Using Prosodic Features for Predicting Phrase Boundaries

Caroline Kaufhold, Elmar Nöth

Chair of Pattern Recognition, University of Erlangen-Nuremberg, Germany

noeth@informatik.uni-erlangen.de

Abstract

Spoken input of address data in modern GPS units is typically done by filling one information slot after another. To fill-in multiple slots at once, the particular slot information contained in the input utterance has to be extracted. We employ phrase boundaries to separate the speech signal into certain slots. In our evaluation, several types of input utterances differing in the number of slot information and their order are thoroughly examined. For each type, a set of twenty strong prosodic features is trained. By incorporating supporting a-priori features, an F-measure value of 93.0% is reached for a typical use case.

Index Terms: prosody, phrase boundary detection, multislot input modality

1. Introduction

A general objective of speech driven applications is improving the human-machine-interface. One major goal is a more natural input modality. Applications in the automotive environment serve as a good example. To give directions, a GPS navigation unit needs to ask for destination information, which typically consists of more than one piece of information. These pieces are organized in multiple slots. Currently the user input is done by filling one slot after another. However, rather than

Please give the city-name.
- Berlin
Please give the street-name.
- Hauptstraße
Please give the number.
- 10

a more natural way of inputting data like

Please give the address.
- Berlin, Hauptstrasse 10

would be desirable. Such an approach would allow the user to fill-in multiple slots at a time. Only one speech utterance is needed while the particular parts may be freely combined. As a first step towards this, we consider input that contains the content of two slots. Thereby *city-name* and *street-name* (which may occur in combination with a *number*) are taken as two form entries. Also, filler words, e.g. "I would like to go to . . .", are not taken into account.

GPS units in general provide only few computational resources. Working with spoken input poses an extraordinary challenge. The system has to deal with a huge vocabulary of street and city names. For this reason, state-of-the-art GPS units which allow multiple-slot filling are bound to a simplified recognition strategy. Only city-name and street-name may be combined and are recognized in a two stage process. At first, the phoneme sequence corresponding to the input signal is hypothesized. Afterwards appropriate city-names and street-names are hypothesized by simultaneously working through the phoneme-hypotheses from left and right. Resulting are two lists of city- and street-names, as found within the left beginning of the phoneme-hypotheses and the right beginning of the phoneme-hypotheses, respectively. Entries of both lists are combined in a manner such that they build possible addresses. After various rating steps the most probable combination is given. However, these lists often hold up to thousands of combinations and significant resources are allocated to the detection of the most probable one. In this paper we follow on the idea of detecting the phrase boundary (PB) between city-name and street-name in the spoken input. This can help to reduce the list of probable address combinations significantly. If the known PB of certain combinations does not match with the computed PB, these combinations are given a lower ranking. For detecting the PB, prosodic information is used. In [1] it is shown that adding prosodic information in limited recognition tasks leads to higher word recognition accuracy and error-rate reduction. To evaluate the impact of prosody for detecting the PB in our application, we compute the twenty most informative prosodic features. Furthermore, we examine the detection of PB using a combination of prosodic features and a priori knowledge.

The paper is organized as follows. In Section 2 the speech database is described. In Section 3 we explain the approach followed in this study. The Erlangen Prosody Module is discussed, which is used to compute the prosodic features. Section 4 deals with the methods used to select and evaluate particular features. In Section 5 the results are analyzed. Also, further improvements considering the inclusion of prior knowledge are presented. Conclusions and outlook are presented in Section 6.

2. Speech Database

Our database consists of the different ways of producing addresses used by German speaker in order to enter them in navigation systems. 97 speakers (48 female and 49 male) were recorded. Each produced a set of 150 different addresses of the type *city-name x street-name* and, in addition, 50 different addresses which also contained the house number (*city-name x street-name x house number*). 12 speakers (3 female and 9 male) produced the 150 input signals the other way round, *street-name x city-name*, and additionally 50 addresses containing also the house number (*street-name x house number x city-name*). Thus the following utterances occurred: *City PB Street*; *City PB Street, Number*; *Street PB City*; and *Street, Number PB City*. In addition, phoneme hypotheses as well as an estimation of the PB position for all recordings were computed (based on forced time-alignment).

3. Phrase Boundaries Detection

In our approach, the PB is detected using prosody, while working independently of already existing recognition systems and their processing. Therefore the feature vectors used in classification are built on an independent segmentation of the speech signals. The prosodic features and aforementioned segmentation are computed using the Erlangen Prosody Module[2], as described below.

3.1. Erlangen Prosody Module

The Erlangen Prosody Module was implemented at the Chair of Pattern Recognition at the University of Erlangen-Nuremberg in the course of the Verbmobil project [3][4].

At first, *basic prosodic features* – like fundamental frequency (F_0) and short time signal energy – are computed for short speech frames with a typical frame rate of 10 ms. These are tied to the voiced/unvoiced decision [5]. As prosodic phenomena last for a larger amount of syllables or words, the frame-wise computed features have to be considered in a larger context. Therefore, a segmentation is performed using the voiced/unvoiced decision. For each voiced segment, features are computed which model the suprasegmental characteristics, also based on the surrounding segments. An example is the maximum F_0 value, which can be computed on the voiced segment in question or a combination of the actual segment and neighboring segments. All features are computed within a context of at most ± 2 voiced segments. In the end, for each voiced segment a feature vector with 187 prosodic features is computed. The feature computation is described in detail in [6].

3.2. Phrase Boundary Model

In order to extract the information pieces contained in the speech signal, we need a way to associate each part of the spoken input with them. Since the particular parts of the address are listed consecutively, we only need to find the boundary separating these chunks. In general, items in spontaneously produced lists are realized such that they are perceptually salient (as described in [7]). According to this, an item-separating boundary can be detected by examination of the prosodic structure of the utterance at its position.

PBs are probable to occur at speech pauses. During the voiced/unvoiced decision they are recognized as voiceless speech parts. The prosodic changes describing the transition from one information chunk to the next can be modeled using prosodic information from the voiced speech parts that are surrounding an unvoiced period. The approach followed in this study is therefore to concatenate the prosodic feature vectors of two consecutive voiced segments, yielding a vector with 374 elements. Based on these features, unvoiced segments representing PB are classified as follows.

4. Selection and Evaluation of Features

As the concatenated feature vector consists of all the prosodic information of two succeeding voiced segments, it is rather large. High-dimensionality is not advantageous in terms of computation. Therefore, a selection step is introduced to select the most distinctive features with highest entropy. For this, we use the *Weka Framework*[8].

4.1. Feature Selection

Features are selected which improve the classification regarding overall recognition rate (RR). For the classification, a $J48$ Decision Tree is trained on the recordings of 30% of the speakers. For testing, the recordings of the remaining 70% are used. The RR is defined as

$$RR = \frac{truePositive_{PB} + trueNegative_{PB}}{2}$$

where $truePositive_{PB}$ denotes the fraction of PB that were correctly classified, and $trueNegative_{PB}$ denotes the fraction of non-PB unvoiced segments that were correctly classified.

A turn-based algorithm is employed, starting with an empty feature set. In each turn, one feature is added that maximizes the RR in combination with the existing feature selection. This is done until 20 features are selected.

4.2. Feature Evaluation

We observe a significantly higher amount of feature vectors that describe no PB compared to feature vectors describing a PB. The overall RR does not compensate for this mismatch. Hence, for evaluation of the selected features regarding recognition performance, a better measurement is needed. We employ the *F-measure*, which employs the *precision* (true positives related to false positives) and *recall* (fraction of found PB) terms:

$$F = \frac{2 \cdot (precison \cdot recall)}{(precision + recall)}$$

where $precision$ and $recall$ are computed as follows:

$$precision = \frac{TP_{PB}}{TP_{PB} + FP_{PB}}\ ,\ recall = \frac{TP_{PB}}{TP_{PB} + FN_{PB}}$$

5. Experimental Results

To gain an understanding of the performance of prosodic features to detect PB in our application, several experiments were undertaken. Input complexity plays a key role in the recognition performance. Hence, we conducted several experiments which apply different constraints to the input data.

In the following, different sequences of *city-name*, *street-name* and *house number* are examined. We present the F-measure of the selected features and the best found features are discussed.

5.1. Most General Input Modality

The most general form of input is for the user to decide which fields are filled-in and in what order. All possible combinations of *city-name*, *street-name* and a potential *house number* to form a valid German address were presented in Section 2. However, as this work is about multiple slots, the input modality of entering only one input value at a time has not been considered.

Figure 1 shows evaluation results for the twenty features which were automatically selected as described in Section 4.1. By successively enlarging the vector of prosodic features, the F-Measure converges within the twenty best features towards 83%. See Table 1 for a listing of the first four selected features as the rest of the features is almost all energy features. The most informative feature is the length of silence of the unvoiced segment itself. In terms of PB detection, this is rather expected. During the production of listings, perceived pauses after each element are commonly used to separate the elements from each

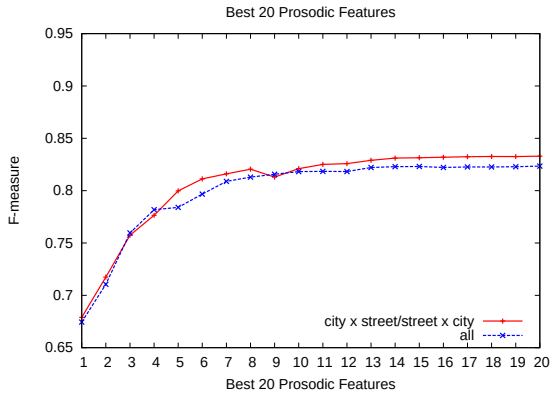

Figure 1: *Improvement of F-measure for successively selected prosodic feature for totally unknown input combinations.*

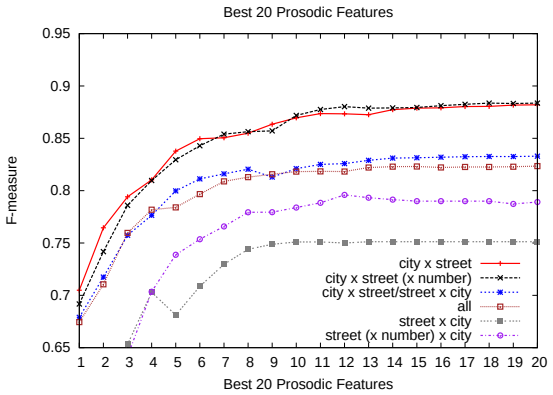

Figure 2: *Improvement of F-measure for successively selected prosodic feature for two input values of known order.*

other. The second and the third selected features are durational features. They take into account the length of the voiced segment to the left, and the combined length of third and second left voiced segment, respectively. A potential reason for this selection is the examination of final lengthening. Final lengthening is also commonly used to emphasize PBs. As a fourth feature, the minimum value of the fundamental frequency (F_0) within the third and second left voiced segments was selected.

Table 1: *The first four selected features for unknown input combinations.*

no.	prosodic feature expresses
1	Length of silence in current segment
2	Duration of left (voiced) segment
3	Duration of left and second left segment
4	Minimum F_0 within third and second left segment

5.2. Restricted Input

In order to reduce the complexity of input signals, the user may be asked to specifically enter the city-name first and then the street-name, or vice versa. Again, for both ways of producing addresses the twenty best features were automatically selected.

As can be seen in Figure 2, the recognition of PBs using prosodic features is better for input signals of type *city-name x street-name*. A possible explanation for this is the fact that the combination *street-name x city-name* is rather unusual for entering addresses in German. This may be a reason why the prosodic boundary is not that emphasized. In contrast, PBs in input signals where the city-name is followed by the street-name are modeled even better using prosodic features. In this case, an F-measure of 88.2% was achieved.

The first four selected prosodic features for the address combination *city-name x street-name* are shown in Table 2. The length of silence in the currently considered unvoiced segment yet again describes the most informative feature. As in the case of totally unknown input sequence, durational features as well as F_0 features are selected. However, the focus is not particularly on durational features. Energy features are rated much higher this time. In addition even more features considering F_0 values of both sides of the current unvoiced segment are selected. The stronger emphasis of F_0 can be an indication for the occurrence of the *continuation-rise* intonation pattern, which is generally used in enumerations or listings. The pitch rises at the end of an item if another one is following; afterwards the onset is at a comparatively lower pitch value. In contrast to the more general input case, the features selected to represent the PB in the restricted case take even shimmer into account. This may hint at the fact that speakers lower their voice at the end of a list item and therefore stronger variations in energy are more frequent. Considering these specific input cases, shimmer may be better characteristic feature for PBs than in the more general input case. The features for the address combination *street-name x city-name* consider more F_0 based features. Their poor performance, however, is in our opinion due to the absence of prosodic information because the user is not familiar with this kind of entering addresses.

Table 2: *The first four selected features of* city-name x street-name *combinations.*

no.	prosodic feature expresses
1	Length of silence in current segment
2	Mean F_0 of left segment
3	Duration of left segment
4	Absolute normalized energy of left and second left segment

In case of restricted input, the information contained in the input signal and its order are known. For further improvement, the impact of a-priori knowledge in combination with prosodic features was examined. Therefore the following additional features were included in the selection process next to the 374 prosodic features.

5.2.1. Position

As an additional feature we use the probability that a PB is observed given a particular position in the utterance. For all recordings of the same structure the position of the PB is computed relatively to the length of the utterance. These relative positions are then summed up. The resulting frequency table gives information about how often a PB occurs at a particular

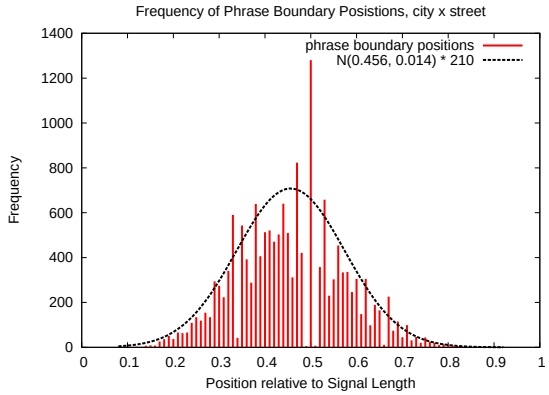

Figure 3: *Bell-shaped curve to represent the probabilities of phrase boundary positions.*

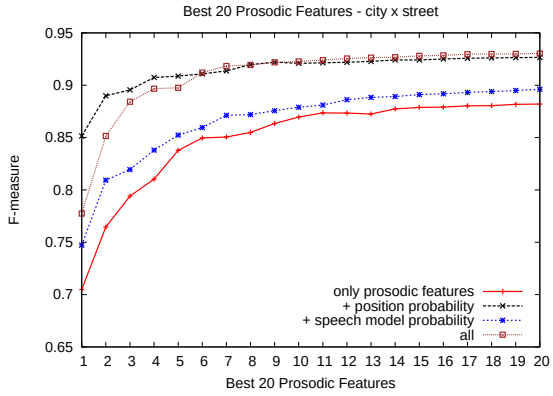

Figure 4: *Improvement of F-measure for successively selected prosodic features and certain additional a-priori knowledge of* city-name x street-name *combinations.*

position in the utterance. To use this information, a Gaussian distribution is fitted to the frequency of the particular positions. For the *city-name x street-name* recordings, it is shown in Figure 3.

5.2.2. Speech Model

In addition, a 4-gram speech model was computed based on the phoneme hypotheses. For every phoneme, the probability is given that a PB follows. For each unvoiced segment, the corresponding phoneme hypotheses are analyzed. According to their probability given by the speech model, the highest probability occurring is added to the feature vector.

5.2.3. Results

Figure 4 shows the improvement gained by taking the described additional knowledge into account. The solid curve shows the improvement per-feature when using only prosodic features for the recognition of PBs. The two curves ending above the solid curve are read in terms of taking the particular additional feature as basis and successively enlarging the feature vector by adding one prosodic feature after another in order to improve the overall recognition rate. As shown, the 4-gram probability of the speech model only slightly improves the F-measure. This is in contrast to the probability based on the position of the unvoiced segment, which increases the F-measure to 92.6%. The curve which is denoted as *all* describes the improvement gained by selecting features out of a set comprising of all prosodic features and both a-priori features. The first three selected features are then the probability given by the position, the length of the silent pause within the unvoiced segment and the probability given by the 4-gram speech model. With this setting, we achieve a remarkable F-measure of 93.0%.

6. Conclusion

In this work, we introduced the PB as a valuable cue to multiple information extraction from single speech signals. A recognition task like spoken input to a GPS unit may be solved more efficiently by using the PB to help separation of input slots. The PB is found by exploiting prosody. The approach followed on in this work considers the impact of prosody on words which were produced in chain opposite to free speech used in the Verbmobil Project. In contrast to the study of Gretter and Seppi, prosody was applied to detect PBs in German utterances which are specifically tailored for the needs of GPS units. Furthermore, the computation of PBs is done independently of the speech recognizer (except the computation of the speech model which is based on the phoneme hypotheses) and can be, therefore, applied to any existing recognition system.

It is shown that prosodic information is a sufficient cue to reliably detect PBs in the signal. This is especially the case when the order of information given is known beforehand; it may also be related to how natural this order is to the speaker. The introduction of supporting a-priori features gives further improvement to the recognition performance. Our examination furthermore revealed almost all types of prosodic information are covered by the automatic selection process.

7. References

[1] R. Gretter and D. Seppi, “Using prosodic information for disambiguation purposes”, in Proc. EUROSPEECH, pp.1821-1824, Lisboa, 2005.

[2] Kießling, A., “Extraktion und Klassifikation prosodischer Merkmale in der automatischen Sprachverarbeitung”, PhD Thesis, Chair of Pattern Recognition, University Erlangen-Nuremberg, 1996.

[3] Wahlster, W. (Ed.)., “Verbmobil: Foundations of Speech-to-Speech Translation”, Heidelberg, Germany: Springer, 2000.

[4] A. Batliner, J. Buckow, H. Niemann, E. Nöth, and V. Warnke, “The Prosody Module”, In: W. Wahlster, Ed., Verbmobil: Foundations of Speech-to-Speech Translations, pp. 106-121, Springer, New York, Berlin, 2000.

[5] A. Maier, F. Hönig, V. Zeissler, A. Batliner, E. Körner, N. Yamanaka, P. Ackermann, and E. Nöth, “A Language-Independent Feature Set for the Automatic Evaluation of Prosody”, In: Proc. Interspeech, Brighton, 2009

[6] F. Hönig, A. Batliner, K. Weilhammer, and E. Nöth, “Islands of Failure: Employing word accent information for pronunciation quality assessment of English L2 learners”, In: Proc. Interspeech, Brighton, 2009

[7] Selting, M., “Lists as embedded structures and the prosody of list construction as an interactional resource”, In: Journal of Pragmatics, 39, pp. 483- 526, 2007.

[8] http://www.cs.waikato.ac.nz/ml/weka/, last visited: Nov. 10, 2009.