10th International Conference on Speech Prosody 2020
25-28 May 2020, Tokyo, Japan

Uptalk interpretation as a function of listening experience

Y. Asano[1], *C. Yuan*[2], *A.-K. Grohe*[3], *A. Weber*[4], *M. Antoniou*[5], *A. Cutler*[5]

[1]Research Institute of Media and Communication, Hokkaido University, Japan
[2]Hefei Institute of Technology, China
[3]Nuance, Germany
[4]English Department, University of Tübingen, Germany
[5]The MARCS Institute for Brain, Behaviour and Development, Western Sydney University

yuki.asano@imc.hokudai.ac.jp

Abstract

The term "uptalk" describes utterance-final pitch rises that carry no sentence-structural information. Uptalk is usually dialectal or sociolectal, and Australian English (AusEng) is particularly known for this attribute. We ask here whether experience with an uptalk variety affects listeners' ability to categorise rising pitch contours on the basis of the timing and height of their onset and offset. Listeners were two groups of English-speakers (AusEng, and American English, henceforth AmEng), and three groups of listeners with L2 English: one group with Mandarin as L1 and experience of listening to AusEng, one with German as L1 and experience of listening to AusEng, and one with German as L1 but no AusEng experience. They heard nouns (e.g., flower, piano) in the framework "Got a NOUN", each ending with a pitch rise artificially manipulated on three contrasts: low vs. high rise onset, low vs. high rise offset and early vs. late rise onset. Their task was to categorise the tokens as "question" or "statement", and we analysed the effect of the pitch contrasts on their judgements. Only the native AusEng listeners were able to use the pitch contrasts systematically in making these categorisations.

Index Terms: Uptalk, F_0, experiences, perception

1. Introduction

Speech prosody conveys many different types of information, at all levels of linguistic structure and from many non-linguistic dimensions. One of the latter dimensions, relatively little-studied to date, is information concerning the dialectal or varietal background of the talker, to which the phenomenon belongs that we investigate in this study, termed "uptalk". Although it is not a new linguistic development, its name is fairly new and attracted research interests in recent decades. Uptalk, as the name suggests, refers to a prevalence of intonational rises in speech. Intonational rises have linguistic functions, of course; signalling a question as opposed to a statement is particularly common across languages, as is the junctural function of signalling a phrase boundary, e.g., [1, 2]. Uptalk, however, denotes rises used when no such linguistic function is being served; utterances may finish on a rising intonation even when the utterance is an ordinary matter-of-fact statement. Particular dialects of English (such as Belfast English, [3]) have long been known to exhibit this pattern, and it can vary across dialects in its realisation, or it can be suggested to be typical of particular talker groups (e.g., "Valley Girl" talk in Calfornia, [4]); [5] reviews the patterning of uptalk across English varieties. For our purposes, a relevant aspect of the uptalk phenomenon is that it is highly common in AusEng, where its use is very widespread, such that it is observed both across generations and across different parts of the country.

There is a growing literature on the precise realisation of uptalk [6, 5, 7], and on the speaker-related variables (sex, age, ethnicity) which are correlated with the use of this style [5]. Remarkably, however, to our knowledge there has been extremely little work on the perception of utterances with uptalk. This means that even basic knowledge of the role of uptalk in speech communication is as yet lacking. Do listeners who are unfamiliar with uptalk become misled when they are presented with uptalk speech by a "native" talker of the same first language (henceforth L1) but an uptalk variety (e.g., might they perceive questions where there are none)? Do uptalk users, correspondingly, become confused if uptalk is missing (e.g., interpret non-rising statements as rudeness)? Do speakers of different varieties develop differing sensitivity to F_0, just as tone language users, in whose languages F_0 cues tell lexical items apart, can have different expectations of intonational cues to syntactic structure than those of non-tone language users [8, 9]? And can uptalk be adapted rapidly by users of another variety of the same L1, and, importantly, by L2 learners?

One study [10] presented AusEng listeners with short utterances that had been artificially manipulated to have final rises that were shorter or longer, and began earlier or later. The listeners were able to develop biases in their responses in that they were more likely to hear longer rises as questions and shorter rises as statements, which is in keeping with the actual distribution of final rises in their language variety. This very usefully suggests that users of an uptalk variety, at least, can adapt to unusual realisations of F_0 in short utterances, and can categorise them systematically. This performance may be crucially based on the perceptual experience they have built up, but since the study in question included no other listener groups, we cannot be certain about that. Perhaps any listener with some experience of the variety in question can do it, or even any listener with the same L1; perhaps listeners who are certain to be sensitive to F_0 can do it; indeed, across the course of such an experiment, it is perhaps possible for anyone at all to set up a functional distinctive categorisation response.

Building on this foundation, we here investigate these further questions. In a way similar to that used in [10], we manipulate the F_0 applied to three-word English phrases that could potentially function as either a statement or a question. We then ask listeners to categorise each utterance accordingly. The reference test group are speakers of AusEng; on the basis of [10] we expect these listeners to successfully achieve a firm categorisation. The remaining groups test effects of differing types of experience: (a) little L2 experience

10.21437/SpeechProsody.2020-150

with uptalk variety (German learners of L2 English with little AusEng experience), (b) little L1 experience with uptalk variety (AmEng listeners from the Eastern-seaboard USA); (c) extensive L2 experience with uptalk variety and L1 experience with F_0 (Mandarin listeners in Sydney); (d) extensive L2 experience with uptalk variety and little L1 experience with F_0 (German-native listeners in Sydney).

2. Experiment

2.1. Methods

2.1.1. Participants

Twenty-eight AusEng listeners in Sydney (f = 16, m = 12, aged between 19 and 36, mean age = 25.6), 18 German listeners in Sydney who had extensive L2 experience with uptalk variety and little L1 experience with F_0 (henceforth German-with, f = 10, m = 8, aged between 22 and 40, mean age = 28.1), 24 German listeners who had little L2 experience with uptalk variety (henceforth German-without, f = 22, m = 2, aged between 18 and 26, mean age = 20.4), 24 Mandarin listeners in Sydney who had extensive L2 experience with uptalk variety and L1 experience with F_0 (f = 14, m = 10, aged between 21 and 40, mean age = 27.3), and 33 AmEng listeners from the Eastern-seaboard USA who had little L1 experience with uptalk variety (f = 20, m = 13, aged between18 and 26, mean age = 29.8) took voluntarily part in the experiment for a small fee. AusEng, Ger-with and Mandarin listeners were recruited and tested at the Western Sydney University in Australia, Ger-without listeners at the University of Tübingen in Germany and AmEng at the University of Maryland in the U.S.A. The minimum length of stay in Australia for Ger-with and Mandarin listeners was eight months (mean length = 8.4 years, ranged between 8 months and 15.5 years). None of the participants studied Musicology. They were all unaware of the purpose of the experiment. None of the participants had any self-reported speech or hearing deficits.

2.1.2. Stimuli

First, 19 sentences with the same elliptical syntactic structure "Got a *NOUN*", but each with a different noun (*NOUN* = *animal*, *banana*, *dinosaur*, *fireman*, *flower*, *glasses*, *highway*, *house*, *jar*, *juice*, *lemon*, *lolly*, *moon*, *onion*, *piano*, *raisin*, *sun*, *mayonnaise*, *elephant*) were created that could be interpreted either as a statement "I have got a *NOUN*." or as a question "Have you got a *NOUN*?", such as in a card game situation [11]. The sentences were recorded by a female AusEng speaker. The speaker was instructed to produce the sentences with a rising contour end as uptalk. The total durations of the sentences ranged between 600 and 947 ms with the average value of 762 ms.

A range of manipulations were then performed using *Praat*. The original F_0 information for each sentence was replaced by a set of predetermined rising contours that were manipulated to change the pitch level of the nuclear rise onset (low vs. high), the pitch level of the nuclear rise offset (low vs. high), and the time of the nuclear rise onset (early vs. late). The two former variables were used in [6] and reported to affect the perception of high rising terminals by Australian L1 listeners and the latter variable reported in [12, 13, 7]. In this study the pitch height of the nuclear rise onset manipulated from relatively low to high pitch in two steps (+ 0 Hz vs. + 50 Hz to the original value), and the time of the rise onset from relatively early to late (+ 0 ms vs. rise onset at the end of the stressed syllable), resulting in four different pitch rise onset for each sentence by varying the time and pitch height of the rise onset resulting in low-early, low-late, high-early and high-late). The pitch of the rise onset varied between 150 Hz and 290 Hz with the average value of 202 Hz. As for the nuclear rise offset, the pitch height was manipulated from relatively low to high pitch in two steps (+ 0 Hz vs. + 50 Hz to the original value). The pitch of the rise offset varied between 171 Hz and 465 Hz with the average value of 297 Hz. In total, 152 trials (19 nouns x 8 intonation contours) were presented to each participant in a random order.

2.1.3. Procedure

A speeded judgement task was conducted to categorize whether the auditory stimuli were question or statement sentences. The experiment took place in an experimental laboratory at respective universities. Four randomized lists were created presenting all stimuli (N = 152). The following randomization criteria were applied: 1. There should be at least 2 trial distance between the same word (in different manipulation) was allowed, but not *banana sun banana*). The experiment was programmed in *Presentation* (Neurobehavioral Systems).

Auditory stimuli were presented via headphones (Sony MDR-CD570). Each trial began with a sinusoid beep of 44100 Hz (= 500 ms). After an 1000 ms of silence, the auditory stimulus was presented without any visual presentation. After the offset of the stimulus, participants were then given a maximum of 3000 ms before timeout. The intertrial-interval was 1000 ms that started after participants pressed a button to answer or the timeout. The next trial was indicated with a visual presentation of the word "next". After each 19 trials, participants could take a pause for how long they needed before continuing the experiment. No feedback was provided during the experiment. Before starting, participants were given a short description of the experiment and the procedure on a piece of paper written in English. It was described that they would hear sentences and they should give an answer as soon as possible whether the heard sentences was a question or a statement by pressing one of the buttons of a button box. The aim of the study was not communicated to the participants. After reading the description, they sat in front of the computer, then clicked a button to start. The experiment lasted approximately 15 minutes. All answers and reaction times were recorded using the button box. Participants used their dominant hand for an "yes" response and their non-dominant hand for a "no" response. After the experiment, participants filled out a questionnaire form to provide their personal language backgrounds, e.g. place of birth, place in which they grew up, dialects, history of learning other languages, length of stay in another English-spoken countries and length of stay in Australia.

2.2. Results

In total, 19307 data points were recorded (127 participants x 152 trials). From these, 257 data points were discarded due to timeout, as were 254 data points for the words 'elephant' and 'mayonnaise' which sounded unnatural with certain contour combinations.

Overall, AusEng users made 36% statement responses. With this as the intercept, a generalized linear mixed-effects regression models (glmer) was built with binary responses as a dependent measure, *participant group* as a fixed factor and *participant* and *word* as random factors including random

slopes for the fixed factors [14, 15] in order to obtain a global picture of the participant performance. The model selection was carried out by eliminating factors that were insignificant as long as this elimination did not weaken the fit of the model. Fitting of the model then proceeded with backward elimination based on log likelihood ratio tests. The best model was validated by removing data points with residuals that lay beyond 2.5*SD* from the mean and the model was refitted. P values were calculated using the Satterthwaite approximation in the R-package `lmerTest`. The values in the following plots are extracted from the best-fit glmer model, and the following multiple comparisons of the model predictions were carried out using the R-package `lsmeans`. German-without listeners produced the highest proportion of statement responses (50%, β = -.68, *SE* = .19, z = -3.6, p <.001), followed by AmEng listeners (51%, β = -.46, *SE* = .13, z = -2.7, p <.001) and then by German-with (41%, β = -.35, *SE* = .22, z = -1.62, p = .1). The lowest proportion was shown by Mandarin listeners (28%, β = .74, *SE* = .16, z = 4.6, p <.001). Thus the experienced German listeners did not differ in overall proportion of statement responses from the AusEng L1 group, while the other groups did. Nevertheless, German-with, AmEng and German-without groups neither differed from one another (β = .33, *SE* = .23, z = 1.5, p = .1 between the German groups,β = .22, *SE* = .18, z = 1.2, p = .2 and β = .11, *SE* = .21, z = .5, p = .6 between AmEng and Germans without and with AusEng experience respectively). Further, these non-AusEng groups' mean responses did not differ significantly from chance level, see Fig. 1.

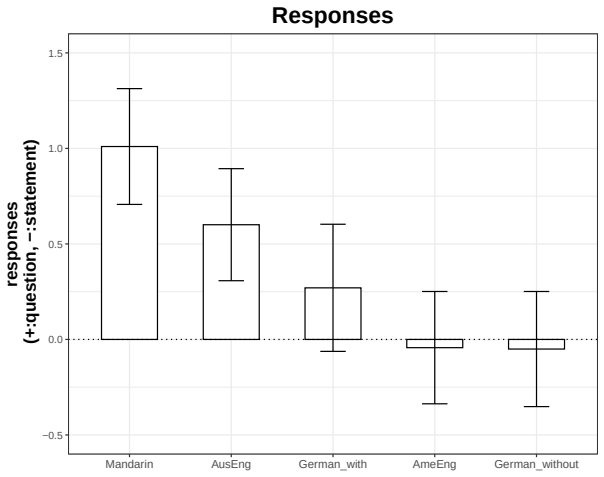

Figure 1: *Mean binary responses computed from the generalized linear mixed-effects model.*

Furthermore, a learning effect was analyzed using glmer with binary responses as a dependent measure, *participant group* and *quartile* (1–4) as fixed factors and *participant* and *word* as random factors including random slopes for the fixed factors. The results revealed an interaction between *participant group* and *item number*: Except for Mandarin listeners, all other participant groups showed some learning effect (more statement responses with an increased item number). For a better understanding of complex interactions, the data was split for each participant group. While Mandarin listeners did not show any significant differences between the quartiles (β = .02, *SE* = 2.4, z = .01, p = 1.0 in the 2. quartile, β = -.15, *SE* = 2.6, z = -.06, p = 1.0 in the 3rd quartile, β = -.17, *SE* = 2.60, z = -.07, p = 1.0 in the 4th quartile, all compared to the 1st quartile), AusEng listeners showed a large learning effect in the 4th quartile (β = .24, *SE* = .10, z = -2.3, p <.03 in the 4th quartile compared to the 1st one). AmEng listeners showed a constant increase of statement responses in the course of the experiment (β = -.32, *SE* = .09, z = -3.5, p <.001 in the 2. quartile, β = -.36, *SE* = .09, z = -3.8, p <0.001 in the 3rd quartile, β = -.38, *SE* = .1, z = -3.8, p <.001 in the 4th quartile). Some learning effect was also found for the two L2 groups (German-with group: β = -..23, *SE* = .11, z = -2.3, p <.05 in the 2nd quartile, β = -.28, *SE* = .11, z = -2.7, p <.03 in the 3rd quartile, β = -.21, *SE* = .11, z = -2.4, p <0.03 in the 4th quartile, German-without group: β = -.23, *SE* = .10, z = -2.3, p <.03 in the 2. quartile, β = -.06, *SE* = .10, z = -.67, p <0.5 in the 3rd quartile, β = -.22, *SE* = .11, z = -2.1, p <.05 in the 4th quartile), see Figure 2.

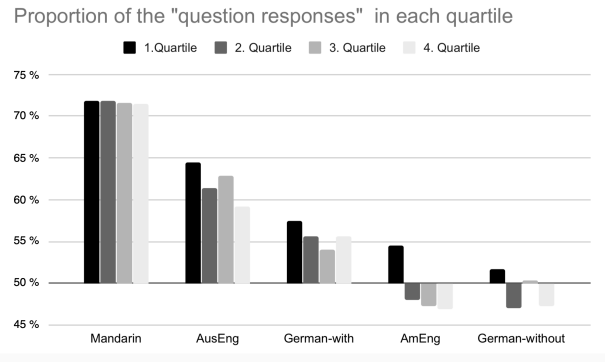

Figure 2: *Proportions of "question responses" in each quartile.*

Analyses of the binary phonetic variables (low vs. high rise onset and offset, early vs. late rise onset) revealed that only AusEng listeners gave more statement responses in the low-rise onset condition compared to the high one, see Fig. 2. This pattern corroborates [16, 6] in that low-rise onsets most strongly determined the perception of uptalk. AusEng L1 differed from the four non-AusEng groups in this: vs. Germans-with, β = .36, *SE* = .11, z = 3.2, p <.01; vs. AmEng, β = .58, *SE* = .10, z = 6.2, p <.001; vs. Germans-without, β = .29, *SE* = .10, z = 2.9, p <.001; vs. Mandarin, β = .57, *SE* = .10, z = 5.5, p <.001.

3. General Discussion

Our experiment examined the effect of listening experience on the interpretation of uptalk. Participants with differing L1 and L2 backgrounds, and differing experience with listening to an uptalk variety, heard rising contours of several types, and categorised them either as question or statement. The participant group with the most extensive experience was AusEng listeners. They showed that even with the stripped-down audio with which they were presented here, they were able to form categories in a systematic manner. Their responses relied principally on a low rise onset as evidence of an uptalk instance (i.e., a statement), which is in line with preferences exhibited in previous studies [6]. The use of uptalk in their native variety AusEng (that is: their extensive exposure to it as native listeners) has enabled them to interpret the fine-grained structure of F_0 contours, even though the precise realisations here were simple abstractions from the natural productions on which their experience was based.

Our following cross-group comparisons allowed us to

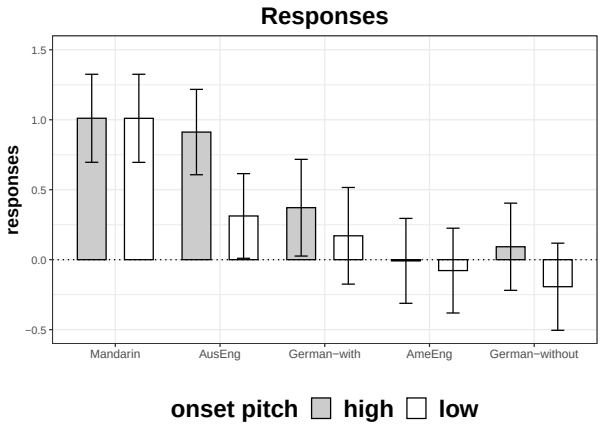

Figure 3: *Mean binary responses computed from the generalized linear mixed-effects model.*

answer a number of further questions concerning this ability, and to begin to build up a comprehensive picture of such perceptual skills.

First, we examined whether listeners whose L1 is a tone language, and hence requires extensive (lexical) use of F_0, could make use of pitch cues of the kind presented here. This proved not to be the case. Not only did these listeners not categorise the input in a systematic manner, they were also the only group to display no evidence of learning during the course of experiment. Note that in post-experiment debriefings, these listeners self reported that they chose "question" responses if they heard a rising contour (an unfortunate strategy given that all stimuli in fact had terminal rises). They were indeed the group with the highest proportion of question responses, and in consequence the lowest proportion of statement responses. They also reported not to have been aware of uptalk patterns in English. This is an interesting fact, since it suggests that their expectations of the role of F_0 may be constrained by their L1 experience, in which the role of F_0 is predominantly to convey lexical distinctions. Clearly, having more extensive experience with F_0 in general did not provide the Mandarin group with a processing advantage in the present study, even when they had acquired experience of the AusEng usage of F_0. We turn next to the German group with experience of listening to AusEng. For this group there is also a control group without experience of uptalk, so that it is possible to assess the effects of that listening experience while other experiential factors are effectively held constant. Here, neither group's result exactly resembled the AusEng response pattern; it is not the case that the AusEng exposure that the one German group had received was enough to essentially turn them into accomplished Australian listeners! However, the two German groups also did not display exactly the same pattern. Both groups showed evidence of learning across the experiment, and overall, the group with experience of AusEng produced a pattern that was closer to the AusEng pattern than was the case for the German group without such experience. We do not know how much exposure would be necessary for these listeners to pattern like the native AusEng listeners, but it is clear that, unlike the Mandarin listeners, the German group were indeed able to translate their listening experience into a degree of processing advantage.

The phonological structure of German is a lot closer to that of English than is the phonological structure of Mandarin, of course. But this closeness of phonological structure was not a major factor in the results, given that the group of German listeners without AusEng experience did not pattern like the AusEng listeners at all (if anything, their responses seem to be best described as completely random!). The German listeners with experience, however, did show a small tendency for their responses to differentiate in an appropriate direction. It is the prior experience that is the factor that has enabled that outcome. Note, however, that both the German groups exhibited some learning effect across the experiment, in that their proportion of statement choices gradually increased; in this respect, the German groups differ from the Mandarin group. We do not assign this difference either to the similarity of phonology between English and German, but maintain our interpretation that the absence of the learning factor in the Mandarin data was caused by their functional expectations with respect to F_0 patterns. Here, our final test group, the AmEng listeners, provides confirmatory evidence. This group, with English as their L1, had no advantage over the German listeners in their overall results pattern. In fact, like those Germans without experience, they seem also to have been simply guessing. Like both German groups, and unlike the Mandarin group, they did however show evidence of learning across the experiment. In other words, even when the L1 is the same, it is absolutely necessary to have had prior listening experience of an uptalk variety in order to show rapid ability to categorise uptalk-like pitch contours in a systematic manner.

We conclude, therefore, that only specific experience with a particular variety can help to build precise sensitivity to that variety's pitch contour patterns. Notwithstanding this, there is potentially good news for all language users in the general learning effect that we have observed across the course of the experiment for most listener groups. This pattern suggests that it should be possible for any listener to acquire this ability with sufficient exposure. The only bar to such acquisition arises when the listener's L1 uses the phonological feature in question (here, F_0) for quite different linguistic purposes, as was the case with our Mandarin listener group. This study has therefore taken the first step towards mapping the perception of uptalk by users of languages or varieties without uptalk; there is much more yet to be learned, but the indications are that a varietal feature such as uptalk is not likely to cause widespread communication difficulty.

4. Acknowledgements

This project was funded by an Australia-Germany Joint Research Cooperation Scheme award to Anne Cutler and Andrea Weber, i.e., for SpeechNet BaWü. We further acknowledge support from the China Scholarship Council awarded to Chi Yuan, and from the Australian Research Council awarded to Mark Antoniou and Anne Cutler. Testing of the AmEng group was enabled by the Department of Linguistics at the University of Maryland. The first and second named authors are co-first authors.

5. References

[1] J. B. Pierrehumbert, "The phonetics and phonology of english intonation," PhD Thesis, 1980.

[2] M. E. Beckman and J. B. Pierrehumbert, "Intonational structure in Japanese and English," *Phonology Yearbook*, vol. 3, p. 54, 1986.

[3] C. Ulbrich, "Belfast intonation in L2 speech," in *Proceedings of Speech Prosody 2010*, 2010.

[4] W. Wolfram and N. Schilling-Estes, *American English: Dialects and variation*. Malden, MA: Blackwell Publishers, 2006.

[5] P. Warren, *Uptalk*. Cambridge, UK: Cambridge University Press, 2016.

[6] J. Fletcher and D. Loakes, "Interpreting rising intonation in Australian English," in *Speech prosody 2010*, 2010.

[7] P. Warren and J. Fletcher, "Phonetic differences between uptalk and question rises in two Antipodean English varieties," in *Speech Prosody 2016*, 2016, pp. 148–152.

[8] M. H. K. lp and A. Cutler, "Cue equivalence in prosodic entrainment for focus detection," in *Proceedings of 17th Australasian International Conference on Speech Science and Technology*, Sydney, 2018, pp. 153–156.

[9] ——, "Juncture prosody across languages: Similar production but dissimilar perception," (submitted).

[10] J. Fletcher and D. Loakes, "Patterns of rising and falling in Australian English," in *Proceedings of the Eleventh Australasian Conference on Speech Science and Technology*, Canberra, Australia, 2006, pp. 42–47.

[11] W. F. Heeren, S. A. Bibyk, C. Gunlogson, and M. K. Tanenhaus, "Asking or telling – real-time processing of prosodically distinguished questions and statements," *Language and Speech*, pp. 1–28, 2015.

[12] J. Zwartz and P. Warren, "This is a statement? latencess of rise as a factor in listener interpretation of HRTs," *Wellington Working Papers in Linguistics*, vol. 15, no. 51–62, 2003.

[13] P. Warren, "Sociophonetic and prosodic influences on judgements of sentence type," in *Proceedings of the 15th Australiasian International Conference on Speech Science and Technology*, J. Hay and E. Parnell, Eds., Christchurch, Australia, 2014, pp. 185–188.

[14] D. J. Barr, R. Levy, C. Scheepers, and H. J. Tily, "Random effects structure for confirmatory hypothesis testing: Keep it maximal," *Journal of Memory and Language*, vol. 68, no. 3, pp. 255–278, 2013.

[15] I. Cunnings, "An overview of mixed-effects statistical models for second language researchers," *Second Language Research*, vol. 28, no. 3, pp. 369–382, 2012.

[16] J. Fletcher and J. Harrington, "High-rising terminals and fall-rise tunes in Australian English," *Phonetica*, vol. 58, pp. 215–229, 2001.