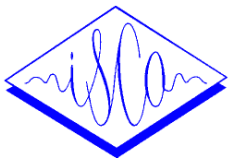

3rd European Conference on
Speech Communication and Technology
EUROSPEECH'93
Berlin, Germany, September 19-23, 1993

EVALUATION OF PROSODY IN THE FRENCH VERSION OF A MULTILINGUAL TEXT-TO-SPEECH SYNTHESIS : NEUTRALISING SEGMENTAL INFORMATION IN PRELIMINARY TESTS

Pascale Nicolas & Pascal Roméas

Laboratoire "Parole et Langage" URA CNRS 261 & Institut de Phonétique, Université de Provence
Aix-en-Provence, France

This study was carried out in the scope of a cooperation between INFOVOX AB, Sweden, and Laboratoire "Parole et Langage" URA CNRS 261, France.

ABSTRACT

Our purpose is to quantify how far the prosody of a Text-to-Speech (TTS) system is perceived from French prosodic structures. It is assumed that neutralising all the segmental information is an important methodological precaution allowing to determine whether it is legitimate to study prosodic parameters of TTS systems regardless to any segmental aspect of speech. In order to test the discriminatory power of merely pitch and pauses in the task of distinguishing between synthetic and natural speech, the spectral information of the original signal is reduced to a steady amplitude synthetic [a] which has the same length and Fo values as the original utterances. The evaluation shows that significantly different scores are assigned to natural and synthetic spectrum-reduced items. The identification of acceptable and faulty synthetic patterns produces a two mode distribution of scores. Yet the discriminatory power of prosodic features varies according to specific TTS applications.

Keywords : *Text-to-Speech, Prosody, Evaluation.*

1. PURPOSE OF THIS STUDY

In this study we propose to add two preliminary tests to the methodology of subjective evaluation of Text-to-Speech (TTS) systems prosody. The purpose of these tests is to quantify how far from French prosodic structures the synthetic items are perceived by listeners. It is assumed that neutralising all the segmental information is an important methodological precaution in order to check whether prosodic parameters of TTS systems really have to be studied regardless to any segmental aspect of speech.

It is widely accepted that prosody contributes to the quality of synthetic speech. Thus it is generally taken for granted that improving TTS prosodic rules will improve the quality of synthetic speech. However the notion of quality is analysable into three dimensions : intelligibility, naturalness, satisfaction, each of which referring primarily either to prosody or to phonetic segments /1/. Moreover, as strong correlations generally appear between each pair of these three axes, it may be difficult to know what is actually assessed in a synthesiser's prosody evaluation. Conversely, in attempts to improve TTS rules, it is not always clear whether durational or even pitch modifications providing better quality of speech must be considered as referring to either prosodic or segmental features of speech. There is not yet any accepted standard methodology for assessment tests involving naive subjects requested to evaluate *only* and *separately* the prosodic component of synthetic speech. Even the cross-synthesisers experiments carried out in the framework of SAM project /2, 3/ do not solve the problem of the strong interaction between segmental quality and prosody in each synthesiser's production /4/. For example, in many languages, assigning a weaker degree of stress to a syllable should consequently reduce the vowel timbre as well as as some consonant cues, which are generally determined by segmental rules.

If on the one hand it seems difficult to know *what* is wrong in a TTS system prosody, on the other hand it may be a first step to work out a specific quantified assessment to determine *how far* prosody is concerned in the quality degree of synthetic speech. Answering this question allows to check the following point : can we be certain that the prosodic component carries a specific responsibility as the users' general satisfaction shows up important quality problems? In other words, is it legitimate to endeavour to improve prosodic aspects of speech separately?

2. METHODOLOGY

In order to test the discriminatory power of merely pitch and pauses in the task of distinguishing between synthetic and natural speech, the spectral information of the original signal was processed so as to be reduced in both cases to a single synthetic vowel [a], thus neutralising all the segmental contrast of speech.

The experiment consisted in two tests. In test 1, listeners did not know that the original material -i.e. before being processed- contained synthesised utterances, the stimuli being just presented as "processed original speech". The task was to assess on a 1-20 gradual scale the probability that the stimulus was a *French* one. In test 2, listeners knew that the original items contained natural and synthesised utterances.

The task was to assess on a 1-20 gradual scale the probability that the stimulus was a *natural* one. Subjects were familiar with the 1-20 evaluation scale because it is commonly used during french scholarship.

The corpus was strictly identical for the natural and for the synthetic version : it consisted in utterances taken from four TTS applications (Air Traffic Control, Bank vocal service centre, Bibliographic Database vocal service centre, Text Reading). The date of the session and random presentation of items were different in test 1 and in test 2. The experiment involved 31 listeners , all french native speakers with normal hearing.

Synthetic stimuli were generated by the french version of the multilingual INFOVOX TTS, based on the rule-guided processing RULSYS program /5/ connected the INFOVOX VOX-PC development board. The board was used as RULSYS front end synthesiser. The sound output was recorded on a SONY DAT and then captured from the DAT to a Masscomp-5400 mini-computer where a program using a Klatt synthesiser allows the spectral reduction process.

Natural stimuli were read by a French native speaker. The native speaker is a professional actor who teaches French oral expression, thus a person used to act in fictive situations. His voice and the synthesiser have a very weak mean Fo difference. Recording took place in an anechoic room using a SONY DAT recorder. This record was then captured on Masscomp where it was processed the same way as the synthetic items.

In both cases, the original samples of speech stimuli were replaced by a steady amplitude synthetic [a] generated by a Klatt synthesiser.

As far as synthetic items are concerned, the Fo values files provided by the RULSYS output trace (one per 10 ms frame) were used as input files to generate the Klatt synthetic [a] with the adequate Fo pattern.

As far as natural items are concerned, the Fo values assigned to the [a] stimulus were the ones detected from the original utterances by the home developped SIGNAIX software.

The [a] stimulus had the same exact duration as the original speech item. The unvoiced part of the original stimuli (where Fo value was null) were reinserted as silent chunks into the [a] signal, with intensity transition slopes of 30 ms. Superimposing the unvoiced sections on to this [a] make it sound closer to speech than an entirely voiced continuum with interpolated Fo values does. This was in fact the only residual of segmental information. Unvoiced segments could easily be distinguished from pauses, since they have smaller durations.

Filtering was not retained as a way of neutralising the segmental information, since a part of the acoustic cues of French still remain even after a 400 Hz low-pass filtering. Moreover this method has another major drawback : as the segmental contrast is still perceived, the information about vowel length -which can be either a prosodic or a segmental information- is not suppressed. On the contrary, in our experiment the only informations that varies as a function of time is voicing (which is a poor segmental cue here) and Fo. It is clear that prosody does not involve exclusively pitch changes, but leaving the durational information brought the risk of leaving as well information on segments.

The test was run in an anechoic room using the SOAP software on an Intel 386 SX 25. The audio interface was assumed by an OROS AU 20 digital / analogic conversion board. Listeners wore high fidelity headphones AKG K240.

A short education program started the performance. An answering grid consisting in a series of twenty green horizontally displayed cells with a score typed at their center (1, ..., 20) was plot on the screen. The mouse movements allowed to place an arrow-shaped cursor into the required green cell. At the end of the session, captured data were exported to a statistics software. Our data consisted in 1798 score assignments : 29 utterances x 2 versions (natural and synthetic) x 31 subjects.

3. RESULTS

Subjects used the scale range (1-20) in a way that is very close to the normal distribution. All scores have been used. The mean value is very close to 10 and the standard deviation is about 4.3 in both parts of the test. A paired Student-t test shows that the distributions are not significantly different ($p>0.79$) from test 1 to test 2, despite the fact that the median is slightly higher in test 2. No effect of rank appears.

The results show that natural and synthetic items (yet both heard as [a]) do not receive identical scores. The standard deviation remains stable ($4<SD<4.5$), but the mean value is significantly different in either cases. The mean value shift is small : the average difference between the score assigned to any item in its natural version and the score assigned to the same item in its synthetic version is 0.97 for test 1 and 1.47 for test 2. The average score is 10.54 vs 9.57 in test 1 and 10.83 vs 9.36 in test 2. Yet a paired t-test between the natural and the synthetic items scores indicates that this small shift of the distribution is significant ($p<0.0001$) in both tests. Moreover, the natural version group has the higher mean value whereas the synthetic version group gets the lower one. In other words, the prosody of natural utterances is significantly assessed as more "french" and more "natural" than the corresponding synthetic utterances which, on the other hand, are assessed as "not french" and "synthetic".

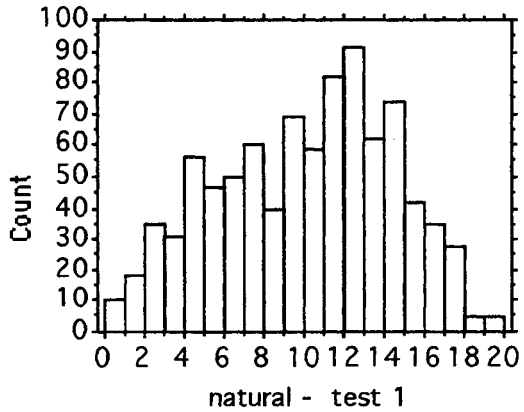

Figure 1 : Distribution of scores assigned to test 1 natural items. Test 1 : segmentally neutralised items, assessment on the basis of the French/not French criterion. Score assignment uses a 1 to 20 stepwise scale ("French" judgements closer to 20, "not French" judgement closer to 1, step value is 1)

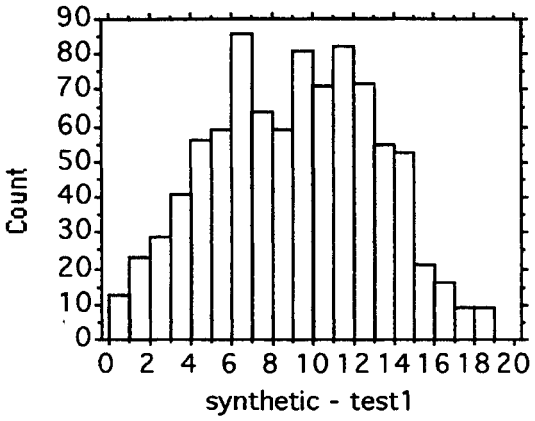

Figure 2 : Distribution of scores assigned to test 1 synthetic items.

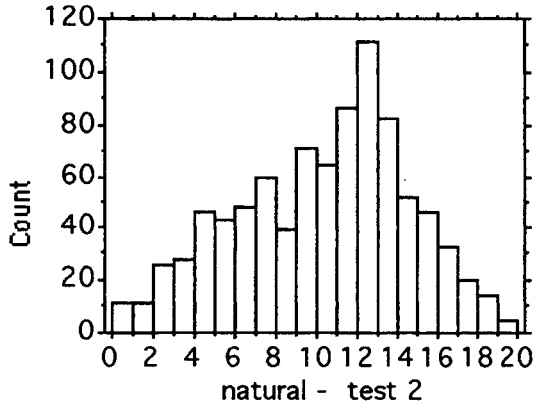

Figure 3 : Distribution of scores assigned to test 2 natural items. Test 2 : segmentally neutralised items, assessment on the basis of the Natural/Synthetic criterion. Score assignment uses a 1 to 20 stepwise scale ("Natural" judgements closer to 20, "Synthetic" judgement closer to 1, step value is 1)

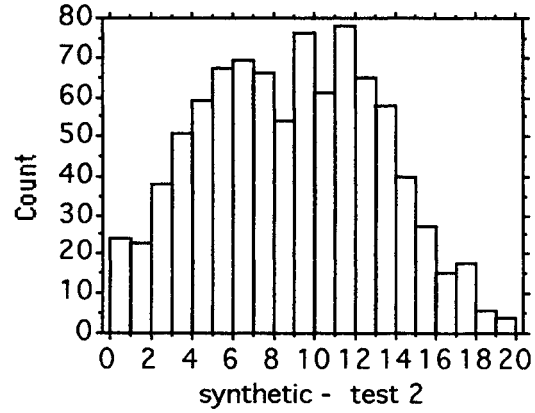

Figure 4 : Distribution of scores assigned to test 2 synthetic items.

This group-specific processing of items is confirmed by intra- and inter-application results. The average score for synthetic items never juts out above 10 except for Bank application in test 2 (10.08), and remains lower than natural items average score in all applications except for Air Traffic Control application in test 1. The difference between the average score for natural items and the average score for synthetic items varies from one application to another. In Air Traffic Control application, which consists in coded messages made of unexpected words arranged in an unexpected syntax, the score difference is very small (0.11) between natural and synthetic items, probably due to the fact that even the human reader had to produce prosodic patterns that were far from usual french utterances. On the contrary, Bibliographic Database application, which consists in a highly predictible standard sequence - namely "authors, year, title, edition", separated by commas -, leaves very little acceptable variation as far as prosodic patterns are concerned, so that the reader produced typical french Fo patterns in a very normative way, whereas the TTS which tended to assign erratic boundary tones on each proper name. These two prosodic strategies were easily discriminated in our experiment since their average score difference is 2.86 in test 1 and 2.77 in test 2. The two other applications represent an intermediate situation from the point of view of prosody predictibility : their average score intergroup difference vary from 0.5 to 1.9, always in favor of natural items.

Another observation still deserves to be mentionned. The scores distribution on the 1-20 scale tends to be *bimodal* in the case of synthetic items : histograms for the two tests and the four applications often have two peaks, generally one around score 7 and another around 13. This is particularly obvious in Bank application, which has one mode around 8 and another one above 13 in both tests, whereas this application clearly shows only one mode around 14 for its natural items scores distribution. This trend appears in all applications, even if weaker in some of them. It was checked that this was not an artefact due to the choice of the histogram interval. Otherwise this tendency seldom appears in the case of natural items. Thus it can be hypothesised that the synthetic items may actually be parted into two groups : some of them sufficiently correct to be identified as French (test 1) and not synthetic (test 2), some others showing obvious prosodic errors which lead the subjects to assign them a score around 7.

4. CONCLUSION

These results show that it is possible to make a significant discrimination between natural and synthetic items processed so as to have no segmental information available. The identification of adequate and non-adequate prosodic patterns is carried out on the basis of a French / not French judgment criterion as well as on the basis of the natural / synthetic judgment criterion. Moreover, erratic synthetic items and acceptable ones seem to have distinct distributions of scores. The discriminative power of merely pitch and pauses is sensitive to the category of TTS application : it seems to be more difficult to distinguish between natural and synthetic speech as the application

requires a high degree of specificity as far as lexicon, syntax, and punctuation are concerned. This is probably one of limits of this methodology. On the contrary, applications that involve texts presenting highly predictible prosodic patterns are liable to be assessed using this methodology. In such cases, our results show that it is legitimate to endeavour autonomous improving of the prosodic component of the TTS system.

REFERENCES

/1/ Pavlovic, C., Rossi, M., Espesser, R., 1990, "Use of magnitude estimation technique for assessing the performance of text-to-speech synthesis system", *Journal of the Acoustical Society of America*, 87, 373-382.

/2/ Benoît, C., Emerard, F., Schnabel, B., Tseva, A., 1991, "Quality comparisions of prosodic and of acoustic components of various synthesisers", *Proceedings of the Eurospeech Conference*, Genova, 2, 875-878.

/3/ Grice, M., Vagges, K., Hirst, D., 1991, "Assessment of intonation in a text-to-speech synthesis system - a pilot test in English and Italian", *Proceedings of the Eurospeech Conference*, Genova, 2, 879-882.

/4/ Santi, S., 1992, "Méthodes d'évaluation subjective de la composante prosodique en synthèse vocale", *Aix Seminar on Prosody, 1992 October 20-21*, 36-46 (Prepublication available).

/5/ Carlson, R., Granström, B., 1990, "An environment for multilingual text-to-speech development", *Proceedings of the ETRW on Speech Synthesis*, 2, 73-82.