7605-5/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICASSP39728.2021.9413400

SEMI-SUPERVISED LEARNING FOR SINGING SYNTHESIS TIMBRE

Jordi Bonada, Merlijn Blaauw

Music Technology Group, Universitat Pompeu Fabra, Barcelona, Spain

jordi.bonada@upf.edu, merlijn.blaauw@upf.edu

ABSTRACT

We propose a semi-supervised singing synthesizer, which is able to learn new voices from audio data only, without any annotations such as phonetic segmentation. Our system is an encoder-decoder model with two encoders, linguistic and acoustic, and one (acoustic) decoder. In a first step, the system is trained in a supervised manner, using a labeled multi-singer dataset. Here, we ensure that the embeddings produced by both encoders are similar, so that we can later use the model

for this is that we consider both tasks to have notably different requirements and constraints, and thus prefer to focus on each step individually. We will tackle the task of semi-supervised training of a pitch model in a separate paper.

The principal contributions of this paper are: 1. A semi-supervised method for learning the timbre of new voices from audio data only. 2. A method for controlling a single decoder with embeddings derived from either acoustic or linguistic features. 3. Distinction between long and short scopes in the decoder for tackling the phonetic context of long vowels in