In Proc. TSD '99, Marienbad, pages 193–198

Fast and Robust Features for Prosodic Classification

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

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

Abstract. In our previous research, we have shown that prosody can be used to dramatically improve the performance of the automatic speech translation system VERBMOBIL [5, 7, 8]. In VERBMOBIL, prosodic information is made available to the different modules of the system by annotating the output of a word recognizer with prosodic markers. These markers are determined in a classification process. The computation of the prosodic features used for classification was previously based on a time alignment of the phoneme sequence of the recognized words. The phoneme segmentation was needed for the normalization of duration and energy features. This time alignment was very expensive in terms of computational effort and memory requirement. In our new approach the normalization is done on the word level with precomputed duration and energy statistics, thus the phoneme segmentation can be avoided. With the new set of prosodic features better classification results can be achieved, the features extraction can be sped up by 64%, and the memory requirements are even reduced by 92%.

1 Introduction

The aim of the VERBMOBIL project is to develop a system that translates spontaneous human-to-human speech from a source to a destination language [5]. During this translation process prosodic information is used at various stages [8]. 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

* 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.

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 each word hypothesis, prosodic features are extracted from the speech signal and used as input to a multi layer perceptron (MLP) for each prosodic event. The output of the MLP can be interpreted as *a-posteriori* probability [3].

In our previous experiments, a time alignment of the phoneme sequence of the recognized words was necessary to perform a phone intrinsic normalization of energy and duration features. A phone intrinsic normalization is important because individual phonemes are affected very differently by a change in speaking-rate or loudness [2, 6, 1]. The time alignment was by far the most expensive operation in terms of computational effort and memory requirement.

In this paper, we present a new set of prosodic features. Phone intrinsic variations are taken into account without the need to perform a time alignment of the phoneme sequence. All that is required is the duration of each word hypothesis. The phone intrinsic normalization is done on the word level with the help of precomputed duration and energy statistics. The new features are described in Section 2. We show that with the new set of features we achieve better results for all prosodic classes that are distinguished in the VERBMOBIL system. These results are detailed in Section 3.

2 Feature Extraction

Aim of the extraction of prosodic features is to compactly describe the properties of the 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, energy, pitch, and pausing. The exact interrelation of these prosodic attributes is very complex. Thus, our approach is to find features that describe the attributes as exactly but also as compactly as possible.

At each edge of the WHG, not only the current edge (i.e. the current word interval) is used for feature extraction but also intervals containing several words. These intervals from the beginning of word f to the end of word t are referred to by $I_{(f,t)}$. Intervals that we use are e.g. $I_{(-2,-1)}$ or $I_{(-1,0)}$. At the end of the word *"not"* in the utterance shown in Figure 1 the Interval $I_{(-2,-1)}$ e.g. denotes the time interval from the beginning of the word *"Of"* to the end of the word *"course"*.

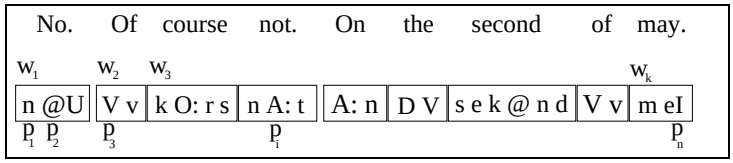

Fig. 1. Utterance *"No. Of course not. On the second of May."* with the phoneme sequence in SAMPA notation.

Each of the features that we used in our experiments (see Section 3) corresponds to an interval as described above. 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 contours. 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 and the optional use of prosodic marking into account.

2.1 Features describing contours

As mentioned above, energy and pitch features are based on the short-term energy and F0 contour, respectively. Some of the features that are used to describe a pitch contour are shown in Figure 2. Additionally, we use the mean and the median as features (not shown in the figure).

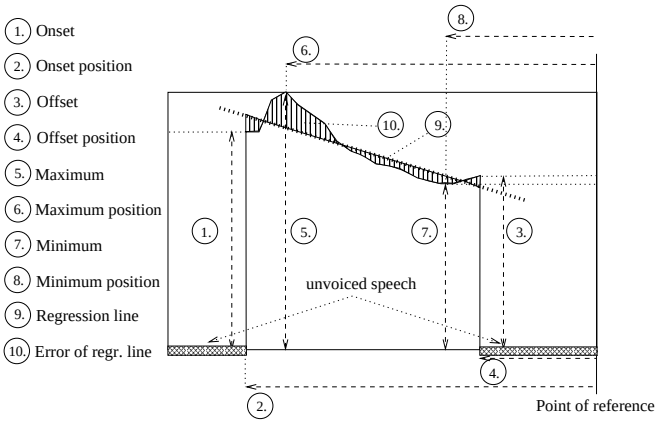

Fig. 2. Example of features used to describe a pitch contour.

2.2 Duration features on the phoneme level

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

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

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

Thus, $\tau_{duration} = \frac{1}{\#I} \sum_{u \in I} \frac{duration(u)}{\mu_{duration(u)}}$ is a measure of how much faster or slower the phonemes in the interval I were spoken compared to their mean duration. This value $\tau_{duration}(I)$ was subsequently used to compute the measure $\zeta_{duration}(J, I)$ (see Equation 2) that we included in our feature vector as normalized speaking-rate for interval J. This value $\zeta_{duration}(J, I)$ is a measure of how much faster or slower the interval J of the utterance was spoken compared to the interval I. This measure takes into account phone intrinsic dependencies as well as the optional use of prosodic marking. The standard deviation $\sigma_{duration(u)}$ and the mean of the duration $\mu_{duration(u)}$ have been computed previously on a large training database.

2.3 Duration features on the word level

A major disadvantage of the approach described in Section 2.2 is the necessity to determine the phoneme intervals. In our feature extraction module the computation of the phoneme intervals requires 92% of the total computation time and 64% of the total memory needed. Therefore, one would prefer to do a normalization on the word level. But for most words w there is not enough training data to get reliable estimates for the $\mu_{F(w)}$ and $\sigma_{F(w)}$. Equation 2 can be interpreted as a transformation of a feature with mean $\tau_F(I)\mu_{F(X)}$ and standard deviation $\tau_F(I)\sigma_{F(X)}$ to a feature with mean 0 and standard deviation 1. If we assume that the $F(u)$ are independent random variables then $\sigma^2_{F(u_1)} + \sigma^2_{F(u_2)} = \sigma^2_{F(u_1)+F(u_2)}$ (see e.g. [4]). Thus, we can compute the mean $\mu_{F(w)}$ and the standard deviation $\sigma_{F(w)}$ for a word $w = (p_1, p_2, p_3, \ldots, p_n)$ with phonemes p_i as shown in Equations 3 and 4 (if $F(w) = F(p_1) + F(p_2) + \ldots F(p_n)$).

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

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

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

2.4 Energy features

In order to describe the short-term energy contour we used only a subset of the features that are shown in Figure 2 because not all of them provide useful information (e.g. onset and offset). Furthermore, we included normalized energy in our feature vector; the same normalization as described in Section 2.3 can be applied here, i.e. $F = energy$ has to be used in Equations 1 and 2.

3 Experiments and Results

In order to evaluate our new feature set we performed several experiments.

1. On a subset of the German VERBMOBIL corpus, we compared the memory requirements and the computation time of the old and the new feature extraction methods. For this experiment, we used a set of 95 pitch, duration, pause, and energy features.
2. On the prosodically labeled German and English subsets of the VERBMOBIL corpus we performed classification experiments for all prosodic events that are used in the system, i.e. phrase boundaries, phrase accents, sentence mood, irregular boundaries, and emotion.

The results of the first experiments are shown in Table 1. As can be seen, the extraction of features could be sped up by a factor of more than 12, while at the same time the memory requirements were reduced almost by a factor of three.

Computation Time		**Memory Requirement**	
old features	**new features**	**old features**	**new features**
216 min	17 min	73 MByte	26 MByte

Table 1. Computation time and memory requirement of the old and new feature extraction methods on 112 min of speech

In Table 2 the recognition results for phrase boundary and phrase accent recognition are displayed ($\mathcal{RR}$ is the absolute, $\overline{\mathcal{RR}}$ the relative recognition rate; see Equations 5 and 6). The recognition did improve, even though the old feature set consisted this time of 276 features based on word, syllable and syllable nuclei intervals, whereas the new feature set comprised only 105 word based features.

$$\mathcal{RR} := \frac{\#\text{ correct classified patterns}}{\#\text{ all patterns}} \tag{5}$$

$$\overline{\mathcal{RR}} := \frac{1}{\#\text{ classes}} \sum_{c \in classes} \frac{\#\text{ correct classified patterns of class c}}{\#\text{ all patterns of class c}} \tag{6}$$

	English		**German**	
	old features	**new features**	**old features**	**new features**
$\overline{\mathcal{RR}}$ boundary	84.0	89.0	84.0	84.7
$\mathcal{RR}$ boundary	86.0	88.5	85.6	86.0
$\overline{\mathcal{RR}}$ accent	77.0	81.4	81.2	81.7
$\mathcal{RR}$ accent	75.0	81.0	80.9	81.0

Table 2. Recognition results for phrase boundaries and phrase accents recognition.

4 Conclusion and Further Work

In our experiments we have shown that our new word based features have at least as much discriminative power as the old features that were based on words, syllables, and syllable nuclei. With the new normalization, recognition results could be improved for all prosodic events. Furthermore, the memory requirements could be reduced by 64% and computation times even by 92%.

In the experiments described in this paper, we have always used an entire utterance for normalization, i.e. in Equation 1 interval I was always an entire utterance. This is a disadvantage if long utterances have to be dealt with. In further experiments we are going to investigate if smaller context sizes can be used.

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. M. Beckman. *Stress and Non–stress Accent.* Foris Publications, Dordrecht, 1986.
3. C.M. Bishop. *Neural Networks for Pattern Recognition.* Oxford University Press, NY, 1995.
4. I.N. Bronstein and K.A. Semendjajew. *Taschenbuch der Mathematik.* Verlag Harri Deutsch, Thun und Frankfurt/Main, 24 edition, 1989.
5. 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.
6. Andreas Kießling. *Extraktion und Klassifikation prosodischer Merkmale in der automatischen Sprachverarbeitung.* Berichte aus der Informatik. Shaker Verlag, Aachen, 1997.
7. 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.
8. Ralf Kompe. *Prosody in Speech Understanding Systems.* Lecture Notes for Artificial Intelligence. Springer–Verlag, Berlin, 1997.