An Environment for Word Prominence Classification in Slovenian Language

Janez STERGAR, Bogomir HORVAT
University of Maribor,
Faculty of Electrical Engineering and Computer Science Maribor, Slovenia
E-mail: janez.stergar@uni-mb.si , bogo.horvat@uni-mb.si

ABSTRACT

Besides phrasing, prominence is one of the most important parameters of speech prosody to model. The so called data driven approaches nowadays seem to be the appropriate solution for prosody modeling in current text to speech (TTS) systems. They allow prosodic regularities to be automatically extracted from a prosodic database of natural speech. In this paper we'll present an evaluation of suitability for automatic word prominence classification of Slovenian language with a hierarchical approach. We classified the prominence of words into two groups, characterized by pitch movements (pitch accent) and stress (stress accent). Pitch movements have been detected from the interpolated syllable pitch contour, while syllable stress was classified from the quantity of energy in the high band of vowel spectra. We'll also present an examination of the correlation between the hand labeled prominent words and the extracted prosody features of the mentioned two classes.

1. INTRODUCTION

Prominence – referring to the strength of relation between elements within a given domain – is one of the most important parameters of speech prosody (besides phrasing) to model. The so-called corpuses driven approaches seem to be the appropriate solution for prosody modeling of the next generation of TTS systems. They allow prosodic regularities to be automatically extracted from a prosodic database of natural speech.

The first step towards data driven prosody prediction is the design of a large corpus labeled with appropriate symbolic prosody labels. The labeling of data can be performed (preferably) manual (skilled experts) or automatically. While automatic labeling can be less accurate than hand labeling, prosodic labeling based on perceptual tests is very time consuming and usually inconsistent. Therefore automatically labeling approaches in the design of databases are preferred.

As automatically approaches usually demand some manual examination and eventual corrections (verification stage) it seems to be appropriate approaching the problem of labeling with a semi-automatic method. As we approached to the labeling of a Slovenian database it was very important to use a concept of labeling suitable for more than only one language.

In this paper we'll present an evaluation of suitability for automatic word prominence classification for Slovenian language with a hierarchical approach using automatic methods. We designed a graphic environment with the goal of semi-automatic phrase breaks as well as word prominence labeling supporting the labeler decisions in different levels of prosodic symbols used for labeling [11]. Within a prosodic phrase we classified the prominence of words into two groups; the first group is characterized by pitch movements (pitch accent) and the second group is characterized by prominent words emphasized by means of lexical stress.

Pitch accent can be reliably detected using the overall syllable energy and some measure of pitch variation. As this measure can be extracted from parameters using the TILT scheme, the features for the first class have been determined from the interpolated pitch contour using some kind of the RFC intonation model and TILT parameters [16]. The main correlates of syllable stress indicated in the literature seem to be syllable duration and high quantity of energy in the high band of vowel spectra where the main formants reside [15]. Therefore the second class of features has been extracted from overall syllable energy and duration in the high frequency domain of major formants. We'll introduce a selective approach of prominent words classification with the goal using it as input to our prominent words prediction module of our TTS system.

2. THE DATABASE

An important step during the adaptation of a TTS system to a new language is the design of a suitable database. The database we used consists of 1206 sentences in the Slovenian language (orthography), which equals approximately three hours of speech. The selection of the text was designed to ensure good coverage of the phones in the Slovenian language; therefore clauses were gathered and included from different text styles (e.g., literature and newspaper texts). These texts were not chosen primarily to achieve the best coverage of prosodic patterns (the aim of the corpus design was suitability for concatenative speech synthesis) therefore no intentional balancing of clause types was performed (declarative – interrogative – exclamations). Also dialogue context and syntax were not

 ISBN 1-876346-48-5

considered, and no semantic analysis was performed since only isolated sentences were included. Semantic prosody was irrelevant for text selection. The whole corpus was determined using a selection of clauses from a 31 million word corpus in the Slovenian language from e-newspapers, e-literature, the WWW or CD's. 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). After the described statistical analysis of the four different text corpora the final corpus was generated. The criterion for the final text filtering was based on monophone, diphone, triphone and fivephone (non-uniform units) richness. Considering comprehension and frequency of units, a careful elimination of sentences was performed. Clauses with poor unit comprehension and unit duplicates were eliminated. In the final corpus 1200 sentences remained [9].

Audio recordings

The audio database recordings were created in a studio environment with a male speaker reading aloud-isolated sentences in the Slovenian language. Every sentence was sampled at 44.1 kHz (16 bit). Because the speaker was a professional radio news speaker, the speech contained no disfluencies (i.e., filled pauses, repetitions and deletions) although for this particular speaker there was some evidence of hesitations in the form of pauses and lengthening. Compared to the German corpus [6] used in [8], the percentage of hesitations differed significantly (<0.5% German, >15% Slovenian).

Phonetic transcription

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 (neural networks were used). The module was designed for the support of two approaches in grapheme–to–phoneme conversion. The first part was intended for those cases in which no morphological lexica were available. The first rule based stress assignment was applied, followed by a grapheme-to-phoneme conversion procedure. The step of stress marking before grapheme-to-phoneme conversion is very important for the Slovenian language, since it very much depends on the type and place of the stress. If the phonetic lexicon is available, a data-driven neural network approach, represented by the second part in the module, can be used. Here, the phonetic lexicon was used as a data source for training the neural networks [9].

The data preparation, generation of the training patterns and the training of neural networks were done completely automatically. The transcription was performed in two steps. In the first step the graphemes were converted into phonemes, and syllable breaks were inserted in the phoneme string. In the second step the stress marks were inserted.

The mapping between graphemes and phonemes was performed by generating training patterns for neural networks (NN) as proposed in [5].

Pronunciation was derived from the IPA Alphabet. In order to represent the IPA symbols in ASCII characters the SAMPA phonetic transcription symbols for the Slovenian language were used [9].

Phonetic segmentation and labeling

The spoken corpus was phonetically transcribed using HTK. Along with standard nomenclature, two special markers were used for pauses between phonemes. "sil" denotes the silence before and after a sentence. "sp" denotes the silence between words in a sentence. Both were determined with a ones-state HMM and all phonemes with three-state HMM in the HTK environment

Accent labeling inventory

We decided to distinguish between word accent, phrase accent and sentence (utterance) accent. Word accent is carried by a word emphasized through perceptual prosodic accent or pitch accent, where phrase accent by our definition is carried by a word most prominent within a phrase comprised of one or more accentuated words. The third accent defined in our inventory is the so-called utterance accent, which is (eventually) carried by a word most prominent in the considered sentence (it is not necessary that a distinction of the utterance accent can be made between words being prominent). The classification of specified accents is a complex matter; therefore, an inventory adequate to distinguish among the three accents was chosen.

However, the following experiments concentrated only on the first category defined in our accent-labeling inventory using two labels (WP, NP). In our inventory a phrase is a sequence of words within major/minor boundaries [11].

Distinction between accented and non-accented words was done within a phrase comparing syllable pitch envelope and normalized syllable mean average pitch changes (normalized on syllable mean average pitch changes for the concerned sentence). Energy and mean energy for syllables in each word were also considered. Through acoustic-visual sessions with a graphic tool also a classification in special cases was made where, depending on the accent type, the accented syllables had low average pitch compared to the sentence average. Word prominence was classified according to four classes similar to those used in [2]:

- EA = Emphatic accent,
- PA = Primary accent,
- SA = Secondary accent, and
- NA = No accent.

We considered primary accent as to be assigned to normally accented words – words perceptibly most prominent within

a phrase (lexical stress). Usually one or more words within a phrase carry a primary accent. We considered the secondary accent to be conveyed by an accentuated word within a phrase not carrying a primary accent. Finally, the emphatic accent is reserved for accented and (lexical) non-accented words that are perceived as extremely stressed relative to other words or are carrying an emphatic function [2], [17].

3. LABELING OF PROMINENT WORDS

We defined two classes of prominence on word level:

- perceptual prosodic accents (words being emphasized by stress) and
- pitch accents (words being emphasized by pitch movements) [7].

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

The acoustic parameters

The first acoustic parameter involved in our experiments was bandpass filtered energy. We used a classical FIR with frequency bounds between 500 – 2000Hz. Experiments in [15] for Italian and [10] for American English and Dutch, (both for male speaker) showed that this band of high frequencies is the most suitable. For every utterance we computed RMS of the bandpass filtered energy (E_{RMS_B}). E_{RMS} can be computed in many variations, however in this study we used the widely used:

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

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

Energy variations across different utterances were reduced with normalizing every syllable with mean syllable energy over the concerning utterance.

The second acoustic parameter was fundamental frequency. As the extraction of the pitch contour is a delicate task we used a successful scheme for f_0 estimation. Therefore we used a robust algorithm for periodicity detection in the autocorrelation domain, suggested by [3] an implemented in Praat [4].

We processed every utterance and computed our measure for pitch changes – pitch dynamics (f_D) – for every syllable:

$$f_{D_j} = \sum_{i=1}^{N} \left| x_{i+1} - x_i \right|, \qquad (2)$$

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

Perceptual Prosodic Accents

The main correlates of syllable stress found in literature seem to be syllable duration and energy [1]. However the studies of [12], [13], [14] showed, that high-frequency emphasis seem to be a more reliable acoustic parameter than just normalized energy by itself for stressed word classification. The presence of a high quantity of energy in the high band of vowel spectra, where the main formants reside, is one of the parameters indicating a strong possibility for syllable stress [15].

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 (Figure 1).

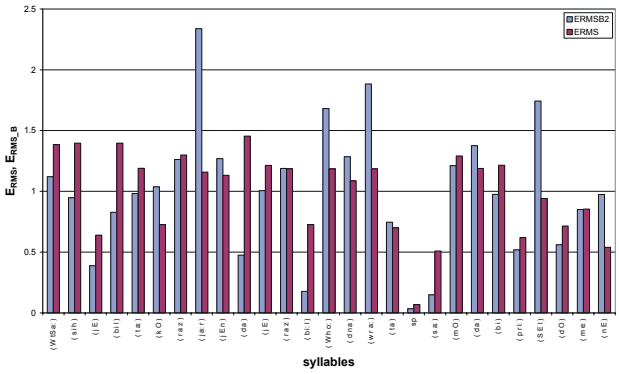

Figure 1: A comparison of E_{RMS} and bandpass E_{RMS_B}.

Therefore we used a simple threshold value for selective distinction of prominent syllables (words). The line of demarcation for every utterance we used, $M(E_{RMS\_B})$, was computed from normal distribution function using mean value and standard deviation (σ):

$$M\left(\tilde{E}_{RMS\_B}\right) \geq \frac{1}{\sigma\sqrt{2\pi}}\, e^{\frac{-(E_{RMS_j}-\mu)^2}{2\sigma^2}} \qquad \mu = 1,\ \sigma = \sigma_M\,, \qquad (3)$$

where M is indexing σ for the concerned utterance.

Pitch Accents

Intonation profiles and accent classification are the major topics of many studies in the last decade [15]. The approaches can be categorized into following categories [18]:

- linguistic systems such as ToBI, which encode events of a linguistic nature and are not so suitable for encoding in automatic systems and
- phonetic systems such as HCLB or INTSINT, which aim only at providing a purely configurational description of the macroprosodic curve without interpretation.

We used the compact RFC representation from the TILT model as one of the parameters (A_{event}) can be considered as

ISBN 1-876346-48-5

a measure of pitch variation [15]. The prominent syllables were automatically classified similarly as the proposed threshold selection for perceptual prosodic accents.

4. RESULTS

We decided to test the correlation of partially hand-labeled prominent words in our database with the automatic labeling approach. Therefore we compared the overall classification of labeled prominent words with the described selective method to a part of the hand labeled database. After combining the two automatically selected classes (correlation of the two classes is less than 7%) and comparing them to the hand-labeled we managed to identify 66% of all prosodic events (prominent/non-prominent syllables) in the hand labeled database (Table 1).

Table 1: Confusion matrix of automatically and hand-labeled prominent words.

	WP	NP	automatically
WP	458	349	807
NP	429	1085	1514
hand-labeled	887	1434	

We also conducted some preliminary tests with our prominence prediction module for perceptual prosodic accents (Table 2):

Table 2: Prediction accuracy of prominent words.

accents	WP	NP	overall
	59,96%	70,38%	67,92%

The preliminary results for prominent word prediction are promising and are expected to improve as a nonselective approach can be applied when the whole database labeled is available. Also the overall prediction accuracy seems to coincide with the covered prominent words in the hand labeled database.

REFERENCES

[1] Bagshaw, P.C. (1993). An investigation of acoustic events related to sentential stress and pitch accents, in English. Speech Communication, 13: 333-342.

[2] Bavarian Archive of Speech Signals. SI1000 (1998). Prosodic Markers Version 1.0 University of Munich, Institute of Phonetics. Munich, Germany.

[3] Boersma P. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. Institute of Phonetic Sciences, University of Amsterdam, Proceedings 17 (1993), pp. 97-110.

[4] Boersma, P. & Weenik, D. (1996). Praat, a system for doing phonetics by the computer. Report 132 of the Institute of Phonetic Sciences, University of Amsterdam.

[5] Hain H.-U. (1999) Automation of the training proce-dure for neural networks performing multilingual grapheme to phoneme conversion. In Proceedings Eu-rospeech 99, vol. 5, pp. 2087-2090. Budapest, Hungary.

[6] Institut für Phonetik und sprachliche Kommunika-tion: Siemens Synthese Korpus - SI1000P, http://www.phonetik.uni-muenchen.de/Bas/.

[7] Müller A. F. Hoffmann R. Accent Label Prediction by Time Delay Neural Networks Using Gating Clusters. Eurospeech01. Aalborg, Denmark. 2001.

[8] Müller A. F., Zimmermann H.G., and Neuneier R. (2000). Robust Generation of Symbolic Prosody by a Neural Classifier Based on Autoassociators. In pro-ceednings ICASSP 00, vol. 3., pp.1285-1288. Istanbul, Turkey.

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

[10] 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.

[11] Stergar J. Horvat B. Labeling of Symbolic Prosody Breaks for the Slovenian Language. International Journal of Speech Technology. Vol. 6, No 3 (May 2003). To be published.

[12] Streefkerk, B M. & Pols, L.C.W. (1996). Prominent accents and pitch movements. In Proceedings of the Institute of Phonetic Sciences (pp. 111-119), Vol. 21, University of Amsterdam.

[13] Streefkerk, B M. (1997). Acoustical correlates of prominence: a design for research. In Proceedings of the Institute of Phonetic Sciences (pp. 131-142), Vol. 20, University of Amsterdam.

[14] Streefkerk, B M., Pols, L.C.W. & ten Bosch, L.F.M. (1999). Acoustical features as predictors for prominence in read aloud Dutch sentences used in ANN's. In Proceedings of the Eurospeech '99 (pp. 551-554), Budapest.

[15] Tamburini F. Automatic detection of prosodic prominence in continuous speech. In LREC02, pp. 301-305. Las Palmas, Spain. 2002.

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

[17] ToporišičJ. (1991). Slovenska slovnica. Založba obzorja Maribor. Slovenija.

[18] Veronis J. Campione E. Towards Reversible Symbolic Coding of Intonation. ICSLP‘98 Sydney, Australia.

ISBN 1-876346-48-5