A METHOD OF UNIT PRE-SELECTION FOR SPEECH SYNTHESIS BASED ON ACOUSTIC CLUSTERING AND DECISION TREES

Christophe Blouin, Paul C. Bagshaw & Olivier Rosec

France Telecom R&D, Lannion, France.

ABSTRACT

This article presents a method to pre-select units for concatenative speech synthesis. The method is based on a procedure of unsupervised acoustic clustering that is coupled with a decision tree for each type of unit. During synthesis, the trees predict the acoustic classes for a given symbolic target and the units belonging to the predicted classes are retained as candidates for the final selection. The units used in this study are phones and diphones, although the methodology is entirely automatic and may be applied to any type of unit or language. The proposed method is evaluated by comparison to a handcrafted method in a formal listening test.

1. INTRODUCTION

1.1. Context

In the field of text-to-speech synthesis, the current trend to generate the signal is by the concatenation of acoustic segments. Many recent systems store a large quantity of acoustic data and select the most appropriate units for the synthesis of a given text. The selection of units breaks down into two stages: the pre-selection and final selection.

The pre-selection searches the data for sets of candidate units, which may be used for synthesis. Generally, each of these sets corresponds to a target phonemic sequence, such as phoneme, diphoneme, or a longer sequence. In this way, the units kept as candidates are those for whom the signals are subjectively the closest to a target signal corresponding to the target phonemic sequence. Following this, the final selection keeps amongst the candidate units only the sequence of units that minimises potential discontinuities at the concatenation of unit signals.

1.2. Terminology

Units are defined as the couples composed of an acoustic element and its associated symbolic element. Here, the acoustic element corresponds to a segment of the speech signal, which is characterised by a set of acoustic parameters. The symbolic element corresponds to a set of symbolic parameters that are predicted by the system to describe the signal of an acoustic element. These symbolic parameters may be of different nature, such as linguistic, phonological or prosodic.

During the synthesis, the system generates a sequence of target symbolic elements (hereon called symbolic target), to which the selection process must associate a signal from the units in the acoustic data. The ideal signal corresponding to a symbolic target is the hypothetical signal that would have been produced had the speaker read the target text (hereon called the target signal).

1.3. The Problem

The fundamental problem in pre-selection resides in the definition of a measure between the symbolic target and the units in the database that reflects the subjective distance between a hypothetical target signal and a unit's speech signal. In theory, the units kept as candidates are those whose signals are the closest to the target signal. In practice, the nature and the combinations of the parameters to be taken into account in such a distance are a subject of research.

The distance measure employed may be defined arbitrarily, such as that based on a user-defined hierarchy of symbolic parameter importance represented by an indexation tree [3]. This approach can be used with a variety of language independent symbolic parameters [2]. However, such a definition is highly dependant on expert knowledge and the pertinence of the chosen symbolic parameters.

Alternatively, a distance measure may be derived automatically [6][1], such as by the creation of decision trees for which the goodness of a split (given a symbolic parameter) is based on the acoustic variance of units in the resultant sub-branches (inspired from [9]). This technique enables units at the terminal nodes to have symbolic parameters that are close, in terms of the derived distance, to those of the symbolic target parameters. Furthermore, it enables the pre-selection to propose candidate units that are mutually acoustically similar. However, the tree hierarchy imposes the parameters of the symbolic element

of candidate units, which are tested on the path to the terminal node, to be identical to those of the symbolic target. Thus, it does not enable units to be selected whose signal could be close to the target signal but whose tested symbolic parameters are different than their target. This problem may be overcome by substituting units in a terminal node with units taken from other terminal nodes of the tree, whereby the surrogate units are acoustically close to the centroid of the set of candidate units at the terminal node and the replaced units are the furthest from the centroid [5]. However, such reinforcement of the homogeneity of the units is indirectly conditioned by the symbolic parameters of each terminal node since the centroid depends upon them.

The method of unit pre-selection presented here attempts to assure the acoustic homogeneity of the candidate units by the use of acoustic clustering that is independent of the imposed hierarchy of symbolic parameters. Furthermore, the method enables the pre-selection of units that are acoustically close to the target signal but that do not necessarily have a symbolic element that is identical to the symbolic target.

2. A METHOD OF UNIT PRESELECTION

2.1. Overview

Two stages are performed prior to the synthesis of a given text. First, a procedure of acoustic clustering is applied to each set of units of the same phonemic type in order to generate homogenous acoustic classes for each phone and diphone. Second, a mapping is established between the parameters of the symbolic element of the units and the acoustic classes obtained. Such a mapping may take the form of a decision tree, a neural network, or other learning machine. Taking the case of a decision tree, this produces a hierarchy of the symbolic parameters, with the most pertinent at the root of the tree.

During the synthesis of a particular text, the decision tree (for example) returns the probabilities for the acoustic classes at a terminal node given a symbolic target. The more probable acoustic classes each contain a set of units whose symbolic elements are similar, but not necessarily identical, to the symbolic target and whose acoustic elements are mutually close.

This method of pre-selection chooses candidate units that are close to the target signal because they are either 1) symbolically close to the symbolic target in terms of the automatically determined relative pertinence of the symbolic parameters (for each type of phoneme), or 2) acoustically close to those symbolically close units.

The members of some acoustic classes have different symbolic elements. During the search for units belonging to classes identified by the decision tree, it is these classes that enable apt acoustic elements to be retrieved even if the unit's symbolic element differs from the target.

2.2. Acoustic Clustering

The goal of the acoustic clustering in this context is different than that in general methods of clustering [7]. Here, the constraint is not to find acoustic classes that best model the distribution of the data, but to derive homogeneous classes that 1) contain sufficient members so as to identify the acoustic similarities between units that are not necessarily symbolically equivalent and, 2) do not contain too many members in order for pre-selection to filter units. The homogeneity of classes is required so as to avoid taking units as candidates if they are acoustically distant from other symbolically equivalent units.

One way to satisfy these constraints is the application of a bottom-up hierarchical agglomeration clustering algorithm [7] based on an inter-unit acoustic distance measure. If the total number of members of the two closest sub-classes is greater than a specified maximum, then the merger of the two sub-classes is forbidden and both are excluded from further possible fusions.

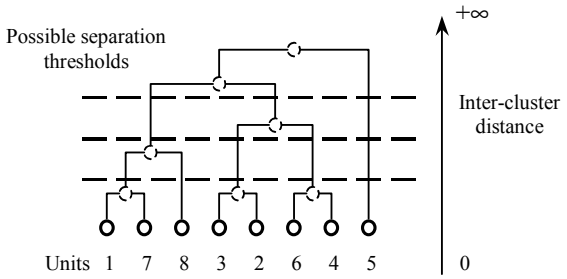

Figure 1: Dendrogram obtained by agglomeration clustering

Figure 1 illustrates the influence of the imposed limit on the number and construction of classes. The horizontal dotted lines indicate the separations obtained with an ad-hoc criterion imposing a maximum limit of 2 (bottom), 3, or 4 (top). The separations result respectively in 5, 4, or 3 classes.

The criterion that inhibits further clustering is not only a simple limit on the maximum number of members per class. It is also based on a measure of inter-class distance variance (consistency coefficient). The derivation of empirical values is discussed in section 2.5.

2.3. Acoustic Distances

The procedure to generate acoustic classes resides entirely on an acoustic measure and its capacity to reflect the subjective distance between two speech signals.

The distance used here is inspired by the results of a previous study on acoustic distances for speech synthesis

[12]. The distance is based on dynamic time warping (DTW) [11][8] with type I constraints, which generates an alignment path without right angles and whose end-points are mutual comparisons of the initial and final frames. The global distance between two signals equals the average of the local distances in the alignment path multiplied by the ratio of the longest duration to the shortest duration. This global distance thus penalises relative differences in duration between the two compared signals.

Local distances in the DTW alignment are defined as the Euclidian distance between 12 Mel-frequency cepstral coefficients (without the first) of 16ms Hann-windowed frames calculated at 8ms intervals (16kHz sampling rate). This distance, which is often used in speech recognition, seems well correlated with a subjective distance [12].

2.4. Decision Trees

A learning machine, such as a decision tree [4], can automatically establish the mapping between the symbolic elements of units and their acoustic classes. As a by-product, a decision tree structure gives an indication of the pertinence of each symbolic parameter to reflect the acoustic characteristics of a unit. This is helpful in the design of the speech synthesis system, which aims to generate the most pertinent symbolic representation possible for a target text.

Many algorithms exist for the generation and application of decision trees, depending on the structure of the trees generated, the type of input parameters and the splitting criteria used. Any such algorithm capable of processing both discrete and continuous input parameters [10] is advantageous, although this study uses only discrete parameters predicted by the linguistic modules of the synthesis system (predicted phone F0, duration, energy are yet unused). Symbolic parameters used are those such as the position of phoneme/syllable/word and Chomsky-Hall phonological features.

The prediction of an acoustic class (given the target symbolic parameters) uses the full decision tree generated by C4.5. No branch pruning is applied apart from the default pruning of outliers.

2.5. Empirical Calibration

A compromise has to be made between the number of classes generated by the acoustic clustering and the ability of the learning machine to predict a class given the symbolic parameters. If each class contains only one member then it is perfectly acoustically homogenous, but this results in a larger number of classes and it will be difficult to accurately predict the class corresponding to a set of symbolic parameters. Conversely, if all units belong to the same class, the learning machine can predict the class without error, but the members of the class are in no way acoustically homogenous. Furthermore, the composition of a given class will also affect the machines predictive ability since a class that is also inadvertently symbolically homogenous will be easier to model.

The limit on the number of members in a class and the inter-class distance variance used in the acoustic clustering control the composition and number of classes. These parameters are empirically varied so as to minimise the error rate with which the decision trees predict the unit classes. As an example, the learning error rate for diphones A_B and P_L decrease respectively from 57% and 57% (one unit per class) to 39% and 31% for the optimised acoustic clustering.

3. EVALUATION

The evaluation of the proposed method of pre-selection is linked to the quality of the set of candidate units it produces. The candidate units are subject to the process of final unit selection and to unit concatenation. The quality of the speech synthesis thereby obtained is evaluated in a formal listening test.

3.1. Final Unit Selection

For the purposes of evaluation, a unit is defined as a consecutive sequence of two demi-phones (i.e. a phone or a diphone). The final selection traverses a trellis of demi-phone candidates by application of a Viterbi algorithm that takes only concatenation costs into account. The weights attributed to concatenation sub-costs are obtained by a multiple linear regression as a function of an acoustic measure of concatenation quality [2]. A sub-cost derived from 12 MFCC is also employed here.

Signal processing, applied for prosodic modification and the concatenation of overlapping units, is realised by the TD-PSOLA algorithm. Finally, signal energy and fundamental frequency are linearly smoothed over several periods around concatenation points.

3.2. Formal Listening Tests

The quality of speech synthesised by the method described here (hereon referred to as AUTO) is evaluated against three other methods: UL and OS, formerly described [2], and EXP, a variant of OS that is given the same symbolic parameters as those exploited in the decision tree generated for AUTO (there is an expert-defined hierarchy of symbolic parameter importance) and that uses the same final selection process as AUTO. The results for only AUTO and EXP are divulged here since the other two methods concern changes other than the pre-selection method.

16 naïve subjects (all native speakers of the target language) listen to 20 randomly ordered, phonetically

balanced sentences, each synthesised by the four tested systems (80 utterances). They are asked to judge the global quality of each utterance on a scale of 1 (bad) to 5 (good).

3.3. Results

The test reveals that subjects prefer the sentences obtained with the EXP pre-selection method (MOS: 3.45) to the version AUTO (MOS: 2.05). There are however three sentences for which the difference in MOS is less than 0.25 and six sentences with a MOS differences of less than 1.0. Although little may be concluded from these figures, it not being a preference test, from the 320 scores differences given by the subjects, 51 indicate no preference between the two methods EXP & AUTO and 28 give a preference for the method AUTO.

4. CONCLUSIONS AND FUTURE WORK

The method proposed uses entirely automatic learning processes and is both language and speaker independent. Decision trees are used to predict acoustic classes from a symbolic target, in which the foresaid classes are obtained by acoustic clustering of unit speech signals and the decision trees are trained from the unit symbolic element. Units belonging to the predicted classes are retained as candidates for the final selection in the synthesis process.

Although the test subjects give a preference for the handcrafted EXP method, it should nonetheless be noted that the proposed method AUTO also gives rise to unquestionably intelligible speech. The principal fault in the sentences obtained by method AUTO lie in the prosody: incoherent rhythm and melody. On the other hand, spectral discontinuities are rare.

It is unsurprising that the prosody of the generated speech is poorly revived. The acoustic distance used takes the spectral envelope and duration of the signal of each unit into account; the distance does not take the dynamics of the duration or fundamental frequency of the unit signals into consideration. Further work is necessary to derive an acoustic distance capable of reflecting the perceptual similarity between two signals for the acoustic clustering. Such a distance will need to take prosodic differences as well as spectral differences into account.

The problems encountered here may be further overcome by making a better choice in the symbolic parameters used by the decision trees. These parameters should be more orientated towards the prosodic-acoustic phenomena that distinguish between the classes that the trees predict.

Furthermore, the segmentation criterion for the dendrogram of the acoustic clustering is restricted by an ad-hoc limit on the number of members of a class. However, such a criterion is perhaps unjustified and it could be more appropriate to optimise this criterion as a function of a unit test set, based on an acoustic distance from an already known acoustic target.

5. REFERENCES

[1] Black, A.W., and P. Taylor, "Automatically clustering similar units for unit selection in speech synthesis," *Eurospeech*, Rhodes, Greece, pp. 601-604, 1997.

[2] Blouin, C., O. Rosec, P.C. Bagshaw, and C. d'Alessandro, "Concatenation cost calculation and optimization for unit selection in TTS", *IEEE Workshop on Speech Synthesis*, Santa Monica CA, USA, 2002.

[3] Breen, A., and P. Jackson, "A phonologically motivated method of selecting non-uniform units," *ICSLP*, Sydney, Australia, 1998.

[4] Breiman, L., J.H. Friedman, R.A. Olshen, and C.J. Stone, *Classification and Regression Trees*, Wadsworth & Brooks/Cole, Monterey CA, USA, 1984.

[5] Bulyko, I., *Flexible Speech Synthesis using Weighted Pinite-State Transducers*, Ph.D. Thesis, University of Washington, USA, 2002.

[6] Donovan, R.E., *Trainable Speech Synthesis*, Ph.D. Thesis, University of Cambridge, UK, 1996.

[7] Jain, A.K., M.N. Murty, and P.J. Flynn, "Data clustering: a review," *ACM Computing Surveys*, Vol.31(3), pp. 264-323, 1999.

[8] Lawrence, R., and J. Biing-Hwang, *Fundamentals of Speech Recognition*, ed. Oppenheim, A.V., Prentice Hall Signal Processing Series, Englewood Cliffs NJ, USA, 1993.

[9] Nakajima, S.-Y., and H. Hiroshi, "Automatic generation of synthesis units based on context oriented clustering," *ICASSP*, pp. 659-662, New York, USA, 1988.

[10] Quinlan, J.R., *C4.5: Programs for machine learning*, ed. Langley, P., Morgan Kaufmann, San Mateo CA, USA, 1993.

[11] Sakoe, H., and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Trans. Acoustics, Speech, and Signal Processing*, Vol.ASSP-26(1), pp. 43-49, 1978.

[12] Wouters, J., and M.W. Macon, "A perceptual evaluation of distance measures for concatenative speech synthesis," *ICSLP*, Syndey, Australia, 1998.