ON THE PERCEPTUAL CLASSIFICATION OF SPONTANEOUS AND READ SPEECH

Eleonora Blaauw
Research Institute for Language and Speech
Utrecht, The Netherlands

ABSTRACT

What prosodic cues do listeners use to classify speech as read or spontaneous? Two different types of spontaneous speech and matching read samples were examined. Results suggest that read speech is characterized by a typical prosodic make-up. Spontaneous speech seems to be characterized by the absence of that typical prosodic make-up; digressions from the 'read' prosody, in any direction, appear to induce listeners to classify the speech as spontaneous.

INTRODUCTION

When we hear someone speak, it is intuitively very easy to tell whether the speaker is talking spontaneously, or whether he is reading a text out loud. Research has confirmed this informal observation, and has shown that prosodic cues are important for the perceptual distinction [1,2].

The spontaneous-read distinction is not as simple and straightforward as it might seem at first glance. One can distinguish many different spontaneous and read styles, for example along a formal-informal dimension, or a careful-casual dimension [3]. These interfering dimensions do not seem to confuse listeners in making spontaneous-read judgments, however. This suggests that the spontaneous-read distinction constitutes a basic and meaningful difference to listeners. In addition, the cues that they use must be retrievable from different types of spontaneous and read speech. In this paper we aim to identify some of these cues. As it has already been shown that prosody plays an important role for the spontaneous-read distinction, we will limit our search to prosodic cues.

In looking for reliable prosodic cues to the spontaneous-read distinction, we should concentrate on characteristics that reflect fundamental underlying differences between spontaneously produced speech and speech read from text. After all, such characteristics should surface in any type of spontaneous and read speech, which renders them reliable.

Spontaneous speech has several fundamental characteristics that distinguish it from read speech, no matter how formal or informal the situation in which it is produced, or how careful or casual the produced speech. Spontaneous speech is produced impromptu, on the spot, which entails that much planning activity is required on the part of the speaker. In addition, it entails that spontaneous speech is highly flexible, and can be optimally adapted to the communicative situation. Read speech is largely prepared beforehand, at the stage where the text is written. The planning required from the speaker is therefore limited. In addition, the possibilities to adapt the speech to the communicative situation is limited; at the actual time of production, read speech is much less flexible than spontaneous speech.

In this paper we will focus on the fundamental difference in flexibility between spontaneous and read speech. It seems plausible that this difference affects the prosodic characteristics of spontaneous and read speech in a distinctive way, so that those characteristics form reliable cues for the perceptual classification of spontaneous and read speech.

The larger flexibility in spontaneous speech production is likely to be reflected in a highly flexible and variable prosody, strongly dependent on the communicative situation in which the speech is produced. The prosodic characteristics of a spontaneous intimate conversation between two close friends are bound to differ greatly from the prosodic characteristics of a spontaneously produced formal speech.

The lesser amount of flexibility in read speech may result in less variable prosody across different samples of read speech. A speaker reading a personal letter out loud to a friend or reading a speech out loud during a formal meeting will probably produce two speech samples with fairly similar prosodic characteristics. In addition, these prosodic characteristics are likely to be 'neutral', positioned somewhere in the middle of all possible spontaneous characteristics. Read speech is typically clear and careful, whereas spontaneous speech can easily digress in any direction. It can be casual and sloppy for example, but it can also be emphatic and highly expressive. It can be slow and hesitant, but it can also be produced at very high rates.

We think that listeners have little trouble in identifying all those different types of spontaneous speech correctly as spontaneous, despite the fact that the prosodic make-up of the speech varies enormously. Or perhaps we should say, thanks to the fact that the prosodic make-up of the speech varies enormously. If read speech indeed shows fairly stable, idiosyncratic prosodic characteristics, telling spontaneous and read speech apart becomes an easy task. Whenever the speech listeners are presented with exhibits those typically read prosodic characteristics, the speech can be classified as read. If the speech shows digressing prosodic characteristics, in any direction, the speech can be classified as spontaneous.

In summary, we hypothesize that spontaneous speech shows vastly different prosodic characteristics, dependent on the situation in which it is produced, whereas read speech shows relatively stable, average prosodic characteristics. Furthermore we hypothesize that listeners can classify different types of spontaneous and read speech correctly. They may do this by classifying speech with typically read characteristics as 'read aloud', and speech with characteristics that digress from the read values as 'spontaneous'.

To test these hypotheses, two speech corpora were selected. The first corpus consisted of casual spontaneous speech produced in an informal interview situation, and matching read speech (i.e. a read version of the interview based on a transcript of the spontaneous speech). The second corpus consisted of careful spontaneous speech, viz. so-called instruction monologues [4], and matching read speech. A selection of fluent spontaneous utterances and their read counterparts was made from both corpora. Classification judgments were elicited for those utterances. Subsequently, the values of four prosodic parameters were established for each of the selected utterances. These prosodic characteristics were then correlated with the percentage of 'spontaneous' judgments obtained from the listeners.

THE INTERVIEW CORPUS

The interview corpus consisted of one and a half hours of spontaneous speech produced by one male speaker, and a read version of large parts of the original interview produced by the same speaker, read from a written transcript. From this corpus 48 fluent spontaneous utterances and 48 matching read utterances were selected. In a perception experiment, each individual utterances was presented to 10 listeners, who were asked to classify each utterance as spontaneous or read. The average classification score was 79% correct (81% for the spontaneous utterances and 77% for the read utterances).

For each utterance we determined mean F0, standard deviation of F0, F0 range (both measures of the amount of F0 variation), and articulation rate. F0 range was defined as the distance between the lowest and the highest F0 value in each utterance. Standard deviation of F0 and F0 range are expressed on a logarithmic scale, in semitones. Articulation rate was defined as the number of syllables per second, excluding pause time.

Table 1: Acoustic characteristics of spontaneous and read utterances selected from the interview corpus; correlation (Pearson's r) between prosodic characteristics and percentage of 'spontaneous' judgments.

	spont.	read	r
mean F0 (Hz)	132	157	-.63
s.d. F0 (ST)	2.0	2.6	-.51
F0 range (ST)	8.2	10.4	-.48
art. rate (syll/s)	7.1	6.3	.52

In addition, Pearson's correlation was determined between the prosodic characteristics and the percentage of 'spontaneous' judgments for each utterance. The results are shown in Table 1.The differences between the two speech styles, as determined with a paired t-test, is significant at the 1% level for all four prosodic parameters. The t-values (df = 47) are 11.5, 5.3, 5.0, and -6.7 respectively. The correlation coefficients are significant at the 1% level as well. Thus, a lower mean F0, a smaller standard deviation of F0, a smaller F0 range and a higher articulation rate are significantly associated with more 'spontaneous' judgments. For more details on the collection and characteristics of this corpus the reader is referred to [5].

THE INSTRUCTION MONOLOGUE CORPUS

Spontaneous instruction monologues were collected from five male speakers. They were asked to give instructions to a listener on how to assemble the front view of a house from a set of cardboard pieces. Both speaker and listener had the same set of building blocks in front of them. They could not see each other, and the speaker did not receive any feedback from the listener. The monologues each lasted about five minutes. The spontaneous monologues were transcribed orthographically, and subsequently each monologue was read aloud by the original speaker. From this corpus 109 fluent spontaneous utterances and the 109 matching read counterparts were selected, divided over the five speakers. For more details on the collection of the corpus, the reader is referred to [6].

In a perception experiment, the selected utterances were presented individually to 21 listeners, who were asked to classify each utterance as spontaneous or read. The average classification score was 77% correct (79% for the spontaneous utterances and 75% for the read utterances).

For the utterances from the interview corpus we also determined mean F0, standard deviation of F0, F0 range, and articulation rate for each utterance. These measures were defined and determined in the same way as for the interview corpus (see above). The results are presented in Table 2. In addition, Pearson's correlation was determined between the prosodic characteristics and the percentage of 'spontaneous' judgments for each utterance. To compensate for absolute differences between the five speakers, z-scores were used for the correlation study. These z-scores were calculated separately for each speaker, across both speech modes. The correlation coefficients are shown in the last column of Table 2.

Table 2: Acoustic characteristics of spontaneous and read utterances selected from the instruction monologues; correlation (Pearson's r) between prosodic characteristics (z-scores) and percentage of 'spontaneous' judgments.

	spont.	read	r
mean F0 (Hz)	128	122	.27
s.d. F0 (ST)	3.2	2.9	.19
F0 range (ST)	13.9	11.7	.31
art. rate (syll/s)	5.2	5.8	-.58

The results show a reversal of the spontaneous-read differences in comparison to the interview corpus for all four prosodic variables The difference between the speech styles, as determined with a paired t-test, is significant at the 1% level for all four prosodic parameters. The t-values (df = 108) are 4.5, 4.2, 5.8 and -7.9 respectively.

A comparison between Tables 1 and 2 leads to the following observations. The characteristics of the read speech samples are, as predicted, fairly similar in both corpora, and intermediate between the values for the spontaneous interview and the spontaneous instruction monologues. Thus, the values for the spontaneous speech samples can be said to digress from the stable 'norm' values for the read speech. Mean F0 forms an exception; its value in the read samples does not lie between the values in the spontaneous samples. This is due to the fact that mean F0 is highly speaker-dependent; we did not use the same speakers in both corpora. In order to make the mean F0 values from both corpora comparable, they should be standardized, for example by expressing them in terms of the distance to the bottom of the speaker's range. We did not have the necessary data to carry out this standardization. For now, we will just assume that, had we used the same speakers, mean F0 would have shown the same pattern as the other prosodic parameters.

Thus, the production results seem to confirm the hypothesis that read speech shows stable prosodic characteristics, whereas the prosodic characteristics of spontaneous speech digress from the read characteristics in any direction.

The correlation coefficients for mean F0, standard deviation of F0 and F0 range are much smaller than they were in the interview corpus. Nevertheless, all four correlation coefficients are significant at the 1% level. Moreover, they are all reversed in comparison to the interview corpus. Thus, in this corpus, a higher mean F0, a larger standard deviation of F0, a larger F0 range and a lower articulation rate are associated with more 'spontaneous' judgments.

CONCLUSION AND DISCUSSION

The experiments described in this paper showed that listeners are able to identify different types of speech correctly as spontaneous or read. The prosodic characteristics of the two spontaneous samples showed large differences, whereas the two read samples both showed more or less the same average prosodic characteristics. The stronger the prosodic characteristics of an utterance digressed from these 'read' values, the larger was the percentage of 'spontaneous' classification judgments from the listeners.

It would be premature to conclude that all read speech shows typically read values on a whole range of prosodic parameters, by which a listener can recognize the speech as read. First of all, we only looked at a few prosodic parameters. Second, the method by which the two read samples used in the present study were collected biases the results towards this conclusion. Although the texts were based on different types of spontaneous speech, the settings in which the read samples were recorded were similar. Both read samples were examples of straightforward laboratory readings of a coherent running text. This is inevitable when one wants to collect matching spontaneous and read speech. However, future research should include read speech collected outside the laboratory, in different communicative settings. Possibly, such read samples will show larger prosodic differences than the present read speech samples. We maintain, however, that the lack of flexibility in read speech production will seriously limit the possible variation in prosodic characteristics. In some special cases this limitation may be overcome, for example when 'reading' a thrilling story to a child. In such a speech sample the prosodic characteristics will digress strongly from the average read values. However, we imagine that in a listening test such speech material would not be classified as read, but as enacted, or in some cases even as spontaneous speech.

REFERENCES

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

[2] Laan, G. & Van Bergem, D. (1993), "The contribution of pitch contour, phoneme durations and spectral features to the character of spontaneous and read aloud speech", *Proceedings Eurospeech '93, Berlin*, Vol.1, pp. 569-572.

[3] Eskénazi, M. (1993), "Trends in speaking style research", *Proceedings Eurospeech '93, Berlin*, pp. 501-505.

[4] Terken, J. (1984), "The distribution of pitch accents in instructions as a function of discourse structure", *Language and Speech,* Vol. 27, pp. 269-290

[5] Blaauw, E. (1992), "On the perceptual difference between read and spontaneous speech", In: M. Everaert, B. Schouten and W. Zonneveld (eds.), *OTS Yearbook 1992*, Utrecht: LEd, pp. 1-16.

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