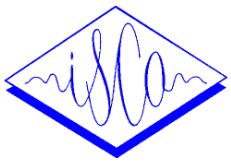

Speech Input/Output Assessment
and Speech Databases
Noordwijkerhout, the Netherlands
September 20-23, 1989

INTELLIGIBILITY TEST FOR THE ASSESSMENT OF FRENCH SYNTHESISERS USING SEMANTICALLY UNPREDICTABLE SENTENCES

C. BENOIT

Institut de la Communication Parlée
U.A. CNRS N° 368 - INPG/ENSERG - Université STENDHAL
BP 25X 38040 GRENOBLE Cedex. FRANCE

1. INTRODUCTION

Within the research field of synthesis assessment methodologies in which the ESPRIT-SAM project is involved, three tests were simultaneously defined and ran in three European laboratories, evaluating English, French and German languages [see related communications of Hazan & Grice and of Jekosh in this workshop]. Similar corpora were used in these languages, following SAM decisions on "Semantically Unpredictable Sentences" (SUS) [van ERP and Grice, 1989].
The French test here presented involved twenty SUS per syntactic structure. Hundred sentences were generated under five conditions : two coding techniques, both under two prosodic models, and natural speech. The four synthesizers used the same diphones dictionary, obtained from the voice of one speaker which represented the reference natural speech.
We were therefore able to compare the "same voice", synthesized with the same diphones concatenating method under varying aspects :
- two coding techniques, both with constant "flat" prosody ;
- a first prosodic modelling vs. constant "flat" imposed prosody (both using CNET's "PSOLA-KDG" wave-form synthesis ;
- a second prosodic modelling vs. constant "flat" imposed prosody (both using ICP's formant-coded synthesis.
Besides, to evaluate the linguistic influence of the corpus on this test methodology, we also compared listeners' answers on "handily semantized" sentences vs. randomly generated ones and on "feed-forwarded" (e.g. preknown) sentences vs. unknown ones.

Four subjects could ear at the same time the stimuli trough headphones in a sound-proof cubicle. Their task was to write down an orthographic transcription of the sentence they understood. A training session preceded the test. It included an acoustic presentation of 20 similar extra-sentences in natural speech, followed by an orthographic and acoustic presentation of 25 synthesized sentences extracted from the actual corpus test. These last stimuli served as evaluation of the "feed-forward" effect.
The hundred sentences were listened to once by five groups of four listeners during each of the five test sessions. Each session presented five sets of twenty sentences ; each set corresponding to a pair [synthesis condition / syntactic structure] in a given session for a given group. The five groups listened to the five sessions in different orders such as to allow all possible combinations for mean results computation across time, groups, syntaxes or synthesisers (see Tab. 1 & 2 below).

2. INTELLIGIBILITY AS A FUNCTION OF THE CODING TECHNIQUE

To evaluate how a coding technique may influence the intelligibility of a synthesizer, the same phonetic strings and the same "constant" flat prosody were imposed on both synthesizers : Fo = 100 Hz and phone duration = 100 ms. Fig. 1 shows the mean results obtained from 100 answers (e.g. a mixture of the five sets of syntax) by 20 listeners. Each point then corresponds to the percentage of correctly understood sentences among 400 answers for each of the five sessions. The curves represent the PSOLA-KDG processing developed at CNET [Hamon et al., 1989] and the formant coding (of the same dictionary) designed at ICP [Al Dakkak et al., 1987]. The first one used the full wave-form diphones dictionary (at Fs = 8 kHz) while the second one needed a strong information reduction and therefore a fast specialized decoder (Fs = 10 kHz).

The generally low values of observed scores are essentially due to the measurement method. For each answer, around seven semantically independant words had to be correctly written so that the sentence may be considered as correct. A percentage of correct words (strongly related to the percentage of correct sentences ; see below § VI) would actually have given higher scores. In the example on Fig. 1, mean percentages of correct sentences across time are 58.1% and 28.6%, whereas mean percentages of correct words are 88.3% and 71.6%. We, however, chose the first index as it gives more dispersed results when varying the tested conditions. This is especially the case of natural speech which, even degraded, does not give any score lower than 95% of correctly identified words.

3. INTELLIGIBILITY AS A FUNCTION OF PROSODY MODELING

Fig. 2 quantifies the "intelligibility" of the automatic prosody modeling by CNET [Sorin et al., 1987] and ICP [Bailly, 1986]. With CNET and ICP synthesisers, the fully automatic text-to-speech version (in white) is compared to the output obtained when using the "constant flat" prosody (in black) with the five syntactic structures. One point corresponds to the mean percentage across time for 20 sentences correctly identified by 20 listeners ; each group of four listeners providing results for one syntax at a different session. We can see that CNET's prosody globally gives better results than the "reference flat prosody" (4 syntaxes out of 5) whereas the ICP prosody does not increase intelligibility in relation to its "reference". Anyway, great care must be taken in the interpretation of these results as only five syntactic structures have been tested, even if they are considered as the most basic ones ; and it is impossible to predict how prosody could increase any TTS system's intelligibility (as it also depends on its acoustic quality) from a single result obtained under one acoustic condition.
It should be pinpointed that the third syntactic structure is the most frequently misunderstood. This also occurred with natural speech. This effect may be due to the obligatory position of the adverb following the initial verb in the French imperative form. The first word of a sentence is of major importance in the understanding of the entire sentence, as is the verb. This may justify the non-existence of such a phenomenon in the corresponding English or German syntactic structures. It would be of interest to verify it in Dutch which posesses the same order of units as in French. A similar observation would then claim for the rejection of the imperative form and its replacement by a somehow more frequent one in our European set of five syntactic structures.

4. SEMANTIC EFFECT ON SUS

Four sentences out of twenty have previously been "handily" given some sense in each syntactic category. This was obtained by a reorganization of the randomly extracted words from their lexicons during sentence generation. No special methodology was observed. Nor were the obtained "sense" measured. We may only say that some of the obtained sentences have more semantic content than the previous ones by combination of related words like *"chaud" / "brûle"* (hot / burn) for instance. Fig. 3 shows the percentage of correctly understood sentences among the 80 SUS (in black) and the 20 semantically "partially predictable" sentences (in white). These percentages are averaged across the scores of 5 groups for natural degraded speech at a given session and on CNET and ICP synthesizers in two specific sessions. In these latter cases, results obtained from the automatic and the flat prosody conditions are averaged. We may notice that no "semantical effect" can be observed on natural speech whereas it appears in synthetic conditions. This claims for the existence of an "intelligibility threshold" under which the comprehension of a message content is required to counterbalance the acoustic deficiency of a communication channel. And this is obvious, as meaningfull messages are always easier to understand than meaningless ones under degraded conditions ! This result anyhow confirms the efficiency of SUS in intelligibility tests as a way to emphasize differences between synthesisers or between any other tested conditions. Furthermore, the observed variations of the "intelligibility slope" across time (especially in the worst comprehension case, i.e. ICP synthesis) is due apparently to a variability in the training effect : simply speaking, meaningfull sentences are easier to memorize than SUS, and their intelligibility increases with repetition.

5. TRAINING EFFECT ON S.U.S.

25% of the synthetic SUS were primarily read and heard by listeners during the training session. Fig. 5 shows the "feed-forward" effect on words (squares) and sentences (circles) intelligibility across time. One point corresponds to mean results of 20 listeners averaged over the five synthesis conditions at a given session. The pre-read sentences are in white while the "unknown" SUS (partly "semantized" excluded) are in black. It is noticeable that this relatively small effect affects much more sentence intelligibility rather than overall word intelligibility. This means that some sentences containing, a priori, one (or at least two) uncorrect word(s) had been correctly understood when primarily presented to subjects. A very small percentage of words were corrected, however, they contributed very strongly to the correction of a high percentage of sentences by listeners.

6. INTELLIGIBILITY OF WORDS AND SENTENCES

We didn't compare here the intelligibility of sentences with their components by testing isolated words. See Jekosh [this workshop] for details on this point in German. However, it is of interest to compare the average percentages of words with those of sentences obtained under varying conditions or with different speakers.
Fig. 6 was obtained by plotting the percentages of correcly identified words vs. those of sentences. Each dot corresponds to correct answers averaged over 20 speakers x 20 sentences (x 7 words) under 25 conditions : 5 sessions x 5 synthesis conditions. Note that a given pair {session/synthesis} corresponds to the 5 syntactic structures tested by each of the five groups of listeners. The observed strong relationship between these two scoring methods confirms the fact that both of them are equivalent for presentation of results.

7. CONCLUSION

This test was a first trial to adapt the original SUS test by Nye and Gaitenby (1974) to other languages than English. These preliminary observations tend to prove its efficiency for testing overall intelligibility of synthesizers under varying conditions. From now on, such a technique will be of great interest for multilingual assessment purposes, even if further investigations are obviously necessary to evaluate how powerful it is.

ACKNOWLEDEGMENTS

This work has been supported by the European ESPRIT project N° 2589 and by the French "*Centre National de la Recherche Scientifique*".
Special thanks to Véronique RISSOAN for her fruitful help in preparing and running this test, and for typing 10000 sentences...

REFERENCES

Al Dakkak O., Murillo G., Bailly G. & Guérin B. (1987) Using contextual information in view of formant speech analysis. Recent advances in speech understanding and dialog systems, NATO-ASI Series, Bad Windsheim, RFA.
Bailly G. (1986) Multiparametric generation of French prosody from unrestricted text. IEEE-ICASSP.
Erp A. van & Grice M. (1989) Multi-lingual syntactic structures for Semantically Unpredictable Sentences. ESPRIT Project 1541 SAM Ext. Phase Final Report. London, UCL, 43-60.
Hamon C., Moulines E. & Charpentier F. (1989) Diphone synthesis system based on time-domain prosodic modification of speech. IEEE-ICASSP, Glasgow, GB.
Hazan V. & Grice M. (1989) [THIS WORKSHOP]
Jekosh U. (1989) [THIS WORKSHOP]
Nye P.W. & Gaitenby J. (1974) The intelligibility of synthetic monosyllabic words in short, syntactically normal sentences. Haskins Labs. Status Report on Speech Research, 37/38, 169-190.
Sorin C., Larreur D. & Llorca R. (1987) A rythm-based prosodic parser for text-to-speech systems in French. 11th Int. Congress of Phonetic Sciences, Taline, URSS. Vol. 1, 125-128.

Order of syntax presentation

Natural	1	2	3	4	5
Synth. 1 + pros.	2	3	4	5	1
Synth. 1 - pros.	3	4	5	1	2
Synth. 2 + pros.	4	5	1	2	3
Synth. 2 - pros.	5	1	2	3	4
	RANDOMIZATION				
Tape N°	1	2	3	4	5

Tab. 1

Content of the 5 sessions. Each synthesis condition involves the same 5 referenced sets of 20 sentences (5 "basic" syntactic rules). 5 such pairs [synthesis-syntax] provide the stimuli content of 1 tape (1 session for 1 group of 4 listeners). All recordings were randomized in the same way to avoid successive sentences belonging to the same pair synthesiser-syntax].

Order of tapes presentation

Session :	1	2	3	4	5
Group 1	1	2	3	4	5
Group 2	2	3	4	5	1
Group 3	3	4	5	1	2
Group 4	4	5	1	2	3
Group 5	5	1	2	3	4

Tab. 2

For a given session and a given group of four listeners, this table gives the tape subjects had to transcribe. Such a "latin square" permutation allows to average results across time and listeners.

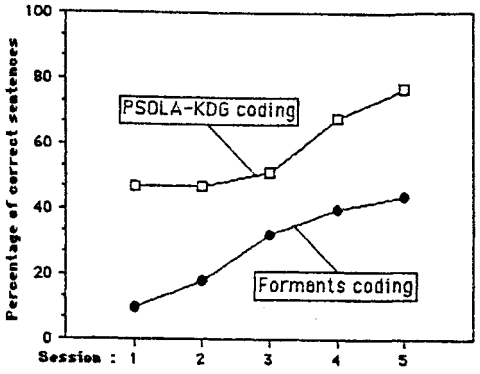

Fig. 1

"coding intelligibility"

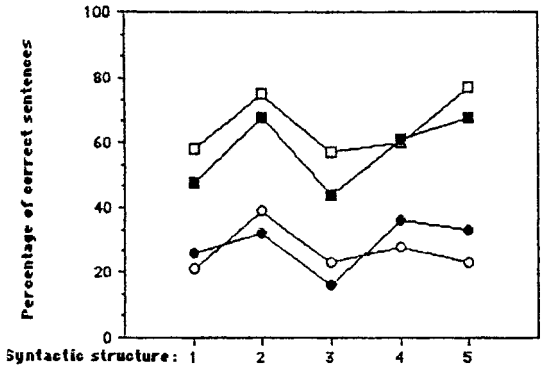

Fig. 2

"Prosody intelligibility"

Squares : CNET synthesiser | black : "constant flat" prosody
Circles : ICP synthesiser | white : automatic prosody

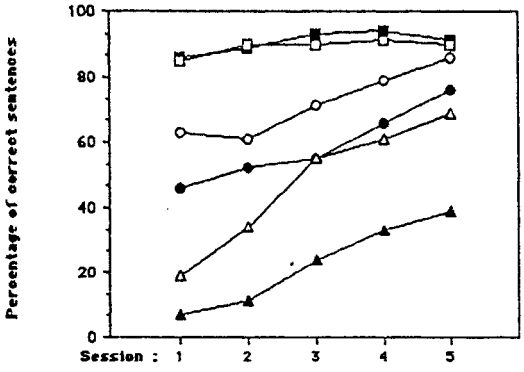

Fig. 3

"Semantic intelligibility"

Squares : Natural speech
Circles : CNET synthesiser
Triangles : ICP synthesiser
White : partly sensed sentences
black : S.U.S.

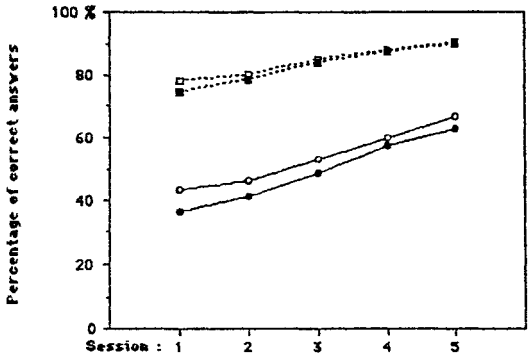

Fig. 4

"training" intelligibility

Squares : correct words
Circles : correct sentences
white : "feed-forwarded" SUS
black : unknown SUS

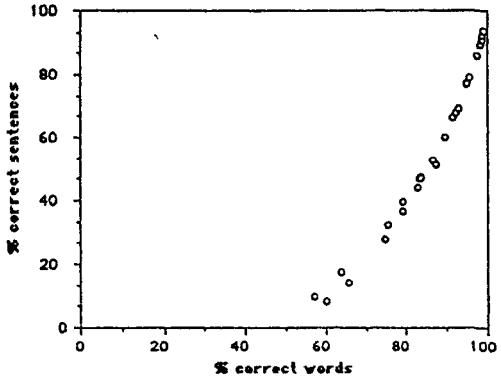

Fig. 5