Perception and Analysis of a Reiterant Speech Paradigm: a Functional Diagnostic of Synthetic Prosody

Albert Rilliard & Véronique Aubergé

Institut de la Communication Parlée 1180, Av. Centrale - 38040 Grenoble Cedex 9 {rilliard, auberge}@icp.inpg.fr

Abstract

A set of perception experiments, using reiterant speech, were designed to carry out a diagnostic of the segmentation / hierarchisation linguistic function of prosody. The prosodic parameters of F0, syllabic duration and intensity of the stimuli used during this experiment were extracted. Several dissimilarity measures (Correlation, root-mean-square distance and mutual information) were used to match the results of the subjective experiment. This comparison of the listeners' perception with acoustic parameters is intended to underline the acoustic keys used by listeners to judge the adequacy of prosody to perform a given linguistic function.

Introduction

The more the quality of TTS systems increases, the more synthetic speech evaluation becomes useful, both for assessment and diagnostic purposes. The better the segmental quality of synthetic speech is, the more evaluation efforts have to focus on prosody. As explained by Pisoni (1997) and Fourcin (1992), new paradigms for prosody evaluation are needed, as just a few are specifically prosody devoted. More precisely, an important goal for high quality synthesiser evaluation (Sonntag & Portele, 1998) is to measure the capability of synthetic prosody to perform linguistic and pragmatic functions (i) as expected for a chosen application, or (ii) by comparison with natural prosody performances (Rilliard & Aubergé, 1999).

A previous perception experiment was held in the longterm aim of calibrating a grid describing how natural prosody "alone" can perform linguistic functions; and of spoting the distribution of synthetic prosody "alone" on this reference grid (Rilliard & Aubergé, ibid.).

In the aim of extracting from the acoustic data the only pertinent features for prosody perception, the study presented here asks the question of how some objective measurements of the stimuli involved in the subjective evaluation can follow the listeners' answers (Hirst et al., 1998).

Methodological Concerns

Prosody isolation methods

A main hypothesis in such "prosody alone" paradigm is the modular functioning of prosody as regards the phonemic continuum. It means that listeners have to directly or indirectly identify some functional values carried by stimuli built without any lexical (that is morphological and syntactic) information. The function studied is the segmentation / hierarchisation, that is the function shared by prosody and syntax.

Several methods have been proposed (for a complete description see Rilliard, 2000):

• incoherent realisations of the function by prosody vs. syntax (by prosody transplantation, Morlec et al., 1998);

• Semantically Unpredictable Sentences (Benoît et al., 1996);

• delexicalised speech, done by filtering the signal (Sonntag & Portele, 1998), substituting the phones with varied syllables (Pagel et al., 1996), or by reiterant speech on a canonical syllable (by human or by synthetic generation (Rilliard, 2000).

This last method has been chosen for the subjective evaluation, mainly because (1) it avoids the artefactual manipulation of filtering (2) the reiterant speech has been tested as an efficient paradigm (Liberman & Streeter, 1978; Larkey 1983) (3) even if none of the three methods are really ecological, it was supposed that reiterant speech generates the lowest cognitive charge for listeners.

The problems raised by reiterant speech are mainly about the synthetic production of /mamama/ sequences. They can be overtaken by analysis-resynthesis methods applied on recto-tono reiterant utterances produced by a human speaker.

Subjective Experience Design & Results

Paradigm

For a complete description of this experiment, see Rilliard & Aubergé (1999)

Shortly described, this experiment was based on a metalinguistic association task between a prosodic utterance (the reiterant signal) and a syntactic structure (text and possibly lexicalised stimuli) which matched the syntax of the original prosody (homogeneous pairs) or which did not match (heterogeneous pairs).

Condition	Reiterant Stimulus	Lexicalised Stimulus
C1	Natural	Text only
C2	Natural	Text + Natural
C3	Synthetic	Text only
C4	Synthetic	Text + Synthetic
C5	Synthetic	Text + Natural
C6	Natural	Text + Synthetic

Table 1: the different experimental conditions

Possible stimuli are either lexicalised or reiterant, in their natural or synthetic version. Combinations of such stimuli in each possible way resulted in a six experimental condition design (summarised in table 1). Each different condition was intended to analyse a specific set of information about the prosodic functionality.

Stimuli

They have been selected under a criterion of a systematic minimal pair opposition of syntactic variations. Sentences are from 5- to 11-syllable long, and proposed a set of noun, verbal and object groups, and clauses structures. All possible combinations of a reiterant stimulus to each same length lexicalised sentences are proposed.

Stimuli are either reiterant or lexicalised, and natural or synthetic. Lexicalised stimuli have been read aloud by a trained speaker. The synthetic ones have been synthesised (fundamental frequency (F0), duration and intensity) by the ICP TTS system, transplanted onto a recto-tono natural sentence with a PSOLA technique. Reiterant stimuli, both natural or synthetic, have been built by transplantation of prosodic parameters onto recto-tono reiterant utterances recorded by the speaker. Using such a procedure for reiterant sentence construction avoids replication effects of the same [ma] sequence.

Subjects

Each condition was performed by 13 different naive listeners (78 for all conditions).

Results summary

An ANOVA analysis was conducted to study two different sources of variation: the syntactic structure of the stimuli and the experimental condition, for each group of stimulus length – from 5 to 11 syllables.

A first result was the great influence of sentence structure on the variance (significant factor for each stimulus length -p<.01).

Next, the effect of the experimental condition is non significant for each length of stimuli, except for the 7-

and 8-syllable ones. Such a result underlined the global coherence of the paradigm and the general good quality of the synthesised prosody, as the answer was similar for synthetic and natural stimuli, and for each kind of stimuli presentation. The 7- and 8-syllable stimuli behaviour has been explained by a more detailed analysis (a post-hoc Tuckey test): the segmentation function produced by the synthesiser was exaggerated, and association / dissociation score was better for synthetic stimuli than for natural ones. On the contrary, synthetic stimuli receive low dissociation scores for hierarchisation divergences: the synthetic prosody is dedicated to the segmentation function and lacks the other indices.

Finally, the interaction between the condition and the structure of the stimuli is highly significant, except for the 6–syllable stimuli (since the synthesis is of good quality for these ones, results are consistent regardless of the condition). This leads us to a more detailed analysis, to point out which pairs of stimuli demonstrated problems on the synthesised stimuli.

Remarkable results underline

• the primary importance of: the placement of the major syntactic boundaries and the hierarchical concordance of syntactic groups;

• the non significant effect of: little (1 to 3-syllables) syntactic group displacement, and a 1-syllable shift in the major boundary placement.

Objective Analysis

Method

The objective analysis is devoted to characterise the acoustic variation of the stimuli as a function of the perceptual measurement: the aim is to diagnose which prosodic parameters, and moreover which specific evolution of parameters can explain the subjective results described in the previous paragraph.

	Condition C2		Condition C4		Condition C5		Condition C6	
	homog.	heterog.	homog.	heterog.	homog.	heterog.	homog.	heterog.
Raw Fo RMS	146	603	168	428	230	508	261	536
Raw Fo Correl.	0,927	0,422	0,932	0,588	0,833	0,489	0,822	0,488
Norm. Fo RMS	56	413	49	276	111	345	124	356
Norm. Fo Correl.	0,927	0,422	0,932	0,588	0,833	0,489	0,822	0,488
BARK Fo RMS	5,38E-03	3,95E-02	4,73E-03	2,65E-02	1,06E-02	3,30E-02	1,20E-02	3,42E-02
BARK Fo Correl.	0,927	0,424	0,932	0,589	0,833	0,490	0,822	0,489
Syl. duration RMS	5177	6972	3002	3347	3464	3907	5794	6203
Syl. duration Correl.	0,475	0,204	0,328	0,222	0,409	0,239	0,211	0,167
GIPC Durat. RMS	4713	12445	3857	4289	4640	7043	8326	9587
GIPC Durat. Correl.	0,743	0,366	0,643	0,544	0,699	0,469	0,499	0,434
Raw Intensity RMS	118	122	145	147	119	122	143	148
Raw Int. Correl.	0,763	0,744	0,746	0,735	0,773	0,752	0,741	0,720
Norm. Int. RMS	13	14	16	16	13	14	15	16
Norm. Int. Correl.	0,763	0,744	0,746	0,735	0,773	0,752	0,741	0,720

 Table 2: average values of root-mean-square distance and correlation between homogeneous and heterogeneous couples of stimuli, for each kind of opposition of the subjective experiment

Parameters

Each stimulus is analysed (lexicalised, reiterant, natural or synthetic), in order to extract the F0, the minimal prosodic unit duration and intensity. F0 and intensity are stylised using 3 dots for each phoneme (extremes and middle) on the original curve. The F0 of vowels only is kept, in order to avoid problems with unvoiced segments, and keep the same number of parameters for each sentence.

Stimuli

A distance between the two stimuli of each pair is calculated. The pairs of stimuli are those presented to listeners during the experimental conditions C2, C4, C5 & C6.

Distance metric

Two metrics are compared to determine which one is better to characterise the perceptive performances. The acoustic parameters are also expressed with different scales which are more or less related to human perception:

• F0 is used in its raw version, is normalised using a logarithmic scale to an average value of 100Hz, and expressed in BARK, MEL & ERB scales (also normalised);

• duration is calculated using two units: either the syllabic duration, or the Group Inter-Perceptual-Centre (also used in the ICP TTS model – Barbosa & Bailly, 94);

• intensity is used in its raw and normalised versions.

Average results are presented in Table 2 (since results do not differ for BARK, MEL and ERB scales, only the BARK scale appears in Table 2). Homogeneous pairs and heterogeneous ones are presented separately.

Comparison with Subjective results

At this stage, there are, for each pair of stimuli, (1) one association result from the subjective experiment, (2) and a distance value for each metric and each parameter. The correlation between these two kinds of data is calculated. This final correlation reflects the relative adequacy between an acoustic parameter and the subjective answer. These results are summarised in Figure 1.

The correlation of the subjective results obtained during experimental conditions C1 & C3 (where the second part of the pair is textual), are calculated using the pairs of stimuli of the "closest" (in terms of nature of stimuli) experimental condition:

• for C1, based on a natural reiterant signal, the couple of objective values used are extracted from the C2 condition;

• for C3, based on a synthetic reiterant stimulus and text, the association / dissociation results are compared to those of C4 and C5, which use the same reiterant stimulus and either a natural or a synthetic lexicalised stimulus. Two kinds of stimuli couples are used because even if C4 is the closest condition, its lexicalised signal (used as a reference for good prosody) is a synthetic stimulus, and the quality of the carried prosody is not known; whereas the lexicalised stimulus of C5 is a natural one, and can be used as a reference.

Results

From the distance analysis, several major points can be listed:

• Homogeneous stimuli are very close together (either for the correlation or the root-mean-square), whereas heterogeneous ones are more deviant;

• From the two metrics used to calculate distances, only the RMS one is sensitive to the F0 scale (either raw data, normalised or BARK)

• F0 is the strongest factor of correlation between homogeneous stimuli – intensity and duration are more deviant;

• GIPC duration is more accurate than syllabic duration to distinguish between homogeneous and heterogeneous stimuli – but it could be a bias of the transplantation model, which is based on GIPC;

• Natural and synthetic stimuli receive similar scores.



Figure 1: Correlation between the association results of the different experimental conditions and the two distance metrics of stimuli couples (Correlation in the left graph and mean -root-square (RMS) in the right one), for each prosodic parameter. The two F0 correlation curves are superimposed, as well as the normalised and Bark RMS ones.

The correlation between distance and subjective scores shows major trends:

• Root-mean-square metric is sensibly less accurate to reflect subjective results than correlation – and is more sensitive to the different scales (Fig. 1);

• The F0 parameter is the most correlated one to association scores;

• GIPC duration is a better predictor of subjective results than syllables;

• Intensity does not seem to be related to experimental results

Differences between natural and synthetic prosody are mainly related to the duration parameter evolution.

Results of condition with only one reiterant stimulus plus text only (C1 & C3) receive higher correlation to the F0 parameters than conditions with a lexicalised reference (C2, C4, C5 & C6). Does the lexicalised reference increase the cognitive load of listener or disturb them? One can also suppose that listeners process (1) a direct acoustic distance in a low level perceptive loop (part of the treatment that allows a speaker to reproduce the acoustic parameters of a listened utterance), plus (2) the metalinguistic task of the perception experiment. At the opposite, the correlation for the duration parameters is equivalent (even better for natural speech) for both kinds of conditions.

The analysis of synthetic conditions underlines some problem of the prosody generation model:

• as it is designed to produce prosody only for the hierarchisation / segmentation function, the learning corpus was modelled to extract only relevant parameters. It results in a very efficient F0 for this function, but only for this function. Correlation score for F0 are exceptionally high between the C3 results and the synthetic parameters. On the contrary, duration parameters are poorly correlated with synthetic parameters, but they fit quite well the reiterant synthetic – lexicalised natural couple parameter. This observation can point out some problems in the generator's duration model.

Conclusions

The acoustic analysis, filtered by the results of the subjective experiment enlightens the analysis of experimental results, and gives a pertinent answer to some remaining questions. First to check the general coherence of the experiment and the capability of listeners to use efficiently the prosodic parameters. Second to isolate more precisely the generation problems of the synthesiser (duration and over-trained F0 production); and at last to hierarchise the relative importance of the different prosodic parameters (F0 first, importance of duration in specific contexts, and little informative intensity).

Such conclusions are encouraging. On the one hand for further works on this paradigm, applied to other functions of prosody; and on the other hand for a more detailed diagnosis, using finer objective tools, like mutual information for example, to characterise locally the pertinence of prosody.

Acknowledgements

We are deeply grateful to Gérard Bailly, Yann Morlec and Marie Cathiard for their tremendous job during the preparation of stimuli, as well as the fruitful advice on the theoretical problems raised by this experiment.

References

- Barbosa P. & Bailly, G. (1994). Characterisation of rhythmic patterns for text-to-speech synthesis. Speech Communication, 15, 127–137.
- Benoit, C., Grice, M. & Hazan, V. (1996). The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using Semantically Unpredictable Sentences. Speech Communication, 18, 381–392.
- Fourcin, A. (1992). Assessment of synthetic speech. In G. Bailly, C. Benoît and T.R. Sawallis (Eds.), Talking Machines: Theories, Models and Designs (pp. 431–434). Elsevier Science, Amsterdam.
- Hirst, D.J., Rilliard A. & Aubergé V., 1998. Comparison of subjective evaluation and an objective evaluation metric for prosody in text-to-speech synthesis. In Proceedings of the third International Workshop on Speech Synthesis, Jenolan Caves, Australia, 1-4.
- Larkey, L.S. (1983). Reiterant speech: an acoustic and perceptual validation. Journal of the Acoustical Society of America, 73(4), 1337–1345.
- Liberman, M.Y. & Streeter, L.A. (1978). Use of nonsense-syllable mimicry in the study of prosodic phenomena. Journal of the Acoustical Society of America. 63 (1), 231–233.
- Morlec, Y., Rilliard, A., Bailly, G. & Aubergé, V. (1998). Evaluating the adequacy of synthetic prosody in signalling syntactic boundaries: methodology and first results. Proceedings of the first International Conference on Language Resources and Evaluation. Granada, Spain, 647–650.
- Pagel, V., Carbonell, N. & Laprie, Y. (1996). A new method for speech delexicalization, and its application to the perception of French prosody. Proceedings of the International Conference on Spoken Language Processing. Philadelphia, USA, 2, 821–824.
- Pisoni, D.B., (1997). Perception of synthetic speech produced by rule: A selective review and interpretation of research over the last 15 years. In J.P.H. van Santen, R.W. Sproat, L.O. Olive, & J. Hirschberg (Eds.). Progress in Speech Synthesis. Springer Verlag, New York, 541–560.
- Rilliard, A. & Aubergé, V. (1999). Prosody diagnostic using reiterant speech. In Proceedings of the International Conference on Phonetic Sciences, San-Francisco, pp. 37–40.
- Rilliard, A. (2000). Mise en place d'un diagnostic des informations linguistiques véhiculées par la prosodie à l'aide d'un paradigme de réitération. DEA report, University Grenoble III.
- Sonntag , G. P. & Portele, T. (1998). PURR A method for prosody evaluation and investigation. Computer Speech and Language. 12, 437–451.