Phonological rules modelling style variations in French Parisian spontanous speech for text-to-speech synthesis

Péan Vincent

LIMSI-CNRS BP133, F 91403 Orsay-Cedex, France

ABSTRACT

Keywords : phonological variability, text-to-phonemes conversion, speaking styles, spontanous speech, speech synthesis.

The study presented in this paper has been carried out in the framework of phonological variability in French, with applications to automatic speech processing in mind. The specific aim of this study is to both characterize and model intra-speaker segmental variants (at and within word boundaries) in two speaking styles. Data have been collected from casual and careful speech corpus. Examples of phonological rules are given here.

INTRODUCTION

When a speaker tries to change his way of speaking from casual to careful speech, intra-speaker segmental variants can be observed which could be modelized by phonological rules. This implies the collect and the study of both casual and careful speech data: the ICY database (for the study of inter and intra-speaker variability and the characterisation of speaking style) is first described.

The methodology used to analyse data is described. A segmental analysis of the data is then provided in terms of statistical quantification. A comparative and qualitative study of segmental strategies both between two styles for a given speaker and between different speakers for one given style is also presented. Finally, we present examples of phonological rules, with consequences of their modification on synthesized speech.

PRESENTATION OF THE DATABASE

Corpus, task and speakers

Style concept implies choice between several possibilities. Thus, in a given setting and for a given speaker, a modification of the speaker's intention could lead to style variation.

The ICY database [1] has been developped to study inter- and intra-speaker variability which occurs when a speaker try to speak more carefully.

ICY has been recorded to collect three different styles of speech: two spontaneous and one read. Spontaneous speech is considered as non read speech. Here a remark may be noted: the structure of speech of a speaker vary with the context of discourse (setting) and with his psychological state. Thus a lot of different spontaneous speech exist, and to collect speech in a laboratory in a specific context for a specific goal gives one of them. With a view to collecting the data (and to generate a modification of the speaker's performance corresponding only with style variation) a methodology has been developped: the speaker's task is a description of two drawings which differ in some of their parts. Each speaker has to describe each object which differed from one drawing to another, with its colors and spatial positions. A lot of phonological contexts (i.e. where phonological variation could occur) are obtained by constraining the speaker to pronounce them in his description: each object which differ in the two drawings may be, with the constraints imposed during the task, described with groups of words which contain phonological context at word boundaries (for example: robe bleue, context of gemination /bb/). The phonological contexts choosen are: gemination, palatalisation, nasalisation, voicing, devoicing, and schwa.

With a view to obtaining the three different styles, a goal is given to the speaker for all of recordings: this consist of making recordings to help hard of hearing children to learn lip-reading. The speaker goes throught the task three times. The casual speech is collected first when the speaker describes the four drawings just to "rehearse". The careful speech is obtained when the speaker does the "real" recording in front of a camera.
The results presented concerned only four speakers: three female (RF, GS, GM) and one male (BP).

THE SEGMENTAL STATISTICAL ANALYSIS OF VARIANTS

The variation studied

The study is about the phonological variation which occurs in two speech styles (casual and careful) between different speakers. The phonological variation considered corresponds to the variation which leads to a complete modification (i.e. insertion, deletion or substitution) of one or more segmental units which constitute a phonological system of reference of Parisian's speech: GRAPHON[2].

The use of a reference is imposed because to make a straight out comparison between two speakers, the linguistic content of their two recordings must be the same, but that is not the case here because the speech is spontaneous. Thus, by first using a comparison with a reference the results obtained on each of the two speakers could be used to compare them.

The methodology

The analysis method is to do ortographic transcription of the recordings for each speaker, and to use it to obtain by GRAPHON and a specific automatic treatment a homogeneous translation grapheme to phoneme with pauses, word boundaries, and syllable boundaries within word, which will be called the "ideal" phonemic transcription for a given speaker. Then a correction of the ideal phonemic string is done according with what the speaker has really pronounced, by listening and using acoustical representation. Then, the corrected phonemic transcription, which will be called the "real" phonemic string, is compared automatically to the ideal transcription. In this way a characterization of the phonological variation as compared to the reference GRAPHON is obtained for each speaker. For example: the speaker says "il y a un nuage jaune sur le dessin de droite"; by GRAPHON and others semi-automatic treatments the ideal phonemic string obtained is:

#IL#I#A#<#N^AJ$#JON$#SYR#LE#D(~S<#DE#DRWAT$#;

the corrections give the real phonemic string:

#Y#A#<#N^AJ$#:ON#SYR#L#D(~S<#D#DRWAT$#.

The string comparison leads to specific information files[3]. The sharp sign '#' mark the word boundaries; the tilda '~' the intraword syllable boundaries; and the '$' the graphemes 'e' corresponding with linguistic E caducs.
Then the resulting information files are semi-automatically analysed as follow.

The data analysis

Different kind of events are obtained from the automatic analysis: insertion, deletion or substitution of one or more segmental units. To each event may correspond a specific phonological event. For example the deletion of 'E' in the monosyllabic word 'DE' is in fact a schwa deletion.

The results present here concern the deletion of schwa (i.e. linguistic E caduc); and the substitution of one consonant by another consonant correspondind to voicing; devoicing; and palatalisation. For example 'grande table', translated in #GR*D$#TABL$#, may lead to #GR*T$#TABL$#; thus we obtain the substitution of 'D' by 'T' which is in fact a devoicing event in a regressive form (the second phoneme influences the preceding one); and between word.

Some results about the schwa have been given [3] but here the name of schwa is given to linguistic E caduc, which corresponds to a graphemic 'e' in the And in the semi-automatic treatment a distinction have been done between different realisations of E.

For example in the utterance **'Euh** ... ce film(e) tchèqu(e)', four timbres of E caduc may be distinguished. **Euh** is an hesitation vowel; (c)e is a 'true' E caduc (i.e. a linguistic E caduc); (film)e is a non-linguistic E caduc; and (tchèqu)e is a E caduc links to the pronunciation of a final consonant before a pause. [4] [5].

The results given here concern only the 'true' E caduc, but a more complete study will be done on different kinds of E caduc. [6].
The devoicing analysis is about the substitution of one of the consonant {/B/, /D/, /G/, /V/, /Z/, /J/} by the unvoiced corresponding one from the set {/P/, /T/, /K/, /F/, /S/, /X/}. This event is considered in all devoicing context between word (i.e. of type: voiced consonant#unvoiced consonant).
The voicing analysis concerns the substitution of one of the unvoiced consonant {/P/, /T/, /K/, /F/, /S/, /X/} by the voiced corresponding one from the set {/B/, /D/, /G/, /V/, /Z/, /J/}. This event is studied in all voicing context between word (i.e. of type: unvoiced consonant#voiced consonant).

The palatalisation analysis concerns the substitution of one dentale fricative consonant /S/ or /Z/ by the palatale fricative consonant /X/ or /J/. This event is observed in all 'palatalisation context' between word (i.e. S#X; S$#X; etc.).

For these three analysis, two different kind of set of contexts have been considered: a) of type consonant$#consonant (where '$' is a potential schwa); and b) of type consonant#consonant.
Thus for one given analysis and one given type of context, six events are considered.

For voicing regressive inter-word context c1$#c2 (e.g. xxxF$#Dxxx): 1)voicing (i.e. c1 is substitued by c1' which is voiced; e.g. F substitued by V); 2) progressive devoicing (i.e c2 is substitued by c2' which is unvoiced; e.g. D substitued by T); 3)[E]-E caduc 'insertion' (i.e. $ is substitued by E; e.g. F$#D becomes FE#D); 4)nothing (i.e. c1$#c2 is not modified); 5)[E:]- hesitation insertion (i.e. $ is substitued by E: which represents an hesitation realized on linguistic E caduc; e.g. F$#D becomes FE:#D); 6)[p]- empty pause insertion (i.e. there is a pause insertion in the context c1$#c2); but this last event may correspond in fact to several sub-events as 6)a)only pause insertion (i.e. c1$#c2 becomes c1$#p#c2; e.g. F$#D becomes F$#p#D); 6)b)[&]- E caduc 'pre-pausal' insertion (cf. above the E caduc of the word 'tchèqu(e)'); (e.g. F$#D becomes F&#p#D); and 6)c)[E4]- hesitation pre-pausal

insertion (e.g. F$#D becomes FE4#p#D). The distiction between the sub-events have not been done: only pause insertion is considered in the three cases.

The same considerations are done for devoicing regressive inter-word context c1$#c2 (e.g. D$#F) except for the points 1) and 2): 1)devoicing (i.e. is substitued by c1' which is unvoiced; e.g. D is substitued by T); 2) progressive voicing (i.e. c2 is substitued by c2' which is voiced; e.g. F is substitued by V).

The two first points change also for the palatalisation analysis with progressive or regressive inter-word context c1$#c2 (e.g. X$#S or Z$#J): 1)progressive or regressive palatalisation (i.e. c1 (or c2) is substitued by c1' (or c2') which is palatale; e.g. X$#S (or Z$#J) becomes X$#X (or J$#J)).

In the same way, for voicing regressive inter-word context c1#c2 (e.g. xxxF#Dxxx): 1)voicing (i.e. c1 is substitued by c1' which is voiced; e.g. F substitued by V); 2) progressive devoicing (i.e c2 is substitued by c2' which is unvoiced; e.g. D substitued by T); 3)[E3]- non-linguistic E caduc (cf above the E of the word 'film(e)') 'insertion' (e.g. F#D becomes FE3#D); 4)nothing (i.e. c1#c2 is not modified); 5)[E2]- non-linguistic hesitation insertion (E2 represents an hesitation realized on non-linguistic E caduc; e.g. F#D becomes FE2#D); 6)[p]- empty pause insertion (i.e. there is a pause insertion in the context c1#c2); but this last event may correspond in fact to several sub-events as 6)a) only pause insertion (i.e. c1#c2 becomes c1#p#c2; e.g. F#D becomes F#p#D); 6)b)[&2]- non-linguistic E caduc 'pre-pausal' insertion (e.g. F#D becomes F&2#p#D); and 6)c)[E5]- non-linguistic hesitation pre-pausal insertion (e.g. F#D becomes FE5#p#D). The distiction between the sub-events have not been done: only pause insertion is considered in the three cases. And again, the same considerations are done for devoicing regressive inter-word context c1#c2 (e.g. D#F) except for the points 1) and 2): 1)devoicing (i.e. is substitued by c1' which is unvoiced; e.g. D is substitued by T); 2) progressive voicing (i.e. c2 is substitued by c2' which is voiced; e.g. F is substitued by V).

The two first points change also for the palatalisation analysis with progressive or regressive inter-word context c1#c2 (e.g. X#S or Z#J): 1)progressive or regressive palatalisation (i.e. c1 (or c2) is substitued by c1' (or c2') which is palatale; e.g. X#S (or Z#J) becomes X#X (or J#J)).

Results

The first results presented concerned the schwa (i.e the linguistic E caduc defined before) and four speakers: three females (GS, GM, RF) and one male (BP). To illustrate the great variability that occurs for the schwa between different in a given style, the percentage of deletion of E caduc obtained by comparison between the real phonemic string of each speaker for each picture and the corresponding ideal phonemic string obtained by GRAPHON have been plotted. (see figures 1 and 2).

figure 1

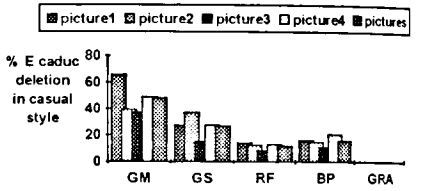
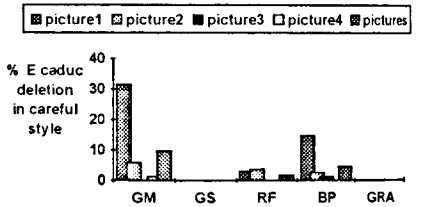


figure 2



The comparison above between two styles for a given speaker shows that a fixed description (given by the rules of GRAPHON here) is not enough to describe speech communication. The GRAPHON rules on the schwa seem to be more appropriate to describe the careful style.

It seems again that the casual style for each speaker is marked by a more important percentage of E caduc deletion than careful style. But the percentage between speakers are very different, thus the phonological rules that will govern this event will be variable.

The second results concerned the voicing defined before. For the contexts c1$#c2 in the two styles (see figures 3 and 4).

figure 3

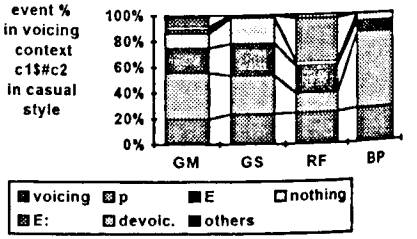
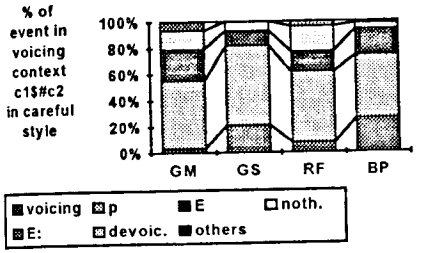


figure 4



Then figures 5 and 6 represent the c1#c2 contexts.

figure 5

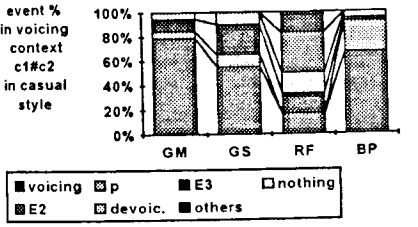
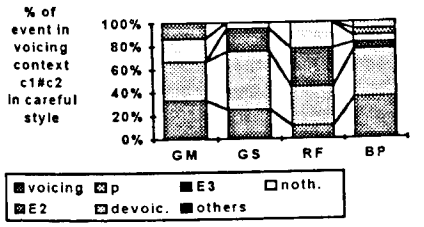


figure 6



The same variability between speakers and styles have been obtained for devoicing and palatalisation contexts.

Phonological rules

The results show in all cases that for a given the phonological behaviour is very different between the two styles studied. Moreover, in a given style, different strategies can be observed.

The role played by E caduc and pauses seem very important to distinguish the two styles. Moreover presence or absence of potential E caduc ('$') implies different strategies for a given speaker in a given style (see figure 3/ figure 4).

These results led us to test on the KTH synthesizer some phonological rules using pause and hesitation insertions and schwa deletions to characterize both a given speaker and a given style. (See figure 7).

Figure 7: examples of phonemic rules.
(1) if casual style
F$#D ---> **VS#D**; F#D ---> **FE2#D**;
(2) if careful style
F$#D ---> **F$#p#D**; F#D ---> **FE3#D**.

CONCLUSION

The study presented here have shown that to generate phonemic rules modelling strategies used by different speakers in different styles it seems to be necessary to take into account phonotactic constraints and specific phonological events.

In futur, perception tests on speech synthesis from the kth synthesizer will be developp to test intelligibility and naturalness involve by this type of rules.

REFERENCES

[1] **Péan V., Williams S., and Eskénazi M.**, 1993, "The design and Recording of ICY, a corpus for the study of Intraspeaker variability and the Characterization of Speaking stYles", Eurospeech, vol 1, pp 627-630, Berlin.

[2] **Prouts B**, 1980, "Contribution à la synthèse de la parole à partir du texte; transcription graphème-phonème en temps réel sur microprocesseur", thesis, France.

[3] **Lacheret-Dujour A., Péan V.**, 1994, "Towards a prosodic cues-based modelling of phonological variability for text-to-speech synthesis", ICSLP, pp 1763-1766, Yokohama.

[4] **Léon P. R.**, 1987, "E caduc : facteurs distributionnels et prosodiques dans deux types de discours", XIth ICPhS, pp 109-112, vol 3, Tallinn, Estonia, USSR.

[5] **Hansen A. B.**, 1991, "The covariation of [ ] with style in Parisian French: an empirical study of 'E caduc' and pre-pausal [ ]", Proceedings of the ETRW 'Phonetics and Phonology of Speaking Styles', pp 30_1-30_7, Barcelona.

[6] **Péan V.**, 1995, "Phonological rules modelling style variations of 'e caduc' in Parisian French spontanous speech for text-to-speech synthesis", Eurospeech, Madrid.