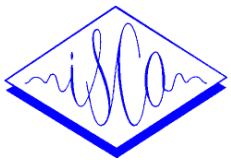

3rd European Conference on
Speech Communication and Technology
EUROSPEECH'93
Berlin, Germany, September 19-23, 1993

THE DESIGN AND RECORDING OF ICY, A CORPUS FOR THE STUDY OF INTRASPEAKER VARIABILITY AND THE CHARACTERISATION OF SPEAKING STYLES

Vincent Péan[1], Sheila Williams[2] and Maxine Eskénazi[1]

[1]*LIMSI-CNRS, BP 133, 91403 ORSAY CEDEX, FRANCE*

[2]*Depts. Psychology, Computer Science, University of Sheffield, Sheffield S10 2UR, GB*

ABSTRACT

This paper describes a corpus for the study of intraspeaker variability and of contrastive speaking styles.
Design accounted for: different speaking styles being compared for the same linguistic content for a given speaker; the speaking styles being clearly produced and perceived to be different without direct prompting or reading; a series of specific predetermined contexts of phonological variation being elicited.
In the "seven errors" task, the speaker describes objects which differ in two drawings and are places where possible phonological variation may occur.
Styles include casual, clear and clear read. The signal from one speaker occupies about 60 MBytes (audio) in all, about 30 minutes of speech, representing an average of about 70 "phonological contexts" per speaker per style, covering: voicing, devoicing, schwa elimination, palatalisation, nasalisation, and geminates. Twenty one speakers of varying origins have already been recorded. The data is being labelled, using semi-automatic labelling techniques and will be perceptually verified to confirm style change.
This methodology has also been used to collect a corresponding corpus of British English.

***Keywords**: speaking styles, variability, speaker characterisation, database*

1. INTRODUCTION

This paper deals with the development of a corpus designed to allow us to study the linguistic behavior of individual speakers for different speaking styles. Various levels may be studied individually and together, among them phonological and acoustico-phonetic variants. The database includes speakers of varying backgrounds (and ages, sexes, etc.) so that: a) we can compare different speaking styles for a given speaker (or a group of speakers); b) the speaking styles studied can clearly be produced and perceived; c) the speakers produce the same linguistic content, that is, predefined contexts where phonological variants may occur, without reading them. Three speaking styles are elicited here: two "spontaneous" non-read and one read.

When collecting examples of various speaking styles, there are several classical problems. They are due to the fact that although it would be best to collect totally free casual speech, this type of speech, by its very nature, is very difficult to use for comparative studies. The pragmatic context may continuously vary, and thus, the speaking style as well (from a very casual beginning with a friend, to a more formal part where there is important travel information for his upcoming trip, for example). Another problem is that if specific linguistic material is to be studied, there will not be enough in a non-directed conversation, and the material will most probably not be found for all speakers and/or several times for the same speaker. A corpus that compares given elements of linguistic content over several speaking styles must therefore be based on elicited speech. It is then important to keep the speech as natural as possible and to elicit speaking styles which are comparable. A comparison of the interrogation of a flight information service with speech read from a computer screen has little sense. The task content must be comparable.
The description below concerns the French database /1/, and the ongoing work on the British English version. Another recording has been started at the University of Provence using the ACCOR recording setup.

2. DATABASE DESIGN: PREDEFINED CONTEXTS

If the phonological behaviour of a given speaker is to be described, there must be a judicious choice of the phonological contexts where variability may occur, and that are to be produced. The sources of phonological variability in French are numerous /2/,/3/,/4/,/5/; a restricted choice of what is to be studied is necessary. We have selected the most frequent variants, presuming that they will also be the ones our speakers will be most likely to pronounce. They are: schwa elision, devoicing, voicing, palatalisation, nasalisation and gemination. In French, assimilation is most frequent a) in regressive form (the second phoneme influences the preceding one, and not vice versa as in English); and b) between words rather than within words /2/,/6/. The contexts were therefore created as groups of words with the variation between them (for example, "chaussure rose").

3. DATABASE DESIGN: THE TASK

With lists of predefined variants in hand, it was then necessary to find a task in which the speaker could be expected to pronounce them without reading them and do so several times, in a different style each time and as naturally as possible. It was decided, from past experience /9/ not to have the subject learn a scenario. The styles obtained needed to have a common context (goal) and to be clearly perceived as different from one another.

In order to preserve a "spontaneous" quality to the speech, it was decided to have the speaker describe objects in a "game of seven errors". The task consists of describing the difference between two drawings which at first look identical, but which, upon closer examination, contain objects which are slightly different. The speaker has to find all of these objects and to describe them. For our purposes, the objects that differed in the drawings were the phonological contexts where variation could occur (for example, "chaussure rose" - "pink shoe", geminate /R/ - on the left and "chaussure rouge" - "red shoe", geminate /R/, too - on the right). The left drawing was considered to be the "reference drawing" from which the description was to start each time. The speaker was also instructed to give the location of the object within the drawing and describe it fully, including its colour, for example.

The speaker's attention was therefore very absorbed in the task. The choice of the precise phonological contexts was complicated by two factors: the fact that the two words had to be pronounced together in a given order ("chaussure rose", not "chaussure" or "chaussure qui est rose") and that the image they were describing had to give a non-ambiguous description (red that could not be mistaken for orange, or a shoe that did not look like a slipper). The proposed drawings were given to a set of four speakers for verification of the objects, and the task instructions. They were then modified and retested before starting the final recordings.

The phonological contexts were also chosen so that they would not all fall in the same syntax structure /7/ (such as noun followed by postposed adjective, as above), and so that they fell within and between phrase boundaries. The structures selected were: noun + adjective; adjective + noun ("longue cravate"); noun + adjective + preposition ("fleur rouge sur"); adjective + noun in a date ("le douze septembre"); noun + conjunction + noun in time ("midi et demie"). The speaker was encouraged to give as full a description as possible. Many "free contexts", contexts of phonological variability which had not been predefined, were also produced, thus increasing the total number of contexts pronounced. Schwa elision contexts were so numerous that they were all taken from free context.

The objects were presented in four different pairs of drawings.

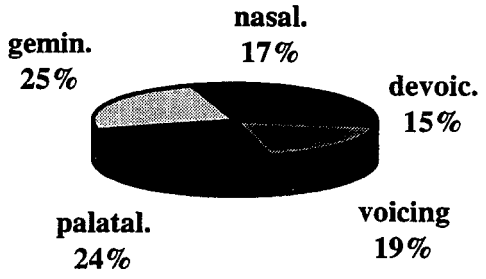

Figure 1. Predefined phonological contexts (any style). In detail: 19 voicing, 9 devoicing, 18 palatalisation, 2 palatalisation + voicing, 7 palatalisation + devoicing, 19 nasalisation, 27 geminate.

Figure 1 shows that approximately equal amounts of the different types of contexts were predefined.

4. SPEAKING STYLES

Once the task and contexts were set up, trials were carried out to perfect the instructions that would make a speaker convincingly produce the selected speaking styles. The three styles were selected to have similar task context - even "reading" is not a style in itself due to the very different nature of the speech produced when, for example, reading to a child, or reading a newspaper aloud; reading when one has the habit of reading aloud or reading when one doesn't and finds it difficult. The situation chosen therefore concerned a purported contract for video recordings to help hard-of-hearing children learn lipreading /8/. The speaker went through this task three times. First there was a "rehearsal" when the video equipment was (the speaker was told) turned off. This furnished the casual style of speech. Then the "real thing" was recorded (the speaker saw gestures of turning the camera on this time). This gave the clear style of speech. The speech was then transcribed and, later in the same day, the speaker read the transcription as another recording that the children could use to learn lipreading. The read speech thus addressed the same task and the same audience as the non-read speech.

5. RECORDING

Recording was carried out in a quiet room. The speaker wore a head-held SHURE SM10 microphone connected to a VECSYS TDS AD/DA board. He was also recorded on a Sony HI-8 video camera. The speech signal was stored at 16KHz on both a DAT and a WORM. During recording, each file contained one whole drawing description for one file. There were therefore four files per style x three styles = twelve files per speaker.

6. FORMATTING THE DATA

The mean disk space occupied by one speaker (for all styles and drawings) is 60 MBytes (varying from 50 to 80 for one very talkative speaker). There are 21 speakers at present and recording continues. There are 11 male and 10 female speakers ranging in age from 19 to 54. They are not all Parisians. Future recordings will enlarge the variety of speakers, especially concerning their linguistic and cultural backgrounds.

7. DATA STRUCTURE

The data was structured into a directory tree structure compatible with the storage media and the type of data we wish to compare most often. It is shown in Figure 2. The description of a whole drawing was cut up into smaller files corresponding to the complete description of one object in the left and right drawings (for example, the comparison of the colour of the shoes on the left and on the right). Dividing the file here was fairly easy as there was always a pause between descriptions of object pairs. This gives an average of 20 speech files per drawing. Each signal file has five associated files: the orthographic transcription, the phonetic transcription, the ideal phonetic transcription in syllable form, the label file (by syllables - see below), and the discourse marker file.

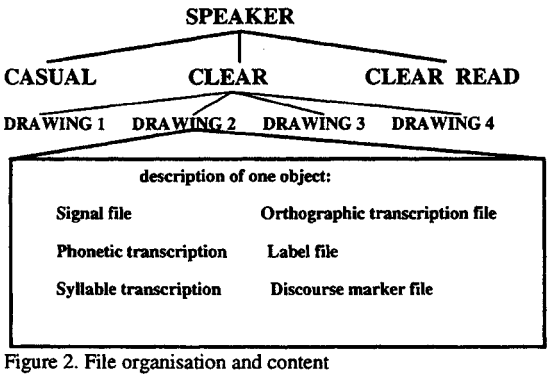

Figure 2. File organisation and content

The video tapes were not synchronised with the audio. Discourse markers are being put into the label file (at a separate level from the phonemes) and into a separate file. Pauses are also carefully transcribed.

8. LABELLING

Labelling is semi-automatic. The orthographic transcriptions are first passed through a grapheme-to-phoneme translation module, GRAPHON, which has been slightly modified to take the spontaneous speech into account. The "ideal" phonetic transcription is then saved, and a copy is automatically divided into syllables (the SYLLA module, developed at LIMSI and again modified for the spontaneous speech). The resulting ASCII file is given to a labeller who, looking at the spectrogram of the signal and listening to it: places all syllable boundaries, using the UNICE software, and "corrects" the "ideal" transcription according to what has really been said. The corrected syllable file is automatically "stuffed" into the hand-marked boundaries and the resulting label file is verified by a labeller using UNICE again.

9. PHONOLOGICAL CONTEXTS

Based on the first 21 speakers, we can see what percentage of the prepared phonological contexts have really been pronounced and how many "free" contexts were also pronounced.

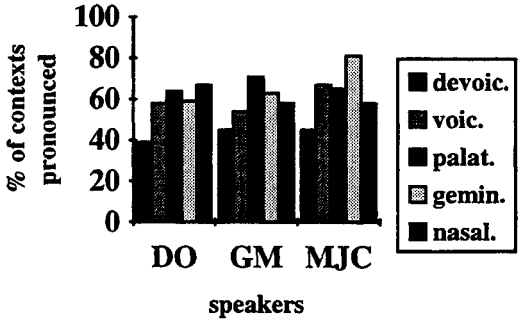

Figure 3. Percentage of predefined phonological contexts actually pronounced (casual style) by three speakers, DO, GM, and MJC.

Figure 3 shows the percent of predefined contexts pronounced by three speakers chosen randomly. In general, from 40 to 81% of the contexts appear. Devoicing tends to be less present than the other variants.

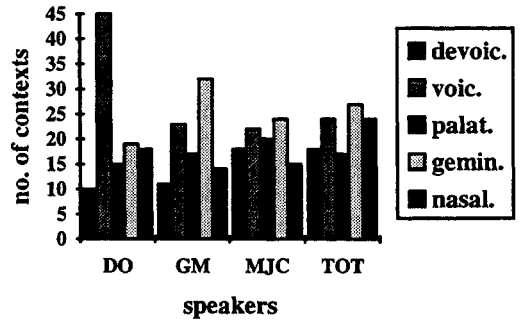

Figure 4. Number of free contexts plus number of predefined contexts by speaker, by the same speakers as in Figure 3, for casual style, compared to the total number of predefined contexts (TOT).

If we now add the number of free contexts to the number of predefined contexts, we see that voicing is well-represented for all speakers. Devoicing remains slightly under expectations, for all speakers. The other variants are fairly well represented. Speaker DO, the least talkative of the three, has a slightly lower number of contexts.
In as far as our dependance on free context to give us all of the schwa deletion examples, we note that DO has 21 contexts, GM 39 contexts, and MJC 17 contexts.

10. ICY IN BRITISH ENGLISH

This methodology has been adapted for the collection of a corresponding corpus in British English. The recording protocol was modified slightly due to the fact that the phonological contexts occur much more often within words in English than in French, therefore the instructions on how to describe the images, used to elicit strings of two words for between-word contexts are relaxed here.
In British English casual speech almost every segment can be subject to assimilation, deletion or reduction in comparison with clear speech. For example, using a pair of drawings depicting beach scenes with five major differences between them, for ten phrases describing the whole drawing, we find a total of 42 predefined and free contexts. These represent 29 distinct contexts in which contrastive variation might be expected. Preliminary investigation with this picture set indicates that on average, at least twenty of these target sequences are reported under each of the two data collection conditions tried, although the current elicitation techniques have not produced contrastive responses in every case. The remaining targets were missed due to the flexibility of the English language which imposes few restrictions on phrase ordering. Our pictures appear to be quite successful at eliciting the predefined object labels and attribute predictors, but choice of phrasing ("a red striped shirt" or "a shirt with red stripes") and sentence structuring varies from speaker to speaker. However, alternative contrastive contexts to those predefined frequently arise from the alternative structure, and so are equally useful for our analyses. Additionally, for the preliminary recordings which have been completed so far, contrasts between the two recording conditions (casual and clear) are limited compared with those between task-completion and task-descriptive speech (really contrastive casual speech is found in words like "picture" and "bottom" in both conditions. We should note, however, that these preliminary recordings were made without the use of video.

11. CONCLUSION

This article has presented the ICY database which uses elicitation to obtain speech that has the same linguistic content for several comparable speaking styles. The first use of this database will be to group speakers having common phonological behaviour when changing styles and to test findings using speech synthesis.
Extensions have already been planned. The speakers will have an interview in which important information concerning their background and their impression of what they were to do during the recording sessions will be elicited. Unlike the usual questionnaires, information will differ according to the main influences in the background of each individual. The coincidence of discourse markers and phonological variants is under study, as is the determination of the linguistic elements which are totally task-dependant. In order to verify the quality of our database, we are also planning to carry out perception tests to see if the style intended by the speaker was indeed perceived by listeners. We are also examining the strategies used by speakers when describing the image - where the description starts, how much in advance the speaker prepares what he is about to say.

This work has been conducted in conjunction with the EEC ESPRIT VOX Project no. 6298.
We wish to thank Alain Marchal, Sandra Whiteside, Sophie Faye, and Guy Chastagner for their participation in various parts of the creation of ICY.

REFERENCES

/1/ *Péan V.*, "Conception de la base de données ICY", Notes et documents LIMSI N°92-10, 1992.

/2/ *Eskénazi M., Vaissière J., Lonchamp, F.*,"Cours sur les indices acoustiques du français".LIMSI,Orsay & CNET, Lannion, 1987.

/3/ *Mariani J.* Introduction au traitement automatique de la parole. Graduate Course Book, LIMSI. 1984.

/4/ *Walter H.* La phonologie du français. P.U.F. Le linguiste., 19 77

/5/ *Liénard J.S.*, Les processus de la communication parlée. Introduction à l'analyse et à la synthèse de la parole. Editions Masson. 1977.

/6/ *Fouché P.* Traité de prononciation française. Ed. Klincksieck. Paris. 1968.

/7/ *Price P.J., Ostendorf M., Shattuck-Hufnagel S., and Fong C.*. "The use of prosody in syntactic disambiguation". J. Acoust. Soc. Am., vol. 90, N°. 6, December 1991.

/8/ *Picheny M.A., Durlach N., Braida L.D.* "Speaking clearly for the hard hearing I: Intelligibility differences between clear and conversational speech". J. of Speech and Hearing Research, vol. 28, p. 96-103, March 1985.

/9/ *Eskénazi M.* "Changing speech styles: strategies in read speech and casual and careful spontaneous speech". Proc. ICSLP Conference, Banff, p. 755-758, October 1992.