A Weighted Superposition of Functional Contours Model for Modelling Contextual Prominence of Elementary Prosodic Contours

Branislav Gerazov$^{1,2}$, Gérard Bailly$^2$ and Yi Xu$^3$

$^1$ FEEIT, University of Ss. Cyril and Methodius in Skopje, Macedonia
$^2$ Univ. Grenoble-Alpes, CNRS, Grenoble-INP, GIPSA-lab, 38000 Grenoble, France
$^3$ Division of Psychology and Language Sciences, University College London, UK

gerazov@feit.ukim.edu.mk, gerard.bailly@gipsa-lab.fr, yi.xu@ucl.ac.uk

Abstract

The way speech prosody encodes linguistic, paralinguistic and non-linguistic information via multiparametric representations of the speech signals is still an open issue. The Superposition of Functional Contours (SFC) model proposes to decompose prosody into elementary multiparametric functional contours through the iterative training of neural network contour generators using analysis-by-synthesis. Each generator is responsible for computing multiparametric contours that encode one given linguistic, paralinguistic and non-linguistic information on a variable scope of rhythmic units. The contributions of all generators’ outputs are then overlapped and added to produce the prosody of the utterance. We propose an extension of the contour generators that allows them to model the prominence of the elementary contours based on contextual information. WSFC jointly learns the parameters of the elementary multiparametric functional contours and their weights dependent on the contours’ contexts. The experimental results show that the proposed weighted SFC (WSFC) model can successfully capture contour prominence and thus improve SFC modelling performance. The WSFC is also shown to be effective at modelling the impact of attitudes on the prominence of functional contours and syntactic relations in French, and that of emphasis on the prominence of tone contours in Chinese.

Index Terms: prosody, model, prominence, neural networks, SFC

1. Introduction

The way speech prosody encodes linguistic, paralinguistic and non-linguistic information via multiparametric representations of the speech signals is still an open issue. Most models of intonation postulate that this encoding is performed by local and salient spatio-temporal patterns such as tones, atoms or breaks inscribed into global gauges such as declinations and is available on GitHub

These function-specific patterns, in fact, emerge from statistical modelling. Given a dataset that contains multiple instances of these patterns, the SFC extracts the shapes and their average contributions thanks to an iterative analysis-by-synthesis training process that consists of training function-specific pattern generators. They are called multiparametric contours because the generated shapes feed a multiparametric score, i.e. one including melody, rhythm, head motion, etc. The SFC has been successfully used to model different functions acting at various linguistic levels, including: attitudes [5], grammatical dependencies [6], cliticisation [4], focus [7], as well as tones in Mandarin [8].

One shortcoming of the SFC model is that it is not sensitive to prominence: prosodic contours are simply superposed-and-added with no possibility of weighting their contributions. In this paper we supplement the SFC architecture with components responsible for weighting the contribution of the elementary contours in the decomposition. The weighted SFC (WSFC) consists in adjoining a weight module to each contour generator: while the contour generator still computes a multiparametric contour for each rhythmic unit of the scope, the weight module computes its contribution given the context of the scope in the utterance.

We assessed the plausibility of the proposed WSFC, and used it to explore two prosodic phenomena: i) the impact of the attitude, and ii) the impact of emphasis on the prominence of the other functional contours in the utterance. The two phenomena were explored in two different languages: French and Chinese. The results show that the integration of weighting in the contour generators is effective, relevant and robust. Also, by adding a degree of freedom to the model, it improves its modelling performance by providing more coherent contours. The whole implementation of the system has been licensed as free software and is available on GitHub$^1$.

2. The SFC contour generators

At the core of the SFC model are neural network based contour generators (CGs) that learn the elementary prosodic contours during the iterations of the analysis-by-synthesis loop [9, 5, 10, 3]. There is a single CG for each communicative function used in the dataset. Within an utterances, instances of these CGs are applied on the different scopes the functions encompass, i.e. the different number of rhythmic units (RUs), e.g. syllables, that the function spans.

An example SFC decomposition of the intonation of a

---

$^1$https://github.com/bgerazov/WSFC
French utterance is shown in the left plot in Fig. 3, where we can see a declaration contour overlapped with one left dependency between the verbal group and its direct object, and two clitic contours cueing articles. The decomposition was performed using the PySFC prosody analysis system [11].

3. The WSFC contour generators

The WSFC is based on the weighted contour generator (WCG) shown in Fig. 1, which is an expanded form of the CGs introduced in the SFC. The WCG includes a module for computing the contour’s weight that is designed to capture prominence based on linguistic context. The global architecture is reminiscent to the mixture of experts (ME) model proposed by Jacobs et al [12] in which predictions of experts, in our case contour generators, are weighted by gates, in our case weighting modules, and added to perform decisions or regressions.

The contour generator module, comprising a single layer neural network, receives the RU’s absolute and relative positions within the function’s scope, and generates the prosodic contour for that particular syllable. The weight module, also comprising a shallow neural network, receives a vector describing the linguistic context that the functional contour appears in. Based on this context vector it outputs a global context-specific weight for all of the RUs within the function’s scope. By imposing a single weight for the whole scope of the contour we force the weighting module to capture the overall prominence of the functional contours dependent on their context. The input context vector can be arbitrarily defined and tailored to the task at hand. In our experiments, we use three different binary encodings of the coinciding functional contours as context vectors to analyse the impact of attitude and emphasis on the prosodic contours; more details are given in Section 5.

The two modules in the WSFC CG are in mutual competition, in the sense that the general amplitude of the contour generated by the contour generator is multiplied by the output of the weighting module. In that sense, the amplitudes of the contours can become arbitrarily large if they are compensated by the weights becoming small. To limit this effect the range output by the weight module is limited to 0 – 2 through multiplying with 2 the final neuron’s sigmoid output. In addition, we apply regularisation to the weights’ mean across the data to avoid the weights becoming small. To limit this effect the range output by the weight module is limited to 0 – 2 through multiplying with 2 the final neuron’s sigmoid output. In addition, we apply regularisation to the weights’ mean across the data towards 1.0. This intuitively corresponds to an average function contour generated by the SFC model. Since training of the CGs is done in batches, the batch size is an important factor to take into account in this regularisation.

4. Objective and goals

Our objective is to assess the efficiency of the WSFC for modelling the prominence of the elementary prosodic contours. We hypothesise that i) the added degree of freedom will increase the modelling performance compared to the original SFC model. We will also use the WSFC to analyse the impact of linguistic context on the prominence of prosody contour realisation in a structured way. In this sense we will test an additional hypothesis: ii) the WSFC is able to capture the impact that function contour context has on its prominence.

5. Experiments

We conducted three experiments to confront the WSFC with empirical data.
square error (RMSE) between the original pitch contour and its reconstruction. Pitch is expressed in semitones and only errors within the vocalic nuclei are considered. We split the data into a training, validation and test sets and trained the WCGs with early stopping. The RMSE results for the test set showed that the WSFC gives a small improvement: 1.83 ± 1.03 for the WSC vs. 1.87 ± 1.03 for the SFC. This was nevertheless found significant using a paired t-test that gave a p-value of 0.037.

5.4. Modelling the impact of attitudes

Morlec et al [5] noted that melodic contours of sentences uttered with attitudes such as doubt, evidence, suspicious irony, suggesting a repetition of “old” information, exhibited reduced modulations by syntactic functions. The WSFC has the capability to contextualise such modulation. For this experiment, we enhanced the context vector to be an overlap indicator, which shows not just the attitude, but also any other functional contour appearing within the scope of the current one. This enhancement allows the WSFC to learn a distribution of weights for the functional contours for each attitude, rather than a single value. Note that in this and the following experiments we do not split the data into a train and test set, as we are analysing the expressive capacity of our model in its ability to capture prominence. The WSFC decomposition of the same French utterance for two attitudes: declaration (DC) and incredulous question (DI) is shown in Fig. 3. We clearly see that the WSFC has captured the quasi-suppression of the contours implementing the DD and XX functions in the context of DI with respect to the full-blown contours in DC. In fact, the weighting factors of XX are reduced by a factor of 5, i.e. from 1.75/1.81 down to 0.3/0.41. This quasi-suppression caused the SFC to average out a smaller clitic contour than the WSFC as is evident in the figure. The same quasi-suppression phenomenon is also present in all of the other attitudes except exclamation, but it is not shown here for brevity.

We compare here two training strategies: full training – performed using Morlec, and pretraining and freezing – pretraining the contour generators on DC first, as we suppose it exhibits the full blown syntactic contours. Then freezing the parameters of the contour generator modules when performing the full training. This mirrors the back-off strategy that was followed when using the SFC. Fig. 4 shows the distribution of weights for two functional contours: clitic and left-dependency, with and without pretraining. We can see that there is a suppression effect by some attitudes on the syntactic contours. We can also see that the WSFC is able to extract a more or less similar distribution of the weights with or without using pre-training. The decomposition proposed by WSFC is thus faithful and robust. There full training exhibits, however, a stronger contrast. The pretraining causes the final weights distributions for DC and EX to be close to 1. On the other hand, the distributions obtained with the full training are further spread out, whilst still maintaining the average value of 1 imposed by the regularisation.

5.5. Modelling the impact of emphasis

The first female speaker subset from Liu is used to analyse the impact of emphasis on the Chinese tones. Since Liu is designed to minimise tonal modulations by using utterances comprising words with identical tones, we use Chen to pretrain the WSFC tone contours with an extended right scope, and then retrain only their strengths on Liu. This is to allow the CGs to capture the well-known carry-over effect in Chinese tones [14], thus increasing modelling performance [15].

For this experiment, we define a further enhanced context vector that is an emphasis indicator. It not only includes information on: attitude, word boundary, and emphasis, it also takes into account at which temporal location the function occurs relative to emphasis. This is done because emphasis in Mandarin involves both on-focus expansion of pitch range and post-emphasis compression of pitch range [13]. Thus the com-
concentrations of attitudes in several other languages.

6. Conclusions

We proposed a prosody model capable of capturing the prominence of elementary prosodic contours that are based on their context of use. The WSFC has been also shown to improve the modelling performance of the SFC due to an added weight mechanism. We have demonstrated its robustness and its usefulness in analysing the impact of attitudes and emphasis on pitch range dynamics [13].

7. Acknowledgements

This work has been conducted with the support of the EU Marie Skłodowska-Curie Actions Individual Fellowship Project H2020-MSCA-IF-2016 “ProsoDeep: Deep understanding and modelling of the hierarchical structure of Prosody”.

Figure 3: Decomposition of the melody of the French utterance “Les gamins coupaient des rondins.” with the SFC (left), and the WSFC for declaration (DC, centre) and incredulous question (DI, right) into constituent functional contours: syntactical dependencies to the left and right (DG, DD), and clitics (XX). Activations of XX, DG and DD are strongly reduced when solicited in the DI context.

Figure 4: Distribution of WSFC weights for clitic (XX) and right dependency (DG) in Morlec without pretrained contours (left column) and with (right column).

An example WSFC decomposition of a Chinese utterance with emphasis is shown in Fig. 5. We can see that the final tone of the second emphasized word “Li” is considerably amplified compared to the preceding and following ones. Moreover, we can see that the tones following the emphasis are conversely reduced and that they are also affected by the word boundaries. Finally, we can see that the emphasis contour features an increase in pitch followed by a post-emphasis lowering.

The distribution of weights for the four tones as a function of their placement relative to the emphasis are shown in Fig. 6. We can see that there is a considerable difference in prominence of the tones located at the final RU in the emphasis. There is also a difference between pre- and post-emphasis weighting for the tones, with weights decreasing for tones 1 and 4 and increasing for tones 2 and 3. It is these changes that, together with the emphasis contour, capture the complex effect of post-emphasis on pitch range dynamics [13].

prominence in French and Chinese. The described methodology can be used to analyse other effects of context on prominence. Moreover, the proposed WSFC architecture allows for task-specific contextual inputs. We are currently exploring realizations of attitudes in several other languages.

Figure 5: WSFC decomposition of the intonation of the Chinese utterance: “Ye Liang hai pa Zhao Li shui jiao zuo meng.”., into component contours: declaration (DC), tone contour 4 (C4), word boundary (WB), and emphasis (EM).

Figure 6: Distribution of WSFC weights in Liu with emphasis context.

The text vector encodes if the RU occurs with no emphasis (None), with emphasis but precedes the final emphasised RU (EMp), is the final emphasised RU (EM), or succeeds it (EMc).

text vector encodes if the RU occurs with no emphasis (None), with emphasis but precedes the final emphasised RU (EMp), is the final emphasised RU (EM), or succeeds it (EMc).
8. References

[1] V. Aubergé and G. Bailly, “Generation of intonation: a global approach.” in EUROSPEECH, 1995.

[2] G. Bailly and B. Holm, “SFC: a trainable prosodic model,” Speech communication, vol. 46, no. 3, pp. 348–364, 2005.

[3] B. Holm and G. Bailly, “Learning the hidden structure of intonation: implementing various functions of prosody,” in Speech Prosody 2002, International Conference, 2002.

[4] G. Bailly and B. Holm, “Learning the hidden structure of speech: from communicative functions to prosody,” Cadernos de Estudos Linguisticos, vol. 43, pp. 37–54, 2002.

[5] Y. Morlec, G. Bailly, and V. Aubergé, “Generating prosodic attitudes in French: data, model and evaluation,” Speech Communication, vol. 33, no. 4, pp. 357–371, 2001.

[6] Y. Morlec, A. Rilliard, G. Bailly, and V. Aubergé, “Evaluating the adequacy of synthetic prosody in signalling syntactic boundaries: methodology and first results,” in Proceedings of the first International Conference on Language Resources and Evaluation, Granada, Spain, 1998, pp. 647–650.

[7] C. Brichet and V. Aubergé, “La prosodie de la focalisation en français: faits perceptifs et morphogénétiques,” Journées d’Etudes sur la Parole, Nancy-France, pp. 33–36, 2004.

[8] G.-P. Chen, G. Bailly, Q.-F. Liu, and R.-H. Wang, “A superposed prosodic model for Chinese text-to-speech synthesis,” in Chinese Spoken Language Processing, 2004 International Symposium on. IEEE, 2004, pp. 177–180.

[9] Y. Morlec, “Génération multiparamétrique de la prosodie du français par apprentissage automatique,” Ph.D. dissertation, 1997.

[10] B. Holm and G. Bailly, “Generating prosody by superposing multi-parametric overlapping contours.” in INTERSPEECH, 2000, pp. 203–206.

[11] B. Gerazov and G. Bailly, “PySFC – a system for prosody analysis based on the superposition of functional contours prosody model,” in Speech Prosody, June 2018.

[12] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton, “Adaptive mixtures of local experts,” Neural computation, vol. 3, no. 1, pp. 79–87, 1991.

[13] F. Liu and Y. Xu, “Parallel encoding of focus and interrogative meaning in Mandarin intonation,” Phonetica, vol. 62, no. 2-4, pp. 70–87, 2005.

[14] Y. Xu, “Contextual tonal variations in Mandarin,” Journal of phonetics, vol. 25, no. 1, pp. 61–83, 1997.

[15] B. Gerazov, G. Bailly, and Y. Xu, “The significance of scope in modelling tones in Chinese,” in Tonal Aspects of Languages, June 2018.