FROM THE PROCEEDINGS OF THE SPECOM CONFERENCE 2000, PAGES 63–66

DETECTION OF PROSODIC EVENTS USING ACOUSTIC-PROSODIC FEATURES AND PART-OF-SPEECH TAGS

Jan Buckow *Anton Batliner* *Richard Huber* *Heinrich Niemann* *Elmar Nöth*

Volker Warnke

University of Erlangen-Nuremberg,
Chair for Pattern Recognition (Computer Science 5),
Martensstr. 3,
D-91058 Erlangen, Germany
{buckow,batliner,warnke,noeth,huber,niemann}@informatik.uni-erlangen.de
http://www5.informatik.uni-erlangen.de

ABSTRACT

Prosody is used to improve the performance of the automatic speech translation system VERBMOBIL [8]. In our earlier work we have developed efficient and robust word-based features that describe F0, energy, speaking rate, and pauses. These features were used to classify prosodic events. We achieved the best recognition results with 95-dimensional feature vectors that describe a context of +/- 2 words [4]. In the experiments presented in this paper we additionally used Part-Of-Speech (POS) flags as features. The POS features are based on a hierarchical POS label system with up to 15 classes. The 95-dimensional acoustic-prosdic feature vectors are augmented with up to 105 POS features that describe a context of up to +/- 3 words. The new features significantly improved the recognition of phrase boundaries, phrase accents and question mood; the recognition errors could be reduced by up to 16.7%. The POS flags allow a neural network (NN) to learn a simple language model. We show that it is important to include this syntactic knowledge during the classification of the acoustic-prosodic features instead of combining it later. This implies that there is some kind of synergy: The POS information helps to correctly classify the acoustic observations. The results presented in this paper provide an effective way to improve the recognition of prosodic events with almost no computational overhead.

1. INTRODUCTION

The research presented in this paper was conducted as part of the VERBMOBIL project. The VERBMOBIL system translates spontaneous human-to-human appointment scheduling dialogues [3]. During the translation process prosodic information is used at various stages. Phrase boundaries, phrase accents, and sentence mood are used to guide syntactic parsing, disambiguate between several possible meanings [7], and improve the naturalness of the synthesis. Irregular boundary markers are used to deal with corrections [9]. Furthermore, some preliminary emotion detection is integrated in order to improve the system behavior in the case of errors [6].

In VERBMOBIL the output of a word recognizer is structured as a word hypotheses graph (WHG). Every edge represents a word hypothesis and every path through the graph a possible acoustic–phonetic interpretation of the observed utterance. The edges in the graph are marked with start and end time, thus making it possible to determine the corresponding segment of the speech signal. In order to make prosodic information available, each edge in the WHG is enriched with probabilities for prosodic events. The probabilities are determined in a classification process. For every word hypothesis, prosodic features are extracted from the speech signal and used as input to multi-layer perceptrons (MLP) for each prosodic event. The output of a MLP can be interpreted as *a-posteriori* probability [2].

In our earlier research, only acoustic information for the classification of acoustic-prosodic phenomena was used. In addition, we provided probabilties determined with statistical language models (LM), i.e. just based on the word hypotheses without considering the actual speech signal. Now, we augmented the set of features for acoustic-prosodic classification with *part-of-speech* (POS) flags for different context sizes. Our aim was to show that syntactic information helps to interpret the acoustic features and thus improves the classification of prosodic events.

Flags for word categories for each word in a context of $+/-n$ words enable a MLP to learn a category-based $(2n+1)$-gram LM. Our choice of word categories was a hierarchical POS labeling system with 2 classes at the highest level (main classes), 6 classes at an intermediate level (cover classes), and 15 classes at the lowest level (base classes). The unique but sometimes underspecified POS labels are not determined via parsing but simply by looking up the POS information in the lexicon. Thus, features based on these labels can be computed very efficiently. The POS labeling system is described in Section 2.

The 95 word-based acoustic-prosodic features that are used in addition to the POS flags have been detailed in [4]. In Section 3 we shortly introduce these features. In Section 4 we present the experiments that were performed in order to investigate the benefits of the POS flags for the classification of phrase boundaries, phrase accents, and sentence mood.

2. *PART-OF-SPEECH* LABELS

The POS of each word was annotated manually in the lexicon which contains all word forms found in the database. Of course, the POS can only be annotated unequivocally if the syntactic context is known. For the isolated word form in the lexicon, we have to find a compromise using the following strategy: If in doubt, we rely on the transliteration, for instance in the case of near–homographs where the initial letter (capital vs. small letter) can tell apart noun from adjective. We use probability in general and the probability in the VM scenario, we specify if possible (unequivocal morphology), and we underspecify if necessary, i.e., if we cannot tell apart different POS. Details can be found in

*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.

part of speech (POS)	cover class	main class
noun	NOUN	CW
proper name	NOUN	CW
auxiliary	AUX	FW
copulative verb	AUX	FW
verb (all other verbs)	VERB	CW
infinitive (or 1./3. pers. plur.)	VERB	CW
participle (pres./past, not infl.)	APN	CW
adjective, not infl., pred./adv.	APN	CW
adjective, infl. (attributive)	API	CW
article, pronoun	PAJ	FW
particle (adv., prep., conj.)	PAJ	FW
interjection	PAJ	FW
character (spelling mode)	NOUN	CW
fragment (of a noun)	NOUN	CW
fragment (¬noun)	API	CW

Table 1: parts–of–speech in the lexicon

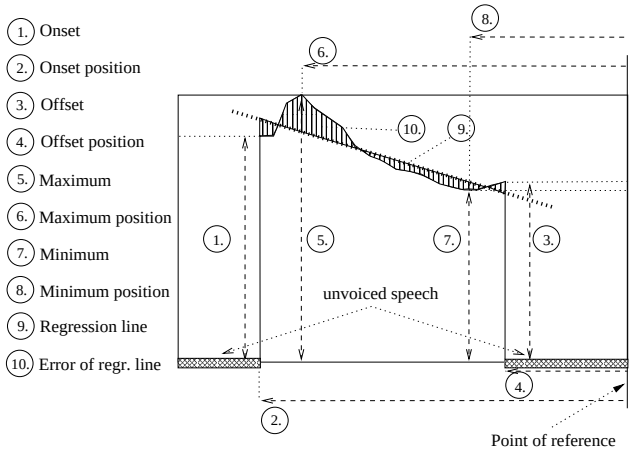

Figure 1: Example of features used to describe a pitch contour.

[1]. In Table 1, each POS label is described shortly and mapped onto its cover class and its main class (content word CW or function word FW). Such an approach to annotate POS in the lexicon yields erroneous results in some cases; we believe, however, that this does not matter very much. E.g., particles that can be either a conjunction at the beginning of an accent phrase or a local adverb somewhere in an accent phrase might be told apart most of the time because of their position in the accent phrase. This lexical POS annotation scheme is particularly useful if one has to deal not with the spoken word chain but with word hypotheses graphs where the left and right context of a word cannot be defined easily - and such a task is, after all, the 'real life' job of automatic speech processing.

3. ACOUSTIC-PROSODIC FEATURES

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 the variations of these properties and the perception of prosodic events is very complex. Thus, our approach is to find features that describe these variations as exactly but also as compactly as possible. The features are then used as basis for classification.

3.1. Feature extraction intervals

The variations of prosodic properties of the speech signals which are relevant for the detection of a prosodic event at a specific time are limited to a certain context. Within this context features which describe the prosodic properties are extracted and used for classification. Experiments have shown that a context of two words surrounding the current word are sufficient to decide if a prosodic event occured. 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.

Therefore, each component of the feature vector that we use in our classification experiments is computed over an interval that consists of at most five words (+/- two words surrounding the current word). The actual features which are included in the feature vector are determined based on phenomena which can be observed at prosodic events, e.g. *phrase final lengthening* or *resetting of the baseline* [10]. Thus, we use one feature that represents the speaking rate of the interval which consists of the current and the previous word. Another feature that we use represents the speaking rate only of the word following the current word. These two features together should, e.g., allow to detect *phrase-final lengthening*. In the experiments described in Section 4, we use a set of 95 acoustic-prosodic features.

3.2. Different kind of features

Prosodic events are perceived based on the variation of speaking-rate, loudness, pitch, and pausing. The features that we extract from the speech signal describe the acoustic correlates of these prosodic properties, i.e. the energy and fundamental frequency (F0) contours, duration and pauses.

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

Features describing contours As mentioned above, energy and F0 features are based on the short-term energy and F0 contour, respectively. Some of the features that are used to describe a pitch contour in a specific interval are shown in Figure 1. Additionally, we use the mean and the median as features (not shown in the figure).

Normalization Variations of speaking-rate or loudness have different effects on individual phonemes. Plosives are e.g. much less affected by changes in speaking-rate than vowels. This phone intrinsic variation leads to word intrinsic variations of loudness and speaking rate, i.e. some words are more effected by changes in speaking rate than others. Thus, we normalize duration and energy features to compensate for this effect.

The normalization that we use is is based on the work of Wightman [11] and shown in the Equations 1 and 2. The reasoning

that leads to the equations is as follows. First, we are interested in capturing how much a feature F (which is $duration$ or $energy$ in our case) varies compared to the "average speaker". For a training database, we compute for each speech unit w the mean $\mu_{F(w)}$ and standard deviation $\sigma_{F(w)}$ for each feature F and speech unit w. The ratio $\frac{F(w)}{\mu_{F(w)}}$ measures how much bigger or smaller the value $F(w)$ is compared to the average $\mu_{F(w)}$. The average of this ratio over an interval I is our measure $\tau_F(I)$, which is defined in Equation 1. The value $\tau_F(I)$ is used to scale the mean $\mu_{F(w)}$ and the standard deviation $\sigma_{F(w)}$ of the feature F computed for a speech unit w. The product $\tau_F(I)\mu_{F(w)}$ can be interpreted as the mean of feature F for speech unit w if uttered with $\tau_F(I)$. This is justified for phoneme duration because Wightman [11] showed that the mean and the standard deviation of the duration of phoneme classes depend linearly on the speaking rate for which $\tau_{duration}$ is an estimate.

The difference $F(w) - \tau_F(I)\mu_{F(w)}$ is negative if $F(w)$ is smaller than the scaled mean $\tau_F(I)\mu_{F(w)}$ of the speech unit w. In the case of duration, a negative difference indicates faster speech; a positive difference indicates slower speech. This value is devided by the scaled standard deviation $\tau_F(I)\sigma_F(u)$ to compensate for speech-sound dependent variations. 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").

We include $\tau_F(I)$ and $\zeta_F(J,I)$ in our feature vector for the features $energy$ and $duration$ and different intervals J. In the case of speaking rate, $\tau_F(I)$ can be interpreted as global speaking rate and $\zeta_F(J,I)$ as normalized mean duration.

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

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

4. EXPERIMENTS AND RESULTS

We performed several experiments in order to investigate if syntactic information combined with acoustic-prosodic features can improve the recognition of prosodic events.

- First, we used only 95 word-based acoustic-prosodic features in our recognition experiments to get some baseline recognition results.
- Then, we augmented the acoustic-prosodic features with 45 POS flag features, i.e. we added to the feature vector flags for each of 15 POS classes for the previous, the current, and the following word (a context of +/- 1 word).
- In order to analyze the influence of the context size, we additionally performed experiments with POS flags for +/- 2 and +/- 3 words.
- The POS labeling system that we used in the experiments is hierarchical (see Section 2). There are 15 POS classes, 6 cover classes and 2 main classes. In addition to the experiments with the 15 POS classes we also performed recognition experiments with 6 cover classes. With these experiments, we wanted to examine how fine-grained the word categories have to be in order to achieve good results.
- Finally, we combined LM classifiers (for words and POS sequences) with the MLPs trained on acoustic-prosodic features (with and without POS flags). POS flag features as well as LM classifiers contain syntactic information, but while the POS flags are added during the acoustic classification the LM information is combined later. With this experiment we examined if LM information and POS flags are redundant.

		CRR	RR
phrase accents	¬ POS	80.6%	80.9%
	+ POS	82.3%	82.6%
phrase boundaries	¬ POS	86.3%	87.8%
	+ POS	88.6%	88.6%
sentence mood	¬ POS	89.5%	90.5%
	+ POS	90.8%	90.5%

Table 2: Recognition results for the detection of phrase boundaries, phrase accents, and sentence mood with and without POS flag features for a context of +/- one word

4.1. The Baseline

As a baseline, we trained MLPs with 95 acoustic-prosodic wordbased features for phrase accents, phrase boundaries, and sentence mood. No word information was used in this experiments. The results are given in Table 2 (rows containing "¬ POS")*.

4.2. Adding POS-Features

In the experiments described here, we added to the feature vector flags for each of 15 POS classes for the current, the following, and the previous word. The new feature set with 140 elements was used to train MLPs for phrase accent, phrase boundary, and sentence mood. The results are shown in Table 2 (lines containing "+ POS"). As can be seen, the POS information improves the recognition of each of the prosodic events. The highest improvements of the *CRR* could be achieved for the recognition of phrase boundaries. This improvement corresponds to a reduction of the classification error by 16.7%.

4.3. Different Context Sizes

In the last section we described the experiments with POS features for a context of +/- 1 word. With such a context, we basically enabled the neural network to learn a simple 3-gram language model. In order to evaluate the effect of larger contexts, we trained MLPs with the 95 wordbased features and POS features for a context of +/- 2 words and +/- 3 words. The results are shown in the upper part of Table 3. Obviously, a larger context does not significantly improve the recognition results. It is not clear, however, if this is due to the limited amount of training data.

4.4. Granularity of POS Classes

Adding POS information as flags to the feature vector enlarges the feature vector significantly. This increases the number of parameters in the neural network, slows down the training, and increases the need for resources like disk space and main memory. Therefore, we investigated the effect if instead of 15 POS classes only 6 cover classes of these POS classes are used. The results are shown in the lower part of Table 3. The recognition results

*RR stands for Recognition Rate, i.e. the percentage of correctly classified patterns, *CRR* denotes the unweighted average of the class-dependent recognition rates. With this notation we follow [7].

15 POS classes				
	context size	CRR	RR	num. feat.
phrase boundaries	+/- 3 words	88.4%	88.9%	200
	+/- 2 words	**88.8%**	88.3%	170
	+/- 1 words	88.6%	88.6%	140
6 POS cover classes				
	context size	CRR	RR	num. feat.
phrase boundaries	+/- 3 words	87.9%	88.1%	137
	+/- 2 words	88.2%	**89.2%**	125
	+/- 1 words	88.3%	87.8%	113

Table 3: Recognition of phrase boundaries with POS features for different context sizes. The upper three results are produced with flags for 15 POS classes. The lower three results are produced with flags for 6 cover classes of the 15 POS classes.

are similar to those with the 15 POS classes. Thus, it is advantageous if only 6 cover classes are used.

4.5. Combining LM and MLP

In previous experiments, it could be show that a combination of language model classifiers and neural networks (trained only on acoustic–prosodic features) yielded better results for prosodic boundary classification than each of the classifiers alone[7]. This shows that syntactic and acoustic knowledge have to be combined in some way to achieve optimal results.

In the experiments described in the previous sections, we have shown that the classification performance of the baseline neural networks can be improved by POS features. Adding POS flags to the acoustic-prosodic features is, basically, another way of combining acoustic and syntactic knowledge. Thus, after those experiments, it still has to be shown that the POS information added during the acoustic-prosodic classification is not redundant to the syntactic information that can later be added by a language model.

We therefore performed several experiments with different combinations of language models and neural networks (with and without POS flag features). The combinations and classification results are shown in Table 4. In the Table, POS-LM denotes a language model trained on the sequence of POS classes instead of the spoken words. LM denotes a language model which was trained on the sequence of words; a very fine-grained category system was used in order to deal with the limited training data. NN means a neural network trained on acoustic-prosodic features alone, whereas NN-POS means a neural network trained on acoustic-prosodic features and POS flags. The knowledge sources are combined linearly. The optimal weighing factors are determined on a training database.

The best results can be achieved with a combination of POS-NN and LM. The result for POS-NN, POS-LM and LM is equally good, but the weighing factor for the POS-LM is 0. Thus, the optimal combination of these three knowledge sources excludes the POS-LM.

5. CONCLUSION

We have shown that an integration of syntactic knowledge during the acoustic classification significantly improves the classification results. Even if a language model is later added, there

Combining syntactic and acoustic knowledge					
POS-LM	LM	NN	POS-NN	CRR	RR
		√		86.3	87.8
	√			82.0	88.2
	√	√		88.6	92.7
			√	88.2	89.2
	√		√	**89.8**	**93.2**
√	√	√		88.9	93.2
√	√		√	(89.8)	(93.2)

Table 4: Recognition results for phrase boundary recognition with different combinations of acoustic and syntactic knowledge.

is still an improvement if POS flags are included in the acoustic-prosodic feature vector. This means that syntactic knowledge enables a classifier to better separate the different prosodic events in the feature space. The presented method of integrating syntactic knowledge into the acoustic-prosodic classification causes almost no computational overhead.

6. REFERENCES

1. A. Batliner, M. Nutt, V. Warnke, E. Nöth, J. Buckow, R. Huber, and H. Niemann. Automatic Annotation and Classification of Phrase Accents in Spontaneous Speech. In EUROSPEECH 99 [5], pages 519–522.
2. C.M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, NY, 1995.
3. T. Bub and J. Schwinn. Verbmobil: The Evolution of a Complex Large Speech-to-Speech Translation System. In *Proc. Int. Conf. on Spoken Language Processing*, volume 4, pages 1026–1029, Philadelphia, 1996.
4. J. Buckow, R. Huber, V. Warnke, A. Batliner, E. Noeth, and H. Niemann. Multi–lingual Prosodic Processing. In *Proc. ESCA Workshop on Dialogue and Prosody*, pages 157–162, Eindhoven, Netherlands, 1999.
5. *Proc. European Conf. on Speech Communication and Technology*, Budapest, Hungary, 1999.
6. R. Huber, E. Nöth, A. Batliner, J. Buckow, V. Warnke, and H. Niemann. You BEEP Machine — Emotion in Automatic Speech Understanding Systems. In *Proc. Workshop on TEXT, SPEECH and DIALOG (TSD'98)*, pages 223–228, Brno, 1998. Masaryk University.
7. R. Kompe. *Prosody in Speech Understanding Systems*. Lecture Notes for Artificial Intelligence. Springer–Verlag, Berlin, 1997.
8. 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.
9. J. Spilker, H. Weber, and G. Görz. Detection and Correction of Speech Repairs in Word Lattices. In EUROSPEECH 99 [5], pages 2031–2034.
10. J. Vaissière. *Language-Independent Prosodic Features*, chapter 5, pages 53–66.
11. C.W. Wightman. *Automatic Detection of Prosodic Constituents*. PhD thesis, Boston University Graduate School, 1992.