INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Intonation Facilitates Prediction of Focus even in the Presence of Lexical Tones

Martin Ho Kwan Ip, Anne Cutler

The MARCS Institute, Western Sydney University, Penrith South, NSW 2751, Australia
ARC Centre of Excellence for the Dynamics of Language, Canberra, ACT 2601, Australia
m.ip@westernsydney.edu.au, a.cutler@westernsydney.edu.au

Abstract

In English and Dutch, listeners entrain to prosodic contours to predict where focus will fall in an utterance. However, is this strategy universally available, even in languages with different phonological systems? In a phoneme detection experiment, we examined whether prosodic entrainment is also found in Mandarin Chinese, a tone language, where in principle the use of pitch for lexical identity may take precedence over the use of pitch cues to salience. Consistent with the results from Germanic languages, response times were facilitated when preceding intonation predicted accent on the target-bearing word. Acoustic analyses revealed greater F_0 range in the preceding intonation of the predicted-accent sentences. These findings have implications for how universal and language-specific mechanisms interact in the processing of salience.

Index Terms: prosody, focus, intonation, speech perception

1. Introduction

In every conversation, listeners not only need to interpret the phonetic sequence that determines what words and utterances they hear, but also the suprasegmental structure that dictates how these utterances are produced. Indeed, as Bolinger [1] pointed out, how a spoken word is processed also depends on its place above the lexical level in the utterance intonation contour. Across languages [2], words that are prosodically highlighted are acoustically clearer [3], are better retained in memory [4], recognised more rapidly [5], and are more likely to direct listeners' attention to new elements of the discourse structure [6].

However, it is still an empirical question whether prosody serves in the same manner across languages as a universal cue to semantic salience. On one hand, findings from speech production research show that languages vary in how different aspects of prosody are used to realise focus. For instance, cross-language experiments comparing English, French, and German found language-specific strategies where only German speakers use duration to enhance new information [7]. Even in languages where prosodic focus is produced in the same way, cross-language variation could nonetheless exist in the degree to which speakers use the different prosodic cues [8]. At the same time, how salience is achieved by means of prosody can also depend on the particular morpho-syntactic structure of the language, as in Wolof [9], where morphological markers are available, so that speakers do not redundantly use intonation, or in Italian [10], where focus contrasts can instead be expressed by word order, such that pitch accents provide less contextual information [11]. Similarly, in Indonesian [12], syntax is used as the primary means for focus marking because words in phrase-final positions cannot be accented. Given the language-specific differences in the resources for marking focus in production, there may in consequence be no universal manner in which prosodic focus is processed in speech perception.

On the other hand, perception of prosodic focus may be based on a common underlying mechanism that is separate from production. One way in which languages may be alike is in listeners' use of utterance-level prosody in their predictions of semantic salience. In a number of experiments, Cutler and colleagues [13, 14, 15] found that listeners could entrain to the preceding intonation contour to predict the location of an accented word in the utterance. In a phoneme detection task, participants heard a series of sentences and responded as fast as they could to a phoneme target (e.g., [d]). Results show that listeners responded faster to the target in sentences where the preceding intonation contour predicted high stress on the target-bearing word than in contexts where the intonation predicted low stress. Importantly, response time was still faster even in contexts where the highly stressed target-bearing word was removed and replaced by a neutral version of the same word. Therefore, beyond the apparent variation in the production of focused words, it is still possible that prosodic contours have a universal function in enabling listeners to navigate speech and locate the semantically most central part of the utterance information structure.

To date, all the evidence on prosodic entrainment to salience has come from English and Dutch. This makes it difficult to reach any conclusions about universality and language-specificity in prosodic perception, since the relation between prosody and focus is essentially the same in these two languages [16]. The experiment we report here aims to better address this issue by examining the same phenomenon in Mandarin Chinese. An investigation with Mandarin listeners could provide us with a unique insight into the language-universality versus specificity question, as Mandarin has features that are similar and features that are different from English and Dutch. Despite their typological distance, all three languages express prosodically salient words in a similar way (i.e., exaggerated F_0 range, increased duration and intensity). However, other differences in their phonological systems could prevent Mandarin from showing the same contour entrainment effect. As a tone language, Mandarin may have a less elaborate intonational system, arguably because much of the F_0 contour is exhausted in the phonetic expressions of contour tones [17, 18]. Since suprasegmental cues to tone also co-specify lexical identity, it may be the case that the exaggeration of prosodic cues used for prosodic focus is only localised on the focused word, with cues in the preceding intonation preempted by tonal movements. Indeed, previous studies suggest that the intonation before focus in Mandarin tends to be similar to a neutral sentence [19]. Intonation may then be less useful, since the same suprasegmental processing space may already be used for tone perception.

http://dx.doi.org/10.21437/Interspeech.2017-264

2. Method

2.1. Participants

We tested 52 native speakers of Mandarin Chinese (M_{age} = 25.59 years, *SD* = 6.10 years; 35 females). All participants were born in Mainland China and had been living in Australia for any period between 23 days to 27 years (*M* = 3.14 years, *SD* = 5.40 years). None reported any hearing impairment.

2.2. Materials

The stimuli were recorded by a female native speaker of Mandarin (age 28 years) who did not know the purpose of the experiment. Twenty-four unrelated experimental sentences were recorded in Mandarin in three versions. In the first version, the target-bearing word received emphatic stress. In the second version, emphatic stress was instead placed on a word that occurred after the target-bearing word, which as a result, received very reduced stress. In the third version, the target-bearing word and the sentence as a whole were produced in a neutral manner. In all of the experimental sentences, the phoneme target was an aspirated [p^h] occurring at the start of the target-bearing word's first syllable. Half of the sentences had the phoneme target occurring on a rising tone (e.g., 葡萄 [$p^hu2\ t\Lambda0$] "grapes") and half had the target on a falling tone (e.g., 骗子 [$p^hjɛn4\ ʂɻ0$] "swindler").

Using Praat [20], the target-bearing words were excised from all three versions of each experimental sentence. The high- and low-stressed target-bearing words from the first and second versions were each replaced by identical tokens of the same target word from the neutral version. This created a context where the prosodic contour leading to a high- or low-stressed target word remained intact, even though the actual target word was replaced by one with a neutral realisation. Therefore, the spliced sentences consisted of two versions of each sentence – one where the intonation contour predicted high stress on the target-bearing word, and one where the contour predicted low stress on that word – with the identical, neutrally realised, target word occurring in both versions.

Two experimental conditions were constructed, each containing one version of each of the 24 spliced experimental sentences, plus an additional set of 24 filler sentences. The experimental sentences with predicted high versus predicted low stress were counterbalanced across the two conditions. All of the sentences were produced at a fast-normal rate.

2.3. Procedures

All participants were tested in a sound-attenuated booth either individually or in pairs of two. The phoneme-detection task was administered using E-Prime software on a laptop computer, with attached to it a set of headphones and a Chronos USB-based device for button pressing. Participants were told that the experiment aimed to examine Mandarin speakers' memory and language comprehension. All participants were told that they had two tasks: first, pay careful attention to the meaning of each sentence, and second, press a button as fast and as accurately as they could whenever they heard a word that began with the target sound [p^h].

At the end, all participants completed a comprehension test in which they were asked to judge whether or not each of the 20 sentences in the list were from the experiment. All participants scored 65 percent or above in the test (*M* = 85.58 percent, *SD* = 10.37 percent, range: 65 – 100 percent).

3. Results

3.1. Response Time and Accuracy

Response times (RT) longer than 2500 milliseconds were excluded from final analyses, because such a delayed response may indicate a reprocessing of the sentence. Of the nine excluded datapoints, three were from sentences with predicted high stress and six came from predicted low stress contexts.

A two-tailed within-subjects t-test with an alpha threshold of .05 was conducted to assess the difference in RT between the predicted high versus low stress utterances. Analyses revealed that participants' RT were significantly faster in predicted high stress sentences (*M* = 541.19, *SD* = 171.71) compared to sentences with predicted low stress (*M* = 567.81, *SD* = 177.31), $t(51) = 2.30$, $p = .026$, $d = .15$ (see Figure 1).

With respect to detection accuracy, we performed a two-tailed binomial sign test to determine whether participants were more likely to miss a button press to the phoneme target in sentences with predicted low stress than in predicted high stress. In total, there were 15 misses in predicted low stress contexts and five misses in predicted high stress contexts, which was statistically different from chance, $p = .041$.

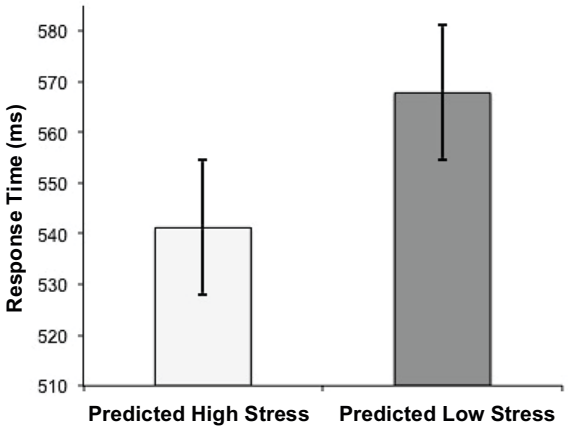


Figure 1: *Response time (ms) as a function of intonationally predicted high versus low stress. Error bars represent the standard error of the mean.*

3.2. Acoustic Analyses

Acoustic analyses of the stimulus recordings were conducted based on simultaneous inspection of the waveform and the spectrogram in Praat [20]. Segments consisting of three to four syllables before the onset of the target-bearing word were annotated and duration, mean F_0, F_0 range, mean rms-intensity, maximum rms-intensity, and rms-intensity range measured. We also measured the duration of the prosodic break, the part of the utterance between the onset of the target-bearing word and the offset of the word before it. Our results show that there was a significant difference in F_0 range between the predicted high and low stress contexts, such that syllables before target-bearing words had greater F_0 range in predicted high stress sentences (*M* = 105.30 Hz, *SD* = 42.61 Hz) than in predicted low stress contexts (*M* = 82.43, *SD* = 33.21 Hz), $t(22) = 4.18$, $p < .001$, $d = .60$. However, no significant differences were observed for mean F_0, or for any of the intensity and duration measures. Examples of a predicted high and low stress stimulus are displayed in Figure 2.

Target: [p^h]

mei2 joʊ3 ɻən2 tsaɪ4 ʈʂʊŋ1 kɔ3 nəŋ2 ɕiɑŋ1 ɕin4 p^hu2 tʌ0 nəŋ2 ʈʂɿ4 tsɑʊ4 ɕjɛn2 ʂweɪ3

没 有 人 在 中 国 能 相 信 葡 萄 能 制 造 香 水

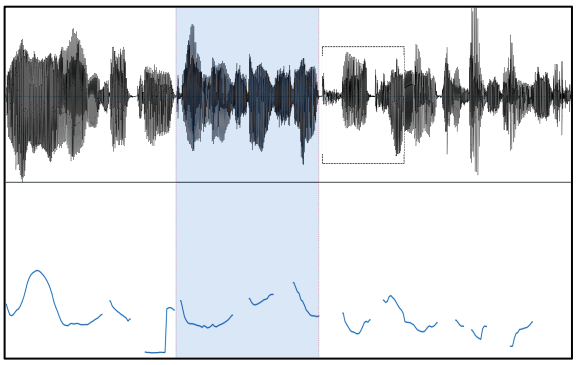
No one in China believes that grapes can be used to make perfumes

(a) 没有人在中国能相信 [**葡萄**] 能制造香水

(b) 没有人在中国能相信葡萄能制造 [**香水**]

(c) 没有人在中国能相信葡萄能制造香水

(a)



(b)

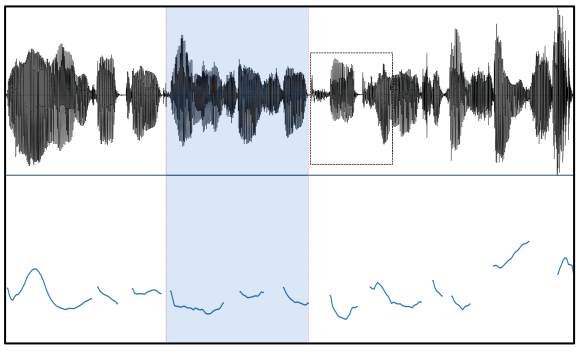


Figure 2: *Waveforms and pitch contours of an example experimental sentence in predicted high (a) and low (b) contexts; text (c) gives the neutral context. The shaded portion – four syllables preceding the target-bearing word – was analysed acoustically (section 3.2).*

As shown in Figure 2, the underlined target-bearing word in (a) and (b) (also shown as waveforms inside the dotted square brackets) was acoustically identical in the predicted high and predicted low sentence. However, the preceding intonational context in the two sentences differs because (a) was originally spoken with emphatic stress on the target-bearing word (i.e., 葡萄 "grapes") while (b) was produced with emphatic stress on a word occurring after the target-bearing word (i.e., 香水 "perfumes"). As revealed in the shaded portions of the waveform and pitch contours, the overall F_0 range expansion was greater in the predicted high stress sentence (106.28 Hz) compared to the predicted low stress sentence (72.07 Hz).

3.3. Control Analysis

As our participants were not fully uniform with respect to how long they had spent in non-Mandarin-speaking environments, an additional analysis was conducted to assess whether participants' RT was related to their exposure to English as a foreign language while living in Australia. Difference scores in RT between the predicted high stress and predicted low stress sentences were calculated for each participant. A Pearson's correctional analysis was performed to calculate the association between participants' difference score in RT and their length of stay in an English-speaking country (measured as days from the date of arrival in Australia to the date of testing), and the result was non-significant, r = -.06, p = .666. (see Figure 3 below). Therefore, the Mandarin participants' RT difference between the predicted high versus predicted low sentences was not related to their amount of exposure to English in an English-speaking country.

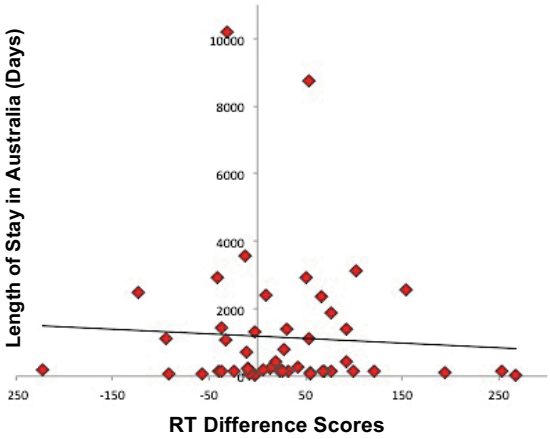


Figure 3: *Non-significant negative correlation between participants' length of stay (i.e., date of testing minus date of arrival) in an English-speaking country (in days) and their RT difference scores.*

4. Discussion

The present experiment offers a useful insight into how language-universal mechanisms may play a part in the sentence comprehension process. Consistent with findings from Germanic languages, we have shown that listeners can entrain to intonation to forecast an upcoming focus, even in a language where the pitch contours are already used for lexical tone perception. Moreover, entrainment occurred even when the actual salient word was replaced by a neutral word, suggesting that processing of salience not only involves attention to acoustic cues from the focused word, but also use of the preceding prosody. In light of these results, the present experiment demonstrates that a universal strategy may still exist in listeners' prefocus entrainment to prosody, despite cross-language differences in the production of focus.

The fact that prosodic cues to salience can co-exist in speech with lexical cues to tone is well known. As demonstrated in previous production studies [8, 19, 21], prosody can be used for producing focused words in Mandarin in ways that do not interfere with tonal identity (e.g., by exaggeration of pitch register). However, what is interesting here is the perceptual reflection of this dual role for F_0: Mandarin listeners could make use of the preceding intonation

contour even before they heard the predicted focused word. According to a number of scholars [see 17, 18], languages with lexical tones tend to have less scope for a complex intonational system, debatably because much of the F_0 contour is preempted by the use of pitch for distinguishing words. On this interpretation, any prosodic cues to focus in Mandarin would be restricted to the focused region of the utterance, and any part of the preceding intonation contour would therefore be uninformative. However, our acoustic analyses reveal that, on the contrary, pitch cues to upcoming focus were present in the preceding intonation, at least in the form of greater F_0 range expansion three or four syllables before the onset of the predicted accent. This indicates that Mandarin listeners could still manage to anticipate upcoming focus by attending to pitch range information in the preceding intonation.

This acoustic finding is noteworthy in light of a recent production study [8] from our laboratory, where a greater degree of F_0 range expansion production of focused words was found for Mandarin speakers compared to English speakers. Given that salience is fundamentally gradient [22], it could be the case that Mandarin speakers start to expand their pitch range quite early in the utterance, in preparation for pitch range exaggeration on the focused word. This may even result from an automatic physiological mechanism; as Bolinger [23] noted four decades ago, the semantically most "interesting" or "important" part of an utterance is associated with heightened arousal, greater respiratory effort, dramatic pitch changes, and more energetic movement. Not only speakers' realisation of focus, but also listeners' entrainment to intonation contours and their faster response times in predicted high stress contexts could thus be due to increasing levels of physiological arousal by each as an acoustically salient word approaches in the speech stream. To test this idea, future research could look at listeners' galvanic skin response as a measure of their arousal level while they perform a phoneme-detection task.

At the same time, however, prosodic entrainment to locate focus may be justified by its value as a comprehension strategy for everyday social interactions. Irrespective of language or culture, holding a conversation presents a number of mental challenges. For one thing, conversational utterances tend to be fragmentary and elliptical [24]. At the same time, there is much uncertainty with respect to how a dialogue will unfold, and listeners often need to constantly organise and update their current discourse model. Entraining to intonation contours to detect the semantically most central part of the utterance may therefore provide a headstart for listeners in navigating the utterance information structure early on, making it a strategy useful for all listeners for maintaining a socially effective conversation.

While this prosodic entrainment is potentially universal, there may nevertheless be some language-specific factors that could modulate the extent to which listeners would rely on it. In the previous reports on English and Dutch, listeners' average RTs were as much as 80 milliseconds faster on the predicted high stress sentences compared to the predicted low stress sentences (22% and 12% of the grand mean in [14, 15] respectively). In the present experiment, the difference in RTs between the predicted high versus low stress sentences, was around 5% of the grand mean (26.62 milliseconds). A reason why Mandarin listeners exhibited this smaller RT difference could be relative lack of variety in prosodic cues in the preceding intonation. In the present experiment, the greater F_0 range found in the preceding intonation contour was the only source of prosodic cue that is available, and this was only for around 750 milliseconds before the onset of the focused word. In previous research such as the Cutler and Darwin study in British English [14], however, pitch was not the only cue that listeners were able to make use of, since listeners still responded faster in predicted high stress sentences when the F_0 contour was rendered uninformative (i.e., artificially levelled out). For this reason, languages may vary in the number of available prosodic cues that listeners can use to predict focus.

A final question that warrants further research is whether some prosodic cues (e.g., F_0, duration) may prove more informative to listeners than others. Although studies in British English [14, 25, 26] have suggested that no single cue is inherently stronger in that language variety, and that what really matters is only that the cues do not conflict, this may not be the case in Australian English in which, for example, F_0 movement has become more profligately used in spontaneous speech ("uptalk"; [27]).

5. Conclusions

Even though Mandarin has lexical tone, whereby F_0 patterns carry a lexical as well as a sentence-level functional load, Mandarin listeners entrain to preceding intonation across utterances to predict upcoming focus. Consistent with data from speech production in Mandarin, acoustic analyses of the stimuli revealed greater F_0 expansion in the preceding intonation of predicted high stress sentences, i.e. the Mandarin listeners' prosodic entrainment was supported to the greatest extent by cues to pitch range, and these cues were located in the utterance portion immediately preceding the focused word. Language-specific factors determine the precise realisation of what appears nevertheless to be a universal listening strategy.

6. Acknowledgements

Financial support was provided by the MARCS Institute and the ARC Centre of Excellence for the Dynamics of Language. We thank Mark Antoniou and Chris Carignan for technical support and advice. We would also like to thank Zhang Yong and Cheng Cheng for their help in participant recruitment.

7. References

[1] D. L. Bolinger, "Around the edge of language: Intonation," *Harvard Educational Review*, vol. 34, no. 2, pp. 282–296, 1964.

[2] D. R. Ladd, *Intonational Phonology*. Cambridge: Cambridge University Press, 2008.

[3] P. Lieberman, "Some effects of semantic and grammatical context on the production and perception of speech," *Language and Speech*, vol. 6, no. 3, pp. 172–187, 1963.

[4] S. Fraundorf, D. G. Watson, and A. S. Benjamin, "Recognition memory reveals just how CONTRASTIVE contrastive accenting really is," *Journal of Memory and Language*, vol. 63, no. 3, pp. 367–386, 2010.

[5] A. Cutler and D. J. Foss, "On the role of sentence stress in sentence processing," *Language and Speech*, vol. 20, no. 1, pp. 1–10, 1977.

[6] C. A. Fowler and J. Housum, "Talkers' signaling of "new" and "old" words in speech and listeners' perception and use of the distinction," *Journal of Memory and Language*, vol. 26. no. 5, pp. 489–504, 1987.

[7] J. F. Hay, M. Sato, A. E. Coren, C. L. Moran, and R. L. Diehl, "Enhanced contrast for vowels in utterance focus: A cross-language study," *Journal of the Acoustical Society of America*, vol. 119, no. 5, pp. 302 –3033, 2006.

[8] M. H. K. Ip and A. Cutler, “Crosslanguage data on five types of prosodic focus,” in *Speech Prosody 2016*, *May 31–June 3*, *Boston, USA*, *Proceedings*, 2016, pp. 330–334.
[9] A. Rialland and S. Robert, “The Intonational system of Wolof,” *Linguistics*, vol. 39, no. 5, pp. 893–939, 2001.
[10] E. Vallduví, *Informational Component*. Garland Publishers, 1992.
[11] M. Swerts, E. Krahmer, and C. Avesani, “Prosodic marking of information status in Dutch and Italian: A comparative analysis,” *Journal of Phonetics*, vol. 30, no. 4, pp. 629–654, 2002.
[12] R. Goedemans and E. van Zanten, “Stress and accent in Indonesian,” in *Prosody in Indonesian Languages*, V. van Heuven and E. van Zanten (eds.). Utrecht: Utrecht University Press, 2007, pp. 35–62.
[13] A. Cutler, “Phoneme-monitoring as a function of preceding intonation contour,” *Perception and Psychophysics*, vol. 20, no. 1, pp. 55–60, 1976.
[14] A. Cutler and C. J. Darwin, “Phoneme-monitoring and preceding prosody: Effects of stop closure duration and of fundamental frequency,” *Perception and Psychophysics*, vol. 29, no. 3, pp. 217–224, 1981.
[15] E. Akker and A. Cutler, “Prosodic cues to semantic structure in native and nonnative listening,” *Bilingualism: Language and Cognition*, vol. 6, no. 2, pp. 81–96, 2003.
[16] C. Gussenhoven, “Focus, mode and the nucleus,” *Journal of Linguistics*, vol. 19, no. 2, pp. 377–417, 1983.
[17] J. Pierrehumbert, “Prosody and intonation,” in *MIT Encyclopedia of Cognitive Science*, R. A. Wilson and F. C. Keil (eds.). Cambridge: MIT Press, 1999, pp. 679–682.
[18] B. Hayes, *Metrical Stress Theory: Principles and Case Studies*. Chicago, Chicago University Press, 1995.
[19] Y. Xu, “Effect of tone and focus on the formation and alignment of f0 contours,” *Journal of Phonetics*, vol. 27, no. 1, pp. 55–105, 1999.
[20] P. Boersma, “Praat, a system for doing phonetics by computer,” *Glot International*, vol. 5, no. 9/10, pp. 1381–3439, 2002.
[21] Y. Chen and C. Gussenhoven, “Emphasis and tonal implementation in Standard Chinese,” *Journal of Phonetics*, vol. 36, no. 4, pp. 724–746, 2008.
[22] E. Flemming, “The role of pitch range in focus marking,” in *Workshop on Information Structure and Prosody*, Studiecentrum, Soeterbeeck, 2008.
[23] D. L. Bolinger, “Intonation across languages,” in *Universals of Human Language II: Phonology*, J. Greenberg (ed.). Palo Alto: Stanford University Press, 1978, pp. 471–524.
[24] S. Garrod and M. J. Pickering, “Why is conversation so easy?” *Trends in Cognitive Science*, vol. 8, no. 1, pp. 8–11, 2004.
[25] A. Cutler, “Components of prosodic effects in speech recognition,” in International Congress of Phonetic Sciences, *August 1–August 7, Tallinn, Estonia, Proceedings,* 1987, pp. 84–87.
[26] A. Cutler and J. M. McQueen, “How prosody is both mandatory and optional,” in *Above and Beyond the Segments: Experimental Linguistics and Phonetics*, J. Casper, Y. Chen, W. Heeren, J. Pacilly, N. O. Schiller, and E. van Zanten (eds.). Amsterdam: Benjamins, 2014, pp. 71–82.
[27] J. Fletcher and J. M. Harrington, “High-rising terminals and fall-rise tunes in Australian English,” *Phonetica*, vol. 58, no. 4, pp. 215–229, 2001.