Prosodic models and speech synthesis: towards the common ground*

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

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

ABSTRACT

Prosodic models have been extensively applied in speech synthesis. However, the necessity of synthesizing prosody has as yet not resulted in a generally agreed upon approach to prosodic modeling. This statement holds for the assignment of segmental durations as well as for generating F0 curves, the acoustic correlate of intonation contours. This paper concentrates on the use and usability of intonation models in speech synthesis. Intonation synthesis can be viewed as a two-stage process, and intonation models differ in terms of the interface they provide between the higher linguistic components and the acoustic prosodic modules. We will review the common ground between intonation models and the constraints imposed by different speech synthesis strategies.

1. Introduction

Prosodic models have been extensively applied in speech synthesis. The situation in this particular branch of applied speech research is thus strikingly different from the one found in automatic speech recognition (ASR) and understanding. In the latter area the use of prosodic models has been rather occasional, for reasons that we have discussed elsewhere [1].

Obviously, there is a 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 modeling. This statement holds for the assignment of segmental

*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.

durations as well as for the generation of F0 curves, the acoustic correlate of intonation contours. This paper concentrates on the use and usability of intonation models in speech synthesis.

Which intonation models have been applied to 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 should be: tones, tunes, uni-directional motions, multi-directional gestures, etc. Modeling 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 [11]. All these major frameworks, as well as a number of more idiosyncratic models, have been implemented in speech synthesis systems.

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.

2. Symbolic representation

In many text-to-speech (TTS) systems the computation of phonological features of intonation is generalized to symbolic prosodic processing, which handles both tonal and temporal properties of speech, and further integrated into the linguistic text analysis component (cf. [9]). Here sophisticated methods developed in computational linguistics, such as syntactic parsing and part-of-speech tagging, are mainly applied in the service of providing sufficient information to drive the acoustic prosodic components of the system, in particular the intonation model but also the duration model.

The intonationally relevant information comprises the sentence mode 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.

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 question 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. 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.

It has become customary to distinguish two major types of intonation models: phonological models that represent the prosody of an utterance as a sequence of abstract units (e.g., tones), i.e. tone-sequence models; and acoustic-phonetic models that interpret F0 contours as complex patterns resulting from the superposition of several components, i.e. 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 [12] 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 applying a set 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 F0 values of tones are computed strictly from left to right, depending exclusively upon the already processed tone sequence and not taking into account any subsequent 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 from 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; moreover, the compatibility

of a Fujisaki-style model with key assumptions of the tone sequence model has been demonstrated [3].

In our paper on the use of prosodic models in speech recognition [1] we have argued that the most appropriate type of intonation model for ASR would be one that provides a functional representation of the positions of accents and phrase boundaries; any intermediate phonological level such as provided by the ToBI annotation convention [8] only introduces a quantization error. In the ToBI notation 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 (apart from sentence mode) used 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 [10].

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 (e.g., [7]). It is obvious that much more information than just the stars and the accents is needed to achieve this kind of improvement to intonation synthesis.

4. 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 [13].

Intonation models need to generate as much of this phonetic detail as possible, and there are several approaches to achieve this task. For instance, the quantitative model of F0 alignment proposed by van Santen and Möbius [13] predicts the temporal alignment not only of the peak of the accent curve but of a series of characteristic anchor points along the accent curve. In this model, the shape of the accent curve depends on the syllabic and segmental composition of the accent group and on the durations of its subcomponents (stressed syllable onset, stressed syllable rhyme, remainder of the accent group). The resulting F0 curve will have the desired complex shape and

precise temporal alignment with the segmental material. The model explains the diversity of surface shapes of F0 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 (e.g., [5]) 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.

5. The common ground

Reflecting the situation found in speech recognition, 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 modeling them. A standardized annotation concept would be an additional advantage, and ToBI has certainly been a step in the right direction. However, in the context of ASR we have argued [1] that ToBI is too much based on one specific intonational phonology and does not generalize across models. We have further argued that it 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 missing a linguistic counterpart.

A mirror image of this situation is evident in the context of speech synthesis. Here again, 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, while at the same time being far too abstract to allow 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 can learn to synthesize these details.

For an 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 [5, 12, 13] offer feasible solutions to the F0 generation task, but their phonological foundations need to be further worked out.

REFERENCES

[1] Batliner, A., E. Nöth, B. Möbius, G. Möhler: Prosodic models and speech recognition: towards the common ground. *Proc. Prosody-2000 (Krakow, Poland).* This volume.

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

[3] Möbius, B. (1995): Components of a quantitative model of German intonation. *Proc. 13th International Congress of Phonetic Sciences* (Stockholm, Sweden), vol. 2, pp. 108-115.

[4] Möhler, G. (1998): Theoriebasierte Modellierung der deutschen Intonation für die Sprachsynthese. PhD thesis, University of Stuttgart.

[5] Möhler, G., A. Conkie (1998): Parametric modeling of intonation using vector quantization. *Proc. 3rd ESCA Workshop on Speech Synthesis* (Jenolan Caves, Australia), pp. 311-316.

[6] Pierrehumbert, J. (1980): The phonology and phonetics of English intonation. PhD thesis (MIT, Cambridge, MA).

[7] Prevost, S., M. Steedman (1994). Specifying intonation from context for speech synthesis. *Speech Communication* **15**, pp. 139-153. [http://www.fxpal.xerox.com/people/prevost/spoken.htm]

[8] Silverman, K., M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, J. Hirschberg (1992): ToBI: A standard for labeling English prosody. *Proc. International Conference on Spoken Language Processing* (Banff, Alberta), vol. 2, pp. 867-870.

[9] Sproat, R. (Ed.) (1998): *Multilingual Text-to-Speech Synthesis—The Bell Labs Approach* (Kluwer, Dordrecht).

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

[11] 't Hart, J., R. Collier, A. Cohen (1990): *A Perceptual Study of Intonation* (Cambridge University Press, Cambridge).

[12] Taylor, P. (2000): Analysis and synthesis of intonation using the Tilt model. *J. Acoustical Society of America* **107** (3), pp. 1697-1714.

[13] van Santen, J., B. Möbius (2000): A quantitative model of F_0 generation and alignment. In: A. Botinis (Ed.), *Intonation: Analysis, Modelling and Technology* (Kluwer, Dordrecht), pp. 269-290.