From: Proceedings of the ISCA Tutorial and Research Workshop on Speech recognition and understanding, Red Bank, NJ, 2001, pp. 23–28

Whence and Whither Prosody in Automatic Speech Understanding: A Case Study.

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

Chair for Pattern Recognition
University of Erlangen-Nuremberg, Germany
batliner@informatik.uni-erlangen.de, http://www5.informatik.uni-erlangen.de

Abstract

The 'case' this paper is dealing with is prosody research at the Chair for Pattern Recognition at the University of Erlangen–Nuremberg during the last fifteen years. We want to show how this mirrors the development of prosody research within automatic speech understanding in general. We sketch the realm of prosody in automatic speech understanding and relate the projects conducted to the research topics. This is illustrated in more detail with experimental results obtained within the last two years. Emphasis is put on the interplay between prosodic information and other knowledge sources.

1. Introduction

In this paper, we want to sketch (1) the development of prosody research in automatic speech understanding (ASU) during the last 15 years and (2) especially during the last two years indicating promising trends for the next years to come. This might be a reasonable thing to do in a workshop on prosody and speech recognition[*] that takes place shortly after the turn of the century – not to speak of the turn of the millennium. A thorough account of this topic that takes into consideration all the work that has been done so far at different sites is, however, beyond the scope of this paper. We therefore want to concentrate on the research on prosody that has been conducted at the Chair for Pattern Recognition at the University of Erlangen–Nuremberg; this is thus the 'case study' mentioned in the title of this paper. We believe that this can be done because this work can be taken as a sort of blueprint for other sites that sometimes had another time table or another focus; the overall tendency is, however, roughly the same.

If one wants to know exactly what's going on, one has to carefully manipulate and select the material, if one wants to check whether it really works one has, alas, to use this information within real-life settings. Thus, knowledge has been always sort of 'lagging behind' in the sense that it is rather precise but obtained for not fully realistic experimental settings. In a very early paper on automatic processing of prosody [21], Lieberman already stressed first the incompleteness of the prosodic features used, and second, that prosody employs redundant information and trading relation between different features at the same time. The material consisted of prosodic minimal pairs and elicited, read speech. This has been (and quite often still is) the usual way to exclude the multifarious intervening factors in real life situations. This approach from basic research was taken over by the early attempts to incorporate prosodic knowledge in ASU. Thus, in the eighties of the last century, we began with carefully selected material – also minimal pairs – and with elicited speech. In the nineties, we started with uncontrolled spontaneous speech, taking into account more and more not only prosodic but also other linguistic parameters.

In the following, we first try to relate the (diachronic) sequence of projects on prosody within ASU listed in section 2 with a (sort of synchronic) overview of relevant topics within this field given in section 3. This relationship turns out to be systematic: we began with topics that represent linguistics proper, as, e.g., boundaries, accents, and sentence modality, broadening the view in a second phase by taking into account pragmatics (i.e., dialogue acts) as well. Then, speech in a wider, paralinguistic meaning became the object of investigation, namely emotions and user states. Very recently, language and speech are no longer the exclusive objects of research but are investigated as some of many means to convey meaning and intention, together with mimic and gesture. In section 4, we present recent experimental results obtained within the last two years and approaches based on these results that illustrate some of the main topics dealt with before: feature evaluation, the combination of knowledge sources for different tasks, and a new module for the recognition of user states. We conclude with some remarks on desiderata and likely trends in the near future.

Note that in this paper, we cannot introduce material, features, and statistic procedures used in more detail; for that, we have to refer to the pertinent papers. We concentrate on the basic approaches and on the experimental results which illustrate the general tendencies outlined in this section and in section 2.

2. The Interplay between Research and Projects

Researchers help defining the goals of projects and projects define the work that has to be done by the researchers. New projects have to introduce new features in order to be attractive enough. Since the mid eighties, our institute was closely connected to or active in prosody research within the following projects; for each project, duration, scenario, speech material, and main topics are given:

- DFG–project[†] **'modus–focus–intonation'** (in Munich), 1984–1989: elicited speech, prosodic marking of accents and sentence modality
- DFG–project **'intonation–register–modus–focus'** (in Munich), 1989–1992: spontaneous speech ('blocks world'), prosodic marking of accents and sentence modality in spontaneous vs. read speech

[*] Terminology is notoriously ambiguous – we want to reserve the term 'speech recognition' for 'pure' word recognition, and use 'speech understanding' in a broad sense, encompassing all levels of speech processing.

[†] DFG is the German Research Council; such grants are given to basic research projects. All the other projects listed in this section were funded by the German Ministry of Education, Science, Research, and Technology (BMBF).

- **speech understanding systems** (‘Sprachverstehende Systeme’), 1985–1990: also known as ‘SPICOS’, elicited speech, automatic prediction of accent in German, mainly used for word recognition
- **ASL** 1991–1992: train information inquiry, elicited speech: automatic prediction of boundaries, accents, question/non-question
- **Verbmobil 1**, 1993–1996: spontaneous speech, appointment scheduling dialogues, push-to-talk, automatic prediction of boundaries, accents, question/non-question, and dialogue acts in German and English
- **Verbmobil 2**, 1997–2000: spontaneous speech, (mostly) appointment scheduling dialogues, barge in, automatic prediction of boundaries, accents, question/non-question, dialogue acts, and speech repairs in German, English (and Japanese); emotion (user’s state)
- **SmartKom**, 1999-2003: spontaneous, multi–modal dialogue system (speech, gestures, facial expressions), automatic prediction of boundaries, accents, question/non-question, and of emotions (user’s states) via speech and mimic

What we see is a broadening of the scope along several dimensions: from read to spontaneous speech, from syntax/semantics to dialogue, from linguistics to emotions to multimodality.

3. The Last 15 Years

Prosody is maybe the phonetic (and by that, acoustic) phenomenon that is most inter–disciplinary, ‘across levels’, and not ‘within (one single) level’ of analysis. This can be illustrated by Table 1 where we try to sketch the realm of prosody in ASU. The numbers to the left simply denote the chronological order in which we want to present the different **main fields** which are, so to speak, the ‘smaller scientific drawers’. The ‘larger scientific drawers’ are denoted with the three headings *LINGUISTICS*, *PARALINGUISTICS*, and *PHYSIOLOGY*. Within linguistics, there are the fields of **phonetics**, **syntax/semantics** (as ‘core linguistics’), **pragmatics**, and those dealing with different **varieties** of language (dialects, sociolects) or with different **languages**. The column ‘primary objects’ denotes the phenomena one mostly has to deal with – the objects of investigation, and ‘primary units’ those units extending over a certain time that mostly are taken as units of investigation. The **means** that are used are **prosodic** and **other** acoustic/linguistic means. Prosodic means can be local and the features are in such case normally macro features, i.e., computed locally (‘structured prosodic features’ [14]), for a linguistically meaningful unit (syllable, word, phrase) and leaving aside microprosodic phenomena. Global units are larger and often given trivially in the database, e.g., turns or whole stories. For these units, we can of course use the same features as for local one, e.g., F0 maxima and minima, duration, aso. At the same time, microprosodic phenomena (jitter, shimmer, aso.) are often computed.[‡] Other acoustic means are most of the time **spectral** features. Other linguistic means are **lexicon** (words, word classes, etc.) and – shallow or deep – **syntactic** structure. In the last column, the main applications are listed; to the right of the table, we indicate the years of our BMBF projects that focused each on one or two aspects of applied prosody.

We use the term *INDEXICAL* for those aspects that (normally) do not change at all or only within a longer stretch of time: individual traits, languages, dialects and sociolects, whereas not *INDEXICAL* aspects are more or less arbitrary (in the sense of de Saussure’s *arbitraire du signe*) and at the same time, local. Typically, the prosodic parameters used for indexical aspects are global, whereas the other ones are local because here, changes happen more frequently.

The prosodic aspects of the main applications that are listed in Table 1 in systematic order (last column) can be sketched – in ‘chronological order’ (numbers to the left of Table 1) – as follows:

1. **word recognition:** A prior classification into +/- accentuated syllables/words supports word recognition [23].
2. **syntactic/semantic parser:** The classification of (phrase) boundaries and accents as well as of sentence modality disambiguates syntactic/semantic readings [17, 24].
3. **dialogue processing:** The same prosodic features as those used for syntactic/semantic parsing can be used for the recognition of dialogue acts and dialogue act boundaries [25].
4. **automatic dialogue systems:** Critical phases in a dialogue between human user and automatic systems can be recognized with the help of prosodic features [5].
5. **language identification:** Languages can be told apart with the help of acoustic – amongst them prosodic – features.
6. **adaptation:** Different varieties of a language can be re–trained with the help of acoustic – amongst them prosodic – features.
7. **speaker adaptation, speaker verification and speaker identification:** For these three tasks, besides acoustic features, prosodic features can be used as well.
8. **diagnostics:** For automatic procedures as, e.g., screening or diagnostics of more or less pathological speaker traits, prosody can be used [22].

We want to stress that such a table cannot cover everything, and that it gives only a very coarse picture, i.e., conveys the main tendencies.[§] The table should thus rather be conceived of as a sort of associative map: if experts are being asked what they associate with *a*, they most likely might answer: *b*. Thus, if experts are being ask what they associate with ”accent” within ASU, they most likely might answer: The use of accent information for syntactic/semantic parsing. If they are asked what they think is most important in connection with more elaborate automatic dialogue systems, as far as prosody is concerned - and if it is not about parsing and dialogue acts - they really might answer: the recognition of emotions/user states in order to find trouble in communication.

A ‘classic’ paper is [18], cf. as well [34]. Prosody research at our institute began in the mid eighties, with the first publications in 1987 and 1988 on sentence mood and accentuation (in German), cf. [7]. Boundary detection was not very important at the beginning, because the sentences were rather short, most of the time read. The focus was on elicited, controlled speech,

[‡] The dichotomies local/global and macro/micro are thus not clear–cut but can be considered as significant tendencies.

[§] For instance, it can of course be doubted whether prosody used within word recognition should only be attributed to phonetics and, the other way round, whether prosody used for syntactic/semantic parsing should only be attributed to pure linguistics and not to phonetics as well.

LINGUISTICS

	main fields	primary objects	primary units	means:			main applications	projects
				prosody		other		
1.	phonetics	accents	syll./words	local	macro	lexicon	word recognition	85–90
2.	syntax/ semantics	accents boundaries sentence mood	syllables words phrases			lexicon syntax etc., spectrum	syntactic/semantic parser	91–03
3.	pragmatics	dialogue acts					dialogue processing	93–00
5.	*INDEXICAL*	languages	everything	local	macro		language identification	
6.	varieties *INDEXICAL*	dialects sociolects		global	micro		adaptation	

PARALINGUISTICS

	main fields	primary objects	primary units	means:			main applications	
				prosody		other		
4.	emotional	emotions attitudes speaker's state	phone turn	global (local)	micro (macro)	lexicon syntax etc., spectrum	automatic dialogue systems	97–03
7.	individual *INDEXICAL*	speaker speech style		local global	macro micro		adaptation, speaker verification/identification	

PHYSIOLOGY

	main fields	primary objects	primary units	means:			main applications
				prosody		other	
8.	biological *INDEXICAL*	speaker's state/ idiosyncrasies	phone turn	global	micro	spectrum	diagnostics
		pathology					

Table 1: The Realm of Prosody – and of Everything else – in Automatic Speech Understanding

on rather unspecific domains, only some few prosodic features were used, and only some 'core'–linguistic phenomena (boundaries, accents, questions) were dealt with.

Eventually, the focus was broadened towards spontaneous speech, specific applications/domains, and a large prosodic feature vector. The features used were not confined to prosody alone, i.e., prosody was used in combination with other knowledge sources; further phenomena, as, e.g., dialogue structure and emotion/user state became more important.

We want to illustrate the second and the fourth field with experimental results obtained during the last two years; a state–of–the–art report on prosody and dialogue processing covering the third field has already been given in [25].

4. The Last Two Years

If we had to characterize shortly the development of prosody research in the last years which surely can be extrapolated into the future, we could say: *'Prosody goes multi':* multi–feature, multi–knowledge, multi–function, multi–modal, multi–lingual, etc.[¶] This development is conditioned by new, more realistic requirements of spontaneous speech databases and more complex tasks within human–computer–communication. By that, the modelling of prosody in ASU is getting less uni–dimensional, and certainly more adequate from a psycholinguistic point of view – not necessarily because prosody works within ASU the same way as psycholinguistic theories assume, but because its contribution to understanding is more fully exploited.

We try to deal with the interesting aspects under several 'multi'–headings; of course, different headings and different topics under the specific headings could be imagined. The experimental results we will present have been obtained during the last two years.

4.1. Multi–Feature

In the first projects during the eighties, we only used some few prosodic features, cf. [7, 23]. In the VERBMOBIL project, a large feature vector was developed comprising up to 276 syllable–based and word–based features modelling F0, duration, energy, and pause, cf. [14, 24]. As phone segmentation cannot be obtained from the word recognition module, we eventually confined ourselves to 95 word–based features only, cf. [2, 1]. It is of course not the exact number of (many) features which is important but the basic approach to use many different prosodic features that often are highly correlated with each other, and to leave it to the statistic classifier to find the appropriate weights for each feature. We certainly do not believe that 95 is 'the magical number' of prosodic features one should use. In our different attempts to select the relevant features we most of the time ended up with a number which is larger than some 20 and less than some 50 features, cf. [3, 4]. For the moment, this might be the realistic range for the 'optimal' number of features within our approach. The feature set we have developed within VERBMOBIL 2 is described in another paper presented at this

[¶] There is another important 'multi'–aspect that we do not deal with in this paper: the Multi–Partner aspect within large, distributed systems. The solution that was chosen for VERBMOBIL is described in [35], cf. as well *http://www.dfki.de/ bert/bellagio-2000-folien/sld001.htm* workshop; thus, we do not need to present the single features here in more detail and only want to refer to [1].

It is possible to boil down the whole feature set onto some two to five principal components (PCs) as predictor variables without a considerable loss of classification performance [4]. By that, we can have a look at the relevance of single features/feature groups for the classification of accents and boundaries in German and English. It turned out that:

- Within the two languages, the impact of PCs is very similar for boundaries and accents.
- Across the two languages, there is a similar impact of PCs.
- The order of relevance is: first comes duration, then energy, then pause, then F0.

Similar results were obtained in other studies, cf. [30, 13], especially, as far as the impact of duration features is concerned. In [4, 1], we try to interpret this ranking. Such an approach that represents the state–of–the–art nowadays can be characterized as a sort of 'shot–gun' approach: a highly redundant feature set is used, and by that, chances are that no information is lost. Our experience throughout is, that results are not markedly worse if the whole feature set is used even if a reduction of the number of features often results in slightly better classification rates.

We believe that we have reached a good level of adequacy, as far as our feature set is concerned; thus during the last years, the basics of our feature set were not dealt with anew. Instead, more weight was put onto the combination of prosody with other linguistic knowledge, cf. section 4.2. It might still take some more time to exploit this combination of knowledge sources, but eventually, it might be necessary to have a second look at the prosodic features themselves, cf. section 4.3.

4.2. Multi–Level, Multi–Knowledge, and Multi-Function

Acoustics alone is not enough in word recognition but has to be combined with language model (LM) information. The same holds for prosody: 'pure' prosody can of course be used alone, cf. [32], but combined with other knowledge sources, better classification results can be achieved. Actually, LMs are almost as good as prosodic information to predict the position of boundaries and accent that are annotated perceptually, without taking into account syntactic information; moreover, LMs are normally markedly better in predicting boundaries and accents that are annotated syntactically, cf. [6, 8]. Still, a combination of different knowledge sources across different levels of analysis almost always yields the best results – and after all, even a small improvement is welcome if performance is already well above 80% correct classification.

The multi–functional aspect is the other side of the story: on the one hand, several knowledge sources – amongst them prosody – contribute to the recognition of a specific phenomenon. On the other hand, prosody is multi–functional, i.e., it contributes to the recognition of different phenomena. In 4.2.1, we present results obtained for a rather 'classic' task, namely the recognition of prosodic boundaries, and show how a combination of prosodic features with part–of–speech (POS) features and LM information improves classification. In section 4.2.2, prosody is used in a sort of preprocessing step for the recognition and subsequent processing of speech repairs. Finally, in section 4.2.3, we present a new approach towards the use of prosodic and other linguistic knowledge for finding critical phases (trouble in communication) in conversations with automatic dialogue systems.

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

Table 2: Recognition results for phrase boundary recognition in German with different combinations of acoustic and syntactic knowledge; CL: class-wise computed recognition rate, i.e., mean of the recognition rates for the two classes, RR: overall recognition rate

4.2.1. Combining Language Models and Neural Networks for the Classification of Boundaries

The prosodic feature set and the POS features used for the experiments described in this section are described elsewhere in several papers, e.g., in [2, 1]; the features model duration, energy, pause, and F0 information for a context of five words, as well as POS information for the same context. In previous experiments, it could be shown that a combination of LM classifiers and neural networks (NNs) that are trained only on acoustic–prosodic features, yielded better results for prosodic boundary classification than each of the classifiers alone [17, 6]. This shows that syntactic and acoustic knowledge have to be combined in some way to achieve optimal results.

In other experiments [9], we have shown that the classification performance of the baseline NNs can be improved by POS features; cf. as well [1]. Adding POS flags to the acoustic-prosodic features is, basically, another way of combining acoustic and syntactic knowledge. Thus, after those experiments, it still has to be shown that the POS information added during the acoustic-prosodic classification is not redundant to the syntactic information that can later be added by an LM.

We therefore performed several experiments with different combinations of LMs and NNs (with and without POS flag features). The combinations and classification results are shown in Table 2. In the Table, POS-LM denotes an LM trained on the sequence of POS classes instead of the spoken words. LM denotes an LM which was trained on the sequence of words; a very fine-grained category system was used in order to deal with the limited training data. NN means an NN trained on acoustic-prosodic features alone, whereas NN-POS means an NN trained on acoustic-prosodic features and POS flags. The knowledge sources are combined linearly. The optimal weighing factors are determined on a training database.

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

4.2.2. Prosody and Repairs

Speech repairs constitute a problem for the parsing of spontaneous speech: they should not be processed as such but rather disregarded. Obligatory parts of a repair are the reparandum –

the ‘wrong’ part of the utterance, and the reparans – the correction of the reparandum. Between these two is the Interruption Point IP which is often marked prosodically. In the utterance *ja ist in Ordnung Montag **IP** hm Sonntag den vierten* (yes it’s ok Monday **IP** uh Sunday the fourth), the result of syntactic analysis should rather be *ja ist in Ordnung Sonntag den vierten* (yes it’s ok Sunday the fourth). In [31], we describe a repair module within the VERBMOBIL system that performs this task. The first step in this module is the localization of the IP with the help of the prosody module. This module classifies each word boundary in the word hypotheses graph as a regular or an irregular boundary. Irregular boundaries are seen as hypotheses for IPs. For each word boundary, a vector with prosodic and POS features is determined. Table 3 shows the problem of a pure prosodic detection. 91% of all IPs are found but there are many false alarms. This is a general problem of binary statistic classifiers in cases where the proportion of the two classes is extreme. So what can be achieved with prosody alone is not a good overall classification but an impressive reduction of the search space: we only disregard some 10% of the IPs and can reduce the number of possible IPs that have to be processed further by the repair module from 33.167 to 18.878! This demonstrates that prosody cannot only be used successfully in a fully integrated approach, cf. section 4.2.1, but in a sequential approach as well. [‖]

	Recognized	
Reference	IP	¬IP
IP	502	57
¬IP	18376	33110

Table 3: Results for prosodic interruption point (IP)–detection

4.2.3. Prosody and Emotion

Automatic dialogue systems used in call-centers, for instance, should be able to determine in a critical phase of the dialogue - indicated by the costumers vocal expression of anger/irritation - when it is better to pass over to a human operator. At a first glance, this seems not to be a complicated task: It is reported in the literature that emotions can be told apart quite reliably on the basis of prosodic features. However, these results are most of the time achieved in a laboratory setting, with experienced speakers (actors), and with elicited, controlled speech.

In a first step, we collected data from a single, experienced, acting person. These data comprise 1240 ‘neutral’ turns produced within the VERBMOBIL scenario that were collected for reasons independent of the aims of this study, and 96 turns in which the speaker was asked to imagine situations in which the VERBMOBIL system was malfunctioning and in which he was getting angry, for instance: *Das ist doch unglaublich!* (That’s really unbelievable!). These data are referred to in the following as ACTOR data. In a second step, data were elicited from 19 more or less ‘naive’ subjects who read 50 neutral and 50 emotional sentences each (the subset of the emotional sentences was a subset of the emotional utterances produced in the ACTOR scenario). These data are referred to as READ data. In a third, more elaborate step, a WOZ scenario [12] was designed to provoke reactions to probable system malfunctions with the following aims:

- The experimental design should elicit speakers’ spontaneous, unprompted, reactions to different kinds of possible system malfunctions.
- The design should enable us to identify changes in the emotional state of the respective speaker, that is, identify problematic phases in the interaction between the human user and the supposed system without reliance on intuitive judgments.
- The design should furthermore allow us to determine which linguistic properties may function as indicators of TROUBLE IN COMMUNICATION.

Table 4 shows all of the phenomena that we use at present for finding TROUBLE IN COMMUNICATION, the number of classes, how we obtained the labels, and which classifier(s) we use to find them. This combination constitutes just a sort of snapshot and is not yet a ‘full–grown’, unified approach. The reference we want to recognize is, for the ACTOR and READ data, the words in those turns that are produced as ‘emotional’, and for the WOZ data, words that are annotated as displaying prosodic peculiarities, as, e.g., lengthening, hyperarticulation, or emphasis; for details, cf. [5].

phenomena	#	**source**
prosodic features *	91	extracted automatically
part–of–speech features POS *	6	annotated in the lexicon by hand
dialogue act features DA	18	LM: trained with VERBMOBIL data, automatic annotation
prosodic peculiarities	10/2	annotated by hand
repetitions	2	annotated automatically (Levenshtein distance)
syntactic–prosodic S boundaries	5	LM: trained with VERBMOBIL data, automatic annotation

Table 4: Concepts used for WOZ classification; starred phenomena: used for ACTOR/READ as well; the first four phenomena are used for the experiments reported on in Table 5.

	Actor	Read	WOZ
# of cases	10316	13053	28649
features	avRec	avRec	avRec
prosodic	95.4	77.4	73.2
POS	72.2	63.0	66.1
POS, only 0	72.4	57.6	64.1
pros./POS	95.7	79.6	73.7

Table 5: LDA, leave-one-out, best classification result in percent with different feature combinations for Actor, Read, and WOZ.

Table 5 shows a comparison of classification results with different feature combinations for ACTOR, READ, and WOZ. Basically, good experimental results could be achieved for the ACTOR scenario, which mirrors most of the results reported on in the literature; for the READ data results were worse; the difference can be traced back to speaker idiosyncrasies and to

[‖] Note that the architecture of the whole VERBMOBIL system is rather sequential than integrated, due to the fact that different modules have been developed by different partners at different sites, cf. [16].

WOZ	
# of case	28649
features	avRec
DA	56.1
POS/DA	66.8
pros./DA	73.4
pros./POS/DA	**74.2**

Table 6: LDA, leave-one-out, best classification result in percent with DA information for WOZ.

the fact that speakers were less experienced. For the WOZ data, which is closest to the 'real-life'-task, classification results were even less convincing. We are thus faced with a well-known problem: The closer we get to the constellation we want to model (dialogue between automatic systems and 'naive' users/customers), the worse our recognition rates will be. The dilemma from our perspective is thus that the closer we get to real life applications, the less visible is emotion, which is why the target needs to be TROUBLE IN COMMUNICATION, and classification has to be based on a combination of different knowledge resources. In Tables 5 and 6, we display results for some of these possible combinations; it turns out that the more knowledge we use, the better the classification will be: in Table 5, it can be seen that normally, better classification can be achieved with the use of more feature classes, in Table 6, it can be seen that for the WOZ data, where dialogue act information is available, this information contributes to classification performance as well.

Until now, emotion is normally processed as a pure acoustic/prosodic phenomenon – as if it were purely indexical; we have seen, however, that such an approach is suboptimal if we have to deal with more natural data and not only with acted speech. Emotion should instead be treated along the same lines as linguistic phenomena, i.e., taking into account all other linguistic knowledge one can get, and eventually, non–linguistic knowledge as well, cf. section 4.3. In the following, we sketch our module **M**_onitoring_ **o**_f_ **U**_ser_ **S**_tate [especially of]_ **E**_motion_ MOUSE which combines these different context-dependent and independent properties in a single model. In the communication between system and user, the user behavior is supposed to mirror the state of the communication. If there are no problems (felicitous communication) or only minor problems (slight misunderstandings) which can be solved, the user behaves neutrally and is not emotionally engaged. If, however, there are severe recurrent misunderstandings (error 'spirals', cf. [20]), that is, if there is TROUBLE IN COMMUNICATION, then the user behavior changes accordingly; it is marked: overt signalling of emotions – changes in prosody, mimic, etc. – and particular, context-dependent strategies, i.e., different strategies to find ways out of these error spirals, can be observed. If there is such trouble, our module MOUSE should trigger an action, for instance, by initiating a clarification dialogue, cf. Figure 2. In such a case, the communication will recover gracefully. If, however, no action is taken, chances are that the user becomes more and more frustrated, and sooner or later he or she will break off the communication (dead end, point of no return).

Figure 1 gives a rough outline of the interaction of MOUSE with a dialogue system: Input into the system is a speech signal which is processed by the word recognizer and the language understanding component. Input into the dialogue manager is a semantic representation which is passed on, together with the speech signal, to MOUSE. If MOUSE recognizes an utterance as neutral, it signals 'no trouble', further normal dialogue processing is initiated, and an answer is generated and synthesized. If, however, MOUSE classifies the utterance as 'indicating trouble', an action as further specified in Figure 2 is initiated, and again, an answer is generated and synthesized.

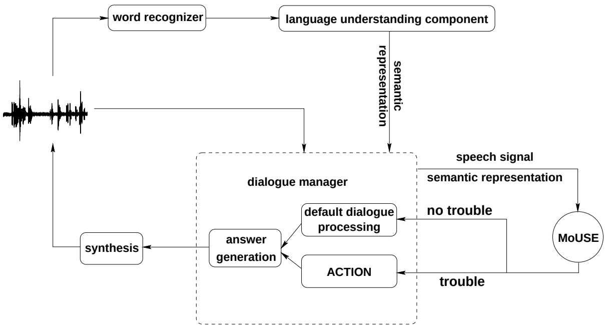

Figure 1: MoUSE: Overview

In Figure 2, the architecture of MOUSE is sketched in more detail. The components that are already implemented are highlighted. Starting point is a user independent training based on data that are as close to the intended application as possible. For training of the 'normal' modules other than MOUSE in an automatic dialogue system, such as word recognition, 'neutral' and 'emotional' data are processed together; for the training of the classifier of TROUBLE IN COMMUNICATION, separate classes have to be trained. For the actual use of this module, it might be advantageous to use a clearly defined neutral phase for the adaptation of the system. For each of the pertaining phenomena that can be found, a separate classifier is used whose output is a probability rating. All probabilities are weighted[**] and result in one single probability that triggers an action if it is above a certain value. This value has to be adjusted to the special needs of the application, for instance, whether one wants to get a high recall or a high precision, or whether both should be balanced. (If the costs of failing to recognize emotions are high – for instance, if important customers may be lost – recall should be high, even if there are many false alarms and by that, precision is low.) Retraining and a different weighting of classifier results may also be necessary for adaptation to different scenarios. The action invoked can at least be one of the following possibilities: Easiest is probably to return to a very **restricted, system–guided dialogue**; a **clarification dialogue** needs more sophistication; to **hand over to a human operator** means to cut off automatic processing but, of course, it is the most secure strategy to yield graceful recovery of the communication. A straightforward way of 'calming down' the user could be to **make the system apologize**, cf. [11].

From a methodological point of view, the strategy is thus the same for the 'core linguistic' phenomena boundaries and accents on the one hand, and for the paralinguistic phenomena emotion/user state on the other hand: not to use prosody alone but to combine it with several other knowledge sources; as for work along comparable lines, cf. [13, 19, 20, 26, 27].

[**] The different scores are weighted, similar to the LM weight used in speech recognition. We use an automatic procedure based on gradient descent for optimization, cf. [37].

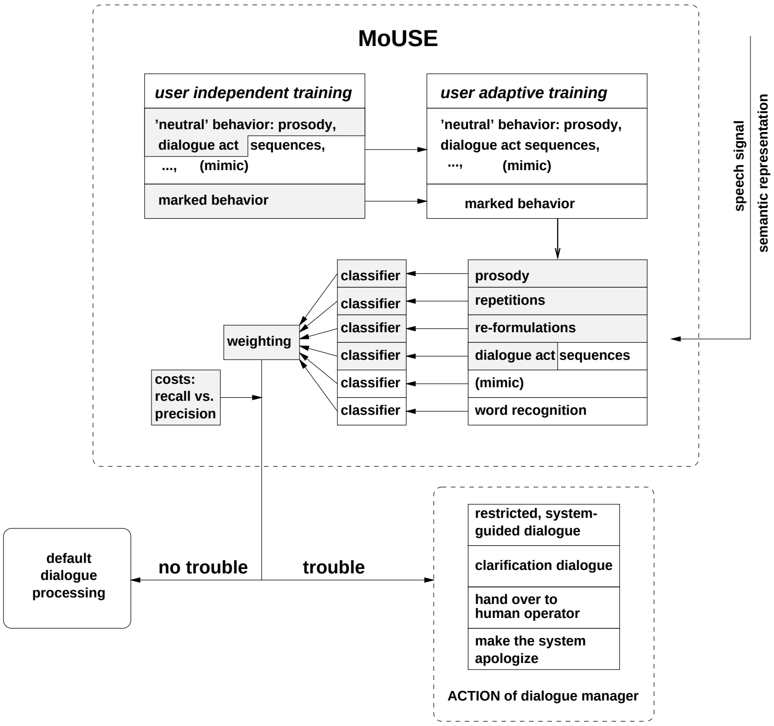
MoUSE
user independent training
'neutral' behavior: prosody, dialogue act sequences, ..., (mimic)
marked behavior
user adaptive training
'neutral' behavior: prosody, dialogue act sequences, ..., (mimic)
marked behavior
speech signal
semantic representation
classifier
prosody
repetitions
re-formulations
dialogue act sequences
(mimic)
word recognition
weighting
costs: recall vs. precision
default dialogue processing
no trouble
trouble
restricted, system-guided dialogue
clarification dialogue
hand over to human operator
make the system apologize
ACTION of dialogue manager

Figure 2: MoUSE: A Sketch of the Architecture

4.3. Multi–Modal

Even advanced state of the art human-machine interfaces are at most to a very rudimentary extent multi–modal, i.e., not as close to human–human–communication as possible. Such an interaction includes, besides the well-established machine input devices and speech, at least the use of the modalities mimic and gesture. Just as in human-human communication, these modalities should be used for transmission of information and, at the same time, to transmit the user state of the communication partner during interaction. By that we mean emotional states such as neutral, anger, or joy, as well as general states (stress, fatigue), and states shedding some light on the communication (surprise, being puzzled). SmartKom is such a multi–modal dialogue system that combines speech, gesture, and mimic as human input and as system output, cf. [36, 28]. In Figure 2, one of the classifiers that is not yet used is the one for mimic; the integration of this knowledge source is one of our main tasks within the SmartKom–project. We concentrate on the transmission of the user's state rather than on the detection of 'linguistically relevant' head movements that transmit information as, e.g., head nodding. The fusion of the information from the different modalities has to take into account different timing of different modalities: a user state can be indicated with one or several modalities, with complete, partial or no overlap.

Currently we are in the phase of integrating and fine-tuning the classifiers for the different modalities, and refrain therefore from giving preliminary results. What could be observed yet is two marked differences to the speech data obtained within the VERBMOBIL project: on the one hand, multi–modality favors elliptic speech because gesture can be used as a sort of anaphoric reference, and new phenomena, for instance off–talk, cf. [28] can be observed, and on the other hand, if we have to use data from a microphone array and not from a head–set microphone, the worse S/N ratio yields a drastic deterioration of the quality of the prosodic features extracted. Thus it might be not only necessary to re–train the classifiers but also to have a second look at feature extraction in general.

4.4. Multi–Lingual

Normally, a prosody module is developed and implemented for only one single language. In a multi–lingual system, however, for instance, a system for flight reservations that tries to encompass some of the main world languages, it is not optimal to stick to such an approach. In this section we want to show that it really pays off in terms of performance if only one multi–lingual prosody module is used.

In the VERBMOBIL system, prosodic information is computed for the three languages *German*, *English*, and *Japanese*; details can be found in [2]. First a prosody module for each of these languages was integrated in the system. Thus a lot of common data and procedures for all languages could not be shared. To reduce the memory requirements we integrated the language dependent modules into one *multilingual prosody module* where other languages easily can be added. The architecture of the multilingual prosodic module is shown in Figure 3.

It is possible to share the feature extraction and classification procedures in a multilingual module because they are language independent. The language dependent data, for instance, duration normalization tables, and specific classifiers are kept in different structures. Via configuration files individual classification parameters for each language, for instance, the different sizes of the n–grams, can be loaded. The prosody module has to deal with different incoming and outgoing data. The communication is done with the *Pool Communication Architecture* (PCA) which is described in [16]. Input into the prosody module is the speech signal and the word hypotheses graph (WHG), output is an annotated WHG, now including additional prosodic information for each word. Furthermore, a set of prosodic features is passed onto the synthesis module. In more detail, processing in the prosody module can be described as follows:

- The control component handles the global behavior of the prosody module, for instance: ‘get the WHG’, ‘start classification’. Furthermore, the language dependent behavior can be configured here, for instance, specific combinations of neural network classifiers and language model classifiers.
- The PCA in VERBMOBIL works event driven. Depending on which data pool first indicates incoming data, the handler for that particular data pool is called. Each data pool gets input from the word recognition module for one language. Thus, the control component selects the corresponding language dependent data, for instance, language–specific normalization tables, which are needed for the feature extraction.
- The WHG component then traverses the WHG. At each node the feature extraction component is called.
- The feature extraction component uses the language dependent data structure, the word hypotheses and word intervals from the WHG. The result is a feature vector which is passed to the classification component.
- The classification component classifies the feature vector using language dependent classifier information. For that we use neural networks which can be combined with language models. The classification result is handed back to the WHG component.
- The WHG component annotates the WHG correspondingly.
- After all edges of the WHG have been processed the annotated WHG is delivered to the output data pool.

The structure of the multilingual module has several advantages. It can be easily extended as mentioned above. In order to add a new language only a few changes to the configuration file have to be made, i.e. the language dependent parameters have to be set. Furthermore, the memory requirement of the multilingual module after some optimization steps (64 MByte) is a lot smaller than the sum of the memory needed for three modules (291 MByte).

It really might be possible to use identical feature sets for related languages; it still has to be proven whether this is possible for unrelated languages, and it might definitely not be possible to treat, for instance, tone and not–tone languages with the same approach.

5. Concluding Remarks

In this paper, we wanted to give an overview of the trends in prosody research for ASU during the last years, exemplified with the work that has been conducted at our institute. We concentrated on one – in our opinion central – aspect: how prosody outgrew the restrictions posed upon it in the laboratory and made some steps into the real world. By that, it surely lost some precision but gained more reality. Emphasis was put on the interplay of prosodic information with other knowledge sources and at the same time, on the work conducted during the last two years; for an overview of prosody and dialogue processing, cf. as well [25].

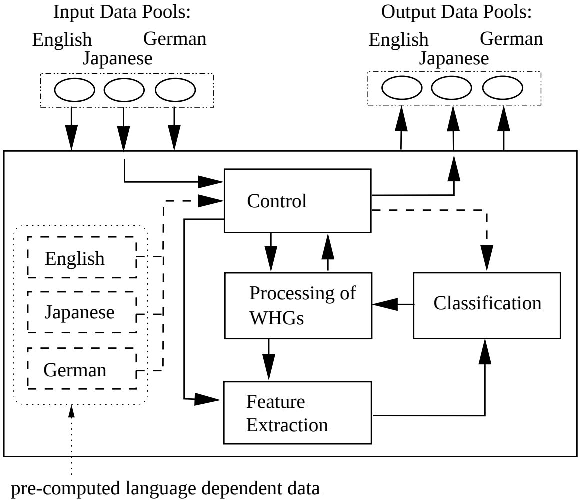

Figure 3: Architecture of the multilingual prosody module for prosodic processing.

If we go back to Table 1, we could say that we are just midways on the full exploitation of prosody as a knowledge source for ASU: topics 1-4 listed in Table 1 and in section 3 represent the core topics of prosody research conducted in larger projects; topics 5 to 8 are only dealt with sparsely until now, cf. [15, 33, 10, 29, 22]. We do not know yet whether this is because main stream research simply was not interested in these topics, or because prosody is not that important for these topics as a knowledge source. There is still plenty to do within those fields that at the moment are object of investigation, but after a while, we surely will come back to the old questions again, as, e.g., prosody and word recognition, or feature computation and selection.

Acknowledgments:
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 701 K5 and in the framework of the SmartKom project under Grant 01 IL 905 K7. The responsibility for the contents of this study lies with the authors.

6. References

[1] A. Batliner and E. Nöth and J. Buckow and R. Huber and V. Warnke and H. Niemann. Duration Features in Prosodic Classification: Why Normalization comes Second, and what they Really Encode. In *Proc. of the Workshop on Prosody and Speech Recognition 2001*, 2001. to appear.

[2] A. Batliner, A. Buckow, H. Niemann, E. Nöth, and V. Warnke. The Prosody Module. In Wahlster [35], pages 106–121.

[3] A. Batliner, J. Buckow, R. Huber, V. Warnke, E. Nöth, and H. Niemann. Prosodic Feature Evaluation: Brute Force or Well Designed? In *Proc. 14th Int. Congress of Phonetic Sciences*, volume 3, pages 2315–2318, San Francisco, 1999.

[4] A. Batliner, J. Buckow, R. Huber, V. Warnke, E. Nöth, and H. Niemann. Boiling down Prosody for the Classification of Boundaries and Accents in German and English. In *Proc. European Conf. on Speech Communication and Technology*, volume 4, pages 2781–2784, Aalborg, Denmark, 2001.

[5] A. Batliner, K. Fischer, R. Huber, J. Spilker, and E. Nöth. How to Find Trouble in Communication. *Speech Communication*, 2002. to appear.

[6] 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.

[7] A. Batliner and E. Nöth. The Prediction of Focus. In *Proc. European Conf. on Speech Communication and Technology*, pages 210–213, Paris, 1989.

[8] A. Batliner, M. Nutt, V. Warnke, E. Nöth, J. Buckow, R. Huber, and H. Niemann. Automatic Annotation and Classification of Phrase Accents in Spontaneous Speech. In *Proc. European Conf. on Speech Communication and Technology*, volume 1, pages 519–522, Budapest, Hungary, 1999.

[9] J. Buckow, A. Batliner, R. Huber, H. Niemann, E. Nöth, and V. Warnke. Detection of Prosodic Events using Acoustic–Prosodic Features and Part–of–Speech Tags. In *Proc. International Workshop SPEECH AND COMPUTER (SPECOM'00)*, pages 63–66, St-Petersburg, 2000.

[10] F. Cummins, F. Gers, and J. Schmidhuber. Automatic discrimination among languages based on prosody alone. Technical Report IDSIA-03-99, 1999.

[11] K. Fischer. Repeats, reformulations, and emotional speech: Evidence for the design of human-computer speech interfaces. In Hans-Jörg Bullinger and Jürgen Ziegler, editors, *Human-Computer Interaction: Ergonomics and User Interfaces, Volume 1 of the Proceedings of the 8th International Conference on Human-Computer Interaction, Munich, Germany.*, pages 560–565. Lawrence Erlbaum Ass., London, 1999.

[12] N.M. Fraser and G.N. Gilbert. Simulating Speech Systems. *Computer Speech & Language*, 5(1):81–99, 1991.

[13] J. Hirschberg, D. Litman, and M. Swerts. Prosodic cues to recognition errors. In *Proceddings of the Automatic Speech Recognition and Understanding Workshop (ASRU'99)*, pages 349–352, 1999.

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

[15] F. Klingholz, R. Penning, and E. Liebhardt. Recognition of Low-Level Alcohol Intoxication from Speech Signal. *Journal of the Acoustical Society of America*, (84):929–935, 1988.

[16] A. Klüter, A. Ndiaye, and H. Kirchmann. Verbmobil from a Software Engineering Point of View: System Design and Software Integration. In Wahlster [35], pages 635–658.

[17] R. Kompe. *Prosody in Speech Understanding Systems*. Lecture Notes for Artificial Intelligence. Springer–Verlag, Berlin, 1997.

[18] W. Lea. Prosodic Aids to Speech Recognition. In W. Lea, editor, *Trends in Speech Recognition*, pages 166–205. Prentice–Hall Inc., Englewood Cliffs, New Jersey, 1980.

[19] G.-A. Levow. Characterizing and recognizing spoken corrections in human-computer dialogue. In *Proceedings of Coling/ACL '98*, pages 736–742, 1998.

[20] G.-A. Levow. Understanding Recognition Failures in Spoken Corrections in Human–Computer Dialog. In M. Swerts and J. Terken, editors, *Proc. ESCA Workshop on Dialogue and Prosody*, pages 193–198, Eindhoven, 1999.

[21] P. Lieberman. Some acoustic correlates of word stress in american english. *JASA*, 32:451–454, 1960.

[22] M. Levit and R. Huber and A. Batliner and E. Nöth. Use of prosodic speech characteristics for automated detection of alcohol intoxination. In *Proc. of the Workshop on Prosody and Speech Recognition 2001*, 2001. to appear.

[23] E. Nöth. *Prosodische Information in der automatischen Spracherkennung — Berechnung und Anwendung*. Niemeyer, Tübingen, 1991.

[24] E. Nöth, A. Batliner, A. Kießling, R. Kompe, and H. Niemann. Verbmobil: The Use of Prosody in the Linguistic Components of a Speech Understanding System. *IEEE Trans. on Speech and Audio Processing*, 8:519–532, 2000.

[25] E. Nöth, A. Batliner, V. Warnke, J. Haas, M. Boros, J. Buckow, R. Huber, F. Gallwitz, M. Nutt, and H. Niemann. On the Use of Prosody in Automatic Dialogue Understanding. In M. Swerts and J. Terken, editors, *Proc. ESCA Workshop on Dialogue and Prosody*, pages 25–34, Eindhoven, 1999.

[26] S. Oviatt, J. Bernard, and G.-A. Levow. Linguistic Adaptations during Spoken and Multimodal Error Resolution. *Language and Speech*, 41(3–4):419–442, 1998.

[27] S. Oviatt, M. MacEachern, and G.-A. Levow. Predicting hyperarticulate speech during human-computer error resolution. *Speech Communication*, 24:87–110, 1998.

[28] R. Siepmann and A. Batliner and D. Oppermann. Using Prosodic Features to Characterize Off-Talk in Human-Computer-Interaction. In *Proc. of the Workshop on Prosody and Speech Recognition 2001*, 2001. to appear.

[29] F. Schaeffler and R. Summers. Recognizing German Dialects by Prosodic Features Alone. In *Proc. Int. Cong. of Phonetic Sciences*, volume 3, pages 2311–2314, San Francisco, Kalifornien, 1999.

[30] 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.

[31] J. Spilker, A. Batliner, and E. Nöth. How to Repair Speech Repairs in an End-to-End System. In R. Lickley and L. Shriberg, editors, *Proc. ISCA Workshop on Disflueny in Spontaneous Speech*, pages 73–76, Edinburgh, Schottland, 2001.

[32] V. Strom and C. Widera. What's in the "Pure" Prosody? In *Proc. Int. Conf. on Spoken Language Processing*, volume 3, pages 1497–1500, Philadelphia, 1996.

[33] A. Thymé-Gobbel and S. Hutchins. On using prosodic cues in automatic language identification. In *Proc. Int. Conf. on Spoken Language Processing*, volume 3, pages 1768–1771, Philadelphia, 1996.

[34] J. Vaissière. The Use of Prosodic Parameters in Automatic Speech Recognition. In H. Niemann, M. Lang, and G. Sagerer, editors, *Recent Advances in Speech Understanding and Dialog Systems*, volume 46 of *NATO ASI Series F*, pages 71–99. Springer–Verlag, Berlin, 1988.

[35] W. Wahlster, editor. *Verbmobil: Foundations of Speech-to-Speech Translations*. Springer, New York, Berlin, 2000.

[36] W. Wahlster, N. Reithinger, and A. Blocher. SmartKom: Multimodal Communication with a Life-like Character. In *Proc. European Conf. on Speech Communication and Technology*, volume 3, pages 1547–1550, Aalborg, Denmark, 2001.

[37] V. Warnke, F. Gallwitz, A. Batliner, J. Buckow, R. Huber, E. Nöth, and A. Höthker. Integrating Multiple Knowledge Sources for Word Hypotheses Graph Interpretation. In *Proc. European Conf. on Speech Communication and Technology*, volume 1, pages 235–239, Budapest, Hungary, 1999.