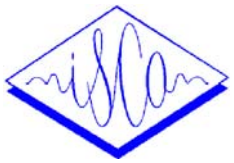
Graph Interpretation

V. Warnke, F. Gallwitz, A. Batliner, J. Buckow, R. Huber, E. Nöth, A. Höthker

Universität Erlangen–Nürnberg,
Lehrstuhl für Mustererkennung (LME),
D-91058 Erlangen, Germany
e-mail: warnke@informatik.uni–erlangen.de
http://www.mustererkennung.de

ISCA Archive
http://www.isca-speech.org/archive

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999



Abstract

We present an integrated approach for the interpretation of word hypotheses graphs (WHGs) using multiple knowledge sources. Commonly, different knowledge sources in speech understanding are applied sequentially. Typically, speech understanding systems, such as the VERBMOBIL speech-to-speech translation system, first use a word recognizer to determine word hypotheses, only based on acoustic and language model (LM) information. The resulting word sequences or WHGs are then segmented according to syntactic and/or prosodic information. Finally, these segments are interpreted by a parser or a stochastic process. Thus, it is impossible to use the knowledge of the syntactic-prosodic process, the parser or any other subsequent process to find the best word sequence. In our new approach we use acoustic, prosodic and LM information to determine the best word chain, to detect syntactic/prosodic/pragmatic phrase boundaries and to classify dialog acts in one integrated search procedure, based on a WHG or a word lattice.

1. Introduction

State–of–the–art speech understanding systems use different knowledge sources to interpret a spoken utterance. In the field of human–human or human–machine dialog processing the most important tasks are the segmentation, classification and interpretation of automatically recognized user utterances using several different knowledge sources [13, 14, 15, 3]. Commonly, these different knowledge sources are applied sequentially. For example, the VERBMOBIL speech-to-speech translation system [15] first uses a word recognizer to determine a word hypotheses graph (WHG, the output of a word recognizer with more than one alternative path) using only acoustic and language model (LM) information. These word sequences are then segmented into syntactic-prosodic phrases using prosodic and LM information. Finally, these already segmented phrases are interpreted by a parser or a stochastic process with the use of several different knowledge sources. Thus, it is impossible to incorporate the knowledge of the syntactic-prosodic process, the parser or any other later process to find the best word chain within the word recognition task.

In our new approach, we integrate multiple knowledge sources into one A^* search to find, for example, the best word chain, the best syntactic-prosodic phrase or dialog act boundaries, and the best dialog act interpretation. Our algorithm can be applied directly to the word lattice (WL) generated during the first pass of our two pass word recognizer or to a WHG. Notice that this approach allows the use of higher level knowledge in the recognition phase. In the case of our integrated word-and-boundary recognizer [4], this WL already contains phrase boundary information. In the case of a standard word recognizer, these phrase boundaries can be determined using a multilayer-perceptron (MLP) with prosodic features and/or a LM using textual information (see section 3.). During the search, the possibility of a dialog act switch is taken into account at each hypothesized phrase boundary. For example, the language model score of the optimal path for the utterance "Good morning, my name is Jones" is determined using the dialog act specific language models for GREETING and INTRODUCTION. This score is combined with the score of the dialog act transition from GREETING to INTRODUCTION, which is calculated using a dialog act sequence LM. During search, the individual cost functions are combined as a weighted sum. Thus, the search procedure implicitly determines not only the best word sequence, but also phrase boundaries and a rough semantic interpretation of the utterance, using all available knowledge sources.

For the A^* search procedure, a reliable estimate of the remaining costs to the end of the utterance is crucial if working with multiple knowledge sources, because there is an exponentially growing number of search nodes even if we use flat word graphs as input. We developed a new method of obtaining optimistic estimates of the trigram scores involved using only a bigram context and a heuristic method to ensure that the best path is determined as efficiently as possible.

The main differences between our previous approach [16] and the new approach are:

- a new A^* expansion procedure was developed
- the new procedure is suitable for direct use within a word recognizer
- the estimation of the remaining costs is improved
- DA sequences are modeled across turn boundaries
- the approach is extended by a prosodic–syntactic segmentation

Furthermore, we present our new results on a large test set, and compare segmentation and DA accuracy based on different sets of knowledge sources.

The remainder of this paper is organized as follows: First, the knowledge sources and methods used for our new approach are briefly described. Next, the A^*–procedure which allows for the integration of the different knowledge sources is presented, together with the methods to efficiently estimate the remaining costs. Finally, experimental results are reported that were achieved on VERBMOBIL spontaneous speech data.

*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 and by the DFG (German Research Foundation) under contract number 810 939-9. The responsibility for the contents lies with the authors.

	prosodic		*textual*	
words	*accent*	*bound.*	*syntactic*	*DA*
two				
o'clock				
in				AC-
the				CEPT
afternoon				
sounds				
fine	A	B	M	D
where				
would				RE-
you				QUEST
like	A			
to				
meet	A	B	M	D

Table 1: An annotated utterance from the English part of the VERBMOBIL-database with phrase accents A, prosodic B, syntactic-prosodic M, dialog act boundaries D and the DA class.

2. Data and Knowledge Sources

The VERBMOBIL-database contains spontaneous–speech dialogs of German, English, and Japanese speakers. For each utterance, a basic transliteration is given containing the spoken words, the lexically correct word form, pronunciation, and several labels for (filled) pauses and non–verbal sounds. In addition to this basic transliteration, large parts of the corpus are annotated with supplemental labels, such as prosodic (B) and syntactic-prosodic (M) phrase boundaries, dialog act boundaries (D), phrase accents (A), and dialog act classes (DA) [2, 6, 1]. For our experiments, the same 18 DA classes are used as in [6]; they are defined by their illocutionary force, such as "GREET, INIT, BYE, SUGGEST, REQUEST, ACCEPT, ...".

An example utterance annotated with A, B, M, and D label is given in Table 1. A high correlation between the different types of boundary labels can be found not only in this example, but also in the rest of the corpus (cf. [2] for a detailed analysis). In the average, one of two M boundaries is also a D boundary, and practically all D boundaries are also M boundaries. That is the main reason why we started to combine our prosodic classifier (Multi–Layer Perceptron) with a text–based classifier (Stochastic Language Model) in previous works [10, 16]. In [16], we presented our first results using an integrated approach with multiple knowledge sources. For our new approach, we use the data from the German part of the VERBMOBIL-database annotated in the manner described above. Because of different amounts of training data available for the different knowledge sources (790 turns for A and B, 12970 turns for M, 5980 turns for D) we have different training and validation sets for each classifier. However, our experimental results were achieved always on the same disjunctive test set with 1683 turns.

3. Methods

3.1. Prosodic Information

We trained multi–layer perceptrons (MLPs) to recognize the prosodic phrase or DA boundaries in a similar way as described in [8, 9]. For each word, a vector of prosodic features is computed automatically from the speech signal. This vector models prosodic properties over a context of five words, taking into account duration, pause, F0 and energy contour. The computation of the feature vector is based on a time alignment of the phoneme sequence corresponding to the spoken words. The MLP has one output node for the DA or prosodic phrase boundary (D, B) and one for the other word boundaries (¬D, ¬B).

We assume that the MLP estimates posterior probabilities. However, in order to balance for the a priori probabilities of the different classes, during training the MLP was presented with an equal number of feature vectors from each class. The best classification result so far (cf. below) was obtained with 95 prosodic features for each word and a MLP with 5/3 nodes in the first/second hidden layer.

3.2. Textual Information

For our textual–information based classifiers we use interpolated stochastic n-gram language models (LMs) [7, 12]. The LMs are used to compute the probability $P(\omega) = P(w_1 w_2 \dots w_T)$ for different kind of symbol sequences, such as normal word sequences (word recognizer, DA classifier), word and label sequences (word–and–boundary recognizer, prosodic–syntactic boundary and DA classification), and sequences of DAs (modeling of DA sequences).

Modeling word sequences

If we use the integrated classifier to find the best word chain in a WL or large WHG, we use a standard trigram LM to model the word sequences during the search procedure. Such a LM can also include phrase boundary information to find the best phrase segmentation. Furthermore, LMs that include phrase boundaries have been shown to yield better word accuracies on spontaneous speech than pure word based models [4].

Classification of boundaries

For the *segmentation* of turns into phrases, we trained LMs to estimate the probability of a boundary occurring after the current word given the neighboring words, cf. [9]. For each word boundary, symbol sequences

$$\dots w_{i-2} w_{i-1} w_i v_i w_{i+1} w_{i+2} \dots$$

are considered, where w_i denotes the i-th word in the spoken word chain and v_i is either (D, M) or (¬D, ¬M).

During the integrated A^*-search, the costs of introducing a boundary after the currently expanded node i are calculated based on the probability

$$P(w_{i-1} w_i v_i w_{i+1} w_{i+2}).$$

Classification of DAs

For classification of DAs, a LM based on a set of 884 disjunctive word categories is trained for each of the 18 DAs on the corresponding word sequences obtained from the hand–segmented and hand–labeled turns. The category system is partly based on manually assigned syntactic or semantic word classes, such as "DAYS_OF_WEEK" and "LASTNAMES".

After training a specific LM for each DA of interest, we are able to compute the probability $P_k(w_i \mid w_{i-N+1} \dots wi-1)$ during the A^*–search for each expanded node i and dialog act k.

Modeling DA Sequences

We also use a stochastic LM to compute the probability for sequences of DAs. It is trained on DA sequences $D_1 D_2 \dots D_m$, where D_i is one of the 18 DAs (e.g., "GREET, INTRODUCE, INIT, SUGGEST, ..."). These can be automatically derived from the hand–labeled training corpus.

For the classification within the A^*-search, the n predecessor DAs are computed, taking into account possible predecessor turns of the same dialog, the probability $P(D_m \mid D_{m-n+1} \dots D_{m-1})$ of the current DA is calcu-

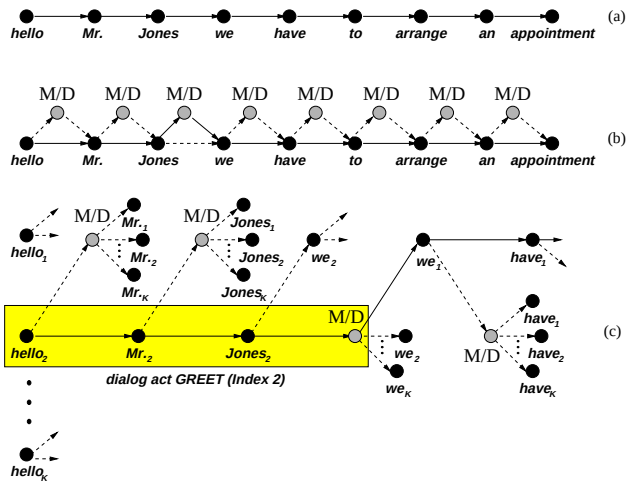


Figure 1: (a) A flat word graph with the spoken or recognized word chain. (b) The expansion procedure for integrated boundary classification. (c) The expansion procedure for integrated boundary and DA classification.

lated. This probability can be used during the search as a priori probability for the currently examined DA.

3.3. A^*–search

In the following we will introduce the search procedure in an informal manner; it is based on the A^*–algorithm [11]. The search proceeds left-to-right through a word graph. The procedure is suitable for any type of word graph, e.g. a complex graph with a high number of word hypotheses, a flat graph containing only the best recognized word chain, or a manually transliterated spoken word chain.

The Expansion Procedure

The main difficulty with integrating several knowledge sources into one A^*–search lies in the expansion procedure. In [16] we modeled the DA boundaries implicitly within the word nodes. In our new expansion procedure each phrase boundary is explicitly modeled as a node of its own. Thus, the costs for inserting boundary can be computed directly, and a boundary node is now required at the end of each DA.

An example for the new expansion procedure is given in Figure 1; the best path is indicated with the solid lines, the dashed lines indicate alternative expansion rules. Figure 1 (a) shows an example utterance produced by a word recognizer (or the manually transliterated word chain) used as input to the search procedure. In Figure 1 (b) the expansion step for the case of integrated word and boundary classification is depicted. After each word, a possible phrase boundary has to be modeled. If the boundary node has a better score than the following word node, the boundary is inserted into the graph, and the word node is expanded after the boundary node.

The complex expansion procedure for integrated boundary and DA classification is shown in Figure 1 (c). At the beginning of a turn, each DA is possible. Thus, we have to start the expansion with K alternative nodes (one for each DA). Now the costs for the different alternatives are computed, and the best scored node is expanded next. In our example, the node $hello_2$ (2 is the index for the DA GREET) achieves the best score. Because the current node is no boundary, there are only two alternatives to continue the search. Either there is a phrase boundary after *hello*, or the phrase continues with the word $Mr._2$. In this case a change to another DA is not possible, because new DAs can only be started if a boundary node is expanded. In our example, this happens at the end of the first DA (GREET) at the boundary after the word *Jones*. Now, all K alternatives for the word *we* have to be generated, and the search again continues with the best scored node. The search is stopped as soon as an explicit goal node is scored best.

Computing the Costs

The costs $score_k$ associated to a node n_k that is generated when node n_j is expanded are computed according to

$$score_k = score_j - \sum_{s=1}^{S} \lambda_s \cdot \log P_k^s,$$

where P_k^s is the probability related to the knowledge source s that is additionally introduced at node n_k (as described above). The different scores are weighted, similar to the language model weight used in speech recognition. We use an automatic procedure based on gradient descent to optimize the λ_s.

Estimating the Remaining Costs

Due to the huge search space involved, a good estimation of the remaining costs to the end of the utterance is crucial to keep the number of expanded nodes within reasonable bounds. This can easily be achieved for costs that do not depend on more than one predecessor node, such as the acoustic scores, the preestimated prosodic scores, and bigram language model scores, by scoring the word graph in backward direction based on a dynamic programming technique, c.f., for example [5]. However, determining the exact remaining costs for higher order n–gram models can only be done by coding all relevant word combinations as separate 'words', which is computationally very expensive.

Thus, bigram–based optimistic approximations of the costs for higher order language models have to be determined, that is, the estimated costs must not be higher than the real n–gram costs; this ensures that the optimal path is found [11]. These estimates are determined as follows: During the initialization of the LM, for each possible bigram the best scored higher order n–gram starting with that bigram is stored. We call the resulting model an *optimistic bigram* model; note that it does not satisfy the conditions of a probabilistic distribution. During the calculation of the remaining costs the optimistic bigram is used instead of the n–gram model.

Including this estimation of the remaining costs into the search procedure yields a drastic improvement in computation time. However, for our current experiments based on word chains the search space is comparably small, so that even a full search can be performed in real–time. We expect that further pruning techniques will have to be implemented for large word graphs.

4. Experiments and Results

All experiments were performed on the data described in Section 2. The manually transliterated word chains were used as input. The aim of the experiments was to examine if the recognition rates for boundaries and DAs can be improved by adding further knowledge sources to the classification procedure.

The measures used for evaluating dialog act classification are 'DA accuracy' (DAA) and 'DA correct' (DAC); DAA takes insertions, deletion and substitutions into account while DAC gives the relative amount of correctly classified DAs. For the boundary (M, D) classification results we give the precision (PR) and the recall rate (RE).

First we used word graphs annotated with D boundaries

λ_s		*DA class.*		D *class.*	
da	*das*	*DAA*	*DAC*	*PR*	*RE*
1.00	0.00	68.3	70.0	100	100
0.50	0.50	59.9	62.0	100	100
0.80	0.20	69.9	71.5	100	100
0.90	**0.10**	**70.8**	**72.6**	**100**	**100**
0.98	0.02	69.6	71.4	100	100

Table 2: Recognition results in % using manually annotated word graphs.

simulating 100% correct boundary classification to show how the recognition rates for DA classification improve, if only the 18 DA LMs (*da*) are used, and if the DA sequences LM (*das*) is added. The results are given in Table 2.

The first line is the baseline system using only the 18 DA LMs and manually segmented word graphs. If we give an equal weight to both classifiers the results worsen, but a weight that compensates for the different value ranges yields improved recognition rates.

Second, we wanted to determine the best D segmentation and DA classification using the MLP (*mlp*) trained on B boundaries, the LM including M boundaries (*lm*), the boundary LM for D boundaries (*bound*), the 18 DA LM (*da*) and the DA sequences LM (*das*). This is done using an automatic optimization procedure to find the best weight–configurations for the λ_s. The optimization procedure minimizes the total costs of the best path for each utterance in a cross-validation set (here, the test set). The costs are accumulated for each knowledge source during the iterations. After each iteration, they are normalized, and the new weights are used as the next configuration, starting with equal weights. We stop if the total costs do not decrease anymore. Using this procedure, we achieved our best results. These are presented in Table 3.

Iteration	*DAA*	*DAC*	*PR*	*RE*
1	45.60	52.36	92	57
5	50.92	59.88	91	60
10	52.10	62.36	89	66
15	52.53	63.59	88	68
20	52.65	64.60	88	69

Table 3: Recognition results in % using an automatic optimization procedure for the weight configurations classifying dialog acts and boundaries.

One can see, that the recognition results for DA classification improve with each iteration. For the D segmentation the recall improves considerably with only a minor loss of precision. The results for DA classification are, of course, somewhat lower than the results shown in Table 2, because those experiments were performed based on manually DA-segmented utterances.

The best result was achieved using all knowledge sources with the following weight configuration:

lm	*da*	*das*	*mlp*	*bound*
0.249	0.272	0.058	0.222	0.199

5. Conclusion and Future Work

The results show that the classification of phrase boundaries and dialog acts based on the spoken word chain and the speech signal can be improved by incorporating additional knowledge sources into an integrated search procedure. We presented an efficient search algorithm for this purpose. Furthermore, an automatic optimization procedure for determining a suitable weight configuration for combining the different knowledge sources was described.

In [4] we have already shown that integrating boundary information into a word recognizer improves the word recognition rates without any computational overhead. The next step will be to combine both approaches to directly optimize the word accuracy and the boundary and dialog act classification rates in one integrated search.

6. References

1. J. Alexandersson, B. Buschbeck-Wolf, T. Fujinami, M. Kipp, S. Koch, E. Maier, N. Reithinger, B. Schmitz, and M. Siegel. Dialogue Acts in VERBMOBIL-2 – Second Edition. Verbmobil Report 226, 1998.
2. 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(4):193–222, 1998.
3. H.U. Block. The Language Components in Verbmobil. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, volume 1, pages 79–82, München, 1997.
4. 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*, Sydney, 1998.
5. F. Gallwitz, E.G. Schukat-Talamazzini, and H. Niemann. Integrating Large Context Language Models into a Real Time Word Recognizer. In N. Pavesic and H. Niemann, editors, *3rd Slovenian–German and 2nd SDRV Workshop*, pages 105–114. Faculty of Electrical and Computer Engineering, University of Ljubljana, Ljubljana, 1996.
6. S. Jekat, A. Klein, E. Maier, I. Maleck, M. Mast, and J. Quantz. Dialogue Acts in Verbmobil. Verbmobil Report 65, 1995.
7. F. Jelinek. Self–organized Language Modeling for Speech Recognition. In A. Waibel and K.-F. Lee, editors, *Readings in Speech Recognition*, pages 450–506. Morgan Kaufmann Publishers Inc., San Mateo, California, 1990.
8. Andreas Kießling. *Extraktion und Klassifikation prosodischer Merkmale in der automatischen Sprachverarbeitung.* Berichte aus der Informatik. Shaker Verlag, Aachen, 1997.
9. Ralf Kompe. *Prosody in Speech Understanding Systems.* Lecture Notes for Artificial Intelligence. Springer–Verlag, Berlin, 1997.
10. M. Mast, R. Kompe, S. Harbeck, A. Kießling, H. Niemann, E. Nöth, and V. Warnke. Dialog Act Classification with the Help of Prosody. In *Int. Conf. on Spoken Language Processing*, volume 3, pages 1728–1731, Philadelphia, 1996.
11. N.J. Nilsson. *Principles of Artificial Intelligence.* Springer–Verlag, Berlin, 1982.
12. E. G. Schukat-Talamazzini. *Automatische Spracherkennung – Grundlagen, statistische Modelle und effiziente Algorithmen.* Vieweg, Braunschweig, 1995.
13. E. Shriberg, R. Bates, P. Taylor, A. Stolcke, D. Jurafsky, K. Ries, N. Cocarro, R. Martin, M. Meteer, and C. Van Ess-Dykema. Can Prosody Aid the Automatic Classification of Dialog Acts in Conversational Speech? *Language and Speech 41*, pages 439–487, 1998.
14. P. Taylor, S. King, S. Isard, and H. Wright. Intonation and Dialogue Context as Constraints for Speech Recognition. *Language and Speech 41*, pages 489–508, 1999.
15. 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.
16. V. Warnke, R. Kompe, H. Niemann, and E. Nöth. Integrated Dialog Act Segmentation and Classification using Prosodic Features and Language Models. In *Proc. European Conf. on Speech Communication and Technology*, volume 1, pages 207–210, 1997.