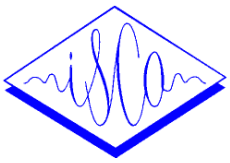

The Third ESCA/COCOSDA Workshop
(ETRW) on Speech Synthesis
Jenolan Caves House, Blue Mountains,
NSW, Australia, November 26–29, 1998

Comparison of subjective evaluation and an objective evaluation metric for prosody in text-to-speech synthesis.

Daniel Hirst, Albert. Rilliard** & Véronique Aubergé**.*

*CNRS LPL, Université de Provence, Aix-en-Provence

**CNRS ICP, Université Stendhal, Grenoble

ABSTRACT

An experimental technique is described for eliciting a subjective evaluation of the prosody of synthetic speech by untrained listeners. The technique makes use of a graphic display time-aligned with the speech signal. Subjects are asked to indicate which parts of a recording are unsatisfactory by clicking on a computer screen with a mouse. The technique was applied to two TTS systems for French. Results obtained using this technique are to be compared with those obtained using an objective evaluation metric for prosodic characteristics, comparing the synthetic versions with a number of different readings by human speakers.

1. INTRODUCTION AND BACKGROUND

The evaluation of the prosody of synthetic speech has been the object of much investigation (see [12], [13], [4] for surveys) but has not yet found very satisfactory solutions. Evaluation techniques in general fall into two categories: subjective techniques, requiring the use of human experts, and objective techniques, which do not. Although objective techniques will no doubt in the long run provide the most efficient type of evaluation they themselves need to be evaluated with respect to subjective techniques. Both subjective and objective techniques can be used to provide either global or local (diagnostic) evaluation of synthetic speech.

1.1 Subjective evaluation

Although subjective evaluation techniques all make use of human experts, the expertise required varies considerably from one technique to another, ranging from untrained native speakers of the language at one extreme to fully trained phoneticians at the other extreme. Robust procedures using untrained listeners would of course be extremely useful since besides the extra cost involved in using trained phoneticians, there is also the danger of experts being influenced by theoretical biases, particularly in such a controversial domain as that of prosody.

Hirst et al. [10] asked untrained subjects to underline unsatisfactory portions of a written text while listening to a version of which the prosody had been modified by re-synthesis. They found that the actual length of underlining (in cm.) was quite highly correlated with the listener's global evaluation of the naturalness of the re-synthesised passage. The underlining had the advantage over the global estimation that it provided a diagnostic evaluation of the portions of the passage where the prosody was judged inadequate.

In this paper we present a technique for eliciting a quantified subjective diagnostic evaluation of synthetic speech from untrained listeners. Results using this technique are to be compared with results from objective analyses of prosodic characteristics carried out using the same material.

1.2 Objective evaluation.

Objective evaluation techniques for the prosody of synthetic speech can be based either on explicit knowledge or on a comparison with a reference utterance. The weakness of such techniques lies obviously in the fallibility of the knowledge or in the potential lack of representativity of the reference utterance.

In this paper we present a corpus-based objective evaluation technique to be applied to text-to-speech synthesis. The basic principle of this technique is to compare the synthetic signal not to one recording but to several. For each prosodic parameter we take the root-mean-square distance between the value measured for the synthetic version and the most similar version of the natural recordings as a dissimilarity rating between the natural and synthetic versions. Separate scores are obtained for segmental duration, fundamental frequency and intensity of the output of the TTS module.

The technique provides both a global and a local evaluation of the synthetic speech. The same technique could be used at various different levels of granularity.

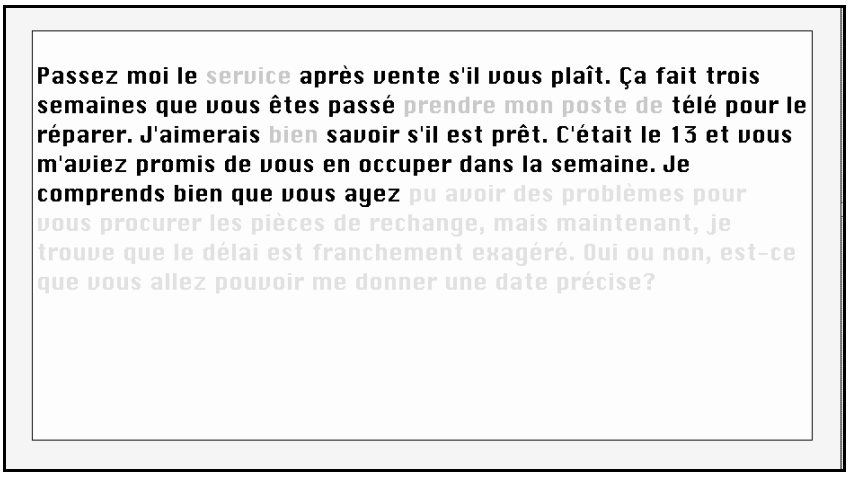

Figure 1. Sample screen displaying one of the texts. The black font corresponds to the text which the listener has already heard, the grey font to that which he has not yet heard. The red font (which appears as a different shade of grey here) corresponds to words which the listener has selected as having unsatisfactory prosody (see text).

2. MATERIAL AND METHODS

2.1 Subjective evaluation

The complete text of a passage is displayed on a computer screen using a grey font. Subjects are asked to listen to a synthesised version of the text and at the same time to follow the written text on the screen. It was expected that the dynamic presentation of the written text while the recording was being listened to would minimise the influence of the individual speech segments by removing the need to decode the individual words. As we note below, this expectation was only partially justified.

Information on the duration of different units of the written text was provided to the computer program in order to synchronise the visual display with the recording. For this experiment, basically for simplicity, the word was taken as the unit of display. The same technique could, however, equally well be used with other sized units such as the syllable.

During the display, directly each unit (here each word) of the text is heard, it is highlighted using a black font, in time with the recording, with the result that if the subjects take no action, at the end of the recording the complete text is displayed in black. This type of dynamic display is somewhat similar to that used in "Karaoke" singing, with which several of our subjects were familiar. Subjects are asked to mark parts of the text when they are not satisfied with the synthetic speech by clicking on the different words with a mouse. The term prosody was not used since we did not wish listeners to make meta-linguistic judgments. Listeners were, however, asked to pay more attention to the way in which the sentences were pronounced rather than listening to the individual speech sounds. 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.

All the actions by 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.

Passages of 5 semantically linked sentences taken from the French version of the Eurom1 multilingual corpus [3] were used for the test. 20 passages were synthesised using the two TTS systems developped in our two institutes. These TTS systems are henceforth referred to as (not necessarily respectively!) system A and system B. Information on the duration of each word was obtained directly from the synthesisers and used to synchronise the graphic display with the speech signal. When this information is not directly available we should need to rely on automatic alignment techniques.

A total of 12 subjects took part in the listening tests. Speech signals presented to listeners were recorded as AIFF sounds at a frequency rate of respectively 16 kHz for synthesiser A and 10 kHz for B. They were presented via headphones at a comfortable level of hearing. Each subject heard a random ordering of the 20 passages produced by one of the two synthesisers and then after a short pause the other 20 simuli synthesised by the other synthesiser. One half of the subjects heard synthesiser A first and then synthesiser B, the other half of the subjects heard the synthesisers in reverse order. At

the end of each passage a global score out of 20 was given by the subjects.

2.2 Objective evaluation

For the experiment we used the same French passages as for the subjective evaluation. For each of the languages in the EUROM1 corpus, a number of readings by different speakers are available for 40 passages of 5 semantically linked sentences. Each recording for four of the languages has been hand-labelled for word boundaries and pauses and the F0 curves for the recordings have been stylised using the MOMEL algorithm [9] with manual corrections.

Word durations and fundamental frequency values of the synthetic speech were obtained directly from the synthesiser output but as mentioned above could be obtained by automatic analysis of the natural recordings. Fundamental frequency curves of the human recordings were stylised using the MOMEL algorithm in order to remove small-scale inaudible differences between pitch-contours, which are likely to interfere with dissimilarity ratings [7], [8]. Fundamental frequency and intensity comparisons were time-adjusted relative to the duration of each word.

3. RESULTS

An analysis of variance on the scores obtained by the two synthesis systems in the subjective evalutaion showed a systematic difference between the two systems whether scored on a global rating or by a count of the number of words highlighted during the "Karaoke" presentation. The result of Hirst et al. [10] was confirmed - there was a significant correlation between the number of words highlighted during the Karaoke test and the global score given for each passage ($R = 0.768$, $p < 0.001$).

The order of presentation was not significant as a simple effect ($F(1; 5152) < 1$) but there was a highly significant interaction ($F(18; 5152) = 2.589$, $p < 0.0003$) between the order of presentation and the system. When system A was presented first, its score was not significantly different from that of system B; when system A was presented second it scored significantly worse than system A, see Figure 2. There was some indication from the listeners' spontaneous commentaries on the test that part of this effect might be due to the better segmental quality of system B. The difference needs however to be interpreted with caution since the subjects who heard the systems in order AB were not the same as those who heard the systems in order BA.

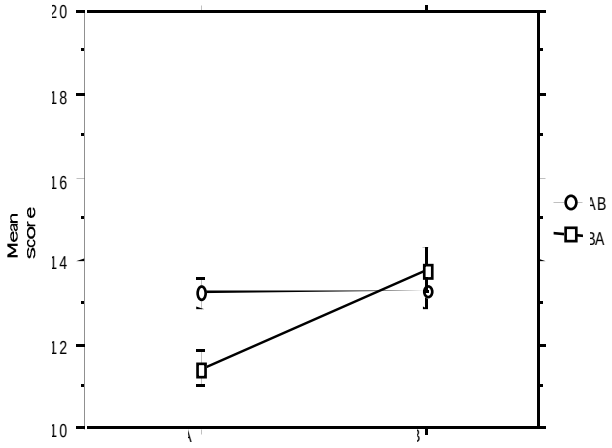

Figure 2. Mean scores for synthesisers A and B with AB order (circles) and BA order (squares).

The prosodic characteristics of the synthetic recordings are to be compared as mentioned above with a number of different natural recordings. These analyses were not yet completed when the text for this paper was submitted. More complete results will be presented during the oral presentation and in future publications.

As a distance metric, the durations of each word were compared both directly and after normalisation with respect to average speaking rate. It is hoped that this double analysis will allow us to factor out speaking rate as a separate criterion. Fundamental frequency values were normalised with respect to the diffent speaker's mean values. Different frequency scales (Mel, Bark, ERB) are being tested in addition to raw Hz values. For each word of the different passages, a score is obtained consisting of the squared difference between the duration, mean F0 or intensity value and the closest value among the available natural recordings. The mean and sum of squared errors is then computed for each passage for each of the two synthesisers.

4. DISCUSSION AND CONCLUSIONS

Although we are only able to report on very preliminary results at this stage we feel that the combination of objective and subjective analyses described here provides a very useful paradigm for the diagnostic evaluation of the prosody of speech synthesis.

The subjective analysis technique makes it possible to obtain from untrained listeners with fairly little human intervention a detailed diagnostic evaluation of the output of a synthesiser which we have shown to be quite highly correlated with global assessments.

The technique of comparing synthetic speech to several different natural human readings will, we hope, provide a further diagnostic tool which avoids the pitfalls of expert

evaluation on the one hand and that of a single version of an utterance on the other.

Acknowledgements

The authors thank AUPELF and COST 258 for supporting this research and they thank Gérard Bailly and Philippe Di Cristo for supplying us with the synthetic versions of the passages.

5. REFERENCES

[1]. Aubergé, V. 1992. Developing a structured lexicon for synthesis of prosody. In G. Bailly and C. Benoît, editors, *Talking Machines: Theories, Models and Designs*, Elsevier B.V 307-321.

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

[3]. 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. 1995. Eurom – a spoken language resource for the EU. Proc. *ESCA Eurospeech '95*, 867-870.

[4]. Di Cristo, A., Di Cristo, P., Campione, E & Véronis, J. (forthcoming.) A prosodic model for text-to-speech synthesis in French. In A. Botinis (ed.) forthcoming.

[5]. Gibbon, D.; Moore, R.; Winski, R.. 1997. *Handbook of Standards and Resources for Spoken Language Systems*. Berlin, De Gruyter.

[6]. Grice, M.; Vagges K.; Hirst D.J. 1991. Assessment of intonation in text-to-speech synthesis systems - A pilot test in English and Italian. *Proc. Eurospeech* '91, 2, 879-882, Genova.

[7]. Hermès, D.J. 1998a. Auditory and visual similarity of pitch contours. *Journal of Speech, Language and Hearing Research* 41, 63-72.

[8]. Hermès, D.J. 1998b. Measuring the perceptual similarity of pitch contours. *Journal of Speech, Language and Hearing Research* 41, 73-82.

[9]. Hirst, D.J., Di Cristo, A & Espesser, R. forthcoming. Levels of representation and levels of analysis for the description of intonation systems. In M. Horne (ed.) *Prosody: Theory and Experiment*. Kluwer Academic Press, Berlin.

[10].Hirst, D.J.; Nicolas, P.; Espesser, R. 1991. 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.

[11].Morlec, Y., Bailly, G., Aubergé. V. 1996. Generating intonation by superposing gestures. In Proceedings of the International Conference on Speech and Language Processing, (Philadelphia) 1: 283-286

[12].Morton, K. 1991. Expectations for assessment techniques applied to speech synthesis. *Proc. Institute of Acoustics*, 13 (2)

[13].Pols, L. and Sam-partners 1992. Multi-lingual synthesis evaluation methods. *Proc. ICSLP* '92. Banff, 1 181-184.