Speech Prosody 2016
31 May - 3 Jun 2106, Boston, USA

Listeners' discrimination of read and spontaneous speech is primed by performance of a prior speech production task

Rosanna Morris-Haynes [1], *Laurence White* [1] *& Sven L. Mattys* [2]

[1] School of Psychology, Plymouth University
[2] Department of Psychology, University of York

rosanna.morris-haynes@plymouth.ac.uk,
laurence.white@plymouth.ac.uk, sven.mattys@york.ac.uk

Abstract

Distinguishing read and spontaneous speech seems intuitively to be a straightforward task, but listener performance in experimental studies is highly variable. Indeed, two recent studies showed chance-level discrimination performance, suggesting that – even with relevant prosodic cues available – listeners' judgements may be heavily mediated by their contextual interpretation. Using lexically-identical map-task and read utterances previously found to be poorly discriminated despite available cues, we asked whether speech style identification could be primed by active familiarisation with the context of the speech production task. A between-subjects design with two conditions (priming vs no priming) was used. In both conditions, listeners completed a forced-choice speech style discrimination task on lexically-identical paired utterances. In the priming condition, prior to the discrimination task, listeners completed a communicative map task in pairs, equivalent to that used to generate the spontaneous speech stimuli. Although cues to speech style were available in the stimuli, performance in the no-priming condition was at chance. Discrimination performance was significantly better for subjects in the priming condition, suggesting that recent exposure to the production context of spontaneous speech promotes engagement of appropriate discrimination strategies. Indeed, subjective judgement data indicated that the priming condition increased listener awareness of relevant speech-style cues.

Index Terms: speech perception, speech style discrimination, read and spontaneous speech

1. Introduction

The prosodic characterisation of any given utterance provides crucial information for the listener alongside the utterance's lexical content. Suprasegmental variations in duration, pitch and amplitude perform a wide variety of functions, and are variously employed in indicating syntactic and semantic relationships between lexical items (e.g. when contrasting information) [1], performing discourse functions (e.g. question-marking) [2], managing speech in interaction (e.g. during turn-taking) [3,4], and signifying speaker attitudes and emotions [5]. The ways in which these functions are prosodically realised may in turn be determined by such factors as the conventions of a given interaction setting [6], the degree of formality of an interaction [7] or the relationships between those involved [8]. Just as such factors may influence prosody in speech production, the decoding of resulting prosodic events by listeners with access to the same contextual information must also be in light of these factors. It is such shared information between speakers and listeners that facilitates accurate interpretation of a given prosodic function, as defined by (and appropriate to) the context in which it occurs [9,10].

When distinguishing between speech styles, listeners must make a judgement on the original production context of a given utterance. Therefore the ability of a listener to accurately interpret the intended function of a given prosodic form may be highly relevant to determining the original production context of that speech. 'Read speech' and 'spontaneous speech' are two frequently contrasted styles, which are often found to differ in their prosodic characteristics [11,12,13,14,15,16], and to be perceptually distinguishable based on prosody alone [14, 16]. Crucially, however, considerable variation in possible elicitation circumstances makes a definitive characterisation of either speech style elusive. For example, across a sample of studies, speech rate has been found to be faster both in spontaneous speech [14,16] and in read speech [17,18], mean pitch to be higher both in spontaneous [11] and in read speech [12,16], and pitch variation to be greater both in spontaneous speech [19] and in read speech [14,11]. Where prosodic characteristics are found to be predictive of speech style - e.g. speech rate [14], mean pitch [11,12], or pitch variation [11], listeners are indeed found to orient to these cues, though no such cues are found to be exclusive predictors of listener perception [11,12,16]. The apparent absence of a robust set of defining prosodic features for read and spontaneous speech styles suggests that listeners' orientation to relevant cues might be somewhat context-specific, since the saliency of such cues is seen to vary from one speech style example to the next. Given such variation, one strategy for listeners may be to actively engage in interpreting prosodic cues relative to their presumed intended function, from which could be extrapolated a judgement regarding the original production context of the speech.

A single prosodic element may be representative of multiple functions. For example, a rising utterance-final pitch movement has a number of roles in interaction, including (but not limited to) a direct request for information or marker of uncertainty [2,11], a forward-looking marker in communicative tasks [20], or as a signifier of topic or turn continuation [4]. A falling final pitch movement, on the other hand, is often associated with declaratives [2], or with topic or turn-finality [4]. Listeners are highly adept at correctly interpreting such cues, doing so in accordance with a range of principles, generally shared by interlocutors [9,10,21,22]. Grice's maxims [9] of 'quality', 'quantity', 'relation' and 'manner' define a set of central rules according to which interlocutors produce and perceive speech, mediated by their

 doi: 10.21437/SpeechProsody.2016-48

shared context, attitudes, knowledge and experience. However, with impoverished access to the context and circumstances of production, passive listeners to pre-recorded speech – as opposed to active interlocutors – are forced to interpret prosodic features based to some extent on their own assumptions about context and speakers, which may in turn be incorrect. For example, naïve listeners are found to erroneously interpret the often flat speech of Parkinson's sufferers as representing boredom or disinterest, as opposed to the physical limitations of the condition [23], as a result of their lack of knowledge about the speaker. Furthermore, *how* listeners interpret prosodic cues when removed from the context of production can also be manipulated. Listeners' interpretation of rising final intonation as representative of a speaker's uncertainty when producing a statement can be influenced by the information the listener is given about the speaker in question [24]. Believing the speaker to be lacking in subject-specific knowledge leads to an interpretation of rising final intonation as indicating uncertainty, but this is not the case when the speaker is believed to be an expert [24]. Listeners' correct identification of the original production context of an utterance might therefore rely heavily on the interpretation they assign to a particular intonation event, such as a marker of continuation in spontaneous speech, or an indicator of uncertainty regarding content in read speech.

The present study investigates whether increasing listeners' awareness of the production context of a given speech style can facilitate their recognition of that speech style in a forced discrimination task. We hypothesise that by taking part in a communicative task (which matches the production context of the spontaneous speech under investigation) prior to completing the style discrimination task, listeners will have greater access to the context-specific knowledge seen to be crucial in the accurate interpretation of prosodic cues and, as a consequence, improve their ability to discriminate read from spontaneous speech. Results from the current context priming condition are compared to those in a no-priming condition, which was previously reported in [25].

2. Method

2.1. Design

A two-condition (priming versus no priming) between-subjects design was used. In the no-priming condition, listeners completed a forced-choice speech style discrimination task, in which they heard pairs of lexically identical read (RD) and spontaneous (SP) utterances and identified which they thought was the SP utterance in each pair. In the priming condition, listeners completed a map task with a partner before undertaking the perceptual task. We compared the discrimination performance of the listeners who had taken part in the map task themselves with that of the listeners who had not.

2.2. Participants

Listeners were 56 native British English speakers (20M, 36F, mean age 37.6 years, SD 17.4) with no reported speech or hearing impairments, and were paid, or received course credits, for their participation.

2.3. Materials

Perceptual task: Speech stimuli were extracted from a larger corpus, described in [26]. SP utterances were taken from a task in which speakers directed a partner around pictograph landmarks on a map. For the RD utterances, speakers later read aloud written transcriptions of their own spontaneous utterances. Speakers were 8 native British English speakers (4M, 4F). Four SP/RD utterance pairs were selected for each speaker. The selected SP and RD utterances in each pair were identical in lexical content, and were free from explicit cues such as laughter or interruptions, as well as disfluencies and colloquial words.

Communicative map task (priming condition): A set of maps was designed which featured pictograph landmarks similar to those used to elicit the spontaneous speech used in the perceptual task. New landmarks were designed to match the original landmarks in number of syllables (either two or three syllables), but were different from those used in the perceptual task, in order to avoid familiarization with those specific landmarks. Each pictograph was made up of an adjective followed by a physical landmark (for example 'red bus stop').

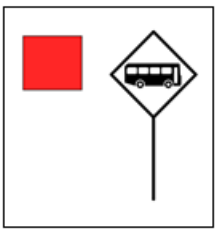

Fig 1. Example of new landmark ('red bus stop')

2.4. Procedure

Participants in the priming condition took part in the communicative map task with a partner prior to completing the perceptual task. Each pair completed the map task four times.

Listeners in both conditions (priming and no-priming) completed a forced-choice speech style discrimination task, in which they heard pairs of RD/SP utterances, and were asked to decide which utterance from each pair came from spontaneous speech. Presentation order of the utterance pairs was random, as was that of the utterances within each pair. In total, there were 32 SP/RD utterance pairs, and for each, there was a pause of 500ms between the two utterances. Following the perceptual task, all participants completed a questionnaire in which they were asked in open-ended questions to detail their strategies for speech style identification.

2.5. Statistical Analyses

Analysis of perceptual results was carried out on the raw response data from the forced choice perceptual task. For the phonetic analysis, measures of pitch and duration were extracted from the speech material. We used a series of mixed effects logistic regression models to investigate prosodic differences between speech styles, using the raw prosodic measures data. We also used mixed effects logistic regression models to investigate which cues best determined listeners'

discrimination between styles, using the raw response data ('correct' or 'incorrect'), including the random factor of utterance pair ('glmer' package in *R*, [27]). Models were compared using log-likelihood χ^2 tests.

Qualitative data in the form of listeners' descriptions of their discrimination strategies was coded according to identified key themes, whenever two or more participants had identified a particular prosodic feature as forming part of their discrimination strategy. Statistical analyses were then carried out on the coded data in SPSS.

2.6. Pitch measures

F0 mean: Mean F0 for each whole utterance.
F0 standard deviation: Mean F0 per vocalic interval was used to calculate F0 SD across the utterance
F0 range (Hz): 80% range calculated based on mean F0 values for each vocalic interval in the utterance
F0 range (ST): 80% F0 range (Hz) calculated in semitones [12*Log2(Hz)-12*Log2(origin)]
Final pitch movement (slope): Diff in Hz between min and max F0 of final stressed vocalic interval/ duration.

2.7. Durational measures

For each utterance, the spectrogram and waveform were inspected in Praat [28], and utterances segmented into consonantal and vocalic intervals. Same category intervals occurring immediately adjacent to one another were treated as one interval. Interval durations (ms) were then used to calculate a range of durational metrics [26]. 'Articulation rate variation' was introduced to address the potential relevance of temporal cues to upcoming highly informative words [13], which may differ according to the focus of the interaction.

SD Voc: Standard deviation of vocalic intervals
%V: Percentage of utterance comprised of vocalic intervals
Mean V: Mean duration of vocalic intervals
Varco V: 100 x std dev. of vocalic intervals/ mean
nFinalV: Duration of final vocalic interval/ mean vocalic interval duration for the utterance
Articulation rate: Utterance duration/ number of syllables in utterance (excluding pauses)
Articulation rate variation: Std dev. of articulation rate calculated over overlapping windows of 5 syllables

3. Results and Discussion

3.1. Acoustic differences between speech styles

To investigate acoustic differences between speech styles, we constructed and compared two mixed effects logistic regression models for each prosodic measure. The first model included only the random factors of speaker and utterance pair, with the prosodic feature under investigation added to the second model. Model pairs were compared using log-likelihood χ^2 tests. Results revealed a significant difference between speech styles in terms of mean f0 (Hz) (M_{RD} 156.5, SD 42.5, M_{SP} 146.7, SD 40.9), χ^2 (1)= 14.03, $p< .001$), and final pitch slope value (M_{RD} .02, SD 1.2, M_{SP} 1.14, SD 1.64), χ^2 (1)= 9.21 p= .002). A trend was seen in final vowel duration difference by speech style (M_{RD} 133.6, SD 46.5, M_{SP} 145, SD 57), χ^2 (1)= 2.88 p= .08). Therefore SP speech had a lower mean pitch than RD speech, and featured more frequent or pronounced final pitch rises, which may also influence the trend towards longer final vowel duration in SP speech. There were no significant differences by speech style (RD/SP) for any of the remaining pitch or duration measures investigated. The current stimuli therefore have a modest number of prosodic differences according to speech style, but remain relatively similar.

3.2. Effect of priming on listener performance

Priming was found to have a significant positive effect on listeners' discrimination performance. Mean number of correct identifications of the SP utterance from a SP/RD pair was not different from chance (14) in the 'no-priming' (NP) condition (M_{NP}= 13.9, *SD*= 2.8), *z*=-0.254, *p*= 0.8. However in the priming condition (P), the mean number of correct responses per utterance pair was significantly above chance (M_P= 19.7, *SD*= 3.6), *z*=2.51, *p*= .01. Discrimination scores in both the priming and no-priming conditions showed a main effect of both speaker ($p <.05$) and listener scores ($p <.05$). This is in line with previous studies in which significant listener variation is found [13,16]. Results therefore indicate that speech style discrimination is improved by prior exposure to comparable speech, though there remains significant variation between listeners and between how well individual speakers' styles are distinguished.

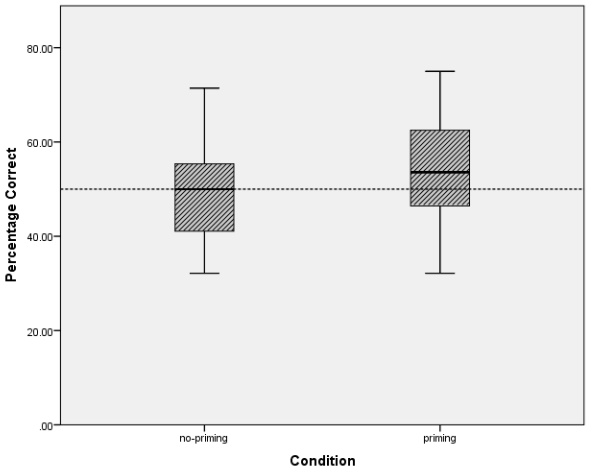

Fig.2. Percentage of correct responses per utterance pair for priming and no-priming conditions

3.3. Listener orientation to the available style cues

Mixed effects logistic regression models were conducted, using observed prosodic cues as predictors, across both conditions and for both speech styles, with the random factor of utterance pair. No relationship was found between listener performance and any combination of the observed prosodic cues (all *p*s>.05). This was true for both SP and RD speech in the no-priming and priming conditions. Therefore whilst listeners' discrimination performance was better following familiarisation with the map task, results do not show a clear relationship between performance and the salient prosodic cues observed. However, investigation of qualitative responses given after the perceptual task in both experiments

did show an increase in listeners' metalinguistic awareness of these salient cues following priming.

3.4. Qualitative responses

Analysis of qualitative responses following the priming condition revealed an increase in the percentage of participants reporting orientation to rising final pitch for SP/ falling final pitch for RD as a cue to speech style. This rose from 36% in the no-priming condition to 44% following participation in the map task. This suggests that increased familiarity with the map task made listeners more overtly aware of this cue, even if orientation to final intonation remained a difficult task, since it wasn't simply the case that a falling final pitch slope indicated read speech and rising indicated spontaneous speech (see Table 1).

	RD	SP
Rise	7	18
Fall	25	14

Table 1. Distribution of positive and negative final pitch slopes according to speech style (based on continuous measures, collapsed either side of '0')

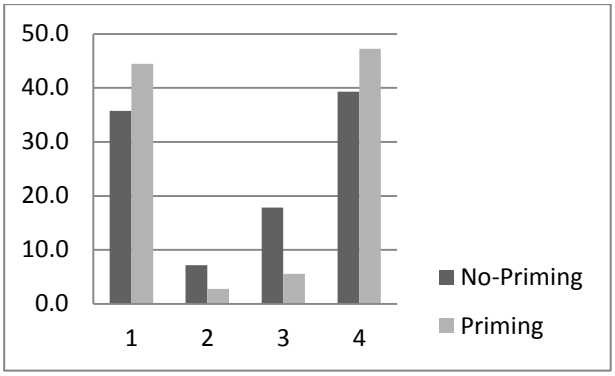

Fig.3. Percentage of participants reporting final intonation as relevant to their discrimination strategy. 1= cue reported as 'SP rising/ RD falling', 2= 'SP falling/ RD rising', 3= cue reported without details, 4= cue not reported

An increase was also seen in listeners reporting to judge speech style based on pitch across the utterance, though this cue was highlighted less frequently than final intonation (4% in the no-priming condition, 14% in the priming condition). Despite this rise however, of those listeners reporting pitch as a salient cue in the priming condition, 57% said they thought pitch was higher in spontaneous speech (the remainder did not give details as to how they oriented to this cue). Crucially, no listeners specifically identified the observed salient cue of mean pitch being higher in read speech than spontaneous. Further cues were identified as predictors by listeners which did not in fact differ between styles. This included degree of pitch variation, which 44% of listeners in the priming condition reported to be greater in spontaneous speech than read. Therefore whilst discrimination performance was improved in the priming condition, listeners also showed evidence of potentially orienting to cues which would not overall have been beneficial for successful discrimination. Finally, listeners in both conditions reported that they considered the intentions of the speaker when interpreting prosodic cues (29% of listeners in the no-priming condition and 36% in the priming condition), including such comments as: '*when they went up at the end of the sentence it was like they were asking a question, so I thought it was spontaneous'*. Therefore at least some listeners were actively engaged in attempting to use the presumed intentions of speakers' prosody choices to infer the original production context, and therefore speech style.

4. Conclusion

The results indicate that priming, in the form of active familiarisation to the context under which our spontaneous speech stimuli were produced, significantly improved performance in a pairwise read *vs* spontaneous speech discrimination task. However, listeners' orientation to the observed style cues – final intonation and mean f0 – remains unclear, since regression analyses did not relate the improved performance to these prosodic features. Previous studies suggest that listeners' speech style identification relies heavily upon the interplay between cues, including pitch, segmental durations, and voice quality [16] (the latter of which was not investigated in the present study), though a definition of this interplay remains elusive. Whilst final intonation alone was not found to determine speech style perception, it was reported by many participants as being important in their discrimination strategies, suggesting its relative saliency may have been dependent upon a range of co-occurring cues. Indeed, priming increased listeners' metalinguistic awareness of the saliency of final intonation as a style cue and so the contribution of this cue, when combined with other prosodic features, cannot be dismissed.

Despite listeners' more successful style discrimination following priming, qualitative results also showed listeners erroneously identifying misleading cues to speech style, as well as cues which were simply not relevant to the current speech. Listeners' impressions of read speech as more monotonous than spontaneous speech mirror findings from a previous study in which speech manipulated to sound monotonous was more often identified as read: this is despite read speech often being found to feature *greater* pitch variation than spontaneous speech [16]. Such findings highlight the potential role of listeners' stereotypical impressions of speech style in their interpretation of prosodic cues. Listeners are found to exhibit stereotypical biases in 'foreign accent' identification [29], for example, with commonly shared beliefs about accent-specific pronunciation traits being held by groups of similar listeners.

Although significantly above chance, the relatively low correct discrimination scores seen in the priming condition indicate that this is still a difficult task for listeners. This has also been seen in other discrimination studies using map task and read utterances [17]. If listeners' stereotypical impressions of speech style do indeed play a role in their orientation to prosodic cues, then it may be that these impressions are misleading when applied to map task speech.

Further work will focus more closely on individual listener differences, given that our results indicate that listeners' mindset and experience are important in their judgements of speech style. In particular, the current study clearly shows that placing the listener in the position of the speaker does provide a boost to their exploitation of relevant cues to the difference between read and spontaneous speech.

5. References

[1] Kurumada, C., Brown, M., Tanenhaus, M.K. 2012. Pragmatic interpretation of contrastive prosody: it *looks like* speech adaptation. *Proc 35th annual meeting of the Cognitive Science Society, Sapporo, Japan 2012, pp 647-652*

[2] Hirschberg, J., Pierrehumbert, J. 1986. The intonation structuring of discourse. *Proc 24th annual meeting on Association for Computational Linguistics*, 136-144

[3] Schegloff, E. 2000. Overlapping talk and the organization of turn-taking for conversation, *Language in Society*, 29 (1), 1-63

[4] Swerts, M., Geluykens, R. 1992. The prosodic structuring of flow in spoken discourse. Proc Workshop on Prosody in Natural Speech, Philadelphia, 221-230

[5] Frick, R.W. 1985. Communicating emotion: The role of prosodic features, *Psychological Bulletin*, 97(3), 412-429

[6] Strangert, E. 2005. Prosody in public speech: analyses of a news announcement and a political interview, *INTERSPEECH*, 3401-3404

[7] Yaeger-Dror, M. 2002. Register and prosodic variation, a cross language comparison, *Journal of Pragmatics*, 34, 1495-1536

[8] Roth, W.M., Tobin, K. 2010. Solidarity and conflict: Aligned and misaligned prosody as a transactional resource in intra- and intercultural communication involving power differences, *Cultural Studies of Science Education*, 5, 805-847

[9] Grice, P. 1975. *Logic and conversation*, from Cole, P., Morgan, J., 1975. *Syntax and Semantics 3: Speech Arts*, New York, Academic Press, 41-58

[10]Gussenhoven, C. 2002. Intonation and interpretation: phonetics and phonology, *Proc Speech Prosody 2002, Aix-en-Provence, France, April 11-13, 2002*

[11] Batliner, A., Kompe, R., Kieβling, Nöth, E., Niemann, H. 1995. Can you tell apart spontaneous and read speech if you just look at prosody? *Speech Recognition and Coding-New Adventures and Trends*, 101-104

[12] Blauuw, E. 1991. Phonetic characteristics of spontaneous and read-aloud speech, *PPoSpSt-1991*, paper 012.

[13] Blauuw, E. 1994. The contribution of prosodic boundary markers to the perceptual difference between read and spontaneous speech, *Speech Communication* 14(4), 359-375

[14] Dellwo, V., Leemann, A., Kolly, M.-J. 2015. The recognition of read and spontaneous speech in local vernacular: The case of Zurich German. *Journal of Phonetics*, 48, 13-28

[15] Levin, H., Schaffer, C.A., Snow, C. 1982. The prosodic and paralinguistic features of reading and telling stories. *Language and Speech*, 25 (1), 43-54

[16] Laan, G. 1997. The contribution of intonation, segmental durations, and spectral features to the perception of a spontaneous and a read speaking style. *Speech Communication*, 22, 43-65

[17] Mixdorff, H., Pfitzinger, H.R. 2005. Analysing fundamental frequency contours and local speech rate in map task dialogs. *Speech Communication*, 46, 310-325

[18] Howell, P., Kadi-Hanifi, K. 1991. Comparison of prosodic properties between read and spontaneous speech material. *Speech Communication*, 10, 163-169

[19] Daly, N., Zue, V. 1992. Statistical and Linguistic Analyses of F0 in Read and Spontaneous Speech, *Int Conf on Spoken Language Proc*, Vol 1, Pages 763-766

[20] Lai, C. 2014 Interpreting final rises: task and role factors, *Proc Speech Prosody 2014*, Dublin

[21] Ohala, J.J., 1983. Cross language use of pitch: an ethological view, *Phonetica, 40*, pp 1-18

[22]Hirschberg, J., (2002). The pragmatics of intonational meaning. *Proc Speech Prosody 2002, Aix-en-Provence, France, April 11-13, 2002*

[23]Jaywant, A., Pell, M.D. 2009 Listener impressions of speakers with Parkinson's disease, *Journal of the International Neuropsychological Society, 16*, 49-57

[24] Tomlinson, J.M., Fox Tree, J.E. 2011. Listeners' comprehension of uptalk in spontaneous speech. *Cognition*, 119, 58-69

[25] Morris-Haynes, R. White, L., Mattys, S.L. 2015. What do we expect spontaneous speech to sound like? *Proc International Congress of Phonetic Sciences 2015, Glasgow*

[26]White, L., Mattys, S.L., Wiget, L. 2012. Segmentation cues in spontaneous speech: Robust semantics and fragile phonotactics. *Frontiers in Psychology*, 3, 375

[27] Baayen, R.H., Davidson, D.J., Bates, D.M. 2008. Mixed-effects modelling with crossed random effects for subjects and items, *Journal of Memory and Language*, 59(4), 390-412

[28] Boersma, P., Weenink, D. 2014. Doing phonetics by computer, http://www.fon.hum.uva.nl/praat/

[29] Neuhauser, S., Simpson, A.P. 2007. Imitated or authentic? Listeners' judgements for foreign accents, *Proc of ICPhS, Saarbrücken, 6-10 August 2007*