Robust text-to-speech duration modelling with a deep neural network

Gustav Eje Henter[1], Srikanth Ronanki[2], Oliver Watts[2], Mirjam Wester[2], Zhizheng Wu[2], Simon King[2]

[1]Digital Content and Media Sciences, National Institute of Informatics, Tokyo, Japan [2]Centre for Speech Technology Research, University of Edinburgh, UK

Outline

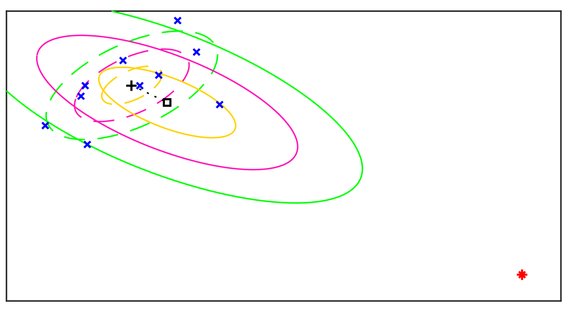
- Statistical speech synthesis is sensitive to bad data and bad assumptions
- We propose using techniques from robust statistics to reduce this sensitivity
  - This will be important on big and found datasets
- Robust techniques synthesise improved durations from found audiobook data
  - Paper [1] presented at ICASSP 2016

Why duration modelling?

- Duration is a major component in natural speech prosody
- Current duration models are weak and unconvincing
- Engineering approach: Throw data and computation (DNNs) at the task!
  - Problem: Big/found speech corpora have poor quality control
  - Problem: Our models are wrong – durations are skewed and non-Gaussian

Sensitivity of conventional approach

- Standard maximum likelihood estimation (MLE) is sensitive to unexpected data behaviour
- Gaussian toy data $\times$ with outlier $*$
  - Outlier can be error or genuine
  - Gaussian fit changes a lot with and without outlier! (solid vs. dashed)



- Robust statistics allow "giving up": Ill-fitting datapoints can be disregarded
  - This gives a better model of the typical case (high-density regions)
  - "Robust speech synthesis"

DNN duration prediction

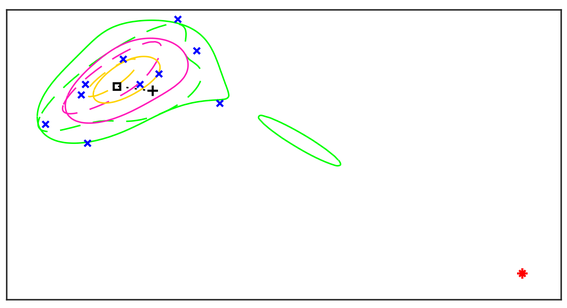
- Assume phone durations $\boldsymbol{d}$ are independent and GMM distributed

$$f_{\boldsymbol{D}}(\boldsymbol{d};\, \boldsymbol{\theta}) = \sum_{k=1}^{K} \omega_k \cdot f_{\mathcal{N}}(\boldsymbol{d};\, \boldsymbol{\mu}_k,\, \mathrm{diag}(\boldsymbol{\sigma}_k^2))$$

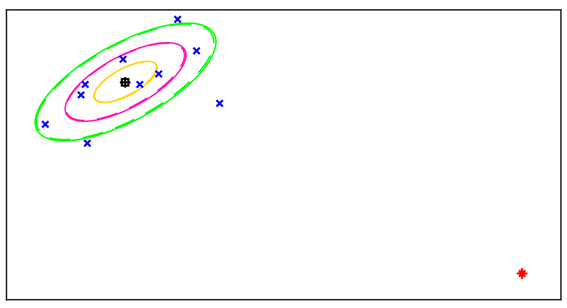
  - Setting $K = 1$ yields a conventional Gaussian DNN duration model
- Distribution parameters $\boldsymbol{\theta} = \{\omega_k,\, \boldsymbol{\mu}_k,\, \boldsymbol{\sigma}_k^2\}_{k=1}^{K}$ depend on linguistic features $\boldsymbol{l}$ through a DNN $\boldsymbol{\theta}(\boldsymbol{l};\, \boldsymbol{W})$ with weights $\boldsymbol{W} \Rightarrow$ a mixture density network (MDN)
- Conventional, non-robust MLE parameter estimation for data $\mathcal{D} = \{\boldsymbol{d}_p,\, \boldsymbol{l}_p\}$

$$\widehat{\boldsymbol{W}}_{\mathrm{ML}} = \operatorname*{argmax}_{\boldsymbol{W}} \sum_{p \in \mathcal{D}} \ln f_{\boldsymbol{D}}(\boldsymbol{d}_p;\, \boldsymbol{\theta}(\boldsymbol{l}_p;\, \boldsymbol{W}))$$

Achieving robustness



2-component GMM



β-estimated Gaussian ($\beta = 1/3$)

- Robust output generation: Component selection in MDNs

$$k_{\max}(\boldsymbol{l}) = \operatorname*{argmax}_{k} \omega_k(\boldsymbol{l})$$

$$\widehat{\boldsymbol{d}}(\boldsymbol{l}) = \operatorname*{argmax}_{\boldsymbol{d}} f_{\mathcal{N}}(\boldsymbol{d};\, \boldsymbol{\mu}_{k_{\max}}(\boldsymbol{l}),\, \mathrm{diag}(\boldsymbol{\sigma}^2_{k_{\max}}(\boldsymbol{l})))$$

  - This is robust if $K > 1$ (some components/data ignored)
  - Conventional approach from [2], but not motivated through robustness
- New, robust estimation principle: β-estimation [3]

$$\widehat{\boldsymbol{W}}_{\mathrm{M}\beta} = \operatorname*{argmax}_{\boldsymbol{W}} \sum_{p \in \mathcal{D}} \left((f_{\boldsymbol{D}}(\boldsymbol{d}_p;\, \boldsymbol{\theta}(\boldsymbol{l}_p;\, \boldsymbol{W})))^{\beta} - \frac{\beta}{1+\beta} \int (f_{\boldsymbol{D}}(\boldsymbol{x};\, \boldsymbol{\theta}(\boldsymbol{l}_p;\, \boldsymbol{W})))^{1+\beta} \mathrm{d}\boldsymbol{x} \right)$$

  - $\beta > 0$ is a tuning parameter, $\beta \to 0$ recovers MLE
  - Statistically robust: only finite penalty if $f_{\boldsymbol{D}}(\boldsymbol{d}_p;\, \boldsymbol{\theta}) = 0$

References

[1] G. E. Henter, S. Ronanki, O. Watts, M. Wester, Z. Wu, and S. King, "Robust TTS duration modelling using DNNs," in Proc. ICASSP, 2016.

[2] H. Zen and A. Senior, "Deep mixture density networks for acoustic modeling in statistical parametric speech synthesis," in Proc. ICASSP, 2014.

[3] 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.

Experiment

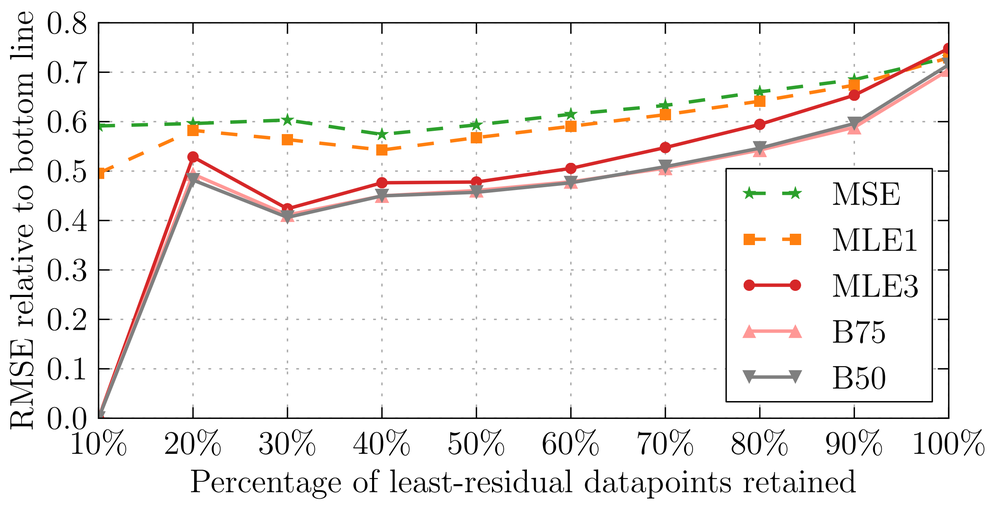
- Data: Vol. 3 of Jane Austen's "Emma" audiobook from LibriVox ($\approx$3 hours)
  - Has transcription errors and other found-data problems
  - Public domain: https://librivox.org/emma-by-jane-austen-solo/
- Input features: 592 binary + 9 continuous features based on Festvox
  - Pause phones inserted based on natural speech (oracle)
- Output features:
  - Duration prediction: 6-vector of (5) state and phone durations
  - Acoustic modelling: 86$\times$3 normalised STRAIGHT features
- DNN: 6 tanh feedforward layers with MDN output, implemented in Theano
- Systems:

Label	Duration prediction	Role	Robust?
VOC	Vocoded speech (waveform)	Top line	-
FRC	Durations from forced alignment	Oracle	-
BOT	Monophone mean duration	Bottom line	$\times$
MSE	Minimum mean-square error	Baseline	$\times$
MLE1	Gaussian-output MDN fit w/ MLE	Baseline	$\times$
MLE3	$K = 3$ Gaussian-component MDN fit w/ MLE	Previous	✓
B75	Gaussian-output MDNs fit w/ β-divergence,	Proposed	✓
B50	tuned to ignore $\approx$25 or 50% of datapoints	Proposed	✓

- All systems (except VOC) used the same DNN acoustic model

Objective results

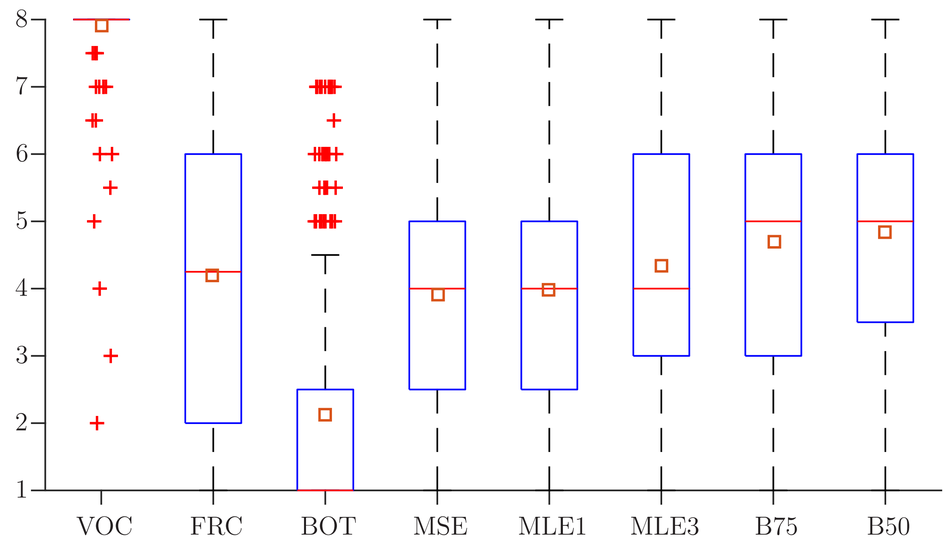
- RMSE (frames per phone) between predicted and forced-aligned durations
  - Measured on progressively larger and less well explained test-data subsets
  - Normalised to place BOT at 1.0



- Conclusion: Robust systems reject outliers and better describe the typical case

Subjective results

- MUSHRA/preference test hybrid
  - 21 listeners ranked 18 (of 21) sentences per system
  - Box plot of aggregate ranks (higher is better; squares are means):



- Nearly all differences (except MSE vs. MLE1) are statistically significant
- Conclusion: Robust duration prediction is subjectively preferred

Learn more

Paper

homepages.inf.ed.ac.uk/ghenter/pubs/henter2016robust.pdf

Audio examples

homepages.inf.ed.ac.uk/ghenter/demo/henter2016robust

Test materials

dx.doi.org/10.7488/ds/1317