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

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

ABSTRACT

The topic of this paper concerns automatic training of prosody for a text-to-speech synthesis system. We propose a method that permits the system to be specialised for a concrete application by training it on a representative corpus of data for that application.
Design, training and evaluation of the system is carried out for an application concerning the reading of telephone directory listings. The format follows the benchmark proposed for the evaluation session of the 3rd international workshop on speech synthesis in Jenolan-Australia.

1. THE PROSODIC MODEL

We hypothesise that:
(1) prosodic structures are characterised by the superposition of global contours (Auberge, 1992) which evolve in a multi-parametric space;
(2) these contours directly encode attributes peculiar to each linguistic level (phrase, group, sub-group, ...).
The implementation following from this global approach consists of a collection of level-specific neural networks, responsible for predicting the expansion of melodic and rhythmic movements (Morlec, 1997). The final prosodic contour is then obtained as the weighted sum of the predictions from the ensemble of modules. Hierarchical training is carried out using a corpus where the application, the set of linguistic levels, and the minimal oppositions between the attributes for each level are given. The training procedure allows us to obtain an application-dependent parameterisation of the model (neuron connection weights and initial activations).

2. THE CORPUS

The corpus corresponds to an application involving dissemination of telephone directory information. It consists of 256 sentences read by a reference speaker. Each directory entry conforms to a predefined syntax corresponding to the following format:
<first name> <last name>, <number>, rue <street name>, <city province>, <telephone number>
Since the syntactic structure is fixed, the construction of the corpus is guided simply by phonotactic variation: the distribution of the number of syllables for each item and variation in phonetic content (three possibilities for each item and length).

The corpus is generated automatically according to these structural specifications. This automatic generation method provides an optimal distribution of morpho-syntactic and syllabic structures within the corpus as in [1]. Generation of the corpus operates on three linguistic levels: sentence, clause (delimited by comas) and group.

3. TRAINING

Training is carried out using the whole corpus. The mapping from the morpho-syntactic and phonotactic description of these messages to their segmental durations and melody is hierarchically learned thanks to a sequential neural network [4] that implements the modular and superpositional model of intonation used in the ICP text-to-speech synthesis system. Learning operates succesively on three linguistic levels: sentence, group (delimited here by commas), and sub-group. Pause generation is part of the segmental duration generation process [2].

4. ASSESMENT

4.1. Principles

[3] The prosodic model is evaluated using a combined Transcription and Mean Opinion Score (MOS).

4.2. Stimuli

The stimuli are composed of 10 telephone addresses extracted from the corpus. Three synthetic versions are evaluated: (1) synthesis with original prosody;

(2) previous COMPOST text-to-speech system, where the training data consisted of utterances structured according to parts of speech below the sentence (clause, noun group, verb group...);

(3) COMPOST text-to-speech system built from the telephone directory database.

Figure 1: The sentence module architecture

4.3. Experiment procedure

4.3.1. Transcription

4.3.2. Mean Opinion Scores

4.4. Results

4.4.1. Transcription mistakes

4.4.2. Mean Opinion Scores

5. CONCLUSIONS

REFERENCES

[1] Véronique Aubergé and Gérard Bailly. Generation of intonation: a global approach. In *Proceedings of the European Conference on Speech Communication and Technology*, pages 2065–2068, Madrid, 1995.

[2] Plínio Barbosa and Gérard Bailly. 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.

[3] M. Cartier, F. Emerard, D. Pascal, P. Combescure, and A. Soubignou. 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, 1999.

[4] Yann Morlec, Gérard Bailly, and Véronique Aubergé. 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.