PROSODIC INFORMATION FOR INTEGRATED WORD-AND-BOUNDARY RECOGNITION

F. Gallwitz *H. Niemann* *E. Nöth* *V. Warnke*

University of Erlangen-Nuremberg
Chair for Pattern Recognition
Martensstr. 3, 91058 Erlangen, Germany
gallwitz@informatik.uni-erlangen.de

ABSTRACT

In this paper, we present an integrated approach for recognizing both the word sequence and the syntactic-prosodic structure of a spontaneous utterance. The approach aims at improving the performance of the understanding component of speech understanding systems by exploiting not only acoustic and syntactic information, but also prosodic information directly within the speech recognition process. Whereas spoken utterances are commonly modelled as unstructured word sequences in the speech recognizer, our approach includes phrase (or clause) boundary information in the language model, and provides HMMs to model the acoustic and prosodic characteristics of phrase boundaries and disfluencies. This methodology has two major advantages compared to pure word–based speech recognizers. First, additional syntactic information is determined by the speech recognizer which facilitates parsing and resolves syntactic and semantic ambiguities. Second, the integrated model yields significantly better word accuracies than the traditional word–based approach.

1. INTRODUCTION

In spoken language, especially in spontaneous speech, prosodic boundaries are of similar importance for understanding an utterance as punctuation marks are in written language. Words which "belong together" from the point of view of meaning are grouped into *prosodic phrases*, and it is widely agreed upon that there is a high correspondence between prosodic and syntactic phrase boundaries [13, 5].

Prosodic boundaries are often marked by silence periods, and sometimes by filled pauses, such as "uh", and they are usually indicated by specific energy and fundamental frequency (F0) contours and by durational variations of the surrounding syllables [4]. Also, as punctuation marks in written language, they are partly indicated by word order.

In automatic speech understanding, this information may be important even in the context of a comparatively simple application, such as an automatic train timetable information system. Consider, for example, the following user utterances:

U1: *Of course not on Monday.*
U2: *Of course not. On Monday!*

This work was funded by the DFG (German Research Foundation) under contract number 810 939-9 and by the German Federal Ministry of Education, Science, Research and Technology (*BMBF*) in the framework of the VERBMOBIL Project under the Grant 01 IV 701 K5. The responsibility for the contents of this study lies with the authors.

The question whether a prosodic phrase boundary occurred after the word "*not*" is of considerable importance for the semantic interpretation of the word sequence and for determining the next system utterance. Depending on the phrasing, one of the following two utterances may be appropriate:

S1: *What day would you like to travel?*
S2: *You would like to travel on Monday?*

Selecting the wrong response (**S1** for **U2**, or **S2** for **U1**) will most certainly annoy the caller and will probably make her/him hang up.

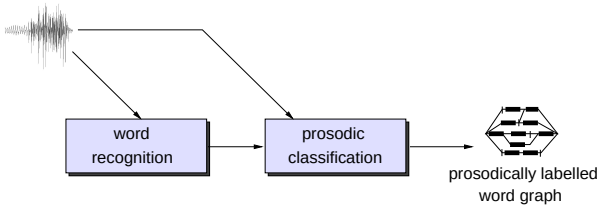
It might be argued that the correct interpretation of the word sequence could also be determined without prosodic information, if the dialogue history is taken into account. Depending on the previous system utterance, at least one of the two above interpretations could be declared illogical. This involves a considerable amount of higher–level knowledge and "intelligent" processing, however, whereas prosodic information in the speech signal can directly resolve the ambiguity. Furthermore, there is no reason to ignore information that may without a doubt contribute to finding the correct semantic interpretation, even if a sufficiently intelligent dialogue module is available [5, Sec. 8.4].

The first speech understanding system to really integrate prosodic information into the understanding process is the German VERBMOBIL speech–to–speech translation system for appointment scheduling dialogues [12, 2]. In the VERBMOBIL prototype, prosodic information is calculated on the basis of the speech signal and the word recognition result. This information is used in various system modules, mainly for resolving syntactic and semantic ambiguities, and has been shown to significantly improve the total system performance [5]. For example, VERBMOBIL is able to provide different English translations for German utterances that contain the same word sequence but are prosodically distinct [5]:

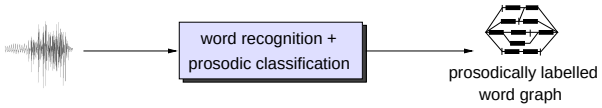
Ja zur Not geht's auch am Samstag.
(*Well, if necessary, Saturday is also possible.*)

Ja. Zur Not. Geht's auch am Samstag?
(*Okay. If necessary. Is Saturday possible as well?*)

Speech recognition and prosodic analysis are performed in two separate modules, however, and the speech recognizer itself is only concerned with finding an optimal sequence of words (or a *word graph*, a graph of competing word hypotheses) that covers the whole speech signal. That is, the prosodic and the syntactic structure of the utterance are neither determined nor taken into account by the speech recognizer. The basic structure of the word recognition and prosody classification modules in VERBMOBIL is



(a) Sequential approach



(b) Integrated approach

Figure 1: The sequential approach to word recognition and prosody classification that has been successfully applied in the VERBMOBIL speech–to–speech translation system [5, 4] and the integrated approach proposed in this paper.

depicted in Figure 1 (a).

We believe that syntactic-prosodic boundary information is also useful in an earlier stage of spontaneous speech processing. It is well known that state of the art speech recognizers are based on two sources of knowledge: acoustic information and language model information. Statistical language models provide the probability of a given word sequence based on a rather simple model: it is assumed that a spoken utterance is an unstructured sequence $w_1, w_2, ...w_n$ of words. Obviously, this is not the case. It is intuitively clear that words at the beginning of a new phrase correlate less strongly with the last word of the preceding phrase than words within the same phrase.

A similar effect has also been found in the neighborhood of filled pauses [9]. As a consequence, a language model for spontaneous speech is proposed in [10], where different types of disfluencies (filled pauses, repetitions, and deletions) are predicted, and probabilities of following words are estimated on the basis of the fluent word sequence that was supposedly intended by the speaker. This approach, however, did not have a significant impact on the recognition accuracy. One of the reasons for this result is noted in [10]: phrase (or clause) boundaries grossly violate the assumptions of the proposed model, because filled pauses strongly correlate with boundaries of linguistic segments. Thus, 'cleaning up' the surrounding words to remove the disfluency can be counterproductive.

In our approach, phrase boundaries are directly integrated into the language model, and filled pauses are allowed to occur in two different functions: Either they are syntactically insignificant and thus ignored in the language model ('clean–up'), or they occur at phrase boundaries. Furthermore, phrase boundaries are also allowed to occur at fluently spoken word–word transitions. The fact that a word is separated from its predecessor by a phrase boundary should contribute a great amount of information when language model probabilities are calculated, while the preceding word is less significant. By integrating models for syntactic-prosodic phrase boundaries into the word recognizer and into the statistical language model, the word recognizer can incorporate information about the structure of the utterance. An integrated model of sequences of words and boundaries allows for a distinction between word transitions across phrase boundaries and transitions within a phrase, which is an obvious advantage.

An entertaining but representative example that clearly shows the advantages of an *integrated* processing of word information and prosodic information as proposed in this paper is given in [7]:

A: *What is that in the road ahead?*
B: *What is that in the road? A head?*

Here, not just the semantic interpretation, but also the *word sequence* depends on the prosodic structure of the utterance. That is, if prosodic information is taken into account in this example, it will be considerably more helpful if it is integrated into the word recognition process.

In phrase boundary recognition experiments based on word recognizer results, it has been shown that prosodic features can significantly improve the detection accuracy of syntactic phrase boundaries compared to a pure language model based approach [5]. This is especially the case with syntactically ambiguous boundaries, as in the above example utterances. In this paper, we investigate how prosodic information can be incorporated into our integrated approach to recognize words and syntactic-prosodic boundaries. In earlier experiments, only the baseline mel–cepstral feature set has been used, and no additional prosodic information has been incorporated [3].

As the feature set used for our separate boundary classifier is not suitable for the system architecture of the integrated word–and–boundary recognizer, we developed new frame based prosodic feature sets that incorporate information on the fundamental frequency and energy contours as well as durational information. These features are used as input to an ANN in order to calculate the prosodic probability of a phrase boundary for each time frame. The resulting probabilities are then utilized as a second input stream to the HMM based recognizer, in addition to the acoustic–phonetic probabilities that are based on a cepstral feature vector and a Gaussian codebook. Thus, the integrated recognizer combines three sources of information: acoustic–phonetic information, prosodic information, and language model information.

The remainder of this paper is structured as follows: In Section 2, we briefly describe the phrase boundary labelling system that was used as a basis of our experiments. In Section 3, the treatment of phrase boundaries during training and recognition in our approach is described. In Section 4, a hybrid HMM–MLP system architecture is presented that incorporates prosodic features into the recognition process. The prosodic feature sets employed in our experiments are described in Section 5. The training procedure of the hybrid speech recognizer is then discussed in Section 6. Finally, experimental results are given in Section 7. The paper closes with a brief summary of the main results.

2. SYNTACTIC-PROSODIC BOUNDARIES

Starting point for the annotation of our material with syntactic–prosodic labels was the assumption that there is a strong – albeit not perfect – correlation between syntactic phrasing and prosodic

phrasing, cf. [7, 11, 8]. This assumption could be corroborated earlier in experiments with German read speech where similar labels could be used successfully for the training of prosodic classifiers, cf. [6]. In order to save time, we annotated these boundaries only using the written word chain. The 'syntactic-prosodic' boundaries relevant for our present purpose – we called them M3-boundaries – are those syntactic boundaries that are expected to be marked prosodically, as can be seen in the following example:

perhaps I should first introduce myself M3 *my name is Lerch*

In the VERBMOBIL data, the average length of a prosodic phrase between two M3-labels is 5.4 words, while the average turn length is 22 words. Details on the data used in our experiments are given in Section 7. More details on our labelling scheme can be found in [1].

3. BASIC APPROACH

The basic idea behind our approach is that phrase boundaries should be treated in the language model (LM) in a similar fashion as words. Thus, we provide a language model category (or word class) for phrase boundaries in the n–gram LM, and we provide HMMs to model the acoustic and prosodic characteristics of phrase boundaries.

In [5], it has been shown that the syntactic-prosodic boundaries often happen to occur in combination with non-verbal noises, pauses or filled pauses. This makes it desirable to exploit the information that is provided regarding the correlation between boundaries on the one hand and pauses, filled pauses, and non–verbals (NV) on the other hand. Thus, we incorporate this information by training suitable LM category emission probabilities for different non–verbal phenomena which occur at phrase boundaries.

Furthermore, we assume that silence periods and non–verbals within phrases and across phrase boundaries can be discriminated based on acoustic and prosodic information. Thus, different HMMs are trained for pauses and non–verbals within phrases, and across phrase boundaries. For example, two different HMMs are used for the two occurrences of the filled pause '*uh*' in the utterance

From Munich uh I want to travel on uh Saturday.

The first '*uh*' is modelled by the specific boundary model named 'M3-uh', which is trained on all occurrences of '*uh*' at phrase boundaries, and the second '*uh*' is trained on all occurrences of '*uh*' within phrases.

We train HMMs for several combinations of boundaries and non-verbals, and include them in the statistical language model according to their syntactic function: Non–boundary models for pauses and non–verbals are skipped in the language model and boundary models are treated like words, both during training and decoding.

A special situation arises in the case of a phrase boundary that does not correspond to a filled pause or a non–verbal. Here, we provide a one–state HMM that always consumes one time frame. By consuming one time frame, the recognizer can incorporate information on the acoustic and prosodic characteristics of this type of phrase boundaries. This HMM is also included in the phrase boundary LM category.

During the word recognizer search procedure, several different situations have to be taken into account at transitions from word w_i to word w_{i+1}. The word recognizer implicitly makes a decision for the most probable alternative, based on the language model scores and on the acoustic and prosodic scores of the word and boundary HMMs involved (In the following, we only consider the bigram scores; the higher order language model scores are calculated accordingly):

1. If no boundary or non-verbal is hypothesized, the bigram score $p(w_{i+1} \mid w_i)$ is used.
2. If a M3 boundary is hypothesized (possibly represented by a M3 silence model or a M3 non-verbal model), the bigram scores $p(\mathrm{M3} \mid w_i)$ when entering the M3-model and $p(w_{i+1} \mid \mathrm{M3})$ when entering w_{i+1} are employed.
3. If no boundary, but a non-verbal (NV) or silence period is hypothesized, the constant unigram probability $p(\mathrm{NV})$ is used when entering the NV–model, and $p(w_{i+1} \mid w_i)$ when entering w_{i+1}. Thus, non-verbals or silence periods that do not mark syntactic boundaries are treated as random events that do not depend on the surrounding word context. Consequently, they are ignored when the probability of the following word is calculated.

The search algorithm of the recognizer (e.g. beam-search or $A^\star$ search) will now determine the optimal sequence (or word graph) containing words and boundaries.

4. SYSTEM ARCHITECTURE

The proposed approach can be used with any state–of–the–art HMM–based speech recognizer, irrespective of the specifics of the HMM topology, the type of density, or the decoding algorithm. Only some slight modifications to the decoding algorithm might be necessary, to allow for the treatment of syntactically irrelevant silence–periods and non–verbals as described above. Even without additional prosodic information, the integration of phrase boundaries into the recognition process has been shown to yield improved word accuracies for spontaneous speech recognition [3].

It is our goal, however, to incorporate additional prosodic information into the approach. Prosodic information, e.g. movements of the F0 contour, can help to improve the detection of phrase boundaries, which implicitly — via the statistical LM — might also improve the word accuracy. Furthermore, ambiguous boundaries can only be reliably classified if prosodic information is taken into account, which is especially important when the occurance of a prosodic boundary has an impact on the semantic interpretation of an utterance.

In preliminary experiments, a direct integration of prosodic features into the feature vector used by the word recognizer did not yield any improvement. Instead, there was even a significant decline in word accuracy. The probable reason for this lies in the complex distributional properties of features that are derived from prosodic parameters, such as the second derivative of the F0, which could not be accurately modelled by the Gaussian distributions employed in our word recognizer.

We have therefore developed a hybrid architecture that independently processes acoustic–phonetic and prosodic information on a level close to the signal. Both streams of information are then

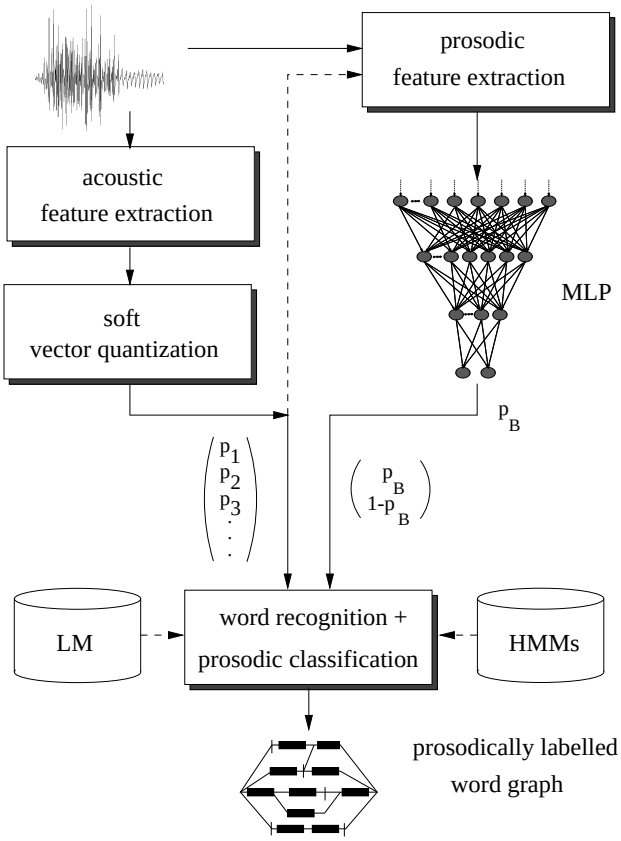


Figure 2: Proposed architecture of an MLP–HMM hybrid system for integrated classification of prosodic boundaries using additional prosodic features.

combined during the recognition process. Acoustic–phonetic information (i.e. mel–cepstral coefficients and their first derivatives) are processed as in our baseline SCHMM recognizer. This involves a soft vector quantization on the basis of a Gaussian codebook. Acoustic–prosodic features are used as input to a multi–layer perceptron (MLP), which estimates the probability of a prosodic boundary in the current frame. The architecture is depicted in Figure 2. The dashed arrow indicates that information extracted from the stream of vector quantization results, e.g. durational information, may be included in the prosodic feature vectors. The two input streams of the word recognizer are treated as stochastically independent during the calculation of the HMM probabilities. A prosodic weight factor (similar to the linguistic weight factor for LM probabilities) is introduced to allow for a balancing of acoustic and prosodic information.

5. PROSODIC FEATURES

The following acoustic parameters are considered to be the most valuable for the classification of prosodic information in ASU [4, p. 67]:

- energy (the acoustic correlate of loudness),
- the fundamental frequency F0 (the acoustic correlate of pitch),
- pause-length,
- and phone duration.

Although there are obviously strong interdependencies between acoustic-phonetic and acoustic-prosodic information, we find it helpful to use the terms *acoustic–phonetic feature* and *acoustic–prosodic feature*. The purpose of acoustic-phonetic features is mainly to incorporate segmental, phonetic information over a short period of time; typically this time is in the order of the mean phoneme duration (about 70ms). Acoustic–prosodic features cover suprasegmental information that is generally included in significantly larger portions of the speech signal, typically one or more syllables or words, or even a whole utterance.

As mentioned above, in the VERBMOBIL system, prosodic features are calculated based on a time alignment of the word recognition result [5]. This approach is now commonly used, because it allows for the incorporation of information about the position of word and syllable boundaries, and for a normalization of the features based on word, syllable, or phoneme information. Unfortunately, this type of feature is not suitable for the integrated approach of recognizing words and prosodic information in one step, for the simple reason that no recognition result can be available before the recognition process even started. Instead, an incremental calculation of features should be possible without having to wait for the end of an utterance. Furthermore, all prosodic features have to be calculated frame–based, and only based on the speech signal (or on information that can be derived from the speech signal efficiently and incrementally, such as the vector quantization result). Thus, we developed a number of frame based suprasegmental features which incorporate specific movements of prosodic parameters. These may indicate the occurrence of prosodic events. For example, it is worth looking at, whether the second derivative of the fundamental frequency, calculated over a fixed–sized window of one or two seconds, gives hints on the position of prosodic boundaries.

One of the prosodic feature sets that was used for our experiments, which is exclusively based on F0 information, is shown in Table 1. Alternatively, we employed a set of 64 features which also included features calculated on the basis of the energy contour, in a similar fashion. As yet, no experiments have been performed which explicitly include durational information. We are currently developing methods for extracting durational information from the result of the soft vector quantization. Obviously, this stream of symbols (with corresponding probabilities) can be used to detect lengthenings and variations in the speaking rate. For this purpose, we calculate a number of values over fixed size windows, such as the average number of frames the best–scoring codebook class stays in the first position. These values, divided by similar values calculated over a significantly larger time interval, can then be included in the prosodic feature set.

6. MLP AND HMM TRAINING

It is not straightforward to define the optimal output of the MLP in the hybrid architecture described above. Ideally, it should provide the phrase boundary probability 1.0 for frames that are associated with a boundary HMM, and zero for non–boundary frames. This is not feasible, however, because prosodic boundaries cannot realistically be associated to one single time frame. Instead, indications for a prosodic boundary should also be expected in the surrounding frames. Restricting the MLP training set for the phrase boundary class to the comparatively small set of time frames which are directly associated to a boundary HMM is certainly sub–optimal.

Our solution to this problem is to define a heuristic goal func-

F_1	$F_0(t)$	F_0 at time t
F_2	$\Delta_t^{10}(F_0)$	delta coefficient at t with context of 10 frames into the future and into the past
F_3	$\Delta\Delta_t^{10}(F_0)$	delta delta coefficient at t with context of 10 frames into the future and into the past
F_4	$\Delta_t^{20}(F_0)$	delta coefficient at t with context of 20 frames into the future and into the past
F_5	$\mathbf{MSE}_t^{20}(F_0)$	mean square error of F_0 and F_0 regression line within a 40 frame context centered at t
F_6	$\Delta\Delta_t^{20}(F_0)$	delta delta coefficient at t with context of 20 frames into the future and into the past
F_7	$\Delta_t^{40}(F_0)$	delta coefficient at t with context of 40 frames into the future and into the past
F_8	$\mathbf{MSE}_t^{40}(F_0)$	mean square error of F_0 and F_0 regression line within a 80 frame context centered at t
F_9	$\Delta\Delta_t^{40}(F_0)$	delta delta coefficient at t with context of 40 frames into the future and into the past
F_{10}	$\Delta_t^{80}(F_0)$	delta coefficient at t with context of 80 frames into the future and into the past
F_{11}	$\mathbf{MSE}_t^{80}(F_0)$	mean square error of F_0 and F_0 regression line within a 160 frame context centered at t
F_{12}	$\Delta\Delta_t^{80}(F_0)$	delta delta coefficient at t with context of 80 frames into the future and into the past

Table 1: A set of 12 F_0-based features used to incorporate suprasegmental information. All features are based on F_0 values that were first transformed on a logarithmic scale and then linearly interpolated in unvoiced parts of the signal. The frame length is 10 ms.

tion for the MLP which is based on the cosine function, as depicted in Figure 3. Each peak corresponds to the first frame of a boundary HMM, when a forced alignment of the transliteration is performed. This goal function is used for training the MLP. During HMM training, the output of the trained MLP is used. In the worst case, the MLP output is not correlated with the occurance of boundaries. This should not degrade the recognition performance compared to a system without prosodic information, however, because the resulting HMM emission probabilities for the boundary and non–boundary classes would be close to 0.5 for all HMM states, which means that the MLP output has no impact on the recognition result. Nevertheless, any correlation of the MLP output with phrase boundaries should improve the recognition results for phrase boundaries, because the HMM emission probabilities are then trained accordingly. The recognition of word HMMs is only affected if certain words typically occur in the neighborhood of phrase boundaries. This effect is expected to have a positive influence on the word recognition performance.

This goal function was also used for evaluating the performance of the hybrid word–and–boundary recognizer in the case of an optimal performance of the MLP boundary classifier (see below).

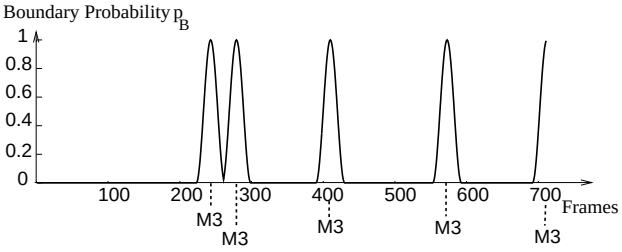


Figure 3: The desired output of the MLP classifier, which was used for training the MLPs, and for the experiments involving 'ideal' boundary classification performance

In this case, the ideal MLP output is used both during training and recognition. The latter, of course, is only possible if the position of phrase boundaries is available for the test sample.

7. EXPERIMENTS AND RESULTS

The experiments reported in this paper have been performed on a subset of the German VERBMOBIL corpus. The training, validation, and test samples are shown in Table 2.

sample	turns	words	M3 phrase boundaries
training	11714	258956	36039
validation	48	1044	137
test	268	4783	768

Table 2: Training, validation, and test data. The figures for phrase boundaries do not contain the trivial boundaries at the beginning or end of a turn.

We use a SCHMM word recognizer with a codebook size of 512 classes. No speaker adaptation is performed and only intra–word subword models (polyphones) are used. A bigram language model is employed in the first pass of the recognition process, and a 4-gram language model in the second pass. The vocabulary size is 2860 words; 6 additional boundary models were used in the experiments involving phrase boundaries.

The word accuracies are calculated based on the word chain, i.e. the boundary labels were removed from the recognizer results. The evaluation of the recognized boundaries is performed in the following manner: First, an alignment based on the minimum Levenshtein-distance criterion is performed between the recognized word chain and the reference transliteration. During this procedure, the boundary labels are treated just like words. Then, all pairs of hypothesized symbols and reference symbols that include at least one boundary are used to calculate precision and recall rates. Note that even a perfect boundary classification will not result in 100% precision and recall if a certain number of word recognition errors is present in the recognition result, because these can lead to a mismatch between the alignment of the reference and the hypothesized word–and–boundary sequence.

The recognition results are given in Table 3. The baseline word recognizer does not include any boundary information; silence periods are ignored in the LM, and filled pauses are treated like words. This setup has been shown to yield optimal performance on this data set when no boundary information is available.

The word error rate for the integrated approach *without* additional prosodic features is about 4 percent lower than that of the baseline system. Furthermore, the boundary information is produced

	WER	Recall	Precision
baseline word recognizer	23.8 %	—	—
integrated w&b recognizer	22.9 %	74.5 %	75.7 %
VM prosodic classifier	—	75.1 %	74.7 %
hybrid w&b recognizer	22.9 %	75.7 %	75.3 %
hybrid with 'ideal' MLP	22.3 %	88.2 %	78.5 %

Table 3: Word error rates (WER), and recall and precision rates for M3 phrase boundaries.

with a precision and recall rate of about 75% for both. For a comparison, we evaluated the prosodic classifier that is integrated into the VERBMOBIL (VM) system (cf. Section 1) on the word chains (after removing the boundary labels) that were produced by the integrated approach[1]. This module uses a MLP classifier based on a set of 276 prosodic features combined with an n–gram language model [5]. The results of this sequential approach are almost identical with that of the integrated approach, which, at this stage, does not make use of any prosodic features.

In the following line, the results for our MLP–HMM hybrid architecture are given, which is based on a set of 64 prosodic features calculated on the basis of the energy and F0 contour. No improvement in word accuracy is obtained, but a slight improvement in the overall M3 classification performance: recall is improved by 1.2 percent, whereas precision is degraded by only 0.4 percent. This improvement is not statistically significant, however.

To evaluate the approach in the case of an ideal MLP classifier for M3 boundaries, we used the goal function for the MLP training both during training and recognition (see Section 6). This result can be regarded as an upper limit for improvements that can be achieved by optimizing the prosodic classifier within the given architecture, and without modifying the word HMMs. The drastic improvement in boundary classification rate is not surprising, because information about the position of phrase boundaries in the test sample is incorporated in this approach. Furthermore, a further relative reduction of word error rate by about 3 percent is obtained. This result indicates that additional knowledge about the position of phrase boundaries does improve word recognition, but it does not do so dramatically.

8. SUMMARY AND CONCLUSION

In this paper, we presented an integrated approach for the recognition of words and prosodic phrase boundaries. Furthermore, we described how prosodic information can be incorporated into the approach by employing a MLP–HMM hybrid recognizer and a frame–based suprasegmental feature set.

The largest relative improvement in word recognition could be achieved without prosodic information, simply by including models for phrase boundaries in the vocabulary, and in the statistical language model. Introducing a MLP classifier for phrase boundaries based on suprasegmental energy and F0 features slightly enhances the boundary classification performance, but does not improve the word accuracy. An experiment with the ideal MLP output indicates that knowledge about the position of phrase boundaries in the test data only slightly improves the word accuracy compared to the integrated word–and–boundary recognizer which does not incorporate this information. There is, however, a large potential of further increasing the boundary classification performance by enhancing the prosodic feature set.

[1] These word chains contain less errors than those produced by the baseline recognizer. We wanted to exclude this source of errors for the VM prosodic classifier, however, to directly compare the M3 classification performance with that of the integrated word–and–boundary recognizer.

9. REFERENCES

1. 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*, 25:193–222, 1998.
2. Thomas Bub and Johannes 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.
3. F. Gallwitz, A. Batliner, J. Buckow, R. Huber, H. Niemann, and E. Nöth. Integrated Recognition of Words and Phrase Boundaries. In *Int. Conf. on Spoken Language Processing*, volume 7, pages 2883–2886, Sydney, 1998.
4. A. Kießling. *Extraktion und Klassifikation prosodischer Merkmale in der automatischen Sprachverarbeitung*. Berichte aus der Informatik. Shaker Verlag, Aachen, 1997.
5. R. Kompe. *Prosody in Speech Understanding Systems*. Lecture Notes for Artificial Intelligence. Springer–Verlag, Berlin, 1997.
6. R. Kompe, A. Batliner, A. Kießling, U. Kilian, H. Niemann, E. Nöth, and P. Regel-Brietzmann. Automatic Classification of Prosodically Marked Phrase Boundaries in German. In *ICASSP*, volume 2, pages 173–176, Adelaide, 1994.
7. W. Lea. Prosodic Aids to Speech Recognition. In W. Lea, editor, *Trends in Speech Recognition*, pages 166–205. Prentice–Hall Inc., Englewood Cliffs, New Jersey, 1980.
8. P.J. Price, M. Ostendorf, S. Shattuck-Hufnagel, and C. Fong. The Use of Prosody in Syntactic Disambiguation. *Journal of the Acoustic Society of America*, 90:2956–2970, 1991.
9. E. Shriberg and A. Stolcke. Word Predictability After Hesitations: A Corpus Based Study. In *Int. Conf. on Spoken Language Processing*, volume 3, pages 1868–1871, Philadelphia, 1996.
10. Andreas Stolcke and Elizabeth Shriberg. Statistical Language Modeling for Speech Disfluencies. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, volume 1, pages 405–408, Atlanta, 1996.
11. J. Vaissière. The Use of Prosodic Parameters in Automatic Speech Recognition. In H. Niemann, M. Lang, and G. Sagerer, editors, *Recent Advances in Speech Understanding and Dialog Systems*, volume 46 of *NATO ASI Series F*, pages 71–99. Springer–Verlag, Berlin, 1988.
12. W. Wahlster. Verbmobil — Translation of Face–To–Face Dialogs. In *Proc. European Conf. on Speech Communication and Technology*, volume "Opening and Plenary Sessions", pages 29–38, Berlin, 1993.
13. C.W. Wightman, S. Shattuck-Hufnagel, M. Ostendorf, and P.J. Price. Segmental Durations in the Vicinity of Prosodic Boundaries. *Journal of the Acoustic Society of America*, 91:1707–1717, 1992.