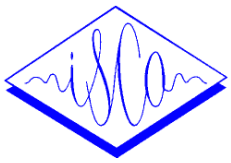

The ESCA Workshop
on Speech Synthesis
Autrans, France
September 25-28, 1990

QUALITY EVALUATION OF FRENCH TEXT-TO-SPEECH SYNTHESIS WITHIN A TASK THE IMPORTANCE OF THE MUTE "e".

Danièle LARREUR an Christel SORIN
Centre National d'études des Télécommunications
22301 LANNION FRANCE

ABSTRACT

A specific test was set up to evaluate how the elision or the pronunciation of the mute"e" inside plurisyllabic words can influence the correct identification of common words and proper names in French synthesis. The results show primarily that in order to obtain identification scores identical to those obtained with natural speech (where the mute"e" is usually elided in this position), the current quality of the CNET TTS system requires the systematic pronunciation of this mute"e". However, the decrease in identification scores observed when the mute"e" was elided in synthetic speech was notably reduced by "doubling" the surrounding consonants. Some consequences are drawn for the design of a new set of speech units, within the context of concatenation-based French synthesis.

INTRODUCTION

After the recent improvements introduced both at the signal processing level (PSOLA Synthesizer) and at the linguistico-prosodic level (CNETVOX90 Software), the CNET diphone-based Text-to-Speech (TTS) system (MOULINES and al, 1990) provides, for French, a synthetic speech quality allowing its use in various general public applications. It now appears important to focus on specific pronunciation problems, which were not carefully investigated beforehand, but can contribute to an improved listening comfort. One of these problems is, in French, the pronunciation or the elision of the mute"e".

Most of the time, in colloquial French, the mute"e" is elided if its presence does not facilitate the pronunciation of a word *(pal(e)frenier)* or of a word sequence *(je te l(e) demand(e) vraiment)*. On the other hand, it is maintained when its omission would create a consonant cluster considered difficult to pronounce by French speakers *(habiterions)*.

1- THE PROCESSING OF MUTE "e" IN THE CNET DIPHONE-BASED TTS SYSTEM

In the previous version of the CNET's TTS system (STELLA 1985), because of the limited speech quality provided by the LPC synthesizer and for guarantying a maximal intelligibility, were considered as mute"e", only the mute"e" appearing at the end of polysyllabic words : all other occurences of mute"e" were never elided and systematically pronounced as /ø/ vowels. This decision has been responsible for the choice of the diphones to be used for synthesizing the mute"e". The set of diphones remained unchanged in the new version of the CNET TTS system used in this study.

1.1- Characteristics of the current set of diphones including a mute"e".

With the exception of 2 diphones [mute"e"-silence], all the mute"e"-diphones were extracted at the boundary between 2 words, the mute"e" being pronounced inside the last syllable of a polysyllabic word. We will note /ə/ this pronounced mute"e". The [ə-silence] diphones were extracted from the logatom *euteute*, pronounced in 2 versions, with a rising or a falling Fo contour.

1.2- Rules for the pronunciation/elision of the mute"e"

It was obvious that the rough processing of the mute"e of the previous version of our TTS system was inadequate : in particular, the systematic pronunciation of every mute"e" inside polysyllabic words created an impression of "stumbling" which increased the lack of fluentness perceived by the listeners. Now, in spoken French, even in the most articulated speaking style (newspaper loud-reading), the proportion of elided mute"e", inside polysyllabic words, exceeds 40 % (LUCCI 1983). With "no elision" inside polysyllabic words, our synthetic speech was clearly "over articulated".

In the current version of our TTS system (MOULINES and al 1990), all the mute"e" are processed by a new set of rules which are applied after the linguistic analysis. This new set of rules has been derived from a first set of 52 rules provided by Professor MARTY from Illinois University (MARTY and HART, 1985). The MARTY and HART's set of rules corresponds to a straightforward implementation of mute"e" pronunciation rules mainly proposed by FOUCHE (1959) and DELATTRE (1968) and reported in several, well known books on French pronunciation (for ex, WARNANT 1987, GREVISSE 1986, LEROND 1980).

However, although this new mute"e" processing notably improved the overall naturalness of the synthetic speech, many cases of elision observed in natural speech sounded "inappropriate", "strange", "too relaxed" and even led to some understanding difficulties when replicated with our TTS system. The initial set of rules was therefore modified in the following way (only the main modification are indicated here) :

1) the mute"e" is always pronounced in monosyllabic words *("je, te, le prends")*,
2) the mute"e" is always pronounced in the first syllable of polysyllabic words *("demander")*,
3) the mute"e" is always pronounced when preceeded by 2 consonants *("barbe rousse")*,
4) the mute"e" is always elided when located at the end of a word followed by a pause,
5) the mute"e" is always elided when located at the end of a word and preceeded by a single consonant *("un(e) poir(e))*,
6) the mute"e" is always elided when located inside a polysyllabic word (first syllable excluded, see rule 2) and preceeded by a single consonant, except if the mute"e" is followed by a consonant cluster containing the phoneme /y/ (*"all(e)mand", pal(e)frenier,* but *"chanterions")*.

Informal listening sessions with various kinds of synthesized texts showed that this new set of rules notably improved the perceived naturalness of the synthetic speech. However, the last rule mentionned above (elision inside polysyllabic words) still leads, in some cases, to intelligibility difficulties, especially when it is applied to proper names.

A formal test was therefore designed to better evaluate the contribution of the mute"e", inside polysyllabic words, to the intelligibility of synthetic utterances : given the current quality of our TTS system, should it be pronounced or elided ?

2- TESTING THE CONTRIBUTION TO INTELLIGIBILITY OF THE MUTE "e" INSIDE SYNTHESIZED POLYSYLLABIC WORDS

2.1- Choice of the testing procedure

In order to take into account all the parameters that can influence the intelligibility of synthetic speech in real applications, we chose a test procedure where the sujects had to carry out a real understanding task. 10 texts, approximately 4 sentences long and only partly semantically predictable, have been designed. The test material consisted in 54 polysyllabic proper names and 28 polysyllabic common words, included in these texts. Every test word contained a unique mute"e", located inside the word (never in the first syllable nor in the last one) and always preceeded and followed by a single consonant (ex : *"foncera", "Richepin"*). After having listened to the texts, the sujects answered, in writing, some questions allowing to check if they correctly identified the specific test-words included in the texts (see Annex 1).

This test was completed by a "classical" intelligibility test where the subjects were presented with lists of isolated words that they had to write down. These lists contained 82 polysyllabic proper names and 124 polysyllabic comon words, each of them presenting the same characteristics as the test-words included in the texts. A total of 136 proper names and 152 common words with an intermediate mute"e" were thus tested.

The goal of the experiment was to compare the identification scores of the test-words in 3 different conditions : natural speech ; synthetic speech with elided intermediate mute"e" in every test-word (application of the rule 6) ; synthetic speech with pronounced intermediate mute"e" in every test-word, all other mute"e" being processed in the same way in both synthetic versions.

2.2- Synthesis of the test material

A first synthesis of the test texts was made using the mute"e" diphones available in our diphone set for every pronounced mute"e". It appeared that both inside monosyllabic words (ex *"le"*) and in the first syllable of polysyllabic words (ex *"regarder"*), the use of these ə-diphones led to a degradation of the perceived quality when compared to the quality obtained in using the ø -diphones. Knowing the way the ə -diphones were extracted (see section 1.1), **this observation suggests that the realisation of the pronounced mute"e" (and/or that of its surrounding phoneme) inside words is rather different than at the end of plurisyllabic words.**

Therefore, 2 sets of mute"e" diphones were used : the "real" ə-diphones for the realisation of the mute"e", when pronounced at the end of plurisyllabic words ; the ø-diphones for all other occurences of pronounced mute"e", while reducing the global duration of the resulting /ø/ vowel to roughly 60 ms.

After several trials it was also decided to synthesize the "elided-mute"e" version of the test material in two different ways : in a first version, the intermediate mute"e" was simply elided and replaced by the coresponding [consonant 1-consonant 2] diphone ; in a second version, the intermediate mute"e" was elided and "replaced by a "doubling" of the surrounding consonants, i.e. by the following sequence of diphones : $[C_1\text{-}C_2]$ + $[C_1\text{-}C_2]$ + $[C_2\text{-}C_2]$. This "consonant-doubling" procedure seemed to notably facilitate the identification of the consonant clusters resulting from the elision of the mute"e". It must be noted that we didn't succeed in obtaining the same improvement by simply lenghtening the cluster consonants.

To summarize, test list and texts were presented in 4 different versions :

- **V_O version** : synthetic speech, all the intermediate mute"e" being elided,
- **V_1 version** : synthetic speech, all the intermediate mute"e" being elided and "replaced" by a doubling of the surrounding consonants,
- **V_2 version** : synthetic speech, all the intermediate mute"e" being pronounced (using the shortened /ø/ vowel),
- **V_3 version** : natural speech, all the intermediate mute"e" of the test words having veen elided spontaneously by the speaker.

2.3- Running the test

32 subjects took part of the test. The test sessions were organized in a way such that the subjects were never presented with 2 different versions of the same test material. The lists of test words were presented only once. The texts were presented 3 times each, with a 4 s pause between sentences, allowing the subject to answer all the questions related to specific informations included in these (semantically complex) messages.

2.4- Results

A test-words was considered wrongly identified if the pronunciation of the corresponding written word produced by the subject differed from that of the presented item (disregarding the pronunciation of the optional intermediate mute"e"). The results are summarized on table 1. They mainly show that :

1) When the optional intermediate mute"e" is simply elided, the correct identification scores obtained with synthetic speech (V_O version) are notably inferior to those obtained with natural speech pronounced in the same way (elided intermediate mute"e"). Averaged over all the test words (texts and lists), the gaps are 6.5 % for common words and 15 % for proper names.

2) the "doubling" of surrounding consonnants in the second "elided-mute"e"" version (V_1) notably improves the identification of all the test-words : the difference between synthetic an natural speech identification scores is then reduced to roughly 4.5 % for both the proper names and the common words.

3) the systematic pronounciation of the intermediate mute"e" in the V_2 synthetic version leads to identification scores which are very close to those obtained with natural speech when the corresponding mute"e" are elided (roughly 83 % for proper names and 93 % for common words).

However, when being asked to give a **preference** rating on a new text, presented at the end of the test in the 3 synthetic versions (V_O, V_1, V_2), 28 out of the 32 subjects who participated to the test ranked the V_2 version (every intermediate mute"e" being pronounced) in the last (worse) position : the systematic pronunciation of every intermediate mute"e" was judged "not fully natural sounding"...

CONCLUSION

This study on how to process the mute"e" in a diphone-based TTS system highlights the distance which still separates top quality synthetic speech from natural speech. It shows that a compromise remains to be made, in the current state of our TTS system, between naturalness and intelligibility : to maintain the same intelligibility level as provided by natural speech, the mute"e" cannot be elided in positions where it is commonly elided in spontaneous speech.

One of the main reasons of this "pronounciation default" of our synthetizer lies in the inadequate quality of the consonant clusters appearing with the elision of the mute"e". In fact, the common elision of the mute"e" in natural speech is responsible for 30 % of the 2-consonant clusters, 50 % of the 3-consonant clusters and 92 % of the 4-consonant clusters (VAN EIBERGEN 1986).

An overall improvement of the quality of consonant clusters will probably be achieved in using, intead of our actual diphones, longer units such as, for example, those proposed by EMERARD (1986). However, it still remains to be determined if the consonant clusters resulting from a mute "e" elision are different from the "normal" clusters occuring inside words (the same remark holds also for clusters occuring at the boundary between 2 words). The relatively good results obtained in doubling the surrounding consonants suggest that such a difference indeed exists and is important enough to be accounted for in concatenation-based synthesis.
This study also showed the necessity of defining a new set of mute"e"-diphones to be used in monosyllabic words and in the first syllable of plurisyllabic words : the mute"e"-diphones extracted at the boundary between 2 words seemed to be inadequate for correctly synthesizing these contexts.
All these observations are taken into account in the work currently underway for defining a new set of speech units to be used in a future version of the CNET TTS system.

ACKNOWLEDGMENTS

Many thanks to Martine BOYER for her help in setting up the test and to Luc MATHAN for proof-reading the English version of this paper !...

REFERENCES

DELATTRE P. (1968) : "Le jeu de l'E instable intérieur en français" in Studies in French and Comparative Phonetics, Mouton, The Hague, 17-27.
EMERAD F. (1986) : "Constitution d'un dictionnaire d'éléments acoustiques et règles de concaténation", CNET Internal Report NT/LAA/TSS/276.
FOUCHE P. (1959) : Traité de Prononciation Française, 2nd edition Klincksieck, Paris.
GREVISSE M. (1986) : Le bon usage. Grammaire Française, 12th edition Duculot, Paris.
LEROND A. (1980) : Dictionnaire de la Prononciation Française, last edition, Larousse, Paris.
LUCCI V. (1983) : "Le E muet", in Etude Phonétique du Français Contemporain à travers la Variation Situationnelle, Editions de l'Université de Grenoble, Grenoble, 105-139.
MARTY F. and HART R.S. (1985) : "Computer program to transcribe French Text into Speech : Problems and suggested Solutions"-Tech. Rep. LLL-T-6-85, University of Illinois, Urbana, USA.
MOULINES E. and al (1990) : "A Real-time French Text-to-Speech System generating High-Quality Synthetic Speech", Proceed. IEEE-ICASSP 90, 309-312.
STELLA M. (1985) : "Speech Synthesis", in Computer Speech Processing, Fallside F. ed., Prentice hall, 421-460.
VAN EIBERGEN J. (1986) : "Le E latent en français", in Bulletin de l'Institut de Phonétique de Grenoble, Vol XV, 75-107.
WARNAND L. (1987) : "Dictionnaire de la Prononciation Française dans sa Norme Actuelle", 4th edition, Duculot, Paris.

	V2 **Synthetic Speech** *(pronounced mute "e")*	V3 **Natural Speech** *(elided mute "e")*	V1 **Synthetic Speech** *(elided mute "e" and doubled consonants)*	V0 **Synthetic Speech** *(elided mute "e")*
		LISTS		
Proper Names	87,6 % (1,6)	83,7 % (5,6)	78,9 % (2,2)	65,3 % (5,7)
Common Words	93,4 % (0,8)	94,5 % (1,2)	90,0 % (1,5)	87,4 % (1,6)
		TEXTS		
Proper Names	82,4 % (3,2)	82,5 % (5,5)	80,0 % (3,0)	71,5 % (2,2)
Common Words	91,0 % (3,9)	92,3 % (2,3)	91,3 % (1,7)	88,4 % (2,4)

TABLE 1

PERCENTAGE OF CORRECTLY IDENTIFIED TEST WORDS

(standard-deviation in brackets)

TEXT 8

Le petit chaperon rouge après avoir mangé le loup de la forêt d'**Eppeville** se sentit l'estomac lourd. Elle cour ut chez sa mère-grand qui fumait sa pipe goulûment et joua **sublimement** une sonate de **Couperin**. Tout en jouant, elle pensait que sa robe zébrée qu'elle avait achetée à **Massevaux** ne lui allait pas. Elle irait la changer et peut-être rencontrerait-elle ce **vaguemestre** si charmant qui appréciait tant La **Rochefoucault.**

The test words are indicated here in boldfaced type on the text.

ANSWER THESE QUESTIONS	*Expected answers*
1) Où est située la forêt ?	*Eppeville*
2) Quel morceau joue le petit chaperon rouge ?	*une sonate de Couperin*
3) Comment a-t-elle joué ce morceau ?	*sublimement*
4) Où avait-elle acheté une robe ?	*à Massevaux*
5) Qui espère-t-elle rencontrer ?	*le vaguemestre*
6) De quel écrivain est-il question ?	*La Rochefoucault*

ANNEXE 1

EXAMPLE OF A TEST TEXT AND RELATED QUESTIONS