Prosodic models, automatic speech understanding, and speech synthesis: towards the common ground

Anton Batliner[1], *Bernd Möbius*[2], *Gregor Möhler*[2], *Antje Schweitzer*[2], *Elmar Nöth*[1]

[1] Chair for Pattern Recognition, University of Erlangen-Nuremberg, Germany
[2] Institute of Natural Language Processing, University of Stuttgart, Germany
batliner@informatik.uni-erlangen.de, moebius@ims.uni-stuttgart.de

Abstract

Automatic speech understanding and speech synthesis, two of the 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 paper we want to discuss the reasons for this state of affairs as well as possible strategies to overcome the shortcomings of the use of prosodic modelling in automatic speech processing.

1. Introduction

The application of prosodic models in automatic speech understanding (ASU) and speech synthesis is strikingly different: in the latter, they 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 paper, we want to discuss the reasons for this state of affairs and possible strategies to overcome the shortcomings of the use of prosodic modelling. The paper consists of two parts: in the first part we deal with the role of prosodic modelling in ASU, in the second part we deal with the role of prosodic modelling in speech synthesis. Due to space limitations, this cannot be carried out as an in-depth treatise but rather as a *set of postulates* intended to provoke discussion.

2. Automatic speech understanding

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/F0.) 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 [1]. This state of affairs might be traced back to the general difficulty of carrying over theoretical work into practice as well as to the well-known differences between the two cultures: on the one hand, humanities, on the other hand, engineering. In the following, 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 want to show the shortcomings of intonation models, seen from an ASU perspective. Then, we will show what can be done to overcome these shortcomings by sketching our own *functional* prosodic model, and we will outline the common ground of prosodic models on the one hand and ASU on the other hand.

2.1. The reasons why (Occams razor still matters)

For prosodic theory, subtle changes in meaning that probably are 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, because they can only be observed in 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 within the Verbmobil system, had to be processed where turns are on the average three times longer than in the information retrieval application [5]. 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 as evident in the other applications.

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—in our opinion—unhappy notion of *pitch accent*, which prevents a more realistic view where all relevant features—be it 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 which one can better be used for the description of a special language or of languages in general. Consider the old debate whether levels or movements, whether local events or global trends, are *the* 'correct' units of descriptions: a speech recognizer does not care whether it is trained with levels or with movements 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 it is an H* at 200 Hz and a following

L* at 100 Hz or whether there is a movement between 200 Hz and 100 Hz.

We fully agree with the view 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. 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, but often results will get worse. Phonological systems like the ToBI-approach [8] only introduce a *quantisation error*: the whole variety of F0 values available in acoustics is reduced to a mere binary opposition *Low* vs. *High*, and to some few additional, diacritic distinctions. This fact alone prevents tone levels (or any other *prosodic phonological* concepts such as, e.g., the one developed within the IPO approach) from being a meaningful step that automatic processing could 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 [1, 5].

The classical phonological concept of the Prague school has been abandoned in these models, viz. that phonemes—be it segmental or suprasegmental—should only be assumed if these units make a difference in meaning. Such a functional point of view gave way to more formal criteria such as, for instance, economy of description. Thus, it was not differences in meaning that decided upon the descriptive units but formal criteria, and only afterwards were functional differences sought that can be described with these formal units. In [3] 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 consists of. If phonological concepts could be motivated from theoretical reasons, it was supposed that ASU should use them, cf. [10], p. 182—irrespective of whether they really make sense as units of ASU or not: this can only be decided upon empirically, not by theoretical considerations.

In conclusion, *Occams razor* (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.

2.2. A functional prosodic model

In this section, we sketch an alternative model that puts emphasis on *function*, not on phonological *form* – actually, every other working approach towards using prosodic information in ASU we know of is along these lines, cf. [7, 5] 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 should—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. The phonetic form is modelled directly with a large feature vector which uses all available information on (appropiately normalized) F0, energy, and duration; other linguistic information on, 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 [1, 5].

2.3. The common ground

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. Only 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 normally not available afterwards: the output is a word hypotheses graph with word boundaries only. An additional computation of phone segment boundaries would mean a considerable overhead. Thus intonation models where an exact alignment with phones is necessary cannot be used. Therefore, we only use word boundaries in the new version of our prosody module in Verbmobil [1]—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 more easily be convinced that the integration of prosody indeed pays off. Later, it will be simply a matter of conquer 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.

3. Speech synthesis

Prosodic models have been extensively applied in speech synthesis, simply because there is an obvious need for every speech synthesis 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 F0 curves, the acoustic correlate of intonation contours. This section concentrates

on the use and usability of intonation models in speech synthesis. 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 [6], superposition of underlying phrase and accent curves [2], and concatenation of line segments [10].

Intonation synthesis can be viewed as a two-stage process, the first aiming 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. At the same time, different application scenarios for speech synthesis may require different interface designs. We will review the common ground between intonation models and the constraints imposed by different speech synthesis strategies.

3.1. Symbolic representation

In many text-to-speech (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 the 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. Other synthesis strategies offer more immediate interfaces between symbolic and acoustic representations of intonation. Concept-to-speech (CTS) systems, in particular, provide a direct link between language generation and acoustic-prosodic components. A CTS 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. Yet, 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 (in other words: the definition of a semantics-syntax-prosody interface) is a hot research topic.

3.2. F0 generation from symbolic input

The task of the acoustic-phonetic component of an intonation model in speech synthesis is to compute continuous acoustic parameters (F0/time pairs) from the symbolic representation of intonation. A large variety of models have been applied in speech synthesis 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 of abstract units (e.g., tones), viz. tone-sequence models; and acoustic-phonetic models that interpret F0 contours as complex patterns resulting from the superposition of several components, viz. superposition models. Besides these prevalent models at least three other approaches have been taken, viz. perception-based, functional, and acoustic stylization models.

All of these approaches rely on a combination of data-driven and rule-based methods: they all systematically explore natural speech databases, but they vary in terms of what is derived from the analysis to drive intonation synthesis. For instance, *acoustic stylization models* represent intonation events either by continuous acoustic parameters [11] or as events that are related to phonological entities such as tones or register [4]. The abstract tonal representation provided by *phonological intonation models* is converted into F0 contours by means of phonetic realization rules. The phonetic rules determine the F0 values of the (H and L) targets, based on the metric prominence of the syllables that they are associated with, and on the F0 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 F0 contour by additively superimposing phrase and accent curves and a speaker-specific F0 reference value. Phrase and accent curves are generated from discrete commands, the parameter values of which are usually derived by generalization of values that were statistically estimated from speech databases. 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.

In section 2 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 [8] such a functional representation would consist only of the location of accents (the stars) and phrase boundaries (the percents). In practice, the situation in intonation synthesis appears to be similar. In many TTS systems the only symbolic prosodic information used (apart from sentence mood) is the location of accents and boundaries. It has been demonstrated, however, that models which use more precise input information, such as ToBI *accent type* labels in addition to accent location, can generate F0 contours that are perceptually more acceptable than models which use accent location alone [9]. Phrasing and accenting are surface reflections of the underlying semantic and syntactic structure of the sentence. Computing detailed intonational features such as accent type from text is difficult and unreliable. Thus, relying only on accent location is not a judicious design decision but one bowing to necessity. The potential improvement to synthesized prosody can be illustrated by manually marking up the text, or by providing access to semantic and discourse representations. It is obvious that much more information than just the stars and the percents is needed to achieve this kind of improvement to intonation synthesis.

3.3. Intonation synthesis and phonetic detail

F0 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, F0 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 F0 peak occurs significantly later in polysyllabic accent groups than in monosyllabic ones [12]. Intonation models need to generate as much of this phonetic detail as possible. The quantitative model of F0 alignment proposed by van Santen and Möbius [12], for instance, explains the diversity of surface shapes of F0 contours by positing that accents belonging to the same phonological (and per-

ceptual) 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 (e.g., [4, 11]) also synthesize F0 contours from a small number of prototypical patterns. They learn, and predict, phonetic details of F0 movements from a set of features comprising segmental, prosodic and positional information. While the F0 prototypes are defined as being phonetically distinct, they are also intended to be related to phonological intonation events.

3.4. The common ground

Recent advances in speech synthesis may 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 F0 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 speech synthesis system, a complete intonation model needs to provide a mapping from categorical phonological elements to continuous acoustic parameters. Quantitative models such as those presented recently [4, 11, 12] offer feasible solutions to the F0 generation task, but their phonological foundations need to be further worked out.

4. Conclusion

We have illustrated that the basic problems connected with the use of prosodic models in speech processing are similar for ASU and speech synthesis. 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 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 speech synthesis, where ToBI lacks the required granularity.

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 phonological level—precisely the type of model that is widely used in intonation synthesis. 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 speech synthesis, in contrast, more detailed input information is required to generate F0 contours that are perceptually more acceptable than those based on accent and phrase boundary location 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.

We believe that no intonation model equally appropriate for both tasks, ASU and speech synthesis, 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.

Acknowledgment

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 contents lies with the authors.

5. References

[1] A. Batliner, A. Buckow, H. Niemann, E. Nöth, and V. Warnke. "The prosody module", In W. Wahlster, editor, *Verbmobil: Foundations of Speech-to-Speech Translations*, 106–121. Springer, New York, Berlin, 2000.

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

[3] J. Hirschberg and J. Pierrehumbert. "The intonational structuring of discourse", In *Proc. 24th Annual Meeting of the ACL (New York, NY)*, 136–144, 1986.

[4] G. Möhler and A. Conkie. "Parametric modeling of intonation using vector quantization", In *Proc. 3rd ESCA Workshop on Speech Synthesis (Jenolan Caves, Australia)*, 311–316, 1998.

[5] 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. ASSP*, 8:519–532, 2000.

[6] J. Pierrehumbert. "Synthesizing intonation", *J. Acoust. Soc. Am.*, 70:985–995, 1981.

[7] 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*, 41:439–487, 1998.

[8] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg. "ToBI: a standard for labeling English prosody", In *Proc. ICSLP–1992 (Banff, Alberta)*, 2:867–870, 1992.

[9] A. Syrdal, G. Möhler, K. Dusterhoff, A. Conkie, and A. Black. "Three methods of intonation modeling", In *Proc. 3rd ESCA Workshop on Speech Synthesis (Jenolan Caves, Australia)*, 305–310, 1998.

[10] J. 't Hart, R. Collier, and A. Cohen. *A Perceptual Study of Intonation*, Cambridge UP, Cambridge, MA, 1990.

[11] P. Taylor. "Analysis and synthesis of intonation using the Tilt model", *J. Acoust. Soc. Am.*, 107:1697–1714, 2000.

[12] J. P. H. van Santen and B. Möbius. "A quantitative model of F0 generation and alignment", In A. Botinis, editor, *Intonation—Analysis, Modelling and Technology*, 269–288. Kluwer, Dordrecht, 2000.