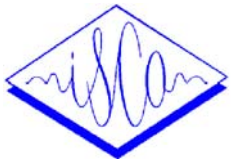

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

TRAINING AN APPLICATION-DEPENDENT PROSODIC MODEL CORPUS, MODEL AND EVALUATION

Yann Morlec, Gérard Bailly et Véronique Aubergé
Institut de la Communication Parlée INPG/U3/CNRS
46, av. Félix Viallet 38031 Grenoble CEDEX FRANCE
e-mail: (morlec,bailly,auberge)@icp.inpg.fr

ABSTRACT

A model of intonation is trained here in order to capture stylistic factors for an application: reading of telephone directory listings. The system was designed to carry out one of the evaluation tasks of the 3rd International Workshop on Speech Synthesis in Jenolan-Australia and the input to the system conforms to the format of the listings defined there. The resulting synthetic prosody is fed into the ICP concatenative synthesis system and compared to natural prosody and prosody obtained from text reading material.

1. INTRODUCTION

Most prosodic models used in current Text-to-Speech Systems (TTS) are tuned on text reading materials and designed to fulfil basic tasks devoted to intonation: salience, hierarchicalisation and segmentation. If these functions are also assumed by intonation in other communication tasks - including other reading materials - the phonological structure and their phonetic instantiation in the speech signal may vary considerably. Together with these "stylistic" factors, the re-use of TTS prosodic models in other reading tasks faces the problem of defining a homogenous linguistic description of the input text: it is difficult to determine clearly in "texts" such as lists, listings or e-mails, syntactic structures such as sentences, clauses or groups.

The aim of this paper is to demonstrate that a linguistically-motivated model of intonation (see section 2) is able to capture such stylistic factors while keeping unchanged the strategy anchoring it to the linguistic structure.

2. THE PROSODIC MODEL

Our prosodic model is based on the following hypotheses:

- Prosodic structures are characterised by the superposition of global multi-parametric contours.
- These contours are anchored on the linguistic structure and directly encode attributes peculiar to each linguistic level (sentence, clause, group, sub-group, ...).

The mapping from the linguistic structure of the message to the set of contours is currently implemented as a hierarchy of sequential neural networks [5], each network - or module - taking in charge the generation of a level-specific multi-parametric contour for a given linguistic level. For instance we assume a strong superposition: multi-parametric contours are just added-up.

A key feature of the model lies in the training stage: while most models are based on a detection of salient events (tones, accents, breaks...) and build phonological structures on top of them, our model operates top-down. Starting from the largest linguistic level to the smallest, each module extracts a prototypical contour for each relevant attribute [2]. It is thus the responsibility of the corpus designer to ensure a statistically significant coverage of linguistic levels and features in the training corpus.

Each prototypical contour thus grabs any phonetic event – even if not salient – that recurs in all or most of the prosodic instantiation of a level-specific linguistic feature. So a melodic accent occurring in all instantiations of an incredulous question, or a declination line occurring in all instantiations of a declarative sentence [5], is considered as part of the prototypical prosodic contour of an incredulous question resp. a declarative sentence.

We test here if such a model that imposes the way linguistic levels are encoded and combined in the prosodic parameters is able to capture application-dependent "stylistic" prosodic factors from a pure linguistic description of the message.

3. THE CORPUS

3.1. Design

The corpus corresponds to the dissemination of telephone directory information. It consists of 256 sentences containing between 20 and 50 syllables and read by a male speaker. Each directory entry conforms to a predefined syntax corresponding to the format proposed for the evaluation session of the 3rd International Workshop on Speech Synthesis in Jenolan-Australia:

<first name><last name>, <number> rue <street name>, <city>, <province>, <telephone number>.

Example: Nathan Montserrat, 52, rue Saint, Mâcon, Morbihan, 02 20 02 12 19.

Since the syntactic structure is fixed, the design of the corpus is only governed by phonotactic distribution: we vary systematically the number of syllables of each item and its phonetic content. Each item has between 1 and 6 syllables with missing intermediate lengths. These missing prosodic units will be properly predicted by the model thanks to the interpolation ability of the network. Former studies [5] also showed that our model is able to appropriately extrapolate prosodic contours of units

10.21437/Eurospeech.1999-373

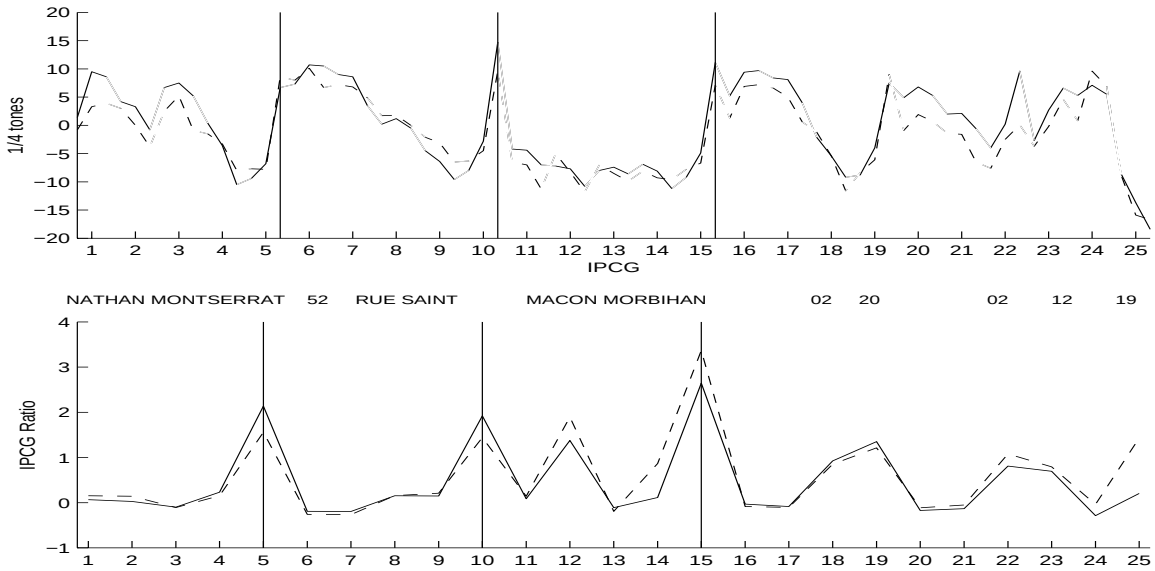

Figure 1: Melodic (Top) and rhythmic (bottom) contours for a 25 syllable utterance. Original prosody is plotted in dotted line and predicted prosody in solid line. Note the 3 major rhythmic lengthenings that separate the 4 principle fields by the emergence of pauses.

containing a few more syllables than the longest unit of the corpus. The whole corpus comprises 256 sentences between 20 and 50 syllables.

People and *street names* are invented and do not correspond to any known person or place.

For the field *<city province>*, each item corresponds to an actual city (among a list of 39076 French cities) and province name but each combination is illicit in order to avoid "clichés" during reading and the use of implicit knowledge by listeners during the evaluation.

Telephone numbers contain 10 digits. The first corresponds to the operator selection (for instance "0" for France Telecom, "7" for Cégétel), the second determines the region (for example "4" for the south-east region). The next 8 digits are usually grouped to form 4 numbers.

3.2. Generation process

3.2.1. Training

Training is carried out using the whole corpus. Learning operates successively on three linguistic levels: sentence, clause (delimited here by commas), and group (each number and lexical item is considered as a group).

3.2.2. Linguistic labelling

At each level, we encode the dependency relations at the boundary between adjacent units [5]: clauses are considered as independent and groups follow the relations adopted for running text.

3.2.3. Pause generation

Pause generation is part of the prediction of segmental duration [3]: pauses are not determined a priori by any linguistic cue or punctuation marks. They can occur in any syllable and emerge from a process limiting the lengthening factor of each segment.

3.2.4. Melody

Melody is characterised by 3 F0 values per vocalic nucleus. Expressed in quarter-tones they are linearly interpolated. A micromelodic contour stored in the polysound dictionary is added during the concatenation/synthesis process.

3.2.5. Prediction

We obtain a very good fit between predicted prosodic contours and original ones (see figure 1). Whatever the number of syllables of the fields, the model is able to capture essential features of this specific task:

- The sentence contour covers long units and globally follows a declination line.
- Clauses are characterised by a final lengthening resulting in a pause, a declination line and an initial accent characterising this particular speaking style, where the speaker try to be as intelligible as possible. Non-final clauses end with a continuation rise.
- The speaking style is also characterised by a speech rate slower than those observed for other corpora produced by the same speaker.
- The model tested in interpolation preserves these features.

We test next whether the prosodic features learned by the model are perceptually relevant by comparing the performance in intelligibility and perceived quality between natural and predicted prosody. We incorporated in the test the prosody trained on text reading material and used in the current text–to-speech system, in order to show that our model is able to generate application-specific prosody without any additional text processing or phonological representation.

4. PERCEPTUAL EVALUATION

In order to simulate a real application, synthetic and natural prosody were fed into the same concatenative speech synthesiser [1] instead of the analysis-synthesis of flat versions used in [5].

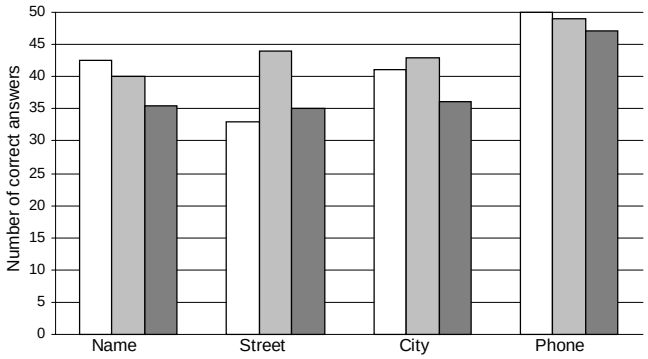

Figure 2: Number of correct transcriptions (maximum 50) for each field and for stimuli with: Original Prosody (in white), Application-dependent TTS system (in light gray) and Standard TTS system (in dark gray).

4.1. Principles

Prosodic models are evaluated using a combined *transcription* and *MOS* (Mean Opinion Score) test. This perceptive multi-criteria evaluation - including a quality and an intelligibility test - follows most of the protocol used by Cartier et al. [4].

The transcription task is essential for this application where listeners asked for precise information and have to catch it in an utterance containing several fields.

Mean opinion scores give indications about the overall quality and intelligibility.

4.2. Stimuli

The stimuli are made of 10 relatively-short telephone directory entries from the corpus containing between 24 and 32 syllables. Three synthetic versions are evaluated:

(1) original prosody (*ORG*).
(2) built from the telephone directory database (*TEL*).
(3) standard TTS prosody (*STD*) delivered by feeding the TTS prosodic model with the same input description.

The *STD* model was trained previously using the same neural network implementation and modular architecture. It differs from the *TEL* system only in the training data: *STD* was trained on utterances structured according to parts of speech below the sentence (clause, noun group, verb group...).

4.3. Experimental procedure

Twenty naïve subjects participated in this experiment. Five sessions with 4 participants were conducted in a quiet listening room. The stimuli were played on a loudspeaker at a comfortable listening level. The 30 stimuli (10 stimuli times 3 synthetic versions) were played in a random order.

Each stimulus was played twice. After the first listening, subjects had to transcribe a part of the message (see below). After the second, they had to give their opinion about the quality and the intelligibility of the utterance.

A short training test containing 2 messages was carried out by the subjects so that they would be accustomed with the synthesis quality and the questionnaire.

4.3.1. Transcription

After the first listening of each message, subjects were asked to write down one of the 4 following fields:

- surname, name
- Number in the street, name of the street
- city, province
- last six telephone digits

In this procedure, listeners knew what field they had to transcribe before listening to the utterance. In a session, each of the 4 listeners transcribed one of the 4 fields of the utterance. They never transcribed consecutively the same field. The a priori knowledge of the field they had to transcribe aimed at avoiding post-processing: they could capture information online.

4.3.2. Mean Opinion Scores

Three questions and 5 corresponding choices of answer were proposed to subjects after the second listening:

Q1: Give your general opinion about the global quality of utterance.
A1: very good, good, okay, bad, very bad.
Q2: Do you have difficulty in understanding words?
A2: never, rarely, sometimes, often, always.
Q3: Is speech rate convenient?
A3: yes, a bit too slow, a bit too fast, too slow, too fast.

4.4. Results

4.4.1. Transcription errors

Following [4] each phonetic error in word transcriptions was counted as a mistake. Figure 2 gives the number of phonetically correct transcription for the 4 fields.

It shows that our TTS system performs well at the segmental level whatever the prosodic module in use. The mean of correct transcriptions is 41.3 (compared with 50).

We can notice that *STD versions* always achieve lower scores than *ORG* and *TEL*. This result is coherent with the MOS results. (see below).

ORG and *TEL* scores are similar except for the street field for which *TEL* is curiously better. Errors concern street names and are uniformly distributed among 6 items.

The telephone field obtained a better score than the 3 others (means=48.6). It could be due to:

- An order effect, since the last six telephone digits are located at the end of the utterance.
- An easier identification of numbers (from 0 to 99) compared to the transcription of unpredictable persons, street or city names.

4.4.2. Mean Opinion Scores

Each opinion scale has 5 possible answers. As in [4] these answers are coded from 5 to 1: from the best one to the worst. For Q3 about speech rate, 2 scores are derived:

- A quality score (5, 3.5, 3.5, 2, 2).

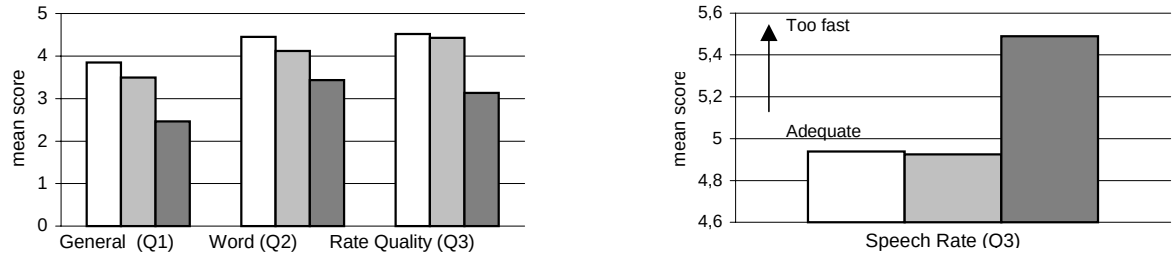

Figure 3: Mean opinion scores (MOS) for Q1, Q2 and Q3. For Q3 2 scores are deduced: quality of speech rate (left) and speed of speech rate (right).

- A speed score (5, 4, 6, 3, 7).

We made no attempt to normalise the scale for each listener and between questions.

The MOS for Q1, Q2 and Q3 are given in figure 3. Both graphs reveal a clear hierarchy between the 3 synthetic versions: *ORG* obtains the best scores, *TEL* is lower but stays very close, *STD* is clearly worse.

Q1 scores (general opinion about the quality of the utterance) are lower (means=3.26) than Q2 and Q3 (means=4 for both). Q1 rates the quality of synthetic speech versus the quality expected from natural speech.

Q2 and Q3 scores seem to take into account that all stimuli are synthetic. However they have very few problems with the understanding of words (this results is correlated with transcription scores) and speech rate is noted as "good" for *ORG* and *TEL*.

The right graph of Figure 3 represents the perception of speech rate (5 represents an adequate speech rate, above region is too fast, below region is too slow). We can see that *ORG* and *TEL* have an adequate speech rate whereas *STD* is too fast in this particular application.

4.4.3. Interpretation

Transcription scores and MOS for Q2 show that:

- Our TTS system produces a good speech quality at the segmental level.
- When prosodic predictions are not quite adapted to the current application, intelligibility may decrease rapidly.

MOS results show that the corpus designed for an application concerning the reading of telephone directory listings has been properly designed:

- The speaker's prosodic realisations were adequate.
- The prosodic model gives accurate predictions that keep essential features of the original prosody.

The biggest difference between the pair (*ORG,TEL*) and *STD* concerns speech rate. Speech rate for *STD* is too high in that application although adequate for the reading of standard sentences. We will examine which of speech or articulation rate is responsible for this.

5. DISCUSSION & CONCLUSION

Thanks to a perceptual experiment combining transcriptions and MOS, we evaluated the performance of our prosodic model trained - specialised - for a given application.

The results show a significant preference for the *TEL* versions over the *STD* ones. Nevertheless, these results have to be moderated since a direct comparison between *TEL* and *STD* is not quite realistic. Indeed the prosodic module generating *STD versions* was trained on well-formed sentences (as in most of TTS systems) whereas *TEL versions* have a completely different and specific format. The matching between *STD* prosodic units and *TEL* prosodic units has to be discussed. The second remark is that the "standard" prosodic module was trained with utterances often shorter and with fewer of clauses in each sentence than those proposed in the new corpus. This means that *STD* versions are often the results of extrapolation that degrades the quality of synthetic speech.

However, we can claim that our model is versatile and can be adapted to any application thanks to the optimal design of an application-dependent corpus used to train our connectionist prosodic model.

6. ACKNOWLEDGEMENTS

This work was supported by COST258 "Naturalness of Synthetic Speech" and AUPELF ARC B3.

7. REFERENCES

[1] Alissali, M. & Bailly, G. COMPOST: A client-server model for applications using text-to-speech. In *Proceedings of the European Conference on Speech Communication and Technology*, volume 3, pages 2095-2098, Berlin, Germany, 1993.

[2] Aubergé, V. & Bailly, G. Generation of intonation: a global approach. In *Proceedings of the European Conference on Speech Communication and Technology*, pages 2065-2068, Madrid, 1995.

[3] Barbosa, P. & Bailly, G. Generation of pauses within the z-score model. In Jan P.H. Van Santen, Richard W. Sproat, Joseph P. Olive, and Julia Hirschberg, editors, *Progress in Speech Synthesis*, pages 365-381. Springer Verlag, New York, 1997.

[4] Cartier, M., Emerard, F., Pascal, D., Combescure, P. & Soubignou, A. Une méthode d'évaluation multicritère de sorties vocales. Application au test de 4 systèmes de synthèse à partir du texte. In *Actes des XIXèmes Journées d'Etude sur la Parole*, pages 117-122, Bruxelles, Belgique, 1992.

[5] Morlec, Y., Bailly G., & Aubergé, V. Synthesising attitudes with global rhythmic and intonation contours In *Proceedings of the European Conference on Speech Communication and Technology*, volume 1, pages 219-222, Rhodes - Greece, 1997.