Learning to predict Pitch Accents using Bayesian Belief Networks for Greek Language

Panagiotis Zervas, Manolis Maragoudakis, Nikos Fakotakis, George Kokkinakis

Wire Communications Laboratory
Department of Electrical and Computer Engineering
University of Patras
26500 Rion, Patras, Greece
{pzervas, mmarag, fakotaki, gkokkin}@wcl.ee.upatras.gr

Abstract

Any text-to-speech (TTS) system that aims at producing understandable and natural-sounding output needs to have a module for predicting prosody. In natural speech, some words are said to be stressed, or to bear Pitch Accents. Errors at this level may impede the listener in the correct understanding of the spoken utterance. Regarding the performance of data driven methods, the scale and quality of the corpus are important. Since there is no suitable corpus available for modeling Modern Greek prosody, we created a corpus consisted of 5.500 words, distributed in 500 paragraphs. In describing pitch accent in particular and intonation features in general, we use Pierrehumbert's theory adopted for MG. In the present study, we try to predict four categories of PA H*, L*, L+H* and unaccented.

1. Introduction

It has been an ongoing topic in science of linguistics as well as in applications to speech synthesis and speech interpretation the association between prosodic variations, semantic, syntactic and discourse features of utterances. Finding rules to associate the prosodic choices speakers make and the structure and meaning of the utterances they generate, concurrently with the context in which they are produced, can help to create more natural sounding synthetic speech and to interpret the full meaning of natural utterances. These associations even when well understood, it is often difficult to acquire the information needed to produce them in real life applications such as text-to-speech or speech recognition systems. Of late however, scientists employ machine learning techniques in order to extract such rules automatically from prosodically-labeled corpora using a finite set of symbols, e.g. ToBI (Silverman et al, 1992).

Human speakers use pitch contours to convey part of the overall meaning of their speech. In varying contour, they are also varying pitch accent choice and placement, as well as deciding how to chunk up' words into levels of prosodic phrasing (Pierrehumbert, 1980). For American English, tones are defined as the phonological abstractions for the target points obtained after broad acoustic stylization (Pierrehumbert, 1981).

In natural speech, some words appear more intonationally prominent than others. Such words are said to be stressed, or to bear pitch accents. Although pitch accent is a perceptual phenomenon, words that hearers identify as accented tend to differ from their deaccented versions with respect to some combination of pitch, duration, amplitude, and spectral characteristics. Pierrehumbert distinguishes only two tones, a high tone (H) and a low tone (L), which is contrasted against each other: H is higher in the speaker's range than L would be in the same place. Sequences of H and L tones are restricted by a finite-state grammar, figure 1, which in turns distinguishes four categories of tones on the basis of their distributional properties: initial boundary tones, pitch accent tones, phrase accent tones, and final boundary tones.

An assortment of algorithms have been investigated for predicting prosodic patterns, including Hidden Markov Model (HMM) (Conkie et al, 1999), neural network (Muller, 2001), dynamical system (Ross & Ostendorf 1995), decision trees (Hirschberg, 1993), and ensemble machine learning techniques like bagging and boosting (Sun, 2002). In the present paper we try to predict pitch accent tones by training Bayesian Networks with a GRToBI (Arvaniti & Baltazani, 2000) annotated text corpus.

GRToBI is a tool for the annotation of Greek speech corpora that encodes intonational, prosodic and phonetic information. It was designed for the variety of Greek spoken in Athens. In terms of design, GRToBI is similar to the original ToBI system for American English, but it has been adapted so that prosodic phenomena requiring special attention in Greek, such as sandhi, can be duly transcribed.

2. Data and Prosodic Annotation

As regards the performance of data driven methods, the scale and quality of the corpus are important. Since there is no suitable corpus available for modeling Modern Greek (MG) prosody, we created a corpus consisted 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. We used newspaper articles, paragraphs of literature and sentences constructed and annotated by a professional linguist. The corpus was uttered under the instructions of the linguist, in order to capture the most frequent intonational phenomena of MG language.

In describing pitch accent tones in particular and intonation features in general, we use Pierrehumbert's theory adopted for MG (Arvaniti, 2000). According to this view, three prosodic constituents at and above the word are significant in MG intonational structure: the Prosodic Word (PrWd), the intermediate phrase (ip) and the Intonational Phrase (IP), figure 1.

The PrWd consists of a content word and its clitics, has only one lexical stress, therefore it may bear at most one Pitch Accent (L*+H, H*, L+H*, L*, H*+L) in the fundamental frequency (F0) contour (only PrWds with

enclitic stress, may bear at most two Pitch Accents). The ip must include at least one Pitch Accent and is tonally demarcated by the presence of a Phrase Accent (H-, L-, !H-) at its right end. The IP must include at least one ip and its tonally demarcated by the presence of a Boundary Tone (L%, H%, !H%) at its right end.

Unlike previous approaches that cope the problem of pitch accent placement as a binary task (given a word form in its context, decide whether it should receive a pitch accent or not), our tone layer contain four categories of pitch accents, H*, L*, L+H* and unaccented. For our study we reduced the pitch accent tones categories to the most frequent and important ones. 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).

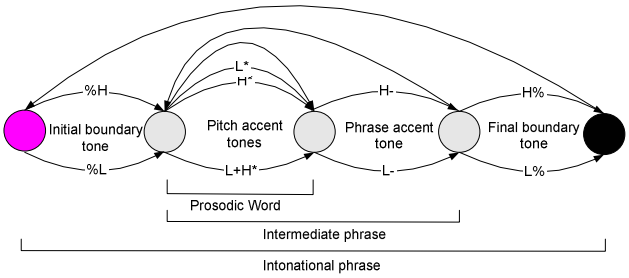

Figure 1: Finite-state grammar for H/L tone sequences

In order to predict the pitch accent tone of a prosodic word, linguistic features were incorporated. Researchers have stressed the important role of syntactic and morphological information for several languages. For our experiments, the POS of the words in an adjustable window varying from -1+2 to -2+2 words was utilized. Taking into account that in real-time pitch accent tones prediction tasks, fully syntactic parsing would be time-consuming and would produce many syntactic trees, as well as that in several languages, including MG, syntactic tools are not freely available, we present a syntactic feature applied to this task, i.e. syntactic phrase boundaries. This information is considered as shallow syntactic information, it is unambiguous and can be extracted rapidly (Stamatatos, 2000). Our feature set for PA prediction also contains the position of the syllable that is bearing the lexical stress, if the word is a prosodic word, the number of syllables and the finally the break index tone. The length of the current sentence measured in syllables was also taken into account for our experiments.

a. Phrase Boundary Detection

Regarding phrase boundary detector, or chunker, is based on very limited linguistic resources, i.e. a small keyword lexicon containing some 450 keywords (articles, pronouns, auxiliary verbs, adverbs, prepositions etc.) and a suffix lexicon of 300 of the most common word suffixes in MG. In a first stage the boundaries of non-embedded, intra-sentential noun (NP), prepositional (PP), verb (VP) and adverbial phrases (ADP) are detected via multi-pass parsing. Smaller phrases are formed in the first passes, while later passes form more complex structures. In a second stage the head-word of every noun phrase is identified and the phrase inherits its grammatical properties

b. Part-Of-Speech Tagging

MG has a complex inflectional system. There are eleven different POS categories. Articles, nouns, adjectives, pronouns, verbs and numerals are declinable while adverbs, prepositions, conjunctions, interjections and particles are indeclinable. For our approach, we used a 2-level morphological analyzer for MG. Table 1 tabulates the POS tags that were used:

POS category	POS-specific features	Common
Adjective (ADJ)	Degree	Gender Number Case
Noun (N)	Common/proper	
Pronoun (PN)	Personal/relative, interrogative, person	
Participle (V)	Sub-category of verb	
Article (ART)	Definite/indefinite	
Numeral (NUM)	Ordinal/cardinal	
Verb (V)	Voice, mood, person, number	
Conjunction (CON)	Coordinating/subordinating	
Preposition(PRE)		
Adverb (ADV)		
Residuals (RES)	Acronym/abbreviation/foreign word	

Table 1: Modern Greek POS labels.

3. Bayesian framework for Pitch Accent Tones Prediction

Our approach is Bayesian given that we start from a causal theory of how the morpho-syntactic features of a sequence affect the intonational attributes of a sentence and reason from observed effects to underlying causes. Our recognizer, similarly to a text-to-speech system, has uncertain a priori knowledge regarding the prediction of a pitch accent tone and during a training phase, it uses evidence provided from partial observations to induce the intonational properties of a newly given case. Furthermore, we adduce Bayesian analysis regarding the impact certain linguistic attributes pose to the task of correctly identifying the pitch accent tones by considering both the naïve Bayes and Bayesian network probabilistic assumptions.

In our approach, we define a probabilistic model for resolving pitch accent tones disambiguation over a search space H*T, where H is the set of possible lexical and labelling contexts {h1,…,hk} or "variables" and T is the set of allowable pitch accent labels {t1,…,tn}. Using Bayes' rule, the probability of the optimal label Topt equals to:

$$T_{opt} = \arg\max_{T \in (t_1,\ldots,t_n)} p(T \mid H) \Leftrightarrow$$

$$T_{opt} = \arg\max_{T \in (t_1,\ldots,t_n)} \frac{p(H \mid T)\, p(T)}{p(H)} \rightarrow \quad (1)$$

$$T_{opt} = \arg\max_{T \in (t_1,\ldots,t_n)} p(H \mid T)\, p(T)$$

Approximations of the probability distributions of (1) deal with the trade-off between computational complexity and efficiency. For a given sequence of observations of variables h1,…,hk, equation (1) becomes:

$$T_{opt} = \arg\max_{T \in (t_1,\ldots,t_n)} p(t_i)\, p(h_1,\ldots,h_k \mid t_i) \quad (2)$$

There are two possible assumptions that can be considered from this point, regarding whether the training features are considered independent of each other or taking into account a specific kind of dependency among all or a subset of them. Returning to equation (2), if we assume that each feature (lexical item) is independent of all others, we adopt the naïve Bayes approach, while in the case of taking into consideration the dependency of them, we apply the Bayesian networks approach.

a. Naïve Bayes

The naïve Bayes classifier is based on the simplifying assumption that the attribute values {h1,…,hk} are conditionally independent given the target value. In other words, the probability of observing the conjunction of attributes given the target value of an instance is just the product of the probabilities for each individual attribute value [8]:

$$P(h_1,\dots,h_k|t_j) = \prod_{i=1}^{k} p(h_i|t_j) \qquad (3)$$

Substituting this into equation (2), the naïve Bayes classifier method is obtained:

$$T_{opt} = \arg\max_{T\in\{t_1,\dots,t_n\}} p(t_j)\prod_{i=1}^{k} P(h_i \mid t_j) \qquad (4)$$

The different terms P(tj) and P(hi|tj) are estimated based on the frequency over the set of available training examples. The set of these estimates is used to classify every new example using equation (4).

b. Bayesian Networks

Given a set of variables H={H1,…,Hk}, where each variable Hi 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 Hi is independent of its non-descendants given its parent nodes". P includes information about the probability distribution of a value hi of variable Hi, given the values of its immediate predecessors in the graph, which are also called "parents". This probability distribution is stored in a conditional probability table. The unique joint probability distribution over H that a network B describes can be computed using:

$$p_B(H_1,\dots,H_n) = \prod_{i=1}^{n} p(H_i \mid parents(H_i)) \qquad (5)$$

The classification task of equation 2 is quite straightforward using Bayesian networks. Applying equation (5) to equation (2), the optimal pitch accent tone T_{opt} equals to:

$$T_{opt} = \arg\max_{T\in\{t_1,\dots,t_n\}} p(t_j)\prod_{i=1}^{k} P(h_i \mid parents(h_i), t_j) \qquad (6)$$

In order to estimate the terms of equation (6), the structure and the parameters of the Bayesian network have to be learned from the training data. Regarding the former, the PC learning algorithm (Heckerman et al, 1995) was applied, while for the latter, we used the EM algorithm (Heckerman et al, 1995). Since learning a Bayesian network is an NP-hard problem (Mitchell, 1997) (i.e. there are 2(n(n-1))/2 possible networks describe n variables), a search strategy had to be followed: initially, the most probable forest-structured network is constructed (i.e. a network in which every node has at most one parent). A greedy search is performed by adding, deleting or reversing the arcs randomly. In case that a change results in a more probable network it is accepted, otherwise cancelled. Throughout this process, a repository of networks with high probability is maintained. When the search reaches a local maximum, a network is randomly selected from the repository and the search process is activated again. It should be noted that in order to avoid the convergence to the previous local maximum the network is slightly modified, meaning that we delete some arcs. Since the training data set is large we also sub-sample the data to speed the network evaluation process up. During the search, the size of the sub-samples is increased. The network complexity is also controlled during the search, so that a limited number of arcs is allowed in the beginning and, as the process progresses, more and more arcs are approved. Recall that given two nodes X and Y of x and y discrete states each, the conditional probability table of a network $X \to Y$ will store at least $x \cdot y$ parameters. It is important to penalize huge tables, corresponding to fully-connected networks, which is the most naïve way of learning.

4. Experimental Results

On the subject of evaluating our Bayesian probabilistic framework, we conducted experiments by applying naïve Bayes and Bayesian networks to varying word junctures and compared the extracted outcome to the performance of the CART algorithm, a machine learning technique that has been previously used with successful results (Hirschberg, et al 1996). The evaluation of the performance was estimated by using the precision, recall and F-measure metrics per each pitch accent tone class, as they have been explained in Section 2. Results were obtained using the 10-fold cross validation method. In figure 2 it is depicted the number of instances of each category of pitch accent tones that are annotated in our prosodic database.

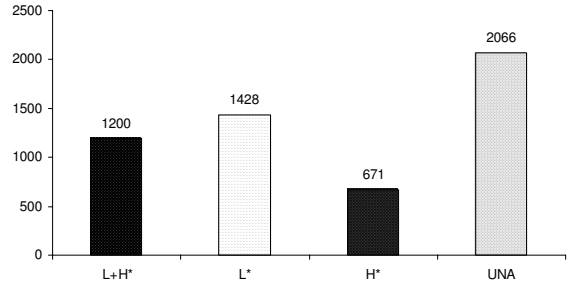

Figure 2: Number of instances of each pitch accent tone category in our database.

Per class precision (Pclass) is defined as the number of correctly identified instances of a class (tp), divided by the number of correctly identified instances, plus the number of wrongly selected cases (fp) for that class:

$$P_{class} = \frac{tp}{tp + fp} \quad (7)$$

Per class recall (Rclass) is 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):

$$R_{class} = \frac{tp}{tp + fn} \quad (8)$$

The F-measure is the harmonic mean of precision and recall, calculated as:

$$F = 1 \Big/ \left(\alpha \frac{1}{P} + (1 - \alpha) \frac{1}{R} \right) \quad (9)$$

where α is a factor which determines the weighting of precision and recall.

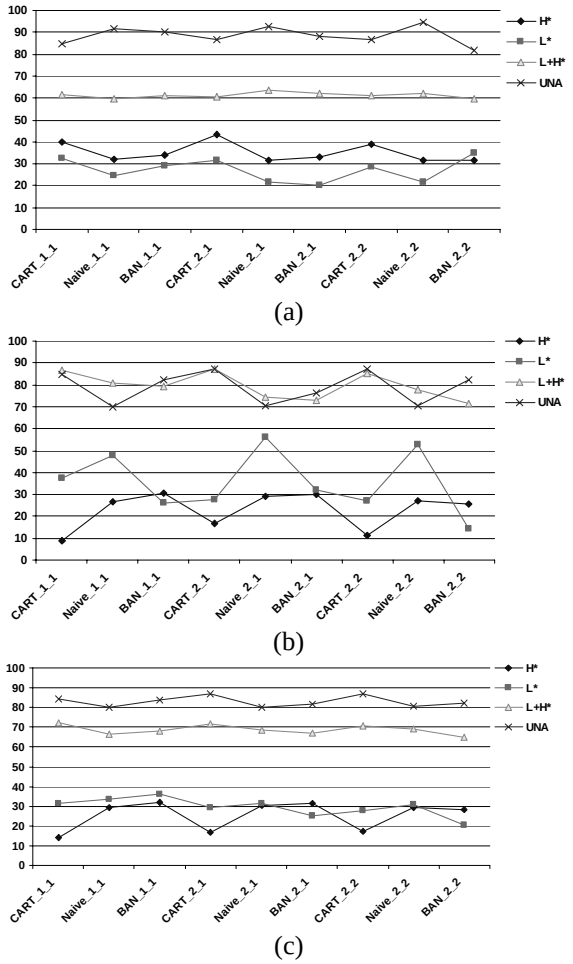

Figure 3: (a) Precision, (b) Recall and (c) F-Measure as a function of window size for each algorithm.

For our experiments we applied three different window sizes. From the results depicted in fig. 3 the following can be concluded. First, Bayesian networks attained the highest F-measure and precision among the other algorithms on inferring the difficult to predict categories such as L* and H* when a [-1,1] window was applied. CART and Naïve Bayes performed well for the [-2, 1] window but it had low precision as regards to the prediction of L* and H* categories. All algorithms revealed good results predicting UNA and L*+H categories. The prediction of L* was the hardest to predict although the number of instances for training were more than L+H* and H*

5. Conclusion

We have described the application of Bayesian learning to pitch accent prediction problem. Naïve Bayes and Bayesian Networks were evaluated against CART algorithm. The evaluation was practiced with the application of different window sizes ranging from [-1,1] to [-2, 2]. Results showed that Bayesian learning can give as good results as CART. Furthermore Bayesian model makes robust predictions in cases of missing or incomplete data (H*, L*, L+H*).

References

Arvaniti, A., Baltazani, M., (2000). GREEK ToBI: A System for the Annotation of Greek Speech Corpora, VOL. II, 555-562, LREC.

Conkie, A., Riccardi, G., and Rose, R. C., (1999). Prosody recognition from speech utterances using acoustic and linguistic based models of prosodic event, Proc. Of Eurospeech, Budapest, Hungary, pp. 523-526.

Heckerman D., Geiger D., Chickering DM., (1995) Learning Bayesian networks: the combination of knowledge and statistical data, Machine Learning, 20, 197-243.

Hirschberg, J., (1993). Pitch accent in context: predicting intonational prominence from text, Artificial Intelligence, 63:305-340.

Mitchell T., (1997). Machine Learning (Mc Graw-Hill).

Muller, A.F. and Hoffmann, R., (2001). A neural network model and a hybrid approach for accent label prediction, Proc. Of the 4th ISCA Tutorial and Research Workshop on Speech Synthesis, Perthshire, Scotland.

Pearl, J., (1988). Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference (San Mateo, CA: Morgan Kaufmann).

Pierrehumbert J., (1981).Synthesizing intonation. Journal of the Acoustical Society of America, 70(4):985-995.

Pierrehumbert, J., (1980), The Phonology and Phonetics of English Intonation, PhD dissertation, MIT, Indiana University Linguistics Club.

Ross, K. and Ostendorf, M., (1995). A dynamical system model for recognizing intonation patterns, Proc. of Eurospeech, Madrid, pp. 993-996.

Silverman K., Beckman M. E., Pitrelli J., Ostendorf M., Wightman C., Price P., Pierrehumbert J., and Hirschberg J., (1992). ToBI: a standard for labeling English prosody. ICSLP-92, volume2, pages 867-870.

Stamatatos, E., Fakotakis N. and Kokkinakis G., (2000). A Practical Chunker for Unrestricted Text. Proceedings of the 2nd International Conference of Natural Language Processing, pp. 139-150.

Sun, X., (2002). Pitch accent prediction using ensemble machine learning, Proc of ICSLP2002, Denver, Colorado, Sept. 16-20.

6. Acknowledgments

The proposed work is supported by GEMINI (IST-2001-32343) EC project.