An End-to-End Network to Synthesize Intonation using a Generalized Command Response Model

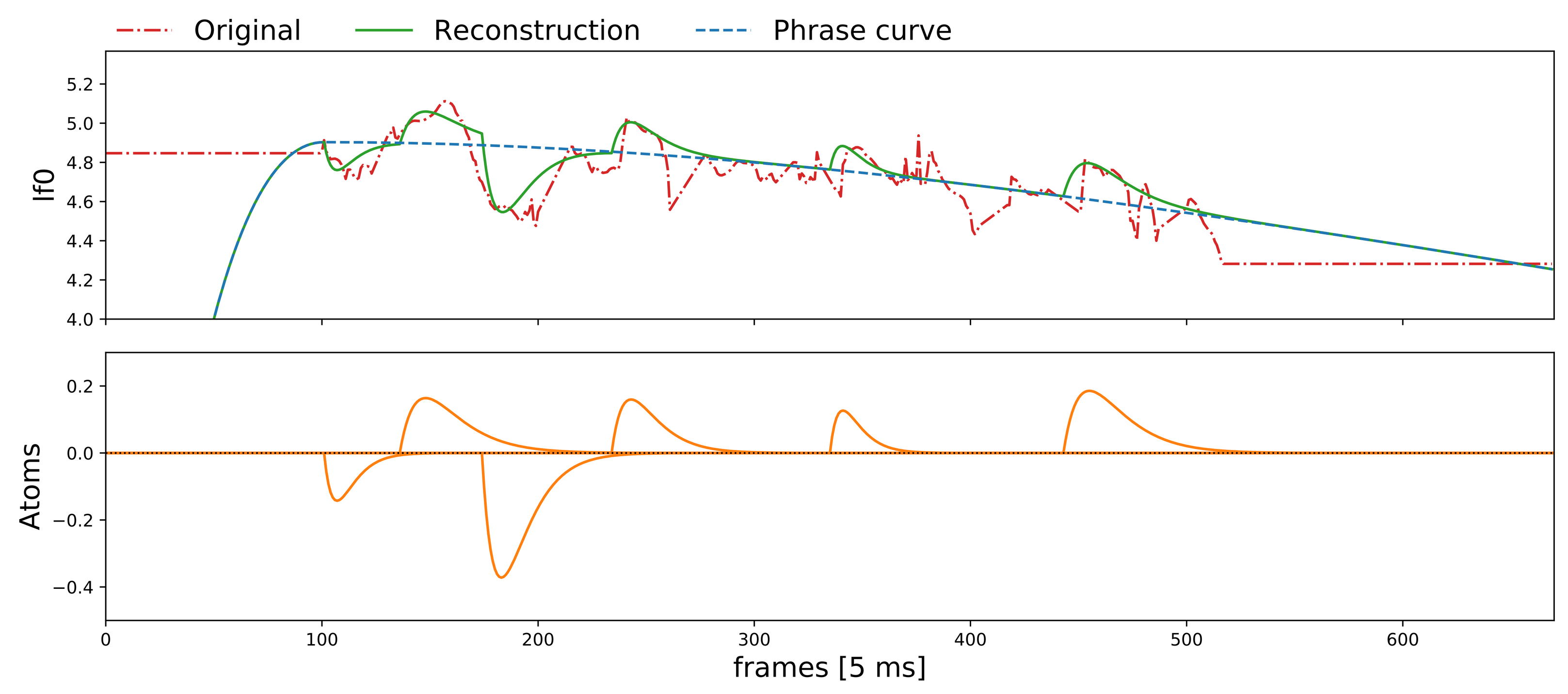
François Marelli[1,2,3], Bastian Schnell[1,2], Hervé Bourlard[1,2], Thierry Dutoit[3], Philip N. Garner[1]

[1]Idiap Research Institute, Martigny, Switzerland [2]École Polytechnique Fédérale de Lausanne (EPFL), Switzerland [3]Université de Mons (UMONS), Institut NUMEDIART, Belgium

{francois.marelli, bastian.schnell, phil.garner}@idiap.ch

Introduction

- Correctly handling intonation is crucial in speech synthesis, for both the perceived naturalness and the conveyed meaning of a sentence.
- The Generalized Command Response Model (GCR) [1] represents the intonation contour (LF_0) as a phrase component and a superposition of muscle responses to spike command signals.
- In this work, we propose an end-to-end neural network trained to synthesize pitch by reproducing the GCR model behaviour.
- We introduce trainable linear second-order recurrent units for muscle modelling, and demonstrate gradient stability under modest conditions.
- The system achieves subjective scores matching a state-of-the-art baseline.



Modelling intonation with GCR: pitch reconstruction (top), impulse responses (bottom).

Why GCR?

- Consistent with Fujisaki's Command-Response Model [2].
- Physiologically inspired from glottal muscles, and interpretable.
- Allows the (cross-language) transfer of emphasis at word-level.

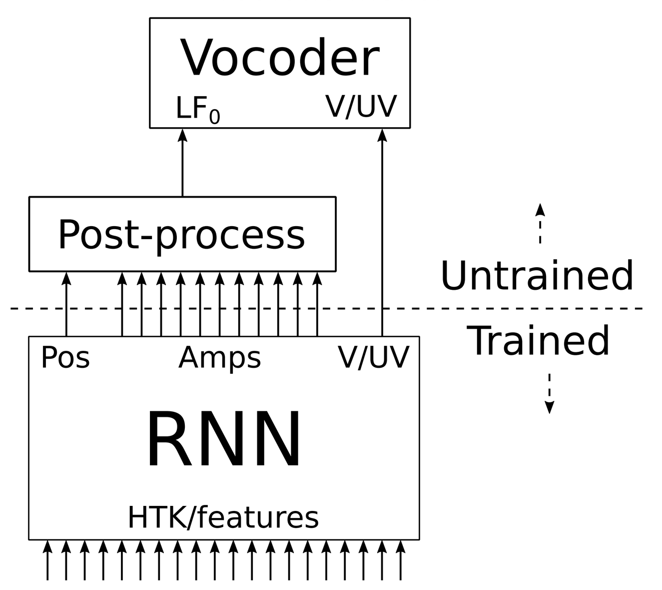
End-to-End Model

Previous work proposes a RNN to emulate the spike generation [3]. This method requires hardcoded post-processing steps and omits the phrase component.

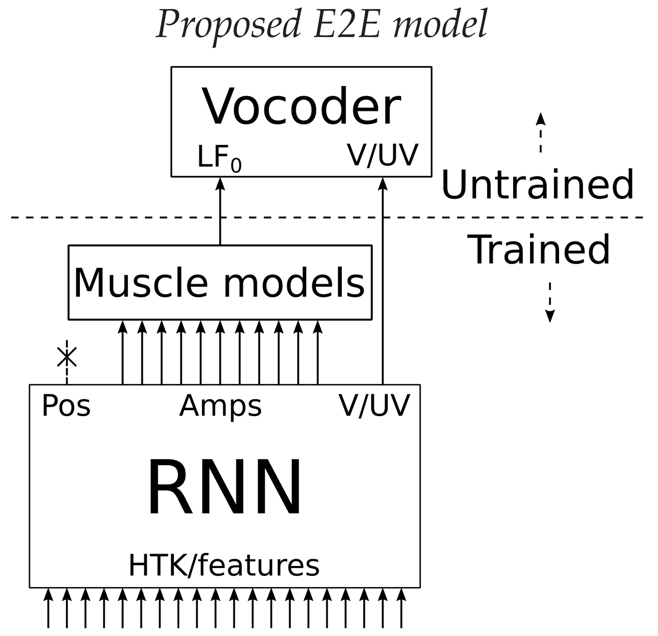
The proposed End-to-End (E2E) architecture offers the following:

- Trainable muscle parameters.
- Phrase component generation.

Previous model

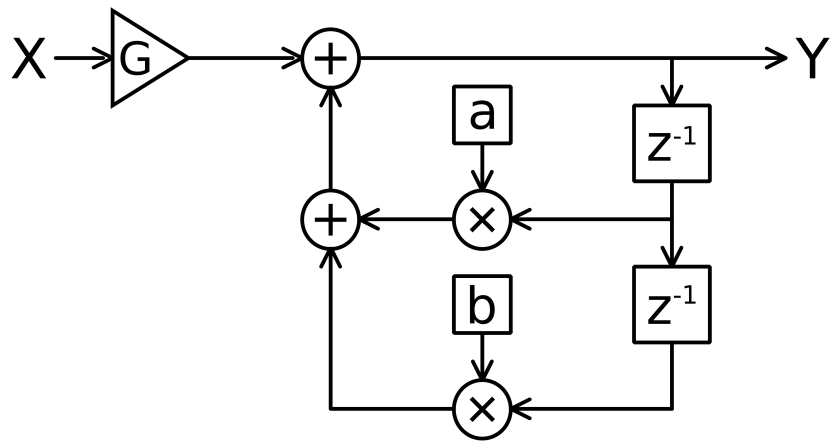


Proposed E2E model



Muscle Models

Muscle responses can be modeled using second-order linear recurrent systems.



Generic discrete transfer function:

$$y_{(k)} = Gx_{(k)} + \alpha y_{(k-1)} + \beta y_{(k-2)}$$

The gradients are computed for training through back-propagation:

$$\frac{\partial y_{(k)}}{\partial \alpha} = \sum_{n=0}^{k-1} [y_{(k-1-n)} \cdot K_n]$$

$$\frac{\partial y_{(k)}}{\partial \beta} = \sum_{n=0}^{k-2} [y_{(k-2-n)} \cdot K_n]$$

$$\frac{\partial y_{(k)}}{\partial x_{(k-n)}} = GK_n$$

$$K_n = \begin{cases} \alpha K_{n-1} + \beta K_{n-2} & \text{if } n > 0 \\ 1 & \text{if } n = 0 \\ 0 & \text{if } n < 0 \end{cases}$$

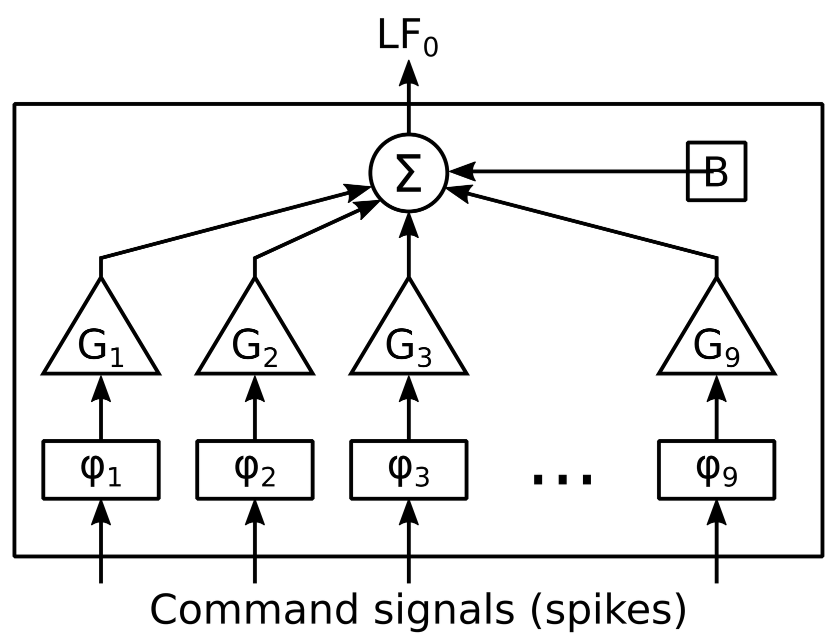
- The recurrence in K_n causes gradient explosion, preventing convergence.
- Under the assumption that muscle models have an under-damped behaviour, the transfer function can be expressed in polar notation. A compressing transform is then used to constrain it and guarantee the gradient stability.

$$y_{(k)} = Gx_{(k)} + 2\rho\cos(\phi)\, y_{(k-1)} - \rho^2\, y_{(k-2)}$$

$$y_{(k)} = Gx_{(k)} + 2\,\sigma(p)\,\tanh(c)\, y_{(k-1)} - \sigma^2(p)\, y_{(k-2)}$$

Model Architecture

- The output layer of the network is composed of a set of muscle models (φ_i).
- Each unit is multiplied by a normalization gain before summation.
- A speaker-dependent bias is added to enable phrase component modelling.

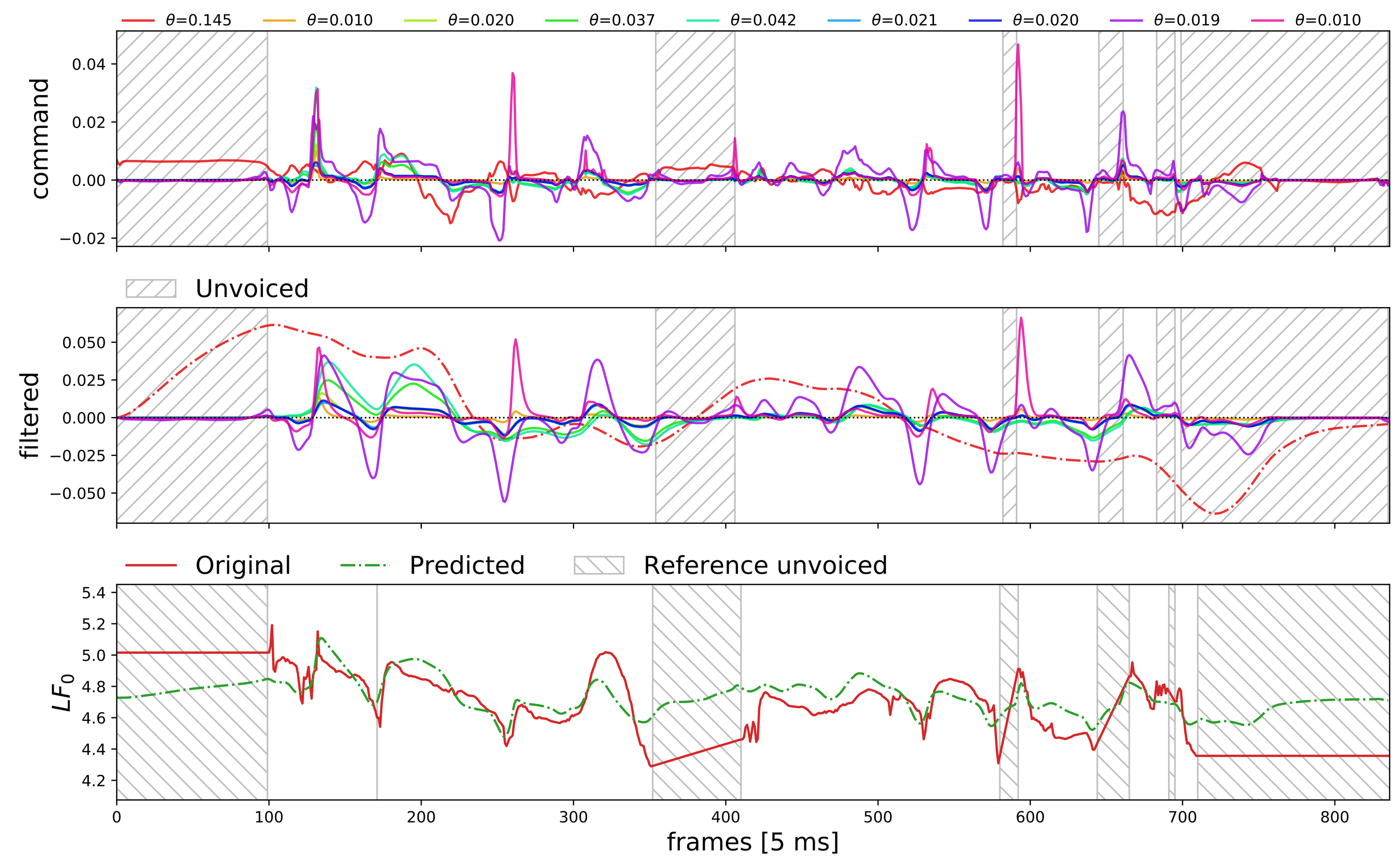


Acknowledgment

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

Results

Trained model behaviour

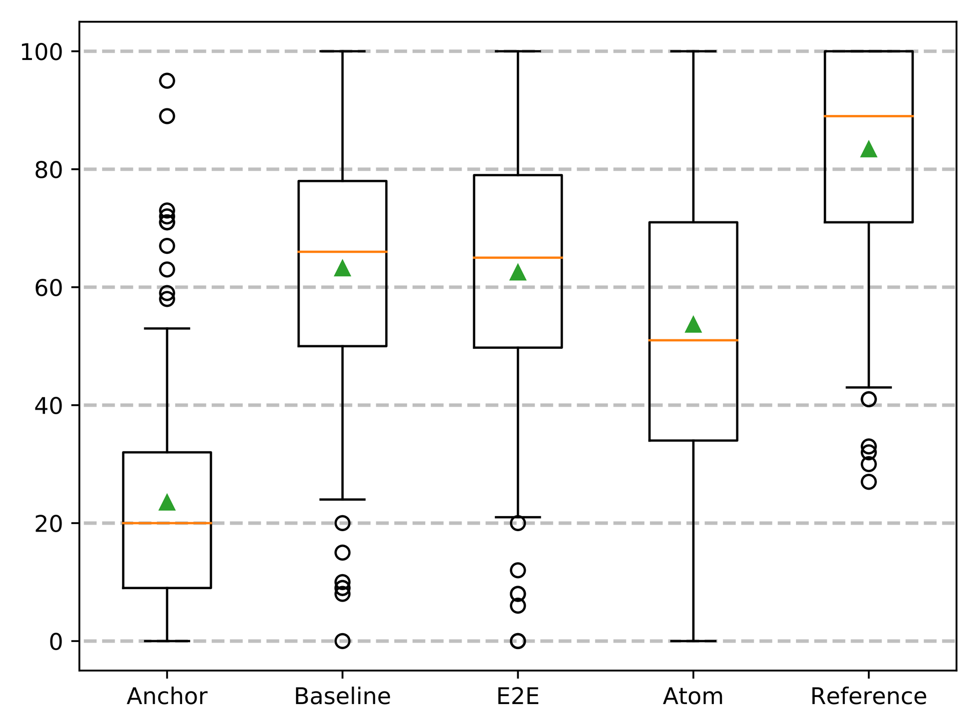


- Spiky command signals are obtained by applying L1 regularization over time on the inputs of the muscle models.
- The filtered commands (muscle outputs) show that the phrase component is modelled by a slow moving filter (red dash-dotted line).
- Objective and subjective scores show that the synthesized LF_0 improves on the previous model (Atom) and matches the quality of a strong baseline.

Objective scores

Model	F_0 RMSE	V/UV error
Baseline	21.3 Hz	10.4 %
Atom	28.8 Hz	14.9 %
E2E	22.3 Hz	10.7 %

Subjective scores



- The proposed model takes advantage of the flexibility of the E2E architecture, while retaining the properties and behaviour of the GCR model.
- Further work would include a psycho-linguistic analysis of the model, and the investigation of its exploitation in emotional speech synthesis.

References

[1] Pierre-Edouard Honnet, Branislav Gerazov, and Philip 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] Hiroya Fujisaki, Sumio Ohno, and Changfu Wang, "A command-response model for F0 contour generation in multilingual speech synthesis," in *The Third ESCA/COCOSDA Workshop (ETRW) on Speech Synthesis*, 1998.

[3] Bastian Schnell and Philip N Garner, "A neural model to predict parameters for a generalized command response model of intonation," *Proc. Interspeech 2018*, pp. 3147–3151, 2018.