Supplementary Material: A Recurrent Latent Variable Model for Sequential Data

Anonymous Author(s)
Affiliation
Address
email

1 Dataset Description

Table 1: Data statistics and evaluation settings

	Blizzard [5]	TIMIT [3]	Onomatopoeia	Accent [8]	IAM-OnDB [6]
Number of IDs	1	630	51	2,046	500
Number of items	2.2M	6300	6738	2046	13040
Average length	8000	49218	25394	8000	628
Predefined splits	False	True	False	False	True
Minibatch size	128	64	64	128	20
Learning rate	0.0003	0.001	0.0003	0.001	0.0003

1.1 Speech Modeling

Speech synthesis is an area of increasing interest. Along with automatic speech recognition (ASR), text to speech (TTS) has been a popular task in machine learning and signal processing. Conventional TTS systems consist of a database storing the pieces of speech and combine them according to given phonetic labels. Training a generative model on speech data without any side information has not been studied well, but can be of great interest for certain types of speech. There are also few apporaches using neural networks to directly work on raw time-signals. In this paper, we train our models without any conditional information, with the idea that each model should learn the distribution of speech using its own representational power. The idea of modeling raw speech is largely orthogonal to the idea of adding conditional information or language level modeling, and understanding the necessary components to build good unconditional models should only improve future conditional models.

Blizzard We use data made available as a part of the text-to-speech (TTS) challenge known as the Blizzard Challenge 2013. This dataset consists of approximately 300-hours of English speech spoken by a single female speaker. Given the extreme length of these audio files, we process the data so that each effective sample duration is $0.5s$, obtaining a total of $2.2M$ datapoints with which to train our models. $0.5s$ is a rather short duration for true speech segments, and to compensate we use *truncated backpropagation through time*, initializing the next initial hidden state with the previous last hidden state and resetting the initial hidden state to a zero-vector every four updates. This gives an update window of approximately $2.0s$, which should be sufficient for analyzing and generating small snippets of realistic speech. We shuffled and divided the dataset into train/validation/test splits using a split fraction of $0.9/0.05/0.05$ See [5] for more information.

TIMIT TIMIT [3] is one of the most widely used datasets for benchmarking speech recognition systems. It contains $6,300$ English sentences spoken by 630 speakers representing 8 different dialects. This is more speaker variability in this data than in Blizzard, and should allow for an analysis of how the proposed model handles this type of variability in comparison to pure language level concerns. There are predefined train/validation/test splits for this task.

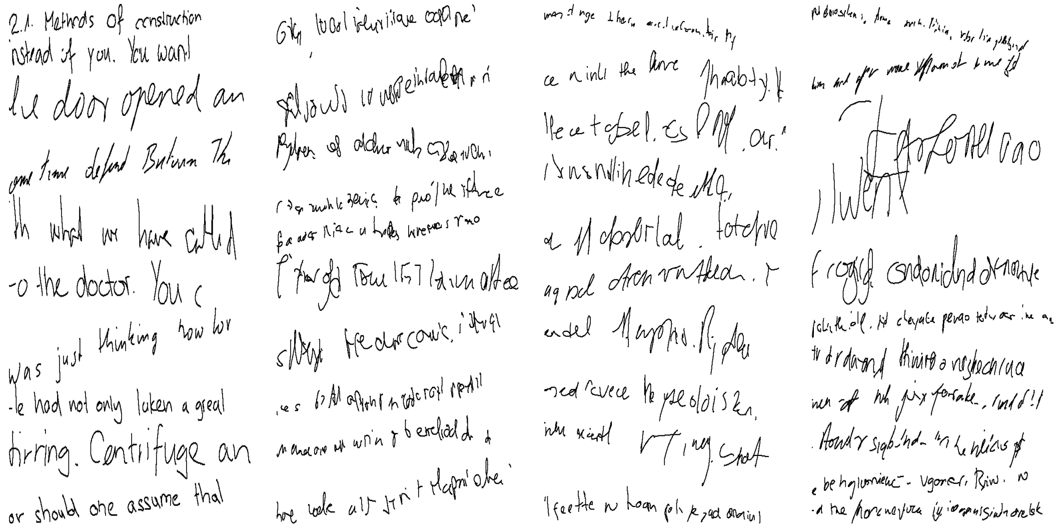
Onomatopoeia Unlike Blizzard and TIMIT, Onomatopoeia is a set of non-linguistic human-made sounds such as coughing, screaming, laughing and shouting. It features 6,738 datapoints from 51 voice actors, comprising sounds for 52 scripted and AI game characters. This dataset has been provided by Ubisoft[1] because it would be useful in open video games to be able to generate new such sounds on the fly, and avoid the impression of canned sounds. Unlike speech, there are no known segments equivalent to phonemes which could be used to condition the generative process, making it important to develop a good unconditional generative model (or one that is only conditioned on static information such as character and onomatopoeia type). We shuffled and divided the set into train/validation splits using the split fraction $0.9/0.1$.

Accent We use the GMU speech accent archive [8] to build a dataset of English accented speech, referred to as the Accent dataset. This dataset contains English paragraphs read by both native and non-native English speakers. There are currently $2,046$ readings, each from a different speaker but the dataset is actively maintained and growing. This is an extreme case of speaker variability, and may be the dataset which best shows variability ”in the wild” since each example is user submitted from around the world. Our experimental procedure was nearly identical to Blizzard, using TBPTT with $0.5s$ samples and resetting every four updates. We split the set with $0.9/0.1$ fractions chosen randomly, as in Onomatopoeia.

1.2 Handwriting Generation

The IAM-OnDB dataset contains $13,040$ handwritten lines written by 500 writers [6]. Each sequence consists of (x, y)-coordinates as well as binary indicators which tell whether the pen touches the screen or not. We follow the pipeline of Graves [4] to process the train and validation splits.

2 Handwriting Modeling: Samples

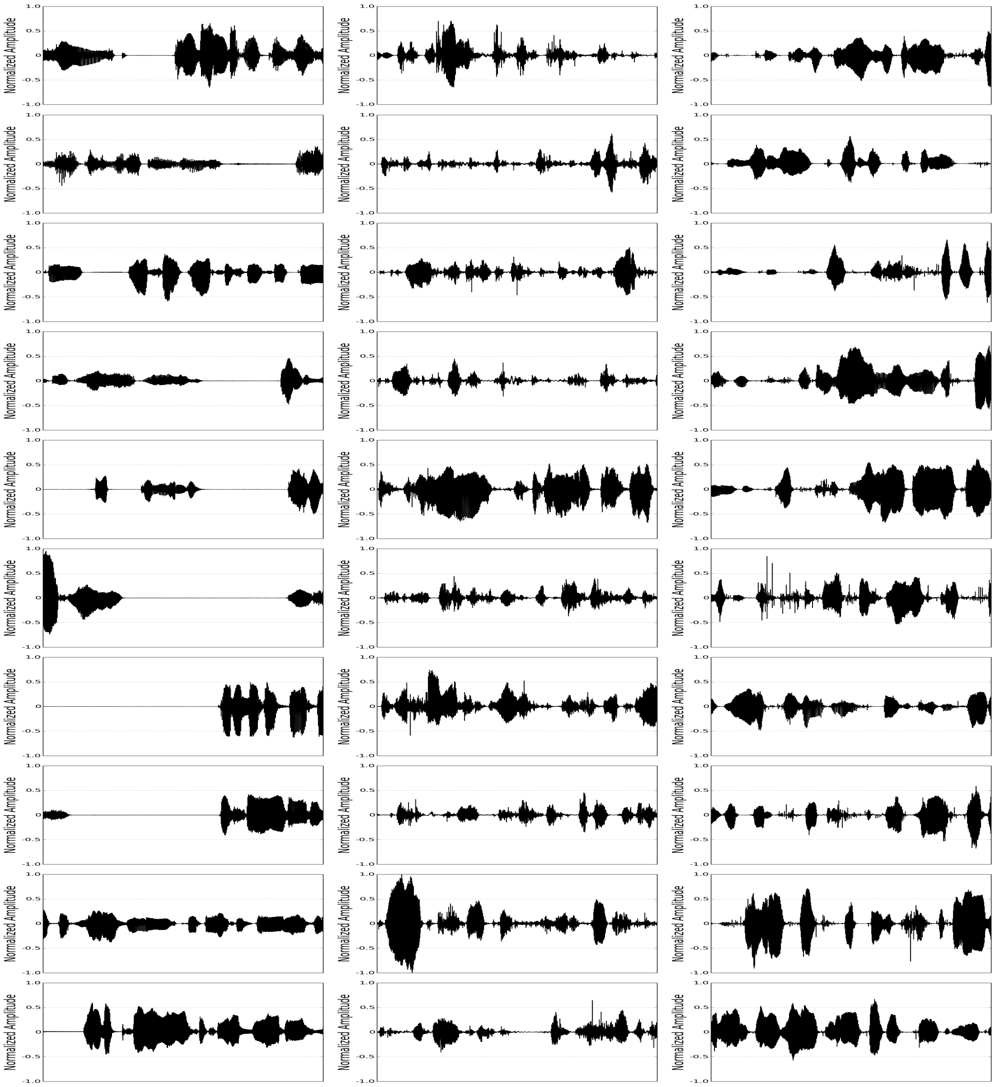

2.1. Methods of construction
instead of you. You want
he door opened an
one time defend Britain. Th
th what we have called
-o the doctor. You c
was just thinking how lov
-e had not only taken a great
drring. Centrifuge an
or should one assume that

(a) Ground Truth (b) RNN-Gauss (c) RNN-GMM (d) VRNN

Figure 1: Selected ground truths and unconditionally generated handwritings from RNN-Gauss, RNN-GMM and VRNN. VRNN retains the writing styles from the beginning until the end while RNN-Gauss and RNN-GMM tend to change the style during the generation.

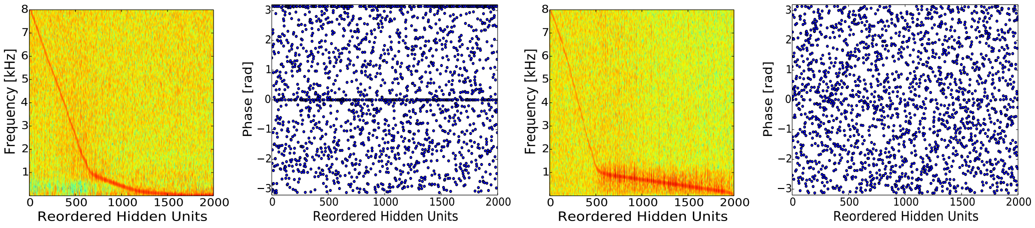
[1] `http://www.ubi.com/`

3 Speech Modeling: Samples



(a) Ground Truth (c) RNN-GMM (d) RNN-VAE

Figure 2: (1)-(9) rows show the waveforms of training examples and generated samples from RNN-Gauss, RNN-GMM and RNN-VAE.



(a) $\varphi_\tau^{\text{enc}}$ (b) $\varphi_\tau^{\text{dec}}$

Figure 3: (a) The frequency and phase responses of the encoder and (b) the same of the decoder filters learned by the VRNN which was trained on TIMIT. Best viewed in color.

Learned Speech Filters Speech processing is rarely done on high-dimensional windows of raw waveforms, typically using instead lower-dimensional responses to predefined frequency filters. Since we train the model directly on raw waveforms, the model *learns* its own frequency filters for speech.

Here we visually inspect the first-layer weight matrices, or the filters, of the VRNN trained on TIMIT (both from $\varphi_{\tau}^{\text{enc}}$ and $\varphi_{\tau}^{\text{dec}}$). To do so, we plot the frequency and phase responses of each weight matrix in Fig. 3 following [7].

As shown in Fig. 3 (a) and (b), each filter of the weight responses to only a certain frequency band, and as a group they cover the entire frequency band. There is a logarithmic dependency between the number of filters and frequency which is similar to Mel-filter banks (see, e.g., Fig. 1 (b) in [1]). This can also be seen in existing results using convolutional networks from Dieleman and Schrauwen [2]. Furthermore, the phase responses show that none of the filters are redundant (right slates of Fig. 3 (a) and (b)), meaning the model is able to effectively utilize hidden units in the first layer.

Bibliography

[1] A. Bertrand, K. Demuynck, V. Stouten, et al. Unsupervised learning of auditory filter banks using non-negative matrix factorisation. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pages 4713–4716. IEEE, 2008.

[2] S. Dieleman and B. Schrauwen. End-to-end learning for music audio. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 6964–6968, May 2014. doi: 10.1109/ICASSP.2014.6854950.

[3] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett. Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1. *NASA STI/Recon Technical Report N*, 93:27403, 1993.

[4] A. Graves. Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850*, 2013.

[5] S. King and V. Karaiskos. The blizzard challenge 2013. In *The Ninth annual Blizzard Challenge*, 2013.

[6] M. Liwicki and H. Bunke. Iam-ondb-an on-line english sentence database acquired from handwritten text on a whiteboard. In *Document Analysis and Recognition, 2005. Proceedings. Eighth International Conference on*, pages 956–961. IEEE, 2005.

[7] Z. Tüske, P. Golik, R. Schlüter, and H. Ney. Acoustic modeling with deep neural networks using raw time signal for lvcsr. In *Proceedings of the Annual Conference of International Speech Communication Association (INTERSPEECH)*, 2014.

[8] S. Weinberger. The speech accent archieve. `http://accent.gmu.edu/`, 2015.