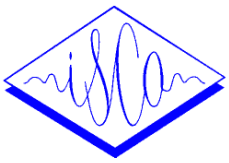


2nd International Conference on
Spoken Language Processing (ICSLP 92)
Banff, Alberta, Canada
October 12-16, 1992

CHANGING SPEECH STYLES: STRATEGIES IN READ SPEECH AND CASUAL AND CAREFUL SPONTANEOUS SPEECH

Maxine Eskénazi

LIMSI - CNRS
BP 133 91403 Orsay Cedex France

ABSTRACT

This paper examines segmental and suprasegmental elements which contribute to an impression of one speaking style as opposed to another. A corpus containing three styles of speech, casual, careful and read, for the same linguistic content was gathered. Examination of global and individual results reveals that: 1) spontaneous styles of speech cannot be considered to be linear modifications of read speech (careful speech is not necessarily faster than read speech, but slower than casual speech, for example), but probably closer to separate types of modifications of casual speech; and 2) the same perceived style is achieved by different speakers in different ways.

1. INTRODUCTION

One of the many sources of the apparent variability of the speech signal, is a speaker's conscious or unconscious change in the style in which he expresses himself [1]. Herein, we define style to be the expression of information about the dialect and socioeconomic background of the speaker, information about the manner in which he is expressing himself (formal, casual, reading, etc.), and information on the image he has of the speaker(s) he is addressing (slowing down for the hard of hearing, or foreigners, etc.). Style may overlap, but does not encompass, the range of a speaker's emotions and attitudes.

In this article, we examine the differences in suprasegmental and segmental elements of speech, such as intensity or devoicing, for different styles of speech.

As for spontaneous speech where many styles of speech may be enumerated, read speech reading a play aloud, a book to a child, and a book to a blind person [2] also involves differing styles. A comparison of one type of read speech with one style of spontaneous speech, except in the case of a specific application ([3], for telephone applications), may not give necessary perspective to the significant elements which contribute to the perception of a given style. We therefore propose to compare two styles of spontaneous speech, *casual conversation* and *careful repetition*, to one style of read speech, *reading a dialogue*, as in reading a play aloud.

The underlying motivation for this study is to try to better understand the variability of speech by looking for the extremes of the variability in given dimensions. A more pragmatic purpose is to eventually ameliorate the quality of synthetic speech with better knowledge of what constitutes careful speech. Results from this type of study could aid in making synthesis more understandable, on the one hand, and more convincingly natural (because it could take individual characteristics into account) on the other hand.

We used a corpus of speech where, in a fairly spontaneous situation, the same linguistic content was obtained both in casual and careful style, and then the transcription of the casual dialogue was reread by the speaker.

The following yardsticks were used to characterise each style: overall intensity, F0 maximum, dynamic range of F0, number of pauses, speaking rate, amount of phonological changes, F1/F2 shift, and the amount of stop bursts. Global values over all speakers have been compared, as have the values for individual speakers, in view of examining their separate, and differing strategies [4].

2. A CORPUS OF THREE SPEECH STYLES: SPOT

A corpus of spontaneous speech was designed, recorded, labelled, and verified [5].

2.1 Corpus design

The speech was elicited using a "wizard-of-Oz"/scenario technique designed to provoke changes in style during the course of the dialogue. Thirteen Parisian speakers (10 m, 3 f) ranging in age from 24 to 35 years old were given a scenario to be acted out over the telephone with the "wizard". All of the speakers, except GA, were given two different scenarios to act out. They were to play the role of a second-year student from a computer science school who wants to do a four-month research project at LIMSI. The names of the research subjects, the person's address, and items in his background all included phonetic contexts where phonological variants might be expected (for example, "les processus schématisés" where palatalisation could occur at word boundary). The wizard was to provoke a change in the speaker's style, from *casual* to *careful* speech, while maintaining the same phonetic context for the new style. This was done by asking, "Comment?" ("What?") twice during the conversation (if it did not become too evident), just after the speaker had used a sentence having one of the desired phonetic contexts. It was expected that the speaker would repeat basically the same message, rewording it, while keeping the same information-carrying words each time. From these five-minute dialogues, the utterances just before and after the "What?" were excised. All measures were taken only on the parts of the utterances before and after the "what?" which had approximately the same phonetic content, roughly the *rheme* of the sentence. Therefore, out of 25 five-minute dialogues, only 44 pairs of sentences were left. Obtaining the same linguistic material in two different styles of spontaneous speech is not easy; our paradigm, in hindsight, although amusing to design and record, is not efficient if we compare the useful amount of speech to total dialogue time.

2.2 Corpus recording

The speakers were asked to sit in an office and use a telephone which had been fitted with a microphone having larger bandpass characteristics (100-5000 Hz) than that of the telephone microphone. Two audio recordings were made: one of the speaker only, and another, of both sides of the conversation, through a tap on the wizard's telephone. The conversations were also recorded on videotape for later work on gestures. The speaker's signal was digitized at 10 kHz and stored on a PC-compatible. This is the signal used for the measurements below.

2.3 Labelling and verifying the corpus

The speech was phonemically labelled and orthographically transcribed. Spectrograms and an F0 analysis were obtained using the UNICE software (LIMSI-VECSYS).

10.21437/ICSLP.1992-231

In order to verify that the speech was also *perceived* as changing in style, and that the second style was judged to be *more careful* than the first, a jury of four listened to the pairs of utterances.

The jury consisted of the speaker, the wizard, the author, and a person not otherwise involved in this database. After listening to a pair, they were asked whether one utterance of the pair reflected an effort to "make oneself better understood" (literal translation of the question as it was asked in French). The majority decision determined which pairs were kept. Only 24 of the 44 original pairs (ten different speakers, 8 m, 2 f) were retained! In view of the results per speaker below, we could still question the inclusion of speaker RS's utterances, since the elements characterising his careful speech are typical of Lombard speech - he would have interpreted the request to repeat as meaning that there was noise on the line.

The speakers were also asked to read the orthographic transcription of their conversations. False starts and hesitations were removed to make reading more fluent. The text was presented in play form. The read style was set by the person who read the wizard's lines as if rehearsing a play. The speech with the same linguistic content as in the other two styles was excised and labelled.

In as far as the total amount of data is concerned, it should be noted that the quantity of data varies from one speaker to another, due to the jury decision. Also, one of the male speakers was no longer available at the time the read speech was recorded.

3. SUPRASEGMENTAL AND SEGMENTAL PARAMETERS STUDIED

The elements used to characterise the speech were: overall intensity, F0 maximum, dynamic range of F0, number of pauses, speaking rate, amount of phonological changes, F1/F2 shift, and the amount of stop bursts. In an earlier comparison of the two spontaneous styles [5], the amount of *empty* words, such as "euh", and the number of incidences of stuttering (as an indication of an effort to better articulate) were also examined. They were not used here due to the fact that they would probably reflect only reading skill in the third style.

Each measure was taken individually per utterance, then grouping all utterances of each speaker, and finally totalled over all speakers. It should be noted that, due to the nature of our paradigm, the material on which our measures are based is the high information content part of the sentence, roughly corresponding to the *rheme* of the sentence. For example, the sentence, "And you announced a project on connectionnist models." before the "what?", would give us " a project on connectionnist models" after the "what?". Interesting studies have pursued the difference between high- and low-information parts of sentences, by labelling the individual words as being of high and low information content [6]. It would be interesting to measure our data in this manner also, to see if style is expressed only in the individual high content words, or over all of the rheme of the sentence.

3.1 overall intensity

The mean intensity (in dB) of each utterance was calculated. In order to also have an idea of the perceived intensity, we asked seven subjects who had not taken part in the experiment so far to listen to the 24 pairs of utterances and to indicate whether the *careful* utterance was "louder" than the *casual* one. This result is expressed as the percentage of the jury who said that the careful version was louder.

3.2 F0 maximum and dynamic range of F0

Two measures using F0 were obtained. First, the mean F0 value of F0 maxima on stable vowels (determined by hand) was calculated. Then, the dynamic range of F0 for a given utterance was determined by subtracting the minimum value of F0 for the whole utterance from the maximum value (again on stable vowels). In order to normalise values over all speakers, it is expressed as a percentage of the F0 maximum value :

$$(F0max - F0min) / FOmax$$

3.3 number of pauses

The number of pauses, irrespective of their durations, was measured for each utterance. It was observed that, in careful speech, pauses often appeared just before and just after the information-containing words of the utterance. It is possible that, as for *empty* words and stuttering, the observations for read speech reflect reading skill, for some speakers.

3.4 speaking rate

Literature on the subject of speech called careful, formal, or clear, typically indicates that speakers slow down when they are trying to be better understood. Speaking rate is expressed as the mean number of phonemes per second.

3.5 phonological changes

Using the label character strings, phonological variations representing voicing, devoicing, schwa deletion, palatalisation, and nasalisation were totalled and expressed as a percentage of all possible contexts where variants could be present.

3.6 F1/F2 shift

The F1 and F2 values of all stable /a/, /i/, and /u/s were measured and the mean value taken for each speaker and for all the speakers together for each speech style.

3.7 presence of stop releases

In French, stop releases are either not present or very low in amplitude, often depending on their place in a sentence. As another means of exploring the effort to better articulate, the number of stop releases present was expressed as a percentage of the total number of stops pronounced (nasals were not counted here).

4. RESULTS

4.1 global results

The table below gives the global results.

MEAN FIGURES - ALL SPEAKERS

	casual:stdev	careful:stdev	read:stdev
Dynamics of FO in %	34.5 - 7.2	38.6 - 8.0	40.8 - 6.6
FO maximum in Hz	189	216	196
Intensity in dB	65.0 - 3.3	66.2 - 3.1	64.9 - 2.7
Sp. Rate (phms/sec)	13.2 - 1.5	12.7 - 1.6	13.6 - 0.9
% phonol. variants	29.5 - 9	25.3 - 11	17.8 - 7
% of all stop bursts	74.6 - 15	86.0 - 6	86.1 - 11
F1 mean value /u/	390	416	375
F1 mean value /a/	542	547	535
F1 mean value /i/	294	309	267
F2 mean value /u/	1028	885	1010
F2 mean value /a/	1570	1593	1593
F2 mean value /i/	2173	2192	2180

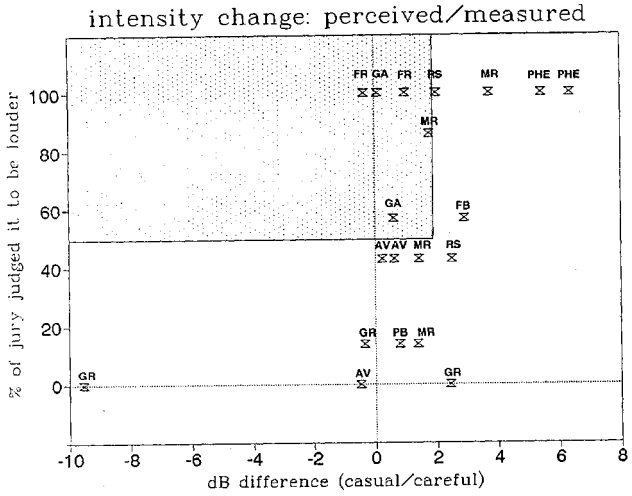
Two observations may be made. First, some studies of style change seem to indicate that one style may be modelled as a degree more (or less) of various segmental and suprasegmental elements than another. Careful style, for example, would simply be slower than casual speech, but faster than read speech. Our data do not agree with this. Read speech, for example, is much more "expressive", as measured by the dynamic range of F0, than careful speech is. This would argue for another manner of studying speech style. Instead of viewing read speech the base upon which modifications are made toward other styles of speech, it would be more logical (although more difficult) to use casual speech as the base, with each speech style being a unique type of modification. There is a parallel here with language acquisition: a child speaks in a casual style until age five or six, when he learns other styles, such as reading or speaking in a respectful way to teachers. Each of these styles is learned separately and may be viewed as a modification of his casual speech. The data in the following sections will be evaluated in this manner.

The second observation concerns the very large standard deviations. Observing only the totals of the data over all speakers implies that we presume that all speakers express style differences in the same manner. The standard deviation values here point to the fact that speakers are in fact using different strategies to achieve the same perceived result.

4.2 Individual results

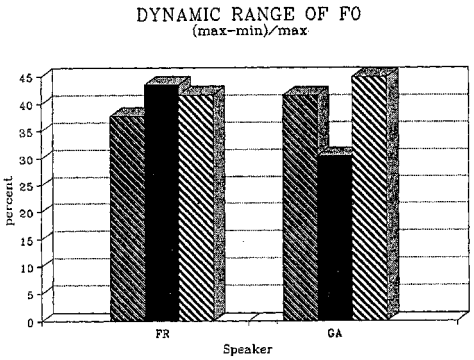
In order to examine the differences in the strategies of the individual speakers, let us look at the results for speakers FR and GA.

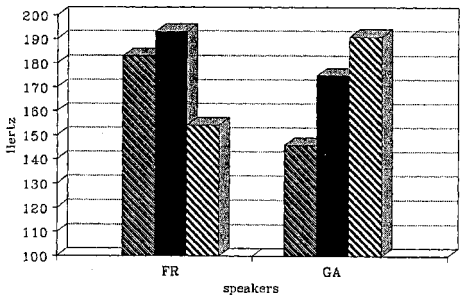
First let us look at these speakers' behaviours in the use of intensity as a means of expressing style change. Literature on the subject of formal (careful) speech typically indicates that speakers increase intensity when they are trying to be better understood. The graph below plots the actual intensity measurements compared to the percentage of the jury that perceived the *careful* speech to be louder than *casual* speech.

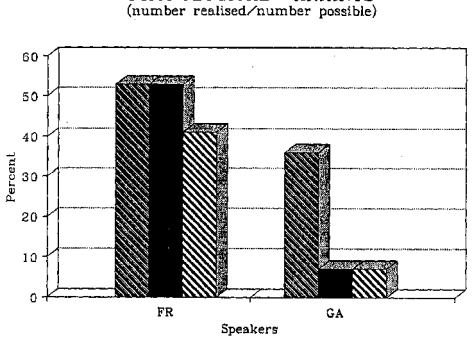


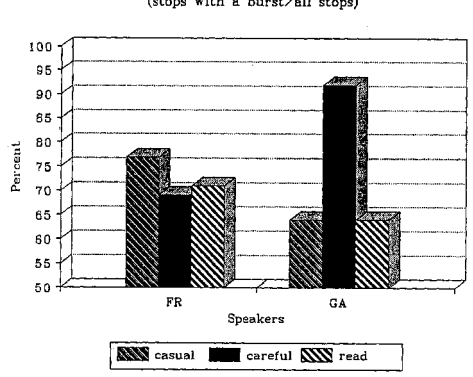
FR and GA, although perceived to be speaking louder by over 50% of the jury, did not actually produce louder speech. Other elements need to be investigated to determine what gave the jury the impression that they were speaking louder and more carefully. If we look at our measures of F0, we have the results at the right.

FR's careful speech has a higher F0, and has a slightly larger pitch range. GA also increases F0 in careful, as opposed to casual speech, but decreases the pitch range. The increase in F0 may account for the impression of louder speech. Further examination, into the segmental elements which may reflect an effort to articulate more precisely, reveals the following results for phonological variants and stop releases:

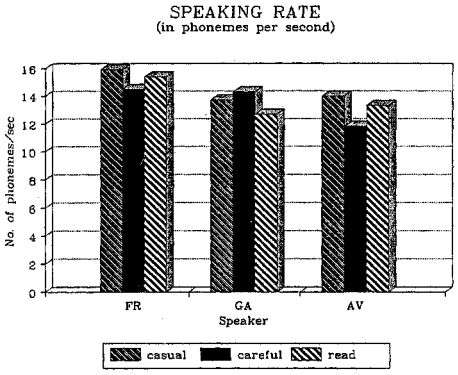








We see a clear effort to articulate better on the part of GA, for careful speech as opposed to casual speech, which is not the case for FR. If we look at the results for read speech, we see that FR and GA have a decreased amount of phonological variants as compared to casual speech, but do not make a particular effort concerning stop releases. GA seems to make an overall effort to articulate better for careful speech, whereas more precise articulation is characteristic of read speech for FR. The results for speaking rate and change in F1/F2 show no statistically significant variation from one style to another for these two speakers. We include the results for speaking rate here to show the difference between the results for FR and GA and those of AV.



Due to lack of space, all of the results for all speakers cannot be shown. Some other significant strategies:

* For AV: She increased maximum F0 and F0 dynamic and spoke more slowly for careful speech, as opposed to casual speech. For read speech, she decreased intensity and the amount of phonological variants.

* For PHE: Read speech was characterised by an increase in F0 dynamic, and a decrease in intensity and the number of phonological variants.

5. INTERPRETATION OF THE RESULTS

The results show that individual speakers do express speech styles in different ways. Although certain elements found in the literature, such as slowing down and speaking louder for a careful style, are used by some speakers, it is not the case for everyone. Other elements, such as an increase in the dynamic range of F0, are used by other speakers who are also perceived to be making an effort to be better understood. One of the reasons that careful speech is not always characterised by an increase in intensity may be (in the case of AV for example, who is an assistant professor) that teachers tend to lower their voices rather than raise them to get students to listen more closely.

Some of the results for read speech may be artifacts related to reading skills. An assessment of the reading skills of each speaker should help clarify this point.

Certain style changes are marked by statistically significant data, but this is not the case for all speakers and all styles. It is probable that the expression of, for example, careful speech, was not as strong for all speakers. The familiarity of certain speakers with the wizard and the desire to play the role as well as possible may be the causes here. It is certainly possible to imagine different degrees of a given style, such as *careful speech with a complete stranger, careful speech with a good friend, careful speech with a small child.* In this case it might be possible to place the change on one same axis, the speech slowing down when speaking to a small child rather than a good friend.

A model of a real voice rather than a composite one for high quality synthetic speech must take these individual strategies into account if it is to be convincing. Speech recognition may also benefit from better understanding of the strategies used here, being able to predict a number of variations for a given speaker from a small sample of his speech.

6. CONCLUSIONS

We have gathered and analysed casual, careful, and read speech from several speakers. Our results on this data show that each style is characterised by different elements. Classically, studies have considered read speech as a conservative starting point on a linear axis where casual speech would be at the other end. Casual speech seems to us to be a better base from which to find those elements than read speech does.

We have also shown that speakers use different strategies to achieve the same perceived result.

As mentioned above, this type of data is difficult to obtain. We are presently recording another database for style comparison that includes a large number of speakers. The paradigm used furnishes more usable speech per recording, and will be used to confirm our findings.

We also intend to soon test our findings by manipulating synthetic speech and evaluating the perception of the results.

Many more studies, involving other style changes, relations between elements of different nature (phonological and prosodic, for example), and languages other than French also need to be carried out in order to aid our comprehension of the limits of individual variability.

References

[1] Shockey, L. 1983, Phonetic and phonological properties of connected speech, Ohio State Working Papers in Linguistics.

[2] Granström, B., 1991, The use of speech synthesis in exploring different speaking styles, ESCA ETRW on the Phonetics and Phonology of Speaking Styles, Barcelona, Catelonia, September, 1991.

[3] Silverman, K., Blaauw, E., Spitz, J., Pitrelli, J., 1992, Towards using prosody in speech recognition/understanding systems: differences between read and spontaneous speech, Fifth DARPA Workshop on Speech and Natural Language, Harriman, N.Y., February, 1992.

[4] Eskénazi, M., Lacheret-Dujour, A., 1991, Exploration of individual strategies in continuous speech, Speech Communication, vol. 10 P. 249-264.

[5] Eskénazi, M., Isard, A., 1991, Characterising the change from casual to careful style in spontaneous speech, Journal of the Acoustical Society of America, Vol. 90, Houston.

[6] Koopmans-van Beinum, Florien, 1992, The role of focus words in natural and in synthetic continuous speech: acoustic aspects, to appear in: Speech Communication, vol. 11, 1992.