Published in: Interspeech, Sep 2006, Pittsburgh, United States. pp.2366-2369.

Generating German intonation with a trainable prosodic model

Gérard Bailly *Jan Gorisch*

Institut de la Communication Parlée, 46, av. Félix Viallet, 38031 Grenoble Cedex, France
{gerard.bailly,jan.gorisch}@icp.inpg.fr

Abstract

A trainable prosodic model called SFC (Superposition of Functional Contours), proposed by Holm and Bailly, is here confronted to German intonation. Training material is the publicly available Siemens Synthesis Corpus that provides spoken utterances for high-quality speech synthesis. We describe the labeling framework and first evaluation results that compares the original prosody of test sentences of this corpus with their prosodic rendering by the proposed model and state-of-the-art systems available on-line on the web.
Index Terms: speech synthesis, prosody, evaluation

Introduction

The trainable prosodic model SFC (Superposition of Functional Contours) has been developed by Holm and Bailly [1-3]. It implements a theoretical model of intonation initially sketched by Aubergé [4, 5] that promotes an intimate link between phonetic forms and linguistic functions: metalinguistic functions acting on different discourse units (thus at different scopes) are directly implemented as global multiparametric contours. These metalinguistic functions refer to the general ability of intonation to demarcate phonological units and convey information about the propositional and interactional functions of these units within the discourse. This trainable prosodic model has been confronted to speech styles (from read speech to spoken maths) and different languages including French, Galician or more recently Chinese [6]. German is of most interest because of its rich morphology and its potentially deep recursive syntactic embedding. Analysis of German prosody notably induces Schreuder and Gilbers [7] to question the Strict Layer Hypothesis [8] and claim for the existence of recursive prosodic phrases. While most quantitative models of German intonation that have been so far applied to speech synthesis use a phonological representation with few levels when not limited to prosodic phrases [9, 10].

We describe here our first efforts in confronting the SFC - that may potentially capture rich embedded performance structures [11] – to German intonation. Our first parameterization of the SFC using limited training material is evaluated against state-of-the-art text-to-speech systems available on the web.

1. The SFC prosodic model

In this section we sketch briefly the main features of this model. For more details please refer to Bailly et al [1].

The SFC trainable prosodic model *directly* encodes metalinguistic functions by phonetic events – i.e. overlapping multiparametric contours – without any intermediate surface representation.

Input. These metalinguistic functions refer to the general ability of prosody to segment, structure, emphasize or encode semantic or pragmatic cues associated with speech units. As emphasized by Aubergé [12], other linguistic agents (morpho-syntax, semantics, etc) collaborate with prosody to encode this information. These metalinguistic functions apply to units of variable sizes (discourse, sentence, clause, group, word, syllable, phoneme). The set of these metalinguistic functions [see intonation and its uses in 13] is quite open and most of the parameterization of the SFC resides in the identification of the metalinguistic functions used in the training material and the speech units they apply to.

Prosodic contours. The SFC postulates that these metalinguistic functions are encoded via multiparametric contours. The extend of each contour equals to the scope of the functions, i.e. the contours are coextensive to the speech units carrying the functions. Since the same metalinguistic function (e.g. segmentation) may apply to speech units of various sizes - and potentially embedded – the elementary multiparametric contours associated with each unit and each function (one unit may carry several functions, such as a word being emphasized and having a particular role in the syntactic structure that merits a specific spotlight) overlap. The parallel encoding of these several metalinguistic functions by overlapping contours is simply done by superposing and adding these elementary contributions by parameter-specific operators (see illustration in Figure 1): for f0, addition in the log-domain; for duration, addition of z-scores of rhythmic units, etc.

Mapping functions to contours. Considering prosodic contours as the superposition of elementary contours is a many-to-one ill-posed problem that requires regularization schemes. The Fujisaki model [14], for example, imposes constraints on the shape of these elementary contours (exponential responses of second-order filters to impulses and square waves). The SFC model does not impose such low-level constraints, but relies only on the consistency between different instantiations of the same discourse function on different units of different sizes within the corpus.

Contour generators The instantiation of a given function on a unit - the calculation of one elementary multiparametric contour - is performed by so-called *contour generators*. A *contour generator* generates thus a family of contours, a set of prosodic *clichés* [15] that develop on units of different size but encode the same metalinguistic function. General-purpose contour generators have been developed in order to be able to generate a coherent family of contours indexed only by their scopes. These contour generators are implemented as simple feed-forward neural networks and an original analysis-by-synthesis method has been developed to train these networks that best predict the target prosodic stylization given the metalinguistic labels [1].

2. Training the model

2.1. The prosodic database

We used speech data from the speaker *ai* of the SI1000P database commercially available from ELRA. The SI1000P recordings were done to provide material for high quality concatenate speech synthesis. It contains 1000 newspaper sentences read by two German professional broadcasting announcers in studio quality together with the laryngographic signal and the glottal pulse stream. Parts of the corpus were labeled and segmented phonemically (SAM-PA) and prosodically (borders+ accents). For this experiment we labeled carefully 80 utterances of the corpus. 70 sentences

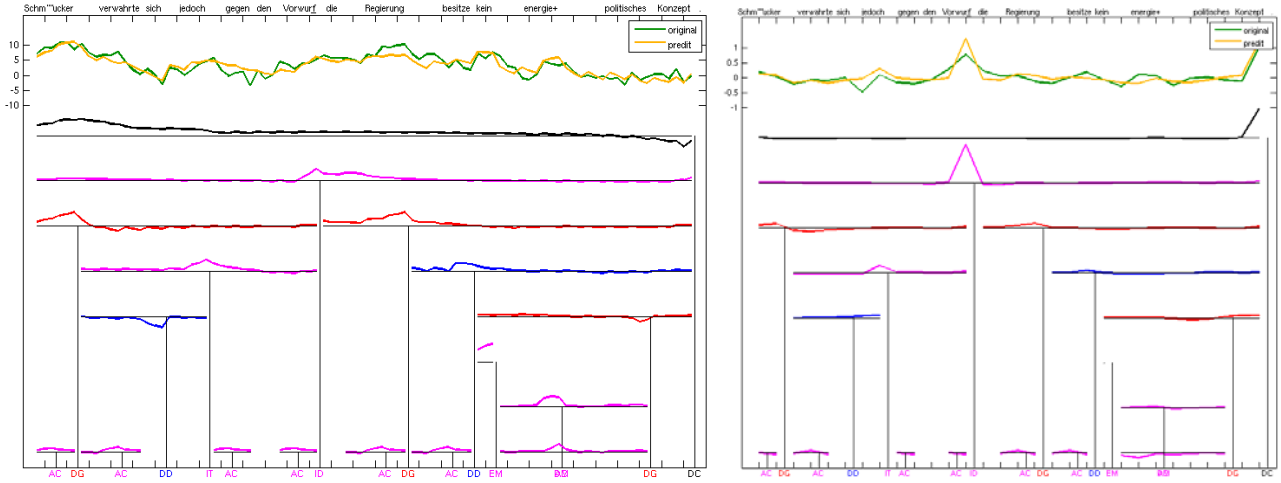

Figure 1. Comparing original and predicted prosodic contours. Left: f0. Right: syllable lengthening. For each caption: top: superposition of predicted (yellow) and original (green) contours; bottom: elementary contours predicted for each discourse function used to encode the linguistic structure of the utterance; the prediction is obtained by superposing and adding these elementary contours. Horizontal axis represents the syllable count

were used as training material. The 10 remaining sentences were used for evaluating the trained prosodic model (see §3).

2.2. Annotating utterances with metalinguistic functions

The first metalinguistic function is sentence modality (we have shown elsewhere [16] that prosodic attitudes in general are characterized/encoded by prosodic clichés). Its scope is the whole sentence. In our corpus all sentences are declarative and markers thus cue beginning and ending of each sentence. For instance, the sentence shown in Figure 1 is first annotated with the mark DC (for declaration) as below:

[Schmücker verwahrte sich jedoch gegen den Vorwurf die Regierung besitze kein energiepolitisches Koncept.]$_{DC}$

In our work, we always consider metalinguistic functions responsible for giving cues to the syntactic structure of sentences in the discourse. We thus annotate dependency relations between chunks [see also 17, 18, 19, for dependency structure analysis and prosodic correlates of attachment/branching of syntactic constituents]. We consider four kinds of dependency relations that may link constituents (words, groups, phrases, clauses): left dependency (DG, *dépendence à gauche*) linking the head of a sub-tree (the "governor" or "mother") with its immediately linearly preceding dependent unit ("sister"), right dependency (DD, *dépendence à droite*) linking the governor with its immediately following dependent unit, interdependency (IT) linking two adjacent units headed by the same governor and independency (ID) when none of the preceding simple relations can be identified. The syntactic parse we use is thus very simplified and can be accomplished using a chink and chunk technique [20, see also 21, for the use for syntax/prosody mapping]. For instance, the sentence shown in Figure 1 is further parsed as below:

[[Schmücker]$_{DG}$ *[[[verwahrte sich]*$_{DD}$ *[jedoch]]*$_{IT}$ *[gegen den Vorwurf]]]*$_{ID}$ *[[die Regierung]*$_{DG}$ *[[besitze]*$_{DD}$ *[kein [energie-politisches]*$_{DG}$ *[Konzept]]]]*

We then added markers for morphological decomposition. In the example, the word "energiepolitisches" is then parsed as:

[[energie]$_{AM}$ *[politisches]]*

German has lexical stress. So markers are also added to signal lexical stress position within each morpheme. In the example, the words "Schmücker" and "Regierung" are further marked as below:

[[Schmü]$_{AC}$*[cker]]* and *[[Regie]*$_{AC}$*[rung]]*

The last metalinguistic function is emphasis. Some demonstrative (e.g. "dieser"), negations (e.g. here "kein") or newly introduced proper nouns (e.g. "Schmücker") receive emphatic stress. In the example, the word "kein" receives the marker EM (for emphasis):

[kein]$_{EM}$

The Figure 1 displays the elementary melodic and rhythmic contours generated by the 8 contour generators (responsible for generating multiparametric contours encoding DC, DD, DG, IT, ID, AM, AC and EM) on the different scopes. It also displays the results of the superposition and addition of these elementary contours in comparison with the original training material.

2.3. Prosodic stylization

We analyze and generate*multiparametric prosodic contours*, i.e. we model themelody *and* rhythmic organization of the utterance. These contours capture the prosodic characteristics of the syllables of each utterance. Each syllable is characterized by a melodic movement [stylized by three F0 values on the vocalic nucleus as initially proposed by 22] and a lengthening factor (that will stretch or compress all phonemic segments of that syllable using z-scoring, see [23]).

2.4. Mapping metalinguistic functions to elementary multiparametric contours

The SFC (i.e. the eight contour generators) is then trained using phonetic and associated annotations from 70 utterances. As stated above, the iterative training consists in adjusting the contour generators so that their combined outputs best predicts observed multiparametric contours. Prosodic stylization of 10 test utterances (see Appendix) is then predicted using the trained SFC. Prediction performance on training and test material is given in Table 1. Such numbers are difficult to find in the literature… They could be compared to the correlation coefficients published by Mixdorff & Jokish [24]: 0.55 for F0 parameters and 0.82 for durations.

Table 1: RMS errors (correlation coefficients) pour training and test material. F0 mean is set to 80 Hz for that speaker.

Parameter	Training	Test
F0 (Hz)	13.8 (0.76)	17.3 (0.69)
F0 (semitones)	2.3 (0.77)	3.1 (0.7)
Lengthening factor	0.19 (0.79)	0.19 (0.46)
Durations(ms)	18.5 (0.71)	21.8 (0.74)
Nb. of phonemes	6485	252

Proser	***http://www.atip.de/german/technologie/tts/proseronline.htm***
AT&T	***http://public.research.att.com/~ttsweb/tts/demo.php***
Festival CSLU	***http://cslu.cse.ogi.edu/tts/demos/***
Cepstral	***http://www.cepstral.com/demos/***
DFKI Mary	***http://mary.dfki.de/online-demos/online-speech-synthesis/speech_synthesis***
Loquendo	***http://actor.loquendo.com/actordemo/default.asp?language=en***

Table 2. Names and web sites of the systems used in the evaluation experiment (output gathered on 10 March 2006).

3. Subjective evaluation

In order to situate this preliminary prosodic model with reference to available implementations, we collected prosodic characteristics of outputs of 6 state-of-the-art text-to-speech synthesizers (see Table 2): these systems will be anonymously named Alien1..6 in the following. The 10 test sentences were submitted to each online text-to-speech server and the synthetic audio files collected. Automatic f0 detection and phonemic alignment was then corrected by hand. The same prosodic stylization as used by SFC was finally performed to gather 6 alternative prosodic contours for each test sentence. Note that we compensate summarily for differences in voice registers of the different systems: we compute the mean f0 for each alien system and scale it to the mean f0 of our target speaker *ai* using a simple rule of three.

3.1. Evaluation procedure

We compared the synthetic prosody computed by our 7 different prosodic systems with the natural prosody using TD-PSOLA resynthesis of the natural test signals (procedure similar to [25]). Note that the stimuli driven natural prosody is synthetic: prosody for all systems is characterized by segmental durations and three F0 values on the vocalic nucleus. We do add automatically residual micromelody as proposed by Monaghan [26] and implemented as an option in the SFC [1]

The evaluation paradigm combines advantages of Mean Opinion Score (MOS) ratings and preference tests. We ask our Subjects to position our synthetic stimuli – identified as colored icons – in a geometric plane whose abscissa is the MOS scale. As they can listen to stimuli as many times as they want, they can compare stimuli by pairs, position stimuli already ordered in some part of the plane and further refine their judgment (this procedure has been used by Pfitzinger [27] for studying perceived speech rate of short segments). When they are satisfied with their ranking, the stimuli are once again played in decreasing MOS order before confirmation. Each subject ranks thus the ten sets of 8 stimuli arbitrarily thrown on 10 successive test planes.

3.2. Results

We launched the evaluation campaign on April 10th. By the time of the submission, 24 German listeners participated in the MOS test. 9 listeners have experience with synthetic speech and prosody while 15 have none. The preliminary results are shown in Figure 3. Four groups emerge from this first evaluation: the SFC has a statistically significant different mean from the natural prosody and alien systems 1, 5 and 6 (Anova analysis using anova1 and multcompare with hsd option *M*atlab® procedures). Experience with synthesis/prosody has no significant influence on the results. For instance, the proposed system generates an acceptable prosody that lies in the heading set of state-of-the-art systems we were able to input.

These results should be interpreted with caution. On one hand, we compare prosody computed by raw state-of-the-art TTS systems with one computed by a prosodic system which receives enriched text. Although one can imagine that most of theses systems compute some accentual, morphological and syntactic information, the linguistic front-end involved in the computation of this information is not prone to errors and delivers impoverished data compared to the hand-labeled data delivered to the SFC. On the other hand, these preliminary modeling and evaluation results aim at confronting the SFC to a new language and helping us identifying theoretical or modeling flaws.

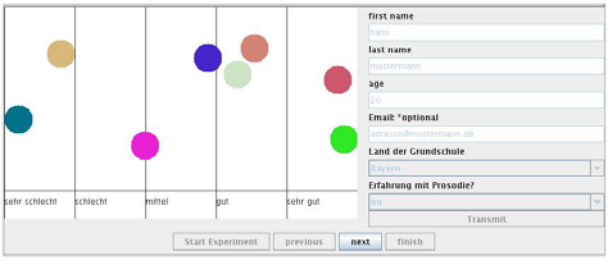

Figure 2. Java interface for subjective evaluation. Subjects should position stimuli (the natural utterance + synthetic renderings by 7 different prosodic models, including the one proposed here) on a MOS scale. Subjects can listen to each stimulus as many times as required (simple-click) for taking their decision. Positioning is done by drag and drop.

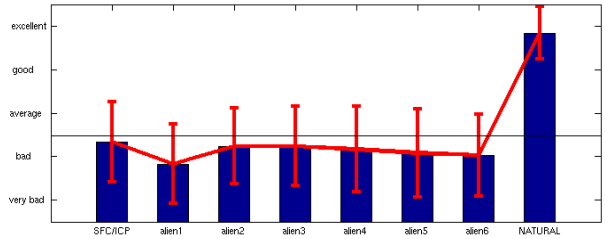

Figure 3. Comparative results of the MOS test. The proposed model lies in the heading set of state-of-the-art systems.

4. Conclusions

We described a first confrontation of a trainable prosodic model, the SFC (recently proposed by Holm and Bailly) to German intonation. The rich morphological productivity of the German language and its highly embedded syntax fit quite well our theoretical framework. The first results are objectively and subjectively encouraging but the generated prosody is still far from natural prosody – as the ones generated by the state-of-the-art systems we collected. Of course the evaluation procedure we used allows and perhaps favors such discrimination.

Potential improvements of the generated prosodic contours are: large differences occur because SFC only produces one prototypical cliché per metalinguistic function. A preliminary analysis of gross deviations between predicted and natural prosodic parameters suggests that several patterns can concurrently be used by speakers to encode the same function. Although this can be handled within the SFC by adding more metalinguistic functions, we want first to understand if this choice is arbitrary or contextual. The strict independence between the elementary contours should also be questioned. This is certainly especially true for the anchoring of contours

encoding metalinguistic functions on larger units than the word. This anchoring should probably include lexical stress positions.
Despite its crude assumptions and potential refinements, the SFC is flexible enough to automatically capture essential prosodic regularities of the language it observes given general assumptions on metalinguistic functions of intonation.

5. Acknowledgements

We warmly thank Harald Hoege at Siemens and Valerie Mapelli at ELRA for helping us getting rapidly this precious database.

6. Appendix

Test sentences:
0017: es sei falsch diese Sorge zu verniedlichen.
0169: ich bin Arbeiter.
0454: seine Aufrichtigkeit ist unbestritten.
0479: dennoch verlaufen alle Arbeiten planmäßig.
0814: die Regierung teile diese Auffaßung nicht.
0839: pro Kopf und Tag werden etwa achtzehn Mark bezahlt.
0980: es geht um die Kirche des Jahres 2000.
0625: die Polizei hat 6000 Mann.
0751: er genießt darum besondere Sympathien.
0508: er ist verheiratet und hat zwei Töchter.

References

[1] Bailly, G. and Holm, B. (2005) *SFC: a trainable prosodic model.* Speech Communication, **46**(3-4): p.348-364.
[2] Holm, B. (2003) *Implémentation d'un modèle morphogénétique de l'intonation. Application à l'énonciation de formules mathématiques*. PhD Thesis. Institut National Polytechnique: Grenoble - France. 239 pages.
[3] 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.
[4] 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.
[5] 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.
[6] Chen, G.-P., Bailly, G., Liu, Q.-F., and Wang, R.-H. (2004) *A superposed prosodic model for Chinese text-to-speech synthesis*. in *International Conference of Chinese Spoken Language Processing*. Hong Kong. p.177-180.
[7] Schreuder, M. and Gilbers, D. (2004) *Recursive patterns in phonological phrases*. in *International Conference on Speech Prosody*. Nara, Japan. p.341-344.
[8] Selkirk, E.O. (1984) *Phonology and syntax: the relation between sound and structure*. Cambridge, MA: MIT Press.
[9] Heuft, B. and Portele, T. (1996) *Synthesizing prosody: a prominence-based approach*. in *ICSLP*. Philadelphia, PA. p.1361-1364.
[10] 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.
[11] 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.
[12] Aubergé, V. (2002) *A Gestalt morphology of prosody directed by functions: the example of the step by step model developed at ICP*. in *Speech Prosody*. Aix-en-Provence - France. p.151-155.
[13] Bolinger, D. (1989) *Intonation and its Uses*. London: Edward Arnold.
[14] 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.
[15] Fónagy, I., Bérard, E., and Fónagy, J. (1984) *Clichés mélodiques.* Folia Linguistica, **17**: p.153-185.
[16] Morlec, Y., Bailly, G., and Aubergé, V. (2001) *Generating prosodic attitudes in French: data, model and evaluation.* Speech Communication, **33**(4): p.357-371.
[17] Bachenko, J. and Fitzpatrick, E. (1990) *A computational grammar of discourse-neutral prosodic phrasing in English.* Computational Linguistics, **16**: p.155–167.
[18] Bailly, G. (1989) *Integration of rhythmic and syntactic constraints in a model of generation of French prosody.* Speech Communication, **8**: p.137-146.
[19] Pynte, J. and Prieur, B. (1996) *Prosodic breaks and attachment decisions in sentence parsing.* Language and Cognitive Processes, **11**(1): p.165-191.
[20] Balfourier, J.-M., Blache, P., and van Rullen, T. (2002) *From shallow to deep parsing using constraint satisfaction*. in *Coling*. Taipei, Taiwan. p.36-42.
[21] Di Cristo, A., Di Cristo, P., Campione, E., and Veronis, J. (2000) *A prosodic model for text to speech synthesis in French*, in *Intonation: Analysis, Modelling and Technology*, A. Botinis, Editor. Kluwer: Amsterdam. p. 321-355.
[22] Tournemire, S.D. (1997) *Identification and automatic generation of prosodic contours for a text-to-speech synthesis system in french*. in *Proceedings of the European Conference on Speech Communication and Technology*. Rhodes, Greece. p.191-194.
[23] Barbosa, P. and Bailly, G. (1997) *Generation of pauses within the z-score model*, in *Progress in Speech Synthesis*, J.P.H.V. Santen, et al., Editors. Springer Verlag: New York. p. 365-381.
[24] Mixdorff, H. and Jokisch, O. (2001) *Implementing and evaluating an integrated approach to modeling German prosody*. in *ETRW on Speech Synthesis*. Perthshire, Scotland
[25] Boula de Mareüil, P., d'Alessandro, C., Raake, A., Bailly, G., Garcia, M.-N., and Morel, M. (2006) *A joint intelligibility evaluation of French text-to-speech systems: the EvaSy SUS/ACR campaign*. in *Language Ressources and Evaluation Conference (LREC)*. Genova - Italy. p.2034-2037.
[26] Monaghan, A.I.C. (1992) *Extracting microprosodic information from diphones -- a simple way to model segmental effects on prosody for synthetic speech*. in *International Conference on Speech and Language Processing*. Banff, Canada. p.1159-1162.
[27] Pfitzinger, H.R. (1998) *Local speech rate as a combination of syllable and phone rate*. in *International Conference on Spoken Language Processing*. Sydney, Australia. p.1087-1090.