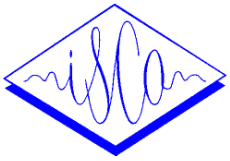

ETRW on Dialogue and Prosody
Veldhoven, The Netherlands
September 1-3, 1999

MULTI-LINGUAL PROSODIC PROCESSING

Jan Buckow *Richard Huber* *Volker Warnke* *Anton Batliner* *Elmar Noeth*
Heinrich Niemann

University of Erlangen-Nuremberg,
Chair for Pattern Recognition (Computer Science 5),
Martensstr. 3,
D-91058 Erlangen, Germany
{buckow,huber,warnke,batliner,noeth,niemann}@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 [9]. The methods to classify prosodic events have been developed on the German sub-corpus of the VERBMOBIL speech database. In this paper we describe how the methods that we developed on the German sub-corpus can be applied to other languages. Preliminary experiments show that these methods are suited for English and Japanese, as well. Efficiency problems are addressed and a new set of features that eliminates most of these problems is presented. The new set of features facilitates a multi-lingual module for prosodic processing. We present an architecture for such a multi-lingual module and discuss the advantages of this approach compared to an approach that uses separate modules for different languages. This multi-lingual module and the new feature set are evaluated w.r.t. computation time, memory requirement, and classification performance. Preliminary results show that the memory requirement can be reduced by at least 70%, 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 dialogs [6]. During the translation process prosodic information is used at various stages. Phrase boundaries, phrase accents, and sentence mood are used to guide syntactic parsing, disambiguate between several possible meanings [10], and improve the naturalness of the synthesis. Irregular boundary markers are used to deal with corrections [12]. Furthermore, some preliminary emotion detection is integrated in order to improve the system behavior in the case of errors [7].

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 a MLP can be interpreted as *a-posteriori* probability [4].

As the importance of prosody for the system performance could be shown on a German sub-corpus of the VERBMOBIL data [9] we investigate the applicability of our approach for the other VERBMOBIL languages. These experiments are described in Section 4.1. In these experiments, 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 [13, 3, 8, 1].

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 multi-lingual prosodic processing.

We show that with the new set of features and a multi-lingual 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 4.2 and 4.3). But first we give a few examples of how and why prosody is used in VERBMOBIL in Section 2.

2. PROSODY AND DIALOG

Dialog processing in VERBMOBIL is very complex and prosody is used at various stages during the translation process [10]. 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 the basis for later syntactic and semantic parsing as described in [12]. Important prosodic information in the context of syntactic/semantic parsing is:

*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.

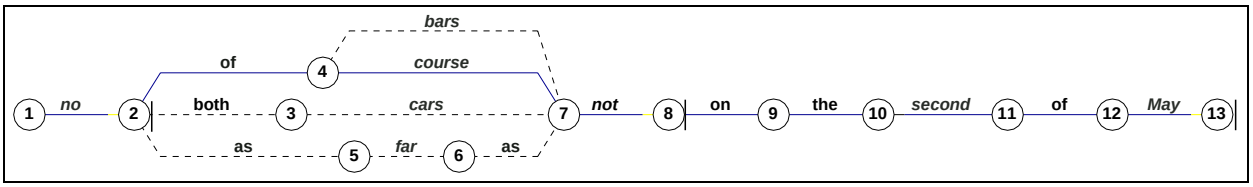

Figure 1: Word lattice produced by the English VERBMOBIL word recognizer. The utterance was *"No. Of course not. On the second of May."* The word graph is shown after prosodic annotation. Boundary hypotheses are displayed as vertical lines and phrase accent positions are indicated by slanted characters.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

1. Which words of an utterance carry a phrase accent?
2. Where in an utterance are prosodic boundaries?
3. What is the sentence mood at the prosodic boundaries?

This information does not only speed up 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 and 2 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 dialog. 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 dialog 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. Furthermore, a subset of the prosodic features is transmitted to the synthesis module. 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 [12].

Emotion: In the VERBMOBIL domain only *anger vs. not anger* is distinguished. Anger indicates that the dialog goes astray. In such circumstances strategies to recover from error might be employed [7].

Speech synthesis: A subset of prosodic features is sent to the synthesis module. These features can be used to adapt the synthesized speech to the speaker and to make the output sound more natural.

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

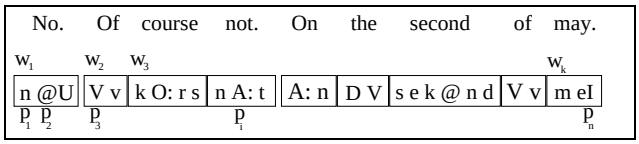

Figure 2: Utterance *"No. Of course not. On the second of May."* with the phoneme sequence in SAMPA notation.

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 [2].

3. FEATURE EXTRACTION

Aim of the extraction of prosodic features is to 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 find features that describe the attributes as exactly but also as compactly as possible. These features are then used as basis for classification.

3.1. Feature extraction intervals

In order to decide at some point of an utterance if a prosodic event occurred or not, some context is necessary. In preliminary experiments a context of two words to each side of the current word proved to be sufficient; larger contexts did not improve classification results. In our classification experiments each component of a feature vector is determined over some interval; e.g. the regression coefficient of the energy contour is used as a feature and computed over several intervals $I_{(f,t)}$ (time interval from the beginning of word f to the end of word t). Intervals that we use are e.g. $I_{(-2,-1)}$ or $I_{(1,2)}$. At the end of the word *"not"* in the utterance shown in Figure 2 the interval $I_{(-2,-1)}$ e.g. denotes the time interval from the beginning of the word *"Of"* to the end of the word *"course"*.
Prosodic processing in VERBMOBIL is performed on WHGs (see Figure 1). Begin and end time of each edge are given by the word recognizer. Word sub-units intervals (e.g. phonemes intervals), however, can not be produced by the word recognizers in VERBMOBIL. The best matching sequence of acoustic models, that would allow to determine phoneme intervals, is hidden

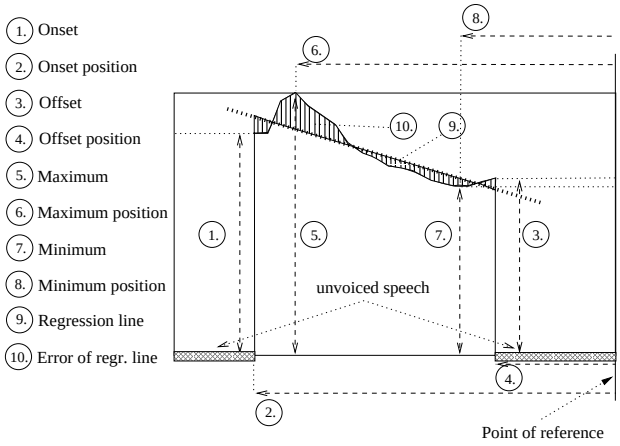

Figure 3: Example of features used to describe a pitch contour.

in temporary back-pointer tables. If a phoneme segmentation has to be computed (e.g. in order to perform a normalization on the phoneme level; see below) a subsequent processing step has to align acoustic models for phoneme units with the speech signal and thus determine phoneme intervals. This processing step is expensive in terms of memory requirement and computation time.

3.2. Different kind of features

As mentioned above, prosodic events are perceived based on the perception of the prosodic attributes speaking-rate, loudness, pitch, and pausing. The features that we extract from the speech signal describe the acoustic correlates of these attributes, i.e. energy and fundamental frequency (F0) contour, duration and pauses.

The pause features are easily extracted: These are simply the duration of *filled pauses* (e.g. "uhm", "uh", ...) and *silent pauses*. Energy and pitch features are based on the short term energy and F0 contour, respectively. Duration features should capture variations in speaking-rate and are based on the duration of speech units. A normalization of energy, duration, and pitch features can be performed in order to take phone intrinsic variations into account.

Features describing contours

As mentioned above, energy and F0 features are based on the short-term energy and F0 contour, respectively. Some of the features that are used to describe a pitch contour in a specific interval are shown in Figure 3. Additionally, we use the mean and the median as features (not shown in the figure).

Normalization

Variations of speaking-rate or loudness have different effects on individual phonemes. Plosives are e.g. 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. The position of the word accent also has some effect on the variability. These considerations have led to the normalization that is described in the following paragraphs.

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 [13]. First, we are interested in capturing how much faster or slower an utterance was produced compared to the "average speaker". For a 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 of Wightman in [13]. He showed 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 could be used to detect strong deviations from the scaled mean duration; the disadvantage 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 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.

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

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

We include $\tau_{duration}(I)$ and $\zeta_{duration}(J, I)$ in our feature vector as global speaking rate and normalized local speaking rate.

In our first experiments (see Section 4.1), a time alignment of the phoneme sequence was performed and $\tau_{duration}(I)$ was computed according to Equation 1 (with $F = duration$, I being some interval and $\#I$ denoting the number of units u in the interval I). The units u were phonemes in this case.

Duration normalization on the word level

A major disadvantage of the normalization described in last paragraph is the necessity to determine the phoneme segments during classification. In our feature extraction module the computation of the phoneme segments requires 92% of the total computation time and 64% of the total memory needed. Therefore, one would prefer to normalize on the word level and thus avoid the time alignment. Equations 1 and 2 are applicable to word intervals, as well. But for most words w there is not enough training data to get reliable estimates for the $\mu_{duration(w)}$ and $\sigma_{duration(w)}$.

Equation 2 can be interpreted as a transformation of a fea-

ture $F(u)$ with mean $\tau_F(I)\mu_{F(X)}$ and standard deviation $\tau_F(I)\sigma_{F(X)}$ to a feature with mean 0 and standard deviation 1. If we assume that the $F(u)$ are independent random variables then $\sigma^2_{F(u_1)} + \sigma^2_{F(u_2)} = \sigma^2_{F(u_1)+F(u_2)}$ (see e.g. [5]). Thus, we can compute the mean $\mu_{F(w)}$ and the standard deviation $\sigma_{F(w)}$ for a word $w = (p_1, p_2, p_3, \ldots, p_n)$ with phonemes p_i as shown in Equations 3 and 4 as long as $F(w) = F(p_1) + F(p_2) + \ldots F(p_n)$.

$$\mu_{F(w)} = \sum_{i=1}^{n} \mu_{F(p_i)} \quad (3)$$

$$\sigma_{F(w)} = \sqrt{\sigma^2_{F(p_1)+F(p_2)+\ldots F(p_n)}} = \sqrt{\sum_{i=1}^{n} \sigma^2_{F(p_i)}} \quad (4)$$

In case of $F = duration$ this means that if we assume the durations of the phonemes are independent random variables then the word duration statistics can be deduced from the phoneme duration statistics. Thus, if during recognition a normalization on the word level has to be performed according to Equations 1 and 2 then either word duration statistics $\mu_{F(w)}$ and $\sigma_{F(w)}$ can be used if reliable estimates exist or the estimates can be deduced according to Equations 3 and 4.

This normalization on the word level can be performed without time alignment of the phoneme sequence. Such a time-alignment is only necessary to compute the word and phoneme duration statistics. This can be done offline, so that precomputed tables can be used during recognition. Thus a very significant reduction of memory requirement and computation time can be achieved (see Section 4).

Energy features

In order to describe the short-term energy contour we used only a subset of the features that are shown in Figure 3 because not all of them provide useful information (e.g. onset and offset). Furthermore, we included normalized energy in our feature vector; the same normalization as described in the last paragraph can be applied here, i.e. $F = energy$ has to be used in Equations 1 and 2. As in the case of duration, we included $\tau_{energy}(I)$ and $\sigma_{energy}(J, I)$ in our feature vector.

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. compare the normalization based on phoneme segments with the normalization on the word level,
3. determine the reduction in memory requirement and 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 ourselfs to phrase accents and phrase boundaries. 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.

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.

	German	**English**	**Japanese**
data set	GER	ENG	JAP
dialogs	33	37	31
minutes	$\approx$ 112	$\approx$ 38	$\approx$ 80

Table 2: Data sets used in the classification experiments.

The VERBMOBIL corpus momentarily consists of 30 CDROM with high quality speech recordings. Only a small subset of the CDROMs has yet been prosodically labeled. While the data sets GER and ENG have been labeled by trained personnel, the data set JAP has been labeled by students in an effort to obtain some data for the experiments that are described below.

4.1. Classification with old feature set

With the experiments described here we wanted to determine the applicability of the methods that we developed on German data to other languages. Therefore, we performed classification experiments on the data sets given in Table 2. (1.) We split each data sets in training and test sets. (2.) A time-alignment of the words and the phoneme sequence of each word was performed. (3.) Features were extracted. (4.) MLPs were trained and tested.

As described in in Section 3.2, a time-alignment of the phoneme sequence was performed in order to be able to normalize on the phoneme level. For English and German a word recognizer trained on VERBMOBIL data was available. For Japanese we mapped the Japanese phoneme set to the German phonemes. Then we build a recognizer for Japanese *mora* (word sub-units; see [11]). The acoustic models of the *mora* units were constructed from the German acoustic models for the phonemes of that mora.

Features were extracted as described in Section 3. A normalization was performed on the phoneme level. A context of two words to each side of the current word was used. Since the time alignment of the phoneme sequence allows also to determine syllable and syllable nuclei intervals, we included features for these intervals in our feature vector. Thus, we ended up with a set of 276 features for English and German. For Japanese we had only 170 features because instead of word and syllable intervals we

	German		English		Japanese	
	$\overline{\mathcal{RR}}$	$\mathcal{RR}$	$\overline{\mathcal{RR}}$	$\mathcal{RR}$	$\overline{\mathcal{RR}}$	$\mathcal{RR}$
Boundaries	84.0	85.6	84.0	86.0	86.1	93.4
Accents	80.9	81.2	77.0	75.0	61.0	73.2

Table 3: Classification results with features using normalization on the phoneme level, i.e. with time alignment of the phoneme sequence.

	German		English	
	$\overline{\mathcal{RR}}$	$\mathcal{RR}$	$\overline{\mathcal{RR}}$	$\mathcal{RR}$
Boundaries	84.7	86.0	89.0	88.5
Accents	81.0	81.7	81.4	81.0

Table 4: Classification results with features computed using a normalization on the word level (without time alignment of the phoneme sequence).

only had *mora* intervals.

The classification results are shown in Table 3. $\mathcal{RR}$ and $\overline{\mathcal{RR}}$ denote absolute and average recognition rate, respectively. $\mathcal{RR}$ is the fraction of correctly classified patterns of all patterns. $\overline{\mathcal{RR}}$ is the average of the recognition rates for each class.

4.2. Classification with new feature set

The classification results presented in Section 4.1 indicate that we can use our approach for the classification of prosodic events in German, English and Japanese. In a multi-lingual dialog system like VERBMOBIL we are faced with several problems, though: The feature extraction that was used in the experiments described in Section 4.1 requires a time alignment of the phoneme sequence. Thus, acoustic models for each language have to be part of the component that extracts features. This requires a large amount of memory. Furthermore, Equations 1 and 2 show that the quality of the normalized duration and energy features depends very much on the quality of the time alignment. This means on the one hand, that the acoustic models have to be accurate enough to yield a good alignment. On the other hand, this means that the classification performance is likely to drop if this requirement is not met, i.e., these features are not very robust.

As a consequence, we developed the normalization on the word level as described in Section 3.2. These features are more robust: As long as the word recognition performs well, the normalization is accurate. With a feature set that uses this normalization we performed classification experiments. In this case no phoneme segments were available, and therefore, only word intervals could be used. Thus, our feature set consisted only of 105 word based features. The results for English and German are shown in Table 4.

While the recognition results on German data improved only slightly, the improvement for English data is significant. This can be explained with the amount of training data used to train the recognizers. While the German recognizer has been trained with approximately 30 hours of speech, the English recognizer was trained with only 8 hours of speech.

The new feature set is a sub-set of the old feature set. The only difference is the word-based instead of the phoneme-based normalization. Despite of that, we get an improvement rather than a degradation of performance with the new feature set. An explanation for that might be that the word based normalization is more robust than the phoneme based normalization, whereas the syllable and syllable nuclei features in the old feature set provide no additional information.

Computation time		Memory requirement	
old features	**new features**	**old features**	**new features**
216 min	17 min	73 MByte	26 MByte

Table 5: Computation time and memory requirement of the old and new feature extraction methods on 112 min of speech

4.3. Efficiency

As a last experiment we measured the computation time and the memory requirement during feature extraction on the data set GER, 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 sub-set of word-based features that is currently used in the VERBMOBIL system. We chose this subset as 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 5.

5. MULTI-LINGUAL ARCHITECTURE

During the development of the three mono-lingual modules used for classification in Section 4 the task of integrating these three different modules in the VERBMOBIL system became increasingly difficult. Due to the large number of system parameters which are controlled via parameter files error prone replication was necessary. As a result, we started to develop an architecture for a multi-lingual module for prosodic processing. The structure is shown in Figure 4. The structure is very simple:

- A control module controls the global behavior of the prosody component of the VERBMOBIL system. Furthermore the language dependent behavior can be configured here.
- Communication in VERBMOBIL is event driven. Depending on which data pool first indicates incoming data, the handler for that particular data pool is called. Each data pool is connected to a word recognition component for one language. Thus, the control module selects the corresponding language dependent lexicon, phoneme set, and word duration and energy statistics (needed for the normalization as described in Section 3.2). The WHG is transmitted to the WHG component along with the language dependent data.
- The WHG component then traverses the WHG. At each node the feature extraction component is called.

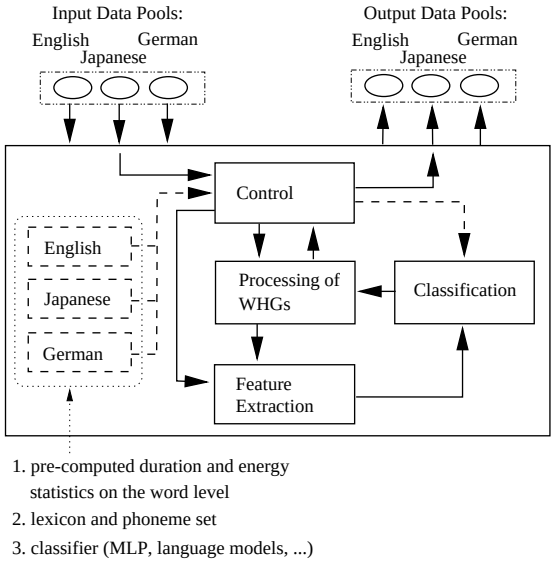

Figure 4: Architecture of the multi-lingual module for prosodic processing.

- The feature extraction component needs energy and duration statistics, words 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. Here we usually use MLPs sometimes in combination with language models (LM). 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 for the correct language.

The structure of the multi-lingual module has several advantages. It can be easily extended. In order to add a new language only a few changes to the configuration file have to be made; i.e. the language dependent parameter files like lexicon, phoneme set, duration statistics and classifier parameters have to be set. Furthermore, the memory requirement of the multi-lingual module (171 MByte) is far 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 is 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, while the memory requirement could be reduced by almost a factor of three. 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 multi-lingual module for prosodic processing was described and the advantages of this architecture were discussed. Currently, a module with that architecture is integrated in the VERBMOBIL system. The memory requirement of a multi-lingual module compared to three single mono-lingual modules (with the new feature set) is further reduced by 41%. In combination with the reduced size of the feature extraction component an overall reduction of more than 75% was achieved.

7. REFERENCES

1. 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.
2. 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.
3. M. Beckman. *Stress and Non–stress Accent*. Foris Publications, Dordrecht, 1986.
4. C.M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, NY, 1995.
5. I.N. Bronstein and K.A. Semendjajew. *Taschenbuch der Mathematik*. Verlag Harri Deutsch, Thun und Frankfurt/Main, 24 edition, 1989.
6. T. Bub and J. Schwinn. Verbmobil: The Evolution of a Complex Large Speech-to-Speech Translation System. In *Int. Conf. on Spoken Language Processing*, volume 4, pages 1026–1029, Philadelphia, 1996.
7. R. Huber, E. Nöth, A. Batliner, J. Buckow, V. Warnke, and H. Niemann. You BEEP Machine — Emotion in Automatic Speech Understanding Systems. In *Proc. of the Workshop on TEXT, SPEECH and DIALOG (TSD'98)*, pages 223–228, Brno, 1998. Masaryk University.
8. Andreas Kießling. *Extraktion und Klassifikation prosodischer Merkmale in der automatischen Sprachverarbeitung*. Berichte aus der Informatik. Shaker Verlag, Aachen, 1997.
9. 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.
10. Ralf Kompe. *Prosody in Speech Understanding Systems*. Lecture Notes for Artificial Intelligence. Springer–Verlag, Berlin, 1997.
11. P. Ladefoged. *A Course in Phonetics, Second Edition*. Hartcourt Brace Jovanovich, New York, 1982.
12. 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 *Int. Conf. on Spoken Language Processing*, Sydney, 1998.
13. C.W. Wightman. *Automatic Detection of Prosodic Constituents*. PhD thesis, Boston University Graduate School, 1992.