What do we expect spontaneous speech to sound like?

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

[1]School of Psychology, Plymouth University, UK., [2]Department of Psychology, University of York, UK.
rosanna.morris-haynes@plymouth.ac.uk; laurence.white@plymouth.ac.uk; sven.mattys@york.ac.uk

ABSTRACT

Listeners have been shown to distinguish text read aloud from spontaneous speech, with a range of prosodic features suggested as cues to speech style. However, significant variation is seen across studies, both in speech elicitation methods and in the nature of listeners' orientation to prosodic cues. We asked whether listeners could distinguish spontaneous 'map task' speech from lexically identical read utterances. Experiment 1 found that, although our spontaneous speech differed prosodically from read speech, listeners did not appear to use available cues to distinguish styles. Experiment 2 found that, even when matched spontaneous and read utterances were presented consecutively, listeners still did not reliably discriminate between styles despite available cues. We suggest that listeners' ability to distinguish between speech styles derives from the interaction of expected and available cues, including prosody, mediated by listeners' interpretation of such cues as being representative of speech context and the intentions of the speaker.

1. INTRODUCTION

'Read speech' and 'spontaneous speech' are terms broadly used to refer to two typically contrasted speech styles, though the terms themselves do not refer to any fixed set of conventions or inherent set of prosodic features. Nevertheless, read and spontaneous speech are often described in terms of their differences, at syntactic, lexical and prosodic levels [12,15,18], and are suggested to be perceptually distinguishable based on prosody alone [2,15]. A typical conceptual contrast made between these two styles is of spontaneous speech as informal, dynamic and unrehearsed (as from a casual conversation) versus read speech as scripted and formal (as from a news reader) [13]. However, as is often highlighted [4,8,13], if one considers the read speech of an actor or the spontaneous speech of a political public speaker, it is clear that read and spontaneous speech regularly deviate from such 'typical' incarnations. A range of tasks, topics, and relationships between interlocutors serve to define versions of read or spontaneous speech for specific within-study comparison. Given such differences in elicitation [3,8,13,17], the prosodic characteristics associated with either style naturally vary between studies. For example, faster speech rate has been found in both spontaneous speech [13,8], and read speech [17,11]. Similarly, higher mean pitch has been seen in both spontaneous speech [2] and read speech [3,13], and pitch variation has been found to be greater in both spontaneous speech [6] and read speech [8,2]. Furthermore, significant between-speaker variation in these contrasts is seen within studies [2,4,7,8,13,17].

Despite this variation, listeners are typically able to distinguish read and spontaneous speech [2,3,4,8,13,15]. This is true, even when listeners are presented with lexically-identical utterances in the two styles, potentially putting the focus for discrimination on prosodic cues [13,8]. However, how listeners actually make use of these cues appears rather complex. In discrimination experiments in which prosodic characteristics are found to differ significantly between speech styles - e.g. speech rate [8], mean pitch [2,3], pitch variation [2] or boundary marking behaviour [4], these cues have indeed been found to contribute to listeners' perception of speech style. Crucially however, individual cues are not found to be exclusive determiners of listeners' perception [2,3,13] and the reliance on such cues varies greatly between listeners. Furthermore, prosodic differences may be found between speech styles which seem to be ignored by listeners within the perceptual task [8]. In addition it has been found that speech style discrimination by an automatic classifier is less successful when trained on listeners' (above chance) speech style judgements than when trained on prosodic features present in the same speech [2], suggesting that listeners do not make full use of the style cues available to determine speech style.

Finally, effective listener discrimination of speech styles is not universally observed [17], suggesting that some examples of read and spontaneous speech are not sufficiently different for listeners to tell them apart. Mixdorff and Pfitzinger [17] suggested their read utterances were not 'typical' of the style, being embedded within a dialogue, which may have prompted speakers to return "into the mood of the original interaction" when reading [17]. Indeed, many studies refer to the notion of 'typical, 'good' or 'poor' examples of read or spontaneous speech, with deviation from ‘typical’ examples of either speech style often attributed to the details of the elicitation technique [8,13,17].

Whether considered ‘typical’ or not, it is clear that there is significant variation in the incarnations of ‘read aloud’ speech and speech produced spontaneously. That listeners experience variable difficulties in distinguishing between styles may point to the involvement of a top-down approach, in which listeners anticipate specific cues based on a conceptualised ‘typical’ version of either style. Furthermore, listeners are found to apply different interpretations to the same speech events based on knowledge or assumptions about the speaker or production context [1,21]. It may be that this sensitivity to speaker context also forms part of a top-down approach to distinguishing speech style, in which listeners also interpret cues as representative of the assumed intentions or circumstances of the speaker. When attempting to determine speech style therefore, listeners may simultaneously be attending to available cues, listening for *expected* cues, and applying interpretations to cues in relation to presumed speaker or production context.

This study investigates whether listeners are able to tell apart examples of read and spontaneous speech which might not be considered ‘typical’ of each style, and whether significant prosodic differences exist between these variants. We also investigate whether prosodic characteristics influence listeners’ perception of style, and may therefore point to particular expectations about the characteristics of read and spontaneous speech.

2. EXPERIMENT 1

2.1. Method

2.1.1. Participants

Listeners were 76 university students (62 F, 14 M, mean age 20 years (*SD* 3.3, range 18-26), native British English speakers with no reported speech or hearing impairments.

2.1.2. Materials

Speech data was extracted from a larger corpus, described in White, Mattys and Wiget [23]. Spontaneous (SP) utterances were taken from a task in which speakers directed a partner around landmark pictographs on a map. For the read (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 items in each pair were identical in lexical content, and were free from explicit cues such as laughter or interruptions from the dialogue partner (in the SP corpus).

2.1.3. Procedure

Participants were instructed to rate a series of individual utterances on a 7-point Likert scale according to how likely it was that the utterance had come from SP or RD speech (1= extremely likely ‘spontaneous’, 7= extremely likely ‘read’). All 32 RD and 32 SP utterances were heard in pseudo-random order (not more than three consecutive utterances of the same style), with a different order of presentation for each participant.

2.1.4. Statistical analysis

Analysis of perceptual results was carried out on the raw response data (rating of 1 to 7 for each utterance). ‘Correct response’ scores were obtained by collapsing ratings 5-7 given to RD speech as ‘correct’, and ratings 1-3 given to SP speech as correct. ‘Neutral’ ratings of 4 were not included in correct response analyses. For the phonetic analysis, measures of pitch and durational metrics were extracted from the speech material.

2.1.5. 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.

Pitch measures were converted to z-scores to normalise between speakers, using the formula:

(1)

$$z = \frac{X-\mu}{\sigma}$$

2.1.6. Durational measures

For each utterance, the spectrogram and waveform were inspected in Praat [5], and boundaries between consonantal and vocalic intervals identified. Same category intervals which occurred immediately adjacent to one another were treated as one interval.

Interval durations (ms) were then used in the calculation of the following durational metrics [23].

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

2.2. Results and Discussion

2.2.1. Listener Performance

Overall, listeners were above chance in correctly identifying read or spontaneous utterances, though performance was relatively poor, with a mean correct score by utterance of 55.1% (*M* 41.9, *SD* 14.4, *t*(63)= 2.2, *p*= .032). Previous studies have found disfluencies and colloquial wording to be amongst the more effective markers to speech style [15,17], therefore the effect of such cues on utterance ratings was checked before analysing the contribution of prosodic characteristics. Indeed, some utterances did feature either disfluencies (mispronunciations, false starts, etc., N=12/64 (RD=2/32, SP=10/32)), or the colloquial, concatenative ‘gonna’ (N=8/64 (RD=4/32, SP=4/32)). Mann-Whitney U tests showed that utterances that featured ‘gonna’ were rated as ‘more spontaneous’ (*N*=8, mean rank=17) than those that did not (*N*=56, mean rank=34.71, *U*=100.00, *z*= -2.52, *p*= .012), and that utterances which featured disfluencies were rated as significantly ‘more spontaneous’ (*N*=12, mean rank=19.04) than those that did not (*N*=52, mean rank=35.61, *U*=150.5, *z*=-2.78 *p*= .005). Disfluencies and ‘gonna’ thus appear to influence listener perception of style and so such utterances were removed from further analysis. Following this step, performance with the remaining utterances was no longer above chance (*M* 40.1, *SD* 14.0, *t*(46)= 1.0, *p*= .306), and no difference was found between speech styles in correct responses (M_{RD} 43.1, *SD* 13.8, M_{SP} 36.4, *SD* 13.5), *t* (45) = 1.7, *p*= .105). The failure to distinguish the two styles does not reflect typical findings [2,3,4,8,13,15], although poor listener discrimination has been found in another study which compared spontaneous utterances, elicited using a map task, with their transcribed read counterparts [17].

2.2.2. Acoustic cues to differences between speech styles

Analysis of z-score-normalised pitch measures revealed higher mean F0 in RD speech (M_{RD} .4, *SD* .8, M_{SP} -.4, *SD* 1.0), *t* (45) = 3.3, *p*= .002) and a positive mean final pitch slope in spontaneous speech (M_{RD} -.4, *SD* .9, M_{SP} .4, *SD* .9), *t* (45) = -3.2, *p*= .003). There were no significant differences by speech style (RD/SP) in any of the durational metrics investigated, including articulation rate, which has previously been found to differ between styles [8]. The nature of the map task is such that in order to facilitate understanding and therefore task completion, speakers may adopt a slower, more careful style for the benefit of their partner [16]. Therefore, despite being produced ‘spontaneously’, utterances in the present study may feature some of the characteristics of clear speech [9,19], which could serve to narrow the perceptual gap between styles.

A stepwise forward logistic regression with speech style as dependent variable and all acoustic measures as factors found mean F0 and final pitch slope to be significant predictors of actual speech style ($\chi^2(2)$= 16.931, *p*<.001). However, a further stepwise forward logistic regression with *perceived* speech style as the dependent variable found no significant predictors of listeners’ judgements of style. Therefore whilst there were pitch cues to speech style available, listeners were not reliably orienting to these cues. Dellwo et al. [8] similarly found that although F0 variability (SD F0) differed between their examples of read and spontaneous speech, degree of F0 variation did not account for listener performance. Likewise, Mixdorff and Pfitzinger [17] found a greater incidence of (rising) ‘non-terminality’ and ‘establishing contact’ final pitch movements in spontaneous speech, yet successful discrimination between styles was largely attributed to the presence of non-linguistic markers such as fillers (which we eliminated from this analysis, as noted above). This failure to exploit potentially useful cues may have been exacerbated by the presentation of utterances in random order, with substantial variation between the examples from which listeners could attempt to calibrate their judgements. We hypothesised therefore that presenting RD/SP utterance pairs side-by-side might facilitate recognition, as it would give listeners the chance for a more direct comparison between styles.

3. EXPERIMENT 2

3.1. Method

3.1.1. Participants

Listeners were native British English speakers (11M, 17F, mean age 38.4, SD 16.2) with no reported speech or hearing impairments, and were paid £4.

3.1.2. Materials

The speakers used in Experiment 1 contributed 4 SP/RD utterance pairs each. Within each pair, the SP and RD utterances were lexically identical. All were free from interruptions, pauses, disfluencies and colloquialisms. To achieve this, 13 of the 32 utterance pairs from the Exp 1 set were replaced with new utterances from the same corpora.

3.1.3. Procedure

Listeners heard pairs of SP/RD utterances, randomly presented in either order, with a pause of 500ms between the two utterances, and were required to decide which utterance from each pair was from spontaneous speech. There were 32 SP/RD utterance pairs, pseudo-randomly ordered.

3.2. Results and Discussion

3.2.1. Listener performance

Overall, identification of the SP utterance from a SP/RD pair was at chance (*M*= 13.9, *SD*= 2.8), *t* (63) = -.4, *p*= .724). The opportunity to compare one utterance directly against its counterpart of the opposing style did not make the task any easier for listeners.

2.2.2. Acoustic differences between speech styles as produced and between perceived speech styles

As in Experiment 1, there were no differences according to speech style for any of the durational metrics. Investigating z-score-normalised pitch measures revealed differences between RD and SP speech for final pitch slope, with a positive mean pitch slope again seen in spontaneous speech (M_{RD} -.4, *SD* .8, M_{SP} .4, *SD* 1.1), *t* (62) = -3.1, *p*= .003). A linear regression found no significant relationship between the total number of correct responses per pair and the final pitch slope value for either read (*p*>0.05) or spontaneous utterances (*p*>0.05). Thus, whilst pitch slope again differed between styles (based on a partially overlapping sample), listeners did not use this cue to discriminate styles, even when able to make a direct pairwise comparison between utterances.

4. CONCLUSION

The poor performance in the current study suggests that listeners were unable to develop successful strategies to determine speech style based on the cues available. As has been discussed, the task may have been made more difficult due to a narrowing of the potential differences between speech styles (and therefore a reduction in potential style cues), as a result of speakers adopting a speaking style more akin to clear speech [19] during the map task. However, final pitch slope was found to differ between speech styles. Beyond marking overt questions in interaction, a rising utterance-final pitch movement may be used to signify uncertainty [22], to mark topic or turn continuation [20], or to aid co-ordination across a communicative task [14]. A falling final pitch movement on the other hand may often be associated with declaratives [10], or with topic or turn-finality [20]. Current differences in final pitch slope between styles may have occurred for a number of reasons. SP speech was taken from a context – direction-giving – in which speakers required regular feedback in order to complete the task, whereas the read speech did not form part of an interaction. Furthermore, SP utterances were often medial in longer speaker turns, increasing the potential for utterance-final rising pitch movements relating to both turn and topic continuation. RD utterances were always presented as single sentences ending with a full stop, in a non-communicative environment, thereby implying ‘turn’ finality. Differences in final pitch slope did not, however, appear to be a salient cue to speech style for listeners. Notably, map task speech has been found to feature a higher prevalence of rising final pitch slopes than conversational speech [14], therefore orienting to this cue might not have been an obvious strategy for listeners. Whilst the spontaneous and read speech examples used in the current experiments could be considered ‘typical’ within their respective production contexts, to listeners who had no knowledge of the circumstances of speech elicitation the resulting prosodic characteristics may have been relatively obscure as cues to speech style. The poor discrimination performance seen in the current study may therefore be related to multiple challenges faced by listeners - namely, being presented with perceptually similar, ‘non-typical’ speech style examples and having no access to speech production context from which to recalibrate their expectations.

6. REFERENCES

[1] Arnold, J.E., Hudson Kam, C.L., Tanenhaus, M.K. 2007. If you say thee uh you are describing something hard: the online attribution of disfluency during reference comprehension. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 33, 914-930

[2] 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

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

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

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

[6] 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

[7] Dellwo, V., Leemann, A., Kolly, M.-J. 2012. Speaker idiosyncratic features in the speech signal. *Proc Interspeech 2012*, 1584–1587

[8] 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

[9] Hazan, V., Baker, R. 2010. Does reading clearly produce the same acoustic-phonetic modifications as spontaneous speech in a clear speaking style? *DiSS-LPSS Joint Workshop 2010*

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

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

[12] Kowal, S., Bassett, M.R., O'Connell, D.C. 1985. The spontaneity of media interviews. *Journal of Psycholinguistic Research*, 14, 1-18

[13] 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

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

[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] Lindblom, B. 1990. Explaining phonetic variation: A sketch of the H&H Theory. In: Hardcastle, W., Laver, J. (eds), *Speech Production and Speech Modelling*. Dordrecht: Kluwer, 403-439.

[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] Remez, R.E., Rubin, P.E., Nygaard, L.C. 1986. On spontaneous speech and fluently spoken text: Production differences and perceptual distinctions. *Journal of the Acoustical Society of America*, 79, S26

[19] Smiljanic, R., Bradlow, A.R. 2008. Temporal organisation of English clear and conversational speech. *Journal of the Acoustical Society of America*, 124, 3171-3182

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

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

[22] Ward, G., Hirschberg, J. 1985. Implicating uncertainty: the pragmatics of fall-rise intonation. *Language*, 61, 747-776

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