AUTOMATIC PROSODIC LABELING OF 6 LANGUAGES

Halewijn Vereecken [1], Jean-Pierre Martens [1], Cynthia Grover [2], Justin Fackrell [2] and Bert Van Coile [1,2]

[1] ELIS, University of Ghent, Sint-Pietersnieuwstraat 41, B-9000 Ghent, Belgium
[2] Lernout & Hauspie Speech Products NV, Sint-Krispijnstraat 7, B-8900 Ieper, Belgium

ABSTRACT

This contribution describes a method for the automatic prosodic labeling of multi-lingual speech data. The prosodic labels are word boundary strength and word prominence. The speech signal and its orthographic representation are first transformed to feature vectors comprising acoustic and linguistic features such as pitch, duration, energy, part-of-speech, punctuation, word frequency and stress. Next, the feature vectors are mapped to prosodic labels via a cascade of multi-layer perceptrons. Experiments on 6 different languages demonstrate that combining acoustic with linguistic features yields a better performance than obtainable on the basis of acoustic features alone.

1. INTRODUCTION

It is well known that high quality speech synthesis can only be achieved by incorporating accurate *prosodic models* to detect prosodic phrase structure, to identify phrasal prominence and to determine phoneme durations. The ultimate goal of a prosody module is to improve the naturalness and, to a lesser extent, the intelligibility of synthesized speech. The prosodic models are often derived from large speech databases which are *labeled* both at a *phonetic* and a *prosodic* level. As manual labeling suffers from some major drawbacks, we aim to use automatically labeled databases for that purpose. In this paper we will deal with the *automatic prosodic labeling* of multi-lingual speech data. The automatic phonetic segmentation and labeling (annotation) is dealt with elsewhere [7, 8].

The prosodic events we are concerned with are prosodic phrasing and phrasal prominence. Prosodic phrasing refers to the grouping or separating of words within a sequence of spoken words, and phrasal prominence refers to the relative importance of the words in a prosodic phrase. Following the findings of Portele *et al* [4], and of many others before them (see [4] for an overview), it was established that both phrasing and prominence are *gradual* phenomena. The disjuncture or coherence between two words is expressed by means of a **prosodic boundary strength** (PBS) between 0 and 3: 0 refers to ordinary word boundaries, and values 1, 2 and 3 refer to weak, intermediate and strong boundaries respectively [2]. Phrasal prominence is labeled by assigning to each word a **prominence** (PROM) value between 0 and 9, with 0 being used for words which are not at all prominent and 9 being used for most prominent words.

In the next section we will review two successful approaches to automatic prosodic labeling that have been reported in the literature. Our system, described in section 3, was inspired by these efforts. Basically, the speech signal and its orthography are mapped to a series of acoustic and linguistic features, which are then mapped to prosodic labels using multi-layer perceptrons (MLPs). The acoustic features include pitch, duration and energy on various levels; the linguistic ones are part-of-speech labels, punctuation, word frequency, etc. In section 4, we demonstrate that the linguistic prosodic features are to some extent complementary to the acoustic ones, especially for word prominence. We also show that the prosodic labeling performance is better when the phonetic annotation was done manually, but that the degradation obtained by using an automatic annotation remains sufficiently small.

2. SOME EXISTING SYSTEMS

Often, automatic prosodic labeling is viewed as a standard recognition problem involving first feature extraction and then classification. The feature vector extraction maps the speech signal and its orthography to a time sequence of feature vectors that are ideally good discriminators of prosodic classes. The goal of the classification component is to map the sequence of feature vectors to a sequence of prosodic labels. If some kind of language model describing acceptable prosodic label sequences is included, an optimization technique like Viterbi decoding is used for finding the most likely prosodic label sequence.

This idea is elaborated thoroughly by Wightman and Ostendorf [9]. *Intonational labeling* is performed at the syllable level, with each syllable being marked as either prominent, carrying a boundary tone, both prominent and carrying a boundary tone, or neither prominent nor carrying a boundary tone. In addition, word boundaries are labeled with a 7-scale break index (*break index labeling*). In essence, feature vectors are mapped to posterior probabilities via decision trees, and these are combined with a Markov model of the prosodic label language. The feature vectors in [9] comprise continuously-valued duration, pitch and energy measures, and some categorical features such as a flag indicating whether or not the word was followed by a breath.

The success of the above approach was further demonstrated in the framework of the German VERBMOBIL project (see e.g. [1]). The scope there was to study different reference labels (syntactic-prosodic labels obtained automatically during text generation, hand-marked syntactic-prosodic labels, or the more perceptual prosodic labels), different feature vectors, different classes to distinguish (e.g. combinations of boundary labels, combinations of accent labels, and combinations of boundary and accent labels), different classifiers (MLPs, Gaussian distribution classifiers, polynomial classifiers), as well as different language models (e.g. a 5-gram language model of the orthographic word chain separated by boundary labels). Each feature vector was composed of a large number of acoustic features (duration, pitch, energy) and a few simple linguistic features such as a flag indicating whether or not a syllable carries primary lexical stress. Syntactic/semantic features, if used at all, were mostly used in combination with the *output* of the classifiers.

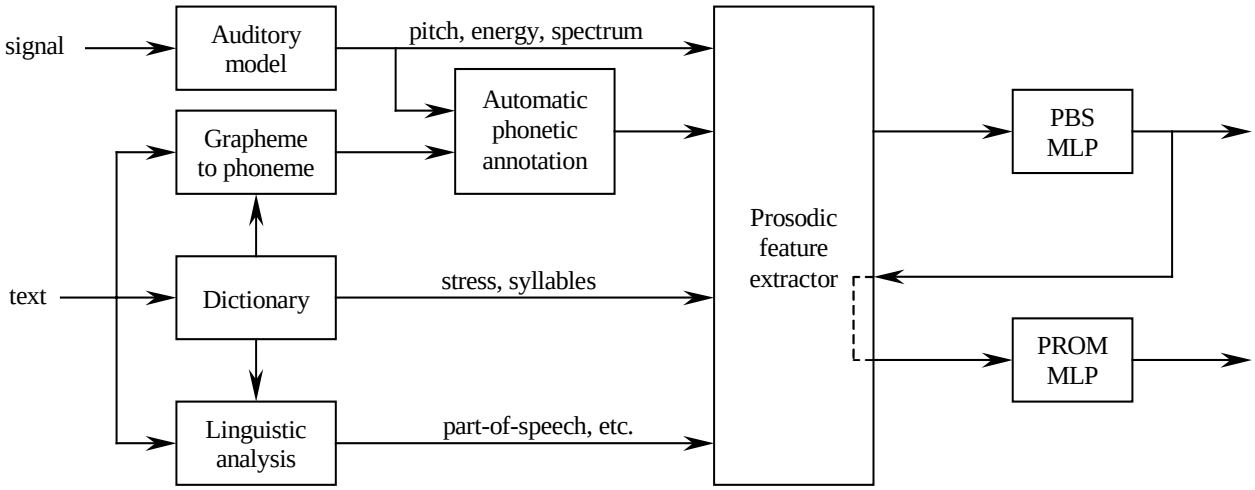

Figure 1: Prosodic feature extraction and classification.

3. SYSTEM OUTLINE

3.1. Introduction

Following the successful approaches described above, we also view automatic prosodic labeling as a recognition problem. We could not find a prosodic label language model that caused a sufficiently large reduction in perplexity to justify the increased complexity due to the need for a Viterbi decoder. Therefore we have decided to skip the language model in our present system. The prosodic labeling is thus reduced to a 'static' classification problem, involving feature extraction and classification (Fig. 1).

3.2. Feature Extraction and Classification

For the purpose of obtaining *acoustic* features, the speech signal is analyzed by an auditory model [6]. The corresponding orthography is supplied to the grapheme-to-phoneme component of a TTS system, yielding a phonotypical phonemic transcription. Both the transcription and the auditory model outputs (including a pitch value every 10 ms) are supplied to the automatic phonetic annotation tool, which is described in detail in [7, 8]. The phonetic boundaries and labels are used by the prosodic feature extractor to calculate pitch, duration and energy features on various levels (phone, syllable, word, sentence). A linguistic analysis is performed to produce *linguistic* features such as part-of-speech information, syntactic phrase type and word frequency. Syllable boundaries and lexical stress markers are provided by a dictionary. Both acoustic and linguistic features are combined to form one feature vector for each word (PROM labeling) or word boundary (PBS labeling).

The classification component of the prosodic labeler starts by mapping each PBS feature vector to a PBS label. Since phrasal prominence is affected by prosodic phrase structure, the PBS labels are used to provide phrase-oriented features to the word prominence classifier. Both classifiers are fully connected MLPs of sigmoidal units. The PBS MLP has 4 outputs, each one corresponding to one PBS value. The PROM MLP has 1 output only. In this case, PROM values are mapped to the [0:1] interval. The automatic labels are rounded to integers.

3.3. Prosodic Features

Trying to come up with a compact set of features that is suitable for every speaker and language is an endless task. The cues for signaling boundaries and accents are to a certain extent *speaker dependent*. The speaker has at her or his disposal a number of different cues, such as presence/absence of pauses, pause length, pitch contour position, pitch excursion, amplitude, syllable duration. A good speaker will use all of these cues in a regular fashion; a poor speaker will use only some of them and irregularly. Consequently, one has to adjust what counts as a strong and a weak boundary/accent depending on the speaker. What is more, the cues for signaling boundaries and accents are definitely *language dependent*. It is, for example, not realistic to expect that French boundaries are signaled in the same way as Dutch ones. The word-initial glottal stops so common in Germanic languages will not be present in Romance languages to break up the speech stream. To get around these difficulties we simply extract as many features as we can possibly think of, and leave it to the classifier to determine the relevant ones for a particular speaker and language. The reader is referred to [1, 4, 9] for a phonetic and empirical motivation of the features.

Acoustic PBS Features. Similar to Batliner *et al* [1] our acoustic features describe the speech segments representing the three words (syllables) preceding and the two words (syllables) following the boundary to classify. These 5 words and 5 syllables are characterized by 110 parameters representing the pitch and energy contours over the corresponding speech intervals, and by 18 parameters providing durational information. Also provided are the durations of the pauses surrounding the preboundary word, and of the glottal closure of the postboundary word. Pitch and energy values are normalized with respect to the mean pitch and energy of the sentence. Segmental durations are normalized per phone type by subtracting the mean and dividing by the standard deviation of the type [9].

Note also that we do not use any information concerning breaths (contrary to [9]). In fact, breaths were replaced by pauses prior to the prosodic feature extraction with the purpose of improving the automatic phonetic annotation (see [7]).

If a word starts with an unvoiced stop, there is some discussion as to whether the silence preceding the burst is just a closure or an intended pause. Similar to Sanderman ([5], p. 17), such silences are labeled pauses if they are longer than 100 ms.

Linguistic PBS Features. For each word boundary, the following linguistic features are derived for the preboundary word: number of words and position of the word in the sentence, distance in words to the previous and next punctuation, type of the next punctuation (e.g. colon, period,…), word frequency, number of syllables, letters and capital letters in the word, position of the primary stress relative to the word start and ending, most likely part-of-speech of the word, accentability of the word (something like the content/function word distincton), etc. This yields about 70 features, depending on the language (not all languages have the same number of part-of-speech labels). Since syntactic phrase types were not yet available for all languages, we have not incorporated them so far.

Acoustic PROM Features. Prominence features were derived by assuming that the perception of word prominence is triggered by the primary stressed syllable. Contrary to what was found for PBS, it did not help to include acoustic features describing the words surrounding the one to be scored. We thus restricted to the acoustic PBS features describing the word, its stressed syllable, the pauses surrounding the word, and the glottal closure of the succeeding word.

Linguistic PROM Features. Same as the PBS features.

Additional PROM Features. As indicated before, the PBS labels provide additional features, such as the PBS before and after the word, and the position of the primary stressed syllable in the prosodic phrase. Prosodic phrases are obtained by interpreting the highest PBS values as phrase breaks.

4. EXPERIMENTAL EVALUATION

4.1. Prosodic Databases

We evaluated the prosodic labeling tool on 6 databases corresponding to 6 different languages: Dutch, American English, French, German, Italian and Spanish. Each database contains about 1450 *isolated* sentences representing about 140 minutes of speech. The sentences include a variety of text styles, syntax patterns and sentence lengths. The recordings were made with professional native speakers (one speaker per language). All databases were carefully hand-marked on a prosodic level by a native or near-native labeler. Further details on the design of these corpora are given in [3].

4.2. Prosodic Labeling Results

In this section we present labeling performances using (1) only acoustic features, and (2) acoustic plus linguistic features. Results using only linguistic features will not be presented here since they boil down to prosodic *modeling* and not *labeling*.

Each database is partitioned into a training set (75%), a cross-validation set (10%) and a test set (15%). The labeling performance is measured by calculating on each data set the correlation, mean square error and confusion matrix between the automatic and the hand-marked prosodic labels. Hence, we measure the performance of the automatic labeler against a human labeling of the same data. The error-backpropagation training of the MLPs proceeds until a maximum performance on the cross-validation set is obtained. Both MLPs have 1 hidden layer; the number of hidden nodes is chosen so that the number of parameters in each MLP is about 2000.

The test set results for PBS and PROM are shown in Tables 1 and 2 respectively. Since the database contains sentences, the PBS predictions apply only to within-sentence boundaries. As the majority of the word boundaries have PBS=0, we have also included the performance of a baseline predictor always yielding PBS=0. There appears to be a correlation between the complexity of the task (measured by the performance of the baseline predictor) and the labeling performance. For PROM we give the exact identification ± 1, so as to compare PBS and PROM labeling accuracy on a similar scale.

Adding linguistic features does improve the prosodic labeling performance significantly. The PROM labeling is improved dramatically; the improvements for PBS are smaller, but taken as a whole they are significant too. Hence, there seems to be some vital information contained in the linguistic features. This could indicate that the manual labelers were to some extent influenced by the text, which is of course inevitable. We can not compare our results with those mentioned in the literature as nearly everybody utilizes different corpora and different prosodic labels.

Language	'PBS=0'	AC	AC + LI
Dutch	70.1	76.4 (0.79)	78.4 (0.82)
American	60.5	74.6 (0.79)	70.2 (0.72)
French	75.2	77.4 (0.74)	78.7 (0.78)
German	70.0	79.0 (0.84)	81.7 (0.87)
Italian	79.6	87.7 (0.88)	88.5 (0.90)
Spanish	86.9	91.6 (0.84)	92.6 (0.86)

Table 1: PBS labeling performance of the baseline predictor (PBS=0), an MLP labeler using acoustic features (AC) and an MLP labeler using acoustic plus linguistic features (AC+LI): exact identification (%) and correlation.

Language	AC	AC + LI
Dutch	79.1 (0.81)	80.6 (0.82)
American	69.7 (0.82)	76.7 (0.87)
French	76.1 (0.75)	81.7 (0.81)
German	73.6 (0.80)	79.1 (0.84)
Italian	74.6 (0.80)	84.1 (0.89)
Spanish	80.2 (0.83)	92.6 (0.92)

Table 2: PROM labeling performance: exact identification ± 1 (%) and correlation.

4.3. Influence of Phonetic Annotation

It is obvious that the underlying phonetic annotation is of crucial importance to the prosodic labeling performance. If phone boundaries are wrong, so are the acoustic features derived from these boundaries: pitch, duration and energy will be calculated over the wrong signal parts. In order to assess the influence of

the quality of the phonetic annotation, about 20 minutes of speech of three language database were manually segmented and labeled on a phonetic level. For these 20 minutes, the phonetic segmentation and labeling supplied to the feature extractor consisted of either (1) the hand-marked segments and labels, (2) the labels and segments emerging from the automatic annotation system when supplied with the hand-marked labels, or (3) the labels and segments emerging from the automatic annotation when supplied with the phonotypical phonemic transcription. The prosodic labeler was trained on 15 minutes and tested on the remaining 5. No cross-validation was used this time; instead, performance was optimized on the test set. *Only acoustic features* were used, together with 2 simple lexical features (for PBS: number of syllables in the words before and after the boundary; for PROM: number of syllables in the word and position of the primary stressed syllable in the word). The test set results are given in Tables 3 and 4.

Language	**(1)**	**(2)**	**(3)**
American	76.4 (0.84)	76.8 (0.84)	77.4 (0.83)
French	76.7 (0.76)	74.6 (0.75)	71.4 (0.72)
Spanish	95.6 (0.93)	94.2 (0.92)	94.0 (0.92)

Table 3: PBS exact identification (%) and correlation for different phonetic annotations supplied to the prosodic feature extractor. Situations (1), (2) and (3) are explained in the text.

Language	**(1)**	**(2)**	**(3)**
American	74.4 (0.81)	76.4 (0.81)	71.3 (0.79)
French	78.6 (0.81)	74.7 (0.80)	70.7 (0.77)
Spanish	84.1 (0.85)	82.7 (0.84)	82.7 (0.83)

Table 4: PROM exact identification ± 1 (%) and correlation for different phonetic annotations supplied to the prosodic feature extractor.

Given the small amount of test examples, the differences between situations (1), (2) and (3) are not significant (except for French PROM). As a rule, the prosodic labeling improves as the phonetic annotation gets better. When comparing the automatic annotation with the manual one for situation (3), we can see that on average 3% of the manual phone boundaries is omitted, 5% of the automatic boundaries is inserted between 2 manual boundaries, 13% of the automatic boundaries differs by more than 20 ms from the corresponding manual boundary, and 7% of the automatic phone labels differs from the manual ones (see [8]). Our acoustic features turn out to be rather robust against these errors. Perhaps the boundary deviation error criterion of 20 ms used for evaluating an automatic phonetic annotation is too sharp. We argue that one should not manually correct the phonetic annotation prior to the prosodic labeling, but correct the automatic prosodic labels instead.

Situation (3) was used for obtaining Tables 1 and 2. One may notice that the labeling results in Tables 1 and 2 (column AC) are sometimes worse than those mentioned in Tables 3 and 4 (situation 3), even though less training data was available in the latter experiments. The reasons for this are that (a) the 140 minutes are more complex than the 20 minutes subset, (b) the feature sets differ slightly, and (c) we did not use cross-validation for obtaining Tables 3 and 4.

5. CONCLUSION

In this paper, a system that automatically labels prosodic boundary strength and word prominence is described. The system comprises a feature extractor and a cascade of multi-layer perceptrons. Boundary strength and word prominence are labeled on a gradual scale. Our feature vector not only comprises acoustic features (pitch, duration and energy), but a whole series of linguistic cues as well (e.g. part-of-speech labels, word frequency, punctuation, stress,…). We have evaluated our system on 6 different languages. The labeler achieves approximately the same performance for each of the languages. By comparing the prosodic models extracted from the automatic and the manual labels respectively, it will be possible to evaluate the validity of our automatic labeling strategy.

6. ACKNOWLEDGMENT

This research was performed with support of the Flemish Institute for the Promotion of the Scientific and Technological Research in the Industry (contract IWT/AUT/950056).

7. REFERENCES

1. Batliner, A., Kompe, R., Kieβling, A., Mast, M., Niemann, H., and Nöth, E. "M = Syntax + prosody: a syntactic-prosodic labelling scheme for large spontaneous speech databases", *to appear in Speech Comm*.

2. Grover, C., Heuft, B., and Van Coile, B. "The reliability of labeling word prominence and prosodic boundary strength", *Proceedings ESCA Workshop on Intonation*, 165–168, 1997.

3. Grover, C., Fackrell, J., Vereecken, H., Martens, J.-P., and Van Coile, B. "Designing prosodic databases for automatic modelling in 6 languages", *to appear in Proceedings ESCA Synthesis Workshop*, 1998.

4. Portele, T., and Heuft, B. "Towards a prominence-based synthesis system", *Speech Comm.* 21, 61–72, 1997.

5. Sanderman, A., *Prosodic phrasing*, Ph. D. dissertation, IPO, Eindhoven, The Netherlands, 1996.

6. Van Immerseel, L., and Martens, J.-P. "Pitch and voiced/unvoiced determination with an auditory model", *J. Acoust. Soc. Am.*, Vol. 91, No. 6, 3511–3526, 1992.

7. Vereecken, H., Vorstermans, A., Martens, J.-P., and Van Coile, B. "Improving the phonetic annotation by means of prosodic phrasing", *Proceedings Eurospeech*, 179–182, 1997.

8. Vorstermans, A., Martens, J.-P., and Van Coile, B. "Automatic segmentation and labelling of multi-lingual speech data", *Speech Comm.* 19, 271–293, 1996.

9. Wightman, C., and Ostendorf, M. "Automatic labeling of prosodic patterns", *IEEE Transactions on Speech and Audio Processing*, Vol. 2, No. 4, 469–481, 1994.