Embedding Context-Dependent Variations of Prosodic Contours using Variational Encoding for Decomposing the Structure of Speech Prosody

Branislav Gerazov[1,2]
Gérard Bailly[2]
Omar Mohammed[2]
[1]FEEIT, UCMS, Skopje, Macedonia
[2]GIPSA-Lab, Grenoble, France

Yi Xu
UCL, London, UK

Philip N. Garner
Idiap, Martigny, Switzerland

Prosody in speech is used to communicate a variety of linguistic, paralinguistic and non-linguistic information via multiparametric contours. The Superposition of Functional Contours (SFC) model is capable of extracting the average shape of these elementary contours through iterative analysis-by-synthesis training of neural network contour generators (CGs) (Bailly and Holm, 2005). An example prosodic decomposition of the intonation contour for the French utterance "Son bagou pourrait faciliter la communauté." based on the annotated linguistic functions is shown in Fig. 1.

The Weighted SFC (WSFC) model is an extension to the SFC that can capture the prominence of each functional contour in the final prosody (Gerazov et al., 2018b). It does so through expanding the CGs with a weighting module that outputs a scaling factor based on their linguistic context. The WSFC has been shown to be able to successfully capture the impact of attitude and emphasis on prominence.

While the WSFC successfully captures gradience, the true spatio-temporal variance of these prosodic contours is multidimensional. To this effect, we recently proposed a Variational Prosody Model (VPM) that is able to capture a part of this variance (Gerazov et al., 2018a). Its variational CGs (VCGs) use the linguistic context input to map out a prosodic latent space for each contour. This two-dimensional latent space can then be used to visualise the captured context-specific variation. Since the VCGs are still based on synthesising the contours based on rhythmic unit position input, the mapped prosodic latent space is amenable for exploration only for short contours, such as Chinese tones or clitics, shown in Fig. 2.

Here we propose an extension on the VPM based on variance embedding and recurrent neural network contour generators (VRCGs). In our new approach, we use a variational encoder to embed the context-dependent variance in a latent space that is used to initialise a long short term memory (LSTM). The LSTM then uses rhythmic unit positions to generate the prosodic contour. This approach decouples the prosodic latent space from the length of the contour's scope, thus it can now be readily explored even for longer contours. Fig. 3 shows the embedded variance in the prosodic latent space of the left-dependency contour solicited in 6 different attitudes. We can clearly see that the declaration and especially exclamation attitudes give a full contour realisation, while the other induce its suppression.

References

[Bailly and Holm2005] Gérard Bailly and Bleicke Holm. 2005. SFC: a trainable prosodic model. *Speech communication*, 46(3):348–364.

[Gerazov and Bailly2018] Branislav Gerazov and Gérard Bailly. 2018. PySFC – a system for prosody analysis based on the superposition of functional contours prosody model. In *Speech Prosody*, June.

[Gerazov et al.2018a] Branislav Gerazov, Gérard Bailly, Omar Mohammed, Yi Xu, and Philip N. Garner. 2018a. A variational prosody model for the decomposition and synthesis of speech prosody. In *ArXiv e-prints https://arxiv.org/abs/1806.08685*, June.

[Gerazov et al.2018b] Branislav Gerazov, Gérard Bailly, and Yi Xu. 2018b. A weighted superposition of functional contours model for modelling contextual prominence of elementary prosodic contours. In *INTERSPEECH*, Septembre.

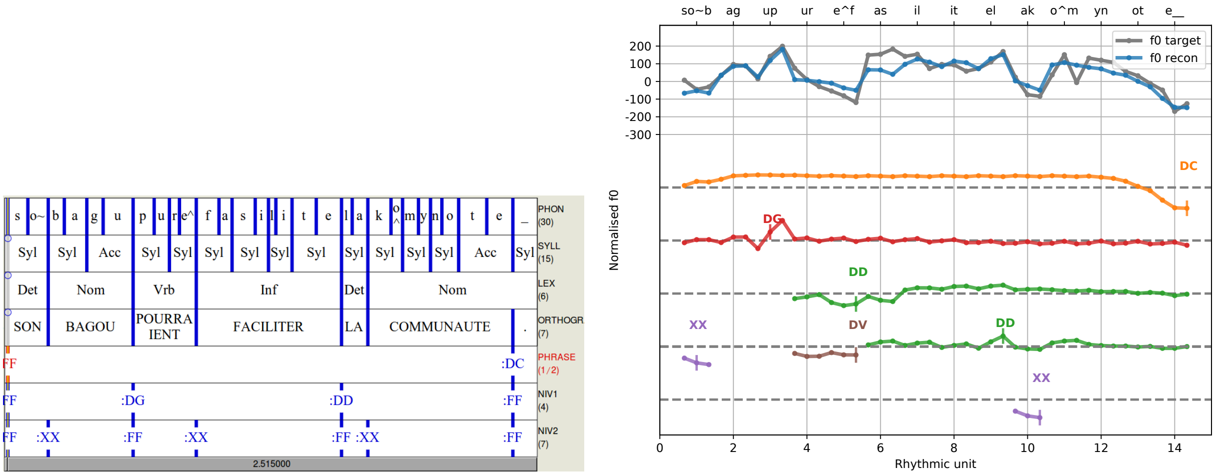

Figure 1: Example Praat annotation (left) and SFC decomposition (right) of the intonatino of the French utterance: “Son bagou pourrait faciliter la communauté.” The example shows the extracted elementary contours for the annotated linguistic functions: declaration (DC), dependency to the left/right (DG/DD), and cliticisation (DV, XX). Decomposition was done using the PySFC system, and the figures are taken from (Gerazov and Bailly, 2018).

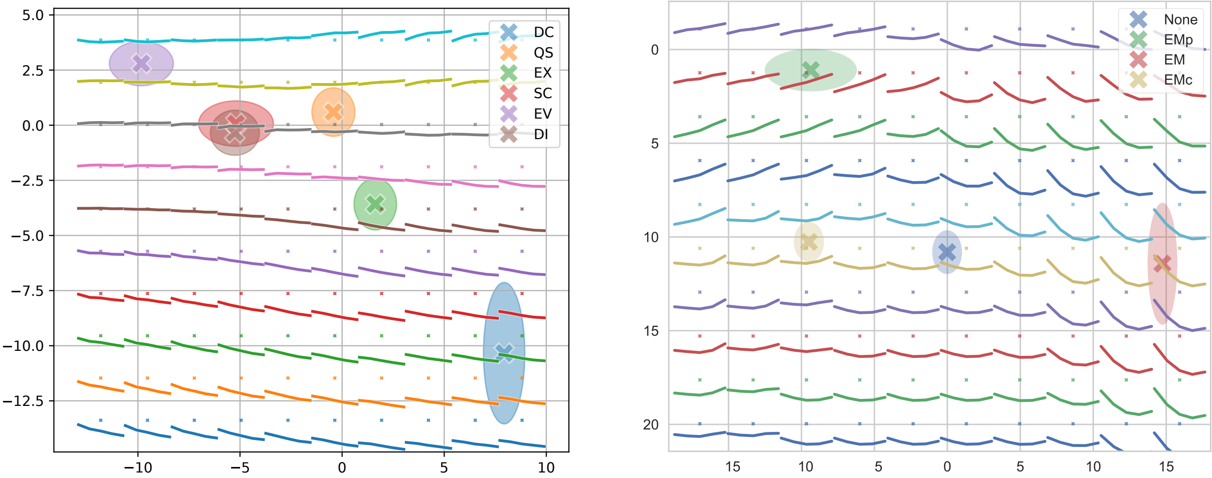

Figure 2: Structure of the prosodic latent space for the French clitic function contour XX dependent on the attitude context (left): declaration (DC), question (QS), incredulous question (DI), evidence (EV), suspicion (SC), and exclamation (EX); DC and EX only elicit a full-blown contour. Prosoodic latent space of Chinese tone 3 dependent on the emphasis context (right): no (none), pre- (EMp), on- (EM), and post-emphasis (EMc); on-emphasis the tone has pronounced prominence. Figures taken from (Gerazov et al., 2018a).

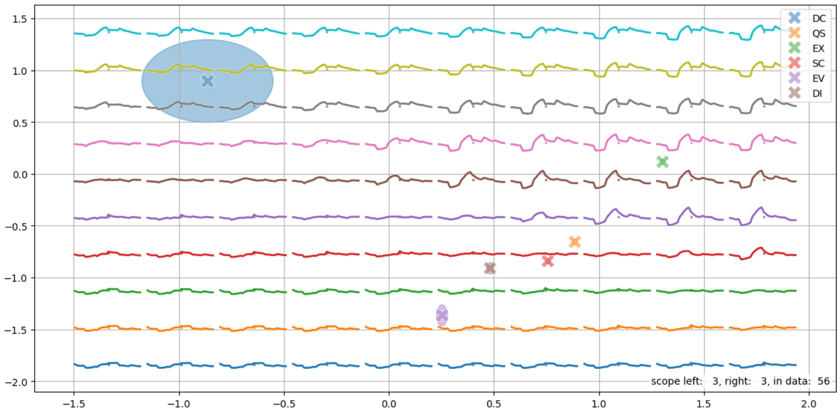

Figure 3: Prosodic latent space of left-dependency function contour (DG) structured based on attitude context with attitude codes same as in Fig. 2; again DC and EX elicit full-blown contours, with EX inducing larger contour prominence.