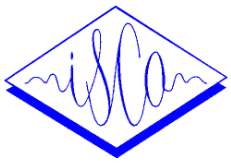


3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

Towards a Prosodic Cues-Based Modelling of Phonological Variability for Text-to-Speech Synthesis

Anne Lacheret-Dujour *$ Vincent Péan * +

* *LIMSI-CNRS BP133, F 91403 Orsay-Cedex, France*
$ *ELSAP, F-14032 Caen Cedex, France*

0. ABSTRACT

Keywords : phonological variability, prosody, syntax, text-to-phonemes conversion

The study presented in this paper has been carried out in the framework of phonological variability in French, with applications in automatic speech processing in mind. The specific aim of our study is to model inter-speaker phonemic variants at word boundaries: global strategies and strategies specific to different classes of speakers. Data have been collected from casual speech corpus. Variants are defined through their relationship with the prosodic structure of a given utterance, This structure depends on syntactic organization of discourse. One example of this phenomena is given here with the pronunciation of the schwa at words' frontiers.

1. INTRODUCTION

Two rules-based text-to-phoneme(s) conversion systems, taking into account phonological variability in French, have been developed at LIMSI. The first one, GRAPHER (8), was developed to correct segmentation errors (i.e. deletions and insertions of segments) in the ESOPE speech-recognition system (6); it constitutes one of the first attempts to process segmental variability in speaking French. The second, VARION (4), deals with the most frequent phonological variations which occur in Parisian dialect (including phonological substitutions) according to speech rate (slow or fast). VARION's tests on phonetician's transcriptions of recorded speech showed the lack of consistency in the variants distribution, while too many combinations were produced. Therefore a new system must be developed, taking into account the distributional consistency of phonological occurrences in a given style. In such a dynamic system the variants are not processed as autonomous events, they are generated according to their phono-syntactic relationships (5).

The development of such a system first implied the study of segmental variation in casual speech (i.e. non reading speech): the ICY database is first described. A segmental analysis of the data is then provided in terms of statistical quantification. A qualitative study of both segmental and suprasegmental strategies is also presented. Prosodic cues that are used for segmenting an utterance in intonation units are given. They are based on the following hypothesis: the melodic continuum is organized on prosodic events first according to discursive and syntactic constraints: these events reflect the more or less strong syntactic dependency which links words and sequences of an utterance. We show how this dependency has direct consequences on segmental strategies. Finally, we present examples of phonological rules, with consequences of their modification on synthesized speech.

2. PRESENTATION OF THE DATABASE

2.1 Corpus, task and speakers

The ICY database (9) was developed to study the inter- and intra-speaker variability that occurs when a speaker tries to speak more carefully.

+ *Names in alphabetic order*

ICY was recorded to collect three different styles of speech: two spontaneous and one read. Spontaneous speech is considered be "non-read" speech: the structure of speech of a speaker vary with the context of discourse and with his psychological state. Thus a lot of different types of spontaneous speech exist, and speech obtained in a laboratory in a specific context for a specific goal is only one of them.

To collect the data (and to generate a modification of the speaker's performance linked only to style variation) the following paradigm was developed: the speaker's task is a description of two drawings which differ in some of their parts. Each speaker has to describe each object that differed from one drawing to the other, with its colors and spatial positions. Many phonological contexts (i.e. where phonological variation could occur) are obtained by constraining the speaker to mention these objects in his description: each object which differs in the two drawings may be, with the constraints imposed during the task, described using groups of words which contain a phonological context at word boundaries (for example: robe bleue, geminate context /bb/). The phonological contexts chosen are: gemination, palatalisation, nasalisation, voicing, devoicing, and schwa.

To obtain the three different styles, a goal is given to the speaker for all of recordings: we are making recordings to help hard of hearing children to learn lip-reading. The speaker goes through the task three times. The style of spontaneous speech studied here is collected first, when the speaker describes the four drawings just to rehearse: it is the casual speech. Later a "real" recording provides clear speech, and a subsequent rereading of the transcription of the first session provides read speech.

Presently there are 21 speakers recorded, the data is stored on WORM, WARM and DAT: each speaker occupies about 60MBytes. The results presented concerned only four speakers (BP, RF, GS, GM).

2.2 The segmental statistical analysis of variants

The variation studied

The study concerns the phonological variation which occurs in a casual style between some speakers and a reference of Parisian's speech. 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 are part of the system of reference of Parisian's speech: GRAPHON(11), a grapheme-to-phoneme module conversion.

The use of a reference is imposed because to straight out comparison between two speakers necessitates the same linguistic content in their two recordings. 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 can be compared.

The methodology

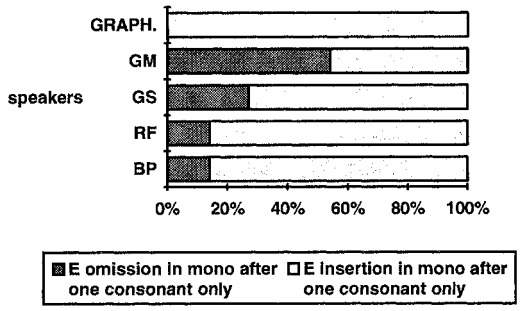
The analysis method begins with the ortographic transcription of the recordings for each speaker from GRAPHON. This automatic process provides a homogeneous grapheme-to-phoneme translation with pauses, word boundaries, and syllable boundaries within words, which will be called the "ideal" phonemic transcription for a given speaker.

Then a correction (labelling) of the ideal phonemic string is carried out according to what the speaker really pronounced, by listening and using an acoustical representation. The corrected phonemic transcription, which will be called the "real" phonemic string, is then automatically compared 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 gauche"; after GRAPHON, the ideal phonemic string obtained is:
#IL#I#A#<#N^AJ#JON#SYR#LE#D(~S<#DE#DRWAT# corrections give the following real phonemic string:
#Y#A#<#N^A#JON#SYR#L#D(~S<#D#DRWAT#;
the string comparison leads to a file with the following elements:
o.E/img.D-->E<--D.emd/#DE#DRWAT#, which means there is deletion (o) of the segmental unit E in a 'D' intra-word left context (img.D) and in a 'D' inter-word right context (D.emd) with the lexical context 'de droite' (#DE#DRWAT#). The sharp sign '#' mark the word boundaries and the tilda ~ marks the intra-word syllable boundaries.
Then the resulting file is used to automatically generate a file which contains different aspects of information about the event concerned. The structure of each element of this file is [1)¦2)¦3)¦4)¦5)¦6)¦7)] where: 1)is the event {i or e or s} for **i**nsertion, **o**mission or substitution; 2) is the element(s) concerned {x or [x...x]/ x or [x...x]} by the event; 3) is {i or e/{I or F}}, **i**ntra-word or inter-word position of the element(s), and in the last case at **I**nitial or **F**inal word; 4) gives {l_c/r_c}, the left phonemic context (l_c) and the right phonemic context (r_c); 5) is {M or P/{I/{nb_syll} or M/{nb_syll/{num_syll}} or P/{nb_syll}}}, information about the word, **M**ono- or **P**oly-syllabic, in this last case the syllabic position of the element, **I**nitial syllable or **F**inal syllable or **M**edian syllable, and the number of syllables (nb_syll) in the word, and in the case of median syllabic position, the numero of the syllable (num_syll); 6) gives the word(s) which contains the element(s); 8) is the occurrence of the complete event.
The automatic processing of this file permits easy acces to any information needed: the information in each step is more and more precise.

Results

The results presented concern the schwa and four speakers.
The comparison between the real phonemic string of each speaker and the corresponding ideal phonemic string obtained by GRAPHON shows that the percent of omissions of the phoneme [E] in mono-syllabic word is very variable: some speakers seems to have the same strategy and some others have very different strategies (see figure 1).

fig.1



Our modified GRAPHON generates poly-syllabic words which never have a phoneme, [E], in the final position. So to compare speakers we have to consider the insertion of [E] in final words again by comparison with GRAPHON (see figures 2a and 2b). These phenomena shows also groups of phonological behaviour. The phonotactic constraint of mono- or poly-syllabicity, or the consonant group preceding the final [E], are correlated with the insertion of [E]: for example BP, in mono-syllables systematically inserts [E] after a consonant group (liquid,non_liquid) and GM does the same, but after a (liquid,non_liquid) group.

Fig.2a

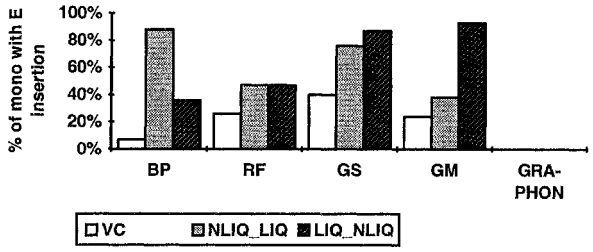
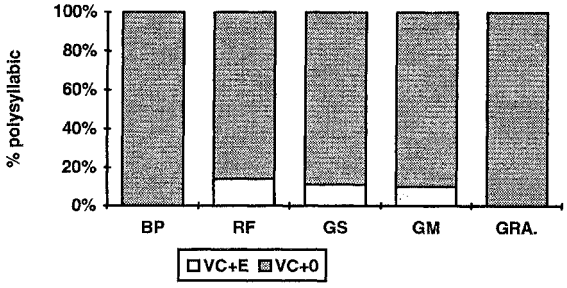


Fig.2b



Thus the comparison above shows that a fixed description is not enough to describe speech communication, there is not one form of Parisian speech but one Parisian speech in a given context.
But the study above, which brings out the variability of the speakers' strategies, is based on the [E] element after consonants without distinction between filled pauses (i.e. the grapheme 'euh') and schwa ('e'). If the distinction is in mono- and poly-syllabic words (see figures 3a and 3b), the results show that the strategy of each speaker is different and controlled by phonotactic constraints.

Fig.3a

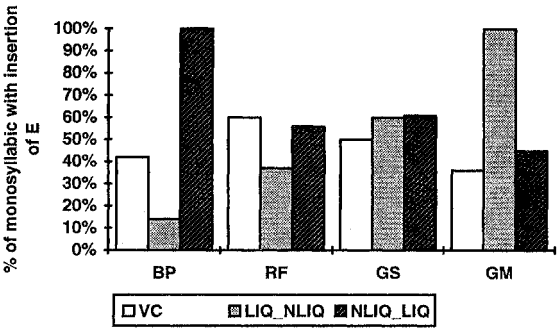
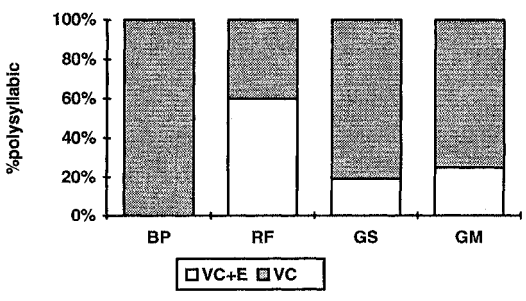


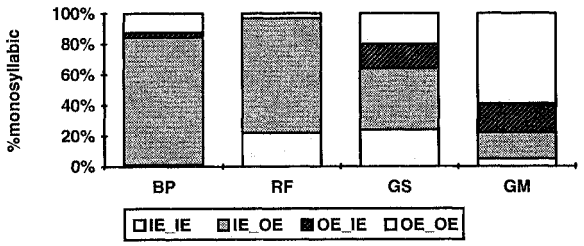
Fig.3b



The insertion of [E] after a vowel was studied. These corresponds to a filled pause ('euh') placed after a vowel; this phenomenon is not frequent in this corpus.

The studies above show that rules for insertion or deletion of schwas have to take into account phonotactic constraints. Moreover the variation which is generated in a string of sounds is not independant of what happened before. To illustrate this, we studied the correlation between the pronunciation of schwa in a group of two successive words: the first word is from the set {le/te/me/se/ce/ne/je le/de} and the second word has a final grapheme, 'e' (figure 4).

Fig.4



It appears that the deletion of schwa in the second word (x_OE) is linked to the deletion of the schwa of the .first word (OE_x).
A rule can be developped to illustrate this for a given speaker: #word1#word2#, if word2=mono-syllabic then: if word1 is from {le,te,me,se,ce,ne,je,le,de} then the final ('e') of word2 is [0] or [E]; else if word1 is from {l(e),t(e),m(e),s(e),c(e),(e),j(e),l(e),d(e)} then the final ('e') of word2 is [0].
Some examples of rules which model one given speaker and take phonotactic constraints and coherence into account have been tested on a synthesizer and give more natural speech, apparently easier to listen to.

3. QUALITATIVE STUDY OF THE DATABASE

3.1 Prosodic events: cues of grammatical dependency

The recordings were analysed using the pitch estimator of the UNICE signal analysis software developed at LIMSI (2). Our study is therefore based on acoustic data (the signal is analysed as it is produced) and not on perceptual criteria. A perceptual analysis must be the second stage of our work.
A set of phonological rules has been developed on two speakers and tested on two others. They are used to segment the melodic continuum into intonation units (IU). An intonation unit is a string of tones, a tone being either high (H) or low (L). The rules are the following (see figure5):

Rules for tone labelling

- The continuum is labelled in words and syllables. Every F0 value is taken into account, including those of the non final syllables of words and of functional words.
- A pi+1 point is labelled 'L' when it is less or equal than the point p; it is labelled 'H' in other cases.

Rules for segmentation in intonation units

- The first syllable of a IU is either 'H' or 'L'; the last syllable is always 'H' (the acoustic correlate with the phrase stress in French).
- H tones which do not end lexical words will never end a IU (they are labelled Hi).
- When we have a series of several high tones, the last one will end the IU.

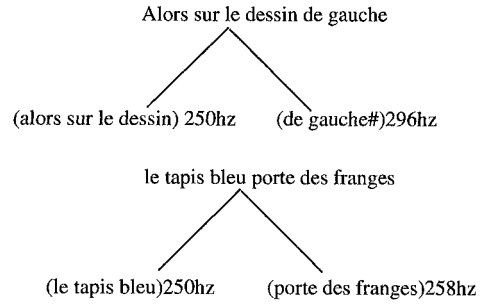
A	((alors	sur	le	dessin)	(de	gauche))		
	((242-205	205	118	242-250)	(222	296))		P= 456ms
	((H-L	L	L	H-H)	(L	H))		
	RI-LI	**RD-LI**	**RI-LD**	**RI-LD**	**RD-LD**	**RI-LD**		
B	((le	tapis	bleu)	(porte	des	franges))		
	(222	229-229	250)	(222-205	205	258))		
	(L	Hi-L	H)	(L-L	L	H))		
	RD-LI	**RD-LD**	**RI-LD**	**RD-LI**	**RD-LD**	**RI-LD**		
C	((alors	que	sur	le	dessin)	(de	droite)	
	(211-211	$	205	205	216-229)	(211	250))	P= 715ms
	(L-L		L	L	Hi-H)	(L	H))	
	RD-LI	**RD-LD**	**RD-LI**	**RD-LI**	**RI-LD**	**RD-LD**	**RI-LD**	
D	(le	tapis	est	bleu)				
	(211	190-200	207	250				
	(L	L-H	H	H)				
	RD-LI	**RI-LD**	**RD-LI**	**RI-LD**				

Legend
'R' = right; 'L' = left; 'I' = syntactic independency; 'D' = syntactic dependency; 'P' = silence;'$' = deletion of vocalic nucleus -> no F0 value.

Figure 5 : example of tone labelling & segmentation in Intonation Units

Following the concepts of prosodic fitting and/or autonomy proposed for spoken French in (7), the linear analysis of tonal sequences allows the identification of two classes of IU: autonomous units and units fitted in larger IU. These melodic fittings seem to be the mark of a specific discourse hierarchy. Spotting a prosodic fitting obeys to the following principle: two IU are fitted in a larger one when the terminal tone of the first IU is lower than the last tone of the second one (see figure 6).

Figure 6: example of prosodic fittings

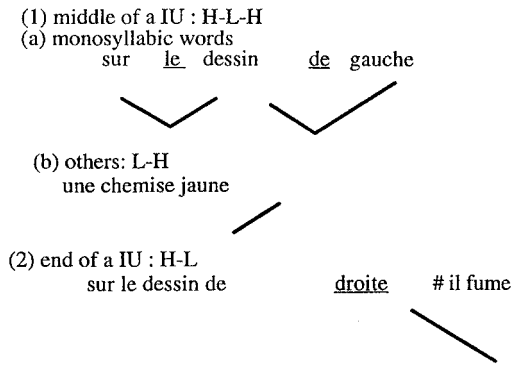


Legend
= Silence

Autonomy between two segments shows a grammatical disjunction. This is all the more important since a IU terminal tone is high in relation to the register of a speaker and to the initial tone of next IU. This disjunction is often marked by a silence on the signal: the pause is then a redundant cue for identifying a grammatical frontier (see figure 5 & 6).
On the contrary, fittings reveal a grammatical linkage. For the same speaker, they may change from one production to the other: in figure 5 - occurrence B -, the speaker produces two fitted IU, while in occurrences 1D and 1E only one IU is realized. Nevertheless, in all cases the grammatical structure is: nominal group (NG) + verbal group (VG). While the nature of the NG (pronoun) explains occurrence 1E, it is not the case for 1D (noun). These examples highlight the inadequacy of grammars based on automatic identification of syntactic constituants to show the relationships between prosody and syntax. Other grammars based on dependency relationships between segments (10), as they occur in casual French, must be provided, as well as rythmic models.

The most frequent melodic patterns, showing the dependency relationships, for the 4 speakers studied are summarized in figure 7.

Figure 7: grammatical dependency & prosodic patterns



3.2 Segmental consequences of intono-syntactic dependency - an example: the pronunciation of the e caduc

The prosodic organization in terms of IU and autonomous or fitted patterns has direct consequences on segmental pronunciations: one example is given with the different pronunciations of the schwa between words which can be predicted by cues of dependency.

For the realization of this phoneme, we propose phonological rules which must take into account the following criteria in order to generate coherent phonological ways for different strategies of discourse (see figure 8):

- left and right phonemic context,
- phonotactic constraints (number of syllables of a group),

and also the following grammatical constraints which influence prosodic patterns:

- the grammatical class of a word (functional or lexical word),
- syntactic distribution of a word in a given occurrence.

Figure 8: examples of phonemic rules: the different pronunciations of the schwa in a monosyllabic word, when it is preceded and followed by one consonant

Functional Word

(1) if BIU: 'E' ->e
ex: *#le dessin...*
(2) if MIU: E -> (e)
ex: *#sur le dessin de gauche...*

Lexical Word

(3) If EIU : 'E' -> (e)
ex: *#il n'y a pas de chaise#le dessin qui est...*
(4) if MIU: 'E' -> $
ex: *#une chemise jaune#*

Legend

BIU = beginning of an intonation unit
MIU = middle of an intonation unit
EIU = end of an intonation unit
# = frontier of a IU which marks syntactic and prosodic independence
'E' = grapheme 'E'
e = mandatory generation of the schwa
(e) = variation in the generation of the schwa: pronunciation & deletion
'$' = deletion of the schwa

According to the grammatical category of a word (functional or lexical), phonological mechanisms are different. The possible deletion of the schwa in functional words expresses a prosodic & syntactic dependency, its obligatory pronunciation marks the beginning of a IU & marks its independence in relation to the preceding unit. In lexical words, the obligatory deletion of the schwa indicates grammatical dependency, the variants (deletion or pronunciation) express grammatical autonomy and indicate the end of a IU.

4. CONCLUSION

We have presented a new approach to processing phonological variants in casual Parisian Frenchwhich is summarized as follows: the probability of the occurrence of a given phoneme is calculated not only according to phonemic constraints; we also take into account morpho-syntactic information (types of words -functional or lexical - and syntactic distribution) which governs dependency relationships between segments. These relationships may vary from one speaker to another. Prosodic strategies highlight this phenomenon: they show how a speaker dynamically constructs his discourse and organizes dependency networks. This study has been done on 4 speakers, it must be enlarged to cover the other subjects of the database (19) in order to validate our work statistically.

Other constraints must be taken into account in order to avoid the generation of too many variations and to reproduce consistently speakers' phonological strategies for a given style. Thus, a better knowledge of grammatical category and gender of a word must lead to a finest grapheme-to-phonemes conversion. The pronunciation of the schwa in the middle of a IU illustrates this fact: the rules are different between the determiner *une* and the preposition *de*: the 'e' is frequently deleted in the determiner, its pronunciation is often linked to an hesitation (lexical encoding), therefore, this variant is not generated in our system; on the other hand, in the preposition the pronunciation of the 'e' is very variable *(ex: il a une cigarette/ -> /yn/, au dessus de sa tête -> /d¶/ or /d/).*

Acknowledgements

We would like to thank Björn Granström and the members of the Department of Speech Communication and Music Acoustic laboratory of KTH in Sweden for allowing us to use their synthesizer; and Martine Garnier for her collaboration.

REFERENCES

(1) Encrevé P., 1988, "La liaison avec et sans encha"nement - Phonologie tridimensionnelle et usages du français", Seuil, Paris.

(2) Gauvain J.L., 1989, "Mode d'emploi de Unice" , Internal Report - LIMSI-CNRS.

(3) Labov W., 1972, "Sociolinguistic Patterns", University of Pennsylvania Press.

(4) Lacheret-Dujour A., 1989, "Automatic Generation of Phonological Variations", *EUROSPEECH,* vol 2, pp. 376-379, Paris.

(5) Lacheret-Dujour A., 1991, " Phonological Variation in Read Speech, Reduction Phenomena and Speaker Classes: Do Allophonic Choices Represent Speaking Styles?", Proceedings of the ESCA Workshop on *Phonetic and Phonology of Speaking Style*, pp. 38-1-38-10, Barcelone.

(6) Mariani J., 1982, "ESOPE: un système de compréhension de la parole continue", Thesis, France.

(7) Morel M.A., Rialland A., 1992, "Emboîtements, autonomies, ruptures dans l'intonation française", in *Travaux de Linguistique du CERLICO n¡5: Subordination;* Presses Universitaires de Rennes II, France.

(8) Néel F. & al., 1986, "Module de traduction phonétique avec variantes", Proceedings of the GRECO Workshop on *Lexique et traitement automatique des langages* , pp. 129-138, Toulouse.

(9) Péan V., Williams S., 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.

(10) Tesnière L., 1959, "Eléments de syntaxe structurale", Klincksieck, Paris.

(11) Prouts B., 1980, "Contribution à la synthèse de la parole à partir du texte; transcription graphème-phonème en temps réel sur microprocesseur", thesis, France.