WHAT DO SPECTRAL AND PERCEPTUAL ANALYSES REVEAL ABOUT SPONTANEOUS SPEECH IN DIALOGUES OF DIFFERENT STYLE?

Lioba Faust
Institut für Kommunikationsforschung und Phonetik
Poppelsdorfer Allee 47, 53115 Bonn, Germany
Phone: +49-228-735641 E-Mail: lfa@asl1.ikp.uni-bonn.de

ABSTRACT

The analyses of segmental changes, vowel durations of [i:, ɪ] and [a, ɑ:], and listeners' classifications of utterances of three different variants of German dialogues reveal that there are no crucial characteristics to classify speech as either spontaneous or read. The casual spontaneous style shows the strongest phoneme and syllable reductions and is generally classified correctly. For the careful spontaneous and for the read utterances listeners' classifications vary strongly with respect to the speakers.

INTRODUCTION

In automatic speech recognition a shift of interest can be observed from read speech towards spontaneous speech. The variability of the speech signal is expected to be higher in spontaneous than in read speech and even higher in casual than in careful spontaneous speech. Nevertheless, linear modifications between the styles cannot be assumed [2]. Furthermore, read speech must not be considered as a contrast to spontaneous speech, but may show as much stylistic variation as spontaneous speech [3]. Therefore, the purpose of the experiment was to have a look at variability in three *natural types of conversation* together with as little restrictions as possible on the controlled elicitation of speech, which is both of great phonetical interest and indispensable for an automatic treatment of everyday conversation. For the same reason, we accepted the fact that speech material would be linguistically and phonetically different and rejected the restriction on relatively small units of speech, e.g. sequences of sounds or isolated sentences.

METHOD OF EXPERIMENTATION

Corpus design and recordings

Four female German students between 23 and 27 years of age participated in an experiment of dialogue recordings of the following different speaking styles: 1. casual speech: totally free conversation, 2. careful spontaneous speech: time-scheduling negotiation dialogue using the formal mode of address with the relevant dates given in a calendar, 3. read speech: rereading of the transcribed utterances of the second dialogue variant. Hesitations, word repetitions and repairs that had been transcribed were generally dropped for the copy to be reread. The most important issue for the rereading copy was to preserve the dialogue structure and, in general, the grammatical structure of the utterances.

The speakers were sitting in the institute's speech laboratory, in two neighbouring rooms separated by a glass pane. The communication was performed using headset microphones (Sennheiser HMD 414-6). Speech was digitally recorded on separate channels.

The casual dialogues were recorded first without the speakers' knowledge of being recorded. While the supervisor went away under a pretext for about three minutes, the speakers where left on their own. This led to very different dialogue structures. Moreover, the acoustic conditions such as a constant distance between the speakers' mouth and the microphone could no longer be controlled. This was accepted, since speech under close to natural conditions was desired. For the same reason, we could not use the headset, therefore two condenser microphones

(Neumann KM 140) were used. For spectral analyses as well as for the listeners' experiment acoustically "useful" utterances had to be very carefully selected.

Listening experiment: Design and performance

From each speaker and each variant three utterances of phrasal length were selected. The utterances were grammatically sentence-like, with different intonational patterns, and they contained a certain amount of pauses, hesitations and repairs. From the careful spontaneous and the read versions, identical utterances were collected. The utterances were presented in random order, and each utterance was played twice.

The listeners had to classify the utterances as "spontaneous" or "read" in a forced choice task. Furthermore, the listeners were asked to rate the degree of reliability for their decision on a five-point scale from very safe to very unsafe. Moreover, the essential linguistic or phonetic features underlying the listeners' decisions had to be specified. The given criteria were: syntactical structure, speech fluency, repairs, articulation, intonation, and speech rate.

The classifications were made by ten phonetically educated listeners who were members of the institute and twelve naive listeners (beginning students).

Spectral and segmental analyses

All dialogue utterances were orthographically transcribed, manually segmented and labelled and marked with phrasal accents. The spectral analyses that were necessary for the examination of segment durations and phoneme productions were carried out using a PC programme for speech labelling developed at the IKP (SONA) [5].

Vowel duration

The duration of the vowels [i:, ɪ] and [ɑ:, a] was measured, and the mean value was calculated for each speaker and each speaking style using SPSS for PC. First, two groups were built for each vowel: phrase-final and non phrase-final. Then, each group was subdivided into phrase-accented and non-accented vowels. Finally, the groups for non-accented vowels were divided into function words and content words. This grouping was made with regard to the prosodic variation that was observed for the different speaking styles.

Segmental reduction

Reduction of segments was measured by comparing the string of labelled speech with the canonical phoneme symbols that were obtained from the orthographic transcription using the PC programme P-TRA developed at the IKP [6]. Segmental variations were grouped into deletion of syllables (including contraction of words), deletion of sounds, substitutions and insertions. The number of misarticulations was calculated for each speaker and each speaking style (SPSS).

RESULTS

Listening experiment

Table 1: Number of "spontaneous" and "read" classifications per speaker and style (Sp. = Speaker; cas. = casual; car. = careful; r = read)

	Utterances classified as						
	spontaneous			read			
Sp.	cas.	car.	r.	cas.	car.	r.	Total
BP	37	52	30	06	14	35	174
SO	64	62	25	02	04	41	198
VB	58	41	37	06	25	29	194
VG	59	58	27	06	08	39	197
Total	218	213	119	20	51	144	763

As illustrated in Table 1, the casual utterances were correctly identified in almost all cases. Wrong decisions were in all cases reached by naive listeners. The appearing clear correct decision for this style may also be seen as a fact of the restricted acoustic conditions that had been present and that was certainly more easily perceivable by the educated listeners.

Having a look at the results for careful and read speech, correct decisions seem to be predominant. A closer look at the individual speakers concerning the careful style shows a high degree of misclassifications for speaker VB, whereas for SO only 4 wrong decisions - reached by naive listeners - occur. For VB, both in careful and in read speech, educated listeners come to even more wrong decisions than naive listeners. For all speech styles, the right decision is significant (p=.0011). However, significance is no longer maintained by the educated listeners (p=.0957).

As to the reliability of listeners' decisions, listeners are generally very or rather safe about reaching a correct decision. This result is significant (p<.01), but depends on the speakers and the spealing style. Looking at the read utterances by speaker VB, educated listeners are in most cases undecided, regardless whether their decision is right or wrong, whereas naive listeners are in most cases rather safe even when their decision is wrong and rather safe or undecided when they are right. The results are similar looking at the careful utterances of the same speaker.

The obvious supposition is that educated listeners are more careful in classifying perceived utterances as "read" or "spontaneous" as they are more used to carefully listening and more conscious of the variability in speech utterances than naive listeners.

For the correct decisions, the distribution of the phonetic or linguistic features was examined. Concerning the casual style, all features are mentioned more or less frequently by the listeners with only "speech fluency" being significant (p=.0003). "Fluency" is also significantly often mentioned in careful style (p<.01). Other criteria here are "repairs" (p=.0021), "intonation" (p=.0140) and "articulation" (p<.01 for the group of educated listeners). For the read speech style "fluency", "articulation", "intonation", and "syntactical structure" yielded significant results (p<.05).

Vowel duration

Results on vowel duration show a great variability depending on word category and phrasal accentuation and are of course very speaker dependent. For the individual speaker a clear trend concerning all vowel groups cannot be observed. This means that a single style is to be seen as a unique expression of speech, and vowel duration alone cannot generally indicate a specific speech style. The results show that casual speech is not obviously faster, and read speech, as it might be most carefully articulated, is not necessarily slower than careful spontaneous speech.

For [ɪ], results (cf. Table 2) support the classifications of the listening experiment, i.e. speaker SO slows down as she speaks more carefully whereas for VB mean durations become shorter. The results of BP and VG correspond to the listeners' decisions as duration in most cases does not change very strongly. Moreover, the large standard deviation in all cases indicates that the variation of speech rate differs to a great extent within a single phrase. Results are similar for the other vowel groups that had been analysed. They are not listed for reasons of space.

Table 2: Mean and standard deviation (ms) of vowel durations for non-final, non-accented [ɪ] in function words

	casual	careful	read
Sp.	Mean; std	Mean; std	Mean; std
BP	56,7; 23,1	60,1; 16,1	53,6; 16,3
SO	55,9; 24,6	56,7; 21,5	65,8; 19,4
VB	60,3; 21,9	48,5; 17,0	54,6; 16,8
VG	58,1; 24,1	55,7; 18,1	47,2; 12,5

Segmental reduction

For all speakers, the greatest difference in articulation compared to the canonical form is found in casual style. Here we find a large amount of syllable deletions, but also sound deletions (in most cases final stops) and substitutions. The only sound changes that are found less frequently than in careful and read style are insertions.

The difference between careful and read style is exemplified in Figure 1. For speaker SO an impressive decrease of deletion can be noted, for VB instead, an increase towards read speech. For the speakers BP and VG sound changes are not so heavy. These results explain the listeners' decisions in the listening experiment.

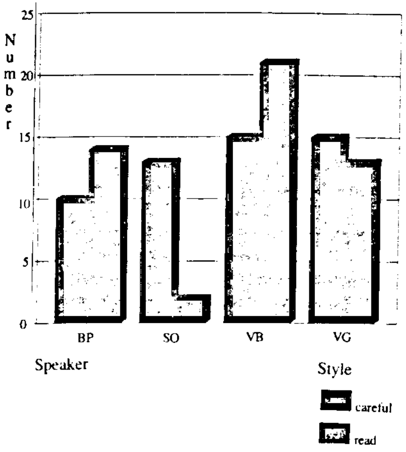

Figure 1: Sum of syllable deletions

CONCLUSION AND DISCUSSION

Listeners' decisions, especially for the speakers SO and VB, show that there exists a clear concept of what is expected both for read and for spontaneous speech. Obviously, read speech is expected to be very clear, i.e. without any misarticulations or repairs, and fluent, which means absence of pauses and hesitations. The appearance of hypercorrect articulation is related to read speech whereas a certain amount of segmental reduction is associated with spontaneous speech. If this concept cannot be recognized clearly, utterances cannot be unambiguously classified.

Of course, apart from the selection of utterances and the acoustical conditions, reading skills may influence the results, especially concerning the listening experiment. If reading skills are to be defined as the ability of perfectly transferring visual print patterns into acoustic patterns, then the utterances of speaker SO should be considered as a perfect reading. Listening to VB, however, utterances are very fluent and sound more natural than those of SO.

As a conclusion, these results suggest that the way listeners expect read speech to be does not match reality for all speakers. Moreover, the degree of acceptance of a speaker's reading style might be low. Read speech should rather be considered as having as much stylistic variation as spontaneous speech. Casual speech, on the other hand, seems to be a kind of slurred variation of spontaneous speech. This means that any expression of spontaneous or read speech is adapted to the given particular communicative situation, which is performed speaker-specifically.

Further examination of speech rate within a phrase, dynamic range, range of F0 and listeners' classifications of manipulated utterances, e.g. inserting pauses and hesitations into the signal of read utterances is called for. Furthermore, the classification of utterances related to a specific style may yield interesting results.

ACKNOWLEDGEMENT

This research study was partly supported by the German Federal Ministry of Education, Science, Research and Technology (BMBF).

REFERENCES

[1] Blaauw, E. (1992): Phonetic differences between read and spontaneous speech. In: Proc. ICSLP, Vol. 1, Banff, 751-754.

[2] Eskénazi, M. (1992): Changing speech styles: Strategies in read speech and casual and careful spontaneous speech. In: Proc. ICSLP, Vol. 1, Banff, 755 - 758.

[3] Eskénazi, M. (1993): Trends in Speaking Styles Research. In: Proc. ESCA Eurospeech, Vol.1, 501-508. ESCA ETRW, Barcelona, 36.

[4] Swerts, M./Collier, R. (1992): On the controlled elicitation of spontaneous speech. In: Speech Communication 11, 463-468.

[5] Stock, D. (1994): SONA. http://asl1.ikp.uni-bonn.de//sona.html.

[6] Stock, D. (1992): P-TRA - Eine Programmiersprache zur phonetischen Transkription. In: Beiträge zur angewandten und experimentellen Phonetik. Steiner Stuttgart.