**AN END-TO-END NETWORK TO SYNTHESIZE INTONATION USING A GENERALIZED COMMAND RESPONSE MODEL**

FRANÇOIS MARELLI¹23, BASTIAN SCHNELL¹2, HERVE BOURLARD¹2, THIERRY DUTOIT³, PHILIP N. GARNER¹

¹Idiap Research Institute, Martigny, Switzerland ²École Polytechnique Fédérale de Lausanne (EPFL), Switzerland ³Université 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 to synthesize pitch by reproducing the GCR model behaviour.
- We introduce trainable 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.

**Muscle Models**

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

\[
\frac{\partial y_{(i)}}{\partial x} = \sum_{k=0}^{N} K_{i,k} \frac{\partial y_{(i)}}{\partial K_{k,i}} \quad \text{if } n \neq 0
\]

\[
\frac{\partial y_{(i)}}{\partial x} = \sum_{k=0}^{N} K_{i,k} \frac{\partial y_{(i)}}{\partial K_{k,i}} \quad \text{if } n = 0
\]

\[
\frac{\partial y_{(i)}}{\partial x} = G K_{i} \quad \text{if } n < 0
\]

- The gradient is computed for training through back-propagation:
- 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.

\[
y_{(i)} = G \phi(0) \phi(y_{(i-1)} - p) y_{(i-2)}
\]

\[
y_{(i)} = G \phi(0) \phi(y_{(i-1)} - p) y_{(i-2)}
\]

**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.

**Model Architecture**

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

**Results**

- 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.

**REFERENCES**


**ACKNOWLEDGMENT**

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