Prosodic models and speech recognition: towards the common ground.[1]

A. Batliner[$], E. Nöth[$], B. Möbius[*], and G. Möhler[*]

[$]University of Erlangen-Nuremberg, Chair for Pattern Recognition, Erlangen, F.R.G.
[*]University of Stuttgart, Institute of Natural Language Processing, Stuttgart, F.R.G.
batliner@informatik.uni-erlangen.de

ABSTRACT

In spite of the claim made by many researchers that prosody is a valuable source of knowledge in speech recognition in particular and in automatic speech understanding (ASU) in general, it has not been used up to now to a considerable extent. Partly, this might be due to the fact that its role is more important in more elaborated speech whereas until recently, the main emphasis was on dictation systems or on rather simple dialogue systems. In our opinion, a second, maybe even more important point is that mainstream prosodic models are not designed for use in ASU which means that they are often not well suited for this task. ASU needs a functional representation, i.e., a genuine phonological representation; units should only be modelled if they denote a clear-cut difference in linguistic meaning. In addition, the prosodic features that classification is based on should be *flat*, i.e., close to the surface and not too much influenced by theoretical considerations; further clustering should be left to the classifier. If we consider two of the most influential prosodic models from this point of view, the ToBI model and the IPO model, then both are too much *in between*: both introduce a special layer of representation, which, on the one hand, is not abstract and functional enough because quite often, a unit has no clear-cut functional linguistic counterpart; on the other hand, the units of description are too abstract, too far away from phonetic reality and from the signal itself and thus, they are not the best features a classifier should use. This statement holds for other prosodic models as well; it is actually corroborated by the recent use of prosodic information for automatic dialogue systems by different research groups. The common ground in all these models and in ASU as well is thus simply, in terms of the ToBI approach, the stars and the percents, i.e., a functional modelling of accent and boundary position - plus, of course, of some other phenomena such as questions vs. non-questions.

[1] This work was funded by the German Federal Ministry of Education, Science, Research and Technology (BMBF) in the framework of the Verbmobil project under Grant 01 IV 102 H/0 and in the framework of the SmartKom project under Grant 01IL905K7 and Grant 01IL905D. The responsibility for the contents lies with the authors.

1. Introduction

In this paper, we want to deal with one of the pivotal questions put forth in the description of this workshop: In the last two decades, a growing number of work on intonation/prosody in general and on intonational modelling in particular has been conducted.[2] Researchers on these topics agree that ASU[3] would benefit from the integration of this work. Whereas in speech synthesis, intonation models have been extensively applied, cf. [4], this does not hold for ASU: only in the last few years, prosody really began to find its way into ASU, most of the time, however, within *off-line* (i.e., *in vitro*, laboratory) research. The only existing end to end system where prosody is seriously used is, to our knowledge, the Verbmobil system, cf. [1], [5], [6]. This state of affairs might be traced back to the general difficulty to carry 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. In a second part, we will show what can be done to overcome these shortcomings by sketching our own *functional* prosodic model. In the final part, we will outline the common ground of intonation/prosodic models on the one hand and ASU on the other hand. In this paper, all this cannot be done as an in depth treatise but rather as a *set of postulates* intended to provoke discussion.

2. The reasons why (Occam's 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, prompted speech, etc. 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, cf. [6]. This shows that segmentation is more important in the relatively new field of automatic

[2] We use *prosody* for all phenoma above the segmental level, whereas *intonation* only deals with pitch/F0.

[3] *Speech recognition* deals mostly with phones and words, *ASU* covers higher linguistic levels as well; *prosody* is somewhat in-between or better, right across these levels but most useful for the higher linguistic ones.

processing of rather free dialogues; the contribution of prosody is not as evident in the other applications.

The title of this workshop *Intonation Modelling and Prosodic Transcription* illustrates the situation in a nice way: 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 are 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 the *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 be better used for the description of a specific language or of languages in general. Consider the old debates 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 correctly annotated. 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.

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, cf. [3]. We fully agree with this view if it is about phonological and prosodic *knowledge*, 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 which 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.[4] After all, if such a mapping can be done automatically, this means that 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 *B* will do no harm, but often, results will get worse. To put it bluntly, phonological systems like the ToBI approach, cf. [8], only introduce

[4] Here, we speak of classification performance, not of theoretical interest or adequacy.

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 like 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 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 *off-line* classifications of such phonological prosodic concepts in the laboratory; this has to be distinguished from the successful *integration* into an existing end to end system, as we have shown within the Verbmobil Project, cf. [1], [5], [6].

The classical phonological concept of the Prague school has been abandoned in these models that phonemes - be it segmental or suprasegmental - should only be assumed if these units make a difference in meaning. Such a rather functional point of view gave way to rather 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 functional differences that can be described with these formal units were sought. In [2] 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 automatic speech processing should use them, cf. [9], p. 182 - irrespective of whether they really make sense as units of ASU or not; this can only be decided empirically, not by theoretical considerations.

In conclusion, *Occam's razor* (law of economy) should 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.

3. A functional prosodic model

In this section, we want to sketch an alternative model that puts emphasis on function, not on phonological form.[5] 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

[5] Actually, every other working approach towards using prosodic information in ASU we know of is along these lines, cf. [6], [7], and the references given in these papers.

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 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], [6].

4. The common ground

Mainstream ASU nowadays means statistical processing. For this approach, large databases are needed. For that, standardization of different annotation concepts is a necessary step. 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 models* 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 consideration: models should take into account the requirements and limitations of speech processing modules. For instance, even if word recognition normally computes phone segment boundaries, these are not available afterwards: output is a word hypotheses graph with word boundaries only.[7] An additional computation of phone segment boundaries would mean a considerable overhead. Therefore, we only use word boundaries in the new version of our prosody module in Verbmobil, cf. [1], without any drop in performance!

The two cultures 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 that 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.

---

[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] This means that intonation models where an exact alignment of prosodic event with phones is necessary cannot be used.

REFERENCES

[1] Batliner, A., Buckow, J., Niemann, H., Nöth, E., Warnke, V. (2000). The Prosody Module. In: W. Wahlster (Ed.), *Verbmobil: Foundations of Speech-to-Speech Translations*. Springer, New York, Berlin, pp. 106 - 121.

[2] Hirschberg, J., Pierrehumbert, J. (1986): The intonational structuring of discourse. *Proceedings of the 24th Annual Meeting of the ACL*. New York, pp. 136-144.

[3] Ladd, D.R. (1996): Introduction to Part I. Naturalness and Spontaneous Speech. In: Y. Sagisaka, N. Campell, N. Higuchi (Ed.), Computing Prosody. Approaches to a Computational Analysis and Modelling of the Prosody of Spontaneous Speech, Springer, New York, Berlin, 3-6.

[4] Möbius, B., Möhler, G., Schweitzer, A., Batliner, A., Nöth, E. (2000): Prosodic models and speech synthesis: towards the common ground. This volume.

[5] Nöth, E., Batliner, A., Warnke, V., Haas, J., Boros, M., Buckow, J., Huber, R., Gallwitz, F., Nutt, M., Niemann, H. (1999): On the Use of Prosody in Automatic Dialogue Understanding. *Proceedings of the ESCA Workshop on Dialogue and Prosody*, Eindhoven. pp. 25-34 (to appear in *Speech Communication*).

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

[7] Shriberg, E., Bates, R., Taylor, P., Stolcke, A., Jurafsky, D., Ries, K., Cocarro, N., Martin, R., Meteer, M., Van Ess-Dykema, C. (1998): Can Prosody Aid the Automatic Classification of Dialog Acts in Conversational Speech? *Language and Speech 41*, pp. 439-487.

[8] Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., Hirschberg, J. (1992): TOBI: A Standard for Labelling English Prosody. *Proceedings of* ICSLP'92, Banff, pp. 867-870.

[9] 't Hart J., Collier, R., Cohen, A. (1990): *A Perceptual Study of Intonation - An Experimental-phonetic Approach to Speech Melody.* Cambridge: Cambridge University Press.