A Demonstration of the Merlin Open Source Neural Network Speech Synthesis System

Srikanth Ronanki, Zhizheng Wu, Oliver Watts, Simon King

The Centre for Speech Technology Research, University of Edinburgh, United Kingdom

Abstract

This demonstration showcases our new Open Source toolkit for neural network-based speech synthesis, Merlin. We wrote Merlin because we wanted free, simple, maintainable code that we understood. No existing toolkits met all of those requirements. Merlin is designed for speech synthesis, but can be put to other uses. It has already also been used for voice conversion, classification tasks, and for predicting head motion from speech.
Index Terms: Merlin, speech synthesis, deep learning.

1. Introduction

This paper is the companion for a demonstration of our new neural network speech synthesis toolkit, named Merlin, described in the full SSW paper[1].

2. Design philosophy

Merlin[2] is not intended to include all functions necessary to construct a text-to-speech system. It requires a front-end and a vocoder, but is not hardwired to any particular ones. To train a DNN also requires data with aligned labels.

Front-end Merlin requires an external front-end, such as Festival or Ossian, to provide the input for the DNN. Any front-end can be used. Its output needs to be formatted as HTS-style labels with either phone-level or state-level alignment. The toolkit includes functions to convert such labels into sequences of vectors of binary and/or continuous features. These features are derived from the label files using HTS-style questions, with a small extension to enable continuously-valued features to be derived. It is also possible to directly provide already-vectorised input features if this HTS-like workflow is not convenient.

Vocoder Currently, Merlin supports two vocoders: STRAIGHT and WORLD. The Merlin distribution includes a modified version of the WORLD vocoder, with separate analysis and synthesis executables. Merlin supports both fixed and variable (e.g., pitch synchronous) frame rates.

Data HTK or HTS can be used to obtain state-level alignments for the training data. Merlin can also use only phone-level alignments, which can be found using other toolkits such as festvox clustergen.

Duration modelling Merlin models duration using a separate DNN to the acoustic model. The duration model is trained on the aligned data, to predict phone- and/or state-level durations. At synthesis time, duration is predicted first, and is used as an input to the acoustic model to predict the speech parameters.

3. ARCTIC recipe

Merlin currently comes with one 'recipe', but we will add over time, and we hope users will also contribute. In the spirit of Kaldi, we intend recipes to accompany published work, making it easily reproducible.

The first recipe is a voice demo using the 'slt' data from ARCTIC. The recipe uses entirely Open Source components: CMU dictionary, radio phoneset, WORLD vocoder. It uses a feed-forward DNN architecture of 6 layers each comprising 1024 nodes with tanh activation. The same architecture is used in both the duration acoustic models, for simplicity.

It is easy to change the architecture, and recurrent layers are supported (e.g., RNN, LSTM, and GRU). Maximum likelihood parameter generation (MLPG) using pre-computed variances from the training data is applied to the output features for synthesis, and post-filtering is applied to the resulting trajectories. These things are user-configurable.

4. Testing and benchmarking

The current version of Merlin is already in use by several collaborators and has been used (i.e., tested!) in 10+ Masters projects[3], combined with either Festival or Ossian, and STRAIGHT or WORLD. It also been tested with multiple languages including US & UK English, Polish, Italian, Chilean Spanish, the Jinhua dialect of Chinese, Hindi, Telugu and Tamil.

Blizzard 2016 Merlin was used as the DNN benchmark system for the 2016 Blizzard Challenge, with the Ossian front-end and WORLD vocoder (both Open Source) to provide an easily-reproducible system[4]. In the results for naturalness, speaker similarity and intelligibility, Merlin outperformed the HMM-based benchmark (Festival, HTS and STRAIGHT).

The next release of Festival should include a DNN-guided hybrid version of its Multisyn unit selection engine, which uses Merlin. This system was used for CSTR's Blizzard Challenge entry, and gave very satisfying performance.

Acknowledgements: This research was supported by EPSRC Programme Grant EP/I031022/1, Natural Speech Technology (NST). The NST research data collection may be accessed at `http://hdl.handle.net/10283/786`.

[1] Zhizheng Wu, Oliver Watts, Simon King, "Merlin: An Open Source Neural Network Speech Synthesis System" in Proc. 9th ISCA Speech Synthesis Workshop (SSW9), September 2016, Sunnyvale, CA

[2] Like Festival, Merlin is named after a Scottish beer. The beer is named after Merlin the wizard who, like DNNs, can do magic.

[3] Thanks to our MSc Speech & Language Processing students.

[4] Except that this year, we used our own proprietary dictionary. In future, we will select an Open Source dictionary.