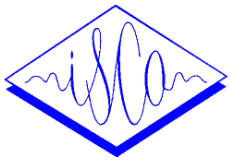

Fourth ISCA ITRW on
Speech Synthesis (SSW-4)
Perthshire, Scotland
August 29 - September 1, 2001

A selection/concatenation text-to-speech synthesis system: databases development, system design, comparative evaluation

Romain Prudon & Christophe d'Alessandro

LIMSI-CNRS
BP133 - Université Paris XI, F91403 Orsay, France
Romain.Prudon@limsi.fr, christophe.D'Alessandro@limsi.fr

Abstract

This paper describes the development of a new text-to-speech synthesis system in French. The system is based on selection and concatenation of natural speech segments, taken in large annotated speech data bases. In a first part the databases design, content and annotation procedures are presented. It appeared that about 1 hour speech databases are large enough for a building a TTS system. In a second part, the system architecture is described. A key feature of the present system is that only 4 simple and efficient selection criteria are proposed. A formal comparative evaluation procedure is described in the third part. The experiments show that the new system is preferred along all the evaluation categories to the previous system, which is based on diphone concatenation and synthesis by rules of the prosody. The most significant improvements brought by the new system seems to be for voice pleasantness and overall impression.

1. Introduction

This paper describes the development of a new corpus-based text-to-speech (TTS) synthesis system for the French language, which is based on speech segments selection and concatenation. Along the lines proposed for Japanese and English by e.g. [8, 4, 5, 3, 6], following the pioneering work of [11], non uniform speech segments are used for synthesis. The present synthesis system does not use any rules either at the segmental or suprasegmental levels. It is entirely based on optimal selection and concatenation of natural speech segments.

The system can be split into tree main components. A preliminary phase consists in choosing or recording a given speech database, together with the associated texts. The task of the first (off-line) component consists in automatic analysis and labeling of the speech and text databases. Then two (on-line) components are used for synthesis: a text analysis and target generation component, and a signal selection and concatenation component. The goal of the text analysis/target generation component is to convert the input text into a phonological description consisting in a phoneme chain associated to some sort of prosodic and accentual description. The goal of the selection/concatenation component is to find in the database possible speech segments, according to the desired target, and to select an optimal sequence of segments. Then this optimal sequence is smoothly concatenated to give birth to synthetic speech.

Both a phonological description of the database (e.g. phonemes, break indices, syllables, ldots) and a phonetic description of the database (e.g. segment durations, pitch . . .) are needed for segment selection. These annotations are generated almost automatically, using automatic speech alignment, and automatic prosodic analysis and stylization. Compared to other selection synthesis systems, our system makes use of very few criteria for segment selection, and is based on a very simple data structure. This allows for a relatively small system, which runs in real time on a standard personal computer.

The speech quality obtained has been checked using a subjective overall quality test [12]. The new system has been compared to our diphone TTS, and demonstrated a significant improvement along certain lines, like voice pleasantness.

This paper is organized as follows. In section 2, the database design and development is described. Section 3 presents the architecture of the synthesis system, the data structure, the selection criteria, the optimal selection algorithm, and the concatenation procedure. Section 4 presents the perceptual assessment test and discusses its results. Section 5 is a brief conclusion.

2. Database structure and development

2.1. Voices, lexical and segmental content

A main advantage of corpus-based synthesis is the possibility to change quickly and efficiently the system voice, and then to offer multiple voices, or even multiple voice qualities depending on the content of the speech database.

For this project, we started with a publicly available large database of read speech, containing speech samples and texts for 120 male and female speakers [7]. This database has been carefully designed for training speaker independent speech recognition systems. It contains for each voice all the phones, and most of the diphones and triphones that are actually used in French read speech. Several speakers from BREF have been used as sources for the synthesis system. This database contains phonemes strings aligned with the input orthographic text and the acoustic signal.

In a second stage, we recorded 3 new speech databases. These bases, that are used in the evaluation experiment reported in section 4, are only medium sized, about 20 or 30 times the size of a diphone database. The texts already used for BREF are read by a female and two male speakers. Speech signal is recorded in a sound insulated booth, on the two tracks of a DAT recorder: acoustic signal is recorded on the first track, and the Electroglottographic signal is recorded on the second track. For each speaker, about 1 hour to 1,5 hours of speech is recorded at 48 kHz, and then downsampled at 16 kHz. For instance for the database **sel1**, 550 sentences are recorded (about 75 mn of speech). These 550 sentences contain 10784 phonetic words, corresponding to 2923 different phonetic words. Phonetic word means the phonetic transcription of an orthographic word. Despite this relatively little number of different phonetic words, the data base coverage in terms of phonetic words is rather

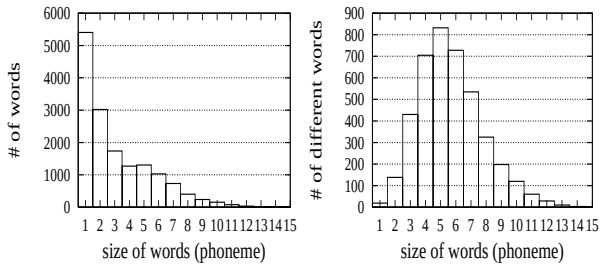

Figure 1: *Phonemic content of phonetic words. Left panel: number of occurrence of phonetic words as a function of their number of phonemes. Right panel: Number of different phonetic words, as a function of their number of phonemes.*

good. When taking at random 10000 words in the same type of text corpus (same newspaper texts), our experiments showed that more than 73 % of the corresponding phonetic words are present in the database. When performing the same measure using a list of 100000 words, the same result is obtained. Of course this result may be subject to changes according to the particular domain which is aimed at for synthesis. In the present study, we envisaged only newspaper texts.

In terms of segmental units, the 75 mn database contain of course all the phonemes, about 927 diphones (i.e. 75 % of the theoretically possible diphones), and about than 2200 different (phonetic) syllables. The database coverage in terms of syllables is very good. Like for the words, when taking 10000 syllables at random in the same type of text corpus, one find almost 95 % of the phonetic syllables in the 75 mn database.

Figure 2.1 illustrates the phonetic content of the phonetic words (these figures were computed on the same 75 mn database). The average length for different phonetic words is 5 phones (see right panel). However this figure does not take into account the frequency of occurrence for phonetic words. On the left panel, it is shown that most of the phonetic words observed in the database (and thus presumably in texts to be synthesized also) are mono- or diphones.

Finally, it must be pointed out that the 75 mn speech database results in about 144 Mbytes of raw data: it is 28 times larger than the size of our diphone database for French synthesis.

In summary, even a relatively "small" large corpus covers most of the diphones, most of the syllables, and a large majority of the phonetic words that are likely to occur, at least when synthesizing texts similar to those used for building the speech database.

2.2. Labeling

After signal recording, the database must be enriched with phonemic, phonetic and prosodic information. The speech signal and phonetic transcription are aligned automatically using the LIMSI speaker-independent speech recognition system [1]. For alignment of read texts, this system is reliable enough for our speech synthesis purpose, as it make very few errors. The output of the speech/phoneme alignment stage is very simple: for each speech signal file, an associated phone file containing the first and last samples for each phonemic label is produced. In the following, phoneme means the abstract segmental category, and phone any occurrence of a phoneme. Note also that the silence is considered as a phoneme. It means that sentence

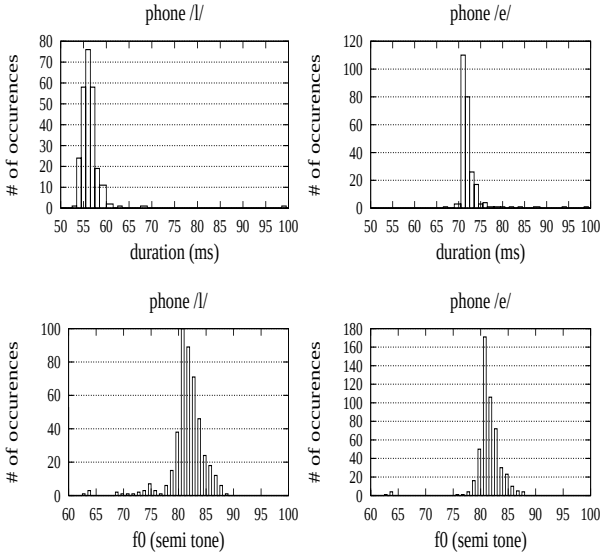

Figure 2: *Histograms of durations (top panels) and pitch (bottom panels) for the phoneme /l/ (left panels) and /e/ (right panels), for all diphones /le/ in the* **sel1** *database.*

initial and sentence final phonemes are implicitly stored in the database.

The following information is derived from automatic phoneme labeling, and stored in the database:

1. phone labels, and phone position in the speech file.
2. position of the phone in the phonetic word (initial, median, final)
3. phonetic syllables
4. right context of the phone

Some prosodic analysis is also performed on the speech signals. Pitch and durations are computed for each phone. The pitch value that is retained is not the raw pitch value, but the perceptually-stylized pitch value, according to the Weighted Time Average Model (WTAM) of syllabic pitch perception [2]. Statistics of pitch and durations are also computed. The following information is derived from automatic F0 analysis and stylization, and stored in the database:

1. WTAM pitch for each phone.
2. duration for each phone.
3. histograms of WTAM pitch for all phonemes.
4. histograms of duration for all phonemes.

Figure 2.2 shows the repartition of durations (top panels) and pitch (bottom panel) for two examples of phoneme (/l/ left panels, and /e/ right panels), taken in the diphone /le/. Note that there are 289 occurrences of the diphone /le/ in the **sel1** database . These histograms are used before the synthesis stage for estimation of average values for each phoneme, and for database pruning. To avoid too much intonation discontinuity in the synthetic signal, phones with extreme values are filtered out of the data base, and are not used for synthesis.

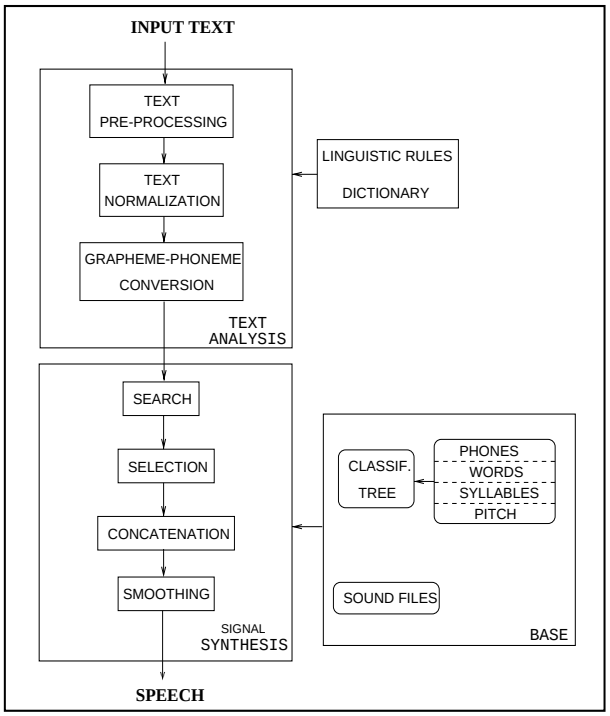

Figure 3: *Architecture of the Selection/Concatenation synthesis system*

2.3. Database structure

We tried to keep organization of the database as simple as possible, using a 2-layer classification tree, including phonemes and diphonemes. Each cell of the first layer correspond to one phoneme (35 phonemes are used in this system for the French language): the cell is pointing to all the occurrences that are present in the database, together with the associated information. This first layer is pointing to the second layer, which contains diphone informations. For each diphone, the cell is pointing to all the occurrences that are present in the database, together with the associated information. This does not mean that diphone units are playing a specific role at the synthesis stage. On the contrary, all types of segment, ranging from one phones to chains of several phones, are actually selected for synthesis. Clearly, any chain of N phones can be retrieved using a diphone chain: therefore it is not necessary to store long units, like 3-phone, 4-phones, 5-phones and so one.

3. Synthesis system

An overview of the synthesis system is given in Figure 3. In addition to the database that has been discussed above, the system can be split into two main functional components. The first component is the text analysis system, and the second component is the selection and concatenation engine, that will be reviewed in more detail below.

3.1. Text analysis

The first part of the synthesis system is shared by the selection/concatenation and by the diphone synthesis systems. It is based on previous work on text preprocessing, grapheme-to-phoneme conversion, morphosyntactic analysis and prosodic phrasing in French. The output of the text analysis stage is a target phonological representation of the speech utterance to be synthesized. This target representation is defined in terms of a phoneme chain, enriched with prosodic boundaries and prosodic marks. The goal of the selection algorithm is to find in the database the optimal chain of speech segments that can be considered as a particular production of this phonological target representation.

The phonological target is expressed as a 3-layer chain. The first layer corresponds always to phonemes (including silence, that is considered as a phoneme). The second layer contains the position of the phoneme in the word (initial=1; median=0; final=2; monophone=3). A third layer contains the position of the phone in the syllable. With this mechanism, function words and contents words are very likely to be selected at the good places. Also, word accents are very likely to be selected at their places, because (in French) they depend mainly of the initial and final syllables. Consider for instance the the part of sentence:

. . . tout en demeurant le . . .

Then, the corresponding tree-layers phonological target will represent both the phonemes and their position in phonetic words and phonetic syllables:

t	**u**	**t**	**A**	**d**	**x**	**m**	**x**	**r**	**A**	**l**	**x**
1	0	2	3	1	0	0	0	0	2	1	2
1	2	1	2	1	2	1	2	1	2	1	2

3.2. Selection criteria

The aim of the selection algorithm is to find the optimal chain that may be considered as a particular production of the target chain. In this system, we tried to reduce the selection criteria as much as possible, contrary to other systems, that are using 20 or 30 criteria [8]. This is because on the one hand it is very difficult to understand the meaning of a criterion among many criteria. On the other hand few criteria result in a smaller and faster system.

The first part of the selection algorithm search in the data base all the possible combinations of segments that may represent the input target chain (i.e. the good phones in the good positions). Then the optimal chain among the array of possible chains is searched for using a dynamic programming algorithm. This algorithm makes use of a distance, or cost, between chains, that are weighted according to selection criterion. Only four selection criteria are used in the current version of the system:

Target criterion 1: find the good phone with the good position in the corresponding phonetic word.

Target criterion 2: find the good phone with the good position in the corresponding phonetic syllable.

Concatenation criterion 1: take adjacent phones in the same speech segment.

Concatenation criterion 2: avoid too large pitch differences between adjacent phones.

For each candidate chain, the associated cost C is computed as the sum of the costs for each sub-chain that is composing the chain C_i in the chain: $C = \sum_i C_i$. The cost of each sub-chain, is computed as a weighted sum of two components for each segment u_i. The first component (with the associated target weight ω_t) is the target cost $C_t(u_i, u_t)$ corresponding to the target criteria. The second component is the concatenation cost (with

the associated concatenation weight ω_c), according to the concatenation criteria, between u_i and u_{i-1}, namely $C_c(u_i, u_{i-1})$. Then the cost for a sub-chain is given by:

$$C_i = \omega_c C_c(u_i, u_{i-1}) + \omega_t C_t(u_i, u_t)$$

The concatenation cost in turn is decomposed into a chain continuity cost C_{cc} (with the associated continuity weight ω_{cc}), according to the first concatenation criterion, and a pitch continuity cost C_{cp} (with the associated continuity weight ω_{cp}) corresponding to the second concatenation criterion:

$$C_c = \omega_{cp} C_{cp}(u_i, u_{i-1}) + \omega_{cc} C_{cc}(u_i, u_{i-1})$$

The target cost is decomposed into a syllable position cost C_{ts} (with the associated weight ω_{ts}), according to the first target criterion, and a word position cost C_{tw} (with the associated continuity weight ω_{tw}) corresponding to the second target criterion:

$$C_t = \omega_{ts} C_{ts}(u_i, u_t) + \omega_{tw} C_{tw}(u_i, u_t)$$

For the moment all the empirical weights have been chosen by a simple trial-and-error procedure. The best values we found are as follows:

ω_c = 0.6		ω_t = 0.4	
ω_{cp} = 0.15	ω_{cc} = 0.85	ω_{ts} = 0.5	ω_{tw} = 0.5

Then, more weight is given to the target criterion than to the concatenation criterion, and more weight to adjacent phones than to pitch continuity.

This very simple mechanism has several advantages. There are few criteria, therefore it is quite easy to test the effects of these criteria, and two search the weights accordingly. A main effects of these criteria is to favour adjacent phones, therefore to find the longer chains in the database. A second main effect is also to favour phones that have the in good prosodic positions. If a chain of phone is chosen in the good position, it is very likely that is will have also an acceptable accentuation, both at the word level and at the phrase level. Then, these effects are not searched for directly, but are obtained as by-products of the search procedures. finally, the selection algorithm is very simple, and then is able to select a speech chain in real-time on a standard desktop computer.

The last step of the synthesis procedure is segment concatenation. For the moment, this is achieved by a simple smoothing of the segments edges, in order to avoid clicks. No additional signal processing is used at this stage.

3.3. Results

Figure 3.3 shows a typical example of segments selection for the sentence "Salut les amis!" ("Hello, friends!", /salylezami/). For this sentence, all the segments are made of 3 or 4 phones, two full words are chosen ("les" and "ami") and the phrase level and word level accentuation is also correct. In such a situation, the synthetic speech is as good as recorded natural speech.

Some statistics on the synthetic chains have been computed. For a typical text containing about 20 sentences, each sentence containing about 70 to 90 phonemes, one find an average length of 3.72 adjacent phones per selected phonetic segment. This figure is to be compared to the average length of phonetic words, reported in Figure 2.1. An impressive result is that about 94 % of the synthesis phones are found in the good position in the corresponding words. Finally, without taking into account monophone words, on average 28 % of the selected phonetic segments are actually phonetic words. Then, a typical sentence contains between a quarter and a third of full words that are present in the database. This figure is of course much better if one takes also into account monophone words, that are all present in the database.

phone chaine	context
.sal	. salarimen
ly l	plus l'apanage
lez a	les assayants
ami.	son ami.

Table 1: *Example of selection for the sentence "Salut les amis!". On the left are the selected phonetic segments, and on the right the pieces of text where these segments were found.*

The general informal impression when listening the selection/concatenation synthesis system may vary a lot according to the input sentence. For some sentences, the result may not be distinguished from recorded natural speech. For other sentences, attention is triggered either by a prosodic discontinuity (e.g. pitch jump, incoherent accentuation) or by a segmental error (e.g. missing segment, too short segment). However, the voice quality is generally very good. In order to assess speech quality of the new system, we performed a formal subjective quality test. The Selection/Concatenation system is compared to our other TTS system, which is based on diphones, and to recorded natural speech.

4. Overall quality test

4.1. Systems tested

The speech quality of the new selection/concatenation system seems quite different of the sound quality of our diphone-based system. Both systems are still clearly recognized as synthetic speech, provided that few sentences are played (for the selection/concatenation system, it may happen that short sentences reach a natural speech quality). But both systems do not have the same shortcomings. Therefore it is important to check the quality of the new system, with the help of a formal subjective evaluation procedure. In the present experiment, three types of speech are tested:

Recorded speech: sentences are extracted from the BREF database, for 3 speakers.

Selection/Concatenation speech: is made using the the system described in this paper.

Diphone speech: is produced by LIMSI diphone TTS system. The linguistic component of this system is similar to the linguistic component of the selection/concatenation system. There are two main difference between the two systems. On the one hand, the diphone system is based on signal processing of a small speech base, resulting in a rather uniform voice quality. On the other hand, in the diphone system, prosody is computed using a set of syntactico-prosodic rules, which are defining all the details of pitch contours and phoneme durations. An overview of the system is given in Figure 4.

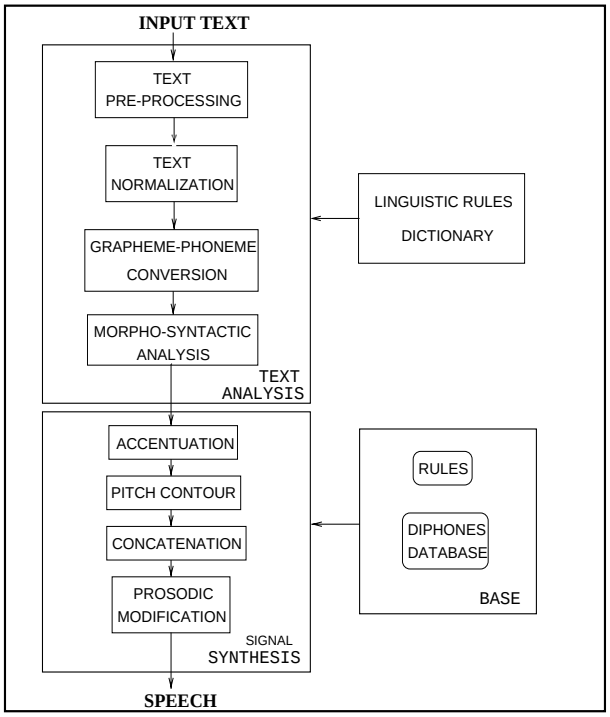

Figure 4: *Architecture of the diphone TTS system*

4.2. Method

Ten subjects participated in the experiments. They were all naive subjects, with no special training in speech synthesis or speech processing. They were also all native speakers of French, students ranging between 22 and 27 years old. Some were member of the laboratory. Stimuli, recorded on an audio CD were presented through a pair of loudspeakers, using Hi-Fi equipment in a sound-insulated quiet listening room.

The subjects listened to 18 extracts, short read paragraphs ranging between 8 and 22 s (either a long sentence, or several short sentences). There were 6 different voices: **dip1** and **dip2** are diphone voices (male voices), **sel1**, **sel2** and **sel3** are selection/concatenation voices (**sel3** is a female voice, **sel1** and **sel2** are male voices). **nat** means natural voice (3 different speakers).

The text material (and also the natural voice examples) are sentences or short paragraphs extracted from news. All the signals were sampled at 16 kHz.

For each sound example, the subjects were asked to fill a same test questionnaire, inspired by the overall quality test ITU-T 1993 proposal [9]. There were 7 types of questions, using a 5 points scale.

voice pleasantness (How would you describe the voice ?) 5: very pleasant, 4: pleasant, 3: fair, 2: unpleasant, 1: very unpleasant.

overall impression (How do you rate the sound quality of what you have just heard ?) 5: excellent, 4: good, 3: fair, 2: poor, 1: bad.

Listening effort (How would you describe the effort you were required to make in order to understand the message ?) 5: complete relaxation possible, no effort required, 4: attention necessary, no appreciable effort required, 3: moderate effort required, 2: effort required, 1: no meaning understood,with any feasible effort.

Comprehension problems (Did you found certain words hard to understand ?) 5: never, 4: rarely, 3: occasionally, 2: often, 1: all the time.

Articulation (Where the sounds distinguishable ?) 5: yes, very clear, 4: yes, clear enough , 3: fairly clear, 4: no, not very clear, 5: no, not at all.

Pronunciation (Did you notice any anomalies in pronunciation ?) 5: no, 4: yes but not annoying, 3: yes, only slightly annoying, 4: yes annoying, 5: yes, very annoying.

Speaking rate (What do you thing of the average speed of delivery ?) 5: much faster than preferred, 4: faster than preferred, 3: preferred, 4: slower than preferred, 5: much slower than preferred.

The subjects listened two times to each example before reporting their responses. Completion of a full test took about 40 minutes, and 30 responses were available for each voice. All the responses were averaged for each voice, resulting in a mean opinion score (MOS) for each voice and each condition.

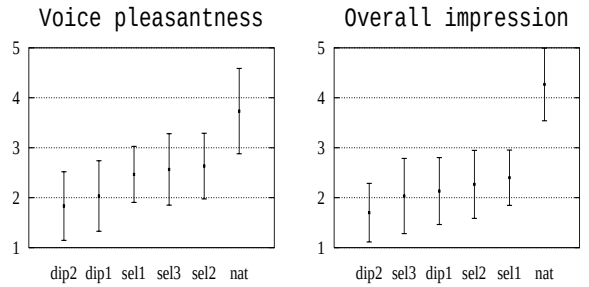

Figure 5: *Subjective evaluation: Voice Pleasantness and Overall Impression.*

4.3. Analysis of results

Results of the test are reported in Figures 5 6 7. The speaking rate has been judged correct for all the voices, and will not be discussed further herein. Generally, **sel1** and **sel2** have the best scores for synthetic voices. Results for natural voice are always clearly separated from those for synthetic voices, for all conditions excepted for the speaking rate. Note that in this test there is no condition made using degraded natural voice. Therefore, it is likely that the subjects recognized natural voice and applied a different strategy for natural voice/synthetic voice evaluation (and indeed several subjects reported that they used such a strategy). In similar experiments using degraded speech conditions, one observe generally less deviation between synthetic and natural speech (see e.g. [10]).

Figure 5 is displaying voice pleasantness and overall impression. As for voice pleasantness, the MOS difference between **sel1**, **sel2**, **sel3** and dip 1 is very significant (Z = 2.62, 3.40, 2.90, $p < 0.01$). Therefore, one can conclude that the selection/concatenation system is significantly more pleasant than the diphone system. As for the overall impression, the difference between **sel1** and dip 1 is just above signification (Z= 1.68, $p < 0.1$), and the difference between **sel1** and **dip2**, and between **sel2** and **dip2** are very significant (Z=4.75, 3.45, $p < 0.01$). One can conclude that there is in general a better overall impression for the new system.

Figure 6 is displaying listening effort and comprehension problems. The new system is always preferred, but it is only

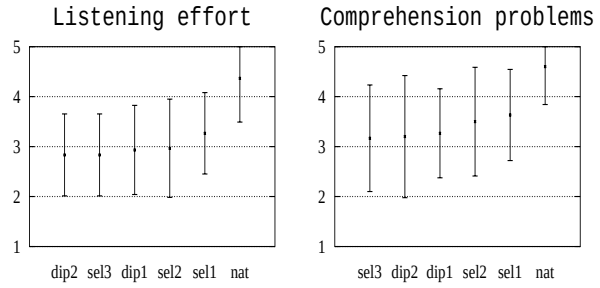

Figure 6: *Subjective evaluation: Listening Effort and Comprehension problems.*

an indication, as the results are not statistically significant (as for listening effort, the difference between **sel1** and **dip1** is just significant).

Figure 7 is displaying articulation and pronunciation. Again, the new system is always preferred, but it is only an indication, as the results are not statistically significant.

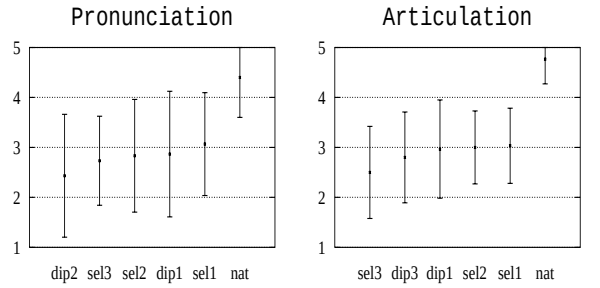

Figure 7: *Subjective evaluation: Ariculation and Pronunciation.*

5. Conclusions

As computer are improving both in power and memory, it appears feasible to process automatically large speech and text databases for building text-to-speech systems based on selection and concatenation of speech segment. This paper describes the development of such a system in French.

The main features of the system is that it describes speech at a phonologic level, using phonetic data only for pruning. Therefore, prosodic and segmental facts are taken as they are realized in actual signals, and not as they should have been realized according to some linguistic or phonetic rules. Only few global selection criteria are used, and concatenation is reduced to a simple segment edges smoothing. Thus the system is fast.

The main advantages of the system is that it allows for rapid and almost automatic development of new voices or new speaking styles, according to the speech database used. Voice quality is far better than in diphone speech. However, some prosodic inconsistencies may be found in the synthetic signal.

A formal evaluation of the selection/concatenation system compared to another system based on diphone concatenation and synthesis by rules of the prosody showed that the new system is preferred along all the evaluation categories, with significantly better results in voice pleasantness, and just significant results in overall impression. However, natural voice is always rated significantly better than synthetic speech.

Future work will be devoted to the development of new voices and speaking style, and to a better prosodic selection. This might imply the development of larger speech databases.

6. References

[1] M. Adda-Decker, L. Lamel, "Pronunciation Variants Across Systems, Languages and Speaking Style," Speech Communication, 29, pp.83-99, 1999

[2] C. d'Alessandro and P. Mertens. Automatic pitch contour stylization using a model of tonal perception. Computer Speech and Language 9, p. 257-288, 1995.

[3] M. Beutnagel, A. Conkie and A. K. Syrdal. Diphone synthesis using unit selection. The 3rd ESCA/COCOSDA Workshop on speech Synthesis, p. 185-190, Jenolan Caves, Australia, 1998.

[4] A. W. Black and P. taylor. CHATR: a generic speech synthesis system. In *Proceeding of COLING-94*, p. 983-986, Kyoto, Japan, 1994.

[5] A. W. Black and N. Campbell. Optimising selection of units from speech databases for concatenative synthesis. Eurospeech, p. 581-584, Madrid, Spain 1995.

[6] A. P. Breen and P. Jackson. Non-uniform unit selection and the similarity metric within BT's Laureate TTS system. The 3rd ESCA/COCOSDA Workshop on speech Synthesis, p. 201-206, Jenolan Caves, Australia, 1998.

[7] J.L. Gauvain, L. F. Lamel and M. Eskenazi. Design consideration and text selection for BREF, a large French read-speech corpus. ICSLP, p.1097-1100, Kobe, Japan, 1990.

[8] A. J. Hunt and A. W. Black. Unit selection in a concatenative speech syntheses system using a large speech Database. ICASSP, p. 373-376, Atlanta, Georgia, 1996.

[9] ITU-T (1993) Draft recommendation P.8S. Subjective performance assessment of the quality of speech voice output devices. Study group 12- contribution 6.

[10] Klaus, H., Fellbaum, K. & Sotscheck, J. (1997). Auditive Bestimmung und Vergleich der Sprachqualität von Sprachsynthesesystemen für die deutsche Sprache. Acta Acustica 83, 124-136.

[11] Y. Sagisaka, N. Kaiki, N. Iwahashi, and K. Mimura. ATR-ν-TALK speech synthesis system. In *Proceeding of ICSLP 92*, p. 483-486, 1992.

[12] van Bezooijen, R., van Heuven V. J. (1995)"Assessment of speech output system", EAGLES (LRE-61-100) report, 80p.

Sound examples for this paper can be found at: <http://www.limsi.fr/Individu/cda/ssw4.html>.