The Prosody Module

Anton Batliner, Jan Buckow, Heinrich Niemann, Elmar Nöth, and Volker Warnke

Lehrstuhl für Mustererkennung, Universität Erlangen-Nürnberg, Germany

Abstract. We describe the acoustic-prosodic and syntactic-prosodic annotation and classification of boundaries, accents and sentence mood integrated in the Verbmobil system for the three languages German, English, and Japanese. For the acoustic-prosodic classification, a large feature vector with normalized prosodic features is used. For the three languages, a multilingual prosody module was developed that reduces memory requirement considerably, compared to three monolingual modules. For classification, neural networks and statistic language models are used.

1 Introduction

In this paper we discuss the use of prosodic information in Verbmobil. Prosodic information is attached to speech segments which are larger than a phoneme, i.e. *syllables, words, phrases,* and *whole turns* of a speaker. To these segments we attribute perceived properties like *pitch, loudness, speaking rate, voice quality, duration, pause, rhythm,* and so on. Even though there generally is no unique feature in the speech signal corresponding to these perceived properties, we can find features which highly correlate with them; examples are the acoustic feature *fundamental frequency* (F0), which correlates to *pitch,* and the *short time signal energy* correlating to *loudness.* In human-human communication, the listener extracts information out of these perceived phenomena, i.e. we can assign certain functions to them. The prosodic functions which are generally considered to be the most important ones are the marking of *boundaries, accents, sentence mood,* and *emotional state* of the user. The processing of emotion in Verbmobil is described in Batliner et al., in this volume. For the other three phenomena, we give examples that are taken from the Verbmobil scenario:

Boundaries:

(1) *Fünfter geht bei mir, nicht aber neunzehnter.* vs.
Fünfter geht bei mir nicht, aber neunzehnter. i.e.
The fifth is possible for me, but not the nineteenth. vs.
The fifth is not possible for me, but the nineteenth would be OK.

Accentuation:

(2) *Ich fahre doch am Montag nach Hamburg.* vs.
Ich fahre DOCH am Montag nach Hamburg. i.e.
I will go on Monday to Hamburg. vs.
I will go on Monday to Hamburg after all.

Sentence mood:

(3) *Treffen wir uns bei Ihnen?* vs.
Treffen wir uns bei Ihnen! i.e.
Do we meet at your place? vs.
Let us meet at your place!

Boundaries and sentence mood:

(4) *Dann machen wir das vielleicht. Ab dem sechsten geht das.* vs.
Dann machen wir das. Vielleicht ab dem sechsten? Geht das? i.e.
Perhaps, we should do that. It is possible after the sixth. vs.
Then let's do it that way. Maybe after the sixth? Is that possible?

Example (4) illustrates one reason why the extraction of prosodic features, their classification into prosodic classes, and the use of these classes in automatic speech understanding is not an easy task: the marking of the boundary between *sechsten* and *geht* interferes with the marking of the sentence mood *question*. All examples highlight an aspect where prosody can probably help the most in spoken dialog systems: especially in spontaneous speech the interpretation of the speech signal becomes an enormous search problem, because spontaneous speech often contains elliptic sentences. As a consequence, when parsing an utterance with a grammar for sentences, after practically each word we have to start a new analysis as well as continue with the old analysis. In order to find all the words which were uttered we have to consider several hypotheses for one spoken word due to recognition errors (typically an order of magnitude more, i.e. 10–20 hypotheses/word). Even if sometimes, semantic/pragmatic constraints might rule out certain readings/meanings, i.e. one of the two reading is implausible in the context of the application/surrounding, prosody can still be very helpful if the computation of constraints from other knowledge sources is more expensive than the computation of prosodic information.

The main reasons, why the use of prosody in dialog systems is not easy, are: First, it is not clear at all how many prosodic classes, e.g., two, three or more boundaries, should be distinguished. Second, segmental (i.e. word chain) and suprasegmental (i.e. prosodic) information influence each other. Third, the different prosodic functions which are realized to a great extent with the same prosodic parameters interfere with each other. Forth, there is a trading relation between prosodic parameters, where the smaller value of one parameter can be compensated by a greater value of another parameter. Fifth, the use of prosodic means is optional: a specific function *can* be expressed with prosody but it does not have to, e.g., when other grammatical means are already sufficient (as in wh-questions). Sixth, the use of

prosodic features is speaker- and language-specific. Thus, even though the number of research projects on prosody in the context of automatic speech recognition/understanding has increased steadily over the past ten years, cf. Price et al. (1991), Wang and Hirschberg (1992), Shriberg et al. (1998), Verbmobil is still—to our knowledge—the world wide first and so far only complete speech understanding system, where prosody is really used, cf. Kompe (1997), Block (1997).

2 Phenomena and Annotation

Consider the following excerpt from a real Verbmobil turn (translated into English), where

<A>	stands for breathing,
w<L>	for unusual lengthening of word *w*,
<P>	for a pause,
B*i*	for acoustic prosodic boundary
D3	for a dialog act boundary, and
M3	for a syntactically motivated boundary:

(see below for details w.r.t. the boundary classes)

(5) ... M3 D3 *well then I'm not present at all* B3 M3 D3 <A> *and in the*<L> B9 <P> *thirty fourth week* B3 M3 <P> <A> *that would be* B3 <P> *Tuesday* B2 *the twenty third* B3 <A> *and Thursday the twenty fifth* M3 D3 <P> ...

2.1 Acoustic-Prosodic Boundaries

Clearly, a classifier which segments this turn based only on acoustic prosodic information, like length of a pause between words, might provide suboptimal prosodic markers for the syntactic and semantic analyses (like the boundary between *in the* and *thirty*). We distinguish therefore between B0: normal word boundary; B2: intermediate phrase boundary with weak intonational marking; B3: full boundary with strong intonational marking, often with lengthening; B9: "agrammatical" boundary, e.g., hesitation or repair. Thus we can distinguish between prosodic boundaries which correspond to the syntactic structure and others which contradict the syntactic structure. However we still have the problem that syntactic boundaries do not have to be marked prosodically. A detailed syntactic analysis would rather have syntactic boundaries irrespective of their prosodic marking, e.g. it needs to know about B9 and B0 in order to favor continuing the ongoing syntactic analysis rather than assuming that a sentence equivalent ended and a new analysis has to be started. Depending on—among other things—the speaker's style, the speaker is sometimes inconsistent with his/her prosodic marking. In the example above, the intermediate boundary between *Tuesday* and *the twenty third* is clearly audible, whereas there is no boundary between *Thursday* and *the twenty fifth*. Syntactic phrasing is—besides by the prosodic marking–also indicated by word order.

2.2 Syntactic-Prosodic Boundaries

For the syntactic boundary classification we have the demand for large training databases, just like in the case of training language models for word recognition. The marking of perceptual labels is rather time consuming, since it requires listening to the signal. We therefore developed a rough syntactic prosodic labelling scheme, which is based purely on the orthographic transliteration of the signal, the so called M system. The scheme is described in detail in Batliner et al. (1998). It classifies each turn of a spontaneous speech dialog in isolation, i.e. does not take context (dialog history) into account. Each word is classified into one of 25 classes in a rough syntactic analysis. For the use in the recognition process, the 25 classes are grouped into three major classes: M3: clause boundary (between main clauses, subordinate clauses, elliptic clauses, etc.); M0: no clause boundary; MU: undefined, i.e. M3 or M0 cannot be assigned to this word boundary without context knowledge and/or perceptual analysis (obviously, only prosodic marking or computationally more expensive knowledge based context modelling can help here in an automatic analysis). For use in the final syntactic and semantic modules in Verbmobil, the 25 M subclasses were mapped onto five syntactic "S" boundary classes which can be described in an informal manner as follows: S0: no boundary, S1: at particles, S2: at phrases, S3: at clauses, S4: at main clauses and at free phrases.

2.3 Dialog Act Boundaries

Even less labelling effort and formal linguistic training is required if we label the word boundaries according to whether they mark the end of a semantic/pragmatic unit. We refer to these boundaries as dialog act boundaries. Dialog acts (DAs) are defined based on their illocutionary force, i.e. their communicative intention, cf. Searle (1969). DAs are, e.g., "greeting", "confirmation", and "suggestion"; a definition of DAs in Verbmobil is given in Jekat et al. (1995), Mast et al. (1995). In parallel to the B and M labels we distinguish between D3: dialog act boundary, and D0: no dialog act boundary. The recognition of these two classes is done in the same way as the recognition of the syntactic classes.

2.4 Irregular Boundaries

In analogy to the other main boundaries, several irregular boundaries marking disfluencies are annotated and mapped onto the two main classes I0: no irregular boundary, and I3: irregular boundary. The processing of this type of boundaries is described in Spilker et al., in this volume.

2.5 Phrase Accents

We distinguish between four different types of syllable based phrase accent labels which can easily be mapped onto word based labels denoting if a word is accented or not: PA: primary accent; SA: secondary accent; EC: emphatic or contrastive

accent; A0: any other syllable (not labelled explicitly). Since the number of PA, SA, EC labels is not large enough, to distinguish between them automatically, we only ran experiments trying to classify "accented word" (A3 = {PA, SA, EC}) vs. "not accented word" (A0). In the Verbmobil domain, the number of emphatic or contrastive accents is not very large. In information retrieval dialogs this could easily change, if there is a large number of misunderstandings and corrections.

In analogy to the syntactic-prosodic M boundaries, phrase accents are also annotated based on the Part of Speech (POS) sequence in a syntactic phrase. For this, we developed a rule-based system which is described in Batliner et al. (1999).

2.6 Sentence Mood

Sentence mood can be marked by means like verb position, words as wh-words, morphology, or prosody. In Verbmobil, we implemented a prosodic classifier for the distinction question Q3 vs. non-question Q0.

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 Kießling (1997).

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 POS flags as features, cf. Batliner et al. (1999). We

use a total of 121 features which can be sub-categorized as follows: 36 F0, 35 energy, 16 duration, 4 pause, and 30 POS features. These 121 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 seven words. Thus the classifier is able to learn a simple 7-gram language model. In section 2 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 (1992). 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 Wightman (1992). 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 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 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 Wightman (1992), Bagshaw (1994), and Kießling (1997). 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 of the recognizer modules. 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.

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 1. 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 1 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.

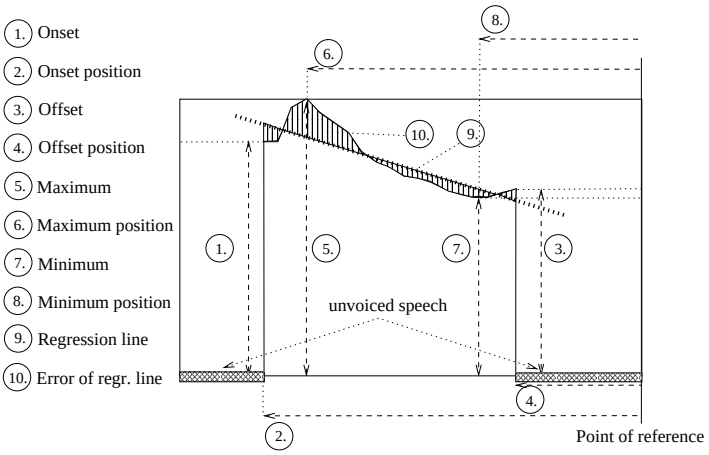

Fig. 1. Example of features used to describe a pitch contour

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 Part of Speech (POS) flag is assigned to each word in the lexicon, cf. Batliner et al. (1999). We include a flag for each of 15 POS classes (for German) or 10 POS classes (for English) and a context of seven 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.

4 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 2.

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 Klüter et al., in this volume. Input into the prosody module is the

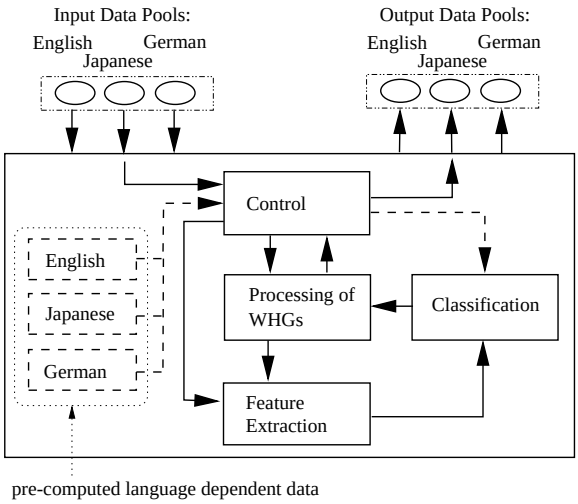

Fig. 2. Architecture of the multilingual prosody module for prosodic processing

speech signal and the word hypotheses graph (WHG), output is an annotated WHG, now including additional prosodic information for each word. Furthermore, a set of prosodic features is passed onto the synthesis module. 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, for instance, specific combinations of neural network classifiers and language model classifiers.
- 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 neural networks which can be combined with language models, cf. section 2. 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 be easily extended as mentioned above. 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).

5 Classification

The classification procedures of the prosody module can be categorized into two classes. The first is the *neuronal network* (NN) using prosodic features as input and the second is the *language model* (LM) depending on textual information as input. Eventually we added POS features to the prosodic feature vectors taking textual information during the NN classification into account.

5.1 Prosodic Classification with NN

In the prosody module a *Multi Layer Perceptron* (MLP) is used as a NN 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., A3, B3 and D3, and their complement, e.g., A0, B0 and D0, see Section 2 for details. 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. Zell et al. (1991b), Zell et al. (1991a). 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 $[0 \ldots 1]$; 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.

5.2 Textual Classification with LM

The second kind of classifier used in the prosody module is a LM classifier. A certain kind of n-gram LM—so called polygrams, cf. Schukat-Talamazzini et al. (1997)—are used for the classification of prosodic events such as syntactic-prosodic phrase boundaries, dialog act boundaries, and phrase accent. Polygrams are a set of n-grams with varying size of n. They are superior to standard n-gram models because n can be chosen arbitrarily large and the probabilities of higher order n-grams are interpolated by lower order ones. The interpolation weights are optimized using the EM algorithm or the unconstrained gradient ascent depending on the used interpolation method. There are several interpolation methods possible for the polygrams, which are described in detail in Schukat-Talamazzini et al. (1997).

For the classification of prosodic events, LMs have to be trained, which model the probability for the occurrence of an event by assigning a label after the current

word given the neighboring words, cf. Kompe (1997). For each word of a spoken word chain, symbol sequences

$$\ldots w_{i-2} w_{i-1} w_i v_i w_{i+1} w_{i+2} \ldots$$

are considered, where w_i denotes the i-th word in the spoken word chain and v_i indicates a prosodic event or no event. Note that theoretically, the sequences

$$\ldots w_{i-1} v_{i-1} w_i v_i w_{i+1} v_{i+1} \ldots$$

should be modeled; experiments showed, however, that this yields worse results. In this case the polygram obviously is not able to cover a sufficiently large word context. The classification of prosodic events as dialog act boundaries D3 vs. normal word boundaries D0 is done by computing the probabilities

$$P(w_{i-2} w_{i-1} w_i \mathsf{D3}\ w_{i+1} w_{i+2})$$

$$P(w_{i-2} w_{i-1} w_i \mathsf{D0}\ w_{i+1} w_{i+2})$$

and adding the probabilities to the WHG. Furthermore it is possible to combine the probabilities of the NN and LM classifier for the prosodic events. Thus recall for these events can be improved (see Section 5.3) when they are combined. The combination is done using empirically estimated weights μ_j:

$$P(v_j) = \mu_j P(v_j \mid NN) + (1 - \mu_j) P(v_j \mid LM)$$

5.3 Experiments and Results

As the effort needed for annotation differs considerably for the different prosodic events, cf. Batliner et al. (1998), 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 Tables 1 to 3, 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 dialog, cf. column "set: dialogs/speaker" in Table 1, train and test speakers for the NN classification were kept disjunct only for German. For the German and English databases used for the NN 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 NN 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

Table 1. NN 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 dialogs, speakers, and cases is given for train and test. Note that for questions only sentence boundaries are considered

Language	set: dialogs/speakers	B3	B0	A3	A0	Q3	Q0
German	# 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
English	# 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
Japanese	# train: 24/20	747	5348	1545	4889	-	-
	# test: 19/18	67	558	165	497	-	-
	% recall – POS	81	89	75	71	-	-

granularity of categorization. The best results we achieved and integrated into the Verbmobil system can be found in Table 1.

Even if it is possible to train NNs with more classes, for the prosodic events A, B and Q, we used only two because more classes yielded worse results due to sparse data. The LM classifiers were trained for the prosodic events M, A and D; results are given in Table 2. Note that here, the reference phrase accent is the rule-based version computed from the POS sequence in a syntactic phrase, cf. Batliner et al. (1999), not the perceptive one used within the NN classifier. If no results are given in Tables 1 and 2, computation was not possible, for instance, due to the small amount of data available. 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 NN except for English questions, where the database is very small. Possibly due to the larger amount of training data, LM classification for German boundaries and accents is better than the NN classification; it might as well be that the "syntactic behavior" of the German speakers is more regular than their prosodic one. For the English boundaries, however, it is the other way round. The NN classification for the German boundaries without POS information is in the same range as reported in Kießling (1997) for an NN classification with the old feature set that used, besides word-based information, phoneme-based information as well.

Due to the large training database, for the syntactic-prosodic M boundaries it was possible to cluster the 25 basic labels into the five classes described above in Section 2. These new classes for German and English are integrated into the Verbmobil system. Results are given in Table 3.

If we combine the output of the NN with the output of the LM, results are slightly better for boundaries and accents. In spite of that, we pass over both results sepa-

Table 2. LM classification: Recall in percent for syntactic-prosodic boundaries M, rule-based accents A, and dialog act boundaries D in the three languages of the Verbmobil system; number of cases is given for train and test. Note that for these different phenomena different amount of training data were available

Language	set	M3	M0	A3	A0	D3	D0
German	# train	26633	126439	102730	174418	14928	99410
	# test	4967	23663	2761	4977	4673	25538
	% recall	86	97	87	92	80	96
English	# train	15826	52656	–	–	–	–
	# test	1813	6069	–	–	–	–
	% recall	83	94	–	–	–	–
Japanese	# train	–	–	–	–	13648	93642
	# test	–	–	–	–	1213	7954
	% recall	–	–	–	–	92	99

rately, because several higher linguistic modules in the Verbmobil system only use either the NN or the LM output.

Table 3. Recall in percent for the five S classes, German: left, English: right

reference		**recognized**				
Label	#	S0	S1	S2	S3	S4
S0	24286	**89**	2	5	2	2
S1	1408	8	**81**	4	2	5
S2	1014	15	3	**69**	3	10
S3	622	8	2	5	**73**	12
S4	3640	4	5	6	6	**79**

reference		**recognized**				
label	#	S0	S1	S2	S3	S4
S0	5771	**89**	1	6	2	2
S1	169	7	**64**	17	0	12
S2	900	5	3	**83**	2	8
S3	145	7	1	7	**71**	14
S4	1066	3	8	9	3	**76**

5.4 Integrated Classification

Several knowledge sources can be used at the same time for classification and segmentation of turns by using an integrated search procedure, i.e., segmentation and classification in one step, rather than a sequential approach, i.e., first segmentation into segments, and second, classification of these segments. For this task it is necessary to build classifiers that can compute the output successively to allow a weighted scoring. For the task of segmentation and classification of dialog acts, a first prototype was presented in Warnke et al. (1999): The segmentation classifiers for syntactic-prosodic M boundaries, dialog act boundaries D and prosodic boundaries B were combined with a dialog act classifier and a classifier for dialog act sequences. The weights for combining the scores of the different classifiers can be optimized automatically with gradient descent. During optimization the precision

of the dialog act segmentation improved from 57% to 69% with a recall of 88%. Recall for 18 dialog acts[1] improved from 52.4% to 64.6%. The results for a sequential approach using the same classifier achieved a segmentation precision of 71% having a recall of 73%; recall for dialog act classification was 62%. It can thus be seen that this integrated classification yields better results for the segmentation and classification of dialog acts.

5.5 Use in the Verbmobil System

The Verbmobil system uses prosodic information in several different modules. The syntactic-prosodic boundaries clustered in the five S classes are used together with the acoustic-prosodic boundaries to segment the best hypotheses into "utterances" used for dialog act segmentation, cf. Reithinger and Engel and Kipp et al., in this volume. The deep syntactic analysis, cf. Kiefer et al., in this volume, reduces the search space of different readings by taking the syntactic-prosodic S boundaries into account. The phrase accent is used in the semantic module, cf. Heine and Bos, in this volume, to disambiguate the different meanings of some particles depending on whether they are accented or not; in the semantic module, the prosodic classification of questions is used if other linguistic information cannot disambiguate between the two different readings question/non-question. For statistic translation, cf. Vogel et al., in this volume, prosodic boundary and question classification is also taken into account.

6 Conclusion

The prosody module used in the Verbmobil system in the first phase of the project is described in Kompe (1997) and Kießling (1997). In the second phase, a new feature set was developed that does not use phoneme-based but only word-based information; thus considerable overload can be avoided without any loss of classification performance because we do no longer need a special time alignment for phones. For the two other languages in Verbmobil besides German, English and Japanese, prosodic feature sets and classifiers were developed; the multilingual architecture reduces memory requirement by a factor of five (290MB to 64MB) in comparison with three monolingual modules. A rule-based system for the annotation of phrase accent position was added; an LM trained with this information yielded considerably better classification results, cf. Batliner et al. (1999). A first prototype was developed for the integrated segmentation and classification of dialog acts.

References

Alexandersson, J., Engel, R., Kipp, M., Koch, S., Küssner, U., Reithinger, N., and Stede, M. Modeling Negotiation Dialogs. In *this volume.*

[1] These 18 dialog acts are clustered and contain all 32 Verbmobil dialog acts (see Alexandersson et al., in this volume).

Bagshaw, P. C. (1994). *Automatic Prosodic Analysis for Computer Aided Pronunciation Teaching.* PhD thesis, University of Edinburgh.

Batliner, A., Huber, R., Niemann, H., Nöth, E., Spilker, J., and Fischer, K. The Recognition of Emotion. In *this volume.*

Batliner, A., Kompe, R., Kießling, A., Mast, M., Niemann, H., and Nöth, E. (1998). M = Syntax + Prosody: a Syntactic-Prosodic Labelling Scheme for Large Spontaneous Speech Databases. *Speech Communication* 25(4):193–222.

Batliner, A., Nutt, M., Warnke, V., Nöth, E., Buckow, J., Huber, R., and Niemann, H. (1999). Automatic Annotation and Classification of Phrase Accents in Spontaneous Speech. In *Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH 99)*, 519–522.

Block, H. (1997). The Language Components in Verbmobil. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 1, 79–82.

Heine, J., and Bos, J. Discourse and Dialog Semantics for Translation. In *this volume.*

Jekat, S., Klein, A., Maier, E., Maleck, I., Mast, M., and Quantz, J. (1995). Dialogue Acts in Verbmobil. Verbmobil Report 65.

Kiefer, B., Krieger, H.-U., and Nederhof, M.-J. Efficient and Robust HPSG Parsing of Word Graphs. In *this volume.*

Kießling, A. (1997). *Extraktion und Klassifikation prosodischer Merkmale in der automatischen Sprachverarbeitung.* Berichte aus der Informatik. Aachen: Shaker Verlag.

Kipp, M., Alexandersson, J., Reithinger, N., and Engel, R. Dialog Processing. In *this volume.*

Kompe, R. (1997). *Prosody in Speech Understanding Systems.* Lecture Notes for Artificial Intelligence. Berlin: Springer-Verlag.

Mast, M., Maier, E., and Schmitz, B. (1995). Criteria for the Segmentation of Spoken Input into Individual Utterances. Verbmobil Report 97.

Klüter, A., Ndiaye, A., and Kirchmann H. Verbmobil from a Software Engineering Point of View: System Design and Software Integration. In *this volume.*

Price, P., Ostendorf, M., Shattuck-Hufnagel, S., and Fong, C. (1991). The Use of Prosody in Syntactic Disambiguation. *Journal of the Acoustic Society of America* 90:2956–2970.

Reithinger, N., and Engel, R. Robust Content Extraction for Translation and Dialog Processing. In *this volume.*

Schukat-Talamazzini, E., Gallwitz, F., Harbeck, S., and Warnke, V. (1997). Rational Interpolation of Maximum Likelihood Predictors in Stochastic Language Modeling. In *Proc. European Conf. on Speech Communication and Technology*, volume 5, 2731–2734.

Searle, J. (1969). *Speech Acts. An Essay in the Philosophy of Language*. Cambridge: Cambridge University Press.

Shriberg, E., Bates, R., Taylor, P., Stolcke, A., Jurafsky, D., Ries, K., Cocarro, N., Martin, R., Meteer, M., and Ess-Dykema, C. V. (1998). Can Prosody Aid the Automatic Classification of Dialog Acts in Conversational Speech? *Language and Speech* 41:439–487.

Spilker, J., Klarner, M., and Görz, G. Processing Self Corrections in a Speech-to-Speech System. In *this volume.*

Vogel, S., Och, F.J., Tillmann, C., Niessen, S., Sawaf, H., and Ney, H. Statistical Methods for Machine Translation. In *this volume.*

Wang, M., and Hirschberg, J. (1992). Automatic Classification of Intonational Phrase Boundaries. *Computer Speech & Language* 6(2):175–196.

Warnke, V., Gallwitz, F., Batliner, A., Buckow, J., Huber, R., Nöth, E., and Höthker, A. (1999). Integrating Multiple Knowledge Sources for Word Hypotheses Graph Interpretation. In *Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH 99)*, 235–239.

Wightman, C. (1992). *Automatic Detection of Prosodic Constituents.* PhD thesis, Boston University Graduate School.

Zell, A., Mache, N., Sommer, T., and Korb, T. (1991a). Design of the SNNS Neural Network Simulator. In *Proceedings of the Österreichische Artificial-Intelligence-Tagung*, Informatik-Fachberichte 287, 93–102. Springer Verlag.

Zell, A., Mache, N., Sommer, T., and Korb, T. (1991b). The SNNS Neural Network Simulator. In *Proceedings of the 15. Fachtagung für Künstliche Intelligenz*, 254–263. Springer Verlag.