A trainable prosodic model: learning the contours implementing communicative functions within a superpositional model of intonation

G. Bailly, B. Holm & V. Aubergé

Institut de la Communication Parlée, UMR CNRS n°5009, INPG/Univ. Stendhal
46, av. Félix Viallet, 38031 Grenoble Cedex, France
{bailly,holm,auberge}@icp.inpg.fr

Abstract

This paper introduces a new model-constrained, data-driven method to generate prosody from metalinguistic information. We refer here to the general ability of intonation to demarcate speech units and convey information about the propositional and interactional functions of these units within the discourse. Our strong hypotheses are that (1) these functions are directly implemented as prototypical prosodic contours that are coextensive to the unit(s) they apply to, (2) the prosody of the message is obtained by superposing and adding all the contributing contours. We describe here an analysis-by-synthesis scheme that consists in identifying these prototypical contours and separating out their contributions in the prosodic contours of the training data. We will show that such a trainable prosodic model generates faithful prosodic contours with very few prototypical movements.

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. In the framework of automatic prosody generation, we aim at computing adequate prosodic parameters carrying that information. We thus consider here prosodic models able to compute automatically prosodic parameters from linguistic, phonological and phonotactic specifications in the context of speech synthesis. In spite of a few systems getting rid of a phonetic description of prosody by incorporating directly the linguistic, phonological and phonotactic specifications in the selection process [30], most speech synthesis systems use a specific prosodic model that computes f0, phoneme durations or energy profile that are used to distort selected units and sometimes to select them [12].

These prosodic models are generally learnt using annotated corpora. Linguistic, phonological, phonotactic and phonetic descriptors are collected for each unit of the utterances (generally the phoneme or the syllable). Model-based (regression trees, HMMs, Neural Networks…) or sample-based (vector quantization, contour selection…) mapping tools are then used to achieve the best phonetic prediction according to a distance metrics (generally RMS). The prediction of prosodic parameters was initially decoupled with separate trainable models for f0 [20, 27, 29, 33], for phoneme durations [8, 19, 24, 26, 35] and – more recently – for the intensity profile [34]. With the development of corpus-based synthesis techniques and powerful mapping tools [9, 36], multiparametric prosodic models [22, 31] tend now to share common mapping models… with a dangerous (?) trend towards theory-neutral models where blind intensive training takes over comprehensive models of intonation [4]. Most trainable prosodic models in fact consider that linguistic, phonological, phonotactic and phonetic descriptors just as possible factors influencing the prosodic realization of a certain phoneme given the speaker and the situation of communication (often reading)… the mapping tools being responsible for evaluating the contributions of these factors within a model of interaction ranging from a additive, multiplicative, sum-of-products [35] models to more complex non-linear models such as neural networks incorporating eventually intermediate predictions made for earlier units though recurrent connections [e.g. 33]

We present here a trainable prosodic model that implements a non theory-neutral model of intonation. This model, initiated by Aubergé [1-3], promotes a intimate link between form and function: discursive functions acting on different discourse units – thus at different scopes – are directly implemented as global multiparametric contours. These contours are co-extensive to the discourse units and are simply superposed and added to generate observable prosodic continuums.

Section 2 describes the phonological model that essentially specifies what are the actual contributions of prosody to the sentence meaning. Section 3 presents the phonetic model used to describe observable prosodic continuums. Section 4 describes the mapping model – known as the SFC model [5, 18] – that essentially consists in training a few contour generators (one per discursive function) using an original analysis-by-synthesis training loop.

2. The phonological model

As stated by Cutler [Cutler, 1991 #300; p.267], "prosody is as much involved as any other aspect of linguistic structure in speakers' efforts to do their part in achieving this goal [maximizing the successful message transmission].. both salience and segmentation figure in prosodic contributions to realization of the speaker-listener contract". Other prosodic contributions to the ease of discourse interpretation include of course communicative values associated with each salient/segmented unit such as contrastive emphasis on phonemes or syllables, lexical stress, emphasis on words, modality or prosodic attitudes on sentences. As stated by Hirst [15], most current account of prosodic function within prosodic annotation systems deal in fact with prominence and boundaries [37] aggregating often under identical symbols very different functions. Within the framework of non-linear phonology, prominence and boundaries apply and delimit embedded constituents such as rhythmic, tonal and intonation

units. This strict layer hypothesis is however questioned by a number of studies that claim for the necessity of incorporating scopes/domains to prominence and boundaries in order to account for the embedded [21] and possibly recursive phonological hierarchy [28].

Instead of considering a posteriori the mapping of linguistic units and such phonological constructs, we consider on the contrary that the general ability of prosody in highlighting and segmenting units goes into the linguistic structure's service in order to help the listener decoding the discourse structure. The domain of action of prosody *is the linguistic domain* and the linguistic structure provides to prosody the specification of its tasks as *triplets* (function; units; importance).

It is then the responsibility of the mapping model to see if all these tasks can effectively be accomplished and how to share the communication channels between tasks according to their importance in the discourse.

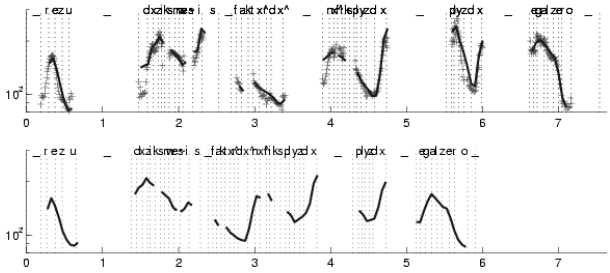


Figure 1: Top Original and stylized f0 curve. Bottom: Predicted f0 curve, the SFC tends to shorten pause durations.

3. The phonetic model

We generate *multiparametric prosodic contours* i.e. melody *and* rhythmic organization of the synthetic message are generated together within the same generation process. In fact each *Inter Perceptual-Center Group* (IPCG) [6] is characterized by a melodic contour (stylized by three F0 values on the vocalic nucleus as initially proposed by de Tournemire [32]) and a lengthening factor (that will stretch or compress the segmental constituents in a nonlinear way).

Melody. A first decomposition of the F0 curve is performed using a stylization procedure similar to MOMEL [14] that factors a smooth macromelodic component and a microprosodic component consisting of microprosodic residual deviations due to the segmental substrate. Contrary to MOMEL stylization that does not imply any a priori synchronization with the segmental chain, we stylize the macromelodic component by sampling it for each IPCG at 10, 50 and 90% of its vocalic nuclei. The mapping model will have in charge with the prediction of this crude approximation, i.e. the *melodic skeleton*. Concatenative synthesis provides however a way to give *flesh* to this skeleton. The same stylization process is in fact performed for the utterances from which the segments are extracted and the residual component (stylization errors + microprosodic component) is stored, retrieved and added at synthesis time (see initial proposal in [23]): Note that this generation process is entirely compatible with a superposition model.

Rhythm. Barbosa & Bailly [6] propose in fact a multi-level timing generation process similar to Campbell [9] but use the IPCG as an intermediate rhythmical unit. Each IPCG is characterized by a lengthening/shortening factor equal to the quotient between the actual duration of the ICPG and an expected ICPG duration. This expected duration is a weighted sum of (a) the sum of the mean values of its constitutive segments (b) an average ICPG duration reflecting a tendency to isochrony. For instance training these weights (see comments in §5) for the prosody of a manual cued speech speaker [13] results in a higher coefficient α for (a) than for the same corpus uttered by a non cuer (α=0.4 vs. α=0.21).

A z-scoring procedure is then applied in order to distribute the actual ICPG duration among its constitutive segments. Pause insertion is obtained by saturating the lengthening factor of the IPCG: the pause duration is computed as the duration loss between the desired lengthening factor and the saturated lengthening factor (for further details please refer to [7]). Thus contrary to prosodic phonology, pause is an emergent process resulting from low-level constraints (overall speech rate, pausing strategy resulting from the control of the saturation curve) and do not determine a priori the performance structure.

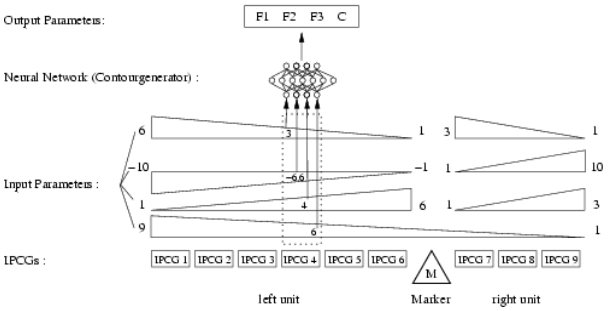
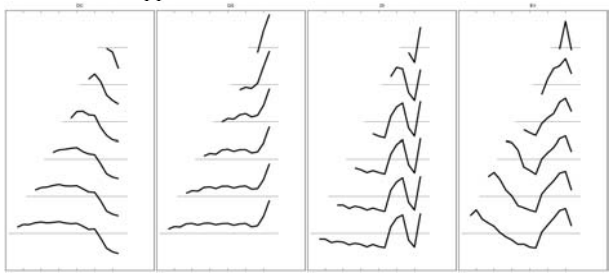
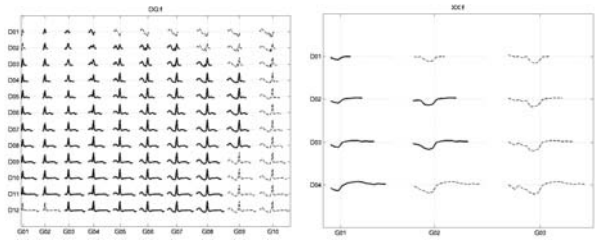


Figure 2: Contour generators are trained in order to deliver instantiations of one metalinguistic function given the size of the units it is applied to.



(a) assertion (b) question (c) incredulity (d) evidence



(e) dependant/governor (f) clitic/content

Figure 3. Expansion of f0 movement for different metalinguistic functions applied to units with increasing size. Top; modalities and prosodic attitudes in French: (a) assertion, (b) question, (c) incredulous question, (d) evidence. Bottom; linking two constituents (abscissa/ordinate: size of left/right unit): (e) between a governor and its left dependant (e.g. major phrase boundary between a subject NG and a VG); (f) between a content word and its preceding clitic word (e.g. f0 dip on the determinant introducing a noun).

4. The mapping model

Considering prosodic contours as the superposition of elementary contours is a many-to-one ill-posed problem that requires regularization schemes. Fujisaki's model [11] imposes for example constraints on the shape of these elementary contours. The SFC model does not impose such low-level constraints. It relies only on the consistency between different instantiations of the same metalinguistic function within the corpus. These instantiations are supposed to be performed by so-called *contour generators*.

Contour generators. As stated in §2, the phonological model delivers *triplets* (function; units; importance). We define the *scope* of a function as the continuous set of words which are concerned with this function and comprises the units this function applies to. Each discourse function is then encoded by a specific prototypical contour anchored to the function's scope by so-called *landmarks*, i.e. beginning and end of the units concerned with this function. As the discourse function can be applied to different scopes, it is characterized by a family of contours – some sort of prosodic "clichés" [10]. General-purpose contour generators have been developed in order to be able to generate a coherent family of contours given only their scope. These contour generators are actually implemented as simple feedforward neural networks receiving as input linear ramps giving the absolute and relative distance of the current syllable from the closest landmarks and delivering as output the prosodic characteristics for the current syllable (see Figure 2). Each network have very few parameters – typically 4 input, 15 hidden and 4 output units = 4*(15+1)+15*(4+1) = 139 parameters – to be compared to the thousands parameters necessary to learn a "blind" mapping between phonological inputs and prosodic parameters such as in [22, 33]. We have shown that our contour generators implement a so-called Prosodic Movement Expansion Model (PMEM) that describes how prototypical contours develop according to the scope (see for example PMEMs of different discourse functions in Figure 3): the set of prototypical contours that a contour generator implementing a certain function actually generates is called in the following a *dynamical prototype*. Note that the choice of the neural networks implementation of the PMEM is not exclusive, but offers an efficient learning paradigm as described below. The final multiparametric prosody is thus obtained by superposing and adding the many contours produced by a few independent contour generators (typically 3 or 4) and parameterized by their variable scopes.

Training contour generators. The problem is now to feed our contour generators with samples of elementary multiparametric contours from raw data. In the case of a superpositional model, the problem is often ill-posed since each observation is in general the sum of several contributions, i.e. here the outputs of contributing contour generators. We thus need extra constraints to regularize the inversion problem, e.g. shapes/equations of the superposed components as in [11]. In our phonetic model, shapes of the contributing contours are *a priori* unconstrained – which we feel to be important in a first time since we have shown that contours may potentially have complex shapes (e.g. those encoding attitudes at the sentence level as in Figure 3). Note however that nothing forbids in the following framework to later add constraints (such as imposing exponential shapes as in the Fujisaki's model) on those contours that are well understood in order to ease the emergence of other contours. The shapes of the contributing contours emerge as a by-product of an inversion procedure that parameterize contour generators in such a way that the prosodic contours predicted by overlapping and adding their contributions in the discourse best predicts observed realizations. For further details on this original analysis-by-synthesis loop - terms as SFC – please refer to [17, 18].

5. Comments

The analysis-by-synthesis procedure presented here gives access to the *hidden structure* of intonation [18]: the phonetic implementation of discourse functions emerges from the automatic parameterization of contour generators. This procedure is data-driven but also model-constrained and thus converges towards optimal prototypical contours that satisfy *both* bottom-up (close-copy synthesis) and top-down (coherent phonological description) constraints.

Contrary to most other trainable models of intonation, the training phase of the model presented here does essentially learn the *shapes* of the contours associated with pre-defined discursive functions. It does not learn for instance how these discursive functions are transmitted in parallel within the prosodic continuum: this is imposed by the superposition hypothesis.

Other non trivial parameters can also be trained during the mapping with raw data. Tuning of parameters such as the weight for the tendency to isochrony used for computing the reference rhythmic units or other global settings for the speaker or the speech style is often rephrased as how does it affects the global convergence of the SFC model. In his PhD thesis [16], Holm showed also how such a model can question the clustering of diverse discursive functions by merging/differentiating their implementations as contour generators.

Conclusions

We demonstrated elsewhere [25] that this model-based comprehensive generation scheme may be compatible with a certain technological efficiency: confronting data-driven models against such thematic databases used here should provide an interesting basis of comparison between models and approaches that we are still looking for.

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. (1997) No future for comprehensive models of intonation?, in Computing prosody: Computational models for processing spontaneous speech, Y. Sagisaka,

N. Campbell, and N. Higuchi, Editors. Springer Verlag. p. 157-164.
[5] Bailly, G. and Holm, B. (2002) Learning the hidden structure of speech: from communicative functions to prosody. Cadernos de Estudos Linguisticos, 43: p. 37-54.
[6] Barbosa, P. and Bailly, G. (1994) Characterisation of rhythmic patterns for text-to-speech synthesis. Speech Communication, 15: p. 127-137.
[7] 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.
[8] Bartkova, K. and Sorin, C. (1987) A model of segmental duration for speech synthesis in French. Speech Communication, 6: p. 245-260.
[9] Campbell, N. (1992) Multi-level timing in speech. University of Sussex: Brighton, UK.
[10] Fónagy, I., Bérard, E., and Fónagy, J. (1984) Clichés mélodiques. Folia Linguistica, 17: p. 153-185.
[11] 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.
[12] Fujisawa, K. and Campbell, N. (1998) Prosody-based unit selection for Japanese speech synthesis. in ESCA/COCOSDA International Workshop on Speech Synthesis
[13] Gibert, G., Bailly, G., Elisei, F., Beautemps, D., and Brun, R. (2004) Evaluation of a speech cuer: from motion capture to a concatenative text-to-cued speech system. in Language Ressources and Evaluation Conference (LREC). Lisbon, Portugal. p. accepted.
[14] 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.
[15] Hirst, D.J. (2003) The phonology and phonetics of speech prosody: between acoustics and interpretation. in International conference on speech prosody. Nara, Japan. p. 163-169.
[16] Holm, B. (2003) Implémentation d'un modèle morphogénétique de l'intonation. Application à l'énonciation de formules mathématiques. Institut National Polytechnique: Grenoble - France.
[17] 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.
[18] 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.
[19] Klatt, D.H. (1979) Synthesis by rule of segmental durations in English sentences, in Fontiers of Speech Communication Research, B. Lindblom and S. Ohlman, Editors. Academic Press: London. p. 287-300.
[20] Ljolje, A. and Fallside, F. (1986) Synthesis of natural sounding pitch contours in isolated utterances using hidden Markov models. TrASSP, 34: p. 1074-1080.
[21] Marsi, E.C., Coppen, P.-A.J.M., Gussenhoven, C.H.M., and Rietveld, T.C.M. (1997) Prosodic and intonational domains in speech synthesis, in Progress in Speech Synthesis, J.P.H. van Santen, et al., Editors. Springer-Verlag: New York. p. 477-493.
[22] 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.
[23] 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.
[24] O'Shaughnessy, D. (1981) A study of French vowel and consonant durations. Journal of Phonetics, 9: p. 385-406.
[25] Raidt, S., Bailly, G., Holm, B., and Mixdorff, H. (2004) Automatic generation of prosody: comparing two superpositional systems. in International Conference on Speech Prosody. Nara, Japan. p. 417-420.
[26] Riley, M. (1992) Tree-based modelling of segmental durations, in Talking Machines: Theories, Models and Designs, G. Bailly and C. Benoît, Editors. Elsevier B.V. p. 265-274.
[27] Sagisaka, Y. (1990) On the prediction of global Fo shapes for Japanese text-to-speech. IEEE International Conference on Acoustics, Speech, and Signal Processing, 1: p. 325-328.
[28] Schreuder, M. and Gilbers, D. (2004) Recursive patterns in phonological phrases. in International Conference on Speech Prosody. Nara, Japan. p. 341-344.
[29] Scordilis, M. and Gowdy, J. (1989) Neural Network based generation of Fundamental Frequency contours. IEEE International Conference on Acoustics, Speech, and Signal Processing,: p. 219-222.
[30] Taylor, P. and Black, A.W. (1999) Speech synthesis by phonological structure matching. in EuroSpeech. Budapest, Hungary. p. 1531-1534.
[31] Tesser, F., Cosi, P., Drioli, C., and Tisato, G. (2004) Prosodic data-driven modelling of narrative style in Festival TTS. in ISCRA WOrkshop on Speech Synthesis. Pittsburgh, USA. p. (submitted).
[32] 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.
[33] 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.
[34] Trouvain, J., Barry, W.J., Nielsen, C., and Andersen, O. (1998) Implications of energy declination for speech synthesis. in ETRW Workshop on Speech Synthesis. Jenolan Caves - Australia. p. 47-52.
[35] van Santen, J.P.H. (1992) Deriving text-to-speech durations from natural speech, in Talking Machines: Theories, Models and Designs, G. Bailly and C. Benoît, Editors. Elsevier B.V. p. 275-285.
[36] van Santen, J.P.H. (2002) Quantitative modeling of pitch accent alignment. in International Conference on Speech Prosody. Aix-en-Provence, France. p. 107-112.
[37] Wightman, C.W., Syrdal, A.K., Stemmer, G., Conkie, A., and Beutnagel, M. (2000) Perceptually based automatic prosody labeling and prosodically enriched unit selection improve concatenative text-to-speech synthesis. in International Conference on Spoken Language Processing. Beijing, China. p. 71-74.