Dovetailing of Acoustics and Prosody in Spontaneous Speech Recognition

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

University of Erlangen-Nuremberg
Chair for Pattern Recognition
Martensstr. 3, 91058 Erlangen, Germany
(buckow,batliner,huber,noeth,warnke,niemann)@informatik.uni-erlangen.de

ABSTRACT

Prosody can be applied to improve the performance of spontaneous speech translation systems like VERBMOBIL. In VERBMOBIL we previously augmented the output of a word recognizer with prosodic information. Here we present a new approach of interleaving word recognition and prosodic processing. While we still use the output of a word recognizer to determine phrase boundaries, we do not wait until the end of the utterance before we start processing. Instead we intercept chunks of word hypotheses during the forward search of the recognizer. Neural networks and language models are used to predict phrase boundaries. Those boundary hypotheses, in turn, are used by the recognizer to cut the stream of incoming speech into syntactic-prosodic phrases. Thus, incremental processing is possible. We investigate which features are suited for incremental prosodic processing and compare them w.r.t. classification performance and efficiency. We show that with a set of features that can be computed efficiently classification results are achieved which are almost as good as those with the previously used computationally more expensive features.

1. INTRODUCTION

In the context of automatic speech recognition and understanding, speaker independent large vocabulary spontaneous speech is exceptionally difficult to deal with, especially if human-to-human communication is involved. Some reasons for that are variations in the speech signal, coarticulation, accents, dialects, hesitations, corrections, filled pauses (e. g. *"em"*, *"hm"*, ...), and ungrammatical utterances. The task of analyzing an utterance gets even harder when the utterance comprises more than one phrase and can be arbitrarily long.

While several of those problems (variations of the speech signal, coarticulation, accents, dialects, filled pauses) can be compensated for by sophisticated feature extraction and acoustic modelling techniques, prosody can be very valuable in order to cope with the other spontaneous speech phenomena. A good overview about the incorporation of prosodic information into automatic speech understanding applications is given in [5].

The approach described there is not directly applicable to all situations though, as it presumes that prosodic analysis can take place after word recognition is completed. Such a sequential processing is not always possible. If a speech understandig application requires to analyze an unlimited stream of speech input, one is interested in a segmentation of that stream into chunks containing phrases, because syntactic and semantic analysis need such segments [4]. On the other hand automatic classification and perception experiments presented in [6] showed that word information is crucial for a reliable determination of phrase boundaries.

Here we have a dilemma. On the one hand we need word information for a reliable determination of phrase boundaries, on the other hand a speech recognizer needs phrase boundaries in order to produce word hypotheses for segments determined by prosodic analysis. In our research, we therefore developed a scheme that enables us to interleave acoustic-phonetic decoding and prosodic processing. This is described in section 3.

When dealing with an unlimited stream of speech input we cannot use global features, e.g. global speaking rate. Furthermore we cannot use normalized features that need global information such as the difference of local energy and global energy. Those features proved to be important for phrase boundary classification, though. In section 4 we describe different types of features that can still be computed in the case of incremental processing. Efficiency issues are addressed in this section as well. The different types of features are evaluated w.r.t. classification performance. Results are given in section 5.

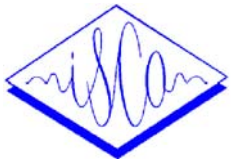
All experiments described in this paper are carried out with data collected in the domain of appointment scheduling, as this was the scenario during phase one of the German VERBMOBIL-project. Therefore, details about the data and the domain are given first in section 2.

2. THE SCENARIO

Phase one of the German VERBMOBIL project (see [2] and [8]) already had very ambitious goals: a translation system for spontaneous human-to-human business appointment scheduling dialogues was to be developed. A first prototype was presented in 1995. Now, in the second phase of the project, several conceptual changes and functional extensions have been envisaged [7]. One new requirement is incremental processing of speech input.

The goal of the research presented in this paper is to develop a scheme that makes an integration of incremental prosodic processing and acoustic-phonetic decoding possible. It is crucial not to lose classification performance and word accuracy. In order to evaluate our incremental feature extraction methods detailed in section 4, we therefore used the same data as in previous, not incremental experiments (see [5]).

[1]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.



5th International Conference on Spoken Language Processing (ICSLP 98)
Sydney, Australia
November 30 - December 4, 1998

Speech samples				
name	dialogues	turns	minutes	word tokens
BS_TEST	3	64	11	1,513
BS_TRAIN	30	797	96	13,145

Table 1: Prosodically labelled speech sample.

In this paper we concentrate on the detection of phrase boundaries that are prosodically marked. Just a small part of the speech data collected throughout the VERBMOBIL-project so far is annotated with prosodic phrase boundary labels. The amount of data used in our experiments is shown in Table 1. Every word of BS_TEST and BS_TRAIN was labelled with Bn ($n \in \{3, 2, 9, 0\}$, see Table 2).

- BS_TEST is used for the test of prosodic classifiers. The dialogues are spoken by three different male and three different female speakers at different recording sites.
- BS_TRAIN contains all turns of the VERBMOBIL speech data that were annotated with prosodic labels until spring of 1998 excluding those in BS_TEST.

Acoustic-prosodic boundary labels	
label	description
B3	prosodic clause boundary
B2	prosodic phrase boundary
B9	irregular boundary, usually hesitation lengthening
B0	every other word boundary

Table 2: Description of acoustic-prosodic boundary labels.

3. DOVETAILING OF ACOUSTICS AND PROSODY

As already mentioned in section 1, syntactic and semantic analysis need segments that correspond to phrases. Phrase boundaries, on the other hand, are most reliably determined if word information is available. A word recognizer, in turn, needs phrase boundaries in order to segment incoming speech.

A solution to this situation is a tight coupling of prosodic processing and acoustic-phonetic decoding. Most HMM based word recognizers use a multi-stage search.

1. In a forward search pass the most likely acoustic models are concatenated leading to most likely word sequences. Usually a beam search with an n-best constraint is used. Bigram language model information is often included.
2. In a subsequent search the resulting sequences of word hypotheses are then combined with language models of contexts ≥ 3. Often pruning takes place at this stage as well.

One way to combine word recognition and phrase boundary detection is to send word hypotheses during the forward search pass from the word recognizer to the module doing the prosodic processing. This module can use the word information when determining phrase boundaries. If a very likely phrase boundary is detected, the position of that boundary is sent back to the word recognizer. The word recognizer can then start to backtrace and combine the acoustic scores with higher order language models from the point of the phrase boundary back to the beginning of the segment.

This coupling is very tight since it requires a close interaction of word recognition and prosodic processing. We, therefore, refer to this coupling as dovetailing (see Figure 1). In the upper part of the figure the processing of the word recognizer is depicted, in the middle the comunication between word recognizer and the prosodic processing module, and in the lower part the prosodic processing module that determines phrase boundary positions.

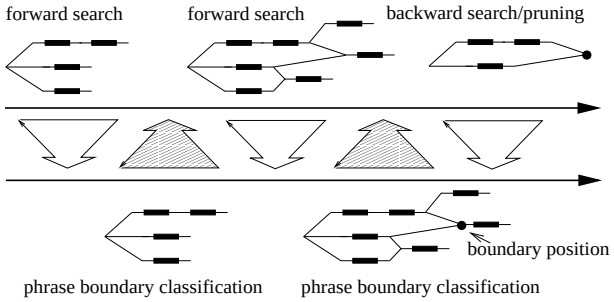


Figure 1: Dovetailing of acoutics and prosody

4. FEATURE EXTRACTION

The necessity to process data incrementally rules out the possibilty of using global features and global normalization as already pointed out in section 1. Additionally, the interception of word hypotheses from the word recognizer during the forward search as described in section 3 brings up another problem: During forward search usually up to two orders of magnitude more word hypotheses have to be dealt with because word recognizers often prune that many hypotheses when combining the language model scores with the acoustic scores.

4.1. Efficiency problems

The approach described in [5] which is also used in the prototype of the VERBMOBIL-system has to be adapted to the situation where word hypotheses are received during the forward search of a word recognizer. Two aspects of the feature extraction described there are not well suited for a high ratio of word hypotheses to words.

1. For every word hypotheses a time alignment has to be performed which is an expensive operation.
2. The vector of features used in [5] comprises several hundred components. If a word hypothesis has to be classified as to whether a phrase boundary follows or not, a context of word or syllable based intervals around the current postion is used to compute those features.

   For example, in order to describe the energy contour of the speech signal at the current position, several energy based features (e.g. the regression coefficient $energy\_Rc$, maximum $energy\_Max$ and minimum $energy\_Min$ of the energy contour) are computed over intervals based on word, syllable or syllable nucleus segments. A typical feature is, e.g., $energy\_Rc$ over the interval of the last, the current and the next syllable.

In VERBMOBIL the output of a word recognizer is structured as a graph which we refer to as word hypotheses graph (WHG). Each edge corresponds to a word hypothesis, each path through the graph corresponds to a complete hypothesis of the spoken word chain.

The efficiency problem becomes apparent when looking at figure 2. In this example, the word intervals w_{-2} and w_{-1} have to be looked at for every feature that uses a context of 2 words to the

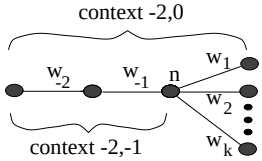


Figure 2: Piece of a WHG showing a node n with all following word hypotheses $w_i, i = 1, \ldots, k$ and a context of two words preceding those word hypotheses in the graph.

left and every word $w_i, i = 1, \ldots, k$. If the density of a WHG increases by a certain factor *c*, the context is processed on average *c*-times more often for *every* node in the WHG. In most cases the computation of a feature requires to look at every frame of the time interval that the feature is computed for. If the density of a WHG is increased by several orders of magnitude this means an enormous increase in computational effort.

A solution to the last problem is to first compute every feature for intervals of only one word and then to use those features to approximate the features with a context of more than one word. In some cases, e.g. maximum of the energy contour in an interval of several words, the exact solution can be computed. In other cases, e.g. the regression coefficient in an interval of several words, an aproximation has to be used.

4.2. Different classes of intervals

There are several ways to describe prosodic attributes at a given position, i.e. after a specific word hypothesis.

1. One can compute features using fixed size windows.
2. Word segments can be used since we get those "for free" from the word recognizer.
3. The word segments from the recognizer could be split into syllable segments. There are several ways to do that. A computationally cheap way is described in section 4.4. We refer to segments that we obtained by such a syllable boundary detection algorithm as syllable-similar intervals.
4. A forced Viterbi alignment with a speech recognizer can be used to get syllable segments. This is an expensive operation.

Depending on the type of interval some features are computable others are not. If e.g. syllable boundaries are not known, the information that a syllable carries the word accent cannot be included in a feature vector. Prosodic means that are used by humans in order to mark prosodic events like accents or phrase boundaries comprise pitch, loudness, speaking rate and pauses. The acoustic correlates of those prosodic means are $F0$, energy, and durational features. Those acoustic correlates are highly dependent on the current phone. Energy for open vowels, e.g., is much higher than for closed vowels. Thus, features that model energy variations should take phone intrinsic statistics into account.

With fixed sized windows, normalization of phone intrinsic values is not possible. If word information is used but no syllable segmentation is performed some kind of normalization can take place which is not as accurate as when a time alignment of the phoneme sequence is available (see subsection 4.3). Syllable based intervals offer the highest degree of information. This situation is summarized in table 3.

Types of features				
intervals	fixed size	word	syllable-similar	syllable
lexicon necessary	no	yes	yes	yes
syllable bound. necessary	no	no	approx.	yes
lexical feat. usable	some	some	yes	yes
norm. durational feat. usable	no	approx.	yes	yes
norm. energy feat. usable	no	approx.	yes	yes

Table 3: Different types of features: What is needed to compute them, what features can be computed.

4.3. Word intervals without time alignment

Feature extraction using word intervals has the advantage that no effort has to be made in order to determine the intervals (in contrast to syllable intervals) because those are given by the word recognizer.

For each word in the word recognizer lexicon the phoneme transcription is given. In addition to that, syllable boundaries and the default position of the word accent are known. If we had a time alignment of the phoneme sequence, we could use phone intrinsic normalisations for durational and energy-based features as described in [1] and [3].

In [1], phoneme duration is supposed to be distributed normally. If we want to avoid to perform a Viterbi alignment we can assume that the durations of the phones of a word are independent random variables. Under this assumption the duration of a word is distributed normally as well, and the mean and the variance of the word duration is simply the sum of the means of the phone durations and the sum of the variances of the phone durations respectively. The same can be done for the energy distribution of the phones of a word.

4.4. Syllable intervals without time alignment

Syllable intervals were used in [5] with great success. One reason for this might be that lexical flags (e.g. a flag if a syllable carries the word accent) can be used in this case. Such information might be helpful for some kind of implicit normalization. Furthermore, the $F0$- and energy contours can be modelled more accurately. A forced viterbi alignment in order to determine syllable boundaries is an expensive operation. Faster ways to reliably determine syllable boundaries are needed.

We trained a multi layer perceptron (MLP) on 72-dimensional feature vectors with standard backpropagation. The vectors were computed every 10msec on frames containing 256 speech samples, i.e. with a duration of 16msec. For every frame 12 mel-frequency cepstral coefficients and 12 delta coefficients for the current speech window as well as for the left and right neighbouring speech windows were concatenated. During training we used the syllable boundaries detected with Viterbi alignment as reference.

In order to determine syllable boundaries we used the positions of the maxima of the MLP output. Given a word hypothesis,

we searched for the "best" segmentation of that word into syllables. As cost function we combined scores for syllable durations (using phone duration statistics) and the output of the MLP. The segmentation yielding the lowest value of the cost function was chosen as "best" segmentation. The performance of that syllable detection algorithm developed on an independent data set is shown in Table 4.

Syllable boundary detection				
precision	recall	#correct	#deletions	#insertions
93%	93 %	7676	558	591

Table 4: Evaluation of syllable boundary detection algorithm.

5. EXPERIMENTAL RESULTS

In this research our aim is to develop a scheme that enables us to incorporate word recognition and prosodic processing if an unlimited stream of speech input has to be analyzed. An architecture which makes such an integration possible was described in section 3. In the sequential approach described in [5], the best results for phrase boundary classification have been achieved with a combination of neural networks and language models.

While the use of language models is not much affected by incremental processing, the acoustic processing, i.e. the feature extraction, has to be adapted as described in section 4. Several experiments have been conducted in order to evaluate different types of features. MLP networks have been trained on the data set BS_TRAIN and evaluated on the set BS_TEST (see section 2). The results are shown in Table 5 ($\overline{\mathcal{RR}}/\mathcal{RR}$ mean/absolute recognition rate).

phrase boundary detection experiments		
experiment	$\overline{\mathcal{RR}}$	$\mathcal{RR}$
FSW-64	76 %	76 %
WSS-274	83 %	83 %
SSS-276	87 %	88 %

Table 5: Phrase boundary detection experiments.

FSW-64 denotes a set of 64 features which describe energy and $F0$ in fixed sized windows. The window sizes correspond to the mean syllable duration and twice the mean word duration determined on a training set. WSS-274 comprises 274 energy, $F0$ and durational features computed using word sized segments with a context of +/- 2 words. No lexical flags for words (e.g. a flag if a word is a function word) were used so far. Phone intrinsic normalisation as described in section 4 has been applied. Finally, SSS-276 is the 276-dimensional feature vector as described in [3]. The computation of that vector requires global information, syllable sized segments and lexical features. Phone intrinsic normalization is performed using phone intervals as determined by a forced Viterbi alignment.

6. CONCLUSION AND FURTHER RESEARCH

We have presented a scheme that enables a combination of acoustic decoding and prosodic segmention. This scheme makes an incremental processing of unlimited speech input possible. Efficiency issues in the context of incremental processing have been addressed. We have described four types of features with different complexity, and developed feature sets for three of those. The features sets have been evaluated according to their phrase boundary classification ability. It could be shown that word interval based features that can be computed very efficiently yield almost as good classification results as the computationally much more expensive feature set described in [3].

As shown in Table 5 the set of features SSS-276 still yields slightly better classification results than WSS-274. Reasons might be the use of syllable based features and lexical flags, global normalization, or the approximation that we used in order to speed up the computation of features with contexts of +/- two words. In further experiments we will use syllable-similar intervals (see section 4) that are fast to compute and make a finer granularity of prosodic modelling possible. As more training data becomes available, we might be able to use word intrinsic instead of phone intrinsic durational and energy modelling.

Obviously, feature set WSS-274 is much faster to compute than SSS-276 since it does not need a Viterbi alignment of the phoneme sequence and features over contexts of several words are computed much more efficiently, especially if a WHG is very dense. It still has to be evaluated how big the gain really is. We have to focus here on WHGs with a high ratio of word hypotheses to spoken words.

Finally, the best suited feature set has to be included in the VERBMOBIL-system and evaluated with respect to word accuracy, phrase boundary classification performance and efficiency.

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. 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.
3. Andreas Kießling. *Extraktion und Klassifikation prosodischer Merkmale in der automatischen Sprachverarbeitung.* Berichte aus der Informatik. Shaker Verlag, Aachen, 1997.
4. 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.
5. Ralf Kompe. *Prosody in Speech Understanding Systems*. Lecture Notes for Artificial Intelligence. Springer–Verlag, Berlin, 1997.
6. V. Strom. Automatische Erkennung von Satzmodus, Akzentuierung und Phrasengrenzen in einem sprachverstehenden System. Dissertation. Rheinischen Friedrich-Wilhelms-Universität Bonn, 1998.
7. W. Wahlster. VERBMOBIL — Erkennung, Analyse, Transfer, Generierung und Synthese von Spontansprache. Verbmobil Report 198, 1997.
8. W. Wahlster, T. Bub, and A. Waibel. Verbmobil: The Combination of Deep and Shallow Processing for Spontaneous Speech Translation. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, volume 1, pages 71–74, München, 1997.