Speech Prosody 2004
Nara, Japan
March 23-26, 2004

Automatic Generation of Prosody: Comparing Two Superpositional Systems

S. Raidt [1,3], G. Bailly[1], B. Holm[1], H. Mixdorff.[2]

(1) Institut de la Communication Parlée Grenoble, France
(2) TFH Berlin University of Applied Sciences Berlin, Germany
(3) Technische Universität Dresden, Germany

{raidt; bailly; holm}@icp.inpg.fr; mixdorff@tfh-berlin.de

Abstract

We face many options when designing a system that automatically generates prosody from linguistic and paralinguistic information. The literature provides several candidate phonetic models, phonological models and mapping tools to actually implement the system. We detail here some dimensions along which these models have to be compared. We show also that systems employing quite similar phonetic models can still have radically different approaches. We present results of a first evaluation comparing two systems using a superpositional model of melody on a common multilingual prosodic database of spoken math formulae. We conclude that prosodic models and intonation theories could certainly benefit from well-defined tasks and fair benchmarks.

10.21437/SpeechProsody.2004-95

1. Introduction

It is a commonly accepted view that prosody crucially shapes the speech signal in order to ease the decoding of linguistic and paralinguistic information by the listener.

The present study compares two automatic prosody generation systems applied to a bilingual corpus. Our aim is not to designate a "winner" but to promote inter-system comparison on common data as valuable aid to shed light on strengths and weaknesses of either system. The systems used here have to face different challenges: the SFC is applied to German (its target language being originally French); the IGM has to deal both with a new type of document (spoken mathematical formulae) and a new language (French). The reader should have in mind that the challenge for the latter demands more invasive adaptation of the system and results are to be considered as preliminary.

After introducing the framework in which we apprehend prosody generation systems, the two systems are briefly described (§3). Objective and subjective evaluation (§4) will enable us to detail some properties of the respective systems (§5).

2. Different system aspects

Automatic generation systems differ along many dimensions: we discuss below some aspects of these.

2.1. Phonetic models

The two systems compared here superpose f0 contours. They differ in the way these contours are parameterized both in terms of *contour shapes* and *segmental anchoring*.

Contour shapes

One popular phonetic model of intonation is Fujisaki's [8, 9]. It describes f0-contours as superposition of two different components to a constant F_b(subject's baseline), representing the two degrees of freedom in movement of the circothyroid muscle.

Other proposals have been made that also consist in superposing predefined shapes. Thorsen [28] for Danish, Gårding [10] for Swedish as well as Aubergé [1-3] for French consider f0 as the superposition of declination lines or more complex constructs. These constructs have the particularity of having embedded scopes thus reflecting a generally embedded syntactic structure.

Few attempts have been reported to learn automatically the actual shapes of embedded contours. Holm et al. [15, 16] propose an analysis-by-synthesis method for decomposing an f0-curve under high-level constraints (§3.2).

Segmental anchoring

Extracted contours are variously connected to the segmental chain. Some phonetic models treat directly the f0 curve without any *a priori* concern with the segments and extract targets (e.g. MOMEL [13]) or commands (e.g. the Fujisaki's model). Others characterize contours (e.g. Traber et al.: 8 f0 values per syllable) or parts of the contours (e.g. the Tilt model [26]) in relation with the syllables.

Phasing contours

When the phonetic model for melody does not *a priori* characterize contours with reference to segmental events such as vocalic/syllabic onsets, the generation model has to take in charge the phasing between the targets/commands and such events *a posteriori*. Mixdorff et al. (§3.1) consider the timing of commands in relationship to the on-/offset of the syllable.

Sampling contours

Other phonetic models characterize prosody of utterances by a constant number of cues per syllable. This strategy is particularly suited for a use of tools for interpolating data (multi-linear or sum-of-products models, decision trees, HMMs, NNs) that require a constant number of input/output parameters. For example, de Tournemire [7] has shown that three f0 values per syllable are sufficient for perceptual equality of the contours. Most of the time the sampling is performed at given percentage of the vocalic nucleus duration or of parts of the syllable [29]. More sophisticated models can be investigated such as the prediction of the relative or absolute phasing of these intra-syllabic targets [30, 31].

2.2. Phonological models

Phonological models should provide the information that determines what (para)linguistic information should be encoded into prosodic structure, what are the constituents of the prosodic structure and they should give information about what are the characteristics of the segmental carrier of this prosodic information.

Projecting phonological description

The basic element of most phonological models used in automatic prosody generation is the accented syllable/mora. The phonological model is thus entirely responsible for tagging each syllable as unaccented vs. accented and generating the appropriate accent type. It often delivers as input to the generation model additional information such as the number of phonemes in the current syllable, its position in the word and in the accentual phrase etc. The accent type may result of the calculation of a deep phonological structure where accentual

phrases are embedded into a prosodic structure, but the final result is often a series of information computed for each syllable of the utterance. This is the strategy followed by the IGM (§3.1). Such a linear sequential characterization is quite suitable for automatic learning algorithms (§2.3). Systems operating by selection and concatenation of signals [27] or prosodic contours [23] also use such a sequential phonological input.

Keeping non-projective phonological structures

Prosody may encode highly nested syntactic relations [11, 12]. This is certainly true for the particular dialog acts – spoken mathematical equations – studied here.

One way to keep trace of these highly structured phonological structures into the phonological input is to convert them into features and numbers: syllabic information may be augmented by accent and boundary types or break indexes that take into account complex hierarchy.

Another way to encode embedded structures is to rely on a phonetic model whose components cue almost one-to-one the units of the prosodic structure. This is the strategy followed by the SFC (§3.2) that superposes as many contours as there are phonological units.

2.3. Generation models

Generation models have in charge the mapping between phonological descriptions and input parameters of phonetic models. They are supposed to generate the variants of prototypical prosodic contours associated with phonological categories if any, to learn the necessary coarticulation effects between adjacent phonological units and interpolate between realizations of phonological configurations to actually generate unseen ones. This latter property of the generation models is of particular interest: when the set of phonological descriptors – features and cues – of syllables is reasonably large, the learning scheme is quickly confronted to data sparsity resulting from combinatory explosion.

If current generation models use automatic learning techniques to map phonological input to observed multiparametric contours, their structure may vary, i.e. in the case of this study: from full connectivity with one big NN in charge of the entire mapping (IGM) to partial connectivity where the interaction of several small NNs' output is left to the phonetic model.

3. Two prosody generation systems

3.1. The IGM

The IGM (Integrated Model) was developed by Mixdorff et al. [19, 20]. They use the Fujisaki model as phonetic model of f0 and a recurrent neural network for learning the mapping between linguistic features and prosodic parameters.

The phonetic model

Once Fujisaki's commands have been determined semi-automatically, they are synchronized with the syllabic string: each command is associated with a syllable and the alignment of the command with respect to the syllable is computed. Each accented syllable is characterized by three accent parameters (amplitude, delay between the onsets of syllable and accent, delay between their offsets) and each phrase-initial syllable by two phrase parameters (amplitude, delay between the onset of syllable and phrase command). In an attempt to preserve the coherence between the different components of prosody, the IGM integrates the prediction of syllable and pause durations as well as their intensity.

The phonological model

In order to find a set of relevant parameters for the phonological description of the underlying text, Mixdorff [19] investigated the correlation of possible parameters with the description of the prosodic characteristics of the acoustic realization and their proportionate influence. He retained 20 parameters concerning the current syllable and four relative to its context, i.e. derived from the parameters having been predicted for the preceding syllable. At the syllable level, these parameters convey information concerning the nature of the phones included in the components of the syllable (syllable, onset, rhyme). At the word level, they forward information about the composition of the word, containing the current syllable, its accentuation and the part of speech. At the phrase level or higher, they convey information about neighboring boundaries and the composition of the unit.

As the IGM had never been applied to mathematical formulae before, the phonological parameters could not be employed without change. Some of the original parameters could be omitted without substitution as the information they mean to convey does not apply to isolated mathematical formulae. A major task was to find an equivalent for the parameter "break index" (BI) and the thereof depending ones as e.g. number of syllables in preceding phrase. It seemed necessary to introduce new parameters because BI is situated between words, whereas the equivalent within a mathematical formula is represented by a word itself (mathematical operator), and therefore demands a decision of placement before or after the word. As no systematic placement could be determined [6], new parameters have been introduced providing information about the mathematical operator and the number of syllables of its operands. This avoids a hard decision for the placement and scaling of the BIs – leaving this task implicitly to the generation model. These modifications resulted in 14 input parameters. We will refer to this modified version of the IGM as "igm".

Training

The mapping between phonological and phonetic parameters is operated by a NN. Mixdorff et al. use a fully-connected NN with two hidden layers (18,12) with log and tan-hyperbolic transfer functions. Four of the output parameters are usually fed back to the input. In the present study, the NN was reduced to one hidden layer of 20 neurons for German using logistic transfer functions only. For French, an enlargement of the NN to 30 hidden neurons seemed to be required. The parameter "intensity" has been omitted – measured values have been used instead.

3.2. The SFC

The SFC (Superposition of Functional Contours) has been developed at the ICP. Its theoretical basis has been settled by Aubergé [1, 3] and assessed by numerous researchers [5, 16, 21].

The phonetic model

The SFC assumes that prosody can be described as superposition of global multiparametric contours anchored to the segmental chain. The parameters implemented for the time being are three f0 values and one lengthening factor per syllable.

The phonological model

The superposing contours are supposed to encode directly (para)linguistic functions. The latter may operate at different scales and on different units in parallel. The size of contours correspond to the scope of functions.

In the current implementation of the model several functions are defined: we consider contours encoding modalities and prosodic attitudes at the utterance level [21] while contours encoding the syntactic structure are indexed by dependency markers [4]. For maths we only distinguish two markers connecting an operator to its left or right operands.

Training

Since the SFC uses units of the same scope both for the phonological and the phonetic model, the relation of both can be

mapped directly. Contour generators (implemented as NN) – one for each (para)linguistic function – are used to produce mulitparametric contours. They are fed with information about the scope of the function and the position of each syllable within the scope. Every generator creates thus a family of contours which all implement a certain (para)linguistic function and whose variation depending on contour-scope may be apprehended as Prosodic Movement Expansion Model [21].

As the shape of the contours is not restricted by the phonetic model, it has to be determined during the training of the system. The problem of recovering individual contributions of the contour generators from their sum is ill-posed. These individual contributions and thus learning patterns for the generators emerge from an analysis-by-synthesis loop [16].

4. Comparative evaluation

4.1. The corpus

A corpus of 134 algebraic formulae has been constructed automatically varying the depth of their syntactic structure and the size of constituents. The oralisation of this type of "sentences" has proven to be fruitful for prosody research due to its high structuring demands [17] and it offers the additional advantage to allow the use of the same (textual) corpus in different languages. The corpus was registered in French and German by two native speakers instructed not to use lexical structure markers (e.g. "open parentheses") but to make use of prosody instead.

30 utterances are chosen at random to be reserved as test corpus. The remaining 104 formulae provide the training data.

4.2. Objective comparison

For objective comparison of the two models different errors were computed, respecting the fact that both of them generate melody and rhythm. In addition information about the generation of pauses is given, as this crucially influences the perception of partition and rhythm of speech. The values have been computed as well for the training as for the test corpus which confirm the respective results (table 1).

For both languages the results concerning fundamental frequency are better for the SFC, whereas it is more obvious for the comparison of the French synthesis. In contrast the IGM achieves better results for syllabic durations.

Table 1: *Prediction errors for the two systems. Lines 2-4 are RMSE; lines 5-7 pause counts*

	German				French			
	Training		Test		Training		Test	
	igm	SFC	igm	SFC	Igm	SFC	igm	SFC
f0 (correlation)	0.63	0.75	0.67	0.69	0.50	0.88	0.36	0.89
f0 [semi-tones]	2.76	1.87	3.06	2.12	5.30	2.23	6.30	2.19
Syllables [ms]	50.4	63.0	53.5	65.9	38.7	48.5	47.0	49.9
Pauses [ms]	133	143	168	139	223	177	296	254
#common pauses	284	262	29	32	389	378	35	34
#extra pauses	70	22	7	3	238	39	23	7
#missing pauses	58	80	8	5	29	40	3	4

4.3. Subjective comparison

From the test corpus we selected the 10 formulae whose synthetic f0-contours differed the most *between the two systems* (RMSE). Note that the differences between synthetic and *observed* contours is *not* taken into consideration for the selection. Synthetic versions (igm and SFC) and a resynthesis version (Org) were build. These 30 stimuli were presented in version-pairs to the subjects in a forced choice preference test. Presenting pairs in both orders (AB BA) the test consisted of 60 pairs to be evaluated.

The subjects were asked to take into account naturalness of rhythm and melody as well as the similarity between the structure of the written form of the formulae and the prosodically forwarded structure. Eleven native German and ten native French speakers with normal hearing attended the test.

For German the evaluations of the synthetic versions are close up. For French the results show a clear preference of SFC over igm. Resynthesis is preferred against synthetic versions – the scores being consistent with the inter-system pairs.

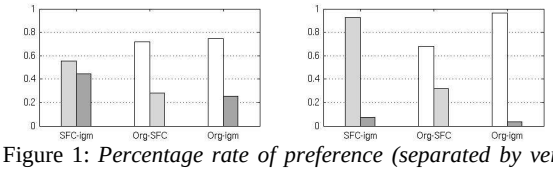

Figure 1: *Percentage rate of preference (separated by version-pairs). Left: German; right: French.*

5. Discussion

The SFC shows rather good results for both languages whereas shortcomings in the rhythm component have to be noted (table 1). For French they are in part due to certain parts of each formula for which no sufficient statistical information was available, that is needed to calculate the phoneme and pause duration from the lengthening factor.

The SFC does not take into account yet that German is a language with lexical accent, whereas French is not. The fact that this seemed to be of minor importance on its performance, might be due to the limited variety of words contained in the formulae and their quite regular appearance.

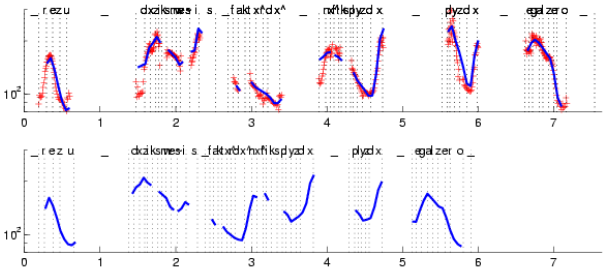

Figure 2: *Top: measured and stylized f0-contour; bottom: prediction. The SFC tends to shorten pause durations.*

The igm performed comparably to the SFC in German, whereas it showed shortcomings for French. First of all this could be due to the used adaptation of the original system which might be improved concerning the labeling of the Fujisaki parameters, the implementation of the NN and the newly defined input parameters.

French utterances seemed to be more resistant to linguistically coherent Fujisaki labeling. Accent commands tend to be very long, i.e. stretching over several syllables. The variation of phrase-contour amplitudes is considerably higher than for the German corpus. This lack of regularity might be responsible for the insufficient results of their prediction.

Another problem, that might occur independently from the language is the phasing of predicted f0-contours. Figure 3 shows an example where the form of the predicted contour is close to the original one (indicating an appropriate prediction of the accent's stretch and the amplitude), but it looses its influence on the intonation nearly completely, because it is shifted to the voiceless part of the syllable.

The scores of the igm for French might have "suffered" from the stimulus selection: the prediction of the SFC being closer to the original, the 10 formulae chosen as most different within the

test corpus show at the same time the greatest differences between igm and original. This is not a bug of the selection procedure but a feature: it chooses utterances that are “difficult” for either system and renders the evaluation more selective.

The training procedure of the SFC describes a loop that is implemented directly in the system. Nevertheless, it can be found in a similar way in the IGM: at first sight, the labeling of Fujisaki parameters, the training of the NN and the subsequent synthesis form a linear process. It becomes a loop though, when the results are examined, and the labeling revised in order to enhance accuracy and predictability of the labeling.

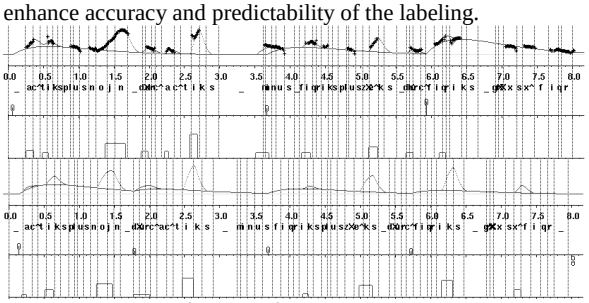

Figure 3: *Example for loss of movement in melody due to a shift of an accent command (around t=5s).*

Conclusions & perspectives

The SFC meets the challenge being applied to German. Using the same phonological input as for French seems – at least for spoken maths – sufficient. The adaptation of the IGM to handle the demanding structure of the used corpus proved to be successful for German whereas its transposition to French still needs further work.

We could present here only global results, but one of the main interests of this small-scaled comparative evaluation is the diagnostic information that detailed analysis of results provides since the proposed stimulus selection procedure tends to choose “difficult” utterances.

References

[1] Aubergé, V., 1992 Developing a structured lexicon for synthesis of prosody, in *Talking Machines: Theories, Models and Designs*, G. Bailly and C. Benoît, Editors. Elsevier B.V. p. 307-321.

[2] Aubergé, V., 1993 Prosody Modeling with a dynamic lexicon of intonative forms: Application for text-to-speech synthesis. *Working Papers of Lund University*, 41: p. 62-66.

[3] Aubergé, V. and Bailly, G., 1995 Generation of intonation: a global approach. in *Proceedings of the European Conference on Speech Communication and Technology*. Madrid. p. 2065-2068.

[4] Bailly, G., 1989 Integration of rhythmic and syntactic constraints in a model of generation of French prosody. *Speech Communication*, 8: p. 137-146.

[5] Barbosa, P. and Bailly, G., 1994 Characterisation of rhythmic patterns for text-to-speech synthesis. *Speech Communication*, 15: p. 127-137.

[6] Campbell, W.N., 1993 Automatic detection of prosodic boundaries in speech. *Speech Communication*, 13: p. 343-354.

[7] de Tournemire, S., 1994 Recherche d'une stylisation extrème des contours de F0 en vue de leur apprentissage automatique. in *Journées d'Etudes sur la Parole*. Trégastel, France. p. 75-80.

[8] Fujisaki, H. and Kawai, H., 1988 Realization of linguistic information in the voice fundamental f requency contour of the spoken Japanese. in *IEEE International Conference on Acoustics, Speech, and Signal Processing*. p. 663-666.

[9] Fujisaki, H. and Sudo, H., 1971 A generative model for the prosody of connected speech in Japanese. *Annual Report of Engineering Research Institute*, 30: p. 75-80.

[10] Gårding, E., 1991 Intonation parameters in production and perception. in *Proceedings of the International Congress of Phonetic Sciences*. Aix-en-Provence, France. p. 300-304.

[11] Gee, J.-P. and Grosjean, F. , 1983 Performance structures: a psycholinguistic and linguistic appraisal. *Cognitive Psychology*, 15: p. 411-458.

[12] Grosjean, Grosjean, F., and Lane, 1979 The Patterns of Silence: Performance Structures in Sentence Production. *Cognitive Psychology*, 11: p. 58-81.

[13] Hirst, D., Nicolas, P., and Espesser, R. , 1991 Coding the F0 of a continuous text in French: an experimental approach. in *Proceedings of the International Congress of Phonetic Sciences*. Aix-en-Provence, France. p. 234-237.

[14] Hirst, D.J., Di Cristo, A., and Espesser, R., 2000 Levels of representation and levels of analysis for the description of intonation systems, in *Prosody: Theory and Experiment*, M. Horne, Editor. Kluwer Academic Publishers: Dordrecht - the Netherlands. p. 51-87.

[15] Holm, B. and Bailly, G., 2000 Generating prosody by superposing multi-parametric overlapping contours. in *Proceedings of the International Conference on Speech and Language Processing*. Beijing, China. p. 203-206.

[16] Holm, B. and Bailly, G. , 2002 Learning the hidden structure of intonation: implementing various functions of prosody. in *Speech Prosody*. Aix-en-Provence, France. p. 399-402.

[17] Holm, B., Bailly, G., and Laborde, C., 1999 Performance structures of mathematical formulae. in *International Congress of Phonetic Sciences*. San Francisco, USA. p. 1297-1300.

[18] Malfrère, F., Dutoit, T., and Mertens, P., 1998 Fully Automatic Prosody Generator for Text-to-Speech Synthesis. in *International Conference on Speech and Language Processing*. Sidney, Australia. p. 1395-1398.

[19] Mixdorff, H., 2002 An integrated approach to modeling German prosody. *Dr.-Ing.habilitatus*. Technische Universität: Dresden.

[20] Mixdorff, H. and Jokisch, O., 2001 Building an integrated prosodic model of German. in *European Conference on Speech Communication and Technology*. Aalborg, Denmark. p. 947-950.

[21] Morlec, Y., Bailly, G., and Aubergé, V., 2001 Generating prosodic attitudes in French: data, model and evaluation. *Speech Communication*, 33(4): p. 357-371.

[22] Pierrehumbert, J., 1981 Synthetizing intonation. *Journal of the Acoustical Society of America*, 70(4): p. 985-995.

[23] Prudon, R., d'Alessandro, C., and Boula de Mareüil, P. , 2002 Prosody Synthesis by Unit Selection and Transplantation on Diphones. in *IEEE 2002 Workshop on Speech Synthesis*. Santa Monica, CA

[24] 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. *International Conference on Speech and Language Processing*, 2: p. 867-870.

[25] t' Hart, J., Collier, R., and Cohen, A., 1990 A perceptual study of intonation: an experimental-phonetic approach to speech melody. *Cambridge: Cambridge University Press.*

[26] Taylor, P., 2000 Analysis and synthesis of intonation using the tilt model. *Journal of the Acoustical Society of America*, 107(3): p. 1697-1714.

[27] Taylor, P. and Black, A.W., 1999 Speech synthesis by phonological structure matching. in *EuroSpeech*. Budapest, Hungary. p. 1531-1534.

[28] Thorsen, N.G., 1983 Standard Danish sentence intonation - Phonetic data and their representation. *Folia Linguistica*, 17: p. 187-220.

[29] Traber, C., 1992 F0 generation with a database of natural F0 patterns and with a neural network, in *Talking Machines: Theories, Models and Designs*, G. Bailly and C. Benoît, Editors. Elsevier B.V. p. 287-304.

[30] van Santen, J.P.H., 2002 Quantitative modeling of pitch accent alignement. in *International Conference on Speech Prosody*. Aix-en-Provence, France. p. 107-112.

[31] van Santen, J.P.H. and Möbius, B., 1997 Modeling pitch accent curves. in *ESCA Workshop "Intonation: Theory, Models and Applications"*. Athens, Greece. p. 321-324.