MODELLING ACOUSTIC FEATURE DEPENDENCIES WITH ARTIFICIAL NEURAL NETWORKS: TRAJECTORY-RNADE

Benigno Uria[1] *Iain Murray*[1] *Steve Renals*[2] *Cassia Valentini-Botinhao*[2] *John Bridle*[3]

[1] Institute for Adaptive and Neural Computation, University of Edinburgh, U.K.
[2] Centre for Speech Technology Research, University of Edinburgh, U.K.
[3] Apple Siri, Gloucestershire, U.K.

ABSTRACT

Given a transcription, sampling from a good model of acoustic feature trajectories should result in plausible realizations of an utterance. However, samples from current probabilistic speech synthesis systems result in low quality synthetic speech. Henter et al. have demonstrated the need to capture the dependencies between acoustic features conditioned on the phonetic labels in order to obtain high quality synthetic speech. These dependencies are often ignored in neural network based acoustic models. We tackle this deficiency by introducing a probabilistic neural network model of acoustic trajectories, trajectory RNADE, able to capture these dependencies.

Index Terms— Speech synthesis, artificial neural networks, acoustic modelling, RNADE, trajectory model

1. INTRODUCTION

One of the main design decisions for speech synthesis systems is whether to use a probabilistic acoustic model or a unit-selection back-end. Probabilistic acoustic models tend to be more amenable to emotion and speaker adaptation, can rely on smaller training datasets, and have a smaller memory footprint in exchange for more computation. However, unit-selection systems with big unit reservoirs (tens of hours) still result in higher quality synthetic speech [1].

Given a transcription, samples from a good conditional probabilistic model of the acoustics should result in plausible speech acoustic realizations. However, samples from current probabilistic models sound noisy and unnatural [2]. For this reason, it is common practice to output the mean acoustic trajectory when synthesising speech. However, mean acoustic trajectories sound muffled due to their unusually high smoothness. To reduce this over-smoothing, postfiltering and generation techniques that take into account the variance of the acoustic trajectory are commonly used [3].

Two conditional independence assumptions contribute to the unnaturalness of samples from acoustic models: conditional independence across time; and conditional independence across acoustic features.

Supported by and EPSRC CASE studentship with Apple, and by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement No 287678 (Simple4All)

Conditional independence across time is usually dealt with by augmenting the acoustic features with dynamical features (finite differences in time) [4]. Using these dynamical features, it is possible to construct a joint probabilistic model of the trajectory across time [5, 6].

In contrast, conditional independence across acoustic features is often taken for granted by neural network based speech synthesis systems [7–9]. However, ignoring the dependency between acoustic features conditioned on a transcription results in lower perceived naturalness as judged using a sensitive MUSHRA test [10].

In traditional decision-tree-tied Gaussian models, independence across features can be relaxed by the use of full [2] or semi-tied covariance matrices [11]. In this paper we attempt to tackle the dependency across acoustic features in systems based on artificial neural networks, where it has commonly been ignored.

2. ARTIFICIAL NEURAL NETWORKS FOR ACOUSTIC MODELLING

Artificial neural networks for regression can be interpreted as conditional probability models. The output of a neural network trained to minimise the mean square error criterion corresponds to the mean of a fixed variance Gaussian distribution [12].

Mixture Density Networks (MDNs) [13, 14] provide an explicit and more powerful model of conditional probability distributions. An MDN uses a neural network to output the parameters of a fixed family of distributions (for example means, variances, and component weights for a mixture of Gaussians) given some inputs. MDNs have been used for speech synthesis [8], modelling the conditional probability of acoustic features conditioned on phonetic labels.

Usually MDNs output the parameters of a one-dimensional mixture of Gaussians for each dimension, in which case, the model assumes conditional independence across dimensions given the input. An MDN could be designed to output a full covariance matrix, for example by outputting the parameters of its Cholesky decomposition [15], but that approach will not scale past a few dimensions, given that the number of outputs would grow quadratically with the dimensionality of the data.

We introduce trajectory RNADE, a conditional generative model based on neural networks capable of capturing the dependence among acoustic features conditioned on phonetic labels. This model works by predicting the trajectory of one acoustic feature at a time. The features that have already been predicted are provided as inputs to the neural network when computing the distribution of the trajectory for each feature.

3. TRAJECTORY RNADE FOR SPEECH SYNTHESIS

Real-valued neural autoregressive density estimators [16], or RNADE, are similar to MDNs. Both use a neural network to predict a distribution of acoustic features conditioned on a set of phonetic labels by outputting the parameters of a real-valued distribution. The critical difference is that RNADE predicts each dimension within an acoustic frame sequentially i.e. the values of features already predicted are also input to the network. This allows RNADE to capture dependencies between the different acoustic features in a frame.

In an RNADE (details can be found in Uria et al. [16]) some ordering of the acoustic features is chosen and the joint probability distribution over the D-dimensional acoustics at time t, $\boldsymbol{x_t}$, given phonetic labels, $\boldsymbol{l_t}$, is factorised into a product of one-dimensional conditional distributions using the product rule,

$$p(\boldsymbol{x_t} \mid \boldsymbol{l_t}) = \prod_{d=1}^{D} p(x_{t,d} \mid \boldsymbol{l_t}, \boldsymbol{x_{t,<d}}). \quad (1)$$

Here $\boldsymbol{x_{t,<d}}$ stands for the d-minus-one-dimensional vector of acoustic features in the t-th frame that precede $x_{t,d}$ in the ordering of variables chosen, that is, that have already been predicted.

In RNADE, all conditional distributions in (1) are modelled by a single neural network using a variable number of inputs. Using a single neural network allows the activations of the first hidden layer to be reused in the computation of each conditional [17] (updating the hidden activations only requires an H-dimensional vector addition). This is an important property of RNADE that gives a one-hidden-layer RNADE its mild computational complexity $O(DH)$ (for both density calculation and sampling), where H stands for the number of hidden units in the autoregressive layer. Although RNADEs with more than one hidden layer can be trained efficiently [18], the computational complexity of sampling from them is $O(DH^2)$, which makes them too slow for speech synthesis, where computational performance is crucial.

The practical limitation to one hidden layer only affects the autoregressive part of RNADE. It is possible to use several hidden layers to compute useful predictive features from the phonetic labels, as shown in Figure 1.

A conventional RNADE model would give a non-Gaussian model of acoustic and delta features within a frame. Even if a Gaussian is used to model each conditional in (1), each conditional depends non-linearly on the value of the previous features. This non-Gaussian distribution cannot be used to obtain a distribution across time with the standard trajectory HMM formalism [6]. To obtain a tractable time-series model, with non-linear dependencies between acoustic features within each frame, we introduce the trajectory RNADE.

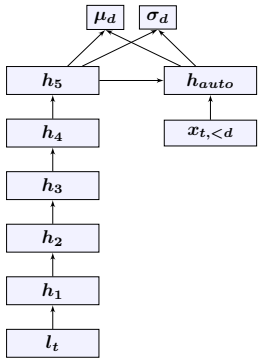

Fig. 1. Connectivity in an RNADE with one hidden autoregressive layer and several conditional feature extraction layers. The diagram only shows the outputs used when predicting the parameters of a Gaussian distribution over the d-th acoustic feature, $x_{t,d}$, conditioned on phonetic labels, $\boldsymbol{l_t}$, and other features already predicted, $\boldsymbol{x_{t,<d}}$. At training time the values of $\boldsymbol{x_{t,<d}}$ are taken from the training dataset, at test time they are samples obtained from the network.

In a trajectory RNADE the trajectory of an acoustic feature over the whole utterance is predicted before the trajectory of the following acoustic feature is predicted. For each acoustic feature, the network outputs the mean and variance for the static, delta and double-delta dimensions of the feature for each frame in the utterance. A joint probabilistic model over the trajectory of the acoustic feature results from the standard trajectory HMM [5]. From this trajectory model we can output a sample (or the highest density trajectory [4]). The trajectory for the feature will be used as an input to the RNADE when predicting the trajectory of the following acoustic feature.

As it is common, during training we model the static and delta features as unconstrained variables, even though the trajectory can only lie on a subspace of the augmented space. This inconsistency between the training criterion and the actual generation procedure causes an underestimation of the variability of the acoustic trajectory [2].

The trajectory HMM formalism corresponds to a product of Gaussians (PoG) [19]. A PoG is a Gaussian distribution whose precision parameter is the sum of the precisions of the n Gaussians multiplied. If we assume the Gaussians multiplied have approximately the same variance, by multiplying each of their variances by n, their product will have the same variance. This leads to a heuristic where we will multiply the variances of the static and delta features by 3. This heuristic has previously been shown to improve the likelihoods of trajectory models [2, 20], and our results in the next section agree with this finding.

Criterion	MDN	Trajectory-RNADE
Avg. log-density x, Δx, $\Delta\Delta x$	0.87	102.60
Avg. log-density x	-32.45	7.18
Avg. log-density x trajectory model	-18.16	23.04
Avg. log-density x trajectory model ($3\times$ variance)	23.12	59.27

Table 1. Average log-density per frame on a held-out test set of 96 utterances, greater numbers are indicative of better models.

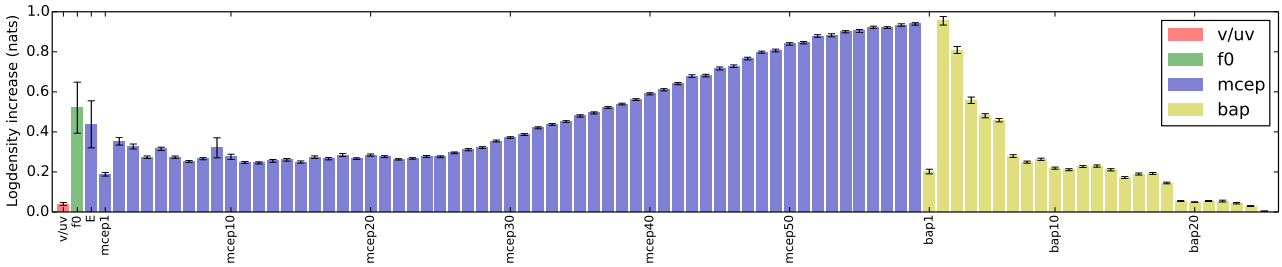

Fig. 2. Increase in average log-density per acoustic feature obtained by a trajectory RNADE with respect to a trajectory MDN. Both models had the variances of the static and delta features multiplied by 3 before calculating the distribution over trajectories. Dimensions are shown in the order they are predicted by the RNADE model.

4. EXPERIMENTAL RESULTS

Our experimental goal is to compare the statistical performance of a trajectory RNADE compared to an MDN.

The systems were trained using the database recorded to build the TTS entry in [21]. It consists of 2 hours of data of a British male voice. The data used here was sampled at 48kHz. We split the dataset into a training subset of 2602 utterances, 98 validation utterances (used for early stopping and selecting hyperparameters during training) and a test set of 99 utterances. We used STRAIGHT vocoding [22] to extract 60 mel-cepstral features, 25 band aperiodicity features, f_0 (linearly interpolated in unvoiced regions) and a voiced/unvoiced feature for a total of 87 dimensions. In order to train both the RNADE and the MDN we used forced-alignments obtained with a standard HMM-GMM system.

The input phonetic labels were the same for both the MDN and RNADE. We included the substate alignment from the HMM-GMM system and its relative position within the substate [9].

The acoustic data was rescaled to the range 0.01–0.99, logit-transformed, augmented with deltas and double deltas, and standardized to mean zero and standard deviation one. Modelling the logit-transformed version of the data guarantees that samples will remain in an acceptable range even when using Gaussian models, which have infinite support.

The MDN used in our experiments had five hidden layers, each had 600 rectified linear units [23]. We used a Gaussian output per dimension (mean and standard deviation) as we found no increase in statistical performance when using a mixture of Gaussians (MoG). Previous work has reported a benefit from mixture models [8], although those models were fitted with ten times more data than here.

The trajectory RNADE in our experiments also had five conditional hidden layers, and one autoregressive hidden layer. Each hidden layer had 600 rectified linear units (see Figure 1). We used a Gaussian output per dimension, again finding no increase in statistical performance when using a MoG. In the RNADE model we chose the following order for the acoustic features: voiced/unvoiced, f_0, mel-cepstral features (0 to 59 in that order), band aperiodicities (1 to 25 in that order).

Both models were trained using AdaDec [24] for 1000 epochs of 1000 updates each, minibatches of size 100 were used. The learning rate was initialized to 3×10^{-4} and decreased by 3×10^{-7} after each epoch.

We are mainly interested in measuring the quality of the two systems as generative models of speech acoustics conditioned on phonetic labels. Samples and mean trajectories from the two models are available online[1]. To measure the statistical performance of the models we report their test-set log-likelihood (average log-density of a frame in the held out test set). Results are shown in Table 1. The first row shows the log-likelihood of each model considering the static and delta features as unconstrained dimensions, this is our training criterion. Note that the first row is not comparable to the rest, as it considers models over the delta-augmented acoustic (261 dimensions) space, while the following rows consider only the acoustic trajectories (87 dimensions). The second row shows the likelihoods of the models by ignoring the delta-features, i.e. each frame is considered independent of the rest conditioned on the labels. The third and fourth rows show the likelihood of

[1] http://www.benignouria.com/permalink/rnade_synthesis

the models using the trajectory HMM formalism to calculate the density of the trajectories. To compensate for the underestimation in variance caused by a training criterion inconsistent with the trajectory model, we also calculated the densities under a trajectory model where the variances of static and delta features have been multiplied by 3 (we also tried 2 and 4, but obtained lower log-likelihoods in both cases). This heuristic increase in variance increases the likelihoods and the global variance of samples.

Trajectory-NADE achieved higher likelihoods under all criteria. This leads us to conclude that it is a better joint model of acoustic feature dependencies conditioned on phonetic labels than an MDN.

Due to the sequential nature of RNADE we can calculate which acoustic features are being predicted to a higher degree of accuracy compared to an MDN. In Figure 2 we investigate the source of the increase in likelihood. The y-axis shows the average increase in likelihood obtained by modelling the trajectory of a feature using a trajectory RNADE instead of a trajectory MDN (both having the variances of statics and deltas multiplied by 3, bottom row in Table 1). The figure shows the features in the order they are predicted by the RNADE model. We expect no increase in prediction accuracy of the voiced/unvoiced feature, as the RNADE uses uses the same predictor variables as the MDN in this case (any difference is caused by the stochastic training procedure). Regarding f_0, we observe an increase in accuracy of about half a nat, which shows that knowing whether a frame is voiced or not is helpful in predicting f_0 (the f_0 values were linearly interpolated in unvoiced regions). A similar increase in log-density is observed in the prediction of the frame energy ($mcep_0$). Regarding the mel-cepstral features, there is a slight increase in log-density of the 1st feature ($mcep_1$), followed by a higher increase in the middle cepstral features ($mcep_2$–$mcep_{28}$) and a greater and growing increase in accuracy in the higher cepstral features ($mcep_{29}$–$mcep_{59}$) that describe the fine structure of the spectrum. These features are very difficult to predict conditioned only on the phonetic labels, but seem to show a high degree of dependency with each other. Regarding band aperiodicity features, we again find a small increase in the log-density of the first feature followed by a higher increase in the following features (which are highly correlated to the first).

We are also interested in measuring the quality of the synthetic speech generated by each system. We performed three forced-preference tests comparing: (1) trajectory samples from an MDN and an RNADE, (2) trajectory means from an MDN and an RNADE (3) trajectory samples and trajectory means from an RNADE. Samples were obtained using the $3\times$ variance heuristic, and means using the MLPG [4] algorithm. All tests were performed by a group of 29 native English speakers using headphones in sound-deadened booths. The participants were asked to choose the higher quality instance from pairs of utterances presented in a random order. The results can be seen in Table 2. The participants showed preference for means and samples generated from the RNADE instead of the MDN. They also showed preference for means instead of samples generated by RNADE.

MDN		Traj. RNADE		n
Sample	Mean	Sample	Mean	
19.4%	-	**80.6%**	-	900
-	-	16.0%	**84.0%**	865
-	33.6%	-	**66.4%**	794

Table 2. Subjective evaluation results. The last column shows the total number of comparisons performed by the 29 participants for each of the tests. For statistically significant results at a 0.99 level in a two-tailed binomial test (indiference null-hypothesis) the preferred system is shown in bold font.

5. DISCUSSION

The use of NADE for speech synthesis was proposed before by Yin et al. [25]. They used fixed-variance real-valued version of NADE to replace the means of an HMM-GMM speech synthesis system; training a different NADE for each state. In contrast, our approach takes full advantage of the neural network formulation of RNADE: it uses a single network that takes phonetic labels as inputs, and outputs means and standard deviations for each acoustic feature.

The trajectory HMM formulism, using deltas to model the dependencies across time, is a limiting one. Calculating the distribution over trajectories is slow. It requires a matrix inversion, making it difficult to optimize the right training criterion. It also limits the family of distributions that can be used to model each frame of the augmented-feature distribution to a Gaussian. In future work we plan using an RNADE model that is autoregressive across the features in a frame and also across time. Autoregressive HMMs are a promising research direction for speech synthesis [20], but in preliminary experiments training fully-autoregressive RNADEs got stuck in local optima that model smooth speech-like acoustic trajectories but ignore the phonetic labels.

We will also investigate the importance of the order in which the acoustic-dimensions are predicted. RNADEs with different orders may model different joint distributions, especially if any of the empirical one-dimensional conditionals in (1) is multimodal.

In conclusion, our experimental results show that trajectory RNADE is better than an MDN at modelling the joint distribution of acoustic features conditioned on phonetic labels. Furthermore, trajectory RNADE also produces higher quality synthetic speech as judged by subjective preference tests.

We would like to thank Gustav Henter, Korin Richmond, Zhizheng Wu, and Simon King for their advice and interesting discussions.

6. REFERENCES

[1] Heiga Zen, Keiichi Tokuda, and Alan W Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.

[2] Matt Shannon, Heiga Zen, and William Byrne, "The effect of using normalized models in statistical speech synthesis," in *Proceedings of the 12th Annual Conference of the International Speech Communication Association, Interspeech*, 2011, pp. 121–124.

[3] Toda Tomoki and Keiichi Tokuda, "A speech parameter generation algorithm considering global variance for HMM-based speech synthesis," *IEICE Transactions on Information and Systems*, vol. 90, no. 5, pp. 816–824, 2007.

[4] Keiichi Tokuda, Takayoshi Yoshimura, Takashi Masuko, Takao Kobayashi, and Tadashi Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2000, pp. 1315–1318.

[5] Heiga Zen, Keiichi Tokuda, and Tadashi Kitamura, "An introduction of trajectory model into HMM-based speech synthesis," in *Fifth ISCA Workshop on Speech Synthesis*, 2004.

[6] Heiga Zen, Keiichi Tokuda, and Tadashi Kitamura, "Reformulating the HMM as a trajectory model by imposing explicit relationships between static and dynamic feature vector sequences," *Computer Speech & Language*, vol. 21, no. 1, pp. 153–173, 2007.

[7] Heiga Zen, Andrew Senior, and Mike Schuster, "Statistical parametric speech synthesis using deep neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 7962–7966.

[8] Heiga Zen and Andrew Senior, "Deep mixture density networks for acoustic modeling in statistical parametric speech synthesis," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2014, pp. 3872–3876.

[9] Yao Qian, Yuchen Fan, Wenping Hu, and Frank K Soong, "On the training aspects of deep neural network (DNN) for parametric TTS synthesis," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2014, pp. 3829–3833.

[10] Gustav E. Henter, Thomas Merritt, Matt Shannon, Catherine Mayo, and Simon King, "Measuring the perceptual effects of modelling assumptions in speech synthesis using stimuli constructed from repeated natural speech," in *Proceedings of the 15th Annual Conference of the International Speech Communication Association, Interspeech*, 2014, pp. 1504–1508.

[11] Mark J. F. Gales, "Semi-tied covariance matrices for hidden Markov models," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 3, pp. 272–281, 1999.

[12] John S. Bridle, "Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition," in *Neurocomputing*, pp. 227–236. Springer, 1990.

[13] Christopher M. Bishop, "Mixture density networks," Tech. Rep. NCRG 4288, Neural Computing Research Group, Aston University, 1994.

[14] Christopher M. Bishop, *Neural Networks for Pattern Recognition*, Oxford University Press, Inc., New York, NY, USA, 1995.

[15] Peter M. Williams, "Using neural networks to model conditional multivariate densities," *Neural Computation*, vol. 8, no. 4, pp. 843–854, 1996.

[16] Benigno Uria, Iain Murray, and Hugo Larochelle, "RNADE: The real-valued neural autoregressive density-estimator," in *Advances in Neural Information Processing Systems 26*, 2013, pp. 2175–2183.

[17] Hugo Larochelle and Iain Murray, "The neural autoregressive distribution estimator," *Journal of Machine Learning Research W&CP*, vol. 15, pp. 29–37, 2011.

[18] Benigno Uria, Iain Murray, and Hugo Larochelle, "A deep and tractable density estimator," in *Proceedings of the 31st International Conference on Machine Learning*, 2014, pp. 467–475.

[19] C. K. I. Williams, "How to pretend that correlated variables are independent by using difference observations," *Neural computation*, vol. 17, no. 1, pp. 1–6, 2005.

[20] Matt Shannon, Heiga Zen, and William Byrne, "Autoregressive models for statistical parametric speech synthesis," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, no. 3, pp. 587–597, 2013.

[21] M. Cooke, C. Mayo, C. Valentini-Botinhao, Y. Stylianou, B. Sauert, and Y. Tang, "Evaluating the intelligibility benefit of speech modifications in known noise conditions," *Speech Communication*, vol. 55, pp. 572–585, 2013.

[22] Hideki Kawahara, Ikuyo Masuda-Katsuse, and Alain de Cheveigné, "Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, no. 3, pp. 187–207, 1999.

[23] Vinod Nair and Geoffrey E Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proceedings of the 27th International Conference on Machine Learning*, 2010, pp. 807–814.

[24] Andrew Senior, Georg Heigold, Marc'Aurelio Ranzato, and Ke Yang, "An empirical study of learning rates in deep neural networks for speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 6724–6728.

[25] Xiang Yin, Zhen-Hua Ling, and Li-Rong Dai, "Spectral modeling using neural autoregressive distribution estimators for statistical parametric speech synthesis," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2014, pp. 3824–3828.