Chapter 1

PROSODIC MODELS, AUTOMATIC SPEECH UNDERSTANDING, AND SPEECH SYNTHESIS: TOWARDS THE COMMON GROUND?

Anton Batliner
Chair for Pattern Recognition, University of Erlangen-Nuremberg, Germany
batliner@informatik.uni-erlangen.de

Bernd Möbius
Institute of Natural Language Processing, University of Stuttgart, Germany
Bernd.Moebius@IMS.Uni-Stuttgart.DE

Abstract Automatic speech understanding and speech synthesis, two major speech processing applications, impose strikingly different constraints and requirements on prosodic models. The prevalent models of prosody and intonation fail to offer a unified solution to these conflicting constraints. As a consequence, prosodic models have been applied only occasionally in end-to-end automatic speech understanding systems; in contrast, they have been applied extensively in speech synthesis systems. In this chapter we aim to make explicit the reasons for this state of affairs by reviewing the role of prosodic modelling in these two fields of speech technology. Subsequently, possible strategies to overcome the shortcomings of the use of prosodic modelling in automatic speech processing are discussed. In particular, the question is raised whether or not there is a common framework for prosodic modelling in automatic speech understanding and speech synthesis systems, and if so, whether any particular model or theory of prosody can serve as a common ground. Finally, a catalogue of tasks in prosody research is proposed that ought to be relevant to both automatic speech understanding and speech synthesis and that might stimulate joint research activities.

Keywords: prosody, intonation model, automatic speech understanding, speech synthesis

Anton Batliner, Bernd Möbius (2005): "Prosodic models, automatic speech understanding, and speech synthesis: Towards the common ground?". In William J. Barry and Wim A. van Dommelen (eds.), The Integration of Phonetic Knowledge in Speech Technology (Springer, Dordrecht), 21-44.

Introduction

The application of prosodic models in automatic speech understanding (ASU)[1] and speech synthesis (TTS)[2] is strikingly different. In the latter, such models have been extensively applied, but there is still no generally agreed upon approach to prosodic modelling. In the former, they have been applied only occasionally, rather in basic research, but almost never within an existing end-to-end system. In this chapter, we discuss the reasons for this state of affairs and possible strategies to overcome the shortcomings of the use of prosodic modelling in automatic speech processing.

This chapter consists of three parts: the first part deals with the role of prosodic modelling in ASU, the second part concerns the role of prosodic modelling in TTS. These two parts are written from an inside perspective and focus on different aspects – simply because the use of models is considerably different in the two branches of speech technology discussed here. In the section on ASU, the argumentation is thus more general, dealing mainly with models as incarnations of theories, whereas in the section on speech synthesis, more details are given, dealing mainly with models as more or less concrete algorithmic formulations of theories. In the third part we present different possibilities for a closer co-operation of ASU and TTS; eventually this might lead to new types of prosodic models that are more adequate for automatic processing than the present ones.

In the title of this chapter, we speak of three different things: models, ASU, and TTS, and of one type of relationship: the common ground. Thus, first we have to know what prosodic/intonation models look like. For obvious reasons, we cannot give a detailed survey of the models that were developed during the last three decades. Instead we sketch common traits and principles that constitute models as such. One important characteristic is that a model is a considerable, sometimes even extreme, reduction of parametric values and, thereby, a mapping of these values onto a small number of units that can be compared with the classic distinctive features on the phone or phoneme level – all other properties may differ. The foundations of the model and, therefore, the philosophy behind it, can be physiological (Fujisaki, 1988), or perceptual ('t Hart et al., 1990), or linguistic (Silverman et al., 1992, "ToBI"), just to mention a few well-known models. ToBI (*Tones and Break Indices*) is by far the most well-known model. There are at least two reasons for this fact: first, it was the first model by way of which researchers from different disciplines attempted to find a common standard and common evaluation procedures; and second, it was developed for (American) English,

Anton Batliner, Bernd Möbius (2005): "Prosodic models, automatic speech understanding, and speech synthesis: Towards the common ground?". In William J. Barry and Wim A. van Dommelen (eds.), The Integration of Phonetic Knowledge in Speech Technology (Springer, Dordrecht), 21-44.

a fact which in itself enhances wide dissemination. The ToBI transcription system is a formalisation of the tone sequence theory of intonation (Pierrehumbert, 1980). It may be characterised as a broad phonemic system, consisting of *High* (H) and *Low* (L) tones and some few, additional diacritics. The phonetic details of fundamental frequency (F_0) contours in a given language have to be established in a second step. ToBI labels, in conjunction with F_0 generation rules, are also frequently used in the intonation components of TTS systems.

Out of the theoretically possible relationships between models, ASU, and TTS, we can imagine four different types:

- Type 1: ASU $\leftarrow$ model $\rightarrow$ TTS
- Type 2: ASU $\leftarrow$ model 1 | model 2 $\rightarrow$ TTS
- Type 3: ASU $\leftrightarrow$ TTS
- Type 4: ASU | TTS

Type 1 meets the ideas of generality put forward in most intonation theories that there is only one model that accounts for all possible applications. Type 2 is a weaker formulation that there have to be models especially tuned for different applications. With Type 3, a direct relationship between ASU and TTS, not mediated by any model, is imagined, and Type 4 (no relationship at all) might not be desirable but mirrors, in fact, the present situation quite closely.

There is a striking difference between ASU (many-to-one) and TTS (one-to-many): in ASU, many speakers/features/feature values have to be mapped onto few units (from parameters to categories), whereas in TTS, it is the other way round: one speaker/category has to be mapped onto many features/feature values (from categories to parameters). It has not been settled yet whether this is a one-way-trip or a round-trip – and by that, whether there is any common ground for these two fields at all, as far as prosody is concerned.

0.1 Caveat and further reading

This chapter is not intended to be an introduction into any of these three topics: models, ASU, and TTS. We hope, however, that it will be useful for experts in one of these fields who wonder why the state of affairs is the way it is. At the same time, we want to provide the readers with a sufficient degree of 'meta-knowledge' without presenting them all the basics. This chapter is thus not written as an in-depth treatise but rather as a *set of postulates* intended to provoke discussion.

Anton Batliner, Bernd Möbius (2005): "Prosodic models, automatic speech understanding, and speech synthesis: Towards the common ground?". In William J. Barry and Wim A. van Dommelen (eds.), The Integration of Phonetic Knowledge in Speech Technology (Springer, Dordrecht), 21-44.

To our knowledge, there is no up-to-date standard introductory textbook on intonation models. A comprehensive review of current intonation models is presented in (Ladd, 1996), albeit from the perspective of a proponent of the tone sequence approach. The language-specific use of some of these models is described in Hirst and Di Cristo's survey of intonation systems (Hirst and Di Cristo, 1998a). The computational analysis and modelling of prosody for the automatic processing of speech is the topic of (Sagisaka et al., 1997). A state-of-the-art account of prosody in ASU is given in (Batliner et al., 2001c), whereas the intonational concepts used in several models, and synthesis approaches based on these models, are dealt with in (Botinis et al., 2001).

1. Automatic speech understanding

For contemporary prosodic theory, subtle changes in meaning that are potentially triggered by prosody are interesting. These are, however, no good candidates to start with in ASU: they will be classified rather poorly because of the many intervening factors, because of sparse data, and because they can only be observed in the laboratory. Therefore, we should start with a clear prosodic marking; the marking of boundaries is probably the most important function of prosody and thus most useful for ASU.

Information retrieval dialogues have been the standard application within ASU for many years. Recently, less restricted dialogues, for instance in the context of the Verbmobil system[3], had to be processed where turns are, on the average, three times longer than in an information retrieval application (Nöth et al., 2000). Segmentation is thus more important in the relatively new field of automatic processing of rather free dialogues – a chance to prove the impact of prosody! The contribution of prosody is not equally evident in other applications.

In the last two decades, a growing body of work on intonation and prosody research in general and on intonational modelling in particular has been conducted. (Note that we use *prosody* for all phenomena above the segmental level, whereas *intonation* only deals with pitch/F_0.) Researchers on these topics agree that ASU would benefit from the integration of this work. However, only in the last few years has prosody really begun to find its way into ASU, most of the time within *offline*, i.e., *in vitro*, research. The only existing end-to-end system that really uses prosody is, to our knowledge, the Verbmobil system (Batliner et al., 2000).

This state of affairs might be traced back to the general difficulty of carrying over theoretical work into practice as well as the well-known

Anton Batliner, Bernd Möbius (2005): "Prosodic models, automatic speech understanding, and speech synthesis: Towards the common ground?". In William J. Barry and Wim A. van Dommelen (eds.), The Integration of Phonetic Knowledge in Speech Technology (Springer, Dordrecht), 21-44.

differences between the two cultures: on the one hand, humanities, on the other hand, engineering. In this section, we want to have a closer look at some of the most important factors that are responsible for this state of affairs, and by that, we want to make this general statement more concrete. First we discuss the shortcomings of current intonation models, as seen from an ASU perspective (section 1.1). Then, we will show what can be done to overcome these shortcomings by sketching our own *functional* prosodic model (section 1.2), and we will outline the common ground of prosodic models on the one hand and ASU on the other hand (section 1.3).

1.1 The reasons why (Occam's razor still matters)

If one speaks of suprasegmental models that meet the standards of a theory, one very often speaks only of *intonation models*, which almost always are *production models*. (Transcription, labelling, and annotation are more down to earth and their topic is thus broader.) Production models might be good for synthesis but not for recognition. Too much emphasis is put on intonation in particular, i.e., too much emphasis on *pitch* in comparison to *other prosodic* features, and too much emphasis on *prosody* in comparison to *other linguistic* features. This is, of course, conditioned by the general approach to constructing intonation models as *stand-alone models*, and by the unfortunate notion of *pitch accent*, which prevents a more realistic view where all relevant features – be they intonational, other prosodic or other linguistic features – are considered in the analysis on the same level.

There is too much emphasis on *theoretical concepts* and on the discussion of which one is better suited for the description of a special language or of languages in general. Consider the old debate pertaining to whether levels or movements, local events or global trends, are *the* 'correct' units of description: a speech recognizer does not care whether it is trained with levels (F_0 maximum, F_0 minimum) or with movements (F_0 range, F_0 slope) as long as the training database is large enough and the labels are annotated correctly. After all, what goes up must come down: it does not matter whether there is an H tone at 200 Hz and a following L tone at 100 Hz or whether there is a movement between 200 Hz and 100 Hz (Batliner et al., 2001a).

Very often it is stressed that one cannot do prosody research or apply prosody within ASU without a 'real' phonological level of description and modelling, and that speech technologists should pay attention to the work of phonologists (Ladd, 1997). We fully agree with the view

Anton Batliner, Bernd Möbius (2005): "Prosodic models, automatic speech understanding, and speech synthesis: Towards the common ground?". In William J. Barry and Wim A. van Dommelen (eds.), The Integration of Phonetic Knowledge in Speech Technology (Springer, Dordrecht), 21-44.

that phonological and prosodic *knowledge* should be used within ASU, but we fully disagree if it is about the direct use of intonation *models* in ASU. All these models introduce a phonological level of description that is intermediate between *(abstract) function* and *(concrete) phonetic form*: tone sequences, holistic contours, etc. It is our experience that one always gets better results if one can do without such an intermediate level, i.e., if one can establish a direct link between (syntactic/semantic) function and phonetic form. (Here, we speak 'simply' of classification performance, not of theoretical interest or adequacy.)

After all, if such a mapping can be done automatically, we can map *level A (phonetic form)* onto *level C (linguistic function)* without an intermediate *(phonological) level B*; with such a level, we have to map A onto B, and B onto C. If this can be done automatically, we do not need B any longer. Sometimes it will do no harm to provide level B, but often results will get worse. Phonological systems like the ToBI approach (Silverman et al., 1992) only introduce a *quantisation error*: the whole variety of F_0 values available in acoustics is reduced to a mere binary opposition L vs. H, and to some few additional, diacritic distinctions. This fact alone prevents tone levels (or any other *phonological prosodic* concepts such 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 large feature vector and to statistical classifiers to find the form to the function. To our knowledge, no approach exists that actually uses such phonological units for the recognition of prosodic events. Of course, there are many studies that describe *offline* classifications of such phonological prosodic concepts; this has to be distinguished from the successful *integration* in an existing end-to-end-system, as we have shown within the Verbmobil project (Batliner et al., 2000; Nöth et al., 2000).

Studies which compare the performance of intonational models for the automatic classification of prosodic events are rare; (Siepmann, 2001) assesses several models on the task of the classification of contrastive accents in German. He finds that classification performance is roughly a function of the number of predictor variables. It increases with the number of these predictor variables made available by a model. These findings fit nicely with our notion of quantisation described above. Evidently, a theoretically and phonologically 'adequate' description – in terms of a minimal inventory of units – on the one hand, and classification performance on the other hand, are simply two conflicting goals.

The difference between phonetic/prosodic knowledge and phonological concepts can be demonstrated with the following example: the prosodic 'default' feature that indicates questions in many languages is a

Anton Batliner, Bernd Möbius (2005): "Prosodic models, automatic speech understanding, and speech synthesis: Towards the common ground?". In William J. Barry and Wim A. van Dommelen (eds.), The Integration of Phonetic Knowledge in Speech Technology (Springer, Dordrecht), 21-44.

final rise (or 'high boundary tone'), even though, at least in English and German, a pronounced accent in a non-final position can disambiguate sentence mood as well (Studdert-Kennedy and Hadding, 1973; Batliner, 1989b). The same holds for Italian, where "[the] primary cue to interrogation in the Southern varieties is the pitch accent: L+H* in Bari Italian and L*+H in Palermo and Neapolitan, after which there is usually a final fall." (Grice et al., in press) This is a very interesting fact in itself, but it is of course not a special tone that is the primary cue but something that can be *described as* a special tone within a special intonational model. This is actually not a nicety but crucial for our argumentation. Thus we want to distinguish between *basic* knowledge about the facts one observes, and knowledge that is *transformed into* and *mediated by* a specific model. The units of such a model might provide a convenient way to make oneself understood. The problem is that, by using such a terminology, one tends to disregard those aspects that are not modelled by this concept; for instance, by using the terminology of a tone *level* model one disregards *movements*, and vice versa, and might end up with a mere *reification* of this concept.[4]

The classical phonological concept of the Prague school has been abandoned in contemporary intonation models, namely that phonemes – be they segmental or suprasegmental – should only be assumed if these units make a difference in meaning. This functional point of view has given way to more formal criteria such as economy of description. Thus, the decision on the descriptive units is not based on differences in meaning but on formal criteria, and only afterwards are functional differences sought that can be described with these formal units. In (Hirschberg and Pierrehumbert, 1986) for instance, the meaning of a tune, which is defined as a structure comprised of accents and tones, can be interpreted compositionally from the meanings of the individual accents and tones that the tune contains. It has been supposed that if phonological concepts could be motivated from theoretical reasons, then ASU should use them[5] – irrespective of whether they really make sense as units of ASU or not: this can only be determined empirically, not by theoretical considerations.

In conclusion, *Occam's razor* (i.e., the law of economy) should thus be followed here as well: *non sunt multiplicanda entia praeter necessitatem (entities are not to be multiplied beyond necessity)*; for *'entities'* read: levels of description or processing.

Anton Batliner, Bernd Möbius (2005): "Prosodic models, automatic speech understanding, and speech synthesis: Towards the common ground?". In William J. Barry and Wim A. van Dommelen (eds.), The Integration of Phonetic Knowledge in Speech Technology (Springer, Dordrecht), 21-44.

1.2 A functional prosodic model

In this section, we sketch an alternative model that puts emphasis on *function*, not on phonological *form* – actually, all other working approaches towards using prosodic information in ASU we know of are along these lines, cf. (Shriberg et al., 1998; Nöth et al., 2000) and the references given in these papers. The prosodic functions that are generally considered to be the most important ones on the linguistic level are the marking of boundaries, accents, and sentence mood; boundaries can delimit syntactic, semantic, or dialogue units. For these phenomena, the first step is the annotation of a large database. Annotation should be as detailed as possible, but more detailed classes can – if necessary – be mapped onto higher classes. We still do not know how many classes are most appropriate for the pertinent linguistic phenomena; it is, however, our experience that quite often, the higher linguistic modules can work fairly well with only two binary classes: present vs. not present.[6] The phonetic form is modelled directly with a large feature vector which uses all available information on (appropriately normalized) F_0, energy, and duration; other linguistic information pertaining to, for instance, part of speech classes is used as well. It is not a theoretical question but one of practical reasoning, availability, implementation, and recognition performance whether all this information is processed sequentially or in an integrated procedure. The model, classification results, and the use of prosodic knowledge in higher linguistic modules are described in (Batliner et al., 2000; Nöth et al., 2000; Batliner et al., 2001c).

1.3 Which common ground for ASU and prosodic models?

Mainstream ASU nowadays means statistical processing. For this approach, large databases and a standardization of different annotation concepts are needed. ToBI has been a step in the right direction but is still too much based on (one specific) phonology; it is not an *across models* but a *within model* approach; cf. the standardization efforts for dialogue act annotations described in (Klein, 1999). Only if they are based on a successful standardization, can the labels of different (intonation) models be used together in order to overcome the sparse data problem. The *primacy of phonology* has to give way to more practical considerations; models should take into account the requirements – and limitations – of speech processing modules. For instance, even if word recognition computes phone segment boundaries, these are often not available afterwards: the output is a word hypotheses graph with word boundaries only. An additional computation of phone segment bound-

Anton Batliner, Bernd Möbius (2005): "Prosodic models, automatic speech understanding, and speech synthesis: Towards the common ground?". In William J. Barry and Wim A. van Dommelen (eds.), The Integration of Phonetic Knowledge in Speech Technology (Springer, Dordrecht), 21-44.

aries would mean a considerable overhead.[7] Thus, intonation models that require an exact alignment with phones cannot be used. Therefore, we only used word boundaries in the final version of our prosody module in Verbmobil (Batliner et al., 2000) – without a decrease in performance!

The two cultures, viz. the humanities and engineering approaches, are still rather remote from each other. As in politics, one should begin with small steps, and with steps that pay off immediately. This means that subtle theoretical concepts are not well suited, but prosodic markers are, which are visible and stable enough to be classified reliably even in a realistic, *real life* setting. Thus it can be guaranteed that prosody really finds its way into ASU, because speech engineers can be convinced more easily that the integration of prosody indeed pays off. Later, it will be simply a matter of conquering or not: if more subtle differences can be modelled with prosodic means and classification performance is good enough, it will be no problem to incorporate them into ASU.

2. Speech synthesis

Prosodic models have been extensively applied in speech synthesis, simply because there is an obvious need for every TTS system to generate prosodic properties of speech if the synthesis output is to sound even remotely like human speech. However, the *necessity* of synthesizing prosody has as yet not resulted in a *generally agreed upon* approach to prosodic modelling. This statement holds for the assignment of segmental durations as well as for the generation of F_0 curves, the acoustic correlate of intonation contours.[8]

Intonation research is extremely diverse in terms of theories and models. On the *phonological* side, there is little consensus on what the basic elements are: tones, tunes, uni-directional motions, multi-directional gestures, etc. Modelling the *phonetics* of intonation is equally diverse, including interpolation between tonal targets (Pierrehumbert, 1981), superposition of underlying phrase and accent curves (Fujisaki, 1988), and concatenation of line segments ('t Hart et al., 1990).

Modelling *speech timing* for synthesis is less diverse. The important role of the syllable as a central processing unit in speech production and perception is widely accepted, but there is an ongoing controversy about how to best implement the pertinent effects in a model of speech timing; cf. the *syllabic timing model* proposed by Campbell (Campbell and Isard, 1991; Campbell, 1992), on the one hand, and the sums-of-products model of *segmental duration* proposed by van Santen (van Santen, 1993; van Santen, 1994), on the other hand.

Anton Batliner, Bernd Möbius (2005): "Prosodic models, automatic speech understanding, and speech synthesis: Towards the common ground?". In William J. Barry and Wim A. van Dommelen (eds.), The Integration of Phonetic Knowledge in Speech Technology (Springer, Dordrecht), 21-44.

In natural speech, tonal and temporal prosodic properties are coproduced, and there is an increasing body of evidence that tonal and temporal as well as spectral properties of speech are jointly planned by the speaker in a way that prosodic events can be optimally perceived by the listener (House, 1990; House, 1996; Dogil and Möbius, 2001). The conventional solution in speech synthesis systems, in contrast, embodies a unidirectional flow of information instead of synergy: first, the duration of speech sounds and syllables is assigned and then the F_0 contour of the utterance is computed.

One pivot in our discussion of prosodic models in automatic speech processing is the relevance of a phonological level of description.[9] This aspect is rather indistinct with respect to models of speech timing. The remainder of this section therefore concentrates on the use and usability of intonation models in speech synthesis.

2.1 Intonation synthesis: a two-stage process

Intonation synthesis can be viewed as a two-stage process, the first aimed at representing grammatical structures and referential relations on a *symbolic* level and the second at rendering *acoustic* signals that convey the structural and intentional properties of the message. Intonation models differ in terms of the interface that they provide between the higher linguistic components and the acoustic prosodic modules.

In many TTS systems sophisticated methods, such as syntactic parsing and part-of-speech tagging, are applied in the service of providing sufficient information to drive the acoustic prosodic components of the system, in particular, the intonation model. The intonationally relevant information comprises sentence mood as well as the location and strength of phrase boundaries and the location and type of accents.

Establishing the relation between syntactic structure and intonational features is among the most challenging subtasks of TTS conversion, and its imperfection contributes to the perceived lack of naturalness of synthesized speech. This shortcoming is unavoidable because TTS systems have to rely on the computation of linguistic structures from orthographic text, a level of representation that is notoriously poor at coding prosodic information in many languages.

The task of the acoustic-phonetic component of an intonation model in TTS is to compute continuous acoustic parameters (F_0/time pairs) from the symbolic representation of intonation. A large variety of models have been applied in TTS systems to perform this task, including implementations of the major frameworks of intonation theory: phonological models that represent the prosody of an utterance as a sequence

Anton Batliner, Bernd Möbius (2005): "Prosodic models, automatic speech understanding, and speech synthesis: Towards the common ground?". In William J. Barry and Wim A. van Dommelen (eds.), The Integration of Phonetic Knowledge in Speech Technology (Springer, Dordrecht), 21-44.

of abstract units (e.g., tones), viz. *tone sequence* models; and acoustic-phonetic models that interpret F_0 contours as complex patterns resulting from the superposition of several components, viz. *superposition* models. Besides these prevalent models, several other approaches have been taken, in particular *perception-based*, *functional*, and *acoustic stylization* models. For instance, the INTSINT system (Hirst and Di Cristo, 1998b) performs an automatic analysis and generation of F_0 curves by deriving a sequence of target points, specified in time and frequency, that represents a stylization of the F_0 curve.

All of these approaches rely on a combination of *data-driven* and *rule-based* methods: they all systematically explore natural speech databases, but vary in terms of what is derived from the analysis to drive intonation synthesis. For instance, there are two different approaches to acoustic stylization modeling. In one approach, continuous acoustic parameters are interpreted as directly representing intonation events (Taylor, 2000); in the other approach, intonation events are related to phonological entities such as tones or register via prototype building (Möhler, 1998). The abstract tonal representation provided by phonological intonation models is converted into F_0 contours by means of phonetic realization rules. The phonetic rules determine the F_0 values of the (H and L) targets, based on the metric prominence of syllables they are associated with, and on the F_0 values of the preceding tones. The phonetic rules also compute the temporal alignment of tones with accented syllables. Fujisaki's classical superpositional model computes the F_0 contour by additively superimposing phrase and accent curves and a speaker-specific F_0 reference value. Phrase and accent curves are generated from discrete commands, the parameter values of which are usually derived by generalization of values statistically estimated from a speech database. While this model can be characterized as primarily acoustically oriented (and physiologically motivated), it is possible to find phonological interpretations of its commands and parameters (Möbius, 1995).

2.2 Intonation synthesis and phonetic detail

F_0 contours as *acoustic realizations of accents* vary significantly depending on the structure (i.e., the segments and their durations) of the syllables they are associated with. For example, F_0 peak location is systematically later in syllables with sonorant codas than in those with obstruent codas (*pin* vs. *pit*), and also later in syllables with voiced obstruent onsets than with sonorant onsets (*bet* vs. *yet*). Moreover, the F_0 peak occurs significantly later in polysyllabic accent groups than in monosyllabic ones (van Santen and Möbius, 2000).

Anton Batliner, Bernd Möbius (2005): "Prosodic models, automatic speech understanding, and speech synthesis: Towards the common ground?". In William J. Barry and Wim A. van Dommelen (eds.), The Integration of Phonetic Knowledge in Speech Technology (Springer, Dordrecht), 21-44.

Intonation models need to generate as much of this phonetic detail as possible. The quantitative model of F_0 alignment proposed by (van Santen and Möbius, 2000), for instance, explains the diversity of surface shapes of F_0 contours by positing that accents belonging to the same phonological (and perceptual) class can be generated from a common *template* by applying a common set of *alignment parameters*. The templates are representatives of phonological intonation events of the type predicted by intonation theories, i.e. accents and boundaries.

Acoustic stylization models (Möhler, 1998; Taylor, 2000) also synthesize F_0 contours from a small number of *prototypical patterns*. They learn and predict phonetic details of F_0 movements from a set of features comprising segmental, prosodic and positional information. While the F_0 prototypes are defined as being phonetically distinct, they are also intended to be related to phonologically motivated intonation events.

2.3 What is the common ground for TTS and prosodic models?

In section 1 we have argued that the most appropriate type of intonation model for ASU would be one that provides a *functional* representation of the positions of accents and phrase boundaries; any intermediate phonological level only introduces a quantisation error. In the ToBI notation (Silverman et al., 1992) such a functional representation would consist only of the location of accents (the stars) and phrase boundaries (the percents). In the following we discuss to what extent, or whether at all, the conclusions drawn for the ASU domain are valid for the TTS domain too; in doing so we consider both the state of the art in intonation synthesis and the feasibility of alternative designs.

In state-of-the-art TTS systems, such as Festival (Black et al., 1999), Bell Labs (Sproat, 1998), AT&T (Syrdal et al., 2000), and others, the only symbolic prosodic information used – apart from sentence mood – is the *location of accents* and *boundaries*. This design can be characterized as the bare-bones minimum of prosodic modelling, because phrase structure and accentual structure are surface reflections of the underlying semantic and syntactic structure of the sentence, and at least a coarse representation of phrasing and accenting needs to be achieved by any self-respecting TTS system.

However, it has been demonstrated that models which use more detailed and more precise input information, for instance ToBI *accent type* labels in addition to accent location alone, can generate F_0 contours that are perceptually more acceptable than models which use accent location alone (Syrdal et al., 1998). The problem is that computing from

Anton Batliner, Bernd Möbius (2005): "Prosodic models, automatic speech understanding, and speech synthesis: Towards the common ground?". In William J. Barry and Wim A. van Dommelen (eds.), The Integration of Phonetic Knowledge in Speech Technology (Springer, Dordrecht), 21-44.

text such detailed intonational features as accent type is difficult and unreliable. It should therefore come as no surprise that even the very same research group that so convincingly demonstrated the importance of detailed input information, came up with the solution (the 'ToBI Lite' approach) of collapsing ToBI accent labels onto merely two categories and of mapping only edge tones marking major phrases onto just one category (Syrdal et al., 2000). Note, however, that strictly speaking, these results are an indication that a greater variation of accent types will result in a higher degree of acceptability; they are no proof that a ToBI-like accent representation is the best or the only possibility of modelling variation.

The degree of potential improvement to synthesized prosody can also be illustrated by manually marking up the text or by providing access to semantic and discourse representations (Prevost and Steedman, 1994). In practice and in existing end-to-end systems, however, the situation in intonation synthesis appears to be similar to the one described for the ASU domain. But it is still worth noting that relying for the most part only on accent and boundary location is not a *judicious design decision* made by speech synthesis researchers but one made by system developers *bowing to necessity*. It is evident that much more information than just the stars and the percents is needed to achieve the kind and degree of improvement to intonation synthesis that has been demonstrated in fragmentary research systems.

Can we do without a phonological representation of intonation in speech synthesis? Certain synthesis strategies beyond the classical TTS scenario offer more immediate interfaces between symbolic and acoustic representations of intonation. *Concept-to-speech* systems, in particular, provide a direct link between language generation and acoustic-prosodic components. A concept-to-speech system has access to the complete linguistic structure of the sentence that is being generated; the system knows *what* to say, and *how* to render it. Such a system may potentially incorporate semantic and discourse representations like those used in the experiment by (Prevost and Steedman, 1994).

Yet, even in concept-to-speech systems it is still necessary to specify the mapping from semantic to symbolic features and from symbolic to acoustic features. The issue of how much, and what kind of, information the language generation component should deliver to optimize the two mapping steps (i.e., the definition of a semantics/syntax-prosody interface) is a hot research topic. Once the two mapping steps are optimized, we may be able to advance one step further and get rid of the intermediate level (i.e., a phonological prosodic representation) just as hypothesized for ASU (see section 1.1).

Anton Batliner, Bernd Möbius (2005): "Prosodic models, automatic speech understanding, and speech synthesis: Towards the common ground?". In William J. Barry and Wim A. van Dommelen (eds.), The Integration of Phonetic Knowledge in Speech Technology (Springer, Dordrecht), 21-44.

The most drastic redesign of intonation synthesis would be to avoid synthesizing intonation in the first place. Consider the early *unit selection* synthesis approach implemented in the CHATR TTS system (Black and Taylor, 1994). Unit selection generates speech by concatenating speech segments of varying length (as short as half-phones and as long as entire utterances) that are extracted at runtime from a large speech database. CHATR follows the strategy of simply *resequencing* speech segments without performing any modifications by signal processing. The underlying assumption is that the listener will tolerate occasional spectral or prosodic mismatches in an utterance if the quality of the output speech in general approaches that of natural speech.

The unit selection algorithm attempts to minimize two types of cost, one for *unit distortion* and one for *continuity distortion*. The former is a measure of the distance of the candidate unit from the desired target, whereas the latter is a measure of the distance between two adjacent units at the concatenation point. Each target is specified by a feature vector that comprises positional, spectral, and prosodic features, and the values of these features for a given target are specified based on some kind of model. In the case of prosodic features, the desired F_0 contour is usually predicted by an intonation model. Thus, even in the most extreme version of corpus-based synthesis, the mapping from a target specification to acoustic-phonetic details of candidate units is mediated by a model that relies on a symbolic representation of intonation, which customarily amounts to a phonologically based or motivated intonation model.

A phonological approach is even advocated explicitly in an interesting recent approach to unit selection termed *phonological structure matching* (Taylor and Black, 1999), where phonological information, such as canonical pronunciation, positional factors and accentuation, is used for unit selection, instead of narrow phonetic transcriptions and absolute duration and F_0 values. The key idea in this approach is that most of the variability in the speech signal is predictable and that units selected from the appropriate context are likely to have the right specifications, including prosodic ones. This means that intonation contours generated by models may not be necessary anymore. But what will still be relevant is the knowledge about the factors and their respective quantitative effects on observed contours; this knowledge can be used to develop powerful unit selection criteria.

Anton Batliner, Bernd Möbius (2005): "Prosodic models, automatic speech understanding, and speech synthesis: Towards the common ground?". In William J. Barry and Wim A. van Dommelen (eds.), The Integration of Phonetic Knowledge in Speech Technology (Springer, Dordrecht), 21-44.

3. Which common ground for ASU and TTS – with or without prosodic models?

We have illustrated that the basic problems connected with the use of prosodic models in speech processing are similar for ASU and TTS. One of these problems is the lack of an appropriate annotation concept. We have argued that ToBI – while representing a step in the right direction – is too much based on one specific intonational phonology and does not generalize across models. We have further argued that in the ASU context, ToBI provides a special layer of representation that is both too abstract (i.e., too far from the signal to be useful as input to classifiers) and at the same time not abstract enough, with some of its notational units lacking a linguistic counterpart. A mirror image of this situation is evident in the context of TTS, where ToBI lacks the required granularity.

3.1 Shared models for ASU and TTS?

In our view, the most appropriate type of intonation model for ASU would be one that provides a functional representation of the positions of accents and phrase boundaries without any intermediate prosodic-phonological level. Actually, such a type of model is widely used in intonation synthesis, albeit with some intermediate prosodic-phonological representation. This apparent similarity between ASU and TTS requirements is brought about by very different motivations. In ASU a finer-grained level of description has not yet been shown to model reliably the linguistic function that it presumably corresponds to. In TTS, in contrast, more detailed input information is required to generate F_0 contours that are perceptually more acceptable than those based on accent and phrase boundary locations alone. While computing such features is extremely hard in a TTS framework, it may be accessible in different speech synthesis strategies such as concept-to-speech.

Recent advances in TTS can be partly attributed to the use of statistical methods for detecting relevant features in large databases, learning them, and modelling them. A standardized annotation concept would be an additional advantage. However, the prevalent annotation convention, viz. ToBI, misses the required granularity: it is too much confined within one type of intonation model; it is too elaborate and specific in terms of its descriptive inventory to lend itself as a generic interface to higher-level linguistic-prosodic analysis; at the same time it is far too abstract to facilitate a computation of the rich phonetic detail and precise alignment that F_0 contours are required to have in order to sound natural. Data-driven intonation models, on the other hand, can learn to synthesize these details. For the integration in a TTS system, a complete

Anton Batliner, Bernd Möbius (2005): "Prosodic models, automatic speech understanding, and speech synthesis: Towards the common ground?". In William J. Barry and Wim A. van Dommelen (eds.), The Integration of Phonetic Knowledge in Speech Technology (Springer, Dordrecht), 21-44.

intonation model needs to provide a mapping from categorical phonological elements to continuous acoustic parameters. Quantitative models such as those presented recently (Möhler, 1998; Taylor, 2000; van Santen and Möbius, 2000) offer feasible solutions to the F_0 generation task. However, it is not clear yet whether these two approaches can be integrated into existing TTS systems without any additional phonological representation.

We believe that no intonation model equally appropriate for both tasks, ASU and TTS, is currently available. The requirements are, for the time being and for some time to come, too different. They might converge in the future, giving rise to a unified solution to prosodic modelling, but we simply do not know when and whether this will be the case.

3.2 Multilinguality

One aspect that we have not discussed in this chapter yet is *multilinguality*. Both ASU and TTS have gone multilingual:

In the Verbmobil system (Batliner et al., 2000), prosodic information is computed for ASU for three languages, viz. German, English, and Japanese. The *multilingual prosody module* facilitates sharing of prosodic feature extraction and classification procedures, which are considered to be language independent. Note, however, that it is not clear yet whether or not the same set of features may be appropriate for typologically different languages, for instance tone and non-tone languages. Language specific data, such as duration normalization tables, are kept in separate structures and are loaded as needed. Similarly, separate classification parameters, such as different n-gram sizes, can be specified by means of configuration files (Batliner et al., 2000).

In remarkably the same spirit, the multilingual intonation component in the Bell Labs TTS system (Sproat, 1998) is used for a number of intonationally quite diverse languages, including American English, French, German, Italian, Japanese, Mexican Spanish, and Russian; this component implements the quantitative model by (van Santen and Möbius, 2000). One of the key assumptions of this model is that phonological accent classes can be mapped onto a corresponding number of distinct F_0 templates by means of alignment parameter matrices (see section 2.2). Language specific adjustment pertains to transformations of these parameter matrices, which can be handled offline and stored in configuration files. Again similar to the ASU prosody design presented above, one of the most intriguing research questions is to what extent the inventory of templates can be shared across languages: notice that Mandarin Chi-

Anton Batliner, Bernd Möbius (2005): "Prosodic models, automatic speech understanding, and speech synthesis: Towards the common ground?". In William J. Barry and Wim A. van Dommelen (eds.), The Integration of Phonetic Knowledge in Speech Technology (Springer, Dordrecht), 21-44.

nese is not currently handled by this multilingual intonation approach (van Santen et al., 1998).

3.3 No panacea: the database argument

Sufficiently large and discourse-rich, (prosodically) *annotated databases* are of course a desideratum (Botinis et al., 2001). They are necessary for a good and robust classification of prosodic events in ASU, and they are necessary for modelling variability in TTS. They are, however, definitely no panacea, especially not for the lack of relevance of intonation models like ToBI for ASU: first, an elaborate prosodic annotation is very *time-consuming* (Batliner et al., 1998) and therefore simply too expensive; this might not be a 'scientific' argument but is nonetheless a decisive obstacle. Second, it is too complicated and thus prone to *low inter-labeller correspondence*, cf. ToBI vs. 'ToBI Lite' (Syrdal et al., 2000). Third, it can be doubted that 'real-life' spontaneous speech is always *prosodically rich* to such an extent that special and/or rare functions are indicated by prosodic means; an extrapolation of constructed examples to spontaneous speech might turn out to be mere wishful thinking. Fourth, in our opinion, *more data* might always be better than less data; but on the other hand, with 'only' 90 minutes of annotated speech material for German, and less than half the amount of data for English, we obtained the following overall classification rates (two-class problems): German boundaries, 87%; German accents, 81%; English boundaries, 92%, English accents, 79% (Batliner et al., 2000). Given the fact that inter-labeller reliability has not been proven to be very high for such tasks, it might not be possible to surpass these results by a considerable extent, even with a much larger training database. The benefit of larger training databases might thus not be the possibility to obtain much better classification rates but the possibility to model variability much better. That means, in turn, that performance will not go down drastically if one has to deal with new tasks, new scenarios, or new applications. (As for the portability of speech recognizers to new tasks, cf. (Lamel et al., 2001)) And finally, and most importantly, more labels cannot be a remedy for the missing link to clear functions, cf. section 1.1.

3.4 A catalogue of shared tasks

A straightforward way for ASU and TTS to co-operate would be to exchange knowledge, concepts, rules, algorithms and special databases between colleagues and research sites. Such a sharing of methods and resources is already a reality in several subdomains of speech processing,

Anton Batliner, Bernd Möbius (2005): "Prosodic models, automatic speech understanding, and speech synthesis: Towards the common ground?". In William J. Barry and Wim A. van Dommelen (eds.), The Integration of Phonetic Knowledge in Speech Technology (Springer, Dordrecht), 21-44.

cf. efficient search algorithms (Viterbi search), signal representations (e.g., HMM), or the use of linguistic or phonetic information (language models, duration models). This kind of exchange and sharing, as well as joint future work, would be a Type 3 approach (cf. the Introduction), a direct link between ASU and TTS, not mediated by traditional phonological models.

It might be argued that the tasks for ASU and TTS differ because TTS normally focuses on a formal speaking style, whereas ASU has to deal with a more casual, informal style. In our opinion, this is not a categorical but only a gradual difference which might diminish in the future: in a more elaborate synthesis, some computer speech will surely be more casual in order to approximate human-human communication. On the other hand, if large-scale content extraction has to be performed automatically from, e.g., radio news, ASU has to deal with formal speech as well.

In the context of prosody, we would propose the following catalogue of shared tasks:

- Inventory of relevant linguistic prosodic functions: marking of accents, phrases, discourse structure, etc. This can be illustrated by the rules for accent assignment that have been developed independently within ASU (Batliner et al., 1999) and TTS (Hirschberg, 1993; Widera et al., 1997), to mention just a few.
- Inventory of relevant paralinguistic prosodic functions: emotions / user states, individual speaker traits, etc. (Batliner et al., 2001c).
- Inventory of structured prosodic features: these features pertain to linguistically relevant units of speech, for instance phonemes, syllables, words, phrases, etc. Structured prosodic features are derived from basic acoustic-prosodic features, such as F_0, energy values, etc. This typology of prosodic features is described in (Kießling, 1997; Nöth et al., 2000).
- Inventory of lexical prosodic features: word accent position, part-of-speech information, etc.
- Inventory of syntactic/semantic prosodic features: sentence mood, syntactic structure and boundaries, positional and counting factors, centres of information.
- Annotation system, oriented towards function (not form), motivated by practical (not phonological) considerations.
- Procedures for detecting, learning, and modelling of prosodic features from speech databases. In state-of-the-art TTS, prosodic

Anton Batliner, Bernd Möbius (2005): "Prosodic models, automatic speech understanding, and speech synthesis: Towards the common ground?". In William J. Barry and Wim A. van Dommelen (eds.), The Integration of Phonetic Knowledge in Speech Technology (Springer, Dordrecht), 21-44.

features are learned from single-speaker databases. It might be feasible to train models on multi-speaker corpora to obtain prototypes via clustering or averaging (Batliner and Nöth, 1989); the prototypes might each represent one possible (virtual) and plausible speaker, but they do not have to represent any particular speaker in the corpus.

- Integration of *all* prosodic parameters and features, not just F_0, in TTS, following the ASU approach and acknowledging the fact of co-production of prosodic features in natural speech by co-modelling them.

A common task for ASU and TTS is to learn the mapping from acoustics to categories. In ASU the direct mapping of acoustic features onto functions without any intermediate phonological level is standard. In TTS, such a direct mapping might be feasible as well, for both offline training and runtime synthesis; hopefully, this will be a research avenue for the near future.

4. Concluding remarks

Coming back to the title of this book, 'Integration of Phonetic Knowledge in Speech Technology', we want to refer to the distinction made in section 1.1 between basic knowledge on the one hand and transformed/mediated knowledge on the other hand. This is, of course, a gross distinction; there might be rather a continuum from pure basic, acoustic, knowledge (e.g., about concrete F_0 values) to transformed, very abstract knowledge. Phonetic knowledge is thus never purely basic but always transformed up to a certain degree. The decisive step is, however, when it comes to a considerable reduction of information in order to achieve a level of 'phonological adequacy', cf. the quantisation error mentioned in section 1.1. In our opinion, phonetic/prosodic knowledge that has not yet crossed this rubicon is of course necessary for speech technology. Actually, it has always been used even if speech engineers might not have been aware of this fact. As for transformed/mediated phonological knowledge, we are not that sure and opt rather for those kinds of co-operation between ASU and TTS as described in the catalogue of shared tasks in section 3.4.

By successively working through this catalogue, we might eventually end up with something that might be called a new type of prosodic model, capable of explaining and predicting variability, and which can connect phenomena and their processing by automatic means more directly than current intonation models do.

Anton Batliner, Bernd Möbius (2005): "Prosodic models, automatic speech understanding, and speech synthesis: Towards the common ground?". In William J. Barry and Wim A. van Dommelen (eds.), The Integration of Phonetic Knowledge in Speech Technology (Springer, Dordrecht), 21-44.

Acknowledgments

The authors wish to thank Gregor Möhler, Elmar Nöth, and Antje Schweitzer, as well as the two editors, for valuable comments and discussion on earlier versions of this chapter. We also acknowledge the helpful suggestions by two anonymous reviewers. This work was funded by the German Federal Ministry of Education, Science, Research and Technology (BMBF) in the framework of the SmartKom project under Grants 01IL905D and 01IL905K7. The responsibility for the content lies with the authors.

Anton Batliner, Bernd Möbius (2005): "Prosodic models, automatic speech understanding, and speech synthesis: Towards the common ground?". In William J. Barry and Wim A. van Dommelen (eds.), The Integration of Phonetic Knowledge in Speech Technology (Springer, Dordrecht), 21-44.

Notes

1. In our understanding, automatic speech recognition (ASR) comprises 'only' word recognition, which is a necessary prerequisite for automatic speech understanding (ASU). Thus, if we have to choose one of these two terms, we prefer ASU because it covers the whole story and not only some part of it. Moreover, 'understanding' is more directly connected with higher linguistic levels such as syntax, semantics, and pragmatics. If we consider the work on automatic processing of prosody conducted so far, it might be the case that the impact of prosody is much stronger for these higher levels, compared to the impact on word recognition.

2. Strictly speaking, TTS is the customary acronym for *text-to-speech*, but in the context of this chapter we have opted to use it for any kind of speech synthesis, disregarding the exact type of input representation (e.g., text, concept, or structured document), unless explicitly indicated otherwise.

3. The Verbmobil system was developed in a large-scale German research project focusing on automatic speech-to-speech translation in appointment scheduling dialogues (Wahlster, 2000).

4. In (Batliner, 1989a) we have discussed the problem of reification from a slightly different point of view. An evident analogy on the segmental level is the famous *rabid/rapid* distinction (Lisker, 1978): it might be possible for a strictly phonological approach to work with only one distinctive feature, whereas for automatic speech processing, this would be a rather suboptimal approach.

5. "Probably, it will be very difficult to detect [automatically] a boundary marker that takes the form of a declination reset. ... [If its identification] in the acoustic signal cannot take place until a close-copy stylization has first been made, and that is the present situation, one can imagine that its automatic detection will only become a possibility once the technique of automatic stylization has been sufficiently mastered." ('t Hart et al., 1990, page 182) That simply means to beg the question – there is ample evidence nowadays, that boundaries can be detected without the help of such phonological concepts as declination (Batliner et al., 1998; Nöth et al., 2000).

6. Of course, linguists would like to get information from prosody for more subtle distinctions; maybe such distinctions can be provided and used successfully in the future, but not with the present state of the art and, especially, of the databases available (sparse data problem).

7. It would of course be no problem in principle for a word recognition module to store computed segment boundaries. In distributed systems, however, if prosody has to use the output of some existing word recognition module, this would mean to rewrite the module accordingly – which could not be done in the Verbmobil system due to project-internal constraints. Instead, in the first phase of the project, phone segments were re-computed in the prosody module, which caused a significant overhead. Thus, in the second phase, we computed only word based prosodic features – without any reduction of recognition performance, cf. (Batliner et al., 2000).

8. Notice that TTS systems do not usually provide a prosodic model for the amplitude profile of the synthetic utterance.

9. We do not argue against any phonological level as such. If we consider the well-established phonological level for word recognition, then there is a clear relationship between distinctive form and function; such a clear relationship, however, has not yet been proven for the prosodic level. Note that a *prosodic* phonological level might still be relevant for language typology, second language learning, etc., even if it might be irrelevant for the automatic processing of speech.

Anton Batliner, Bernd Möbius (2005): "Prosodic models, automatic speech understanding, and speech synthesis: Towards the common ground?". In William J. Barry and Wim A. van Dommelen (eds.), The Integration of Phonetic Knowledge in Speech Technology (Springer, Dordrecht), 21-44.

References

Batliner, A. (1989a). Eine Frage ist eine Frage ist keine Frage. Perzeptionsexperimente zum Fragemodus im Deutschen. In Altmann, H., Batliner, A., and Oppenrieder, W., editors, *Zur Intonation von Modus und Fokus im Deutschen*, pages 87–109. Niemeyer, Tübingen.

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

Batliner, A., Buckow, A., Niemann, H., Nöth, E., and Warnke, V. (2000). The prosody module. In (Wahlster, 2000), pages 106–121.

Batliner, A., Buckow, J., Huber, R., Warnke, V., Nöth, E., and Niemann, H. (2001a). Boiling down prosody for the classification of boundaries and accents in German and English. In *Proceedings of the European Conference on Speech Communication and Technology (Aalborg, Denmark)*, volume 4, pages 2781–2784.

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

Batliner, A., Möbius, B., Möhler, G., Schweitzer, A., and Nöth, E. (2001b). Prosodic models, automatic speech understanding, and speech synthesis: towards the common ground. In *Proceedings of the European Conference on Speech Communication and Technology (Aalborg, Denmark)*, volume 4, pages 2285–2288.

Batliner, A. and Nöth, E. (1989). The prediction of focus. In *Proceedings of the European Conference on Speech Communication and Technology (Paris)*, pages 210–213.

Batliner, A., Nöth, E., Buckow, J., Huber, R., Warnke, V., and Niemann, H. (2001c). Whence and whither prosody in automatic speech understanding: A case study. In Bacchiani, M., Hirschberg, J., Litman, D., and Ostendorf, M., editors, *Proceedings of the Workshop on Prosody and Speech Recognition 2001 (Red Bank, NJ)*, pages 3–12.

Anton Batliner, Bernd Möbius (2005): "Prosodic models, automatic speech understanding, and speech synthesis: Towards the common ground?". In William J. Barry and Wim A. van Dommelen (eds.), The Integration of Phonetic Knowledge in Speech Technology (Springer, Dordrecht), 21-44.

Batliner, A., Nutt, M., Warnke, V., Nöth, E., Buckow, J., Huber, R., and Niemann, H. (1999). Automatic annotation and classification of phrase accents in spontaneous speech. In *Proceedings of the European Conference on Speech Communication and Technology (Budapest)*, volume 1, pages 519–522.

Black, A. W. and Taylor, P. (1994). CHATR: a generic speech synthesis system. In *Proceedings of the International Conference on Computational Linguistics (Kyoto, Japan)*, volume 2, pages 983–986.

Black, A. W., Taylor, P., and Caley, R. (1999). *The Festival speech synthesis system – System documentation.* CSTR Edinburgh. Edition 1.4, for Festival version 1.4.0. [http://www.cstr.ed.ac.uk/projects/festival/manualfestival_toc.html].

Botinis, A., Granström, B., and Möbius, B. (2001). Developments and paradigms in intonation research. *Speech Communication*, 33(4):263–296.

Campbell, W. N. (1992). Syllable-based segmental duration. In Bailly, G., Benoît, C., and Sawallis, T. R., editors, *Talking Machines: Theories, Models, and Designs*, pages 211–224. Elsevier, Amsterdam.

Campbell, W. N. and Isard, S. D. (1991). Segment durations in a syllable frame. *Journal of Phonetics*, 19:37–47.

Dogil, G. and Möbius, B. (2001). Towards a model of target oriented production of prosody. In *Proceedings of the European Conference on Speech Communication and Technology (Aalborg, Denmark)*, volume 1, pages 665–668.

Fujisaki, H. (1988). A note on the physiological and physical basis for the phrase and accent components in the voice fundamental frequency contour. In Fujimura, O., editor, *Vocal Physiology: Voice Production, Mechanisms and Functions*, pages 347–355. Raven, New York.

Grice, M., D'Imperio, M., Savino, M., and Avesani, C. (in press). Towards a strategy for labelling varieties of Italian. In Jun, S. A., editor, *Prosodic Typology: An Approach through Tones and Break Indices.* Oxford University Press, Oxford, UK.

Hirschberg, J. (1993). Pitch accent in context: Predicting intonational prominence from text. *Artificial Intelligence*, 63(1–2):305–340.

Hirschberg, J. and Pierrehumbert, J. (1986). The intonational structuring of discourse. In *Proceedings of the 24th Annual Meeting of the ACL (New York)*, pages 136–144.

Hirst, D. and Di Cristo, A., editors (1998a). *Intonation Systems – A Survey of Twenty Languages.* Cambridge University Press, Cambridge, UK.

Hirst, D. and Di Cristo, A. (1998b). A survey of intonation systems. In (Hirst and Di Cristo, 1998a), pages 1–44.

Anton Batliner, Bernd Möbius (2005): "Prosodic models, automatic speech understanding, and speech synthesis: Towards the common ground?". In William J. Barry and Wim A. van Dommelen (eds.), The Integration of Phonetic Knowledge in Speech Technology (Springer, Dordrecht), 21-44.

House, D. (1990). *Tonal Perception in Speech*. Lund University Press, Lund.

House, D. (1996). Differential perception of tonal contours through the syllable. In *Proceedings of the International Conference on Spoken Language Processing (Philadelphia, PA)*, volume 1, pages 2048–2051.

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

Klein, M. (1999). Standardization efforts on the level of dialogue act in the MATE project. In *Proceedings of the ACL Workshop "Towards Standards and Tools for Discourse Tagging" (University of Maryland)*, pages 35–41.

Ladd, D. R. (1996). *Intonational Phonology*. Cambridge University Press, Cambridge, UK.

Ladd, D. R. (1997). Introduction to part I. Naturalness and spontaneous speech. In (Sagisaka et al., 1997), pages 3–6.

Lamel, L., Lefevre, F., Gauvain, J.-L., and Adda, G. (2001). Portability issues for speech recognition technologies. In *Proceedings of the Human Language Technology Conference HLT-2001 (San Diego, CA)*, pages 9–16.

Lisker, L. (1978). Rapid vs. rabid: A catalogue of acoustic features that may cue the distinction. Haskins Laboratories: Status Report on Speech Research SR-55/56.

Möbius, B. (1995). Components of a quantitative model of German intonation. In *Proceedings of the 13th International Congress of Phonetic Sciences (Stockholm)*, volume 2, pages 108–115.

Möhler, G. (1998). Describing intonation with a parametric model. In *Proceedings of the International Conference on Spoken Language Processing (Sydney)*, volume 7, pages 2851–2854.

Nöth, E., Batliner, A., Kießling, A., Kompe, R., and Niemann, H. (2000). Verbmobil: The use of prosody in the linguistic components of a speech understanding system. *IEEE Transactions on Speech and Audio Processing*, 8:519–532.

Pierrehumbert, J. (1980). *The phonology and phonetics of English intonation*. PhD thesis, MIT, Cambridge, MA.

Pierrehumbert, J. (1981). Synthesizing intonation. *Journal of the Acoustical Society of America*, 70:985–995.

Prevost, S. and Steedman, M. (1994). Specifying intonation from context for speech synthesis. *Speech Communication*, 15(1–2):139–153.

Sagisaka, Y., Campbell, N., and Higuchi, N., editors (1997). *Computing prosody – Computational models for processing spontaneous speech*. Springer, New York.

Anton Batliner, Bernd Möbius (2005): "Prosodic models, automatic speech understanding, and speech synthesis: Towards the common ground?". In William J. Barry and Wim A. van Dommelen (eds.), The Integration of Phonetic Knowledge in Speech Technology (Springer, Dordrecht), 21-44.

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

Siepmann, R. (2001). Phonetische Intonationsmodelle und die Parametrisierung von kontrastiven Satzakzenten im Deutschen. *Forschungsberichte des Instituts für Phonetik und Sprachliche Kommunikation (München), FIPKM*, 38:3–111.

Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., and Hirschberg, J. (1992). ToBI: A standard for labeling English prosody. In *Proceedings of the International Conference on Spoken Language Processing (Banff, Alberta)*, volume 2, pages 867–870.

Sproat, R., editor (1998). *Multilingual Text-to-Speech Synthesis: The Bell Labs Approach*. Kluwer, Dordrecht.

Studdert-Kennedy, M. and Hadding, K. (1973). Auditory and linguistic processes in the perception of intonation contours. *Language and Speech*, 16:293–313.

Syrdal, A., Möhler, G., Dusterhoff, K., Conkie, A., and Black, A. (1998). Three methods of intonation modeling. In *Proceedings of the Third International Workshop on Speech Synthesis (Jenolan Caves, Australia)*, pages 305–310.

Syrdal, A. K., Wightman, C. W., Conkie, A., Stylianou, Y., Beutnagel, M., Schroeter, J., Strom, V., Lee, K.-S., and Makashay, M. J. (2000). Corpus-based techniques in the AT&T NextGen synthesis system. In *Proceedings of the International Conference on Spoken Language Processing (Beijing)*, volume 3, pages 410–413.

't Hart, J., Collier, R., and Cohen, A. (1990). *A Perceptual Study of Intonation – An Experimental-Phonetic Approach to Speech Melody*. Cambridge University Press, Cambridge, UK.

Taylor, P. and Black, A. W. (1999). Speech synthesis by phonological structure matching. In *Proceedings of the European Conference on Speech Communication and Technology (Budapest)*, volume 2, pages 623–626.

Taylor, P. A. (2000). Analysis and synthesis of intonation using the Tilt model. *Journal of the Acoustical Society of America*, 107(3):1697–1714.

van Santen, J. P. H. (1993). Exploring N-way tables with sums-of-products models. *Journal of Mathematical Psychology*, 37(3):327–371.

van Santen, J. P. H. (1994). Assignment of segmental duration in text-to-speech synthesis. *Computer Speech and Language*, 8:95–128.

Anton Batliner, Bernd Möbius (2005): "Prosodic models, automatic speech understanding, and speech synthesis: Towards the common ground?". In William J. Barry and Wim A. van Dommelen (eds.), The Integration of Phonetic Knowledge in Speech Technology (Springer, Dordrecht), 21-44.

van Santen, J. P. H. and Möbius, B. (2000). A quantitative model of F0 generation and alignment. In Botinis, A., editor, *Intonation – Analysis, Modelling and Technology*, pages 269–288. Kluwer, Dordrecht.

van Santen, J. P. H., Möbius, B., Venditti, J., and Shih, C. (1998). Description of the Bell Labs intonation system. In *Proceedings of the Third International Workshop on Speech Synthesis (Jenolan Caves, Australia)*, pages 293–298.

Wahlster, W., editor (2000). *Verbmobil: Foundations of Speech-to-Speech Translations*. Springer, Berlin.

Widera, C., Portele, T., and Wolters, M. (1997). Prediction of word prominence. In *Proceedings of the European Conference on Speech Communication and Technology (Rhodes, Greece)*, volume 2, pages 999–1002.

Anton Batliner, Bernd Möbius (2005): "Prosodic models, automatic speech understanding, and speech synthesis: Towards the common ground?". In William J. Barry and Wim A. van Dommelen (eds.), The Integration of Phonetic Knowledge in Speech Technology (Springer, Dordrecht), 21-44.