.3 Discussion of Experiment 1

Hypothesis 1 that listeners' discrimination would be relatively heightened for 'peaks' compared with 'valleys' was supported. However, the direct source of this effect is ambiguous, because 'peaks' and 'valleys' differed in two dimensions, the direction of the f0 movement

(fall-rise vs. rise-fall) and the f0 level at which the f0 movements occurred ('peaks' in the range 200–307 Hz, 'valleys' in 130–200 Hz). For 'valleys', it was not only the fall-rise of the f0 movement that could have posed challenges for listeners but also its low level. Comparing the speech results with those for complex tones speaks to the question of whether listeners' stronger bias for the second accent saliency for 'peaks' than for 'valleys' is more likely to be a psychoacoustic or linguistic effect. If it were a linguistic effect, one might suggest that the bias could originate from listeners' expectations about 'declination' in speech, i.e., the downtrend in f0 over an utterance (Collier, 1987), while they may have an opposite bias for 'valleys'. However, this is unlikely to have happened as similar results were observed for complex tones, which were not likely to evoke listeners' linguistic expectations (see Appendix A).

We did not find strong evidence for Hypothesis 2 that the pitch-saliency-enhancing effect of an f0 plateau would be reduced for 'valleys' compared to 'peaks'; the interaction between the accent type ('peaks' vs. 'valleys') and the second accent shape ('sharp turn' vs. 'plateau') was not statistically significant. For the 'plateau' condition, the response functions were shifted up, the PSEs were lower, and the averaged frequency of the 'second accent' responses was higher compared to the 'sharp turn' condition. For a plateau-shaped accent, the f0 reached the maximum for a 'peak' or the minimum for a 'valley' earlier in comparison to the f0 in the accent with a sharp turn. Although the steeper slope of f0 movement towards a plateau could hinder listeners' f0 tracking, the disadvantage seemed to be compensated by the stability that the maximum or minimum f0 was maintained for 100 ms.

Hypothesis 3 was supported in that the utterance type ('item') significantly affected the listeners' responses. For instance, while all items were subject to asymmetry, the 'peaks' vs. 'valleys' asymmetry was more notable for 'Lemmy' and 'Mona' than for 'Nellie' and 'Nina' (Fig. 4). The response functions were relatively flat for 'Mona'.

Despite addressing several hypotheses, there are remaining questions. First, as discussed above, the direct source of the 'peaks' vs. 'valleys' asymmetry remains ambiguous. Second, although the interaction effect between the accent type ('peaks' vs. 'valleys') and shape ('sharp turn' vs. 'plateau') was not statistically significant, the variance across the four utterances might have led to this outcome. Finally, throughout the experiment, the first accent always formed a sharp f0 turn. Such a sharp f0 turn is unlikely to be produced in real speech and the lack of stable f0 could have made pitch tracking challenging for listeners, and consequently, they may have based their judgements on guessing, more biased towards the 'second accent' response for 'peaks' than for 'valleys'. Experiment 2 addressed these issues. The 'accent types' and f0 'levels' were crossed, and we used only one English utterance 'does Nellie know Lenny?' to focus on the effect of f0 manipulation. We improved the stimulus design to approximate speech more closely. In all stimuli, the first accent had a 25 ms plateau while the plateau duration of the second accent varied (25 ms vs. 100 ms). The accent 'peak' or 'valley' was aligned at the end of the stressed vowel.

4. Experiment 2

4.1 Method

The experimental design was: 2 Accent Type (Peaks vs. Valleys) × 2 f0 Level (High vs. Low) × 2 Second Accent Shape (25 ms vs. 100 ms plateau) × 5 Accent Height Difference (−2, −1, 0, +1, +2 ST).

4.1.1 Stimuli

As in Experiment 1, half of the stimuli had an f0 contour with two accent 'peaks' and the other half had 'valleys'. The stimuli were based on the English sentence 'does Nellie know Lenny?'. We intended the two names, Nellie and Lenny, to have the same vowels to minimise

the effect of the intrinsic f0 of vowels. A female native speaker of Standard Southern British English in her 20 s read the English sentence several times at a comfortable speaking rate with two accent 'peaks' and 'valleys' and also monotonously, trying to sound neutral without emphasising any particular part of the sentence. The speech was recorded at a sampling rate of 44.1 kHz using a Sennheiser MKH40 cardioid microphone (Wedemark, Germany) and a MixPre-6 digital recorder (Sound Devices, Reedsburg, USA). The microphone was positioned 20 cm from the speaker's mouth. Recording took place in a sound-attenuated booth in the Phonetics Laboratory, the University of Cambridge.

One monotonously spoken utterance was selected as the base for resynthesis. Ten points were identified for f0 stylisation (see Fig. 5), but the end of the plateau was aligned with the end of the stressed vowel, i.e., the plateau was stretched backward in time rather than being symmetrical about the vowel mid-point as in Experiment 1. This change was made from Experiment 1 because the f0 turn tends to be aligned towards the end of the vowel in Southern British English (Knight & Nolan, 2006). The f0 rise or fall time associated with each accent with a 25 ms plateau, i.e., duration between (2) and (3), (4) and (5), (6) and (8), (9) and (10), was controlled at 140 ms (Figs. 5 and 6). The duration control also allowed the f0 movement at the beginning of the first accent to occur after the voiceless interval in 'does'. When manually identifying (1)-(10), the points (4) and (9), which were the end of the stressed vowel were annotated first. Once these points were fixed, the points (3), (7) and (8) were determined as appropriate for desired plateau duration. Then the points (2) and (5) for the first accent and (6) and (10) for the second accent were marked to keep the duration under each accent curve at 140 ms.

There were two conditions for f0 Level (High vs. Low): for High, the f0 contours were within the range 200–302 Hz and for Low, 132–200 Hz. The first accent was always at the height of 0 ST Accent Height Difference, and the height of the second accent varied in five

one-semitone steps (Table 4). Consequently, the f0 range, the difference between the maximum f0 and the minimum f0, for each Level expressed in Hz was narrower for Low (68 Hz) than for High (102 Hz). The f0 baseline was at 200 Hz for Peaks, High Level and for Valleys, Low Level, which was close to the speaker's value near the beginning of the utterance (approximately 195 Hz). The baseline was at 132 Hz for Peaks, Low Level and at 302 Hz for Valleys, High Level. The root-mean-square amplitude of all stimuli was scaled to 70 dB and they were converted to mp3 files for an online experiment. Praat ver. 6.0.29 (Boersma & Weenink, 2017) was used for all sound editing and (re)synthesis procedures.

Fig. 5. The spectrogram and f0 track of the base utterance for resynthesis, spoken monotonously. The ten points for f0 stylisation, stressed vowels (V) and word boundaries are annotated.

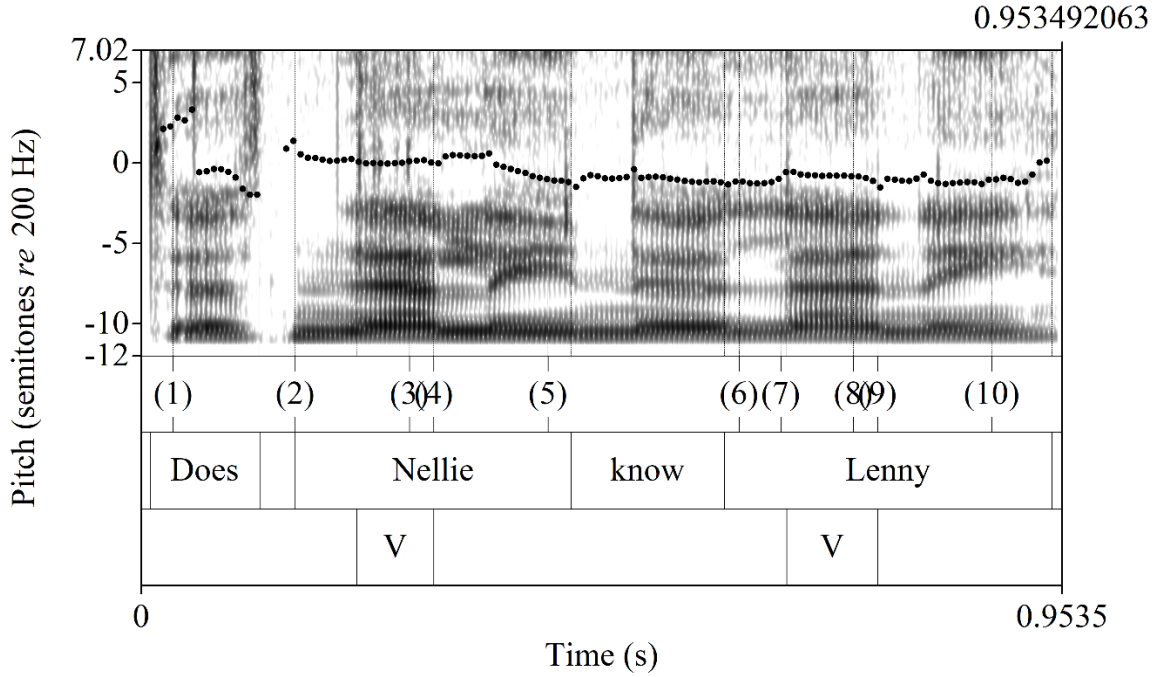


Fig. 6. Example f0 tracks for Peaks vs. Valleys in High vs. Low Levels. Both accents in an utterance have a 25 ms plateau in this figure (Experiment 2).

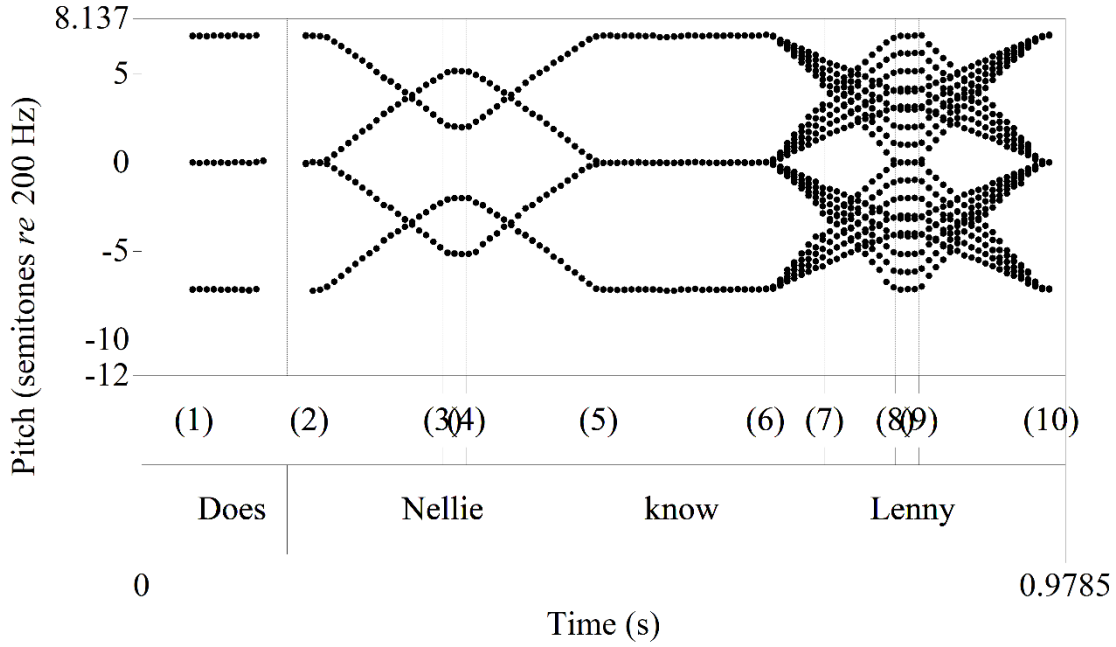


Table 4. The difference between the first accent height and the second accent height. The baseline (in brackets) refers to the f0 at the start of the utterance, at the stable stretch between the two accents, and at the end of the utterance (Experiment 2).

	Accent Height Difference	Hz	from baseline (ST)
Peaks			
High Level (200 Hz)	-2 ST	240	3.16
	-1 ST	254.27	4.16
	0 ST	269.39	5.16
	+1 ST	285.41	6.16
	+2 ST	302.38	7.16
Low Level (132.24 Hz)	-2 ST	158.72	3.16
	-1 ST	168.16	4.16
	0 ST	178.16	5.16
	+1 ST	188.75	6.16
	+2 ST	200	7.16
Valleys			
High Level (302.38 Hz)	-2 ST	251.93	-3.16
	-1 ST	237.79	-4.16
	0 ST	224.44	-5.16
	+1 ST	211.84	-6.16
	+2 ST	199.95	-7.16
Low Level (200 Hz)	-2 ST	166.63	-3.16
	-1 ST	157.28	-4.16
	0 ST	148.45	-5.16
	+1 ST	140.11	-6.16

	+2 ST	132.24	-7.16

4.1.2 Participants

Monolingual native speakers of British English who were born and living in England and aged between 18 and 30 were recruited through Prolific (www.prolific.co). None of the participants had self-reported speech, vision, hearing, or cognitive impairments. In total, 66 listeners participated in the experiment. Data from participants who failed the screening test were excluded (see Section 4.1.3) and data from 57 participants (37 female and 20 male) were analysed (age Mean = 24.26, SD = 3.71). The study was approved by the local ethics committee of the University of Central Lancashire (BAHSS2 0122).

4.1.3 Experimental procedure

The Gorilla Experiment Builder (Anwyl-Irvine et al., 2019) was used to create and host the experiment. Data were collected between 22 June and 11 October 2019. All participants were asked to use a desktop computer and headphones. Before the experiment, they filled in a consent form and questionnaires on their variety of English, gender, and age. They were then asked to wear headphones and adjust the volume to a comfortable level while a ten-second-long pure tone of 1000 Hz at 70 dB was played. Participants took an intensity discrimination task with twelve trials, which identified whether they were wearing headphones or listening in free-field (Woods et al., 2017). Only those wearing headphones could correctly identify the softest tone out of three in each trial and participants who were correct for fewer than ten trials were excluded. In addition, eight catch trials, which were simple mathematical operations with the correct answer either 1 or 2 (e.g., 4 - 3 = ?), were constructed to monitor whether participants were paying attention to the tasks. All participants provided correct answers to all catch trials and therefore no one was excluded on that basis.

The main experiment consisted of eight blocks (2 Accent Type × 2 Level × 2 Stimulus Type). The details of the complex tone stimuli and results can be found in Appendix B. Each block had five stimuli from the five Accent Height Difference conditions in the same Accent Type, Level, and Stimulus Type. Each stimulus was presented three times in each block. Within each block, the presentation order of the experimental and catch trials was randomised for each participant. The order of blocks was counterbalanced across the eight groups.

The instructions for participants were the same as in Experiment 1. There was a practice session before the main experiment. The practice session consisted of eight trials (2 Accent Type × 2 Level × 2 Accent Height Difference [-3 ST, +3 ST]) with the stimulus presentation order randomised for each participant. In the practice session, listeners were given feedback on their choice and they could repeat the practice session if they wanted. (Note that feedback was not provided in the practice session in Experiment 1 which was carried out in a laboratory. In Experiment 2 which was conducted online, participants did not have an opportunity to seek immediate clarification from the experimenter. Instead, feedback was provided to ensure that participants understood the experimental task.) No feedback was provided in the main experiment.[4]

The stimulus was automatically played 0.5 seconds after the onset of each trial. Participants could repeat the stimulus presentation as often as they liked, up to 20 times. In each trial, the question 'which one sounds higher?' or 'which one sounds lower?' appeared at the top of the screen as relevant to Accent Type, together with two buttons labelled 'first' and 'second'. Participants indicated their choice by clicking the appropriate button with a mouse.

[4] An anonymous reviewer commented that the feedback in Experiment 2 could have had a priming effect influencing listeners' performance. Although we cannot ascertain to what extent the methodological differences between the two experiments contributed to the results, the findings in both experiments are similar (discussed in Section 5), indicating that the results are replicable. For instance, the mean frequency (%) of listeners' 'second accent' responses was similar between the experiments, confirming that the methodological differences did not bias listeners in a particular way (Experiment 1, 'peaks', mean = 61.92, SD = 39.52, 'valleys', mean = 49.58, SD = 37.16; Experiment 2, 'peaks', mean = 58.16, SD = 40.49; 'valleys', mean = 46.99, SD = 36.37).

The experiment automatically progressed to the next trial when the participant pressed one of the response buttons. The experiment lasted approximately 30 minutes.

4.1.4 Analysis

There were in total 6,840 data points (2 Accent Type × 2 Level × 2 Second Accent Shape × 5 Accent Height Difference × 3 repetitions × 57 participants). The analysis methods were the same as in Experiment 1.

4.2 Results

As in Experiment 1, we present the response frequency (%) functions for all experimental conditions (Fig. 7) and the averaged response frequencies (Table 5).

Accent Type showed a significant effect ($p < 0.001$) in that listeners had a stronger bias to perceive the second accent as salient in pitch for Peaks (Mean = 58.16, SD = 40.49) than for Valleys (Mean = 46.99, SD = 36.37). Figure 7 shows that the response function for Peaks was shifted up compared to Valleys in all panels.

Second Accent Shape also had a statistically significant effect ($p < 0.001$) and it did not interact with Accent Type or Level. That is, the longer plateau had a consistent pitch-saliency-enhancing effect. In addition, the 100 ms plateau reduced the PSE compared to the 25 ms plateau in all Accent Types and Levels (Fig. 7). This means that in order to perceive the pitch of the two accents as equivalent, listeners required a smaller height difference between the accents for the 100 ms plateau than for the 25 ms plateau when the second accent was lower than the first in f0 for Peaks and higher for Valleys.

Level did not have a statistically significant effect.

The overall effect of Accent Height Difference ($p < 0.001$) as shown in the response functions (Fig. 7) was that listeners perceived the second accent more salient as the Accent Height Difference level increased.

There was no significant Accent Type × Second Accent Shape interaction effect. That is, the Peaks vs. Valleys asymmetry was not dependent on the plateau duration. However, it is worth noting that the PSEs show that listeners needed the larger excursion associated with the second accent for perceptual equivalence only for Valleys with the 25 ms plateau. When there was a 25 ms plateau, the PSEs were around 0 ST for Peaks for both Levels, but for Valleys, the PSE was close to 1 semitone. On the other hand, for the 100 ms plateau, the pitch-saliency-enhancing effect of the longer plateau led the PSE to be around -1 semitone for Peaks for both Levels. However, for Valleys, the PSEs were around 0 ST for both Levels.

Accent Type interacted with Level ($p < 0.001$). For Peaks, the 'second accent' response frequency was slightly higher for Low (Mean = 59.06, SD = 45.85) than for High Level (Mean = 57.25, SD = 48.13), whereas for Valleys, it was higher for High (Mean = 48.13, SD = 36.03) than for Low Level (Mean = 45.85, SD = 36.70). However, the difference in the response frequency between Peaks and Valleys for each Level was marginal and the response functions did not reveal any striking trends supporting the effect of Accent Type dependent on Level. The source of the interaction effect is probably the relatively flat response functions for the experimental condition combining Low Level, 25 ms Second Accent Shape, and Valleys. This interpretation is supported by the nearly significant three-way Accent Type × Second Accent Shape × Level interaction effect ($p = 0.06$). In Table 5, the response frequency for both Peaks (High, Mean = 63.63, SD = 40.23; Low, Mean = 64.33, SD = 38.69) and Valleys (High, Mean = 52.87, SD = 36.43; Low, Mean = 53.33, SD = 38.07) seems unaffected by Level for the 100 ms plateau. However, for the 25 ms plateau, for Peaks, listeners' bias towards the second accent saliency was slightly stronger for Low (Mean = 53.80, SD = 41.22) than for High Level

(Mean = 50.88, SD = 41.26). On the other hand, for Valleys, listeners' second accent saliency bias was reduced for Low (Mean = 38.36, SD = 33.71) compared to High Level (Mean = 43.39, SD = 35.05).

The significant Accent Type × Accent Height Difference interaction was shown by the steeper response functions, i.e., heightened perceptibility of the f0 change of the second accent, for Peaks than for Valleys. The significant Second Accent Shape × Accent Height Difference interaction suggests that a change of one semitone in the second accent height led to different consequences for the 25 ms and 100 ms plateaux. This interaction is probably ascribed to the flattened response function for +1 and +2 semitone Accent Height Differences for the 100 ms plateau. That is, the effect of a one-semitone increase between +1 and +2 Accent Height Difference was more pronounced for the 25 ms plateau than for the 100 plateau. In particular, the 'second accent' response frequency reached the ceiling for Peaks with a 100 ms plateau. Finally, the Level × Accent Height Difference interaction was not statistically significant.

To summarise the results, listeners showed heightened discrimination ability to 'peaks' than to 'valleys'. The long plateau (100 ms) had a consistent pitch-saliency-enhancing effect. The long plateau at 'peaks' could clearly lead listeners to perceive the second accent, which was physically higher in f0 than the first, as more salient than the first, as demonstrated by the ceiling effect. On the other hand, the ceiling effect was absent for 'valleys'. The accent type ('peaks' vs. 'valleys') interacted with the f0 level, because of listeners' markedly reduced pitch discrimination for 'valleys' when combined with a short plateau (25 ms) for the low level.

Table 5. Frequency (%) means and standard deviations of listeners' 'second accent' responses (Experiment 2).

		Level					
		High		Low		Total	
Shape	Type	Mean	SD	Mean	SD	Mean	SD

25 ms	Peaks	50.88	41.26	53.80	40.22	52.34	40.73
	Valleys	43.39	35.05	38.36	33.71	40.88	34.45
	Total	47.13	38.43	46.08	37.88	46.61	38.14
100 ms	Peaks	63.63	40.23	64.33	38.69	63.98	39.43
	Valleys	52.87	36.43	53.33	38.07	53.10	37.22
	Total	58.25	38.72	58.83	38.74	58.54	38.71
Total	Peaks	57.25	41.21	59.06	39.78	58.16	40.49
	Valleys	48.13	36.03	45.85	36.70	46.99	36.37
	Total	52.69	38.95	52.46	38.82	52.57	38.88

Fig. 7. The averaged frequency of 'second accent' responses (%) across all participants by Level, Second Accent Shape (25 ms vs. 100 ms plateaux) and Accent Type in Experiment 2.

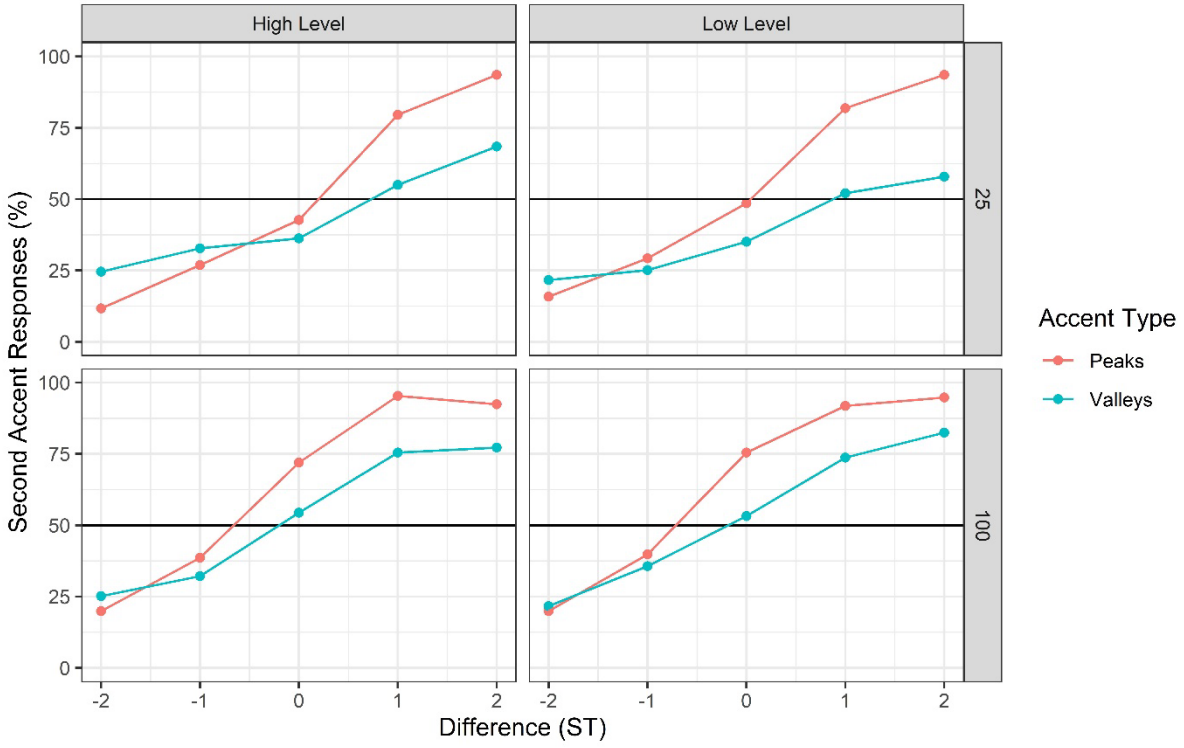


Table 6. Results of the model comparisons ($\alpha = 0.05$). The reference level for all models was Accent Type-Peaks, Second Accent Shape-25 ms and Level-High (Experiment 2).

	χ^2	df	p
Accent Type	115.17	1	p < 0.001 ***
Shape	131.16	1	p < 0.001 ***
Level	0.051	1	0.82
Difference	1742	1	p < 0.001 ***
Accent Type × Shape	0.03	1	0.85
Accent Type × Level	3.91	1	0.04*
Accent Type × Difference	159.59	1	< 0.001***
Shape × Level	0.62	1	0.43
Shape × Difference	12.93	1	< 0.001***
Level × Difference	0.07	1	0.8
Accent Type × Shape × Level	3.41	1	0.06
Accent Type × Shape × Difference	2.91	1	0.09
Shape × Level × Difference	1.43	1	0.23

4.3 Discussion of Experiment 2

The hypothesis about listeners' heightened discrimination for 'peaks' was supported (Hypothesis 1). We did not find strong evidence for the hypothesis that listeners' reduced capacity in perceiving 'valleys' would in turn reduce the pitch-saliency-enhancing effect of a long plateau compared to 'peaks' (Hypothesis 2). However, there was a nearly significant three-way interaction between the accent type, the second accent shape, and the f0 level in line with the hypothesis. A long plateau associated with the second accent could lead to listeners' 'second accent' responses reaching the ceiling for 'peaks' when the second accent was physically higher in f0 than the first by one or two semitones, but this was not the case for 'valleys'. For the 'valleys', the pitch-saliency-enhancing effect of a plateau seemed to be constrained by listeners' reduced capacity to perceive variation in f0. In addition, the PSEs were around 0 semitones for the 'valleys' with a long plateau, in contrast to the negative PSEs that were observed for 'peaks' (Fig. 7). This indicates that listeners needed the long plateau in the second accent in order to perceive pitch equivalence between two 'valleys' with the same minimum f0.

Finally, as hypothesised, the f0 level interacted with the accent type (Hypothesis 4). The interaction was demonstrated for a short plateau. The response function for 'valleys' in the low level was notably flat when the second accent formed a 25 ms plateau. This indicates that listeners' capacity to perceive f0 changes was extremely reduced in the experimental condition for the 'valleys' combined with the 25 ms plateau and the low f0 level. Similar results were found in the complex tone trials (Appendix B).

5. General discussion

The plethora of existing research on the perception of intonation has a strong focus on pitch rises or high accents, while falls or low accents have been neglected. As we endeavour to understand the relationship between the acoustic forms and linguistic interpretations, examining listeners' perception of various pitch shapes at the pre-linguistic level will inform the methodological choices. The present study offers a useful basis for further investigation on how to model intonation taking account of the perception and also on how listeners in different backgrounds, such as foreign language speakers or those with hearing loss, process intonation differently.

In two experiments, the auditory stimuli were question utterances with either two 'peak' or two 'valley' accents aligned to the stressed syllable. The f0 properties associated with the first accent were kept constant throughout the experiment and those of the second accent were varied in f0 contour shape and height. Listeners carried out a two-alternative forced-choice task judging whether the first or the second accent was more salient in pitch ('higher' for 'peaks' and 'lower' for 'valleys'). The judgment of relative prominence would be a more linguistically relevant task than the judgment of pitch height. However, the prominence judgment carries a number of potential problems including the risk that listeners might judge the word prominence rather than the local pitch properties of a vowel or syllable, and the fact

that individual listeners rely on different cues for prominence judgement (Turnbull et al., 2017; Baumann & Winter, 2018). By using psychoacoustic rather than linguistic instructions we hoped to reduce between-listener variance.

The two experiments differed in the stimulus and experimental design. In Experiment 1, the first accent in the stimulus always formed a sharp turn, while the second accent shape varied between forming a sharp turn and a 100 ms plateau. The f0 turn or plateau was aligned to the mid-point of the stressed vowel. The stimuli were based on four English utterances ('Lemmy', 'Nellie', 'Mona', and 'Nina', see Section 3.1.2). For Experiment 2, all auditory stimuli were resynthesised from one utterance. We approximated the acoustic properties of the stimuli to those in real speech more closely than in Experiment 1, by aligning the f0 accent to the end of the stressed vowel and replacing the sharp f0 turn used in Experiment 1 with a 25 ms plateau. The plateau duration for the second accent was varied (25 ms vs. 100 ms). The experiment was conducted online.

Our findings are robust given that similar results were observed between the two experiments. To summarise the findings, listeners exhibited heightened discrimination for 'peaks' compared to 'valleys' (Hypothesis 1). Listeners' reduced discrimination for 'valleys' constrained the plateau effect which enhances pitch saliency in general (Hypothesis 2). One difference in the results was that the interaction between the accent shape and the accent height difference was found only in Experiment 2. This is probably because in Experiment 2, the variation associated with multiple utterances was removed. The long plateau led listeners' 'second accent' responses to ceiling when the second accent was physically higher in f0 than the first for 'peaks' (Fig. 7), while this effect was not consistent across the utterances in Experiment 1 (Fig. 4). It is also possible that the relatively small size of the response data set (2,400 data points from 20 participants) in Experiment 1 compared to that in Experiment 2 (6,840 data points from 57 participants) was not sufficient for the interaction effect to emerge.

The hypothesised utterance type ('item') effect was observed (Hypothesis 3), but the effects of the accent type ('peaks' vs. 'valleys') and shape of the f0 turn (i.e., plateau duration) were generalised across the items. The interaction effects involving the item are difficult to account for, because the four utterances differed in segmental composition, loudness contours and the overall spectral properties which affect perceived pitch. In manipulating f0, the mid-points of the accent and the stressed vowel were aligned, and as a consequence, different portions of the rise or fall were allocated relative to the segmental string across the items (see House, 1996 and Barnes, et al, 2021**,** for discussion on the perceptibility of f0 movements occurring in high sonority regions). For instance, for 'does Mona know Nina? ('Mona')', particularly for 'valleys', listeners were notably biased for the first accent saliency. This may be because the stressed syllable in 'Mona' had higher acoustic energy than that in 'Nina', and also because of the intrinsic f0 of vowels. The stressed vowel [i] in 'Nina' may have sounded higher than [o] in 'Mona' when they were equal in f0, because of the higher intrinsic f0 of [i] (e.g., Lehiste, 1970, section 3.4.1; Whalen & Levitt, 1995). Finally, the hypothesis on the interaction between f0 level (f0 events occurring at 200–302 Hz vs. 132–200 Hz) and accent type ('peaks' vs. 'valleys') was supported in that listeners' discrimination was significantly reduced for 'valleys' at the low level (Hypothesis 4).

5.1 Perceptual asymmetry between 'peaks' and 'valleys'

The experiments established that listeners' heightened capacity to perceive f0 'peaks' compared to 'valleys' is largely due to the direction of f0 movement, the rise preceding the fall in a 'peak'. The perceptual disadvantage of falls in f0 observed with short stimuli or complex tones in previous studies (see Section 1) is not merely a local effect of an auditory stimulus presented in isolation, but the disadvantage persists when the fall forms the early part of an accent in an utterance context. The asymmetry seems to be a psychoacoustic effect, as shown by similar results from reiterated speech and complex tone trials (Appendices A, B). The

auditory mechanism responsible for the asymmetry is out of the scope of our discussion, but it may be linked to how the pitch movements are linguistically coded and tendencies in speech production (see Section 5.3).

Although f0 valleys and low pitch accents are present in speech cross-linguistically, their acoustic shape tends to be asymmetrical to those of peaks and high accents. For instance, low pitch accents tend to be realised as a long plateau (see Asu & Nolan, 2007 and references therein), and this could be due to both articulatory 'ease' and perceptual need. High pitch is produced by speakers' forceful articulation involving muscular tensioning and increased rates of airflow through the vocal folds (Lieberman et al., 1969; Baer, 1979; Titze, 1989; Alipour & Scherer, 2007), whereas pitch falls are less effortful to produce. Meanwhile, the long plateau in the low accent may allow listeners to compensate for the intrinsic perceptual challenges of the falling or low pitch (see Baumann, 2014, for discussion on longer duration for syllables with low than high accents in German).

The perceptual asymmetry could also play a role in categorising pitch accents into low prominence (L*) or high prominence (H*) in such a way that the same degree of f0 excursion creates less ambiguity for H* than for L*. Theoretically, the difference between H* and L* could be about which tonal target falls on the stressed syllable (cf. Pierrehumbert, 1983). For instance, for a f0 rise followed by a fall, H* indicates that a high target is located in the stressed syllable, while L* indicates that a low target is located in the stressed syllable (H+L*). However, the location of the f0 turning point of a 'peak' or a 'valley' is not an unambiguous cue (see Ladd, 2008, Sections 2.2 and 5.1). In our stimuli, the 'peaks' were probably unambiguously identified as high prominence (H*), but the 'valleys' could have created ambiguity when the height difference between the two accents was not clearly perceivable with a short plateau, which could potentially be interpreted as H* preceded by a low tone (i.e., L+H*) rather than L*. The categorical ambiguity of a 'valley' is probably not specific to our

stimuli. Even when the f0 minimum in a valley-shaped contour is located within a stressed syllable, the accent may still be identified as H* depending on the precise alignment of the adjacent peak and the slope of the f0 rise. For example, in 'there's an a*nó*maly in it' (in Barnes, Brugos, Shattuck-Hufnagel & Veilleux, 2012), even when the local minimum of a rise-fall contour is in the stressed vowel, the f0 peak occurring in the stressed syllable or in the early portion of the following syllable (e.g., *nó* or *ma* in 'a*nó*maly') leads to a L+H* percept, while for L*+H the f0 peak would be aligned later. For unambiguous categorisation of an f0 'valley' as L*, the minimum f0 probably needs to be close to the bottom of the speaker's voice range, the f0 minimum might need to form a long plateau, and/or there should be a steep f0 rise from the local minimum (see Dilley & Heffner, 2013 on the complications in distinguishing H* from H+L*). Indeed, the low prominence seems to be somewhat unnatural and hard to process. For instance, in Zahner, et al. (2019), native German speakers were more accurate in judging the location of lexical stress when the stressed syllable had high pitch than when the peak preceded or followed the stressed syllable with low pitch. Low accents (e.g., H+L*, which is formed by a valley-shaped f0 contour with the f0 minimum associated with the stressed syllable, or L* followed by a high boundary tone) inhibited lexical processing while high f0 aligned to the stressed syllable had advantages, and the rise leading to a low pitch accent in H+L* tended to be mistaken as high prominence. On the other hand, high peaks seem to be perceptually salient, potentially serving as a cue to word segmentation in the language acquisition process (Zahner, et al., 2016; Zahner & Brown, 2018).

5.2 Pitch scales in quantifying 'peaks' and 'valleys'

It is unlikely that the perceptual asymmetry between f0 rises and falls is an artefact of our choice of the semitone scale, as the asymmetry has been reported in studies using the Hertz

scale. For instance, Kishon-Rabin et al. (2004) showed that when the reference frequency of a pure tone was at 200 Hz, the pitch discrimination threshold was smaller for an 'increments' paradigm (threshold mean = 3.23 Hz) where the comparison tone increased in f0 by 0.5 Hz in the range of 200.5–210 Hz compared to the 'decrements' paradigm where the f0 of the comparison tone was lowered in 0.5 Hz steps in the range of 190–199.5 Hz (threshold mean = 4.09 Hz). Turner et al. (2019) also showed that listeners achieved higher accuracy and a lower frequency threshold for categorising speech stimuli into rises (just noticeable difference at 35 Hz) compared to falls (just noticeable difference at 40 Hz) when the stimulus onset was at 250 or 350 Hz with the magnitude of change in 5–50 Hz.

The logarithmic semitone scale is commonly used for normalising between-speaker variation in production data (Graddol, 1986; Traunmüller & Eriksson, 1995; Nolan, 2003; Carlson et al., 2004). We intended to implement an equivalent magnitude of f0 change between a rise and a fall, and also across f0 levels, as the change of, for instance, three semitones moving either upwards or downwards forms a musical interval of equal size. The semitone, which is useful in expressing perceptual equivalence in the melodic contour when transposed to different keys, was deemed a reasonable choice. However, the 'peaks' vs. 'valleys' asymmetry would be diminished if the relatively linear Hertz or ERB-rates (equivalent rectangular bandwidth rates) scales were to be used. In the frequency range relevant to human speech (below 500 Hz), the semitone scale is logarithmic, the Hertz scale is linear, and the ERB-rate scale (Patterson, 1976) is in between (see Nolan, 2003 for comparisons).

Our results imply that the semitone scale does not result in equal salience in perceived pitch for falls vs. rises or peaks vs. valleys. When a constant baseline for height comparisons is given, as in the present study, a (close-to) linear scale might be a fairer choice in addressing the perceived equivalence. For instance, Rietveld & Gussenhoven (1985) reported that Hertz may be better than semitones for measuring excursion size in relation to accent strength. In Hermes

& van Gestel (1991), the excursion sizes of the accents in the low and high level were equal when expressed on the ERB-rate. When expressed in semitones, the excursion size was overestimated in the high level but underestimated in the low level, while the opposite problem was noted when it was expressed in Hertz (cf. Terken & Hermes, 2000, who reviewed aforementioned studies and favoured the ERB-rate).

Variation in f0 levels did not have an overarching effect on listeners' pitch height judgements in general. Our results were that a change of one semitone had perceptually similar consequences across the f0 levels (high, 200–302 Hz, range 102 Hz = 7.13 semitones; low, 132–200 Hz, range 68 Hz = 7.19 semitones) other things being equal. However, f0 level interacted with accent type ('peaks' vs. 'valleys'). As long as the f0 accent formed 'peaks', listeners' capacity to perceive changes in one-semitone steps was not affected by variation in the f0 level. The main source of the interaction seemed to be that the 'valleys' occurring at a low frequency was not perceptually weighted as much as the same magnitude of f0 changes in other conditions.

For the 'valleys' in the low-level condition, the results may be partially a by-product of our choice of the semitone scale. The one-semitone step (Table 4) was only 8–9 Hz for the 'valleys' in the low level, while it was equivalent to 12–14 Hz for the 'valleys' in the high level, 9–11 Hz for the 'peaks' in the low level, and 14–17 Hz for the 'peaks' in the high level. In other words, the f0 change on the linear Hertz scale for the 'valleys' in the low level was less perceivable compared to the other conditions. However, as discussed above, using a linear scale for f0 manipulation is unlikely to remove the perceptual challenges altogether.

There are further implications of the choice of pitch scale on perceptual modelling of intonation and the f0 level variation. As discussed in Section 1, the identification of f0 turning points associated with the tonal target is a crucial step in the acoustic analysis of intonation (see

Ladd, 2008, Chapters 2 and 5). Then the saliency-enhancing effect of a plateau raises a question about the identity of the target – the 'plateau' effect suggests that the precise shape of f0 movement affects the perceived pitch. Rather than the scaling and alignment of the f0 maximum or minimum point, the overall acoustic prominence of the syllable may be a better correlate of the speaker's intended target. This issue is not new and alternative models of pitch accent prominence have been proposed, such as Segerup & Nolan's (2006) Pitch Integral, and Barnes, Veilleux, Brugos, & Shattuck-Hufnagel's (2012) Tonal Centre of Gravity. Both integrate f0 and temporal information to calibrate the overall acoustic prominence, but they are based on the 'area under the f0 curve' for peak-shaped accents. Given the perceptual asymmetry between 'peaks' and 'valleys', further consideration should be given in applying these models to low accents.

For demarcating a plateau-shaped peak, for instance, House et al. (1999) and Knight & Nolan (2006) suggest including all the contours within 4% or 6% of the maximum frequency in Hertz as this approximates the threshold of perceptual equivalence ('t Hart, 1981; Rosen & Fourcin, 1986). Using these criteria, the 4% cut-off for a peak with a maximum f0 at 200 Hz is 8 Hz, and the 6% cut-off value is 12 Hz. However, for a low pitch accent with a minimum f0 at 150 Hz, the 4% cut-off value is 6 Hz, and the 6% cut-off value is 9 Hz. That is, using this percentage criterion underestimates the plateau duration and perceived prominence of a low accent. Our results (Fig. 7) showed that when there was a 100 ms plateau at the f0 turn for a 'peak', then the second accent had a smaller excursion than the first accent when the two accents were perceived as equivalent in pitch height. However, for a 'valley', listeners needed the long plateau to perceive the second accent to be as low as the first for the two accents that were equivalent in physical f0.

Finally, our results confirm that the pitch discrimination or processing difficulties found at particular f0 levels are direct consequences of the acoustic properties of the auditory stimuli.

In previous studies, the difficulties tended to be ascribed to the unnaturalness of the auditory stimuli. For instance, Rietveld & Gussenhoven's (1985) study on the relationship between the f0 excursion size and perceived accent prominence used stimuli which resynthesised a female voice into a low male-voice pitch range, and this could have caused some unnaturalness of the stimuli. In Hermes & van Gestel's (1991) study on the perception of f0 excursion at different levels, they used a falsetto voice in the high level (see Section 5.1 for a summary of these studies). In our stimuli, the female voice in the low-level condition was not within the typical range of a female voice but close to that of male one (Fairbanks, 1940; Peterson & Barney, 1952). However, similar trends found with complex tone stimuli (Appendix B) allow us to rule out the explanation based on listeners' expectations of speaker gender and other speech-specific properties.

5.3 Future directions

So far, we have discussed the perceptual advantages of the f0 peak and rises in the context of pitch height discrimination, categorisation, and processing of pitch accent types. The source of the peak vs. valley asymmetry in the human voice frequency range could be a general property of human auditory mechanism or an outcome of continuous exposure to high prominence in speech. Corpus-based studies suggest that high prominence is more common than low prominence (Grabe, 2004, for British English, and Dainora, 2006, for news readers' speech in American English), and the bias towards high prominence may be learned and enhanced. In Kutscheid et al. (2021) where native German speakers identified either low or high prominence in a word or sentence after 3 minutes' exposure to utterances with only high accents (L+H*) or low accents (L*+H, L*, or H+L*), listeners who were exposed to high accents were more likely to identify high accent with prominence than those who were exposed to low accents.

If the association between high information load and high or rising pitch forms part of the top-down knowledge of listeners, the strength of such association, which presumably differs across languages, may have a direct influence on listeners' pitch-processing capacity. Although there is extensive literature on differences in pitch processing between speakers of tonal and non-tonal languages (see Maggu et al., 2018 and references therein), the question of whether the mapping between particular pitch contours and high-information sites in languages affects pitch perception is yet to be investigated. One fruitful avenue for further research is to examine the perceptions of listeners whose native language has frequent low pitch accents, such as Donegal Irish (Dalton & Chasaide, 2005) and Glaswegian English (Smith & Rathcke, 2020).

6. Conclusions

Two experiments established that listeners' pitch discrimination is reduced for f0 'valleys' compared to 'peaks' when judging the height of two 'accents' within an utterance. The f0 turn at the maximum of a 'peak' or the minimum of a 'valley' forming a long plateau increased the perceived pitch saliency compared to a sharp turn or a shorter plateau, making a 'peak' sound higher and a 'valley' lower. However, this effect was constrained by pitch perceptibility; the pitch-saliency-enhancing effect of a long plateau was reduced for 'valleys'. Listeners' discrimination for the 'peaks' was high regardless of the f0 level (low, 132–200 Hz; high, 200–302 Hz), but 'valleys' in the low level posed challenges. In addition, the relationship between f0 and perceived pitch may be linear rather than logarithmic in the human voice range, when listeners judge height from the same baseline. The effects related to the 'peaks' vs. 'valleys' contrast and the f0 level are not specific to human speech, as similar results were observed with non-speech stimuli. Listeners' insensitivity to 'valleys' in the low level may lead listeners to drop their attention in speech processing, while the effect of the listeners' native language deserves further investigation.

Acknowledgements

This work was supported by the Modern Languages and Linguistics research funds 2017/18 and the Culture and the Creative Industries Faculty Research Grants 2018/2019 at the University of Central Lancashire. Antje Heinrich was supported by the NIHR Manchester Biomedical Research Centre. Portions of this work were presented at the Speech in Noise Workshop, Ghent, Belgium, 10–11 January 2019, the British Society of Audiology, Basic Auditory Science meeting, Newcastle University, UK, 3–4 September 2018 and the LabPhon17 (virtual conference), 6–8 July 2020, the University of British Columbia and Simon Fraser University, Canada. We thank Sarah Knight for having recorded her speech for us to create experimental stimuli for Experiment 1. Last but not least, we would like to thank Francis Nolan and Bob Ladd for useful discussion in designing the present project.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 19: 716–723, doi:10.1109/tac.1974. 1100705.

Alipour, F. and Scherer, R. C. (2007). On pressure-frequency relations in the excised larynx. *The Journal of the Acoustical Society of America* 122: 2296–2305, doi:10.1121/1.2772230.

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N. and Evershed, J. K. (2019). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods* 52: 388–407, doi:10.3758/s13428-019-01237-x.

Astruc, L., Payne, E., Post, B., Mar Vanrell, M. del and Prieto, P. (2012). Tonal targets in early child English, Spanish, and Catalan. *Language and Speech* 56: 229–253, doi:10.1177/0023830912460494.

Asu, E. L. and Nolan, F. (2007). The analysis of low accentuation in Estonian. *Language and Speech* 50: 567–588, doi:10.1177/00238309070500040401.

Baer, T. (1979). Reflex activation of laryngeal muscles by sudden induced subglottal pressure changes. *The Journal of the Acoustical Society of America* 65: 1271–1275, doi:10.1121/1.382795.

Barnes, J., Brugos, A., Veilleux, N., & Shattuck-Hufnagel, S. (2011). Voiceless intervals and perceptual completion in F0 contours: Evidence from scaling perception in American English. *International Congress of Phonetic Sciences 17*, 17-21 August 2011, Hong Kong, pp. 108- 111.

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

Barnes, J., Brugos, A., Veilleux, N., Hufnagel, S.S. (2014) Segmental Influences on the Perception of Pitch Accent Scaling in English. *Proc. 7th International Conference on Speech Prosody 2014*, 20-23 May 2014, Dublin, Ireland, pp. 1125-1129, doi: 10.21437/SpeechProsody.2014-214.

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. *Journal of Phonetics* 85, 101020. doi: 10.1016/j.wocn.2020.101020.

Barnes, J., Veilleux, N., Brugos, A. and Shattuck-Hufnagel, S. (2012). Tonal center of gravity: A global approach to tonal implementation in a level-based intonational phonology. *Laboratory Phonology* 3, doi:10.1515/ lp-2012-0017.

Barr, D. J., Levy, R., Scheepers, C. and Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language* 68: 255–278, doi:10.1016/j.jml.2012.11.001.

Bates, D., Mächler, M., Bolker, B. and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67: 1–48, doi:10.18637/jss.v067.i01.

Baumann, S. (2014). The importance of tonal cues for untrained listeners in judging prominence. In the *Proceedings of the 10th International Seminar on Speech Production (ISSP)*, pp. 21–24. Cologne, Germany, 5-8 May 2014.

Baumann, S., & Röhr, C. (2015). The perceptual prominence of pitch accent types in German. In the *Proceedings of the 18th International Congress of Phonetic Sciences (ICPhS XVIII)*, paper 298, pp. 1-5, Glasgow, UK, 10-14 August 2015.

Baumann, S. and Winter, B. (2018). What makes a word prominent? Predicting untrained German listeners' perceptual judgments. *Journal of Phonetics* 70: 20–38, doi:10.1016/j.wocn.2018.05.004.

Binns, C. and Culling, J. F. (2007). The role of fundamental frequency contours in the perception of speech against interfering speech. *The Journal of the Acoustical Society of America* 122: 1765–1776, doi:10.1121/1. 2751394.

Boersma, P. and Weenink, D. (2000). Praat: doing phonetics by computer [computer program]. available at http://www.praat.org/.

Cangemi, F., Albert, A., & Grice, M. (2019). Modelling intonation: Beyond segments and tonal targets. In *Proceedings of the International Congress of Phonetic Sciences* (ICPhS 2019), Melbourne, Australia, pp. 572–576.

Carlson, R., Elenius, K. and Swerts, M. (2004). Perceptual judgments of pitch range. In *Speech Prosody*. Nara, Jaoan, 689–692.

Clopper, C. G. (2013). Modeling multi-level factors using linear mixed effects. In *Proceedings of Meetings on Acoustics*, 19. ASA, 060028, doi:10.1121/1. 4799729.

Collier, R. (1987). F0 declination: The control of its setting, resetting, and slope. In Baer, T., Sasaki, C. and Harris, K. S. (eds), *Laryngeal Function in Phonation and Respiration*. Boston: College Hill, Little, Brown and Company, 403–421.

Collins, M. J. and Cullen, J. K. (1978). Temporal integration of tone glides. *The Journal of the Acoustical Society of America* 63: 469–473, doi:10. 1121/1.381738.

Dainora, A. (2006). Modeling intonation in English: A probabilistic approach to phonological competence. In L. Goldstein, D. Whalen, & C. Best, *Laboratory Phonology 8* (pp. 107–132). doi: 10.1515/9783110197211.1.107.

Dalton, M. and Chasaide, A. N. (2005). Tonal alignment in Irish dialects. *Language and Speech* 48: 441–464, doi:10.1177/00238309050480040501.

Dawson, C., Aalto, D., Simko, J. and Vainio, M. (2017). The influence of fundamental frequency on perceived duration in spectrally comparable sounds. *PeerJ* 5: e3734, doi:10.7717/peerj.3734.

Dilley, L. C. and Heffner, C. C. (2013). The role of f0 alignment in distinguishing intonation categories: Evidence from American English. *Journal of Speech Sciences* 3: 3–67.

Evans, J. P. (2015). High is not just the opposite of low. *Journal of Phonetics* 51: 1–5, doi:10.1016/j.wocn.2015.05.001.

Fairbanks, G. (1940). Recent experimental investigations of vocal pitch in speech. *The Journal of the Acoustical Society of America* 11: 457–466, doi:10.1121/1.1916060.

Fletcher, H. and Munson, W. A. (1933). Loudness, its definition, measurement and calculation. *The Journal of the Acoustical Society of America* 5: 82–108, doi:10.1121/1.1915637.

Fry, D. B. (1958). Experiments in the perception of stress. *Language and Speech* 1: 126–152.

Gordon, M. and Poeppel, D. (2002). Inequality in identification of direction of frequency change (up vs. down) for rapid frequency modulated sweeps. *Acoustics Research Letters Online* 3: 29–34, doi:10.1121/1.1429653.

Grabe, E. (2004). Intonational variation in urban dialects of English spoken in the British Isles. In P. Gilles & J. Peters, *Regional Variation in Intonation* (pp. 9–31). Tübingen, Germany: Niemeyer.

Graddol, D. (1986). Discourse specific pitch behavior. In John-Lewis, C. (ed.), *Intonation in Discourse*. London: Routledge, 221–237.

Green, T., Faulkner, A. and Rosen, S. (2004). Enhancing temporal cues to voice pitch in continuous interleaved sampling cochlear implants. *The Journal of the Acoustical Society of America* 116: 2298–2310, doi:10.1121/ 1.1785611.

Grice, M. (1995). *The Intonation of Interrogation in Palermo Italian*. De Gruyter, doi:10.1515/9783110932454.

Grice, M., Baumann, S. and Benzmüller, R. (2005). German intonation in autosegmental-metrical phonology. In *Prosodic Typology*. Oxford, UK: Oxford University Press, 55–83, doi:10.1093/acprof:oso/9780199249633.003.0003.

Gussenhoven, C. (2004). The Phonology of Tone and Intonation. Cambridge: UK, Cambridge University Press.

Gussenhoven, C. (2005). Transcription of Dutch intonation. In *Prosodic Typology*. Oxford, UK: Oxford University Press, 118–145, doi:10.1093/acprof:oso/ 9780199249633.003.0005.

Gussenhoven, C., B. H. Repp, A. R., Rump, W. H. and Terken, J. (1997). The perceptual prominence of fundamental frequency peaks. *Journal of the Acoustical Society of America* 102: 3009–3022, doi:10.1121/1.420355.

Gussenhoven, C., and Rietveld, T. (1998). On the speaker-dependence of the perceived prominence of f0 peaks. *Journal of Phonetics* 26: 371–380, doi:10.1006/jpho.1998.0080.

Gussenhoven, C., and Rietveld, T. (2000). The behavior of H ∗and L∗ under variations in pitch range in Dutch rising contours. *Language and Speech, 43*(2): 183–203. doi: 10.1177/00238309000430020301.

Gussenhoven, C., and Zhou, W. (2013). Revisiting pitch slope and height effects on perceived duration. In C. F. F. Bimbot C. Cerisara, the *Proceedings of INTERSPEECH 2013: 14th Annual Conference of the International Speech Communication Association* (pp. 1365–1369). Lyon, France.

't Hart, J. (1981). Differential sensitivity to pitch distance, particularly in speech. *The Journal of the Acoustical Society of America* 69: 811–821, doi:10.1121/1.385592.

't Hart, J. (1991). F0 stylization in speech: Straight lines versus parabolas. *The Journal of the Acoustical Society of America* 90: 3368–3370, doi:10. 1121/1.401396.

't Hart, J. Collier, R. and Cohen, A. (1990). *A Perceptual Study of Intonation: An Experimental-Phonetic Approach to Speech Melody*. Cambridge, UK: Cambridge University Press.

Heeren, W., Coene, M., Vaerenberg, B., Avram, A., Cardinaletti, A., Bo, L. del, Pascu, A., Volpato, F. and Govaerts, P. J. (2012). Development of the AE test battery for assessment of pitch perception in speech. *Cochlear Implants International* 13: 206–219, doi:10.1179/1754762811y.0000000035.

Hermes, D. J. and van Gestel, J. C. (1991). The frequency scale of speech intonation. *The Journal of the Acoustical Society of America* 90: 97–102, doi:10.1121/1.402397.

House, D. (1996). Differential perception of tonal contours through the syllable. In the *Proceedings of Fourth International Conference on Spoken Language Processing (ICSLP 96)* pp. 2048–2051, PA, USA, 3-6 October. doi: 10.1109/icslp.1996.607203.

House, J., Dankovieová, J. and Huckvale, M. (1999). Intonation modelling in ProSynth: An integrated prosodic approach to speech synthesis. In the *Proceedings of the International Congress of Phonetic Sciences*. San Francisco, 2343–2346.

Hsu, C.-H., Evans, J. P. and Lee, C.-Y. (2015). Brain responses to spoken f0 changes: Is H special? *Journal of Phonetics* 51: 82–92, doi:10.1016/j. wocn.2015.02.003.

ISO (2003). 226. *Acoustics – Normal Equal-loudness Contours*. Geneva: International Organization for Standardization.

Jeon, J. Y. and Fricke, F. R. (1997). Duration of perceived and performed sounds. *Psychology of Music* 25: 70–83, doi:10.1177/0305735697251006.

Kishon-Rabin, L., Roth, D. A.-E., Dijk, B. V., Yinon, T. and Amir, O. (2004). Frequency discrimination thresholds: the effect of increment versus decrement detection of frequency. *Journal of Basic and Clinical Physiology and Pharmacology* 15, doi:10.1515/jbcpp.2004.15.1-2.29.

Klatt, D. H. (1973). Discrimination of fundamental frequency contours in synthetic speech: Implications for models of pitch perception. *The Journal of the Acoustical Society of America* 53: 8–16, doi:10.1121/1.1913333.

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

Knight, R. A. and Nolan, F. (2006). The effect of pitch span on intonational plateaux. *Journal of the International Phonetic Association* 36: 21–38.

Kohler, K. J. (2008). The perception of prominence patterns. *Phonetica* 65: 257–269, doi:10.1159/000192795.

Kutscheid, S., Zahner-Ritter, K., Leemann, A., & Braun, B. (2021). How prior experience with pitch accents shapes the perception of word and sentence stress. *Language, Cognition and Neuroscience*, 1–17. doi: 10.1080/23273798.2021.1946109.

Ladd, D. R. (2008). *Intonational Phonology*. Cambridge, UK: Cambridge University Press, 2nd ed.

Ladd, D., Schepman, A., White, L., Quarmby, L. M. and Stackhouse, R. (2009). Structural and dialectal effects on pitch peak alignment in two varieties of British English. *Journal of Phonetics* 37: 145–161, doi:10.1016/ j.wocn.2008.11.001.

Lehiste, I. (1970). *Suprasegmentals*. Cambridge, MA: MIT Press.

Lehiste, L. (1976). Influence of fundamental frequency pattern on the perception of duration. *Journal of Phonetics* 4: 113–117.

Lieberman, P., Knudson, R. and Mead, J. (1969). Determination of the rate of change of fundamental frequency with respect to subglottal air pressure during sustained phonation. *The Journal of the Acoustical Society of America* 45: 1537–1543, doi:10.1121/1.1911635.

Maggu, A. R., Wong, P. C., Antoniou, M., Bones, O., Liu, H. and Wong, F. C. (2018). Effects of combination of linguistic and musical pitch experience on subcortical pitch encoding. *Journal of Neurolinguistics* 47: 145–155, doi:10.1016/j.jneuroling.2018.05.003.

McPherson, M. J. and McDermott, J. H. (2017). Diversity in pitch perception revealed by task dependence. *Nature Human Behaviour* 2: 52–66, doi:10. 1038/s41562-017-0261-8.

Miller, S. E., Schlauch, R. S. and Watson, P. J. (2010). The effects of fundamental frequency contour manipulations on speech intelligibility in background noise. *The Journal of the Acoustical Society of America* 128: 435–443, doi:10.1121/1.3397384.

Moore, B. C. J. (2008). Basic auditory processes involved in the analysis of speech sounds. *Philosophical Transactions of the Royal Society B: Biological Sciences* 363: 947–963, doi:10.1098/rstb.2007.2152.

Niebuhr, O. and Winkler, J. (2017). The relative cueing power of f0 and duration in German prominence perception. In *Interspeech 2017*. ISCA, doi:10.21437/interspeech.2017-375.

Nolan, F. (2003). Intonational equivalence: An experimental evaluation of pitch scales. In Solé, M. J., Recasens, D. and Romero, J. (eds), In the *Proceedings of the 15th International Congress of Phonetic Sciences*. Barcelona, Spain, 771 – 774, August 3-9 2003.

Patterson, R. D. (1976). Auditory filter shapes derived with noise stimuli. *The Journal of the Acoustical Society of America* 59: 640–654, doi:10. 1121/1.380914.

Patterson, D., and Ladd, D. R. (1999). Pitch range modeling: Linguistic dimensions of variation. *Proceedings of the XIVth International Congress of Phonetic Sciences*, pp. 1169–1172. Berkeley: University of California.

Peterson, G. E. and Barney, H. L. (1952). Control methods used in a study of the vowels. *The Journal of the Acoustical Society of America* 24: 175–184, doi:10.1121/1.1906875.

Pierrehumbert, J. (1979). The perception of fundamental frequency declination. *The Journal of the Acoustical Society of America* 66: 363–369, doi:10.1121/1.383670.

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

Pierrehumbert, J. B. (1983). Automatic recognition of intonation patterns. In the *Proceedings of the 21st annual meeting on Association for Computational Linguistics*, pp. 85–90, Cambridge, Massachusetts, USA, doi: 10.3115/981311.981328.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rietveld, A. C. and Gussenhoven, C. (1985). On the relation between pitch excursion size and prominence. *Journal of Phonetics* 13: 299–308.

Rosen, S. (1977). The effect of fundamental frequency patterns on perceived duration. *Quarterly Progress and Status Report* 18: 17–30, KTH Royal Institute of Technology, Stockholm.

Rosen, S. and Fourcin, A. (1986). Frequency selectivity and the perception of speech. In Moore, B. (ed.), *Frequency Selectivity in Hearing*. London: Academic Press, 373–488.

Šimko, J., Aalto, D., Lippus, P., Włodarczak, M., & Vainio, M. (2015). Pitch, perceived duration and auditory biases: Comparison among languages. In the Proceedings of the 18th International Congress of Phonetic Sciences (pp. 0575.1-5). University of Glasgow, UK.

Segerup, M. and Nolan, F. (2006). Gothenburg swedish word accents: A case of cue trading? In Horne, G. B. . M. (ed.), *Nordic Prosody: Proceedings of the IXth Conference*. Frankfurt am Main, Germany: Peter Lang, 225–233.

Shen, J. and Souza, P. E. (2017). Do older listeners with hearing loss benefit from dynamic pitch for speech recognition in noise? *American Journal of Audiology* 26: 462–466, doi:10.1044/2017 aja-16-0137.

Shen, J. and Souza, P. E. (2018). On dynamic pitch benefit for speech recognition in speech masker. *Frontiers in Psychology* 9, doi:10.3389/fpsyg. 2018.01967.

Shen, J. and Souza, P. E. (2019). The ability to glimpse dynamic pitch in noise by younger and older listeners. *The Journal of the Acoustical Society of America* 146: EL232–EL237, doi:10.1121/1.5126021.

Shriberg, E., Ladd, D. R., and Terken, J. Modeling intra-speaker pitch range variation: predicting F0 targets when “speaking up”. *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP 96*, pp.650–653. doi: 10.1109/icslp.1996.607445.

Sluijter, A. M. C. (1991). Een perceptieve evaluatie van een model voor alinea-intonatie met synthetische spraak (a perceptual evaluation of a model for paragraph intonation with synthetic speech). In *Internal Report 801*. Institute for Perception Research, Eindhoven, The Netherlands.

Smith, R. and Rathcke, T. (2020). Dialectal phonology constrains the phonetics of prominence. *Journal of Phonetics* 78: 100934, doi:10.1016/ j.wocn.2019.100934.

Studdert-Kennedy, M. and Hadding, K. (1973). Auditory and linguistic processes in the perception of intonation contours. *Language and Speech* 16: 293–313, doi:10.1177/002383097301600401.

Tang, C., Hamilton, L. S. and Chang, E. F. (2017). Intonational speech prosody encoding in the human auditory cortex. *Science* 357: 797–801, doi:10.1126/science.aam8577.

Terken, J. (1991). Fundamental frequency and perceived prominence of accented syllables. *Journal of the Acoustical Society of America* 89: 1768–1776.

Terken, J. (1994). Fundamental frequency and perceived prominence of accented syllables. II. Nonfinal accents. *Journal of the Acoustical Society of America* 95: 3662–3665.

Terken, J. and Hermes, D. (2000). The perception of prosodic prominence. In M. Horne (ed.), *Prosody: Theory and Experiment, Studies Presented to Gösta Bruce*. Dordrecht, Netherlands: Kluwer Academic Publishers, 89–127, doi:10. 1007/978-94-015-9413-4 5.

Titze, I. R. (1989). On the relation between subglottal pressure and fundamental frequency in phonation. *The Journal of the Acoustical Society of America* 85: 901–906, doi:10.1121/1.397562.

Traunmüller, H. and Eriksson, A. (1995). The perceptual evaluation of f0 excursions in speech as evidenced in liveliness estimations. *The Journal of the Acoustical Society of America* 97: 1905–1915, doi:10.1121/1.412942.

Turnbull, R., Royer, A. J., Ito, K. and Speer, S. R. (2017). Prominence perception is dependent on phonology, semantics, and awareness of discourse. *Language, Cognition and Neuroscience* 32: 1017–1033, doi:10. 1080/23273798.2017.1279341.

Turner, D. R., Bradlow, A. R. and Cole, J. S. (2019). Perception of pitch contours in speech and nonspeech. In *Interspeech 2019*. Graz, Austri: ISCA, 2275–2279, doi:10.21437/interspeech.2019-2619, September 15 – 19 2019.

Whalen, D. and Levitt, A. G. (1995). The universality of intrinsic F0 of vowels. *Journal of Phonetics* 23: 349–366.

Woods, K. J. P., Siegel, M. H., Traer, J. and McDermott, J. H. (2017). Headphone screening to facilitate web-based auditory experiments. *Attention, Perception, & Psychophysics* 79: 2064–2072, doi:10.3758/ s13414-017-1361-2.

Yu, A. C. L. (2010). Tonal effects on perceived vowel duration. In B. C. Fougeron, M. Kuehnert, M. D'Imperio, and N. Vallée, *Laboratory Phonology 10* (pp. 151–168). BerlinNew York: Walter de Gruyter.

Zahner, K. & Braun, B. (2018). F0 peaks are a necessary condition for German infants' perception of stress in metrical segmentation. In the *Proceedings of the 17th Speech Science and Technology Conference (SST 2018)*, pp. 73–76, Sydney, Australia.

Zahner, K., Kutscheid, S. and Braun, B. (2019). Alignment of f0 peak in different pitch accent types affects perception of metrical stress. *Journal of Phonetics* 74: 75–95, doi:10.1016/j.wocn.2019.02.004.

Zahner, K., Schönhuber, M. & Braun, B. (2016). The limits of metrical segmentation: intonation modulates infants' extraction of embedded trochees. *Journal of Child Language* 43(6): 1338-1364. https://doi.org/10.1017/S0305000915000744.

Appendix A: Stimulus Type Effect in Experiment 1

Reiterated Speech and Complex Tone Stimuli

Experiment 1 had additional trials with Reiterated Speech and Complex Tone stimuli. The English speech, Reiterated Speech and Complex Tone triplets had an identical f0 contour, loudness contour, and duration. Reiterated speech was considered intermediate between complex tones and English speech in terms of its acoustic and linguistic complexity (cf. Pierrehumbert, 1979; Terken, 1991; Gussenhoven & Rietveld, 1998). Sample sound files are available as supplementary materials.[5]

To create the base stimuli for Reiterated Speech, the speaker (see Section 3.1.3) recorded a sequence of *na* syllables alternating in stress, at two speaking rates (normal and slow) several times (e.g., ná-na-ná-na-ná-na-ná-na-ná-na). One reiterated utterance as a base was resynthesised by editing one stressed syllable *ná* and one unstressed syllable *na* out from the middle of the recorded utterance spoken at a slow speaking rate (stressed *ná* 363 ms, unstressed *na* 400 ms). These two syllables were not adjacent to each other in the original utterance, and the slowly spoken ones were chosen because shortening rather than lengthening in the resynthesis process resulted in more natural-sounding stimuli. The *ná* and *na* syllables were concatenated, with a 10 ms overlap to avoid clipping, as an utterance *nanánananána* with the same stress pattern as the English sentences. The root-mean-square amplitude of all base utterances was scaled at 70 dB before further resynthesis.

Four resynthesised base tones for the Complex Tone stimuli were created as harmonic complexes with an f0 of 200 Hz and all harmonics present up to 6 kHz with a sampling rate 51.2 kHz in the sine phrase. The four tones shared spectral properties but differed in duration

[5] expt1_English_peaks_sharp.wav, expt1_English_valleys_sharp.wav, expt1_English_peaks_plateau.wav., expt1_English_valleys_plateau.wav, expt1_reiterated_peaks_sharp, expt1_reiterated_valleys_sharp.wav, expt1_tone_peaks_sharp.wav, and expt1_tone_valleys_sharp.wav.

so that they were equal in length to each of the four English utterances. The tones were filtered so that the spectral slope decreased by 6 dB/octave above 200 Hz. The same resynthesis procedure for duration, intensity, and f0 was applied to the Complex Tones as in the English Speech and Reiterated Speech stimuli. The duration and intensity properties of the English utterances were copied onto the reiterated utterances for each annotated interval using a Praat script (written by Kyuchul Yoon) which automatically performed resynthesis for duration and intensity respectively (PSOLA for duration, non-PSOLA for intensity). The script automated the process of taking the duration and the intensity contour from each annotated interval from one utterance and embedding it in the corresponding interval of the other utterance. Then the pitch tier which had the linearly stylised f0 track (Fig. 3) was superimposed to create each stimulus.

Experimental Procedure

The main experiment consisted of three Stimulus Type blocks (Complex Tone, Reiterated Speech, and English Speech). The presentation order of the blocks was counterbalanced, and within each block the stimulus presentation order was randomised for each participant.

Analysis and Results

We used the modelling methods described in Section 3.1.5. Here we report only the results pertaining to Stimulus Type (Table A.1). Data in Figure A.1 were collapsed over Second Accent Shape, which did not interact with Stimulus Type.

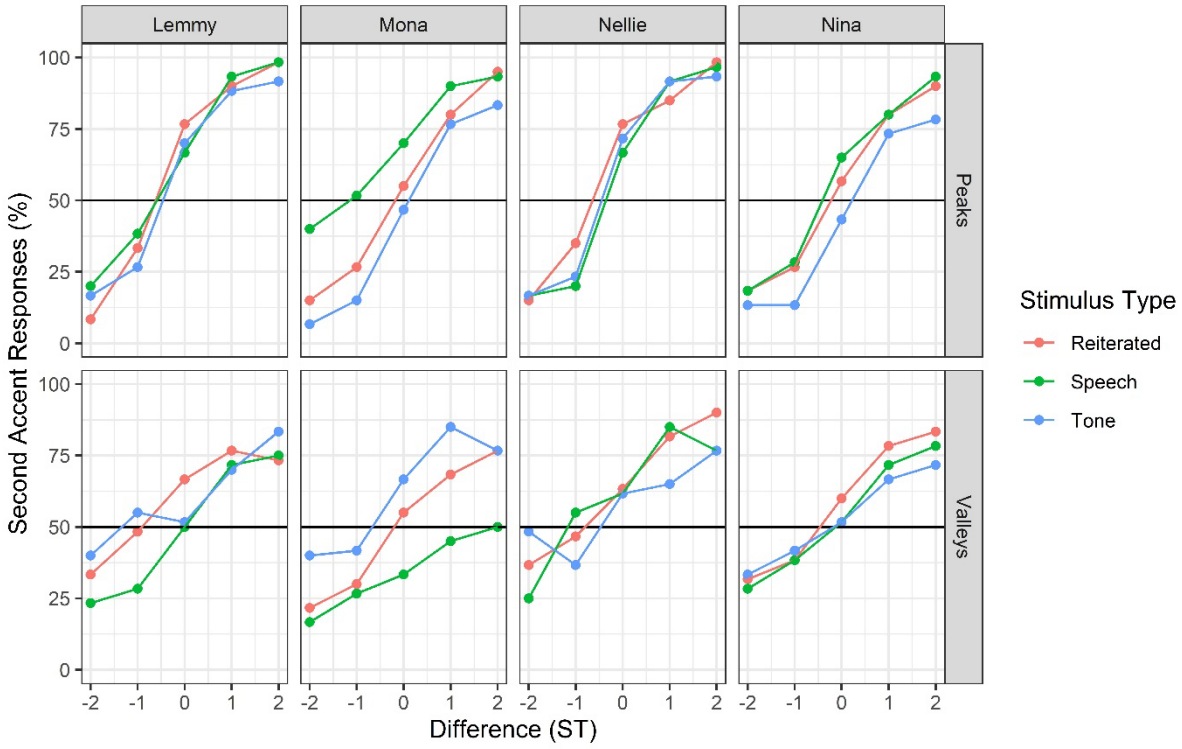
Stimulus Type had a significant main effect ($p < 0.05$) and it interacted with Accent Type ($p < 0.001$). Stimulus Type and Accent Type also interacted with Item (Stimulus Type × Accent Type × Item, $p < 0.001$). Figure A.1 shows that for three out of four Items (Lemmy, Mona, and Nellie), the differences between Stimulus Types were larger for Valleys than for

Peaks. Stimulus Type interacted with Accent Height Difference ($p < 0.05$), showing that the response function slope differed across the Stimulus Types.

Table A.1 Results of the model comparisons ($\alpha = 0.05$).

	χ^2	df	p
Stim Type	6.22	2	0.04*
Stim Type × Accent Type	59.07	2	< 0.001***
Stim Type × Shape	1.13	2	0.57
Stim Type × Item	12.32	6	0.06
Stim Type × Difference	9.21	2	0.01*
Stim Type × Accent Type × Shape	0.47	2	0.79
Stim Type × Accent Type × Item	68.52	6	< 0.001***
Stim Type × Accent Type × Difference	0.88	2	0.65

Fig. A.1 The averaged frequency of 'second accent' responses (%) across all participants by Item, Accent Shape, and Stimulus Type.



Discussion

Fig. A.1 shows the perceptual asymmetry between 'peaks' and 'valleys' for reiterated speech and complex tones. Some previous studies reported that listeners generally show

reduced sensitivity to acoustic variation in speech compared to pure or complex tones (e.g., Klatt, 1973; 't Hart et al. 1990, p. 7; Green et al., 2004; Moore, 2008; Heeren, et al., 2012; Turner et al., 2019). However, our 'item' effect suggests that this is not always the case and the extent of divergence in listeners' responses related to stimulus type depends on its precise acoustic shape. When four utterance items were used, in some cases, the response functions overlapped across the stimulus types (English Speech, Reiterated Speech, and Complex Tones). When the acoustic complexity of the non-speech stimuli or the task complexity increases from the traditional psychoacoustic experiment on discrimination between syllable-length sounds, we may not observe significant differences between speech and complex tones (cf. Studdert-Kennedy & Hadding, 1973; Tang et al., 2017).

Appendix B: Stimulus Type Effect in Experiment 2

Complex Tone Stimuli

Experiment 2 used Complex Tone stimuli in addition to Speech stimuli. The tones were harmonic complexes created with an f0 of 200 Hz and all harmonics up to 6 kHz, at a sampling rate of 44.1 kHz. The complex tones were equal in duration to the English utterance (0.94 s) and they had an identical f0 contour to the speech stimuli. The base tone was filtered so that the spectral slope decreased by 6 dB/octave above 200 Hz before further resynthesis. The intensity rose gradually at the beginning and fell gradually at the end over a 50 ms duration.

Experimental Procedure

The main experiment consisted of eight blocks (2 Accent Type × 2 Level × 2 Stimulus Type). Each block had five stimuli for the five Accent Height Difference levels. The order of blocks was counterbalanced across the eight groups. Four groups started with Speech, then the order of Accent Type and Level was counterbalanced (i.e., Speech-Low Level-Valleys, Speech-Low Level-Peaks, Speech-High Level-Valleys, Speech-High Level-Peaks, etc.). The other four groups started with Complex Tone. Listeners were randomly allocated to the eight groups (8 listeners × 2 groups, 7 listeners × 5 groups, 6 listeners × 1 group).

Analysis and Results

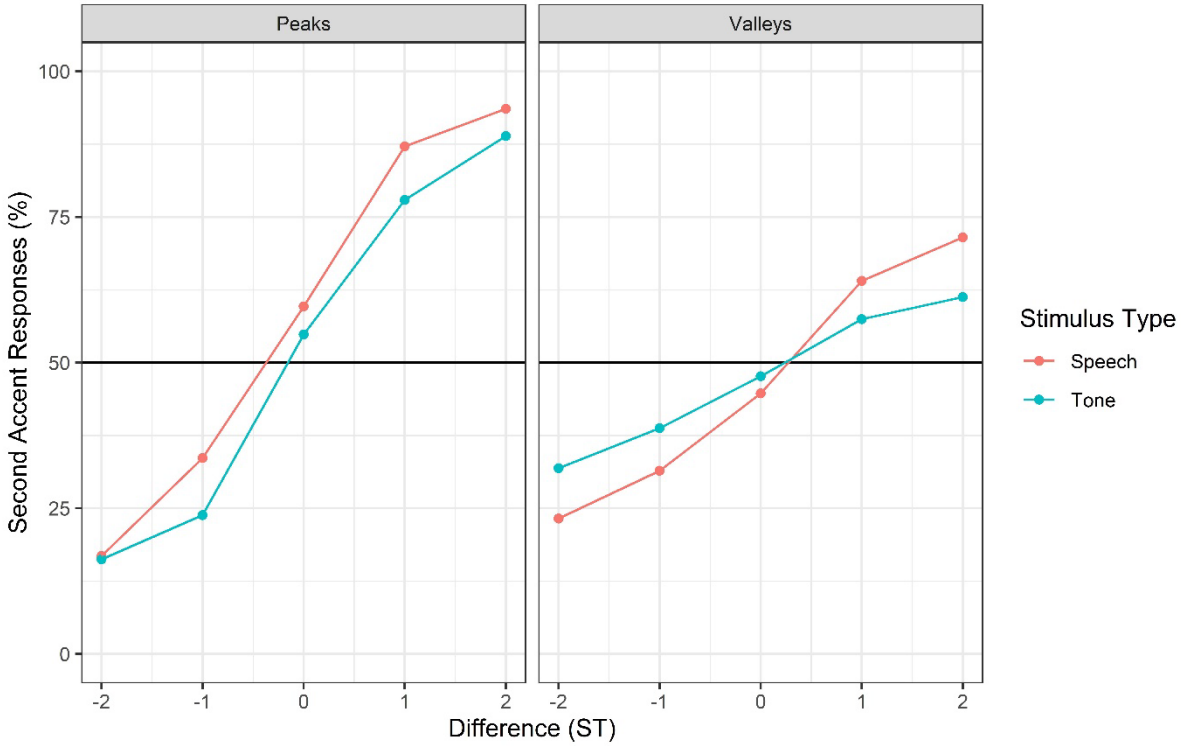
The same analysis methods were used as in Appendix A. Stimulus Type had a significant main effect ($p < 0.001$). It was also part of the Stimulus Type × Accent Type interaction effect. The interaction effects Stimulus Type × Accent Height Difference ($p < 0.001$) and Stimulus Type × Accent Type × Accent Height Difference interaction ($p < 0.05$) were also significant. Therefore, we interpret the three-way Stimulus Type × Accent Type × Accent Height Difference interaction here.

Data in Figure B.1 were collapsed over Second Accent Shape and Level, which did not interact with Stimulus Type. Figure B.1 shows that for Peaks, the response functions for Speech and Complex Tones did not markedly differ, although listeners had a stronger 'second accent' saliency bias for Speech, with a steeper slope of the response function. For Peaks, the PSE was lower for Speech than for Complex Tone, indicating that a smaller excursion for Speech than for Complex Tone led to perceived equivalence in pitch between the two accents. On the other hand, for Valleys, response functions for both Speech and Complex Tone were flat. The slope was slightly steeper for Speech, suggesting that listeners' discrimination was relatively reduced for Complex Tone compared to Speech. The PSEs for Speech and Complex Tone overlapped between zero and one semitone. For both Speech and Complex Tone, perceived equivalence between the two 'valleys' was achieved when the second accent was slightly lower in f0 than the first.

Table B.1 Results of the model comparisons ($\alpha = 0.05$)

	χ^2	df	P
Stim Type	12.82	1	< 0.001***
Stim Type × Accent Type	17.11	1	< 0.001***
Stim Type × Shape	0.05	1	0.82
Stim Type × Level	1.89	1	0.17
Stim Type × Difference	37.11	1	< 0.001***
Stim Type × Accent Type × Shape	0.45	1	0.5
Stim Type × Accent Type × Level	3.18	1	0.07
Stim Type × Accent Type × Difference	4.23	1	0.04*
Stim Type × Shape × Level	1.59	1	0.21
Stim Type × Shape × Difference	0.53	1	0.47
Stim Type × Level × Difference	0.5	1	0.48

Figure B.1 The averaged frequency of 'second accent' responses (%) across all participants by Accent Type and Stimulus Type



Discussion

The perceived asymmetry between 'peaks' and 'valleys' was replicated with complex tones at different f0 levels. The source of the Stimulus Type × Accent Type × Accent Height Difference is probably the markedly flat response function for the complex tones with the f0 'valleys', suggesting listeners' reduced discrimination. The results show that listeners' discrimination was not heightened for the complex tones compared to speech (see Appendix A for similar discussion). It is possible that listeners relied more on the spectral cues than f0 in when judging pitch height when they found the task challenging with the 'valleys', and the richer spectral cues in speech than in complex tones could have resulted in the steeper response function (cf. see McPherson & McDermott, 2017 for different mechanisms for pitch perception).