Evaluation of Corpus Based Tone Prediction in Mismatched Environments for Greek TtS Synthesis

P. Zervas[1], G. Xydas[2], N. Fakotakis[1], G. Kokkinakis[1], G. Kouroupetroglou[2]

[1]Electrical and Computer Engineering Dept., University of Patras, Greece
e-mail:{pzervas,fakotaki,gkokkin}@wcl.ee.upatras.gr

[2]Department of Informatics and Telecommunications, University of Athens, Greece
e-mail:{gxydas,koupe}@di.uoa.gr

Abstract

One of the main aspects in Text-to-Speech (TtS) synthesis is the successful prediction of tonal events. In this work we deal with the evaluation of corpus-based models in operational environments other than the training ones. Two pitch accent frameworks derived by linguistically enriched speech data from a generic domain and a limited domain were initially evaluated by applying the 10-fold cross validation method. As a second step, we utilized the cross domains data validation. Due to the heterogeneity of the data, we further employed three machine learning approaches, CART, Naive Bayes and Bayesian networks. The results demonstrate that the limited domain models achieve in average 10% improved accuracy in self-domain evaluation, while the generic models preserve a their performance regardless the domain of application.

1. Introduction

A significant problem in Text-to-Speech (TtS) synthesis is the accurate prediction of pitch accents. The target sequence constitutes part of the specification of a prosody model that will either drive a signal processing module or a unit selection algorithm in the TtS module chain. The task of pitch accent prediction concerns the placement of the appropriate tone label within a synthetic speech utterance, leading to more or less natural-sounding prosody. Errors at this level may impede the listener to understand correctly the synthesized utterance.

Recent works have shown that corpus based prosody modeling can yield natural-sounding prosodic effects in TtS synthesis [1][2][3]. However, their performance heavily depends on the quality and richness of the training data, in terms of linguistic annotations. Moreover, well-designed undersized databases should quantize low-frequency appearances and maintain a more concrete and normalized set of features, preserving a reasonable level of successful prediction. For example, standard prosody models in English voices of Festival speech synthesis system uses only 5 ToBI pitch accents that group all accent varieties [4].

As far as domain specific application is concerned, further improvements can be achieved by exploiting the limited, by the nature of the task, linguistic phenomena. Thus, a more concrete set of analysis data can be built for the prediction of intonational events.

An assortment of algorithms regarding the building of prosodic models have been investigated, including Hidden Markov Models (HMM) [5], neural networks [6], dynamical systems [7], decision trees [8], and ensemble machine learning techniques like bagging and boosting [9]. In this work we focused on testing and evaluating the robustness of pitch accent prediction models based on Naive Bayes, Bayesian networks [10] and CART [11] approaches operating in mismatched environments. Results showed that Bayesian approaches gave small but consistent advantage as regards precision and recall of predicting pitch accent categories for generic domain trained models.

2. Corpora Description

For the purposes of the evaluation, we have used two speech corpora provided by (a) the Wire Communication Laboratory of the University of Patras (WCL), offering the generic corpus and (b) the Speech Group of the University of Athens (UoA), offering the domain specific one. Professional speakers uttered both corpora in Athens dialect.

2.1. Generic Corpus

The WCL generic corpus consists of 5.500 words, distributed in 500 paragraphs, each one of which may be a single word utterance, a short sentence, a long sentence, or a sequence of sentences. For the corpora creation we used newspaper articles, paragraphs of literature and sentences constructed and annotated by a professional linguist. The corpus was recorded under the instructions of the linguist, in order to capture the most frequent intonational phenomena of the Greek language.

The originally annotated feature set of this database consists of: part of speech, shallow syntactic information, number of syllables in word, index of stressed syllable in word, break index, pitch accents and boundary tones [12], [13] and word frequency factor on the basis of a large and disjoint corpus of about 128 Mb of newspaper text. All the features mentioned above were applied to word level.

2.2. Museum Corpus

The UoA museum corpus includes museum exhibit descriptions. It consists of 5380 words, distributed in 516 utterances. The original corpus includes enriched linguistic information provided by a Natural Language Generator. A professional speaker was used to capture the spoken expressions of a museum guided tour. This corpus was cross annotated by three postgraduate computational linguists.

The originally annotated feature set of this database consists of: part of speech, syntactic tree, break index, pitch accents, phrase accents, boundary tones, newness information, mentioned counter, plus other morphological features

extracted by DEMOSTHeNES [14] (syllabic measures concerning the prosodic structure). These annotations have been applied to the syllable level [15].

2.3. Corpora feature set adaptation

As our task was focused on the cross-domain evaluation of pitch accent models, we did not target to the optimization of a feature set for best performance. Our objective was the comparison of the prediction results produced by a common feature set, under generic and limited domains. Previous works have shown the optimized performance of both models using their full feature set [15],[13] and [12] in predicting prosodic phrase breaks, pitch accents and endtones. To facilitate the evaluation of the two prosodic models, we adapted both databases according to this selected set:

1. part of speech
2. shallow syntactic information
3. number of syllables in word
4. index of stressed syllable in word
5. break indices
6. pitch accents
7. boundary tones
8. number of words from previous major break
9. number of words until next major break

The above features were applied to the word level. Both adapted databases were further cross-checked between the two institutions in order to be validated for their annotation consistency. For our experiments the feature set of each word in a window varying from -2+1 to -2+2 words was utilized. The results showed in this paper were obtained from -2,+1 window utilization since it performed better.

2.4. Categories of intonational events

In describing intonational events, we used Pierrehumbert's theory adapted for the Greek language [16]. According to this view, three prosodic constituents at and above the word are significant in Greek intonational structure: the prosodic word (PrWd), the intermediate phrase (ip) and the Intonational Phrase (IP). The PrWd consists of a content word and its clitics, has only one lexical stress, therefore it may bear at most one Pitch Accent in the fundamental frequency (F0) contour. As the frequency of some marks is low in both corpora, we have grouped them, while they can be useful when more data is available. Phenomena like downstep, accented clitics and tonal crowding have been merged to the most appropriate main pitch accent tone categories (e.g. !H* and H*+L have been fused to the H* category).

Table 1: PA Categories Distribution

		L*	H*	L*+H	L+H*	H*+L	UNA
WCL	5500	301	670	1296	291	214	2728
	%	5.47	12.18	23.56	5.29	3.89	49.60
UoA	5380	332	439	1175	976	676	1782
	%	9.23	12.20	32.66	27.12	18.79	33.12

Thus, our tone layer contains 6 pitch accent categories: L*+H, H*, L+H*, L*, H*+L and unaccented (UNA). In table 1 is tabulated the distribution of the grouped pitch accents in both the WCL and UoA corpora.

3. Classification Framework

To tackle the problem of pitch accent prediction we applied the windowed data described above to a decision tree inducer (CART) [11]. Furthermore, we adduce Bayesian analysis regarding the impact certain linguistic attributes pose to the task of correctly identifying the pitch accent categories by considering both the Naive Bayes and Bayesian network probabilistic assumptions [10]. Decision trees have been among the first successful machine learning algorithms applied to predicting pitch accent and prosodic boundaries for TtS [10], [6] and [13]. On the other hand Bayesian methods make robust predictions in cases of missing or low-frequency appearing data. In the following section a brief description of the above approaches is presented.

3.1. Classification and regression trees (CART)

The Regression trees induced by the CART method are a statistical approach for predicting data from a set of feature vectors. In particular, a CART is a binary branching tree with questions about the influencing factors at the nodes and best predicted values at the leaves. CART contains yes/no questions regarding the features and provides either the probability distribution or a mean and standard deviation. Decision trees are obtained by finding the question that splits the data minimizing the mean "impurity" of the partition; while the "impurity" is small when the items are similar. In our experiments, we used the wagon program from the Edinburgh Speech Tools [4].

3.2. Naive Bayes Rule Generator

naive Bayes is a rule generator (classifier) based on Bayes rule of conditional probability. It uses all attributes and allows them to make contributions to the decision as if they were all equally important and independent of one another, with the probability denoted by the equation:

$$P_r[H \mid E] = \frac{Pr[E_1 \mid H] \times Pr[E_2 \mid H] \ldots Pr[E_n \mid H]}{P_r[E]} \quad (1)$$

Where, $Pr[A]$ denotes the probability of event A, $P_r[A|B]$ denotes the probability of event A conditional on event B, E_n is the *nth* attribute of the instance, H is the outcome in question, and E is the combination of all the attribute values.

3.3. Bayesian Networks

A Bayesian network is a special type of diagram (called a graph) together with an associated set of probability tables. Given a set of variables H=$\{H_1,\ldots,H_k\}$, where each variable H_i could take discrete values from a finite set, a Bayesian network describes the joint probability distribution over this set. Formally, a Bayesian network is an annotated Directed Acyclic Graph (DAG) that encodes a joint probability distribution. We denote a network B as the pair B=$<S,P>$ [10] where S is a DAG whose nodes correspond to the variables of H. P refers to the set of probability distributions that quantify the network. S embeds the following conditional independence assumption: "Each variable H_i is independent of

its non-descendants given its parent nodes". *P,* includes information about the probability distribution of a value hi of variable H_i, given the values of its immediate predecessors in the graph, which are also called "parents".

4. Evaluation

Our task was to examine the behavior of the corpus based pitch accent prediction models, in mismatched environments. In order to increase the evaluations' validity we applied three well established machine learning approaches [1],[3]. As a result, six prosodic frameworks were built. The configurations of the training datasets and the corresponding machine learning algorithm applied are shown in Table 2.

Table 2: Prosodic Frameworks

	Train Dataset Domain	ML Method
WCL_NB	Generic	Naive Bayes
WCL_BAN	Generic	Bayesian Net
WCL_CART	Generic	CART
UOA_NB	Museum	Naive Bayes
UOA_BAN	Museum	Bayesian Net
UOA_CART	Museum	CART

Our evaluation plan was divided into two parts. First we measured the systems prediction performance in their training domain by applying 10-fold cross validation. The second experiment was the evaluation of the previously trained models by feeding their inputs with different domain data than those they have been trained. The performance was estimated by using the recall metric per each pitch accent class, as they have been explained in Section 2. Per class recall (R_{class}) is estimated as the number of correctly identified instances of a class (tp), divided by the number of correctly identified instances plus the number of cases the system failed to classify for that class (fn):

$$R_{class} = \frac{tp}{tp + fn} \quad (2)$$

4.1. Evaluating with 10-Cross Validation

The results of our first experiment are depicted in Fig. 1 and 2. We observed that CART models perform well when more data is available, while the Naive Bayes ones are more efficient in cases with few observations. As it was expected, models trained with domain data provided in most cases better results. Particularly in predicting H*+L and H* categories, where the distributions of both databases were almost the same, limited domain data scored double prediction recall. As regards L+H* category, UOA models provided a mean recall among the various machine learning approaches of 57,26% while generic equivalent gave 11,63%.

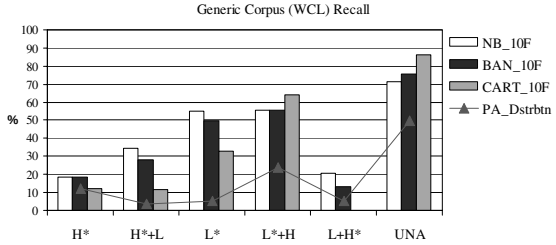

Figure 1: Pitch accent prediction results for generic domain

The total failure of CART approach in predicting L+H* category is due to the sparsity of this tone in the generic training set. All approaches regardless of the training domain performed well in predicting L*+H and UNA categories. The distributions of both categories in our experiment corpora were comparable.

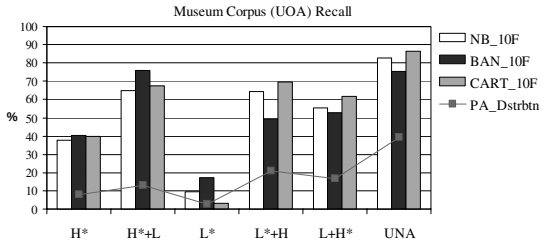

Figure 2: Pitch accent prediction results for specific domain

4.2. Evaluating with Cross Test Sets

The recall of generic and museum models performance with the application of cross data as input is depicted in Fig. 3 and 4.

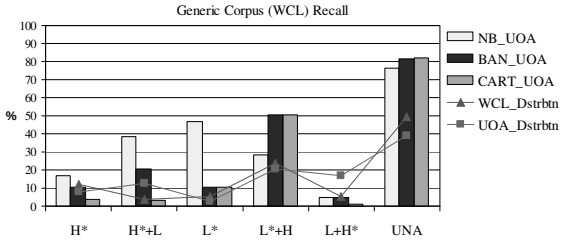

Figure 3: Pitch accent prediction results with museum test set

The task of predicting UNA and L*+H categories had the highest results for both generic and museum models. Performance in prediction of L+H* category was analogous to the number of instances in the training data. As a result, museum models had a 30% mean recall while generic models performance was very low.

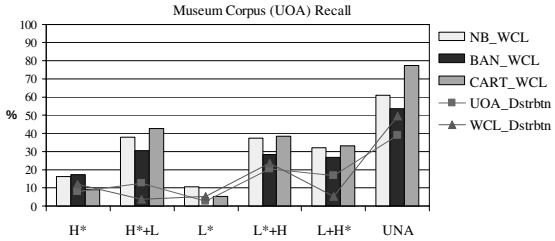

Figure 4: Pitch accent prediction results with generic test set

We have to point out here the robustness of Naive Bayes algorithm for both domain models in the prediction of all pitch accent categories. It performed equally well in cases of missing data such as L+H*, L*, H*+L for generic models and L*, H* and H*+L for museum case.

4.3. Evaluating the Overall Performance

The results from the evaluation regarding total accuracy, average precision and recall for both domain models are depicted in Fig. 5 and 6.

By inspecting Fig. 5 it is clear that that total accuracy of the system decreases with the application of specific domain data with a factor of 12% for Naive Bayes, 10,08% for BAN and 23,03% for CART. However, the average precision and recall are not affected, as they almost preserve they performance, mainly because of the good prediction of L*+H and UNA (which constitutes the 65,78% of the UOA corpus) by the WCL models (Fig. 1).

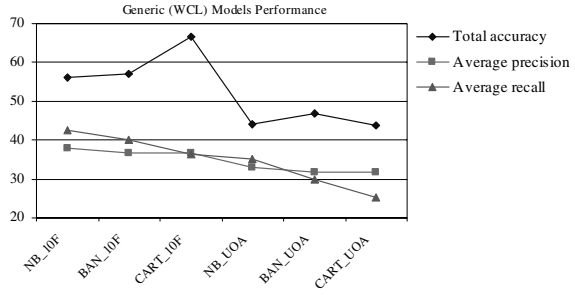

Figure 5: Total accuracy, average precision and recall of generic models.

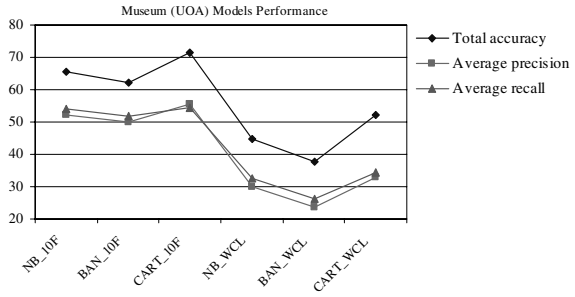

Figure 6: Total accuracy, average precision and recall of museum models.

Regarding the accuracy of the algorithms for the same domain corpus we can see a slightly improved performance for Bayesian networks. As regards precision, it is almost stable for generic domain data and reduced but stable again for limited domain data. Naive Bayes approach seems to have better recall values in both generic and specific PA prediction.

5. Conclusions

In summary, we have described the application of CART and Bayesian learning approaches for the evaluation of intonational models that have been trained with a generic and a specific domain (with museum data) linguistically annotated corpora. It was shown from the evaluation that all algorithms performed equally well in the cases were the domain of the corpus used for testing was the same with the training dataset. As regards cases of missing data, Naive Bayes and Bayesian networks performed better than CART in all evaluation experiments. Regarding cross test set validation, Bayesian approaches gave small but consistent advantage as regards precision and recall of predicting pitch accent categories for generic domain trained models.

6. Acknowledgments

This work was supported by the "Infotainment management with Speech Interaction via Remote microphones and telephone interfaces" - INSPIRE project (IST-2001-32746).

7. References

[1] Dusterhoff K. E., Black A. W. and Taylor P., "Using Decision Trees within the Tilt Intonation Model to Predict F0 Contours", *in Proceedings of Eurospeech'99*, pp. 1627-1630, 1999.

[2] Taylor P. and Black A. W., "Speech Synthesis by Phonological Structure Matching", *in Proceedings of Eurospeech'99*, pp. 623-626, 1999.

[3] Syrdal A. K., Moehler G., Dusterhoff K., Conkie A., Black A. W., "Three Methods of Intonation Modeling", *in Proceedings of 3rd ESCA Workshop on Speech Synthesis*, pp. 305-310, 1998.

[4] The FESTIVAL Speech Synthesis System homepage hhtp://www.cstr.ed.ac.uk/projects/festival/

[5] Conkie, A., Riccardi, G., and Rose, R. C., "Prosody recognition from speech utterances using acoustic and linguistic based models of prosodic event", *in Proceedings of Eurospeech '99*, Budapest, Hungary, pp. 523-526, 1999.

[6] Muller, A.F. and Hoffmann, R. "A neural network model and a hybrid approach for accent label prediction", *in Proceedings of the 4th ISCA Tutorial and Research Workshop on Speech Synthesis*, Perthshire, Scotland, 2001.

[7] Ross, K. and Ostendorf, M. "A dynamical system model for recognizing intonation patterns", *in Proceedings of Eurospeech '95*, Madrid, pp. 993-996, 1995.

[8] Hirschberg, J. "Pitch accent in context: predicting intonational prominence from text", *Artificial Intelligence*, 63:305-340, 1993.

[9] Sun, X., "Pitch accent prediction using ensemble machine learning", *in Proceedings of ICSLP '02*, Denver, Colorado, Sept. 16-20, 2002.

[10] Mitchell T., *Machine Learning*, Mc Graw-Hill, 1997.

[11] Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J., *Classification and Regression Trees*. Chapman Hall, New York, USA, 1984.

[12] Zervas, P., Maragoudakis, M., Fakotakis, N., Kokkinakis, G., "Learning to predict Pitch Accents using Bayesian Belief Networks for Greek Language", *in Proceedings of LREC '04*, (to appear), 2004.

[13] Zervas P., Maragoudakis M., Fakotakis N., Kokkinakis G., "Bayesian Induction of intonational phrase breaks", *in Proceedings of Eurospeech '03*, 8th European Conference on Speech Communication and Technology, Geneva, Switzerland, Sept. 1-4, pp. 113-116, 2003.

[14] Xydas G. and Kouroupetroglou G., "The DEMOSTHeNES Speech Composer", *in Proceedings of the 4th ISCA Tutorial and Workshop on Speech Synthesis*, pp. 167-172, 2001.

[15] Xydas G., Spiliotopoulos D. and Kouroupetroglou G., "Prosody Prediction from Linguistically Enriched Documents Based on a Machine Learning Algorithm", *in Proceedings of the 6th International Conference of Greek Linguistics*, (to appear), 2003.

[16] Arvaniti, A., Baltazani, M., "GREEK ToBI: A System for the Annotation of Greek Speech Corpora", *in Proceedings of LREC '00*, VOL. II, 555-562, 2000.