A Neural Model to Predict Parameters for a Generalized Command Response Model of Intonation

Bastian Schnell[1,2], *Philip N. Garner*[1]

[1]Idiap Research Institute, Martigny, Switzerland
[2]Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland
{bastian.schnell, phil.garner}@idiap.ch

This abstract summarizes our paper accepted in the main conference with the same title.

Index Terms: Speech synthesis, prosody modelling, recurrent neural network, Fujisaki model

1. Abstract

We are interested in general in speech to speech translation, and specifically in transfer of paralinguistics from one language to another. For instance, if a speaker expresses emotion or emphasis in an input language, we would like those features to be present in the synthetic speech resulting from machine translation of speech recognition output. In previous work with colleagues [1], we studied a model of prosody (actually intonation, F_0) based on the Command-Response (CR) model of Fujisaki [2]. By contrast to the CR model, this Generalised Command Response (GCR) model can be extracted easily from an intonation contour using a matching pursuit algorithm [3]. The time-local nature of its constituent *atoms* was shown (by design) to lend itself to transfer of emphasis. In particular, sections of intonation contours can be replaced with others that carry different meaning, all whilst retaining naturalness. We report on an investigation into how to use GCR to generate longer intonation contours for more general contour models. Of course, such contours can be generated by any modern Text-to-Speech (TTS) system. However, we hope to retain the transfer capability of the GCR, which also enables analysis of the underlying physiological process.

Given that GCR atoms approximate (groups of) muscle responses to neural spikes, it would make sense to use a Spiking Neural Network (SNN) to generate these spikes. The generated spikes would be filtered by muscle responses to generate the pitch contour. However, the choice of a spiking network paradigm is not obvious. Rather, given the authors' familiarity with conventional back-propagation based deep learning algorithms and toolkits, we emulate a SNN. In this work we use a bidirectional GRU-based Recurrent Neural Network (RNN) which is capable of generating spikes, hence atoms, for a given text. This in turn allows us to introduce a loss function for the training of spiking outputs which is inspired by losses in SNNs.

1.1. Relation to prior work

Numerous approaches to modelling prosody by the superposition of multiple F_0 contours exist. A common drawback of models like Tilt [6], INSINT [7], ToBI [8], or SFC [9] is, that they are not based on observations of the physiological production aspect. The proposed GCR model is a physiologically based intonation model which has the same representative power as the CR model of Fujisaki [2]. It generates the Log-F_0 (LF_0) contour by a superposition of impulse responses to critically damped second order systems modelling muscle responses in the glottis. The impulse response of a critically damped second order system is a gamma kernel

$$G_{k,\theta}(t) = \frac{1}{\theta^k \Gamma(k)} t^{k-1} e^{-t/\theta} \quad \text{for} \quad t \geq 0 \qquad (1)$$

with k being the system order, Γ being the gamma function, and θ determining the length of the kernel. For a critically damped second-order system as well as the CR model $k = 2$, however previous research [10] has found that $k = 6$ gives better approximations of the original LF_0 contour. A phrase additionally consists of a phrase atom (phrase command in CR model) which models the general shape of the contour and is correlated mainly to the physics of the speakers' lung volume. The work closest to our approach is that of Hojo et. al. [11] where the CR model is represented by a constrained HMM, and a Neural Network (NN) predicts the posteriori probability of its states. A Viterbi-like algorithm extracts the most probable sequence based on the posteriors. The LF_0 generation based on the sequence is then readily synthesised [12].

1.2. Atom Prediction

Rather than use an explicit spiking paradigm such as leaky integrate and fire (LIF), we instead emulate such a network using a conventional backpropagation network. This is achieved using a bidirectional RNN as described in [13]. Rather than use LSTMs with peepholes as in that paper, we use the GRU of [14] where peepholes are moot. For a regression task which targets spiking output of varying amplitudes the commonly used Mean-Squared-Error (MSE) is not an appropriate loss function as it does not consider any temporal information of spikes. The problem breaks down to measuring the distance between two spike trains. Various methods exist to compute such a distance such as the Victor-Purpura metric [15], and others [16, 17, 18, 19, 20, 21]. In general we are interested in a learning rule that uses such temporally-aware measurements to compute losses during training. We have not found a suitable learning rule in the literature for feed-forward NN or RNNs but instead in the field of SNNs. The closest precedent to the learning rule we propose is the SPAN method [5]. In SPAN, each spike is convolved with an "alpha" kernel which adds temporal information of the spike to all surrounding/succeeding frames. MSE can be used as the learning rule on the resulting continuous output. The authors of the SPAN method state that other kernel functions are possible such as Gaussian, linear and exponential kernels [22] . The choice of kernel in the literature is driven by the supposed shape of the post-synaptic potential of neurons in

This research was supported by SNSF Project number 165545, MASS: Multilingual Affective Speech Synthesis, http://p3.snf.ch/Project-165545.

the human brain. However, the spikes we are interested in represent muscle impulses with responses modelled by a gamma kernel as described above. We therefore use the gamma kernel as the kernel function. The length θ of the kernel is by no means obvious. While the desired length of correctly placed spikes is known, no ground truth is available for incorrectly placed spikes. We found that a single short kernel with $\theta = 0.01$ for all convolutions adds the required temporal information to each spike.

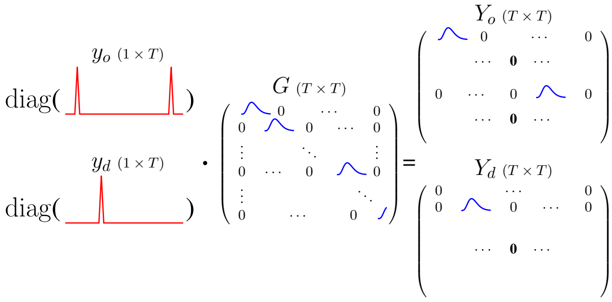

Figure 1: *Frame-wise convolution of NN output y_o and desired output y_d.*

Let us define the matrix G which has the coefficients of the gamma kernel on its leading and above leading diagonals with size $(T \times T)$ where T is the number of frames in a training sample. Further define y_o as the output of the NN and y_d as the desired output each of size $(T \times 1)$. All spikes can be convolved independently from each other with the kernel function by $\text{diag}(y) \cdot G = Y$ (compare Figure 1). We denote $\tilde{y}_d$ as the desired enveloped output given by the sum of all rows of Y_d which corresponds to a superposition of envelopes. The error at each time step t is computed by

$$err_t = \sum_{i=t}^{t+\Delta t} (Y_{o,t,i} - \tilde{y}_{d,i})^2 \qquad (2)$$

with $Y_{o,t,i}$ being the t-th row and i-th column of Y_o, and $\tilde{y}_{d,i}$ being the i-th entry in $\tilde{y}_d$. Δt is given by the length of the gamma kernel used to convolve each spike and represents the number of frames where a spike takes effect. To limit the interval of the sum to $[t, t + \Delta t]$ is critical so that the error is not affected by succeeding parts of the sequence where the spike cannot take effect. Note that the error is computed frame-wise without the superposition of the enveloped NN output, which means that neighbouring spikes cannot interfere. When allowing the interference of spikes two problems arise:

- The NN learns to represent a single target spike by multiple smaller spikes.
- The NN predicts many spikes with opposite amplitude which cancel out.

The former problem is an acceptable variation to our model when assuming that a muscle response is not triggered by a single nerve impulse but multiple ones. However, the latter problem gives clearly unintended and physiological implausible behaviour. In future research the addition of an L1 regularisation term over time on the NN outputs could solve the second problem. For the current model we use the frame-wise atom loss hereafter. A NN trained with the above learning rule gives an activation around each spike position and thus requires a post-processing step to identify the peaks.

Besides position, amplitude and length θ are required for every atom. A single position flag trained with the loss function introduced above gives good estimates of the position of spikes ($y_{d,t} \in \{-1, 0, 1\}$) but cannot predict amplitude and length at the same time. Unfortunately, we were not able to train a NN to predict a θ value directly. Instead, besides the position flag, the NN is trained with MSE to predict one amplitude per θ for a fixed set of θs. The set of θs needs enough values to allow an approximation of the target LF_0 contour with low error, but is limited, which corresponds to the limited number of articulators in the human larynx. To compensate the highly unbalanced training set we convolve each amplitude spike by a normal distribution in time and increase the weight of the loss on frames with non-zero target output, while decreasing it on all others. Our network also predicts a flag for Voiced/Unvoiced (V/UV) LF_0. The target V/UV flag is used to decrease the weight of both losses (atoms and amplitudes) on unvoiced frames, so that the network spends less effort on improving parts which are silent after synthesis.

1.3. Experiments

In the experiments, we mean to test the hypothesis that the basic procedure described above is a plausible approach to generate natural sounding intonation. A-priori we do not expect it to generate state-of-the-art intonation contours; rather, we simply aim to validate that the approach merits further research. We test our proposed model on a subset of the English speech database released for the 2008 Blizzard Challenge [23]. We only use those samples which can be represented by a single phrase atom. For all experiments the original durations, Mel Frequency Cepstral Coefficients (MFCC)s, and band aperiodicity (BAP)s are used as we are only interested in the impact of different LF_0s on naturalness. In our proposed model, the spike position flag is post-processed to identify its peaks which results in a value of $\{-1, 0, 1\}$ per frame. Atoms are constructed by taking the maximum of the nine predicted amplitudes for positive spikes and the minimum for negative spikes respectively. The θ value is implicitly given by the index of the selected amplitude within the nine outputs. LF_0 is reconstructed by superposition of all predicted atoms and the original phrase atom. We plan to predict the phrase atom as well in the future.

On objective scores (RMSE of F_0, V/UV error rate) our model preforms slightly worse than the baseline system (44.46 Hz, 5.43% to 49.89 Hz, 5.94%), but certainly close enough to validate our hypothesis that the approach is plausible. We additionally measured the naturalness of the synthesised speech by a MUSHRA test, where we compared our model, a baseline, and the speech produced by the vocoder with the original acoustic features. The results support our assumption that our and the baseline system are not significantly different.

1.4. Conclusions

We have shown that the combination of an emulated spiking network, a dictionary of atoms representing muscle responses, and a SPAN-inspired training algorithm can generate reasonable intonation contours. Although "reasonable" is open to interpretation, the algorithm produces subjective results that are not significantly different from an accepted baseline. The proposed training algorithm for spiking targets enables the use of DNNs in other research fields currently dominated by SNNs. Future work includes the prediction of phrase atoms, exploiting the capabilities of the GCR model to produce / transfer affect, and reducing the number of heuristics identifying hyper-parameters.

2. References

[1] P.-E. Honnet, B. Gerazov, and P. N. Garner, "Atom decomposition-based intonation modelling," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 4744–4748.

[2] H. Fujisaki, S. Ohno, and C. Wang, "A command-response model for F0 contour generation in multilingual speech synthesis," in *The Third ESCA/COCOSDA Workshop (ETRW) on Speech Synthesis*, 1998. [Online]. Available: http://www.isca-speech.org/archive_open/archive_papers/ssw3/ssw3_299.pdf

[3] S. G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Transactions on signal processing*, vol. 41, no. 12, pp. 3397–3415, 1993.

[4] Z. Wu, O. Watts, and S. King, "Merlin: An open source neural network speech synthesis system," *Proc. SSW, Sunnyvale, USA*, 2016.

[5] A. Mohemmed, S. Schliebs, S. Matsuda, and N. Kasabov, "Training spiking neural networks to associate spatio-temporal input–output spike patterns," *Neurocomputing*, vol. 107, pp. 3–10, 2013.

[6] P. Taylor, "Analysis and synthesis of intonation using the tilt model," *The Journal of the acoustical society of America*, vol. 107, no. 3, pp. 1697–1714, 2000.

[7] D. Hirst, A. Di Cristo, and R. Espesser, "Levels of representation and levels of analysis for the description of intonation systems," in *Prosody: Theory and experiment*. Springer, 2000, pp. 51–87.

[8] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg, "Tobi: A standard for labeling english prosody," in *Second international conference on spoken language processing*, 1992. [Online]. Available: http://www.isca-speech.org/archive/icslp_1992/i92_0867.html

[9] G. Bailly and B. Holm, "SFC: a trainable prosodic model," *Speech communication*, vol. 46, no. 3-4, pp. 348–364, 2005.

[10] B. Gerazov, P.-E. Honnet, A. Gjoreski, and P. N. Garner, "Weighted correlation based atom decomposition intonation modelling," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015. [Online]. Available: http://publications.idiap.ch/index.php/publications/show/3145

[11] N. Hojo, Y. Ohsugi, Y. Ijima, and H. Kameoka, "DNN-SPACE: DNN-HMM-based generative model of voice f0 contours for statistical phrase/accent command estimation," *Proc. Interspeech 2017*, pp. 1074–1078, 2017.

[12] H. Kameoka, K. Yoshizato, T. Ishihara, K. Kadowaki, Y. Ohishi, and K. Kashino, "Generative modeling of voice fundamental frequency contours," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 6, pp. 1042–1053, 2015.

[13] F. A. Gers, N. N. Schraudolph, and J. Schmidhuber, "Learning precise timing with lstm recurrent networks," *Journal of machine learning research*, vol. 3, no. Aug, pp. 115–143, 2002.

[14] K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," *arXiv preprint arXiv:1409.1259*, 2014.

[15] J. D. Victor and K. P. Purpura, "Metric-space analysis of spike trains: theory, algorithms and application," *Network: computation in neural systems*, vol. 8, no. 2, pp. 127–164, 1997.

[16] M. v. Rossum, "A novel spike distance," *Neural computation*, vol. 13, no. 4, pp. 751–763, 2001.

[17] S. Schreiber, J.-M. Fellous, D. Whitmer, P. Tiesinga, and T. J. Sejnowski, "A new correlation-based measure of spike timing reliability," *Neurocomputing*, vol. 52, pp. 925–931, 2003.

[18] J. D. Hunter and J. G. Milton, "Amplitude and frequency dependence of spike timing: implications for dynamic regulation," *Journal of neurophysiology*, vol. 90, no. 1, pp. 387–394, 2003.

[19] R. Q. Quiroga, T. Kreuz, and P. Grassberger, "Event synchronization: a simple and fast method to measure synchronicity and time delay patterns," *Physical review E*, vol. 66, no. 4, p. 041904, 2002.

[20] J. Dauwels, F. Vialatte, T. Rutkowski, and A. S. Cichocki, "Measuring neural synchrony by message passing," in *Advances in neural information processing systems*, 2008, pp. 361–368. [Online]. Available: http://papers.nips.cc/paper/3322-measuring-neural-synchrony-by-message-passing.pdf

[21] C. V. Rusu and R. V. Florian, "A new class of metrics for spike trains," *Neural Computation*, vol. 26, no. 2, pp. 306–348, 2014.

[22] F. Ponulak and A. Kasiński, "Supervised learning in spiking neural networks with resume: sequence learning, classification, and spike shifting," *Neural computation*, vol. 22, no. 2, pp. 467–510, 2010.

[23] V. Karaiskos, S. King, R. A. Clark, and C. Mayo, "The blizzard challenge 2008," in *Proc. Blizzard Challenge Workshop, Brisbane, Australia*, 2008. [Online]. Available: http://www.festvox.org/blizzard/bc2008/summary_Blizzard2008.pdf