ROBUST TTS 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, U.K.

gustav.henter@ed.ac.uk

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 expressive speech are easy to acquire for training them. Unfortunately, found datasets are seldom subject to the quality-control that traditional synthesis methods expect. Common issues likely to affect duration modelling include transcription errors, reductions, filled pauses, and forced-alignment inaccuracies. To combat this, we propose to improve modelling and prediction of speech durations using methods from *robust statistics*, which are able to disregard ill-fitting points in the training material. We describe a robust fitting criterion based on the density power divergence (the β-divergence) and a robust generation heuristic using mixture density networks (MDNs). Perceptual tests indicate that subjects prefer synthetic speech generated using robust models of duration over the baselines.

***Index Terms*—** Speech synthesis, duration modelling, robust statistics

1. INTRODUCTION

Despite steady improvements over many decades of research, computer-generated synthetic speech is still not convincingly natural [1]. A particular Achilles heel is the *prosody* of the artificial speech – encompassing signal aspects such as fundamental frequency, signal energy, and speech-sound durations – which makes machine speech come across as unnatural, inappropriate, and unappealing. This failure makes synthetic speech unsuitable for many attractive applications, for example generating expressive and conversational speech for audiobooks. In this paper, we focus on statistical techniques for improved duration modelling, as a key step towards the overarching goal of more natural and appropriate synthetic speech.

The recent rise of deep neural networks (DNNs) has brought an increase in performance in both automatic speech recognition (ASR) and statistical text-to-speech (TTS) technology. DNN techniques, however, require substantial computational power and large datasets to realise their full potential.

Fortunately, there has also been a revolution in the availability of data, and now virtually limitless amounts of speech, in an endless variety of voices and languages, can be obtained from many different sources. This cornucopia of speech material comes with an important caveat: most large, found datasets have not been created with speech synthesis in mind. As a consequence, they typically offer little in the way of quality control of audio recordings and text transcripts. The sheer size of these data makes manual correction too expensive. Since using impure data tends to degrade synthesis performance [2, 3], speech synthesis mostly continues to rely on smaller, carefully purpose-recorded datasets.

Here we consider duration modelling for audiobook data, using DNNs. To cope with errors and empirical variability that cannot be modelled with standard set-ups, we introduce estimation procedures from the field of robust statistics to automatically identify and disregard dubious or unhelpful datapoints. To our knowledge, these techniques have not been applied to duration modelling before. As a side benefit, we obtain models that better estimate peak probabilities in the data, which is appealing since standard output-generation methods depend on these peaks for synthesis. Our goal is to produce more natural synthetic-speech rhythm and durations by learning from large and less artificial speech datasets, whilst being robust against the inevitable errors and excess variation such data contains.

2. BACKGROUND

2.1. Duration Modelling

Duration modelling has a long history in speech synthesis. In early, formant-based synthesis systems, duration was predicted by rule [4]. Concatenative synthesis approaches that supplanted these systems do not necessarily require modelling of duration, since the units themselves have intrinsic durations, although durations can be predicted and used in the target cost [5].

The emergence of statistical parametric speech synthesis (SPSS) [6, 7] based on hidden Markov models (HMMs) again created the need for an explicit duration model. The widely-used *HMM-based speech synthesis system* (HTS) [8] defaults to using hidden semi-Markov models (HSMMs) with Gaussian duration distributions. Recent DNN-based systems are trained to minimise mean squared prediction error, which similarly is equivalent to maximum-likelihood estimation under a fixed-variance Gaussian model of duration.

In HSMM-based systems, the modelled state-duration distributions are used both to predict the most likely durations for output generation and to guide (re-)alignment during Baum-Welch estimation [9]. However, for a DNN with frame-level inputs, the dynamic-programming used to make HSMM computations feasible is difficult to apply, making re-alignment challenging. Instead it is commonplace to align the data only once at the start of training, using an HMM tool such as HTK [10]. Because the data is neither re-aligned nor has its phonemisation revisited, errors in the material cannot be corrected, even though the DNN training process iteratively creates improved models of acoustics and duration. Such lingering errors are especially likely on found speech datasets. To overcome this shortcoming, and generate durations less affected by transcription or alignment inaccuracies, we look to the field of robust statistics.

2.2. Robust Statistics

Robust statistics [11, 12] is an umbrella term for statistical methods whose performance degrades gracefully in situations where reality does not match the modelling assumptions made. Since speech is a

Author postprint of paper accepted for presentation at ICASSP 2016, Shanghai.

highly complex signal, known not to conform to all our assumptions (cf. [13]), it is compelling to investigate these methods for TTS.

A traditional, maximum-likelihood estimated (MLE) Gaussian distribution has to cover all the data to keep the likelihood high, even if some points look highly non-Gaussian. This means that a few bad datapoints can have a big impact on the model. The central idea of many robust statistical techniques is to allow the model to give up on the most ill-fitting datapoints, in order to better explain the remaining bulk of material, cf. [14]. This notion of pruning away idiosyncratic observations is not entirely new to synthesis, see, e.g., [15, 16], but has traditionally been motivated heuristically, without invoking ideas from the field of robust statistics.

By downweighting or ignoring outlying points in the data, robust methods can describe high-density regions (peaks) of the data distribution better. On the other hand, they may need more examples to converge on a good description of the training material, and thus lag behind non-robust methods on more well-behaved data, but this is less of a concern in big data applications.

2.3. Robust Duration Modelling

As mentioned in Section 2.1, a Gaussian model of duration is standard. This model is symmetric about the mean, and thus fails to account for the significant skewness in empirical duration distributions, with a heavy tail of long-duration realisations. (On the "Emma" dataset described in Section 4.1, the mean duration is 19.8 frames per phone, while the median duration is 17 frames, indicating a distribution skewed to the right.) Additionally, a Gaussian assigns non-zero probability mass to impossible, negative durations. A robust methodology should be able to partially compensate for these model shortcomings, and better estimate the location of the bulk of the training data, and thus the mode of the distribution.

The forced-aligned training data used to learn duration models may contain many errors, given that phonetic transcriptions derived automatically from text cannot perfectly match the acoustics. This is particularly problematic for spontaneous speech corpora, which exhibit greater prosodic variation (including duration), frequent reductions, filled pauses, and other phenomena not present in carefully recorded read speech [17]. The end result is that some training-data durations are inaccurate, or even assigned to the wrong phones [18]. A robust fitting procedure should be able to largely disregard these incongruences, and yield results similar to regular maximum-likelihood estimation on a dataset where the worst-fitting (and likely erroneous) examples have been removed.

3. ENGINEERING ROBUSTNESS

We now describe a DNN-based mathematical framework for duration prediction, plus two enhancements which make the output statistically robust. To our knowledge, the combination of robust statistics and duration modelling is novel.

3.1. Preliminaries

We treat duration prediction as a stochastic regression problem, in which contextual linguistic features $\boldsymbol{l}$ are used to predict a corresponding vector $\boldsymbol{x}$ of duration values, using a database $\mathcal{D} = \{(\boldsymbol{l}_p, \boldsymbol{x}_p)\}_p$ of per-phone linguistic features and matching durations from a fixed forced alignment on some training data.

For our parametric models, we will assume that durations are distributed according to a K-component Gaussian mixture distribution with diagonal covariance matrices,

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

where $f_{\mathcal{N}}$ is the Gaussian pdf and the component masses ω_k are nonnegative and sum to one. The individual phone durations $\boldsymbol{x}_p$ are assumed to be mutually statistically independent given the set of linguistic features $\{\boldsymbol{l}_p\}_p$ over the training data.

To turn (1) into a regression model, we let the distribution parameters $\boldsymbol{\theta} = \{\omega_k,\, \boldsymbol{\mu}_k,\, \boldsymbol{\sigma}_k^2\}_k$ depend on $\boldsymbol{l}$ through a deep feedforward neural network $\boldsymbol{\theta}(\boldsymbol{l};\, \boldsymbol{W})$ with weights $\boldsymbol{W}$; the resulting construction is known as a *mixture density network* (MDN) [19]. The standard method to estimate the neural network parameters (weights) from data is to maximise the log-likelihood:

$$\widehat{\boldsymbol{W}}_{\mathrm{ML}}(\mathcal{D}) = \operatorname*{argmax}_{\boldsymbol{W}} \sum_{p} \ln f(\boldsymbol{x}_p;\, \boldsymbol{\theta}(\boldsymbol{l}_p;\, \boldsymbol{W})). \tag{2}$$

The minimum mean squared error (MMSE) fitting principle for $\boldsymbol{\mu}$, as used in [20, 21], is recovered from (2) by taking $K = 1$ (i.e., using only a single mixture component) and fixing $\boldsymbol{\sigma}_1^2 = s\mathbf{1}$, where s is any positive constant. Notably, both Gaussian MLE and MMSE are optimised by the sample mean, which weighs all data equally, outlier or not. These estimators are not statistically robust.

3.2. Generation-Time Robustness

To produce more stable durations from potentially faulty data, we first consider an MDN-based heuristic drawing on previous work in acoustic modelling. Similar to duration modelling in early HMM-based synthesis systems, the idea is to improve duration prediction through a mismatch between training and generation principles: while the traditional likelihood (2) optimised during training fits multiple components to the data, only a single mixture component $k(p)$ is selected to generate each phone p at generation time. Datapoints which were attributed to other components during training are effectively ignored during synthesis: the unused components thus act as a garbage model.

In this work, we chose to generate durations based on the mode of the heaviest mixture component (greatest ω_k) for each phone p. This follows a procedure used for MDN acoustic models [22], although that work was not motivated in terms of robust statistics.

3.3. Robust Parameter Estimation

Another route is to incorporate robustness at training time, by optimising a criterion different than the likelihood. This can provide a robust model of duration even when only a single mixture component is used ($K = 1$), i.e., without an explicit model of "bad" vs. "good" observations.

For the present work, we consider minimising the density power divergence of Basu *et al.* [23], also known as the beta-divergence [24]. This leads to the following estimation principle:

$$\widehat{\boldsymbol{W}}_{\beta}(\mathcal{D}) = \operatorname*{argmin}_{\boldsymbol{W}} \sum_{p} \Big(f(\boldsymbol{x}_p;\, \boldsymbol{\theta}(\boldsymbol{l}_p;\, \boldsymbol{W}))^{\beta} - \frac{\beta}{1+\beta} \int f(\boldsymbol{x};\, \boldsymbol{\theta}(\boldsymbol{l}_p;\, \boldsymbol{W}))^{1+\beta}\, \mathrm{d}\boldsymbol{x} \Big), \tag{3}$$

where β is a positive tuning parameter. Unlike the generation-time heuristic introduced earlier, the density power divergence offers a

principled approach to robust estimation with theoretical guarantees on its performance.

It can be shown that the limit $\beta \to 0$ recovers the maximum likelihood principle in (2), while larger β-values yield a robust procedure which still converges on the same estimate if the parametric model f is correct. It is recommended to use β-values significantly less than one [23], otherwise the method could discard large amounts of data, making finite-sample estimation accuracy suffer. In practice, β can be used to seamlessly trade off between robustness and statistical efficiency, in order to reject or keep certain fractions of the data, thus adapting the method to the properties of the application at hand.

4. EXPERIMENTAL SET-UP

4.1. Data

We investigated the benefits of robust duration modelling on audiobook data, specifically chapters 1 through 10 of volume three of the novel "Emma" by Jane Austen, as read by Sherry Crowther on LibriVox [25]. 1739 total utterances with 92,025 non-silent phones were partitioned into a large training set (1660) and smaller development (39) and test (40) sets. The total duration of the material was 175 min, with an average utterance duration of 6.06 s.

4.2. Feature Extraction

The data was segmented automatically using Festvox's `interslice` [26] and subsequently forced-aligned at the state level using HTK [10]. Festvox's `ehmm` [27] was used to insert pauses and silences into the phone-label sequences based on the acoustics. This eliminates the most egregious errors where phone durations are prolonged by acoustic silences and pauses not predicted from the text, and should be particularly helpful for non-robust methods.

The extracted phone-label sequences were coupled with text-derived linguistic features encoding a subset of the questions used by the decision-tree clustering in the standard HTS synthesiser. This produced a vector $\boldsymbol{l}_p$ of 592 binary input features, to which 9 numerical features were appended (as in [21]), all normalised to $[0.01, 0.99]$. In the duration experiments, the binary input features were only used to predict six-dimensional vectors of phone durations $\boldsymbol{x}_p$, comprising five substate durations and the total phone duration (their sum). For acoustic features, the STRAIGHT vocoder [28] was used to extract 60 mel-cepstrum coefficients, 25 band aperiodicities, logarithmic fundamental frequency ($\log F_0$) and their corresponding delta and delta-delta features at a 5 ms frame step.

For both the duration and acoustic data, a per-component mean and variance normalisation was applied prior to model training, with the transformation reversed as part of synthesis.

4.3. Models

We built a number of different robust and non-robust DNN-based predictors of duration. As non-robust baselines, we used a traditional MMSE-optimised DNN, labelled "MSE", and a single-component, maximum-likelihood Gaussian MDN, labelled "MLE1".

Against these standard approaches, we contrast the two robust methods from Sections 3.2 and 3.3. Specifically, we trained a three-component MDN using maximum likelihood, where only the heaviest component was used for synthesis. This system was labelled "MLE3". For the robustly trained models using the power divergence in equation (3), we tried $\beta = 0.358$ and 0.663. These β-values were selected based on the asymptotic variance formula in [23, Sec. 4.2.d], such that approximately 75 or 50% of the original datapoints would be retained in the case of Gaussian-distributed observations. The associated systems were labelled "B75" and "B50", respectively. These settings should provide robustness against the most extreme examples without sacrificing too much statistical efficiency. While we used a simple rule of thumb to set β via the expected fraction of data discarded, the best performance is likely attained by tuning β to each application, e.g., via a grid search.

We also included a topline system using reference durations from forced alignment on the test-set recordings (labelled "FRC"), and a bottom line (labelled "BOT") simply predicting durations to be equal to the corresponding mean monophone duration in the training data. No reasonable duration prediction system should be worse than this bottom line. It is, however, conceivable that the forced-aligned durations contain problems as detailed in Section 2.3, so it might be possible to generate subjectively better-than-reference durations through robust modelling.

4.4. DNN Training and Synthesis

All DNN-based duration predictors used a feedforward DNN with six layers of 256 nodes each. The hidden nodes used `tanh` activation functions, while the output layers were configured as in [22]. Each variance-output component was floored at 10% of its global variance value to prevent degenerate solutions during optimisation. System MSE only used mean-parameter output nodes.

The full network was initialised using small random weights, with no pre-training. Network weights were subsequently optimised using stochastic gradient descent on minibatches of size 64. Each duration prediction system was trained with a fixed learning rate, manually tuned to yield close-to-optimal results on the development set in 100 epochs or less. Early stopping was used to avoid overfitting, by aborting training once the objective function on the development set had failed to improve for five epochs. For the methods based on the density power divergence, the best results were obtained by starting from a reasonable but less robust configuration, and then refining this using the robust criterion, similar to established practice [29]. In particular, B50 was initialised from B75, which was initialised from MLE1.

A single acoustic-model DNN was trained, similar in structure to the duration prediction DNNs but with 1024 hidden nodes in each layer. Meta-parameters such as learning rate, batch size, regularisation criteria, etc. were as in [21]. All DNN training procedures were implemented in Python using Theano [30].

At synthesis time, `ehmm` phone sequences derived from the test data were used as input to each duration prediction model. This corresponds to using an oracle pausing strategy, but with no other acoustics-derived information being provided to the predictors. For each phone, the peak of the Gaussian, $\boldsymbol{\mu}(\boldsymbol{l}_p; \widehat{\boldsymbol{W}})$, was used as the predicted duration. In the case of the multi-component system MLE3, the peak of the heaviest component was used. Since the Gaussian mean may not be an integer, substate boundaries in the predicted duration sequence were quantised to the nearest frame.

Once durations had been predicted, a sequence of frame-level linguistic features was generated using the predicted phone substate durations,[1] and used as input to the acoustic-model DNN, the same for all systems, to generate post-filtered MLPG [31] parameter trajectories as in [21]. Waveforms were then synthesised using the STRAIGHT vocoder [28] and normalised following ITU P.56 [32].

[1] The overall phone duration was predicted but not used, instead acting as a secondary, multi-task learning objective. This configuration produced slightly better objective scores than predicting state durations alone.

Model	BOT	MSE	MLE1	MLE3	B75	B50
Correlation	0.55	0.80	0.79	0.79	**0.81**	0.80
RMSE (100%)	9.17	6.70	6.69	6.86	**6.46**	6.56
RMSE (90%)	5.87	4.02	3.95	3.83	**3.45**	3.50

Table 1. Pearson correlation and RMSE (frames per phone) between predicted and forced-aligned durations. RMSE is reported for all data (100%) and for the 90% of the test data with the smallest prediction residual.

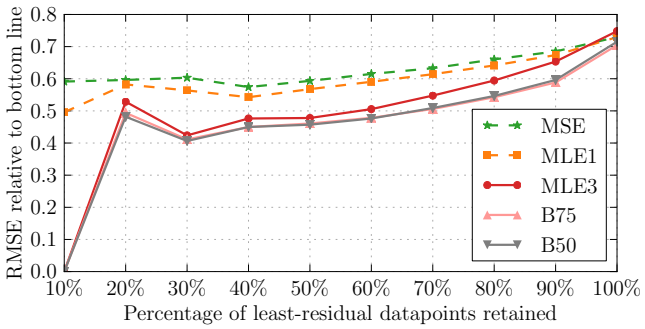

Fig. 1. Relative RMSE (frames per phone) models on progressively larger and less well explained test-data subsets. (BOT is at 1.0.)

5. RESULTS

5.1. Objective Comparison

To evaluate the different duration models, we conducted an objective evaluation using the root-mean-square error (RMSE) in frames per phone (excluding silences and pauses), as well as the closely related Pearson correlation coefficient, for each different system in relation to the reference durations (FRC) from forced-alignment. The results are reported in the first two rows of Table 1.

The advantages of robust methods become clearer if we look at subsets of the data. Since robust methods are designed to ignore extreme examples to describe the remainder of the data better, an interesting measure to consider is the RMSE on the $X\%$ of the test data with the smallest prediction residual. This is plotted in Figure 1, relative to the error of the monophone bottom line, with absolute numbers for $X = 90\%$ provided in the final row of Table 1. It is seen that the robust methods (solid lines), especially the ones based on the β-divergence, consistently outperform the non-robust baselines (dashed lines) for subset sizes less than 100%. This suggests that the robust methods were able to learn better models of the typical durations in the data.

5.2. Subjective Comparison

A perceptual experiment was conducted to evaluate all systems side-by-side using a hybrid between a MUSHRA (Multiple Stimuli with Hidden Reference and Anchor) [33] and a preference test. Listeners ranked several parallel, unlabelled stimuli from the same text but different systems in terms of preference. The scale ranged from 0 (least preferred) to 100 (most preferred). No reference was given as there is no *one* correct prosodic (durational) realisation of a sentence, but rather numerous correct productions. Nevertheless, subjects were told to give at least one stimulus in every set a rating of 100.

21 test-set sentences between 2 and 8 seconds long were used in the evaluation. Each listener rated 18 sets of 8 stimuli each, corresponding to the 6 methods in Table 1 plus FRC and vocoded natural speech (VOC). One sentence was used for a GUI tutorial and the remaining two for a training phase. Sentences were assigned to subjects randomly, presented in random order, but balanced such that all 21 sentences were rated by 18 different listeners each. 21 normal-hearing native English speakers (University of Edinburgh students, remunerated for their time) participated in the test, conducted in sound-insulated booths over high-end Beyerdynamic headphones.

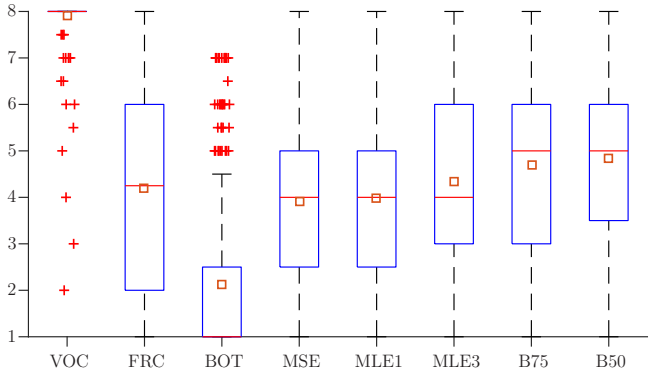

Fig. 2. Aggregated ranks in listening test. Red lines are medians, orange squares denote means; box edges are at 25 and 75% quantiles.

For analysis, each set of eight parallel listener scores was converted to ranks from 1 (lowest) to 8 (highest), with tied ranks set to the mean of the tied position. A box plot of these rank scores aggregated across all prompts and listeners is shown in Figure 2.

The box plot indicates that robust methods perform better than non-robust ones, though a substantial gap remains to the vocoded natural speech. Mann-Whitney U significance tests with Holm-Bonferroni correction [34] applied to keep the familywise error rate below 5% showed most system pairs to be significantly different, except the sets {FRC, MSE, MLE1}, {FRC, MLE3}, and {B75, B50}.

It is clear that robustly predicted durations, particularly those based on the β-divergence, were preferred over their non-robust counterparts. Interestingly, durations generated using the non-robust baselines were not significantly worse than oracle durations based on forced-alignment. Durations based on long-term linguistic and prosodic context (as in the audiobook speech) thus appear to provide little gain when synthesising and evaluating isolated sentences.

It should be noted that speech from robust models is faster than that of other systems: because of the skewness of the empirical data, rejecting extremely long durations noticeably reduces the average duration of certain speech sounds, making robust methods produce more phones per second. As speech rate affects perception [35], this might contribute to the robust methods scoring better than FRC.

6. CONCLUSION

We have described a new application of robust statistical methods to duration prediction in speech synthesis. In an application to audiobook data, the robust techniques were found to predict the vast majority of empirical durations better than non-robust baselines. A subjective evaluation found synthetic speech based on robust durations to be preferred over speech with non-robust durations, as well as over speech synthesised with oracle durations from held-out data.

Acknowledgements: The authors thank Antti Suni at the University of Helsinki for suggesting the use of robust statistical methods specifically for duration modelling. This research was supported by EPSRC Programme Grant EP/I031022/1, Natural Speech Technology (NST). The test stimuli and response data for this paper are permanently available at http://dx.doi.org/10.7488/ds/1317. The full NST research data collection may be accessed at http://hdl.handle.net/10283/786.

7. REFERENCES

[1] S. King and V. Karaiskos, "The Blizzard Challenge 2013," in *Proc. Blizzard Chall. Workshop*, 2013.

[2] J. Yamagishi, Z.-H. Ling, and S. King, "Robustness of HMM-based speech synthesis," in *Proc. Interspeech*, pp. 581–584, 2008.

[3] R. Karhila, U. Remes, and M. Kurimo, "Noise in HMM-based speech synthesis adaptation: Analysis, evaluation methods and experiments," *IEEE J. Sel. Top. Signa.*, vol. 8, no. 2, pp. 285–295, 2014.

[4] D. H. Klatt, "Review of text-to-speech conversion for English," *J. Acoust. Soc. Am.*, vol. 82, no. 3, pp. 737–793, 1987.

[5] I. Bulyko and M. Ostendorf, "Joint prosody prediction and unit selection for concatenative speech synthesis," in *Proc. ICASSP*, vol. 2, pp. 781–784, 2001.

[6] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Commun.*, vol. 51, no. 11, pp. 1039–1064, 2009.

[7] S. King, "An introduction to statistical parametric speech synthesis," *Sadhana*, vol. 36, no. 5, pp. 837–852, 2011.

[8] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. Black, and K. Tokuda, "The HMM-based speech synthesis system (HTS) version 2.0," in *Proc. SSW6*, pp. 294–299, 2007.

[9] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[10] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book (for HTK Version 3.4)*. 2006.

[11] P. J. Huber, *Robust Statistics*. New York, NY: Springer, 2nd ed., 2011.

[12] F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel, *Robust Statistics: The Approach Based on Influence Functions*. New York, NY: John Wiley & Sons, 1986.

[13] G. E. Henter, T. Merritt, M. Shannon, C. Mayo, and S. King, "Measuring the perceptual effects of modelling assumptions in speech synthesis using stimuli constructed from repeated natural speech," in *Proc. Interspeech*, pp. 1504–1508, 2014.

[14] A. Y. Aravkin, T. van Leeuwen, and F. J. Herrmann, "Robust full-waveform inversion using the Student's *t*-distribution," in *SEG Tech. Program Expand. Abstr.*, vol. 30, pp. 2669–2673, 2011.

[15] K. R. Krishnan, "Prosodic analysis of Indian languages and its application to text to speech synthesis," Master's thesis, Department of Electrical Engineering, IIT Madras, India, July 2015.

[16] A. Prakash, A. Baby, A. S. Shanmugam, J. J. Prakash, N. L. Nishanthi, K. R. Krishnan, V. S. Rupak, and H. A. Murthy, "Blizzard Challenge 2015: Submission by DONLab, IIT Madras," in *Proc. Blizzard Chall. Workshop*, 2015.

[17] E. Shriberg, "Spontaneous speech: how people really talk and why engineers should care," in *Proc. Interspeech*, pp. 1781–1784, 2005.

[18] S. Brognaux and T. Drugman, "Phonetic variations: Impact of the communicative situation," in *Speech Prosody*, 2014.

[19] C. M. Bishop, "Mixture density networks," Tech. Rep. NCRG/94/004, Neural Computing Research Group, Aston University, 1994.

[20] H. Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *Proc. ICASSP*, pp. 7962–7966, 2013.

[21] Z. Wu, C. Valentini-Botinhao, O. Watts, and S. King, "Deep neural networks employing multi-task learning and stacked bottleneck features for speech synthesis," in *Proc. ICASSP*, 2015.

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

[23] 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.

[24] S. Eguchi and Y. Kano, "Robustifying maximum likelihood estimation," Tech. Rep. Research Memo 802, Institute of Statistical Mathematics, Tokyo, Japan, June 2001.

[25] J. Austen and S. Crowther, "Emma," in *LibriVox*, 2006. http://librivox.org/emma-by-jane-austen-solo/. Accessed 2015-09-24.

[26] K. Prahallad and A. W. Black, "Segmentation of monologues in audio books for building synthetic voices," *IEEE T. Audio Speech*, vol. 19, no. 5, pp. 1444–1449, 2011.

[27] K. Prahallad, A. W. Black, and R. Mosur, "Sub-phonetic modeling for capturing pronunciation variations for conversational speech synthesis," in *Proc. ICASSP*, vol. 1, pp. I–I, 2006.

[28] H. Kawahara, "STRAIGHT, exploitation of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds," *Acoust. Sci. Technol.*, vol. 27, no. 6, pp. 349–353, 2006.

[29] P. J. Huber, "Robust regression: asymptotics, conjectures and Monte Carlo," *Ann. Stat.*, pp. 799–821, 1973.

[30] J. Bergstra, O. Breuleux, F. Bastien, P. Lamblin, R. Pascanu, G. Desjardins, J. Turian, D. Warde-Farley, and Y. Bengio, "Theano: A CPU and GPU math compiler in Python," in *Proc. 9th Python in Science Conf.*, pp. 3–10, 2010.

[31] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *Proc. ICASSP*, vol. 3, pp. 1315–1318, 2000.

[32] International Telecommunication Union, Telecommunication Standardization Sector, Geneva, Switzerland, *Objective measurement of active speech level*, March 2011.

[33] International Telecommunication Union Radiocommunication Assembly, Geneva, Switzerland, *Method for the subjective assessment of intermediate quality level of audio systems*, June 2014.

[34] S. Holm, "A simple sequentially rejective multiple test procedure," *Scand. J. Stat.*, vol. 6, no. 2, pp. 65–70, 1979.

[35] R. Dall, M. Wester, and M. Corley, "The effect of filled pauses and speaking rate on speech comprehension in natural, vocoded and synthetic speech," in *Proc. Interspeech*, pp. 56–60, 2014.