A STRATEGY TO SOLVE DATA SCARCITY PROBLEMS IN CORPUS BASED INTONATION MODELLING

Valentín Cardeñoso and David Escudero

{valen,descuder}@infor.uva.es
Departamento de Informática
Universidad de Valladolid
47014, Valladolid, Spain

ABSTRACT

Data scarcity in corpus-based intonation modelling for TTS applications is addressed. Multiple model dictionaries are proposed to predict patterns not found in the training corpus. A grouping strategy is proposed to improve models of classes with not enough number of training samples. An experimental study of this strategy shows that better pitch profiles can be predicted this way.

1. INTRODUCTION

Improvements in the naturalness of TTS systems are still an issue, specially because new application fields are emerging which require fast adaptation to new speaker voices and speaking styles and a better set of prosodic rules. Although rule based systems can generate high quality prosody, they are difficult to adapt to changing environments. As far as intonation modelling is concerned, corpus based systems could provide the best engineering solution for these challenges. One of the most important problems of corpus based systems is the scarcity of data available in the corpus. In this work, we provide a strategy to cope with this problem which allows improving intonation modelling following the methodology already presented in previous works [1, 2, 3].

Models of intonation attempt to find out the relationship between a set of linguistic prosodic features (LPFs) of the message and pitch contours (characterised by sets of parameters like in TILT[4] or Fujisaki models[5]). Different approaches to represent intonation and to model the relationship between acoustic and linguistic information can be found in state of the art techniques (as reviewed in [6]). Scarcity problems arise when corpora don't cover all the possible combinations of LPFs (few or no sample for a given combination). Under these situations, it is required to predict intonation for LPFs sets which are not properly modelled if they are at all. Of course, there is still a possibility to redesign a corpus and include more samples. However, it is not allways possible to design and adquire corpora which include the huge number of required combinations. As an example, there are around 27 millions of possible combinations of LPFs reported in [7] (although some of them are obviously absurd) while the corpus used had less than 3000 different combinations.

Despite of its importance, it is not always easy to find references in the bibliography which describe strategies to cope with this problem. Usually, this scarcity is to be avoided delivering better parameter selection procedures for the classification algorithms (neural networks, regression tress, lineal regression, ...) in those special cases where there are not enough training data or anyone at all. In spite of this, there are situations where the number of possible combinations of LPFs which are not properly modelled could represent a high percentage and the naturalness of predicted pitch contours could be highly compromised if a simplistic solution is provided.

In this communication, we will separately face two different problems associated with the scarcity of data: combinations of LPFs with no samples in the corpus and combinations with very few representative samples.

We propose using multiple dictionaries of intonation models to solve the first of these problems. This strategy will be compared with others we already tested in previous works (default pitch patterns [1] and decision trees[3]). We will see that prediction errors can be lowered significantly with the new approach described here.

When the amount of observations of a given class is very limited, the problem is that the associated model won't generalize properly. We propose a class aggregation technique to group together samples belonging to different classes in order to increase the generalization capabilities of the models. We will experimentally show that this aggregation clearly improves the prediction capabilities of the learned models.

We will start summarizing our intonation modelling methodology. Then we describe an experimental procedure to implement multiple dictionaries and the class grouping procedure. In the results section, we will discuss the benefits of applying these strategies. Finally, some conclusions and proposals of future work are presented.

2. INTONATION MODELLING AND REPRESENTATION

The intonation modelling methodology used in this work is schematically depicted in Figure 1. Linguistic analysis of text brings intonation units and provides the set of LPFs associated with each of them. Pitch contours are parameterised using Bézier functions (figure 2). In this way, each intonation unit in the corpus is represented by the set of intonation parameters derived from the control points of the Bézier fitting function and its set of LPFs (see [1] for details).

During the modelling phase, classes of intonation units are built according to some classification criterion and the statistical distributions of the four control points of the Bézier function fit-

This work has been partially supported by Junta de Castilla y León under research grant VA083/03 and by MCYT contract TIC2003-08382-C05-03

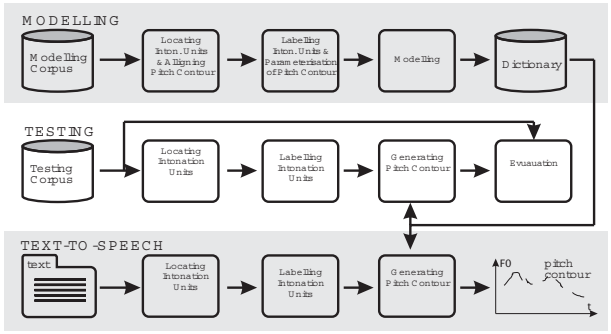

Fig. 1. Intonation modelling and generation of synthetic intonation.

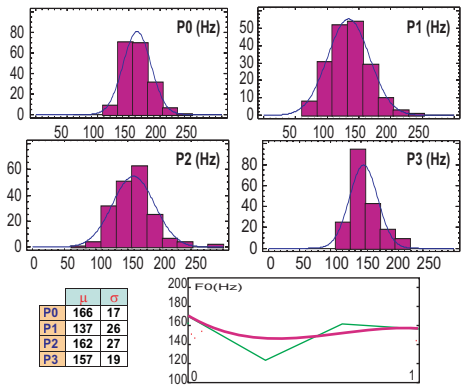

Fig. 2. Statistical model corresponding to one the classes of intonation units in the corpus. Histograms show the distribution of the values of the four control points of the Bézier function. Mean values and standard deviations are gathered in the table and the representation of $F_0(t)$ for the representative pattern (mean values) of the class is drawn beside it.

ting each intonation unit are taken as generative models of the class (see figure 2). We propose to use the set of LPFs as the only grouping criterion of intonation units in a class. Two intonation units are grouped together into the same class if they share the linguistic prosodic features. A dictionary of models of intonation can be built this way. Every combination of LPFs provides an identifier key to lookup a matching model in the dictionary. Each model in the dictionary is represented by the statistical distribution of the intonation profile parameters computed from the corpus for the given class.

In the generation phase, a pitch contour is drawn for each intonation unit, using the statistical distributions of the associated model. Segmentation and labelling of the intonation units is done separately within the linguistic module of the generator. This methodology was showed to be adequate by means of objective and subjective evaluations, which provided intelligibility results comparable to the ones of other approaches and lead to acceptable naturalness qualifications (see [2] for details).

Synthetic pitch contours are generated from the dictionary of classes. Sentences are split into intonation units. We obtain the identifier of the class of every stress group from its LPF. Smooth pitch contours can be synthesized using any of our simulation methods for the control points of the Bézier function. In this communication, we choose the mean pattern of the class to get a smooth pitch contour, since the last aim is to minimize the prediction error.

With this approach, a problem arises when there are not enough samples (or none at all) of a given class in the corpus. In this case, we fail to find a representative model for the class. We focus into this problem in the following sections and describe our proposals to solve it.

3. SOLVING SCARCITY PROBLEMS

3.1. The corpus

We used a corpus designed for concatenative synthesis that contains 4625 stress groups (sequence of syllables within two consecutive stressed words) and 1615 intonation groups (sequence of stress groups within two relevant changes of F0)[1].

The basic intonation unit is the stress group and pitch is evaluated from glottal closing time points. Each stress group is approximated using a third degree Bézier function and the four associated control points are taken as the acoustic intonation parameters of the unit.

The LPFs analysed in this work are inspired in the proposals of previous studies of Spanish intonation carried out by several well recognized authors (a review of the contribution of these authors can be found in [2]). The following features were used: position of the stress group in the intonation group (`posGAenGE`); type of accent (`aceGA`); number of syllables of the stress group (`nSilGA`); number of stress groups in the intonation group (`nGAenGE`); number of syllables of the intonation group (`nSilGE`); position of the intonation group in the sentence (`posGEenFR`); number of stress groups in the sentence (`nGEenFR`). We discarded other LPFs, like symbolic description of the pitch contour trajectory, because we wanted to consider only those features that can be automatically extracted from text. Since there were just a few interrogatives and exclamatives in the corpus, we only used declarative sentences.

In order to get more reliable estimations of the prediction error, our experimental data were made of four different combinations, each one using 75% for modelling and 25% for testing.

3.2. Facing data scarcity

Each combination of LPFs determines one class in the initial dictionary of models. The model of each of the classes is the statistical distribution of the acoustic intonation parameters of the stress groups belonging to the class observed the modelling corpus. If there are few stress groups of certain class, its model will not be characteristic and its use in prediction can be problematic.

To avoid this situation, we propose to iteratively group together pairs of classes. Joining two classes implies creating a new class which includes samples of both of them. A maximum similarity criterion is applied in each step. Thus, grouping two classes implies a precision loss but brings a generalization gain.

To select the candidate classes for grouping, an intra-class similarity metric over the samples is computed. In [2], we described several quality metrics to provide an objective comparison between different classification alternatives, given the corpus. One of such metrics is the sum of the squared error.

[1] Gently provided to us by TALP group of UPC university.

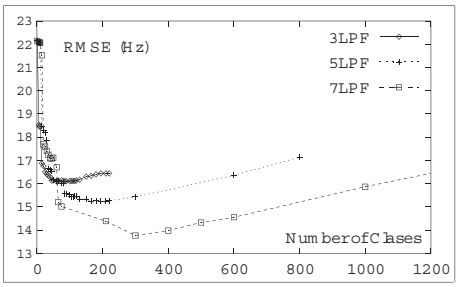

Fig. 3. Prediction error as a function of the number of classes using the 3, 5, and 7 most relevant LPFs, applying the grouping strategy explained in section 3.2.

$$M = \sum_{i=1}^{N_c} \sum_{\bar{P} \in C_i} ||\bar{P} - \bar{\mu}_i||^2 \quad (1)$$

where $\bar{P}$ are the acoustic intonation parameters of the stress groups belonging to class C_i, and $\bar{\mu}_i$ with $i = 1..N_c$ is the mean value vector representing C_i; N_c is the number of classes; $||\bar{P} - \bar{\mu}_i||$ represents the Euclidean distance between vectors $\bar{P}$ and $\bar{\mu}_i$, which gives a self-similarity measurement over the samples of the same class. Once two different classes are merged, a new classification arises and a new value of M can be computed. The candidates to join are the two classes which grouping minimizes the value of M in the new classification. Grouping two classes implies to build a new dictionary. This dictionary can be used to produce synthetic pitch contours. If the prediction error obtained with the new dictionary is smaller than the previous one, then the new classification is better. By repeating the process, we can measure the compromise between precision and generalization obtaining an optimum configuration for the dictionary. The grouping process can be stopped when the loss of precision forces unwanted prediction results.

Three different dictionaries of models will be built using the 3, 5 and 7 most relevant LPFs. The iterative process explained above will be applied separately to the three dictionaries. Doing so, we want to evaluate the effect of the number of LPFs in a compromise between precision and generalization: as the number of LPFs gets higher, the scarcity of samples gets more relevant.

3.3. Empty classes

Some of the classes of the dictionary of models can be void. One class is void if there are no stress groups of such class in the modelling corpus. But, a stress group of any of such void classes can appear when using the dictionary to generate synthetic pitch contours. Then, a strategy is necessary to cope with this situation. In previous works, we have used the typical default pattern solution and the inherent generalization properties of the learning algorithm. Here, we devise an alternative solution which implies using multiple dictionaries.

This strategy implies building several different dictionaries (in our case 1, 3, 5 and 7 LPFs). Each dictionary will use the N most relevant LPFs and all of them are built following the grouping criterion explained in the previous section.

In order to benefit from the use of multiple dictionaries when generating synthetic pitch patterns, we apply the following procedure: most informative dictionary is always used as the first alternative (in this case, the 7 LPFs' dictionary); when a stress group belonging to a void class appears, we recall the dictionary with the higher number of LPFs which classifies the stress group into a non void class. Thus, we ensure that the synthetic pitch contour is associated with the right observations in the corpus, at least partially.

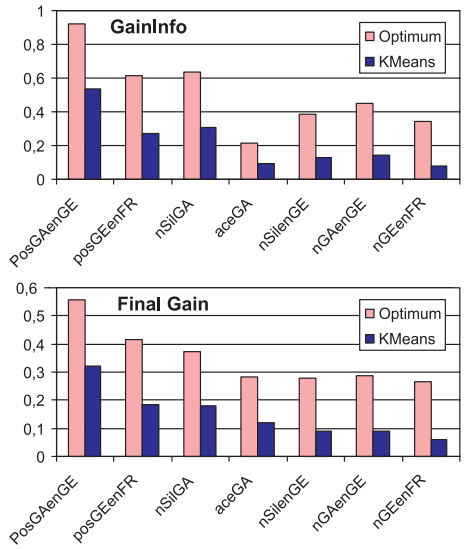

Fig. 4. Ranking the relative importance of the different LPFs. The metrics *GainInfo* and *Final Gain* measure the capacity of the LPFs to justify any given classification of the stress groups of the corpus (see [3] for details). `Kmeans` refers to a classification performed by clustering KMeans K=100. `Optimum` is the classification obtained following the strategy explained in section 3.2 using 7 LPFs and 300 classes.

4. RESULTS

The strategy discussed here will be compared with the use of a default pattern and with the use of a decision tree. The default F0 pattern is the mean vector of acoustic intonation parameters in the modelling corpus. The decision tree is trained with the dictionary of classes. The input of the tree is the LPFs and the output is the identifier of the class in the dictionary. We use the implementation of the algorithm C45 provided with WEKA Tools[2]

The results of these three techniques will be compared in terms of the prediction error. A separate study has been carried out for stress groups associated with void classes and with populated ones.

Figure 3 shows prediction error trends as a function of the number of grouped classes. For a given number of LPFs, an initial dictionary is built with the maximum number of classes. The number of classes is then reduced iteratively grouping them as explained in section 3.2. As can be seen, the prediction error shows a minimum value for a given number of classes. Although a similar behaviour can be found for different numbers of LPFs, it is more pronounced the higher the number of them. An explanation for this is that, before reaching the minimum, the models associated

[2] `http://www.cs.waikato.ac.nz/~ml/weka/`

		Prediction Error: RMSE (Hz)/Correlation		
Classification	% Unseen	Seen	UnSeen	Total
3LPF-DP	1.3	16.07/0.74	17.28/0.58	16.09/0.74
5LPF-DP	9.6	14.89/0.78	18.57/0.69	15.24/0.77
7LPF-DP	17.8	12.99/0.83	17.36/0.71	13.77/0.80
3LPF-DT	1.3	16.09/0.74	17.44/0.56	16.10/0.74
5LPF-DT	9.6	15.42/0.76	18.67/0.68	15.73/0.75
7LPF-DT	17.8	15.41/0.76	18.56/0.68	15.97/0.74
7LPF-MD	17.8	12.99/0.83	15.30/0.78	13.40/0.82

Table 1. Mean prediction errors (RMSE(Hz) / Corr) of the sentences of the testing corpus. *Unseen* is the prediction error of sentences that contain any stress group belonging to a void class in the dictionary. *Seen* is the prediction error of the sentences where all the stress groups have been modelled. % *Unseen* is the percentage of sentences in the testing corpus of the type *unseen*. 3LPF, 5LPF y 7LPF refers to the use of the strategy explained in section 3.2 with 3, 5 or 7 LPFs respectively. DP refers to the use of default patterns. DT refers to the use of a decision tree. MD refers to the use of multiple dictionaries (as presented in this work).

with the classes do not generalize well enough, while after passing through the minimum, the models loose accuracy.

It is remarkable that using a higher number of LPFs does not always guarant better results, unless loss of generalization is not taken into account. Thus, with more than 600 classes, results for 5 LPFs are worse than with 3. This result can be easily explained if we take into account that the higher the number of classes, the lower the number of samples per class and, thus, the less the representation capability of the classes.

Figure 4 compares the ranking of relevance of LPFs obtained here with the ranking obtained in [3]. The value of the relevance is higher now for all features. This means that the new classification reflects better the intonation of the corpus. In relative terms, the ordering in the ranking still remains. Some minor discrepancies in the rankings still arise for the less representative features (`nSilenGE` y `nGAenGE`). This is due to the different number of classes (300 vs 100): the increase of granularity provided by the classification presented here makes these features more relevant.

Table 1 shows the influence of the presence of unseen stress groups on the final results. The main conclusion here is that the use of multiple dictionaries significantly reduces the prediction errors for the unseen configurations and provides better overall results (*Total*) than the other alternatives.

The same default pattern is used in the 3LPF-DP, 5LPF-DP y 7LPF-DP classifications when unseen stress groups are to be predicted. Error values shown in the table are not the same for each of these classifications because they are computed in a sentence by sentence basis. Increasing the number of LPFs has a negative impact since the number of unseen combinations grows but, as a counterpart, it ensures a more accurate prediction of the populated ones.

Comparing the results obtained using decision trees with the ones using default pitch contours, we can conclude that error values are similar in the unseen case, but worse in the case of seen samples. It can also be observed that results get worse as the number of LPFs increases. Incorrectly classified instances cause this: as the number of void classes found in the training stage rises and the number of LPFs inputs to the tree gets higher, the probability of misclassification is bigger. Assigning a wrong class to a stress group might lead to unacceptable prediction error increases.

Informal perceptual tests have been performed to qualify the impact in naturalness of stress groups belonging to void classes. It has been observed that results improve significantly when the method of multiple dictionaries is used.

5. CONCLUSIONS

In this work, we have shown the feasibility of a new strategy based in grouping classes of units of intonation in order to increase the generalization characteristics of the models of intonation obtained from corpus.

A strategy to reduce errors when the units of intonation to predict the pitch belong to a kind that has not been found in the modelling stage has been devised. Results show the benefits of this new approach.

A further step to support our intonation modelling methodology, presented in previous works, has been taken. A thorough testing of alternative units of intonation (not just stress groups) and the impact of more sophisticated parametric representation techniques (instead of control points of Bézier curves) are still to be done in future works. We expect our methodology to properly cope with these challenges too.

6. REFERENCES

[1] D. Escudero and V. Cardeñoso A. Bonafonte, "Corpus based extraction of quantitative prosodic parameters of stress groups in spanish," in *Proceedings of ICASSP 2002*, Mayo 2002.

[2] D. Escudero, C. González, and V. Cardeñoso, "Quantitative evaluation of relevant prosodic factors for text-to-speech synthesis in spanish," in *Proceedings of ICSLP 2002*, Mayo 2002.

[3] D. Escudero and V. Cardeñoso, "Experimental evaluation of the relevance of prosodic features in spanish using machine learning techniques," in *Proceedings of Eurospeech 2003*, September 2003.

[4] P. Taylor, "Analysis and Synthesis of Intonation using the Tilt Model," *Journal of Acoustical Society of America*, vol. 107, no. 3, pp. 1697–1714, 2000.

[5] H. Fujisaki and K. Hirose, "Analysis of voice fundamental frequency contours for declarative sentences of Japanese," *Journal of Acoustics Society of Japan*, vol. 5, no. 4, pp. 233–242, 1984.

[6] A. Botinis, B. Granstrom, and B. Moebius, "Developments and Paradigms in Intonation Research," *Speech Communications*, vol. 33, pp. 263–296, July 2001.

[7] A. Sakurai, K. Hirose, and N. Minematsu, "Data-driven generation of F0 contoures using a superpositional model," *Speech Communication*, vol. 40, pp. 535–549, 2003.

[8] D. Escudero, *Modelado Estadístico de Entonación con Funciones de Bézier: Aplicaciones a la Conversión Texto Voz.*, Ph.D. thesis, Dpto. de Informática, Universidad de Valladolid, España, 2002.