A data-driven adaptation of prosody in a multilingual TTS

Janez Stergar(✦), Caglayan Erdem(❋), Bogomir Horvat(✦), Zdravko Kačič(✦)

University of Maribor
(✦) Faculty of Electrical Engineering and Computer Science Maribor, Slovenia
(❋) Siemens Corporate Technology, Dept. CTIC 5, 81730 Munich, Germany
(janez.stergar, bogo.horvat, kacic)@uni-mb.si, caglayan.erdem@bmw.de

Abstract

Proper accentuation and phrasing make the syntactic and semantic structure of the message more transparent to the listener. Therefore a good modeling of prosody in a TTS system has to be structured into appropriate levels. The implemented prosodic hierarchy should guide the listeners' attention and help in support of the comprehension process. Since prosody functions as a distractor, it is very important to build the prosody module in a TTS system very carefully. With the goal towards improvements of naturalness a concept of a selective hierarchical approach of prominence disambiguation and symbolic modeling will be introduced. The selective statistically based prominence disambiguation and prediction concept will be discussed and the implementation of the neural network (NN) module for prediction of symbolic tags into a multilingual TTS system introduced. We'll conclude with prediction results and a suitability test of the introduced selective approach based on preliminary acoustical tests performed in a multilingual TTS.

Introduction

Accentuation, as one of the important parameters in speech prosody, concerns the assignment of prosodic prominence. As to the factors governing the assignment of accents to words and the influence of syntactic factors, two positions have been advocated. One is that there is no predictable relation between syntactic structure and the distribution of accents (Bolinger, 1972). The second is, according to Chomsky and Halle (1968) that the distribution of accents in English is completely determined by grammatical properties, and the cases to which this does not apply are anomalous, i.e., are not governed by rules.

For the Slovenian language studies of such kind have not yet been performed and the rule-based approach isn't an appropriate (best-suited) solution to choose (Gros, 2000). Therefore – we decided to design an open environment in disambiguation and prediction of accents based on statistical (acoustical) analysis of a large speech corpus – a data driven approach with learning capability was chosen.

Besides phrasing, prominence – referring to the strength of relation between elements within a given domain – is one of the most important parameters of speech prosody to model. A speaker uses prominence to mark those parts that are important in his message, and the listener uses (the perceived) prominence in order to know which parts are of special interest for the perceived message. Structuring the message is not only the benefit of prominence; applying appropriately varying levels of prominence also increases the naturalness and the comprehensibility of speech (Streefkerk, 2002). On the other hand poor prosody is worse than no prosody, since it functions as a distractor (Terken & Collier, 1998). Our concept therefore is based on hierarchy and selectivity to avoid degradation in the comprehension process and increase the quality (naturalness) of produced synthesized speech. In our opinion it is very important to design prosody modeling into a TTS system based on a hierarchical concept, where the more subtle changes in prominence can be decomposed into appropriate levels.

The so-called corpuses (data) driven approaches nowadays seem to be the appropriate solution for hierarchical prosody modeling. They allow prosodic regularities to be automatically extracted from a prosodic database of natural speech and contribute essential to adaptation processes in a multilingual TTS system.

Data-driven approach

One of the most crucial tasks in data-driven prominence prediction (modeling) is the procedure of labeling the corpus with appropriate symbolic tags. Prosodic labeling based on perceptual tests is very time consuming and prone to inconsistency. Therefore automatically approaches in the labeling processes are favorable.

As automatic approaches usually depend on some manual examination and eventual corrections (verification) it seems to be appropriate approaching the problem of labeling with a semi-automatic method. In our approach of selective prominence labeling we applied a graphical environment, which we already designed for the semi-automatic phrase break labeling (Stergar et al., 2003). The tool was designed to simplify the labeler decisions and support the classification of different classes of breaks.

The corpus

The corpus used for prominence modeling consists of app. 1200 sentences in the Slovenian language (approx. three hours of speech). The selection of the text was emphasized for the broad coverage of sentences in the Slovenian language with the main concern towards the best coverage of concatenation segments.

The audio database recordings were created in a studio environment with a male speaker reading aloud isolated sentences in the Slovenian language (44.1 kHz, 16 bit).

The whole corpus was designed using a selection of clauses from a 31 million word corpus in the Slovenian language. The major parts of the clauses covered daily-published news and Slovenian literature; the minority consisted of clauses taken from Slovenian poetry.

First, sentences not shorter then 15 and not longer than 25 words were preselected from the major corpus. Then, four different text corpora were generated and analyzed statistically (approximately 5000 sentences per corpus). The selection of sentences for the final corpus was based on a two-stage process. In the first stage an analysis based on statistical criteria was performed. In the second stage the final text was chosen based on the results of the first stage. In the final corpus 1200 sentences remained.

The statistically analyzed corpora had similar unit statistics, although the distribution of units was not the same. Three of the four corpora included many foreign names (clauses gathered predominantly from newspapers) that we replaced with Slovenian ones, trying to avoid influencing the statistic of non-uniform units. The corpus with the minimum changes of the non-uniform units after foreign name replacements was chosen as our final corpus (Rojc & Kačič, 2000).

Transcription, segmentation and labeling

The phonetic transcription was managed using a two-step conversion module. The first step is realized with a rule-based algorithm. The second step was designed with a data-driven approach (NN were used) (Rojc & Kačič, 2000).

Pronunciation was derived from the IPA Alphabet. In order to represent the IPA symbols in ASCII characters the SAMPA format was widely used. In our grapheme-to-phoneme conversion module the SAMPA phonetic transcription symbols for the Slovenian language were used (Kačič & Zemljak, 1999).

The text corpus was hand-labeled using 13 different classes of part-of-speech tags (POS). All tags were combined in an environment where tracking and correcting tags was simplified for the labelers (Stergar et al., 2003).

The spoken corpus was phonetically transcribed using HTK. Entities “sil” and “sp” respectively, denoting the silence before and after a sentence and between words were determined with a one-state HMM and all phonemes with three-state HMM in the HTK environment.

The prominence disambiguation algorithm

The current prominence disambiguation algorithm relies on adequate phrasing (phrase breaks disambiguation). The dissection of clauses (phrasing) is performed with a semi-automatic procedure introduced in Müller et al. (2002) and is based on prediction of symbolic phrase breaks tags with POS tags on input of the NN prediction module. The method is essentially based on prediction of phrase breaks based on acoustic preprocessing of the appropriate corpus with HTK as depicted in the simplified work flow diagram of the prominence disambiguation algorithm (STEP1- STEP3 in Figure 1) (Stergar et al., 2003). This tagged corpus was the basics for the data-driven approach used in the prominence prediction module.

Marking and classification of tags represent the first three steps in the procedure (algorithm) of prominence disambiguation (Figure 1). Once the phrase breaks tags have been automatically inserted, the prominent words within marked boundaries (inserted tags) have been disambiguated.

Within a prosodic phrase we automatically determine (the most) prominent words and classify their prominence. Yet two major groups are used for classification. Pitch movements (pitch accent) characterize the first group and the second group is characterized by prominent words emphasized by means of stress (perceptual prosodic accent).

Pitch accent can be reliably detected using the overall syllable energy and some measure of pitch variation (Sluijter & van Heuven, 1996). As this measure can be extracted from pitch changes as from the TILT scheme (Taylor, 2000), the features for the first class have been determined from the interpolated pitch contour using pitch dynamics as the main parameter.

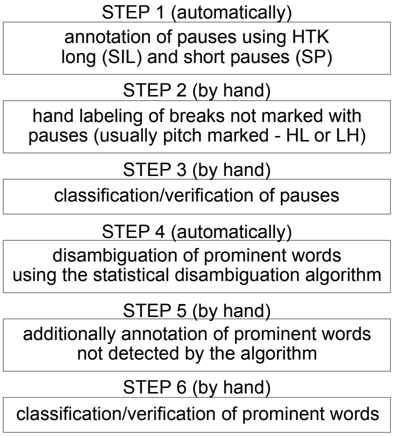
STEP 1 (automatically)
annotation of pauses using HTK long (SIL) and short pauses (SP)
STEP 2 (by hand)
hand labeling of breaks not marked with pauses (usually pitch marked - HL or LH)
STEP 3 (by hand)
classification/verification of pauses
STEP 4 (automatically)
disambiguation of prominent words using the statistical disambiguation algorithm
STEP 5 (by hand)
additionally annotation of prominent words not detected by the algorithm
STEP 6 (by hand)
classification/verification of prominent words

Figure 1: A simplified work flow diagram of the prominence disambiguation algorithm.

A selective prominence disambiguation approach

In the used inventory of prominent words we differentiated two classes of prominence on the word level:

- perceptual prosodic accent (words being emphasized by stress) and
- pitch accent (words being emphasized by pitch movements).

Our aim was the selective detection of both classes automatically. The hand labeling of prominent words of the used database is in progress but is due to a very time consuming process proceeding very slowly.

Pitch dynamics as a measure of prominence

We processed every utterance and computed our measure for pitch changes – pitch dynamics (f_D) – for every syllable (Stergar & Horvat, 2003):

$$f_{D_j} = \sum_{i=1}^{N} \left| x_{i+1} - x_i \right| , \qquad (1)$$

where j is indexing the current syllable and i the concerned samples.

Band pass filtered energy as a measure of prominence

We applied a classical FIR with frequency bounds between 500-2000Hz for band pass filtered energy calculation. Experiments in Tamburini (2002) for Italian and Sluijter & van Heuven (1996) for American English and Dutch, (both for male speaker) showed that this band of high frequencies is the most suitable. Therefore for every utterance we computed RMS of the band pass filtered energy (E_{RMS_B}). E_{RMS} can be computed in many variations, however in our experiments we used the widely used:

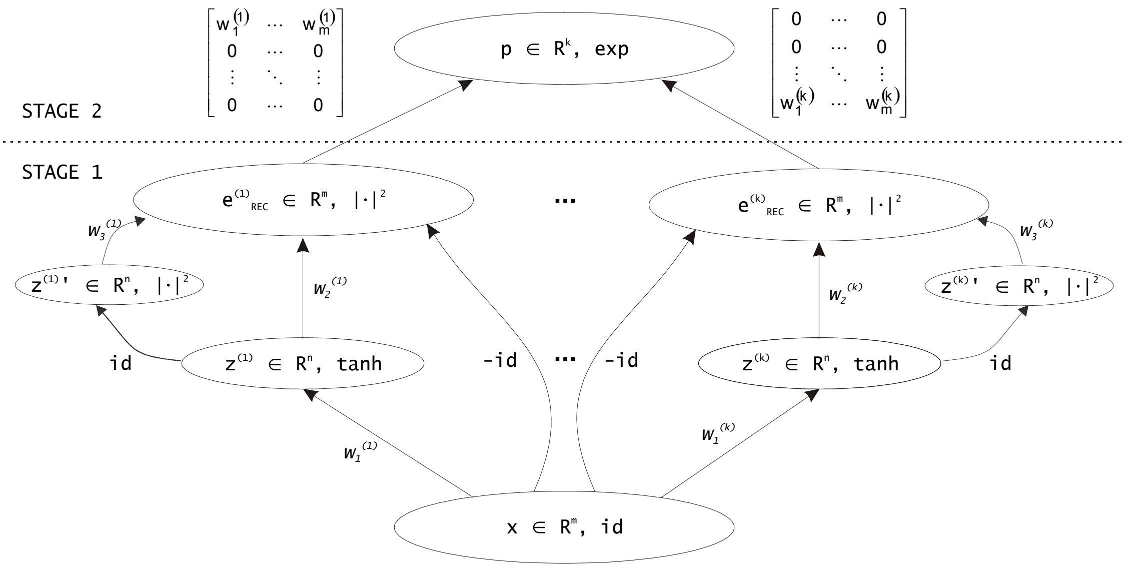

Figure 2: Architecture of the autoassociative NN classifier.

$$E_{RMS_j} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} x_i^2}\ , \qquad (2)$$

where j is indexing the current syllable in the concerning utterance and i the belonging samples.

It is evident (in comparing the distribution of energy and bandpass filtered energy over syllables) that the distance between the energy values being evaluated with the Mahalanobis distance measure in the utterance significantly increases (Stergar & Horvat, 2003).

Statistical selective disambiguation

We used a dynamic threshold value for selective distinction of prominent syllables (words). The line of demarcation for every utterance we used was computed from normal distribution function M(Ψ) using mean value and standard deviation (σ) for f_D and E_{RMS_B} computation respectively:

$$M(\Psi) \geq \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(\Psi_i-\mu)^2}{2\sigma^2}} \quad \mu = 1, \sigma = \sigma_N$$
$$\Psi_i = \begin{cases} E_{RMS_j} \\ f_{D_j} \end{cases} \qquad , \qquad (3)$$

where i is indexing the concerned function, j is indexing the concerned syllable within the sentence N and N indexing σ for the concerned sentence.

NN prediction module

The prediction module we designed is based on a new NN structure based on autoassociative classificators introduced in Müller et al. (2002), (Figure 2). With the used architecture we minimized the problem of unbalanced information flow between the forward and backward path where many inputs are compressed into a single number for classification error. The architecture consists of two stages; STAGE 1 and STAGE 2. The first stage consists of k different autoassociator models for k different classes. Each model is trained only with data from the class it represents. The m-dimensional input vector x is mapped onto a n-dimensional vector z, with n<<m. The NN are trained with the goal that the output vector x' recovers as accurate as possible the original input x. Thus an intermediate representation z of the data in a lower dimensional space is achieved with the compression of x via the matrix $w_1 \in \Re^{n\times m}$ and hence decompression of z via matrix $w_2 \in \Re^{m\times n}$. After training for each autoassociator a reconstruction error e_{REC} is computed. The distance measure ($e_{REC}=(x-x')^2$) is achieved through a squaring activation function of the upper cluster considering the difference between input x and x' achieved using a negative identity matrix -id. The result is a high dimensional error information as input into the classifier.

In STAGE 2 these detailed error information is used to determine which class (model) a given pattern on input x probably belongs to. The classifier is a NN that calculates the class conditional probabilities $p_i = p(x \in class_i)$ from the reconstruction error vectors of the different autoassociator models.

Experiments

First the correlation of partially hand-labeled prominent words in our database with the automatic labeling approach was examined. We compared the overall classification of labeled prominent words with the introduced selective method to one part of the hand labeled database (app. 100 sentences).

After combining the two automatically selected classes (correlation of the two classes was less than 7%) and comparing them to the hand-labeled tags we managed to identify 66% of all prosodic events (prominent/non-prominent words) in the hand labeled database.

Second we conducted some tests with the introduced prominence prediction module for perceptual prosodic accents and pitch accents respectively as well as overall prediction accuracy (Table 1).

We implemented the NN prediction module into a multilingual TTS system to examine the acoustical suitability of our selective disambiguation and prediction framework (Stergar et al., 2004).

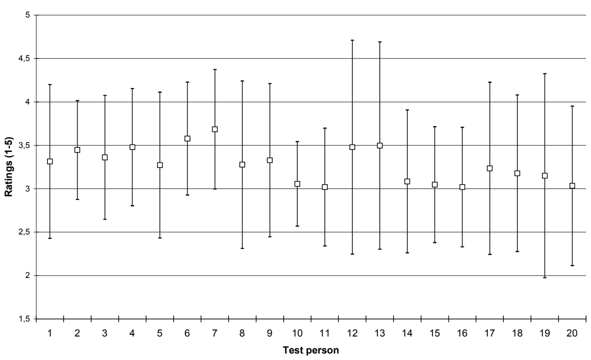

Figure 3: Values and variations of values in the acoustical test ratings per test-person.

Table 1: Overall prominence prediction accuracy (%).

accent type	WP	NP	overall
combined	49,12	80,68	69,17
pitch accent	59,17	69,94	67,95
stress accent	68,33	69,25	69,03

The acoustical results of our adapted multilingual TTS system were presented to a group of 20 non-expert listeners. We generated an inventory of 216 test sentences not used for the training or validation process.

During the listening test each sentence was estimated with marks from 1-5, with 5 denoting the acoustically most pleasant sentence and 1 reserved for unacceptable ones. The test performed during a (approx.) 3-hour session showed that our approach of additional selective modeling of prominence with symbolic tags (using the implemented prosody hierarchy) essentially contributes to naturalness of the synthesized speech. The average ratings (the variances and ratings for each test person are presented in (*Figure 3*) were good-very good.

Conclusion

We introduced a framework for a selective approach in prominence disambiguation and prediction. The prominence disambiguation algorithm was designed on the basis of the automatic phrase break disambiguation. Within the disambiguated boundaries prominent words were marked with a selective statistical approach based on normal distribution and dynamic threshold of preprocessed prosodic parameters: band-pass energy and pitch dynamics. The disambiguation process was dismembered into two steps: disambiguation of perceptual prosodic accents and disambiguation of pitch accents.

We performed experiments in prediction of separated and combined classes with a sophisticated architecture of NN – the autoassociative classificators. The preliminary results for prominence prediction are not so promising, however better results are expected with a nonselective approach when the completely labeled database is available. Also indicated by the experiments is the separate prediction of classes (Table 1).

The overall prediction accuracy is confirmed by the listening tests that were rated with an overall grade of 3,28. It seems that the predicted prominence correlates with the candidates for prominent words (despite not accentuated by the speaker in the training data). Nevertheless this assumption requires some additional manual examination.

We conclude that despite of no state-of-the-art prediction accuracy, rules were extracted with the NN architecture, which enable an accurate prominence prediction for Slovenian language. Additionally no distracting influences to the intelligibility of the synthesized sentences seemed to be perceived.

References

Bolinger. D. (1972). Accent is predictable (if you're a mind reader), Language, vol. 48.

Gros. J. (2000). Automatic Text-to-Speech Synthesis. Linguistica et philologica. ZRC SAZU, Ljubljana.

Terken. J. Collier, R. (1998). Prosody and Intonation in Speech Coding and Synthesis, W. B. Kleijn, K. K. Paliwal edts. Amsterdam: Elsevier.

Kačič. Z., Zemljak, M. (1999). SAMPA - computer readable phonetic alphabet. The WEB portal of Department of Phonetics and Linguistics, University College London (http://www.phon.ucl.ac.uk).

Müller. A. F., Stergar, J., Horvat, B. (2002). Designing prosodic databases for automatic modeling of Slovenian language in a multilingual TTS System. In proceedings of LREC02, Las Palmas de Grand Canaria (Spain), vol. 1, pp. 288-292.

Chomsky, N. Halle, M. (1968). The sound pattern of English. New York: Harper & Row.

Rojc, M., Kačič. Z. (2000). Design of Optimal Slovenian Speech Corpus for use in the concatenative Speech Synthesis System, pp. 321-325. LREC 00, Athens, Greece.

Sluijter, A. & van Heuven, V. (1996) Acoustic correlates of linguistic stress and accent in Dutch and American English. In ICSLP96 (pp. 630-633), Philadelphia, PA.

Stergar, J., Hozjan, V., Horvat, B. (2003). Labeling of Symbolic Prosody Breaks for the Slovenian Language. International Journal of Speech Technology. Vol. 6, No. 3, pp.289-300.

Streefkerk, B. (2002). Prominence. Acoustic and lexical/syntactic correlates. Ph.D. LOT, Graduate School of Linguistics, Netherlands.

Taylor, P. A. (2000). Analysis and Synthesis of Intonation using the TILT Model, JASA. Vol. 107, 3.

Stergar, J., Horvat B. (2003). An environment for word prominence classification in Slovenian language. In proceedings of the ICPhS03, pp. 2087-2090. Barcelona, Spain.

Tamburini, F. (2002). Automatic detection of prosodic prominence in continuous speech. In proceedings of LREC02, pp. 301-305. Las Palmas, Spain.

Stergar, J., Erdem, C., Horvat, B. (2004). Prosody adaptation and modeling in a multilingual TTS system. To appear in Proceedings of Speech Prosody 04. Nara, Japan.