ELSEVIER

Speech Communication 36 (2002) 45–62

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

On the use of prosody in automatic dialogue understanding

E. Nöth [a,*], A. Batliner [a], V. Warnke [a], J. Haas [a], M. Boros [b], J. Buckow [a], R. Huber [a], F. Gallwitz [a], M. Nutt [b], H. Niemann [a]

[a] *University of Erlangen-Nuremberg, Chair for Pattern Recognition (Inf. 5), D-91058 Erlangen, Germany*
[b] *Bavarian Research Center for Knowledge Based Systems (FORWISS), D-91058 Erlangen, Germany*

Received 28 January 2000; received in revised form 27 September 2000; accepted 15 February 2001

Abstract

In this paper, we show how prosodic information can be used in automatic dialogue systems and give some examples of promising new approaches. Most of these examples are taken from our own work in the VERBMOBIL speech-to-speech translation system and in the EVAR train timetable dialogue system. In a 'prosodic orbit', we first present units, phenomena, annotations and statistical methods from the signal (acoustics) to the dialogue understanding phase. We show then, how prosody can be used together with other knowledge sources for the task of resegmentation if a first segmentation turns out to be wrong, and how an integrated approach leads to better results than a sequential use of the different knowledge sources; then we present a hybrid approach which is used to perform a shallow parsing and which uses prosody to guide the parsing; finally, we show how a critical system evaluation can help to improve the overall performance of automatic dialogue systems.

Zusammenfassung

In diesem Aufsatz zeigen wir, wie prosodische Information in automatischen Dialogsystemen verwendet werden kann, und stellen exemplarisch einige vielversprechende neuen Ansätze vor. Die meisten unserer Beispiele sind unseren Arbeiten in dem automatischen Übersetzungssystem VERBMOBIL und im EVAR System (automatisches Zugauskunftssystem) entnommen. Im ersten Teil geben wir einen Überblick und beschreiben Einheiten, Phänomene, Annotationen und unterschiedliche statistische Methoden, angefangen bei der Akustik bis hin zur Verstehensphase im Dialog. Im zweiten Teil gehen wir auf vier unterschiedliche Ansätze ein: Zum einen kann prosodische und andere Information zusammen genutzt werden, um falsche Segmentierungen zu resegmentieren. Zum anderen zeigen wir, dass ein integrierter Ansatz bei gleichzeitiger Nutzung unterschiedlicher Wissensquellen bessere Ergebnisse bringt als ein sequentieller. Weiter stellen wir einen hybriden Ansatz vor, der in einer flachen syntaktischen Analyse prosodische Information im Parser nutzt. Zum Schluss diskutieren wir, inwiefern eine Gesamtevaluation des Systems nicht doch andere Kriterien benutzen muss als etwa eine isolierte Evaluation einzelner Module, wie Prosodie oder Syntax.

Keywords: Dialogue; Prosody; Prosodic labelling; Automatic classification; Spontaneous speech; Large databases; Neural networks; Stochastic language models; Partial parser; A^* search

* Corresponding author.
E-mail address: noeth@informatik.uni-erlangen.de (E. Nöth).

0167-6393/02/$ - see front matter
PII: S0167-6393(01)00025-5

1. Introduction

We describe the present state of the art of using prosody in automatic dialogue systems. By this, we give a rather personal view, exemplified with our own work in the VERBMOBIL domain (Batliner et al., 2000) and in the train timetable information system EVAR (Gallwitz et al., 1998a). Older, well-known surveys on the use of prosody in automatic speech processing are (Lea, 1980; Vaissière, 1988); cf. (Niemann et al., 1998) as well. We will only deal with the *recognition* of prosodic events and the subsequent use of this information in dialogue systems; as for the *synthesis* of prosody, we refer to the paper by J. Hirschberg in this volume. Work on the use of prosody in automatic speech processing in general and in automatic dialogue understanding in particular has been, and is still quite often, 'off-line'; this means that it cannot be used directly in fully automatic systems, because manually corrected features are used, it is based on the spoken word chain, the correct segmentation is assumed, etc. On the other hand, there is an urgent need for 'real life' approaches that could be used in systems which really work and can be applied commercially. This means, in turn, that such real life approaches have to be fully automatic and have to work with word hypotheses graphs (WHG) which are the usual output of word recognition. Manual processing is only 'allowed' while testing the algorithms. In order to meet these requirements, all available knowledge should be used. In our presentation, we sketch those components that are necessary for such a use; this is done in Section 2. In Section 3, we focus on some promising trends.

2. The prosodic orbit: from signal-to-dialogue

In Table 1, we try to sketch those units, phenomena, annotations and statistical modelling methods one normally has to deal with if one tries to use prosody in automatic dialogue systems. By that, we only want to *illustrate* different and possibly alternative procedures; we do not want to present an *exhaustive* overview; of course, a different terminology could be used. Some of the descriptive terms that are used here are intuitively clear, even if a precise description is practically impossible (what is a 'word'?); some of them are rather vague and unclear (what precisely does 'focus' mean?). Still, we believe that all of these terms are well known so that the reader can follow our argumentation. Some interesting topics where prosody can provide valuable information are not mentioned in Table 1, e.g., emotional state of the speaker or speaker identification/recognition. Such topics will be relevant for automatic dialogue systems in the near future.

We do not give every suitable *level of analysis* in Table 1, only those two which are the main topics dealt with in this special issue, i.e., prosody and dialogue, and one rather complex level in between, namely syntax/semantics which is traditionally – and in fact – the mediator between these two levels. We do believe, however, that these levels represent the core of most of the work that has been done in this area. In the first column, under the heading '*prosodic properties*', we deal with the acoustic signal, its perception, and its extraction. Note that the items given under this heading are relevant across all levels of analysis and should not be attributed only to one specific level.

We usually presuppose that somehow the result of a *word recognition* is available. We can use the spoken word chain and by that assume one hundred percent correct word recognition ('cheating') if we want to concentrate on the other phenomena or if we want to determine an upper bound. For a real life task, however, we have to deal with the output of a word recognizer, i.e., with a WHG with several alternative word chains. Sometimes, the WHG does not even contain the spoken word chain. Note that for prosodic processing, a representation of the spoken words is actually not necessary: 'pure' prosody can be used to recognize accentuation or prosodic boundaries, cf. (Strom and Widera, 1996). Afterwards, however, this pure prosody approach has to be combined with word information.

Pitch, loudness, etc. are perceived prosodic properties. Actually, they are given in Table 1 only for 'completeness' because the methods used in automatic speech processing do, of course, not perceive; rather they measure the *acoustic*

Table 1
Units, phenomena, annotations and methods

Prosodic properties	*Level of analysis*		
	Prosody	*Syntax/semantics*	*Dialogue*
Acoustics:	(*Segmental*) *units*		
	Phones/phonemes		
	Syllables	Morphemes	
F0	Words	Words	
energy,	Phrases/sentences	Phrases/sentences	Phrases/sentences
duration,	Turns/utterances		Turns/utterances
.....			
Perception:	*Phenomena*		
	Boundaries/phrasing	Constituents/phrases Clauses/sentences	Dialogue act boundaries
Pitch, loudness, duration,	Accentuation	Focus	Saliency
speaking rate,	Sentence mood	Sentence mood	Dialogue acts ($\approx$ illocution)
...			
Extraction:	*Annotations* (*exemplified with our own approach*)		
	Boundaries: B3, B2, B0, B9	Synt.–pros. M labels (M3, M0) $\rightarrow$ S labels	D3, D0
Automatically extracted/	Accents: EC, PA, SA, NA	A3, A2, A0	–
manually corrected	Questions PQ	Questions SQ	Dialogue acts DA
	Statistical modelling methods		
	NN, DT, LDA, HMM, ...	LM, DT, ...	LM, DT, ...

correlates of perception, i.e., F0, energy, duration, etc. These acoustic correlates have to be computed for a certain time dimension: either a fixed one, if they are measured in fixed time windows, or a flexible one, if they are confined to certain segmental units, such as phones/phonemes, syllables, words, etc. A pure prosody approach has to work with fixed time windows or, e.g., with independently extracted syllable boundaries. The *extraction* of prosodic features in automatic systems is – no wonder – automatic. For a training sample or a test sample that is used as reference, the extraction can be manual as well, or an automatic extraction can be corrected manually afterwards. (This does not happen too often because of the effort needed.)

Note that from an application point of view (i.e. for an automatic system), *perception* units are not 'necessary': if there is a mapping from acoustics onto perception, and again, a mapping from perception onto function, then statistical modelling should be able to directly map acoustics onto function. Of course, knowledge on perception can guide feature selection and feature transformation/normalization. It is, however, our experience that very often, raw feature values rather than transformed or combined feature values should be taken if the database is sufficiently large for the training of the statistical classifier; i.e., we leave it up to the classifier to learn the most appropriate transformation, cf. (Batliner, 1989a).

The same holds for the *phonological* level: to put it bluntly, phonological systems like the well-known ToBI-approach only introduce a 'quantization error'; the whole variety of F0 levels available in acoustics is reduced to a mere binary opposition, Low versus High, and to some few additional, diacritic distinctions. In our opinion, this fact alone prevents tone levels (or any other 'prosodic phonological' concepts as, e.g., the one developed within the IPO-approach) from being a meaningful step that automatic processing should be based on; it seems better to leave it up to a largefeature vector and to statistical classifiers to find the form to the function. Actually, to our knowledge, there is no existing approach which really uses such phonological units for the

recognition of prosodic events. Of course, there are many studies that describe successful *off-line* classifications of such phonological prosodic concepts; however, this has to be told apart from the successful *integration* in an existing end-to-end-system, as we have shown within the VERBMOBIL Project, cf. (Kompe, 1997) and (Batliner et al., 2000). To prevent misunderstandings we want to stress that this caveat does not hold for *phonological knowledge*, which can be a valuable source, but only for the direct use of *phonological theoretical concepts* in automatic speech recognition.

The *segmental units* in prosody can be very short – either a time window or a phone/phoneme – or they can constitute a whole turn/utterance. Larger units are normally only used for comparison/normalization. Dialogue units are higher level units and thus usually longer than those of syntax/semantics.

The *phenomena* we want to deal with are first *phrasing*, i.e., prosodic boundaries that mirror syntactic boundaries which, in turn, mirror dialogue act (DA) boundaries. 'Mirror' means here, that a rather high, albeit not perfect correlation is assumed – otherwise, the use of prosodic information in syntax and/or dialogue would not make much sense. Second comes *accentuation* and, by that, the most important information in a unit, e.g., in a sentence (focus) or in a DA (saliency). Third, prosody can, for certain constellations, disambiguate between different *sentence moods/modalities* and, by that, different illocutionary acts/DAs. For example, prosody can be used to decide whether an elliptic sentence (free phrase) is a statement or a question (Batliner, 1991; Batliner et al., 1993).

In order to know what we are talking about, we have to have *labels* for our phenomena, and in order to know whether we are on the right track or not, we have to *annotate* corpora with these labels which we then can use as training and test data. In Table 1, we give examples of our own work within the VERBMOBIL project which started in 1994 and ended in September 2000. The VERBMOBIL database contains spontaneous speech dialogues of German, English and Japanese speakers. For each utterance, a basic transliteration is given containing the spoken words, the lexically correct word form, and several labels for (filled) pauses and non-verbal sounds. In addition to this basic transliteration, large parts of the corpus are further annotated with prosodic, syntactic and DA labels. All labels are word-based and normally introduced into the spoken word chain to the right of the word they belong to, cf. Table 3. We started with a ToBI-like annotation scheme, cf. (Reyelt and Batliner, 1994; Grice et al., 1996). Because of the caveats mentioned above, we only use the functional boundary tier comparable to the break index tier in ToBI, and the functional accent tier, comparable to the 'starred' tones in ToBI: strong boundary B3, medium boundary B2, no boundary B0, and irregular boundary B9, and phrase accent (primary accent) PA, emphatic/contrastive accent EC, secondary accent SA, and unaccentuated UA; as for details, cf. (Batliner et al., 1998; Kießling, 1997; Kompe, 1997). The boundary labels were used within the syntax modules of VERBMOBIL. Because prosodic boundaries do not always denote syntactic boundaries, we introduced another type of boundaries, the syntactic–prosodic, so-called M boundaries ('M' for language 'M'odel). A total of 25 different sub-classes was mapped onto three main classes: a main boundary class M3 (between clauses, free phrases, etc.), M0 (no boundary), and MU (ambiguous boundary). A detailed description of these M labels, including correlations with other label types and classification results, can be found in (Batliner et al., 1998). Alternatively, the M subclasses were mapped onto five syntactic 'S' boundary classes which can be described in an informal manner as follows: S0: no boundary, S1: at particles, S2: at phrases, S3: at clauses, S4: at main clauses and at free phrases. These S boundaries meet the special needs of some higher linguistic modules in the VERBMOBIL system. Based on the M boundaries and the prosodic-perceptual accent labels as a reference, we developed a rule-based system of accents with primary accent A3, secondary accent A2, and no accent A0 (Batliner et al., 1999b). In addition, syntactic questions SQ are annotated in the basic transliteration. Sentence boundaries annotated with SQ and ending in a high boundary tone H% can be labelled as prosodic questions PQ. We thus have a complete set of boundary, accent and question labels that is based

on the *prosodic form* and an analogous set of labels that is based on *syntactic structure*, i.e. on the surface, on word ordering. DA classes were annotated independently; in this paper, we use the same 18 DA classes as in (Jekat et al., 1995); they are defined by their illocutionary force, such as "GREET, INIT, BYE, SUGGEST, REQUEST, ACCEPT, ...". The criteria for the segmentation of turns into DAs are partly syntactic: for example, all 'material' that belongs to the verb frame of a finite verb belongs to the same DA. By that, we avoided to listen to the turns and could thus reduce the labelling effort. In (Carletta et al., 1997), it is reported that DA segmentation changes only slightly when the annotators can listen to the speech data, but cf. (Shriberg et al., 1998). If we know the DAs, we know their boundaries D3 as well as the complement of these boundaries, D0 (no DA boundary). Table 2 summarizes all types of annotations used in our analyses.

Of course, it is always desirable to have large-scaled annotations of exactly those units one has to deal with; this is not always realistic, however. We thus tried to aim at an *integrated* labelling approach: for example, prosodic, syntactic and DA boundaries are highly correlated with each other; exact figures can be found in (Batliner et al., 1998). If enough material is available, we can use exactly those labels that model the units we are interested in; if not, we can use highly correlated labels. Generally, we try to use *overspecified* labels that are normally not classified as such but are mapped onto some few main classes. For example, we currently do not use D3 labels for the segmentation of DA units in the 'official' VERBMOBIL system, but S4 labels, which in more than 90% correspond to D3 labels. It is, however, no problem to use D3 labels directly in a later stage, if necessary. In analogy, we do not have to annotate (prosodic) saliency in DAs at all, because we can use our prosodic and/or rule-based accent labels instead. Table 3 shows a slightly simplified example from the English VERBMOBIL database with all label types introduced above (The default classes B0, A0, etc., are not shown).

There is, of course, a wide variety of feature extraction algorithms which we do not want to deal with in this paper. Also, there is a wide variety of *statistical modelling methods* for (more or less unsupervised) clustering and subsequent classification of the phenomena. In Table 1, we only mention some of them: Neural Networks (NN) – Multi-Layer-Perceptrons (MLP), a special kind of NN, are used by us to classify prosodic labels; Decision Trees (DT), Linear Discriminant Analysis (LDA), Hidden Markov Models (HMM), and Language Models (LM). Each of these general methods has a variety of sub-methods. Normally, NNs, LDA and HMMs are used for acoustic data (Gallwitz et al., 1999), although categorical labels can be incorporated as well. LMs are used for words (unigrams) and word sequences (bi-, trigrams etc.), and DTs are used for both.

Table 2
Types of annotation and their main classes

Type	Labels for main classes
Prosodic boundaries	B3: strong, B2: medium, B0: no, B9 irregular
Synt.-pros. boundaries	M3: main, M0: no, MU: ambiguous (undefined) (25 detailed classes)
Synt. boundaries	S4: main clauses/free phrases, S3: clauses, S2: phrases, S1: particles, S0: no
Prosodic accents	PA: primary, EC: emphatic/ contrast, SA: secondary, UA: unaccentuated
Rule-based accents	A3: primary, A2: secondary, A0: no
Prosodic questions	PQ
Syntac. questions:	SQ
Dialogue acts	DA (18 classes)
Dial. act boundaries	D3 DA boundary, D0: no DA boundary

Table 3
Example turn with annotations

Turn with types of labels given in Table 1	Dialogue acts
Two o'clock in the afternoon sounds fine PA A3 B3 M3/S4 D3	ACCEPT
Where would you like SA A2 M3/S3 *to meet* PA A3 B3 M3/S4 QBT D3	REQUEST

Practically all studies on the use of prosody in speech processing in general, and in automatic dialogue understanding, in particular, use one or more of the acoustic prosodic features F0, energy, duration, etc., (top left corner of Table 1) and try to recognize the kind of labels given under the heading 'annotations' that represent those events that are given under the heading 'phenomena' in Table 1. This is the common core, everything else differs: number and manner of features extracted, units, phenomena and statistical methods. (Note that this fact makes it virtually impossible to compare classification results across studies in a strict sense!) Classification can be separated and sequential, e.g., first prosodic boundaries, then syntactic boundaries, then DA boundaries, then DAs, and independent from that, accent classification, etc. Classification can be combined and integrated, e.g., one can combine boundary and accent classification, cf. (Kießling et al., 1994), one can integrate DA boundary and DA classification, etc., cf. below and (Warnke et al., 1999), and one can even combine word recognition and boundary classification (Gallwitz et al., 1998b, 1999).

Thus, out of each column and row in Table 1, we can choose one, more or all items we want to use and/or recognize, and this can be done separately, or combined, or integrated. In (Shriberg et al., 1998, p. 446), e.g., it is reported that overall duration is the most important prosodic feature for the classification of DAs: "This is not surprising, as the task involves a seven-way classification including longer utterances (such as statements) and very brief ones (such as backchannels like "uh-huh")." In (Batliner et al., 1999a), it is reported that three durational features alone (word based, syllable based and pause duration) yield an overall recognition rate of 86% for prosodic boundaries. So we *could* use only such duration features but, of course, the more (relevant) prosodic features we use, the better is the classification (Batliner et al., 1999a). In our opinion, this result can reasonably be generalized to all other knowledge sources: the more knowledge sources we employ – and the better they are tuned to each other, the better the classification will be. This, of course, holds only if these knowledge sources are modelled adequately. This means, at least, that enough reliable training data are available, and that the statistical modelling is adequate as well.

3. Some present and future trends

Based on the prosodic orbit put forth in the previous section, we now want to describe some promising trends exemplified with our own material and work. Again, we cannot give a complete overview; for that, we refer to the other papers in this volume. We will concentrate on a *shallow* analysis; as for a *deep* (syntactic) analysis, we refer to (Kompe, 1997).

From a phylogenetic as well as from an ontogenetic point of view, a dialogue with the parents or the peer group is the earliest and most natural way of communication for human beings. If we thus compare automatic dialogue systems with other automatic speech processing applications, we can say that they are 'quite natural', i.e., rather close to the original function of natural, spontaneous language/speech, in contrast to other applications, e.g., automatic dictation systems. This means that we can find parallels between the behaviour of humans in natural dialogues and features that should be incorporated in sophisticated automatic dialogue systems. In a human–human dialogue, the speakers

1. reanalyze, if they went on the wrong track and notice that their analysis will not work;
2. integrate different knowledge sources, and do most certainly not proceed in a strictly sequential manner (Levelt, 1989);
3. pay attention to salient parts of utterances and disregard non-salient ones;
4. are content if and only if they 'get what they wanted', and tolerate non-fatal misunderstandings.

User utterances, e.g. in VERBMOBIL, can be very long. Such utterances are always processed *incrementally* by the listener; that means that the hearer forgets more or less the exact wording of a sentence rather soon, and only stores its meaning and, if necessary, its illocution (Hörmann, 1978, p. 460ff). In analogy, automatic systems should be able to process longer utterances in an incremental

way as well. Otherwise, system responses would be delayed unduly and, by that, user acceptance would be rather low.

We will address all four topics and concentrate on the second and the third topic, where we present methodologies and experimental results for two domains. More details can be found in (Warnke et al., 1999; Nöth et al., 1999). The first and the last topic will be dealt with rather sketchy, since we cannot yet provide substantial results.

3.1. Reanalysis within a sequential approach: the VERBMOBIL system

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 dialogue processing, the most important tasks are the segmentation, classification and interpretation of automatically recognized user utterances using several different knowledge sources (Shriberg et al., 1998; Taylor et al., 1999; Wahlster et al., 1997; Block, 1997; Wahlster, 2000). Commonly, these different knowledge sources are applied sequentially. For example, in the VERBMOBIL speech-to-speech translation system (Wahlster et al., 1997; Block, 1997; Wahlster, 2000), first a word recognizer generates a WHG using only acoustic and LM information. The 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 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 a system like VERBMOBIL, which proceeds in a sequential manner with no back-tracking mechanism, the higher linguistic modules are thus sometimes faced with wrong segmentation. Here, we want to give examples for two different factors that can be responsible for a wrong segmentation.

Consider the following turn containing a repair; note that repairs are most of the time not marked by an edit term in the VERBMOBIL database (Batliner et al., 1995).

Treffen wir uns am Montag B3/S4 *– am Dienstag*
Let's meet on Monday B3/S4 *– on Tuesday*

Let us assume that both NN and LM classify this boundary as B3 and S4. If the syntax module accepts this analysis, the phrase *on Tuesday* has to be interpreted as a free phrase or as a right dislocation (Batliner et al., 1998), and by that, as a sort of – contradictory – addendum or specification of *on Monday*. Cases like these should of course be treated differently from cases like the following where there is no contradiction but a specification:

Treffen wir uns am Montag B3/S4 *– um acht*
Let's meet on Monday B3/S4 *– at eight*

In VERBMOBIL, a repair module is located between prosody and syntax (Spilker et al., 1999, 2000). It uses prosodic information, i.e. looks at boundary locations, to see if a repair occurs in their surrounding. If the repair module can mark the word boundary in question as a possible interruption point, it compares the part-of-speech labels of the constituent to the left and to the right of the end of the reparandum. If it can find the reparandum, these words can be cut out and the correct word chain intended by the speaker can be generated and parsed:

Let's meet on Tuesday

As a second example, consider the following turn with a word-by-word translation into English:

Ja, ich habe am Montag B3/S4 *oder am Donnerstag Zeit*
Well, I have on Monday B3/S4 *or on Thursday time*

Prosody alone might not help because there is a pronounced pause after *Montag*. Here, the analysis window of the LM can be too small, and thus, a wrong segmentation within the verbal bracket can be generated; note that the verbal bracket ('Verbklammer', i.e., a special bracketing for linguistic groupings) is a syntactic phenomenon that does not exist in English. In such cases, the syntax

module will not simply rely on the output of the NN/LM but detect, that the right end of the verbal bracket has not been reached yet, and that a correct analysis can only be generated if this wrong segmentation is discarded (Kasper et al., 1999).

3.2. An integrated approach: the A^* search

The A^* algorithm is an efficient graph search procedure which provides the optimal path through a graph – in our case, the WHG. The procedure is suitable for any type of WHG, 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. During the search, only the currently best path is expanded and used for further processing. After expansion, all candidates are scored and put onto the agenda ordered by quality. The best-scored candidate on the top of the agenda represents now the currently best path and is taken for further processing. The algorithm stops if the last candidate (i.e., the last node of the WHG) is at the top of the agenda. The A^* algorithm is described in more detail in (Nilsson, 1982; Kompe, 1997).

We have seen that in a sequential approach, we sometimes have to repair wrong analyses, e.g., a wrong segmentation, in a subsequent pass within the higher linguistic modules. Another way of combining higher with lower linguistic knowledge is an integrated approach. In such an approach, we integrate multiple knowledge sources into one A^* search to find, for example, the best word chain, the best syntactic–prosodic phrase or DA boundaries, and the best DA interpretation. In VERBMOBIL, DAs are used for a robust template-based translation if the deep syntactic module does not produce any reasonable analysis. Furthermore, knowledge of the correct DA can help choosing the right candidate within the WHG. The phrase boundaries can be determined using an MLP with prosodic features and/or an LM using textual information. During the search, the possibility of a DA switch is taken into account at each hypothesized phrase boundary. For example, the LM score of the optimal path for the utterance "Good morning, my name is Jones" is determined using the DA specific LMs for GREETING and INTRODUCTION. This score is combined with the score of the DA transition from GREETING to INTRODUCTION, which is calculated using a DA 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.

A high correlation between different types of boundary labels can be found not only in the example given in Table 3, but also in the rest of the corpus (cf. (Batliner et al., 1998) for a detailed analysis). On average, one of two M3 boundaries is also a D3 boundary, and practically all D3 boundaries are also M3 boundaries. This is the main reason why we started to combine the MLP of our prosodic classifier with a text-based LM classifier in previous work (Mast et al., 1996; Warnke et al., 1997). For our experiments, 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 prosodic accents and boundaries, 12,970 turns for M3, 5980 turns for D3) we have different training and validation sets for each classifier. Our experimental results, however, were always achieved on the same disjunctive test set with 1683 turns. In (Warnke et al., 1999), the modelling of word and DA sequences and their boundaries is described in more detail. Here, we will concentrate on the A^* search procedure and introduce it in an informal manner. The search proceeds left-to-right through a word graph.

3.2.1. The expansion procedure

The main difficulty with integrating several knowledge sources into one A^* search lies in the expansion procedure. In (Warnke et al., 1997), the DA boundaries were modelled implicitly within the word nodes. In our new expansion procedure, each phrase boundary is explicitly modelled as a node of its own. Thus, the costs of inserting a boundary can be computed directly, and a boundary node is now required at the end of each DA.

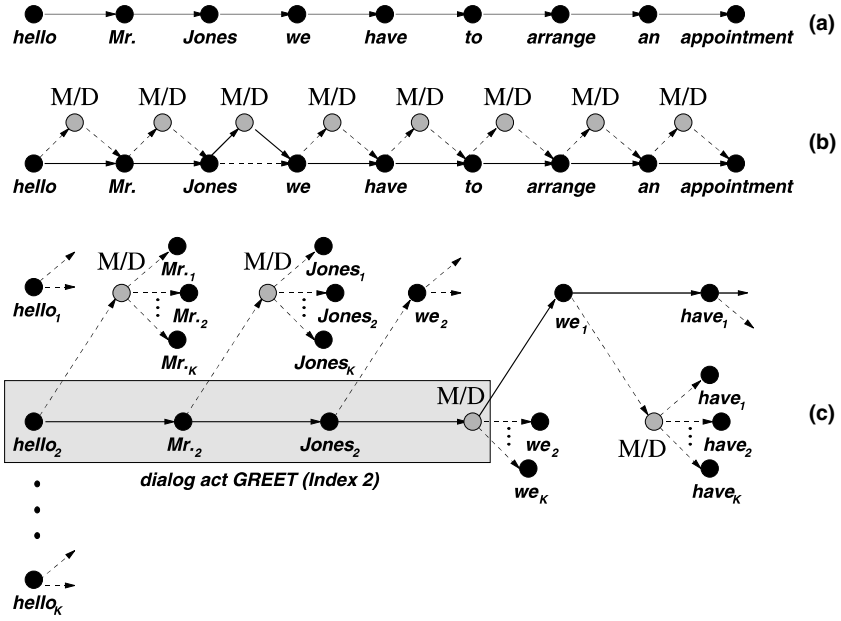

Fig. 1. (a) A flat word graph with the spoken or recognized word chain. (b) The expansion procedure for integrated boundary classification. (c) The expansion procedure for integrated boundary and DA classification.

An example for the new expansion procedure is given in Fig. 1. The best path is indicated with solid lines, dashed lines indicate alternative expansion rules. Fig. 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 Fig. 1(b), the expansion step for the case of integrated word and boundary classification is given. After each word, a possible phrase boundary has to be modelled. 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 Fig. 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. As for the computation of the costs and the estimation of the remaining costs, we refer to (Warnke et al., 1999).

3.2.2. Experiments and results

All experiments were performed using the manually transliterated word chains 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. Analogously to 'word accuracy' and 'word correct', we evaluate the DA classification with '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 (M3, D3) classification results, we give precision (PR) and recall rate (RE).

First, we used word graphs annotated with D3 boundaries simulating 100% correct boundary classification to show how the recognition rates for DA classification improve, if only the LMs for the 18 DA (LM_{DA}) are used, and if the LM for the DA sequence (LM_{DAS}) is added. The results are given in Table 4. 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 D3 segmentation and DA classification. For that, we use all available knowledge sources, i.e. prosodic knowledge encoded in the MLP (*MLP*) trained on the prosodic B3 boundaries, shallow syntactic knowledge encoded in the word sequences used by the LM including the syntactic–prosodic M3 boundaries (LM_{M3}), and knowledge of the DAs and their sequence encoded in the word sequences used by the boundary LM for D3 boundaries (LM_{D3}), the LM for the 18 different DAs (LM_{DA}), and the LM for the DA sequences (LM_{DAS}). This is done using an automatic optimization procedure to find the best weight configuration for 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). Using the automatic optimization procedure, we achieved the results presented in Table 5.

One can see, that the recognition results for DA classification improve with each iteration. For the D3 segmentation the recall improves considerably with only a minor loss of precision. The results for the DA classification are, of course, somewhat lower than the results shown in Table 4, because those experiments were performed based on utterances that were manually segmented into DAs.

The best result was achieved using all knowledge sources with the weight configuration given in Table 6.

It can be seen that all classifiers except the LM_{DAS} contribute approximately to the same extent. This is because they all have the same level of complexity and are based on single words, whereas the LM_{DAS} is based on word *sequences* that constitute always one of only 18 *DAs*. The influence of the LM_{DAS} has thus to be reduced drastically in the optimization phase. It can be seen in Table 4 as well that the weight for the LM_{DAS} has to be reduced for the best DA and DAC classification.

Table 4
Recognition results in % using manually annotated word graphs[a]

λ_s		DA class		D3 class	
LM_{DA}	LM_{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

[a] λ_s is the weight for the two LMs LM_{DA} and LM_{DAS} always summing up to 1.

Table 5
Recognition results in % using an automatic optimization procedure for the weight configurations classifying DAs and boundaries

Iteration	DAA	DAC	PR	RE
1	45.6	52.4	92	57
5	50.9	59.9	91	60
10	52.1	62.4	89	66
15	52.5	63.6	88	68
20	52.6	64.6	88	69

Table 6
Weight configuration

LM_{M3}	LM_{DA}	LM_{DAS}	MLP	LM_{D3}
0.25	0.27	0.06	0.22	0.20

Table 7
Recognition results in % achieved by performing segmentation and classification of DAs sequentially

DAA	DAC	PR	RE
47.3	62.0	71	73

In (Mast et al., 1996), we presented a sequential approach where a turn was first segmented and then the resulting segments were classified into DAs. If we proceed the same way on our new test set and use the same classifiers as for the integrated approach we achieve the results presented in Table 7.

One can see that the integrated approach improves the DAA by over 5% and the DAC by over 2%. Even the segmentation accuracy improves a lot when both tasks are performed in an integrated procedure. These results show that the classification of boundaries and DAs based on the spoken word chain and the speech signal can be improved significantly by an integrated search procedure incorporating a number of knowledge sources.

The most important advantage of the A^* search is, however, not that it yields better results than a sequential approach but the possibility to work directly with the WHG. In a sequential approach, one has to work with the best word chain(s). This might do for basic research but it is, in the long run, not a feasible strategy for 'real life' systems.

3.3. A hybrid approach: prosody, statistics, and partial parsing

In this section, we want to focus on accentuation. Linguistic analysis in spoken dialogue systems has to cope with two main problems. First, spontaneous speech very often is fragmented, ungrammatical or exceeds the system's capacities (e.g. out-of-vocabulary words). Second, word recognition in spoken dialogue systems produces errors, thus rendering utterances ungrammatical on the syntactic as well as the semantic level. In order to cope with these problems, methods of robust parsing have been established. For example, partial parsing methods restrict syntactic analysis to sub-units of utterances only, therefore reducing the above-mentioned problems to these sub-units. Different methods of partial parsing have been successfully employed in spoken dialogue systems, such as the systems described in (Albesano et al., 1997; Aust et al., 1995).

Partial parsing in dialogue systems becomes even more efficient if more sophisticated sources of information, beyond acoustically scored WHGs and DA predictions, can be used to guide the linguistic processor. We concentrated on the integration of prosodic information, extracted from the speech signal, and detected semantic concepts in utterances as additional support for the parser, thus resulting in a hybrid approach to language understanding. The units to be analysed correspond to semantic concepts, e.g., time, date, source or target location for train timetable inquiries, or to DA classes, as, e.g., SUGGEST or ACCEPT, in the VERBMOBIL task. Such units are vital for the correct interpretation of the utterance in the application domain. The parser will identify and analyze these concepts, assigning a semantic representation to each.

For each concept and its possible surface realizations, grammar fragments are defined that may be used by the parser upon request. The parser is guided by prosodic information on phrase boundaries and phrase accents, telling it where to start the partial analysis. Statistical concept detection provides information on which semantic concepts are included by the current utterance, thus helping the parser to choose the appropriate grammar fragments. The use of grammar fragments has the following advantages: the danger of false alarms in parsing is drastically reduced, as well as the time consumed by the parser and the efforts for grammar development.

3.3.1. Accent information in word hypotheses graphs

Here, we want to use prosodic information to determine the salient regions in a phrase. These regions are those parts of a sentence which hold the most important content words, e.g., time

expressions and locations and which most of the time are 'in focus', i.e., are the carrier of the focal accent. To get information for these regions, we use an MLP trained on a part of the VERBMOBIL database.

Using Score(A3|w) and Score($\neg$A3|w) from the output nodes of the MLP for each word w we can estimate the probability P(A3|w) by using the following formula:

$$P(\mathrm{A3}|w) = \frac{\mathrm{Score}(\mathrm{A3}|w)}{\mathrm{Score}(\mathrm{A3}|w) + \mathrm{Score}(\neg\mathrm{A3}|w)}.$$

Note that MLPs do not compute probabilities but only estimates of probabilities. Therefore, the sum of the scores of the output nodes does not sum up to 1.0. In order to be able to use probabilities, the scores are thus normalized in such a way that they sum up to 1.0. Now, we are able to estimate the probability P(A3|w) for each word of an utterance. We decide for a focused region by using a threshold. In Fig. 2, an example is given for a German utterance.

The estimation of accentuated regions in a given utterance offers two possible methods of using this knowledge in combination with the parser:

1. The regions are ranked by their prosodic scores and the ranking list is given to the parser, which has to find the best expression for the given context.
2. A list of possible expressions from the parser is disambiguated using the prosodic score from the NN.

Both methods can efficiently be employed to find the best expression the parser is searching for in the context the concept predictor has estimated. The first way seems to be the better one if working on WHGs, because the parser only has to search in the best-scored paths and thus, search effort is smaller.

For the 18 DA classes sketched above, we estimated the most frequent accentuated words of a subset of the VERBMOBIL database using the method described above. Only those words are considered, that exceed a threshold of 0.8 for the automatically calculated accent probability in more than 80% of their occurrences. In Table 8 the 10 most often observed words are shown, which fulfill this criterion when looking at all DAs. Table 9 shows the five most often observed accentuated words for the most frequent DA classes SUGGEST and ACCEPT. In both tables the words are ranked by their frequency of occurrence in the observed data set.

The results from Tables 8 and 9 show that a successful classification of content words in an utterance is possible through determining the accentuated words. Semantically important information can thus be obtained via the detection of the focused regions. This can be done by only using prosodic-acoustic features. Note, however, that for the classification of DAs, function words that normally are not accentuated are important as well, cf. (Nutt et al., 1999).

3.3.2. Statistical concept detection

As a second additional source of information for the hybrid partial parsing, we apply a statistical approach which uses n-gram LMs as semantic concept predictors. The model has to decide on the occurrence of special semantic concepts in word chains. We show its usability on a corpus collected with the above-mentioned information retrieval

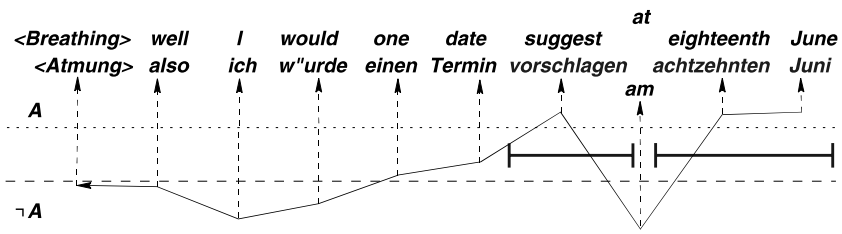

Fig. 2. A German sentence from VERBMOBIL with probability P(A3|w) for each word w and the two focused regions hypothesized (with word-to-word translation).

Table 8
Automatically determined accentuated words for all DAs

$P(A3\|w) > 0.8$		
Rank	% Accentuated	Word (translation)
1	88.57	Freitag (Friday)
2	82.69	Wiederhören (bye)
3	84.31	Donnerstag (Thursday)
4	90.91	Samstag (Saturday)
5	95.35	neunzehnten (19th)
6	81.82	August (August)
7	96.15	vierundzwanzig. (24th)
8	87.50	achten (8th)
9	86.96	wunderbar (marvellous)
10	100.00	sechsundzwanzig. (26th)

Table 9
Automatically determined accentuated words for dialogue acts ACCEPT and SUGGEST

$P(A3\|w) > 0.8$		
Rank	% Accentuated	Word (translation)
ACCEPT		
1	100.00	einverstanden (ok)
2	100.00	Ordnung (all right)
3	100.00	wunderbar (marvellous)
4	85.71	Freitag (Friday)
5	85.71	frei (free)
SUGGEST		
1	82.22	Montag (Monday)
2	87.80	Freitag (Friday)
3	83.33	Donnerstag (Thursday)
4	82.76	Mittwoch (Wednesday)
5	93.10	Samstag (Saturday)

system containing the utterances used for the grammar development. In the following, we present two predictors, one for time expressions and one for date expressions. The predictor should be able to decide whether there appears such a time/date expression in an utterance or not. (Note that predictors for other concepts occurring in our data as, e.g., point of destination/departure, would be rather trivial because they almost always follow certain prepositions, *nach* (*to*) for destination, and *von* (*from*) for departure.)

If we use LMs as semantic concept predictors we have to claim for a word chain $\boldsymbol{w}$ whether the concept we are looking for is expressed in $\boldsymbol{w}$ or not. For this purpose we build two different LMs. The first one is trained with word chains expressing the semantic concept, and the second one with the utterances not expressing it. During analysis the two scores for the incoming word chain are computed – for WHGs, the best word chain in the graph is used – and the predictor with the higher probability is chosen. We apply category based LMs, rational interpolation for the LMs, and a context of three words. The 'Semantic Concept Predictor' results are shown in Table 10 as confusion matrices. One can see that our LM approach to the prediction task performs well enough and can therefore be used as a predictor for the semantic concept analysis, i.e. to select the correct grammar fragment. For the problem of detecting time expressions we obtain a recognition rate of 95.6%; if we measure the performance of the models as being a time expression spotter, we get a recall of 98.4% and a precision of 80.0%. For date expressions we have a recognition rate of 95.7%, a recall of 96.7%, and a precision of 82.4%.

Table 10
Recognition results for time and date expressions in percent

	#	TIME	NOTIME
TIME	1145	98.4	1.6
NOTIME	5657	5.0	95.0
		DATE	NODATE
DATE	1232	96.7	3.3
NODATE	5570	4.6	95.4

3.3.3. The partial parsing algorithm

The partial parser described here is an agenda-driven chart parser, operating as an island parser cf. (Mecklenburg et al., 1995). Our approach restricts the linguistic analysis to the analysis of semantic concepts. Lexicon and grammar of the parser only need to cover the relevant syntactic realizations for each concept, thus resulting in several grammar fragments, and not in one full grammar. Island parsing on the basis of these grammar fragments means that each of the maximal islands, found by the parser, corresponds to one relevant part of an utterance. We coded a grammar fragment for each of the semantic concepts in terms of a context-free phrase structure grammar. Thus, the predictions on the occurrence

of concepts in user utterances can be used to guide the parsing process. This is done by using only those grammar fragments for parsing that correspond to semantic concepts predicted by the concept detection module. In order to further improve efficiency of the parsing process, prosodic information is included into the parsing process. Each word hypothesis contains a prosodic accent score, in addition to the usual acoustic score. This information is used for choosing the initial islands: only those hypotheses which are marked as accentuated are chosen as initial islands.

The chart is initialized with the lexical entries for the hypotheses in the WHG. As not every grammar fragment is used for each parse, many hypotheses are unknown, thus leaving gaps in the chart. In parallel to the chart, two agendas are initialized that guide the flow of the analysis. The first agenda (*seed agenda*) contains all hypotheses that serve as initial islands. The second agenda (*non-seed agenda*) contains the remaining hypotheses. Each hypothesis, whose accent score exceeds a given threshold, is inserted into the seed agenda, the remaining ones into the non-seed agenda. Within both agendas, entries are sorted according to their acoustic score. Agenda entries may not only be used as initial lexical entries (*seed entries*), but also as pairs of chart edges (*non-seed entries*) that comprise pointers to two adjacent chart edges and a list of grammar rules that might combine these two edges to a new one. The following steps are performed until no entries are left in the seed agenda.

1. Take best-scored agenda entry E from seed agenda.
2. If E is a seed entry, then go to 3, else go to 4.
3. For each adjacent chart edge to E, look for rules that can be applied to both, generate an agenda pair for both, and sort it into seed agenda; go to 1.
4. For each grammar rule in E: apply this rule to both edges, insert new edge (if rule can be applied) into chart, generate new agenda pairs for this new edge and insert them into seed agenda; go to 1.

This is done for each of the predicted semantic concepts using the respective grammar fragments. Only if no valid semantic representation for a concept can be found in the chart after parsing, the process is restarted with the non-seed agenda.

First experiments were done for the semantic concepts `time` and `date`. The results are given in Table 11; as for details, cf. (Nöth et al., 1999). The test corpus contained 871 utterances comprising 2761 words in total. We counted the number of parses and the number of island seeds for each concept and evaluated the parsing results by counting the number of deleted (D) or inserted (I) semantic concepts. The first column in Table 11 (NIL) represents the situation where no information is used by the parser (even no lexical information), i.e. every sentence has to be analysed and each word is chosen as an island seed. Taking the lexical information into account (LEX), the number of parses and initial seeds decreases drastically as only words contained in either the lexicon for time expressions or the lexicon for date expressions are considered. The third column reflects the usage of concept predictions delivered by the LM classifiers (PRED). Here the grammar fragments for

Table 11
Number of necessary parses and possible island seeds with different levels of information sources and the number of deletions (D) and insertions (I) for *date* and *time*

	NIL	LEX	PRED	PROS 0.5	ALL
Parses					
Time	871	346	136	219	124
Date	871	292	169	244	133
Seeds					
Time	2761	836	431	439	285
Date	2761	605	416	447	308
D (time/date)	1/0	1/0	1/0	1/0	1/0
I (time/date)	2/2	2/2	3/1	2/2	3/1

each concept are applied only if the respective concept was predicted. Compared with only lexical knowledge being used, the number of parses and island seeds again decreases by almost 50%. When using prosodic information (PROS) instead of concept information, similar results are observed. We used a threshold of 0.5 to decide upon the accentuation of a word and its selection to the seed agenda. The prosodic features used here are our 'usual' prosodic word-based features representing pitch as well as (normalized) energy and duration; for a more detailed account, cf. (Batliner et al., 2000). The last column (ALL) shows results obtained when combining all three sources of information. Again, the number of parses and island seeds can be reduced, without any significant loss in accuracy. This leads to the conclusion, that the more knowledge is used for parsing the less parses and island seeds are needed to obtain the same good results while speeding up the parser. This can be seen in the last two lines of Table 11: the number of deletions and insertions does not increase from column NIL to the other columns.

3.4. Looking back from the end: adequate evaluation and adequate design

In the VERBMOBIL system, each module evaluates and optimizes its analyses and classification results independently from the other modules, and there is an end-to-end-evaluation with the criterion: 'Is the translation approximately correct?' In addition, there is a more or less informal feedback from the higher linguistic modules to the lower ones. If it is not (yet) possible to 'formalize' such a feedback, it should at least be intensified. The criterion should not simply be the *correctness* of the translation, but the *success* of the communication. Sometimes, underspecification will do; this will be shown with the following example.

Let us assume the following utterance with correct word recognition, parse, and subsequent translation:

Um zwei Uhr nachmittags. Wollen wir uns am Berliner Hauptbahnhof treffen?
At two p.m. Should we meet at Berlin main station?

If, however, the boundary between *nachmittags* and *wollen* is not recognized, and if there is no prosodic question PQ at the end of the turn, the parse and the translation would result in:

Um zwei Uhr nachmittags wollen wir uns am Berliner Hauptbahnhof treffen.
At two p.m., we want to meet at Berlin main station.

Here, prosody and especially intonation are irrelevant for the classification of sentence mood because the speaker can produce a final rise or a final fall (Batliner, 1991; Batliner et al., 1993): the prosodic distinction between question and non-question is neutralized. If the turn is translated as a statement, then the *segmentation* is wrong, the *proposition* ($\approx$ the salient words) is right; *illocution* and *translation* are wrong because the sentence mood is not reproduced correctly. The *perlocution*, however, is successful: no matter whether the correct or the wrong translation is generated, if the dialogue partner accepts, e.g., with an *ok*, and if the dialogue partners meet at the given time and place, the communication is felicitous – even if the translation is wrong. This means that there are fatal and harmless errors which should be treated differently in the evaluation. In the design of the system, it might be better to leave such alternatives underspecified. We thus believe that a local optimization – e.g., recognition rates for DA classification – can only be an intermediate step towards the 'ultimate' evaluation within an existing dialogue system.

These arguments corroborate the findings presented in Section 3.3. In a deep linguistic analysis, we thus should leave underspecified certain distinctions, in a shallow analysis, we can concentrate on partial parses. A similar argumentation can be found in (Aust and Schroer, 1998, p. 32); the authors compare the impact of insertions, substitutions and deletions on user acceptance: deletions cause little problems, substitutions are more serious, and most serious are insertions, because "... the systems seems to 'assume' something the caller never mentioned." In analogy, we can say that in our example, the deletion of the marked sentence mood 'question' is less serious than if, the other

way round, the unmarked sentence mood 'statement' was changed into 'question'. This depends of course on the general dialogue structure which always needs a confirmation of time and place, no matter, whether it is required explicitly (via a question) or implicitly (via a statement). This means that in such a dialogue setting, especially turn final statements are always more 'question prone' (Batliner, 1989b) than, e.g., in a monologue, and this means in turn that in such constellations, if the functional load of prosody is very low, one should simply not rely on a prosodic classifier because the costs of an erroneous classification are rather high. For such an approach we need, however, a very close interaction between the different modules of the system.

4. Concluding remarks

In this paper, we focused on the 'What' and on the 'How' concerning the use of prosody in automatic dialogue systems: what we are working with and what we are working on in Section 2, and how we are working, i.e., which methodology we should use, in Section 3. There, we gave some examples for the present state-of-the-art and for promising trends out of our own work. We put a stronger emphasis on *shallow* analysis, automatic learnability and an easy adaptation to new applications. We aim at an integration of all available knowledge sources in a global search procedure; hard decisions should be taken as late as possible. A flexible use of knowledge sources means at least:

1. From a paradigmatic point of view, we should use those units we are interested in, if enough data are available. As a fall back, we can use substitutes that model these units indirectly, e.g., syntactic boundaries instead of DA boundaries.
2. From a syntagmatic point of view, we should use the maximum context available for a given database. As a fall back, we can use less context if this can be modelled more adequately. Thus, for our language model for syntactic boundaries, we use trigrams, for DA classification, we use 4-grams, and for DA sequences, we use bigrams.
3. From a pragmatic point of view, we should concentrate on those parts of an utterance that contain the crucial information (e.g., partial parsing, accentuated words in focus).

We assume that such highly sophisticated methods correspond closely with the strategies of human beings in human–human communication – but this is, basically, yet another story.

Acknowledgements

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 810939-9. The responsibility for the contents lies with the authors.

References

Albesano, D., Baggia, P., Danieli, M., Gemello, R., Gerbino, E., Rullent, C., 1997. A robust system for human–machine dialogue in telephony-based applications. Internat. J. Speech Technol. 2, 101–111.

Aust, H., Schroer, O., 1998. Application development with the Philips Dialog System. In: Proc. 1998 Internat. Symp. Spoken Dialogue (ISSD 98), Sydney, Australia, pp. 27–34.

Aust, H., Oerder, M., Seide, F., Steinbiss, V., 1995. The Philips automatic train timetable information system. Speech Communication 17, 249–262.

Batliner, A., 1989a. Fokus, Modus und die große Zahl. Zur intonatorischen Indizierung des Fokus im Deutschen. In: Altmann, H., Batliner, A., Oppenrieder, W. (Eds.), Zur Intonation von Modus und Fokus im Deutschen, Niemeyer, Tübingen, pp. 21–70.

Batliner, A., 1989b. Wieviel Halbtöne braucht die Frage? Merkmale, Dimensionen, Kategorien. In: Altmann, H., Batliner, A., Oppenrieder, W. (Eds.), Zur Intonation von Modus und Fokus im Deutschen, Niemeyer, Tübingen, pp. 111–162.

Batliner, A., 1991. Ein einfaches Modell der Frageintonation und seine Folgen. In: Klein, E., Duteil, F.P., Wagner, K. (Eds.), Betriebslinguistik und Linguistikbetrieb, Niemeyer, Tübingen, pp. 147–160.

Batliner, A., Weiand, C., Kießling, A., Nöth, E., 1993. Why sentence modality in spontaneous speech is more difficult to classify and why this fact is not too bad for prosody. In: House, D., Touati, P. (Eds.), Proc. ESCA Workshop on Prosody, Department of Linguistics, Lund University, Lund, pp. 112–115.

Batliner, A., Kießling, A., Burger, S., Nöth, E., 1995. Filled pauses in spontaneous speech. In: Proc. 13th Internat. Congress of Phonetic Sciences, Vol. 3, Stockholm, pp. 472–475.

Batliner, A., Kompe, R., Kießling, A., Mast, M., Niemann, H., Nöth, E., 1998. M = Syntax + Prosody: a syntactic–prosodic labelling scheme for large spontaneous speech databases. Speech Communication 25 (4), 193–222.

Batliner, A., Buckow, J., Huber, R., Warnke, V., Nöth, E., Niemann, H., 1999a. Prosodic feature evaluation: Brute force or well designed? In: Proc. 14th Internat. Congress of Phonetic Sciences, Vol. 3, San Francisco, pp. 2315–2318.

Batliner, A., Nutt, M., Warnke, V., Nöth, E., Buckow, J., Huber, R., Niemann, H., 1999b. Automatic annotation and classification of phrase accents in spontaneous speech. In: Proc. European Conf. on Speech Communication Technol., Vol. 1, Budapest, Hungary, pp. 519–522.

Batliner, A., Buckow, A., Niemann, H., Nöth, E., Warnke, V., 2000. The prosody module. In: Verbmobil: Foundations of Speech-to-Speech Translations. Springer, Berlin, pp. 106–121.

Block, H., 1997. The language components in Verbmobil. In: Proc. Internat. Conf. on Acoustics, Speech and Signal Processing, Vol. 1, München, pp. 79–82.

Carletta, J., Dahlbäck, N., Reithinger, N., Walker, M., 1997. Standards for dialogue coding in natural language processing. Dagstuhl-Seminar-Report 167.

Gallwitz, F., Aretoulaki, M., Boros, M., Haas, J., Harbeck, S., Huber, R., Niemann, H., Nöth, E., 1998a. The Erlangen spoken dialogue system EVAR: a state-of-the-art information retrieval system. In: Proc. 1998 Internat. Symp. Spoken Dialogue (ISSD 98), Sydney, Australia, pp. 19–26.

Gallwitz, F., Batliner, A., Buckow, J., Huber, R., Niemann, H., Nöth, E., 1998b. Integrated recognition of words and phrase boundaries. In: Proc. Internat. Conf. on Spoken Language Processing, Vol. 7, Sydney, pp. 2883–2886.

Gallwitz, F., Niemann, H., Nöth, E., Warnke, V., 1999. Prosodic information for integrated word-and-boundary recognition. In: Proc. ESCA Workshop on Dialogue and Prosody, Eindhoven, Netherlands, pp. 163–168.

Grice, M., Reyelt, M., Benzmüller, R., Mayer, J., Batliner, A., 1996. Consistency in transcription and labelling of german intonation with GToBI. In: Proc. Internat. Conf. on Spoken Language Processing, Vol. 3, Philadelphia, pp. 1716–1719.

Hörmann, H., 1978. Meinen und Verstehen. Suhrkamp Taschenbuch Wissenschaft. Suhrkamp, Frankfurt.

Jekat, S., Klein, A., Maier, E., Maleck, I., Mast, M., Quantz, J., 1995. Dialogue acts in Verbmobil. Verbmobil Report 65.

Kasper, W., Kiefer, B., Krieger, H., Rupp, C., Worm, K., 1999. Charting the depths of robust speech parsing. In: Proc. 37th Meeting of the ACL, pp. 405–412.

Kießling, A., 1997. Extraktion und Klassifikation prosodischer Merkmale in der automatischen. Sprachverarbeitung. Berichte aus der Informatik. Shaker, Aachen.

Kießling, A., Kompe, R., Batliner, A., Niemann, H., Nöth, E., 1994. Automatic labeling of phrase accents in German. In: Proc. Internat. Conf. on Spoken Language Processing, Vol. 1, Yokohama, pp. 115–118.

Kompe, R., 1997. Prosody in Speech Understanding Systems. Lecture Notes for Artificial Intelligence. Springer, Berlin.

Lea, W., 1980. Prosodic aids to speech recognition. In: Lea, W. (Ed.), Trends in Speech Recognition. Prentice-Hall, Englewood Cliffs, NJ, pp. 166–205.

Levelt, W., 1989. Speaking: From Intention to Articulation. MIT Press, Cambridge, MA.

Mast, M., Kompe, R., Harbeck, S., Kießling, A., Niemann, H., Nöth, E., Warnke, V., 1996. Dialog act classification with the help of prosody. In: Proc. Internat. Conf. on Spoken Language Processing, Vol. 3, Philadelphia, pp., 1728–1731.

Mecklenburg, K., Heisterkamp, P., Hanrieder, G., 1995. A robust parser for continuous spoken language using prolog. In: Proc. Fifth Internat. Workshop on Natural Language Understanding and Logic Programming (NLULP 95), Lisbon, pp. 127–141.

Niemann, H., Nöth, E., Batliner, A., Buckow, J., Gallwitz, F., Huber, R., Warnke, V., 1998. Using prosodic cues in spoken dialog systems. In: Proc. Internat. Workshop SPEECH AND COMPUTER (SPECOM'98), St-Petersburg, pp. 17–28.

Nilsson, N., 1982. Principles of Artificial Intelligence. Springer, Berlin.

Nöth, E., Boros, M., Haas, J., Warnke, V., Gallwitz, F., 1999. A hybrid approach to spoken dialogue understanding: prosody, statistics and partial parsing. In: Proc. European Conf. on Speech Communication and Technology, Vol. 5, Budapest, Hungary, pp. 2019–2022.

Nutt, M., Batliner, A., Warnke, V., Nöth, E., 1999. Using phrase accent information for dialogue act recognition in spontaneous German speech. In: Proc. ESCA Workshop on Dialogue and Prosody, Eindhoven, Netherlands, pp. 151–155.

Reyelt, M., Batliner, A., 1994. Ein Inventar prosodischer Etiketten für Verbmobil. Verbmobil Memo 33.

Shriberg, E., Bates, R., Taylor, P., Stolcke, A., Jurafsky, D., Ries, K., Cocarro, N., Martin, R., Meteer, M., Ess-Dykema, C.V., 1998. Can prosody aid the automatic classification of dialog acts in conversational speech?. Language and Speech 41, 439–487.

Spilker, J., Weber, H., Görz, G., 1999. Detection and correction of speech repairs in word lattices. In: Proc. European Conf. on Speech Communication Technol., Vol. 5, Budapest, Hungary, pp. 2031–2034.

Spilker, J., Klarner, M., Görz, G., 2000. Processing self-corrections in a speech-to-speech system. In: Verbmobil: Foundations of Speech-to-Speech Translations. Springer, Berlin, pp. 131–140.

Strom, V., Widera, C., 1996. What's in the "Pure" prosody?. In: Proc. Internat. Conf. on Spoken Language Processing, Vol. 3, Philadelphia, pp. 1497–1500.

Taylor, P., King, S., Isard, S., Wright, H., 1999. Intonation and dialogue context as constraints for speech recognition. Language and Speech 41, 489–508.

Vaissière, J., 1988. The use of prosodic parameters in Automatic Speech Recognition. In: Niemann, H., Lang, M.,

Sagerer, G. (Eds.), Recent Advances in Speech Understanding and Dialog Systems, Vol. 46. NATO ASI Series F. Springer, Berlin, pp. 71–99.

Wahlster, W. (Ed.), 2000. Verbmobil: Foundations of Speech-to-Speech Translations. Springer, New York.

Wahlster, W., Bub, T., Waibel, A., 1997. Verbmobil: The combination of deep and shallow processing for spontaneous speech translation. In: Proc. Internat. Conf. on Acoustics, Speech and Signal Processing, Vol. 1, München, pp. 71–74.

Warnke, V., Kompe, R., Niemann, H., Nöth, E., 1997. Integrated dialog act segmentation and classification using prosodic features and language models. In: Proc. European Conf. on Speech Communication Technol., Vol. 1, pp. 207–210.

Warnke, V., Gallwitz, F., Batliner, A., Buckow, J., Huber, R., Nöth, E., Höthker, A., 1999. Integrating multiple knowledge sources for word hypotheses graph interpretation. In: Proc. European Conf. on Speech Communication Technol., Vol. 1, Budapest, Hungary, pp. 235–239.