FROM THE PROCEEDINGS OF THE WORKSHOP ON MULTI-LINGUAL SPEECH COMMUNICATION, PAGES 110–115

A MULTILINGUAL PROSODY MODULE IN A SPEECH–TO–SPEECH TRANSLATION SYSTEM

E. Nöth *A. Batliner* *J. Buckow* *R. Huber* *V. Warnke* *H. Niemann*

Friedrich–Alexander–Universität Erlangen–Nürnberg,
Lehrstuhl für Mustererkennung (Informatik 5),
Martensstr. 3,
91058 Erlangen, Germany
noeth@informatik.uni-erlangen.de
http://www5.informatik.uni-erlangen.de

ABSTRACT

In our previous research, we have shown that prosody can be used to dramatically improve the performance of the automatic speech translation system VERBMOBIL [16]. The methods to classify prosodic events have been developed on the German subcorpus of the VERBMOBIL speech database. In this paper we describe how the methods that we developed on the German subcorpus can be applied to other languages. Experiments show that these methods are suited for English and Japanese, as well. Efficiency problems are addressed and a new set of features is presented. The new set of features facilitates a multilingual module for prosodic processing. We present an architecture for such a multilingual module and discuss the advantages of this approach compared to an approach that uses separate modules for different languages. This multilingual module and the new feature set are evaluated w.r.t. computation time, memory requirement, and classification performance. The results show that the memory requirement can be reduced by 78%, whereas the recognition accuracy does not decrease.

1. INTRODUCTION

The research presented in this paper was conducted as part of the VERBMOBIL project. The VERBMOBIL system translates spontaneous human–to–human appointment scheduling dialogues [21]. During the translation process prosodic information is used at various stages. Phrase boundaries, phrase accents, and sentence mood are used to guide syntactic parsing and disambiguate between several possible meanings [2, 17, 13, 11, 20]. Irregular boundary markers are used to deal with corrections [19]. Furthermore, some preliminary emotion detection is integrated in order to improve the system behavior in the case of errors [3].

In VERBMOBIL the output of a word recognizer is structured as a word hypotheses graph (WHG). Every edge represents a word hypothesis and every path through the graph a possible acoustic–phonetic interpretation of the observed utterance. The edges in the graph are marked with start and end time, thus making it possible to determine the corresponding segment of the speech signal. In order to make prosodic information available, each edge in the WHG is enriched with probabilities for prosodic events. The probabilities are determined in a classification process. For every word hypothesis, prosodic features are extracted from the speech signal (see Section 3) and used as input to multi layer perceptrons (MLP) for each prosodic event. The output of an MLP can be interpreted as an *a–posteriori* probability [8].

As the importance of prosody for the system performance could be shown on a German subcorpus of the VERBMOBIL data [16] we investigate the applicability of our approach for the other VERBMOBIL languages. In [16] a time alignment of the phoneme sequence of the recognized words was necessary to perform a phone intrinsic normalization of energy and duration features. A phone intrinsic normalization is important because individual phonemes are affected differently by a change in speaking–rate or loudness [22, 7, 12, 4].

The normalization has some draw–backs, though, specifically if it is used for several languages simultaneously in one software system. First, in order to compute the time alignment of the phoneme sequence acoustic models for the phonemes of each language have to be trained and used. This requires a large amount of memory. Second, a Viterbi alignment of the phoneme sequence is expensive in terms of computational effort. Third, the features based on phoneme intervals are very sensitive to errors in the time alignment. Thus, we focus on how to overcome these draw–backs and describe a set of features (Section 3) and a system architecture (Section 5) which allow fast and robust multilingual prosodic processing.

We show that with the new set of features and a multilingual system architecture better classification results can be achieved than with the old features and three monolingual modules. At the same time, the memory requirement and computation effort can be reduced significantly (Section 4.1 and 4.2 and 5). Before we look at the feature extraction we give a few examples of how and why prosody is used in VERBMOBIL.

2. PROSODY AND DIALOGUE

Dialogue processing in VERBMOBIL is very complex and prosody is used at various stages during the translation process [2, 17, 13, 11, 20]. Thus, we can only give a few examples of how prosody is used. The word recognition components of the VERBMOBIL system produce lattices of word hypotheses as shown in Figure 1. These lattices are used as input for the modules of the linguistic processing. Important prosodic information in the context of syntactic/semantic parsing is:

1. Which words of an utterance carry a phrase accent?
2. Where in an utterance are prosodic boundaries?

*This work was funded by the German Federal Ministry of Education, Science, Research and Technology (BMBF) in the framework of the VERBMOBIL Project under Grant 01 IV 102 H/0. The responsibility for the contents lies with the authors.

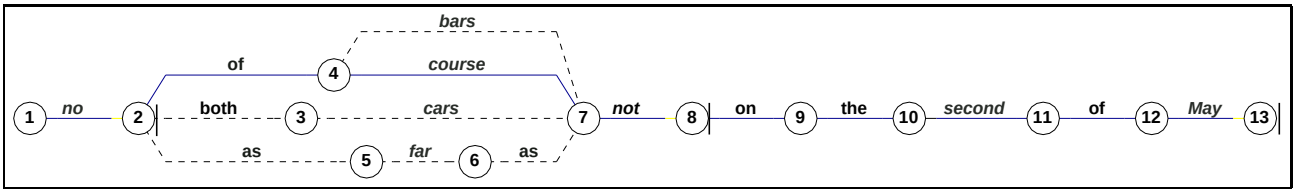

Figure 1: Word lattice produced by the English VERBMOBIL word recognizer. The utterance was *"No. Of course not. On the second of May."* The word graph is shown after prosodic annotation. Boundary hypotheses are displayed as vertical lines and phrase accent positions are indicated by slanted characters. Sentence mood is not shown.

3. What is the sentence mood at the prosodic boundaries?

This information does not only speed up the search during parsing. In some cases prosodic information is necessary in order to disambiguate between several possible meanings. If only acoustic–phonetic information were available many possible readings of the utterance shown in Figure 1 had to be considered, e.g.

1. No . Of course not on the second of May .
vs. 2. No ! Of course not ! On the second of May!
vs. 3. No . Of course not . On the second of May?
vs. 4. No ? Of course not on the second of May ?

Notice that the first two interpretations both make sense in the same context of an appointment scheduling dialogue. Interpretation 1 might be a confirmation that the second of May is not an available date, whereas interpretation 2 expresses the contrary. At this point of a dialogue prosody might help to recover from an otherwise unrecoverable error.

Figure 1 illustrates how the output of a word recognizer can be enriched with prosodic information. For simplicity, in the figure only presence/absence of prosodic events is displayed, whereas in the VERBMOBIL system probabilities are used. In addition to phrase boundaries, phrase accent, and sentence mood, every edge in a WHG is annotated with probabilities for irregular boundaries and emotion. This additional prosodic information is used in the VERBMOBIL system as follows:

Irregular boundaries: Irregular boundary markers are used to detect self–corrections. In spontaneous speech self–corrections are very frequent: A speaker starts a sentence, hesitates/stops, optionally utters an edit term, and then corrects himself. The point of interruption is usually distinctively prosodically marked. A *Part–Of–Speech* analysis before and after the point of interruption often allows to "repair" WHGs of such utterances [18].

Emotion: In the VERBMOBIL domain only *anger vs. not anger* is distinguished. Anger indicates that the dialogue goes astray. In such circumstances strategies to recover from error might be employed [3].

Since manual labeling is very time consuming, only parts of the VERBMOBIL speech database have yet been prosodically labeled. A set of four labels is used for boundary annotation, four levels of accents are distinguished, and sentence mood is labeled at prosodic boundaries as a combination of a question marker and a TOBI–like tonal sequence. Self–corrections are labeled as (1.) begin of *Reparandum* (first word which is corrected), (2.) point of interruption, (3.) *Edit Term* (e.g. "no", "uhm", ...), and (4.) end of *Reparans* (replacement for reparandum). Since there is almost no occurrence of anger in the regular VERBMOBIL speech database, emotional data was collected in *Wizard–of–Oz* experiments. Each word of the data is labeled as *angry/not angry*. Furthermore, a large part of the speech database is annotated with syntactic–prosodic labels [5].

3. FEATURE EXTRACTION

Prosodic features should compactly describe the properties of a speech signal which are relevant for the detection of prosodic events. Prosodic events, such as phrase boundaries and phrase accents, manifest themselves in variations of speaking–rate, loudness, pitch, and pausing. The exact interrelation of these prosodic attributes and prosodic events is very complex. Thus, our approach is to use a number of features in combination which describe these attributes in great detail. These features are then used as a basis for classification. In this paper, we describe those features that are used in the final version of VERBMOBIL; a former version of our feature set is given in [12]. The current feature set is also described in [10], where a comparison between the recognition results with the old and new feature set is given as well.

3.1. Feature Extraction Intervals

The variation of prosodic attributes relevant for the detection of prosodic events is limited to a certain context. Within that context, features which describe the variation are extracted and used for classification of prosodic events. Experiments have shown that a context of two words surrounding the current word are sufficient to decide if a prosodic event occurred. Larger context sizes do not improve the classification performance; this might either be due to the still rather limited size of our training data, or to the fact that a larger context contains only information that is irrelevant for the local events we want to model.

3.2. Different Kind of Features

The features that we extract from the speech signal describe the acoustic correlates of the prosodic–perceptual attributes, i.e. energy and F0 contour, duration and pauses. Furthermore, we use Part of Speech (POS) flags as features, cf. [6, 9]. We use a total of 125 features which can be sub–categorized as follows: 36 F0, 35 energy, 16 duration, 8 pause, and 30 POS features. These 125 features are used for all classifiers except sentence mood, where only a subset of 25 F0 features is used. The lexical POS flags cover a context of five words. Thus the classifier is able to learn a simple 5–gram language model. In section 4 it is shown that this syntactic information improves classification significantly.

Duration Features

Variations of speaking–rate or loudness have different effects on individual phonemes. Plosives are for instance much less affected by changes in speaking–rate than vowels. The variablity of the duration of a phoneme in a syllable depends also on the position of that syllable in the word and the position of the word accent. These considerations have led to the normalization that is described in the following.

Duration Normalization on the Phoneme Level

In order to model local speaking–rate variations we use measures that are based on the work of Wightman [22]. First, we are interested in capturing how much faster or slower an utterance was produced compared to the 'average speaker'. For a large training database, we compute for each phoneme its mean duration $\mu_{duration(u)}$ and standard deviation $\sigma_{duration(u)}$. $\mu_{duration(u)}$ constitutes the duration of unit u spoken by the 'average speaker'. The ratio $\frac{duration(u)}{\mu_{duration(u)}}$ measures how much faster or slower u was produced. The average of this ratio over an interval I is our measure $\tau_{duration}$, which is defined in Equation 1. Note that in the Equations 1 and 2, τ is stated more generally: the feature parameter F can be replaced not only by $duration$ but also e.g. by $energy$.

The value $\tau_{duration}$ is used to scale the mean duration $\mu_{duration(u)}$ and the standard deviation $\sigma_{duration(u)}$ of a speech unit u. The product $\tau_{duration}(I)\mu_{duration(u)}$ can be interpreted as the mean duration of the speech unit u if uttered with speaking–rate $\tau_{duration}(I)$. This interpretation is justified by the experiments in [22]. There it is shown that the mean and the standard deviation of speech–sound categories depend linearly on the speaking–rate.

The difference $duration(u) - \tau_{duration}(I)\mu_{duration(u)}$ is negative if $duration(u)$ is smaller than the scaled mean duration $\tau_{duration}(I)\mu_{duration(u)}$ of the speech unit u. A negative difference indicates faster speech; a positive difference indicates slower speech. This difference can be used to detect strong deviations from the scaled mean duration; the disadvantage of this measure, however, is that the deviation depends on the speech–sound category. If we divide the difference by the scaled standard deviation of the duration $\tau_{duration}(I)\sigma_{duration}(u)$ we get a measure that is normalized w.r.t. speech–sound dependent variation. In Equation 2, $\zeta_F(J, I)$ is defined as the average of that fraction in an interval J (interval I is used as 'reference'). With this approach it is also possible to distinguish between phonemes in accented and not accented syllables, and between phonemes that are in word initial, word final, word–internal syllables, or one–syllable words. This can be achieved simply by using such units u in the Equations 1 and 2.

$$\tau_F(I) := \frac{1}{\#I}\sum_{u\in I}\frac{F(u)}{\mu_{F(u)}} \quad (1)$$

$$\zeta_F(J, I) := \frac{1}{\#J}\sum_{u\in J}\frac{F(u) - \tau_F(I)\mu_{F(u)}}{\tau_F(I)\sigma_{F(u)}} \quad (2)$$

Duration Normalization on the Word Level

The measures $\tau_{duration}(I)$ and $\zeta_{duration}(J, I)$ (computed with phonemes as speech units u), as defined in Equations 1 and 2 can already be used as prosodic features and, in fact, are often used, e.g. in [22], [1], and [12]. These measures have several disadvantages, though. First, during feature extraction the duration of each phoneme has to be determined in order to compute these measures. To compute a phoneme segmentation of the recognized words, however, is time consuming and requires considerable memory resources. The word recognition modules in the VERBMOBIL system cannot provide this segmentation due to architectural constraints. Second, the phoneme segmentation suffers if the audio quality is degraded. This leads to a drop in the recognition accuracy of prosodic events. Furthermore, pronunciation variants can cause the phoneme segmentation to be incorrect and thus lead to erroneous features.

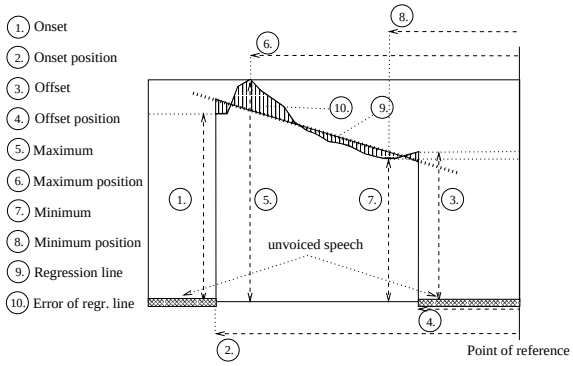

Figure 2: Example of features used to describe a pitch contour.

The normalization according to the Equations 1 and 2 can be used on the word level as well. The word duration statistics $\mu_{duration(w)}$ and $\sigma_{duration(w)}$ for a word w can either be determined directly if enough tokens of this word have been observed in the training data. Otherwise the word duration statistics can be approximated based on the duration statistics of the phonemes that w consists of; this approach is thus time–consuming only during the training. This word based normalization circumvents the disadvantages mentioned above and is, therefore, currently used in the VERBMOBIL system.

Pitch Features

Pitch features are based on the (logarithmic) F0 contour. Examples for features that are used to describe the F0 contour in a specific interval are shown in Figure 2. In addition to the features displayed in this figure, we also use the mean and the median as features.

Energy Features

In order to describe the short–term energy contour we use only a subset of the features that are shown in Figure 2 because not all of them provide useful information (e.g. onset and offset). Furthermore, we include normalized energy in our feature vector. The same normalization as used for the duration normalization on the word level (see above) can be applied; i.e. $F = energy$ has to be used in Equations 1 and 2 with words as speech units u. The measures $\tau_{energy}(I)$ and $\zeta_{energy}(J, I)$ according to these equations are

included in the feature vector that we currently use in the VERBMOBIL system.

Pause Features

The pause features are easily extracted: These are simply the duration of *filled pauses* (e.g. "uhm", "uh", ...) and *silent pauses*.

Part of Speech Features

A POS flag is assigned to each word in the lexicon, cf. [6]. We include a flag for each of 15 POS classes (for German) or 10 POS classes (for English) and a context of 5 words in the feature vector. These POS features can be mapped onto 6 higher categories, as 'noun', 'verb', etc. The 'computation' of these features consists simply of a table lookup and is, therefore, very efficient [9].

4. EXPERIMENTS AND RESULTS

In this section we describe the experiments that we performed in order to

1. investigate if the methods developed on the German subcorpus of the VERBMOBIL data are suited for English and Japanese, as well,
2. determine the reduction in computation time.

As mentioned in Section 2, labeled data sets for phrase accents, phrase boundaries, sentence mood, irregular boundaries, emotion, and syntactic–prosodic boundaries exist. In this paper we restrict ourselves to phrase accents (A), phrase boundaries (B), and questions (Q). Furthermore we do not distinguish all four accent labels and all four boundary labels in our classification experiments, but map these labels to classes as shown in Table 1. For Sentence mood we distinguish between questions and non–questions.

Acoustic–prosodic boundary labels		
label	*class*	*description*
B3	B	prosodic clause boundary
B2	¬B	prosodic phrase boundary
B9	¬B	irregular boundary, usually hesitation lengthening
B0	¬B	every other word boundary
Acoustic–prosodic accent labels		
label	*class*	*description*
PA	A	most prominent (primary) accent within prosodic clause
NA	A	all other accented words carrying secondary accent
EK	A	emphatic or contrastive accent
UA	¬A	unaccented words

Table 1: Description of acoustic–prosodic boundary and accent labels.

The VERBMOBIL corpus momentarily consists of more than 50 CDROMs with high quality speech recordings. Only a small subset of the CDROMs has yet been prosodically labeled. While the data sets for German and English have been labeled by trained personnel, the data set for Japanese has been labeled by students in an effort to obtain some data for the experiments that are described below.

4.1. Prosodic Classification with Neural Networks

In the prosody module a *multi layer perceptron* (MLP) is used as a classifier. The input layer has as many nodes as there are features in the feature vector (see Section 3). The output layer has two nodes corresponding to the prosodic events, e.g., A, B and Q, and their complement, e.g., ¬A, ¬B and ¬Q. The topology of the hidden layers is optimized based on a validation sample. For each word of the WHG a feature vector with a context of two words to the left and to the right is computed. The training is done using the *Stuttgart Neuronal Network Simulator* (SNNS), cf. [24], [23]. During classification in the prosody module, a prosodic feature vector is passed to the MLP, and the scores of the output nodes are normalized to the range of $[0 \ldots 1]$ so that they add up to 1; these scores can thus be interpreted as probabilities. The WHG is then annotated with the probability for the prosodic event and its complement. The probability scores can be extracted by the other modules of VERBMOBIL directly out of the WHG.

As the effort needed for annotation differs considerably for the different prosodic events, cf. [5], the size of the available training data differs accordingly. However, the resulting classifiers yield good recognition rates. Classification errors have different effects depending on whether a prosodic event is not found (miss) or its complement is wrongly classified as a prosodic event (false alarm). Therefore, we consider recall, i.e., $correct/(correct + miss)$, and precision, i.e., $correct/(correct + false\ alarm)$. In Table 2 only recall is given; precision can easily be computed from the numbers provided. Due to sparse data and/or the fact that, especially for English and Japanese, the same speakers were often used for more than one dialogue, cf. column 'dial./speakers' in Table 2, train and test speakers for the MLP classification were kept disjunct only for German. For the German and English databases used for the MLP classification with acoustic–prosodic features, the male/female distribution can be given: German train 38/7, German test 3/3; English train 7/5, English test 3/3 (Japanese: not available).

Several feature vectors and different groups of features in different context sizes were examined to get the best MLP classifier for our prosodic events. Eventually we added POS features, taking textual information during prosodic classification into account. Our final feature set now includes 95 acoustic–prosodic features and a varying number of POS features, depending on the language and the optimized granularity of categorization. (The old feature set used in [12, 16, 15] was based on the phoneme alignment and consisted of 276 features.) The best results we achieved and integrated into the VERBMOBIL system can be found in Table 2. In [10] it is shown that the classification errors are reduced by 3% for German accents and boundaries and by 18% for English boundaries and 24% for English accents when the new feature set is used. The bigger improvement for the English data is due to the fact that the phoneme based time alignment is not as robust as the word based, due to the significantly less training data.

Even if it is possible to train MLPs with more classes, for the prosodic events A, B and Q, we used only two because more classes

	dial./speakers	B	¬B	A	¬A	Q	¬Q
G	# train: 30/45	2310	10964	5140	8134	349	1743
	# test: 3/6	227	1320	697	850	34	240
	% recall – POS	84	88	78	84	88	91
	% recall + POS	89	89	79	86	91	90
E	# train: 33/12	638	4137	1958	2817	47	205
	# test: 4/6	94	611	297	408	4	27
	% recall – POS	97	91	81	78	100	96
	% recall + POS	97	93	82	82	100	85
J	# train: 24/20	747	5348	1545	4889	-	-
	# test: 19/18	67	558	165	497	-	-
	% recall – POS	81	89	75	71	-	-

Table 2: MLP classification: Recall in percent for prosodic boundaries B, prosodic accents A, and prosodic questions Q in the three languages of the VERBMOBIL system; number of dialogues, speakers, and cases is given for train and test.

Computation time	
old features	**new features**
216 min	17 min

Table 3: Computation time of the old and new feature extraction methods on 112 min of speech

yielded worse results due to sparse data. Generally, classification results are good or very good; two overall tendencies can further be observed: first, boundaries can be better classified than accents, and POS information improves the performance of the MLP except for English questions, where the database is very small.

4.2. Efficiency

As a last experiment we measured the computation time during feature extraction on the data set G, using

1. 95 old features that require a time alignment of the phoneme sequence, and
2. 95 new features that do normalization on the word level and therefore do not need a time alignment.

The set of 95 features is the subset of word–based features without the POS flags. We chose this subset as the basis for the experiment in order to get comparable results. Feature extraction with normalization on the phoneme level does not require significantly more computation time or memory if 276 features instead of the 95 features are used. The requirements are dominated by the time alignment. The experiment was performed on the same computer under the same conditions (no load except for the feature extraction process). The result is shown in Table 3. As can be seen, the real time factor drops from 1.93 to .15.

5. MULTILINGUAL ARCHITECTURE

In the VERBMOBIL system, prosodic information is computed for the three languages *German*, *English*, and *Japanese*. First a prosody module for each of these languages was integrated in the system. Thus a lot of common data and procedures for all languages could not be shared. To reduce the memory requirements we integrated the language dependent modules into one *multilingual prosody module* where other languages easily can be added. The architecture of the multilingual prosodic module is shown in Figure 3.

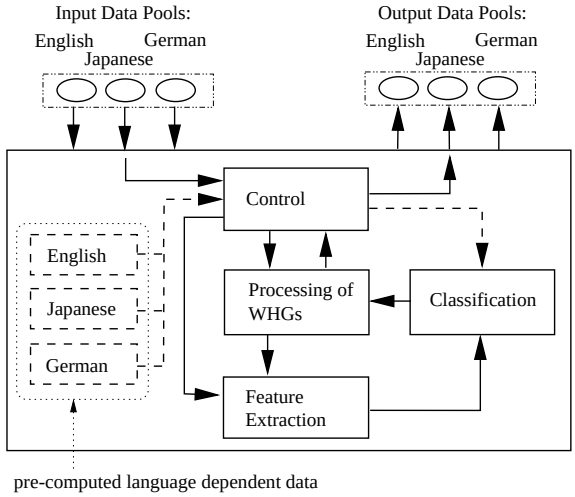

Figure 3: Architecture of the multilingual prosody module for prosodic processing.

It is possible to share the feature extraction and classification procedures in a multilingual module because they are language independent. The language dependent data, for instance, duration normalization tables, and specific classifiers are kept in different structures. Via configuration files individual classification parameters for each language, for instance, the different sizes of the n–grams, can be loaded. The prosody module has to deal with different incoming and outgoing data. The communication is done with the *Pool Communication Architecture* (PCA) which is described in [14]. Input into the prosody module is the speech signal and the word hypotheses graph (WHG), output is an annotated WHG, now including additional prosodic information for each word. In more detail, processing in the prosody module can be described as follows:

- The control component handles the global behavior of the prosody module, for instance: 'get the WHG', 'start classification'. Furthermore, the language dependent behavior can be configured here.
- The PCA in VERBMOBIL works event driven. Depending on which data pool first indicates incoming data, the handler for that particular data pool is called. Each data pool gets input from the word recognition module for one language. Thus, the control component selects the corresponding language dependent data, for instance, language–specific normalization tables, which are needed for the feature extraction as described in Section 3.
- The WHG component then traverses the WHG. At each node the feature extraction component is called.
- The feature extraction component uses the language dependent data structure, the word hypotheses and word intervals from the WHG (see Section 3). The result is a feature vector which is passed to the classification component.
- The classification component classifies the feature vector using language dependent classifier information. For that we use

MLPs which can be combined with language models, cf. section 4. The classification result is handed back to the WHG component.

- The WHG component annotates the WHG correspondingly.
- After all edges of the WHG have been processed the annotated WHG is delivered to the output data pool.

The structure of the multilingual module has several advantages. It can easily be extended to additional languages. In order to add a new language only a few changes to the configuration file have to be made, i.e. the language dependent parameters have to be set. Furthermore, the memory requirement of the multilingual module after some optimization steps (64 MByte) is a lot smaller than the sum of the memory needed for three modules (291 MByte).

6. CONCLUSION

In this paper we have shown that the methods to classify prosodic events that we developed on German speech data are also well suited for other languages. Due to efficiency problems caused by the feature extraction with phoneme–based normalization a new set of features was proposed that avoids these problems. With this new set of features we achieved a speed–up of the feature extraction component by more than a factor of 12. The new features proved to be more robust, and thus, led to significant improvements for English phrase boundary and accent classification.

An architecture for a multilingual module for prosodic processing was described and the advantages of this architecture were discussed. The memory requirement of a multilingual module compared to three single monolingual modules (with the new feature set) is reduced by 78%.

7. REFERENCES

1. Paul C. Bagshaw. *Automatic prosodic analysis for computer aided pronunciation teaching*. PhD thesis, University of Edinburgh, 1994.
2. A. Batliner, A. Buckow, H. Niemann, E. Nöth, and V. Warnke. The Prosody Module. In Wahlster [21], pages 106–121.
3. A. Batliner, R. Huber, H. Niemann, E. Nöth, J. Spilker, and K. Fischer. The Recognition of Emotion. In Wahlster [21], pages 122–130.
4. A. Batliner, A. Kießling, R. Kompe, H. Niemann, and E. Nöth. Tempo and its Change in Spontaneous Speech. In *Proc. European Conf. on Speech Communication and Technology*, volume 2, pages 763–766, Rhodes, 1997.
5. A. Batliner, R. Kompe, A. Kießling, M. Mast, H. Niemann, and E. Nöth. M = Syntax + Prosody: A syntactic–prosodic labelling scheme for large spontaneous speech databases. *Speech Communication*, 25(4):193–222, 1998.
6. A. Batliner, M. Nutt, V. Warnke, E. Nöth, J. Buckow, R. Huber, and H. Niemann. Automatic Annotation and Classification of Phrase Accents in Spontaneous Speech. In *Proc. European Conf. on Speech Communication and Technology*, volume 1, pages 519–522, Budapest, Hungary, 1999.
7. M. Beckman. *Stress and Non–stress Accent*. Foris Publications, Dordrecht, 1986.
8. C.M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, NY, 1995.
9. J. Buckow, A. Batliner, R. Huber, H. Niemann, E. Nöth, and V. Warnke. Detection of Prosodic Events using Acoustic–Prosodic Features and Part–of–Speech Tags. In *Proc. International Workshop SPEECH AND COMPUTER (SPECOM'00)*, page (to appear), St-Petersburg, 2000.
10. J. Buckow, R. Huber, V. Warnke, A. Batliner, E. Nöth, and H. Niemann. Multi–lingual Prosodic Processing. In *Proc. ESCA Workshop on Dialogue and Prosody*, pages 157–162, Eindhoven, Netherlands, 1999.
11. B. Kiefer, H.-U. Krieger, and M.J. Nederhof. Efficient and Robust HPSG Parsing of Word Graphs. In Wahlster [21], pages 280–295.
12. A. Kießling. *Extraktion und Klassifikation prosodischer Merkmale in der automatischen Sprachverarbeitung*. Berichte aus der Informatik. Shaker Verlag, Aachen, 1997.
13. M. Kipp, J. Alexandersson, N. Reithinger, and R. Engel. Dialog Processing. In Wahlster [21], pages 452–465.
14. A. Klüter, A. Ndiaye, and H. Kirchmann. Verbmobil from a Software Engineering Point of View: System Design and Software Integration. In Wahlster [21], pages 635–658.
15. R. Kompe. *Prosody in Speech Understanding Systems*. Lecture Notes for Artificial Intelligence. Springer–Verlag, Berlin, 1997.
16. R. Kompe, A. Kießling, H. Niemann, E. Nöth, A. Batliner, S. Schachtl, T. Ruland, and H.U. Block. Improving Parsing of Spontaneous Speech with the Help of Prosodic Boundaries. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, volume 2, pages 811–814, München, 1997.
17. N. Reithinger and R. Engel. Robust Content Extraction for Translation and Dialog Processing. In Wahlster [21], pages 428–437.
18. T. Ruland, C. Rupp, J. Spilker, H. Weber, and K. Worm. Making the Most of Multiplicity: A Multi-Parser Multi-Strategy Architecture for the Robust Processing of Spoken Language. In *Proc. Int. Conf. on Spoken Language Processing*, volume 4, pages 1163–1166, Sydney, 1998.
19. J. Spilker, M. Klarner, and G. Görz. Processing Self–Corrections in a Speech–to–Speech System. In Wahlster [21], pages 131–140.
20. S. Vogel, F.J. Och, C. Tillmann, S. Nießen, H. Sawaf, and H. Ney. Statistical Methods for Machine Translation. In Wahlster [21], pages 377–393.
21. W. Wahlster, editor. *Verbmobil: Foundations of Speech-to-Speech Translations*. Springer, New York, Berlin, 2000.
22. C.W. Wightman. *Automatic Detection of Prosodic Constituents*. PhD thesis, Boston University Graduate School, 1992.
23. A. Zell, N. Mache, T. Sommer, and T. Korb. Design of the SNNS neural network simulator. In *Proceedings of the Östereichische Artificial-Intelligence-Tagung*, Informatik-Fachberichte 287, pages 93–102. Springer Verlag, 1991.
24. A. Zell, N. Mache, T. Sommer, and T. Korb. The SNNS neural network simulator. In *Proceedings of the 15. Fachtagung für künstliche Intelligenz*, pages 254–263. Springer Verlag, 1991.