# Objective vs. Subjective Evaluation of Synthetic Prosody

**Rilliard Albert**
Institut de la Communication Parlée, Grenoble, France
E-mail: rilliard@icp.inpg.fr

## ABSTRACT

This paper present an experiment designed to test the relative efficiency of two perception methods of synthetic prosody evaluation, compared to the influence of the evaluated stimuli nature. After the description of the possible paradigm, we present the analysis of a perception test and then a comparison of these subjective results to acoustic measurement of the distance between synthetic and natural prosodies.

## 1. INTRODUCTION

Speech prosody is largely described, for various languages (see e.g. [4]). Speech synthesisers are now able to generate quite high quality prosody, adapted to more and more different situations (see [9]). Nevertheless, in order to improve specially the naturalness of synthetic prosody in a given situation, and to go farther in prosody modelling (e.g. study attitudes, emotions, etc.), it is essential to still develop evaluation procedures to check and diagnose the systems, and to analyse how prosodic information is encoded. A major problem with subjective evaluation procedure is that it is a time-consuming process. In order to prevent this default, objective measures of the efficiency of prosodic parameters for a specific task have been proposed and need to be developed further.

With the aim of refining such objective measures of the adequacy of prosody to the segmentation and hierarchisation of speech, we matched the output of subjective tests with objective distances between synthetic and natural prosodic parameters (fundamental frequency, duration and intensity). Two perception tests have already been carried out.

The first one ([7]) proposed passages of speech (of 5 sentences each, extracted from the EUROM1 corpus [2]) to listeners, who had to perform an online evaluation of the prosody of each word, and give also a global rating of each passages. This test is based on a previous experiment performed by [5], with a similar procedure. Then the perceptive results were compared to objective distances between the synthetic stimuli and the natural sentences contained in the EUROM1 corpus.

The second experiment [11] used a set of short delexicalised sentences that listeners had to associate with text, directly matching the pure prosodic content of delexicalised item to the text. The prosody of the sentence used was either natural or synthetic. Then the subjective association scores between pure prosody and text (supposed to reflect the distance between the synthetic prosody and supposed ideal prosody), were compared to the objective distance between natural and synthetic prosody parameters.

Then, for each of the two subjective evaluations subjective scores were compared to the objective distances between (i) the prosodic parameters of synthetic words and (ii) the prosodic parameters of natural words used as references. From this comparison, we aimed at rating the efficiency of objective measures to predict perceptive distances.

But, the comparison of these two experiments [10] shows that subjective results do not fit objective ones with the same efficiency, according to (1) the task proposed to listeners and (2) the length of the stimuli used. The online evaluation paradigm, using long passages of speech, did not fit objective measures with a good efficiency, whereas the association experiment results, based on short sentences, are very close to F0 and duration distances. Therefore, in order to refine objective measures and explore the relative importance of the experimental paradigm and the speech stimuli length, we performed a new perception test, proposed here. This test is based on a similar paradigm as the first experiment, but with the stimuli used in the second one, in order to study the relative influence of the paradigm and of the stimuli length.

We present hereafter the results of the perception test and of the acoustic analysis of the sentences, and then a comparison of both subjective and objective distances.

## 2. EVALUATION EXPERIMENT

### 2.1. SUBJECTS

13 listeners from 18 to 38 years old (mean=23), with no experience of synthetic speech listening served as subject for the experiment. They are either student in linguistic, or working at the lab, and aware of what is prosody.

### 2.2. STIMULUS

22 sentences, from 5 to 11-syllabe lengths, were used as stimuli for the experiment. These sentences are the same as the ones used in [11]. Each sentence is presented both in its natural version and in its synthetic version.


ISBN 1-876346-48-5 

| Length | Sentence | Relevant syntactic structure variation (syllabic length) |
|---|---|---|
| **5** | 1- Ce passant chantait. | Nominal Group (NG) (3) - Verb (V) (2) |
| | 2- Ce pas sera fait. | NG (2) - V (3) |
| **6** | 3- Ce beau passant chantait. | NG (4) - V (2): Adj. (1) / Noun (2) inversion |
| | 4- Ce passant fou chantait. | |
| | 5- On entendait des pas. | Subject (1) - Verb (3) - NG (2) |
| **7** | 6- Ce petit passant chantait. | NG (5) V (2): Adj. (2) / Noun (2) inversion |
| | 7- Ce passant tout fou chantait. | |
| | 8- Son pas doux retentissait. | NG (3) + V (4) |
| **8** | 9- Tu dis que ce passant chantait. | 2 clauses (2 + 6) |
| | 10- Ce passant chantait l'opéra. | NG (3) - V (2) - NG (2) |
| | 11- Je verrai si les enfants jouent. | 2 clauses (3 + 5) |
| **9** | 12- Ce passant chantait tous les six mois. | NG (3) - V (2) - object.(4) |
| | 13- Ce passant chantait, Toto dansait. | 2 independent clauses (5 + 4) |
| | 14- On entendait plus ou moins les pas. | Subject (1) – V (3) - Adv. (3) – NG (2) |
| **10** | 15- L'enfant pleurait quand il était malade. | 2 clauses (4 + 6) |
| | 16- Quand l'enfant pleurait, il était malade. | 2 independent clauses (5 + 5) |
| | 17- Quand il pleurait, l'enfant était malade. | 2 independent clauses (4 + 6) |
| | 18- Ce passant chantait quand Toto dansait. | 2 clauses (5 + 5) |
| **11** | 19- Je vois le marchand de poissons de Paris. | Subject (1) - V (1) - long NG (9) |
| | 20- Ce passant chantait parce qu'il était content. | 2 clauses (5 + 6) |
| | 21- Je mangeais du vin, du Boursin, et du pain. | Variation of the enumeration's length (3 + 2 + 3 + 3) |
| | 22- Même si les enfants jouent, je verrai le chat. | 2 clauses (6 + 5) |

**Table 1:** the corpus of 22 French sentences used in the test, with the syntactic variations proposed to listeners.

As we are only interested in the evaluation of prosody alone, and to avoid an obvious preference score for natural sentences, both stimuli are constructed by using a high-quality TDPSOLA analysis-synthesis of the prosodic parameters (either extracted from the natural signal or from the output of the TTS system) on natural carrier sentences produced with a monotonous intonation (isochronous and with a flat melody).

The sentences are organised by length. Each group of same length sentences presents a set of syntactic oppositions, in term of group length, group nature, or group level (see table 1). These syntactic variations are intended to test the ability of the synthesiser to predict prosodic parameters adequate to the segmentation and hierarchisation function corresponding to each sentence.

### 2.3. PROCEDURE

Perception tests are made on a computer. All sentences are presented once to each subject, in a randomised order, different for each time. The sentences are presented to listener via two modalities: oral and written. The complete text of each sentence is displayed on the screen using a grey font. Subjects have to listen to the sentence and at the same time to follow the text written on the screen. During the display, each word of the text is highlighted using a black font, in time with the recording. This type of dynamic display is somewhat similar to that used in "Karaoke" singing. Subjects are asked to mark parts of the text when they are not satisfied with the prosody by clicking on the different words with the computer mouse. When a word has been clicked on, it is highlighted using a red font. Clicking on words before they are pronounced has no effect on the display. Subjects are, however, allowed the possibility of changing their minds by clicking on words, which they had previously classified as bad, changing the display back to black and to continue to select word after the sentence is over. All the actions of the subjects are recorded on the computer together with the time of the action, which could thus be synchronised with the synthetic recording. At the end of each passage subjects are asked to give a global score reflecting their satisfaction with the way the passage had been read. For French subjects a score out of 20 is a familiar scale used for marking homework and exercises.

### 2.4. RESULTS

Subjective results are expressed either in a local manner (the underlined words) or in a global one (the global scores given for each sentence, and the percentage of underlined words for each sentence).

Local scores are intended to be used for diagnostic purposes, for the model designer. They will not be detailed here.

The analysis of results is performed both on global scores, and on the percentage of underlined word per sentence. As sentences are organised in groups according to their syllabic length, results are analysed separately for each sentence length, from 5 to 11 syllables. These analyses are based on a $S_{13}*P_2*A_X$ experimental design, where S stands for the 13 subjects, P for the two prosodies proposed (natural and synthetic), and A for the X sentences, where:

- X=2 for the two sentences of the 5-syllable length group,
- X=3 for the three sentences of the 6, 7, 8 and 9-syllable length groups,

ISBN 1-876346-48-5  

• X=4 for the four sentences of the 10 and 11-syllable length groups.

| Stimuli length | r | p<0,01 |
|---|---|---|
| 5 | -0,92 | * |
| 6 | -0,98 | * |
| 7 | -0,89 | * |
| 8 | -0,87 | * |
| 9 | -0,97 | * |
| 10 | -0,98 | * |
| 11 | -0,93 | * |

**Table 2:** Correlations between the global scores and the percentage of underlined words. All results are significant with p>0.01.

The first comparison made is the correlation between the global score and the percentage of underlined words: all groups of stimuli length receive of very high score of negative correlation, meaning that the sentence that receive the higher percentage of underlined words have also the worst global scores. This first results confirmed those of [5] and [7], and show s the coherence of listeners' answers.

Then an ANOVA is performed, to study the relative influence of the sentence structure and of the nature of the prosody on subjective results. Results are summarised in the table 3.

The main result of the analysis of variance is the relative good performance of synthetic prosody in comparison to natural prosody: only three groups out of seven receive significantly better score for the natural prosody than for the synthetic one, both for global scores than for the percentage of underlined words. That is:

• the longest sentences (groups of 10 and 11-sylable length), systematically better for the natural;

• for the 8-syllable length group, the percentage of underlined words is higher for synthetic prosody than for natural one (the local analysis shows a greater number of underlined words for synthetic sentences, and each selected word are more underlined for the synthetic version than fir the natural one);

• for the 6-syllable length group, global scores are lower for synthetic speech, and one of the three sentences (sentence #5 on the table 1) receive significantly higher scores than the two others. The syntactic structure of this short sentence (subject / verb / object) is particularly common, and there is no problem for the synthesiser, and for the speaker to produce very high quality prosody for it.

## 3. ACOUSTIC ANALYSIS

The acoustic parameters of prosody (F0, duration and intensity) are extracted from each sentence. Fundamental frequency is measured in Hz, duration correspond to the syllabic duration and to the duration of each Inter Perceptual-Centre Group (IPCG - See [8] and [12] for P-centres, and [1], for IPCG), and intensity in dB.

| Global scores perceptive results | | | | | | | Underlined words percentages perceptive results | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Syll. | Factor | SS | df | MS | F | p | Syll. | Factor | SS | df | MS | F | p |
| 5 | A | 3,25 | 1 | 3,25 | 0,6724 | 0,43 | 5 | A | 0,0192 | 1 | 0,019231 | 1,2203 | 0,29 |
| | P | 5,56 | 1 | 5,56 | 0,7875 | 0,39 | | P | 0,0192 | 1 | 0,019231 | 2,9589 | 0,11 |
| | A*P | 0,17 | 1 | 0,17 | 0,0364 | 0,85 | | A*P | 0,0021 | 1 | 0,002137 | 0,1702 | 0,69 |
| 6 | A | 45,72 | 2 | 22,86 | 8,952 | 0,001 | 6 | A | 0,0833 | 2 | 0,041667 | 4,6956 | 0,02 |
| | P | 25,96 | 1 | 25,96 | 13,823 | 0,003 | | P | 0,0513 | 1 | 0,051282 | 7,8545 | 0,02 |
| | A*P | 5,62 | 2 | 2,81 | 0,483 | 0,62 | | A*P | 0,0278 | 2 | 0,013889 | 1,8947 | 0,17 |
| 7 | A | 9,56 | 2 | 4,78 | 0,732 | 0,49 | 7 | A | 0,0192 | 2 | 0,009615 | 0,6750 | 0,52 |
| | P | 17,55 | 1 | 17,55 | 5,408 | 0,04 | | P | 0,0602 | 1 | 0,060185 | 6,5000 | 0,03 |
| | A*P | 0,33 | 2 | 0,17 | 0,024 | 0,98 | | A*P | 0,0050 | 2 | 0,002493 | 0,1522 | 0,86 |
| 8 | A | 57,77 | 2 | 28,88 | 3,797 | 0,04 | 8 | A | 0,0833 | 2 | 0,041667 | 3,0000 | 0,07 |
| | P | 27,13 | 1 | 27,13 | 8,909 | 0,011 | | P | 0,1154 | 1 | 0,115385 | 14,1898 | 0,003 |
| | A*P | 29,72 | 2 | 14,86 | 1,434 | 0,26 | | A*P | 0,0278 | 2 | 0,013889 | 0,7826 | 0,47 |
| 9 | A | 14,03 | 2 | 7,01 | 1,130 | 0,34 | 9 | A | 0,0178 | 2 | 0,008903 | 0,3906 | 0,68 |
| | P | 57,55 | 1 | 57,55 | 7,764 | 0,02 | | P | 0,0915 | 1 | 0,091168 | 4,1967 | 0,06 |
| | A*P | 24,95 | 2 | 12,47 | 3,183 | 0,06 | | A*P | 0,0691 | 2 | 0,034544 | 1,5039 | 0,24 |
| 10 | A | 51,15 | 3 | 17,05 | 3,163 | 0,04 | 10 | A | 0,1205 | 3 | 0,040153 | 2,4444 | 0,08 |
| | P | 108,04 | 1 | 108,04 | 10,396 | 0,007 | | P | 0,1947 | 1 | 0,194712 | 10,056 | 0,008 |
| | A*P | 22,04 | 3 | 7,35 | 1,083 | 0,37 | | A*P | 0,0329 | 3 | 0,010951 | 0,5928 | 0,62 |
| 11 | A | 20,18 | 3 | 6,73 | 0,535 | 0,66 | 11 | A | 0,0457 | 3 | 0,015224 | 0,2224 | 0,88 |
| | P | 274,62 | 1 | 274,62 | 21,091 | 0,0006 | | P | 1,1285 | 1 | 1,128472 | 13,7324 | 0,003 |
| | A*P | 94,26 | 3 | 31,42 | 6,202 | 0,002 | | A*P | 0,1611 | 3 | 0,053686 | 2,35241 | 0,09 |

**Table 3:** Results of the analysis of variance for both the global scores and the percentages of underlined words per sentences. Results are given for each groups of stimuli length, from 5 to 11 syllables.


ISBN 1-876346-48-5 

In order to fit the perceptual results, we applied some normalisation on this data: F0 was first normalised to a 100Hz starting point, and then expressed in MEL, BARK, and ERB (these normalisations are based on the works of [3] and [6]). Intensity was also normalised to a same mean. Acoustic parameters are calculated only for the vocalic parts of the sentences. Three values of fundamental frequency and intensity are use for each vowel.

Then the objective distances between natural and synthetic prosodies are calculated as:

• The correlation between the natural and synthetic prosodies' acoustic parameters

• The root-mean-square distance between the natural and synthetic prosodies' acoustic parameters.

These distances are calculated for each acoustic parameter, and for each sentence.

| | % underlined words | Global score |
|---|---|---|
| **F0 Norm RMS** | -0,30 | 0,25 |
| **F0 Norm Cor.** | 0,19 | -0,21 |
| **F0 MEL RMS** | -0,31 | 0,26 |
| **F0 MEL Cor.** | 0,19 | -0,22 |
| **F0 BARK RMS** | -0,30 | 0,25 |
| **F0 BARK Cor.** | 0,19 | -0,21 |
| **F0 ERB RMS** | -0,32 | 0,28 |
| **F0 ERB Cor.** | 0,20 | -0,23 |
| **Syllab. duration RMS** | -0,06 | -0,06 |
| **Syllab. duration Cor.** | -0,10 | 0,18 |
| **IPCG duration RMS** | -0,20 | 0,18 |
| **IPCG duration Cor.** | 0,11 | -0,07 |
| **RMS IntNorm** | -0,11 | -0,05 |
| **Norm. Intensity Cor.** | 0,17 | -0,12 |

**Table 4:** Correlations between (1) objectives measures (F0 normalised, and expressed in MEL, BARK and ERB; durations of syllables and IPCG; and normalised intensity), and (2) Subjective scores given to synthetic prosody.

## 4. COMPARISON & CONCLUSIONS

Then objective distances are compared to subjective scores obtained for synthetic stimuli, in order to examine if our objective metrics can fit the perceptual results. In this aim, the correlation between each objective distance (root-mean-square and correlation distance between each acoustic parameters) and subjective results (global scores and percentage of underlined words for synthetic prosody) are calculated (see table 4) for all the 22 sentences produced with a synthetic prosody.

This comparison gives no significant correlation between objective and subjective measures. As in [10], online evaluation of the prosody, even though it gives very useful result for diagnostic purpose on the location of the worst prosodic passages, did not fit objective distances. Conversely, a comparison paradigm [11] with the same sentences gives high correlation with the same acoustic distances. Therefore, the most important factor between the two experimental paradigms seems to be the method used to evaluate the prosody, and not the length or the complexity of the stimuli.

## ACKNOWLEDGEMENTS

We are deeply grateful to Daniel Hirst and Véronique Aubergé for their fruitful advices on the theoretical problems raised by those experiments.

## REFERENCES

[1] Barbosa P. & Bailly G. "Characterisation of rhythmic patterns for text-to-speech synthesis". *Speech Communication*, 15:127-137, 1994.

[2] Chan, D.; Fourcin, A.; Gibbon, D.: Grandstrom, B.; Huckvale, M.; Kokkinakis, G.; Kvale, K.; Lamel, L.; Lindberg, B.; Moreno, A.; Mouropoulos, J.; Senia, F.; Trancoso, I.; in 't Veld, C. "Eurom – a spoken language resource for the EU". Proc. *ESCA Eurospeech '95*, 867-870, 1995.

[3] Hermes D.J. & Van Gestel J.C. "The frequency scale of speech intonation". *Journal of the Acoustical Society of America*, 90 (1), 97-102, 1991.

[4] Hirst D. and Di Cristo A.. *Intonation systems: A survey of twenty languages*. Cambridge University Press, 1998.

[5] Hirst, D.J.; Nicolas, P.; Espesser, R. "Coding the F0 of a continous text in French: an experimental approach". *Proceedings of the XIIth International Congress of Phonetic Sciences*. Aix en Provence 1991, 5: 234-237, 1991.

[6] Hirst, D.J. & Nishinuma, Y. "Automatic scaling for speaker-independent representation of prosody". *Proceedings of the JAPIC First Research Meeting*, National Language Institute, Tokyo, Japan, 1995.

[7] Hirst, D. J., Rilliard, A., and Auberge, V. "Comparison of subjective evaluation and an objective evaluation metric for prosody in text-to-speech synthesis". *Proceeding 3rd Workshop on Speech Synthesis*, Jenolan Caves, Jenolan Caves, Australia, 1-4, 1998.

[8] Marcus S.M.. *Perceptual Centres*. Ph.D. Thesis, Cambridge University, UK, 1976.

[9] Proceedings of the 4th ISCA Tutorial and Research Workshop on Speech Synthesis, Atholl, Scotland, 2001.

[10] Rilliard A. *Vers une mesure de l'intelligibilite linguistique de la prosodie – evaluation diagnostique des prosodies synthetique et naturelle*. PhD Thesis, Institut National Polytechnique de Grenoble, France, 2000.

[11] Rilliard, A. and Auberge, V. "Prosody evaluation as a diagnostic process: subjective vs. objective measurements". *4th ISCA Workshop on Speech Synthesis*, Atholl, Scotland, 2001.

[12] Scott S.K.. *P-Centres in speech – an acoustic analysis*. PhD thesis, University College, London, UK, 1993.