APPLICATION-DEPENDENT PROSODIC MODELS FOR TEXT-TO-SPEECH SYNTHESIS AND AUTOMATIC DESIGN OF LEARNING DATABASE CORPUS USING GENETIC ALGORITHM

O. Boëffard and F. Emerard
France Telecom - CNET
DIH/RCP, 2 avenue Pierre Marzin, 22307 Lannion - France
E-mail: boeffard@lannion.cnet.fr

ABSTRACT

The quality improvement of a Text-To-Speech synthesis system is usually considered as the arduous task of converting any text into speech. This paper is related to the work led at CNET in building application-oriented text-to-speech systems. For a majority of vocal services, the delivered messages have a strong syntactic constraint and use a limited vocabulary. We consider that, with our system, the most hopeful improvements in the overall quality of the speech synthesis signal are linked to the linguistic and prosodic processing. Discarding here segmental problems of the synthetic speech signal, the actual prosodic patterns are judged as too monotonous to allow a great diversity of vocal services. Thus, the actual effort deals with the development of automatic systems to adapt the parameters of statistical prosodic models to a specific speaker's voice under the constraint of a limited amount of different syntactic structures. This work presents an automatic system to build "optimal" training databases used to learn the models' parameters. The formulation of the problem is defined as a set covering problem and is solved using genetic algorithms. Both an objective and a subjective evaluation show the usefulness of this approach.

INTRODUCTION

The availability of effective telecommunications media allows nowadays a great development of services based on vocal technologies. Generally, these services are specialised: for example one can access a meteorological forecasting service, a phone order shopping or an intra-firm phone directory. The restricted field of the vocal service allows an ergonomic interface accepted by the end-user to be defined. Thus, the messages delivered by such systems have a small syntactic variability. Usually, a message is a sentence composed of two parts: one, invariable, is a template for variable fields containing the information for the end-user. Up to now, most vocal services use pre-recorded, compressed and stored natural speech. Obviously, speech output based on this technology has a very good quality. Nevertheless, this technology is unfeasible if the service needs fast updates in the databases: such updates occur in, for example, intra-firm phone directory application. The solution for vocal services requiring continuous updates consists in using a Text-To-Speech synthesis system. This kind of speech synthesis system generates the speech messages sent to the users.

This paper presents an experiment conducted at CNET related to an intra-firm phone directory service using both speech recognition and text to speech synthesis. The overall quality of the CNETVOX Text-To-Speech synthesis system for French is generally judged as acceptable for main text to speech applications. But, for an application where the message has the same syntactic construction, the prosody is too monotonous. It is the reason why an automatic prosodic adaptation system has been developed. The objective of such a system is to learn as well as possible the prosody of one specific speaker uttering messages of the specific application. This application dependent speech synthesis system predicts the prosody with statistical models and needs learning databases to estimate the models' parameters.

The goal of this paper is to demonstrate the crucial work of building an optimised learning database and its influence on the speech output quality. A method to build such an optimised database is proposed. This database is specific to an application and is recorded by a speaker. The speech synthesis system tends to mimic the natural prosody exhibited by the speaker.

In part one, the phone directory inquiry service is presented. Then, in part two, the method for building an optimised learning database is given. Finally, in part three, both an objective and subjective evaluation is exposed.

1- PHONE DIRECTORY INQUIRY SERVICE

Using an intra-firm phone directory inquiry service, a user can automatically obtain, by phone speech recognition, the information on a phone correspondent and ask the system to perform or not the call. The system gives to the user the full name and the phone number of the correspondent. Two kinds of synthetic speech messages are delivered : unchanging messages related to the ergonomic constraints of the application and varying messages giving the information to the user (varying parts or fields are the first name, the family name and the phone number of the correspondent). At CNET, the acoustic level is realised by concatenating diphones and a set of longer units and

processing the speech units with the TD-PSOLA technique [1] : this defines the acoustic string of the message. For the unchanging part of the message, a natural prosody is put on the acoustic string; for the varying parts of the message a model is used to generate the prosody on the acoustic string.

1.1- Prosodic models

For each varying field, the modelised prosodic parameters are the segmental duration (one value for each phoneme) and the F0 pattern (two values for each phoneme). As in lot of statistical prosodic models [2][3], input variables can be of different kinds: there are language- or application-dependent syntactical variables, syllabic variables and phonemic variables. The segmental duration and the F0 patterns are observed through a time dimension. The models, presently described, take into account this time dimension using contextual windows for each variable; the length of a contextual window depends on the nature of the variable. The models used both for duration and F0 prediction are neural networks.

1.1.1- Segmental duration model

Three boolean variables say if the current phoneme is located or not at the end of a field, if the current phoneme is followed or not by a final pause and if it is followed or not by a non final pause. Two real variables give the position of the current syllable inside the current word and the number of syllables in the current word. Two real variables give the position of the current phoneme inside the syllable and the number of phonemes in the current syllable. Finally, three modal variables give the class of the previous, current and next phoneme (a class contains 7 modalities). The output (the duration of the current phoneme) is a real variable taking values on a linear scale and normalised on [0,1]. The neural network used for the duration is a three layered network. The input layer contains 32 cells, the hidden layer contains 15 cells, and the output layer one cell. The activation functions of the cells are sigmoid functions. A modal variable is encoded in a binary way with a 1-in-n technique. A real variable is presented as a real continuous value to the network.

1.1.2- F0 pattern model

A modal variable indicates the field that the current phoneme belongs to. Two boolean variables say if the current phoneme is followed or not by a final pause and if it is followed or not by a non final word. Two real variables give the position of the syllable in the field and the number of syllables in the field. Two other real variables give the position of the current phoneme in the syllable and the number of phonemes in the syllable. Finally, a modal variable indicates the phonemic class of the current phoneme (1 modality over 7). Two output values give the F0 value at the beginning and at the end of the phoneme. These outputs are real variables taking values on a linear scale and normalised on [0,1]. The neural network used for the F0 is a three layered network. The input layer contains 18 cells, the hidden layer contains 10 cells, and the output layer 2 cells. The activation functions of the cells are sigmoid functions. As for the duration model, a modal variable is encoded in a binary way with a 1-in-n technique and a real variable is presented as a real continuous value to the network.

1.2- Learning process of model parameters

The set of learning samples, defined in a training database, contains pairs of input/output variables observed in a natural prosody corpus. This set of learning samples is directly related to the quality of the models: their generality over new unseen inputs and their robustness. A learning database should contain the maximum variability of the phenomenon, but observed through the input variables of the models. Once the learning database is defined, the learning sentences are recorded by a speaker. An automatic labelling system of speech into phonemes (including speech segmentation into phones and the alignment of an automatic phoneme transcription of the text sentence with the speech) and an automatic processing of pitch tracking calculates prosodic information for each phoneme of the sentence. A segmental duration and two F0 values are assigned to each phoneme. This process can also detect an acoustic pause and its duration at a word boundary. The automatic phoneme sequence contains multiple phone sequence depending on phonological information or "regional" variants of pronunciation. This optimal sequence is chosen by the alignment process based on probabilistic criterion [4]

The Aspirin/MIGRAINES software [5] was used to optimise the neural networks parameters. A validation database (natural messages not used in the learning and test databases) is used to find a heuristic threshold in stopping the learning process. This threshold is a maximum of processing iterations on the whole training database defined when the mean square error of a model output increases on the validation database.

2- OPTIMAL LEARNING DATABASE DESIGN

This section presents the solution developed to automatically design the learning database. The system finds a minimal set of sentences which covers the variability of the phenomena (duration and F0) modelised in the text-to-speech system.

2.1- Problem description

With the phone directory inquiry application, only one type of sentence embedding variable fields is defined: "Jean Dupont, poste 12 00". Three different fields are defined: the first name, the family name and the phone number. The application database contains

about 1800 lines. In order to take into account the combination between the fields themselves, the main goal is to build a virtual database instead of finding a subset from the application database. Thus, the three fields are studied separately and can take values in three different sets: 332 utterances for the set of first names, 1343 utterances for the set of last names and 10000 utterances for the set of phone numbers (all combinations of two numbers ranging from 0 to 99). For the three sets, an utterance is associated with a vector of describing variables. The system finds a subset covering as much as possible the entire set of vectors. Once the optimised subsets are found, a virtual database is build generating sentences whose fields take values in the subsets. This "optimal" database contains 100 sentences and will be recorded by a speaker.

The vector variables describing each utterance of a field follows phonological and phonemic properties Three kind of variables are defined: variables describing word lengths, an histogram of phonetic classes (6 classes) and an histogram of syllabic structures (7 classes) of the utterance.

2.2- Problem formulation

For each set; fixed-length vectors are filled according to the analysis of utterances. A set is represented by a two dimensional matrix: the concatenation of all the describing vectors. Each vector variable is encoded into a binary format. A 232x332 matrix is associated to the set of first names, a 232x1343 matrix to the set of family names and a 232x10000 matrix to the set of phone numbers. The densities of these matrices are closed to 20%. The optimal subset is defined as the solution of a set covering problem. The problem constraint formulates that each line has to be covered by one column. A weight is associated with each column, the optimal subset is the set of columns with the minimal cumulative weight and following the problem constraint. The solution of this problem is NP-complete [6].

2.3- Technical solution

The technical solution adopted to solve the previous formulation is based on genetic algorithms. A genetic algorithm is able to find a global optimum for a problem of combination. The algorithm adapted here is based on previous work conducted to solve a bus-driver scheduling problem [7]. All the matrix columns have constant weights. As the cardinal of an optimal subset cannot be directly controlled by the algorithm, an iterative process is achieved. While the cardinal of the optimal subset is under a threshold parameter, the optimisation algorithm is run on the entire matrix discarding the columns of the optimal set. This technique gives the N best set covering vectors in K iterations (N was set to 100 and the experiments give an average of 9 for K). The N sentences built with the three optimal subsets define the optimal learning database.

3- EVALUATION

The optimal database is evaluated under two different evaluations. An objective evaluation consists in looking at the mean squared error from the output of the prosodic models. An overall multicriteria perceptive test conducted on speech synthesis messages is the subjective test.

3.1- Objective evaluation

The optimal database is evaluated with its ability to give to the learning process of the prosodic models well-adapted samples of application sentences. Firstly, an informal listening test has shown that there is a link between a decrease in the mean squared error output of the neural networks and an increase in the prosodic parameters quality. Thus, the criterion defined for an objective measure of the quality of the models outputs is the evolution of the mean squared error through the learning iteration process. Two learning databases are used: a first one, A, is the optimal database as defined in section 2-, a second one, B, contains 100 messages randomly chosen from the application database. Each of the two databases is split into two parts: the first part (subscript L) contains 67% of the whole database for the learning process and the second part (subscript T) contains 33% of the whole database for testing the models. Because of the ordering importance of the optimal database, three partitions are defined for the two learning databases: a random 67% / 33% split (superscript 1), an ordered 67% / 33% split from the beginning of the database, (superscript 2), (for the optimal database, the 67% best samples are used for the learning process) and an ordered 67% / 33% split from the end of the database (superscript 3). Therefore, 6 different learning databases are used for this validation: A_L^1, A_L^2, A_L^3, B_L^1, B_L^2, B_L^3 and 6 different test databases: A_T^1, A_T^2, A_T^3, B_T^1, B_T^2, B_T^3. The convergence curves of the output mean squared error have a different behaviour depending on what learning database was used: A or B. Indeed, for the database B (100 randomly selected sentences), there is a well-known behaviour of error curves: the error curve on the learning partition is under the error curve on the test partition. But, for the database A (defined in section 2-) and for the partition 2 (an ordered split from the beginning of the whole database), the error curve on this learning partition is always above the error curve on the test partition. This behaviour of the error curves confirms the appropriate design of the optimal learning database. A study on the amount of learning messages to keep to the leaning process shows an inversion of the error curves starting when less than 20% of the optimal database is used for the learning partition.

3.2- Subjective evaluation

An AB/BA paired comparison test has been conducted to compare the overall quality of three different prosodic components in the CNET synthesis system. The 8 subjects who participated in the experiment were students who had no experience of listening to synthetic speech. They were aged between 18 and 22.

3.2.1- Assessment method

3 sources of synthetic speech messages are evaluated: the standard CNETVOX Text-To-Speech synthesis system for French (1), the CNETVOX Text-To-Speech synthesis system with the prosodic component built from the optimal database A (2) and the CNETVOX Text-To-Speech synthesis system with the prosodic component built from a random database B (3).

20 different messages chosen from the application database were synthesised by each system at a 8KHz sampling frequency in the 300Hz-3400Hz phone bandwidth. A stimulus contains a pair of two identical messages generated with two different sources.

3 comparisons of the 3 sources are defined: 1-2, 1-3 and comparison 2-3. AB and BA pairs were presented in a random order. They include a one second pause between sentences.

The test has been conducted in a quiet listening room. The stimuli were played on two loudspeakers at a comfortable listening level. A short training period was imposed on the subjects so that they would be habituated to synthetic messages and to transcribe their answers on forms. Each stimulus was listened only once by the subjects; the response "Equivalent" was available but not recommended.

3.2.3- Results

Table 1 shows the results of the AB/BA preference test. A line, in the table 1, is related to a source comparison. The column A indicates the percentage of votes for the first system of the comparison, column B the second system in the comparison and column C the equivalencies.

Table 1: Results of the subjective test

	A	B	C
1-2	32.5%	52.5%	15.0%
1-3	45.0%	55.0%	0.0%
2-3	57.5%	42.5%	0.0%

For the 1-2 comparison, the 5% confidence interval is [47.0%, 57.9%], For the 1-3 comparison, it is [49.5%, 60.4%] and for the 2-3 comparison it is [52.0%, 63.0%]. These confidence intervals show that the preferences are significant.

The results show that the objective difference pointed out in section 3.1 is confirmed by a perceptive difference of the speech synthetic signals. Both prosodic models using statistical models are preferred to the standard prosody generated by CNETVOX. The prosody generated using the optimal database is preferred to the prosody generated with the random database. The two databases contain the same amount of information.

CONCLUSION

In this paper, an automatic system to design training databases used to model prosody for application-dependent text-to-speech synthesis systems is described. Both an objective and a subjective experiment have been conducted to show the validity of this approach. The main contribution of this method is the reduction of the amount of data needed to learn the parameters of the statistical prosodic models. The design of the "optimal" database is clearly related to the adequacy of the statistical predictive models to mimic the natural prosody. Furthermore, the subjective evaluation highlights a significant preference for the statistical models over the standard text-to-speech system.

ACKNOWLEDGMENT

The authors acknowledge the valuable contribution of the MITRE Corparation which provides the Aspirin/MIGRAINES software.

BIBLIOGRAPHY

[1] Bigorgne, D., & al., "Multilingual Psola Text-To-Speech System", proceedings of IEEE-ICASSP, pp 197-190, 1993.

[2] Ross, K.N., "Modeling of intonation fo speech synthesis", PhD dissertation, Boston University, 1995.

[3] Scordilis, M.S., & Gowdy, J.N., "Neural network based generation of fundamental frequency contours", Proceedings of IEEE-ICASSP, pp 219-222, 1989.

[4] Quazza, S., "Predicting durations by means of automatic learning algorithms", IV Workshop of the experimental phonetics group, Turin, Italy, 1995.

[4] Boëffard, O., & al., "Automatic segmentation and quality evaluation of speech unit inventories for concatenation-based, multilingual psola text-to-speech systems", proceedings of EUROSPEECH, pp 1449-1452, 1993.

[5] Leighton, R.R., "The Aspirin/MIGRAINES Neural Natwork Software, V6.0", ftp://pt.cs.cmu.edu/afs/cs/project/connect/code/am6.tar.Z

[6] Grossman, T., & Wool, A., "Computational experience with approximation algorithms for the set covering problem", Technical Report, Los Alamos National Lab., february 1995.

[7] Gonzalez-Hernandez, L.F., "Evolutionary Divide and Conquer for the Set Covering Problem", MSc thesis, DIA dpt., Edinburgh University, 1995.