Robust text-to-speech duration modelling using DNNs

Gustav Eje Henter, Srikanth Ronanki, Oliver Watts, Mirjam Wester, Zhizheng Wu, Simon King

The Centre for Speech Technology Research (CSTR), The University of Edinburgh, UK

gustav.henter@ed.ac.uk

1. Abstract

Accurate modelling and prediction of speech-sound durations is an important component in generating more natural synthetic speech. Deep neural networks (DNNs) offer a powerful modelling paradigm, and large, found corpora of natural and prosodically-rich speech are easy to acquire for training DNN models. Unfortunately, poor quality control (e.g., transcription errors) as well hard-to-predict phenomena such as reductions and filled pauses are likely to complicate duration modelling from found data. To mitigate issues caused by these idiosyncrasies, we propose to improve modelling and prediction of speech durations using methods from *robust statistics*. These are able to disregard ill-fitting points in the training material – errors or other outliers – in order to describe the typical case better. For instance, parameter estimation can be made robust by changing from maximum likelihood estimation (MLE) to a robust fitting criterion based on the density power divergence (a.k.a. the β-divergence) [1, 2]. Alternatively, the standard approximations for output generation with multi-component mixture density networks (MDNs) [3] can be seen as a heuristic for robust output generation.

To evaluate the potential benefits of robust techniques, we used 175 minutes of found data from a free audiobook to build several text-to-speech (TTS) systems, described in Table 1, with either conventional or robust DNN-based duration prediction. The objective results, in Figure 1, indicate that robust methods described typical speech durations better than the baselines. (Atypical, poorly predicted durations may be due to transcription errors, known to exist also in the test data, that make some FRC durations unreliable.) Similarly, subjective evaluation using a hybrid MUSHRA/preference test with 21 listeners, each scoring 18 sets of same-sentence stimuli, found that listeners significantly preferred synthetic speech generated using robust methods over the baselines, as shown in Figure 2.

Label	Duration prediction method	Robust?	Label	Duration prediction method	Robust?
VOC	Vocoded speech (top line waveform)	-	MLE1	Gaussian MLE-fitted DNN (baseline)	no
FRC	Oracle durations from forced alignment against held-out speech	-	MLE3	3-component deep Gaussian MDN only synthesising from the heaviest component	yes
BOT	Monophone mean duration (bottom line)	no	B75	Gaussian DNNs fit using β-divergence, tuned to ignore $\approx$ 25 or 50% of datapoints	yes
MSE	Minimum mean-square error (baseline)	no	B50		yes

Table 1: *TTS systems in evaluation. Except for vocoded speech, all used the same DNN acoustic model but different duration predictors.*

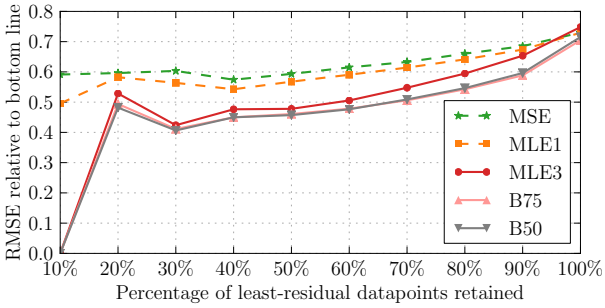

Figure 1: *Relative RMSE (frames per phone) between predicted and forced-aligned (FRC) durations on progressively larger and less well explained test-data subsets. Performance is normalised to place BOT at 1.0. Robust systems (solid) outperform non-robust baselines (dashed) on the majority of datapoints.*

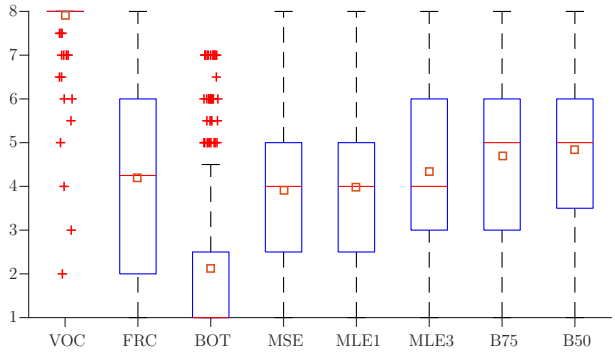

Figure 2: *Aggregated ranks (higher is better) from listening test. Red lines are medians, orange squares means; box edges are at 25 and 75% quantiles. Again, robust methods trump baselines.*

2. References

[1] A. Basu, I. R. Harris, N. L. Hjort, and M. C. Jones, "Robust and efficient estimation by minimising a density power divergence," *Biometrika*, vol. 85, no. 3, pp. 549–559, 1998.

[2] S. Eguchi and Y. Kano, "Robustifying maximum likelihood estimation," Institute of Statistical Mathematics, Tokyo, Japan, Tech. Rep. Research Memo 802, June 2001. [Online]. Available: http://www.ism.ac.jp/~eguchi/pdf/Robustify_MLE.pdf

[3] H. Zen and A. Senior, "Deep mixture density networks for acoustic modeling in statistical parametric speech synthesis," in *Proc. ICASSP*, 2014, pp. 3844–3848.