IberSPEECH 2018
21-23 November 2018, Barcelona, Spain

Multi-Speaker Neural Vocoder

Oriol Barbany[1,2]*, Antonio Bonafonte*[1] *and Santiago Pascual*[1]

[1]Universitat Politècnica de Catalunya, Barcelona, Spain
[2]École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland
oriol.barbanymayor@epfl.ch, antonio.bonafonte@upc.edu, santi.pascual@upc.edu

Abstract

Statistical Parametric Speech Synthesis (SPSS) offers more flexibility than unit-selection based speech synthesis, which was the dominant commercial technology during the 2000s decade. However, classical SPSS systems generate speech with lower naturalness than unit-selection methods. Deep learning based SPSS, thanks to recurrent architectures, surpasses classical SPSS limits. These architectures offer high quality speech while preserving the desired flexibility in choosing the parameters such as the speaker, the intonation, etc. This paper exposes two proposals conceived to improve deep learning-based text-to-speech systems. First a baseline model, obtained by adapting SampleRNN, making it as a speaker-independent neural vocoder that generates the speech waveform from acoustic parameters. Then two approaches are proposed to improve the quality, applying speaker dependent normalization of the acoustic features, and the look ahead, consisting on feeding acoustic features of future frames to the network with the aim of better modeling the present waveform and avoiding possible discontinuities. Human listeners prefer the system that combines both techniques, which reaches a rate of 4 in the mean opinion score scale (MOS) with the balanced dataset and outperforms the other models.

Index Terms: deep learning, speech synthesis, recurrent neural networks, text-to-speech, SampleRNN, time series

1. Introduction

Deep learning has revolutionized almost every engineering branch over the latest years, also being successfully applied to text-to-speech (TTS) where it yields state-of-the-art performance and overcomes classical statistical approaches. The time series problem has been completely leveraged by recurrent neural networks (RNNs) and their variants, making them lead to very good results in the speech synthesis field. Moreover, deep generative models can generate speech sample by sample as first proposed in WaveNet [1], which achieved very fine-grained waveform amplitudes and outperformed previous statistical parametric speech synthesis (SPSS) models as a deep auto-regressive system. This paper exposes two of the proposals presented in the bachelor's thesis of the first author [2] that were applied to better model the speech generated with a multi-speaker deep learning-based model.

In our previous work [3], SampleRNN [4] was adapted to generate coherent speech in Spanish. SampleRNN models the joint distribution of speech samples using a hierarchical RNN, and it can be used to generate speech-like sequences. As the generation is unconditioned (only previous waveform samples are inputs), the generated sequence is a random speech-like waveform with short-time coherent structure that emulates speech but does not resemble proper spoken contents. As in the case of WaveNet [1], we can inject additional controlling features in the input that span long-term structures of the speech, like frames of acoustic information or pitch contours. In [3], a TTS system is presented which consist of two modules. The first one transforms the linguistic features (phoneme, stress, length of sentence, etc.) into an acoustic features sequence. Then this conditions the second module, a SampleRNN, which generates the waveform, hence acting as a neural vocoder. A joint optimization of models improves the generated speech. In this work we focus on the neural vocoder based on SampleRNN. The main motivation is to generalize the system proposed in [3] to many speakers as a shared deep neural network (DNN) structure, as it achieves better results in generating quality speech than learning the parameters of a single isolated speaker [5, 6]. As in our previous work, it transform acoustic features (MFCC, F0, etc.) into speech waveform. However, the speaker code is added as a new feature to control the identity. Our final goal is to control the phonetic content using the acoustic features and the speaker characteristics by means of the speaker id feature. This would allow to generate different synthetic voices using limited labelled data, as the neural vocoder is trained on unlabelled data. It could also be applied to voice conversion [7]. In this case, the input to the vocoder are the acoustic features of one speaker and the speaker id of other speaker. While the capability of changing the speaker identity influences the architecture and the proposals of this work, this paper is focused on how to apply the acoustic features to get the best quality, without changing the speaker identity with respect to the speaker associated with the input acoustic features.

This work proposes two contributions: speaker dependent normalization of the acoustic features, and an acoustic-features look ahead mechanism. The first proposal aims to perform voice conversion. The acoustic features fed to the network and used to condition the generated speech, such as pitch or Mel-frequency cepstral coefficients (MFCC), depend on the speaker, which means that the input to indicate the speaker identity is redundant. This redundancy is identified by the network, which assigns low weights to the speaker identity and make it irrelevant. The speaker-dependent normalization aims to give importance to the speaker identity by isolating the features from the speaker and thus forcing the network to use the speaker identity to generate natural speech for each user.

The look ahead approach questions the causality of time series modeling, which is not needed unless input features are extracted at real time and therefore not known beforehand. In the case of TTS systems, the text which will be uttered is known before hand and therefore, the acoustic features of all the signal are known or can be predicted before generating the speech waveform. Moreover, in natural speech, the phonemes sound different depending on the context and thus can change depending on future phonemes (co-articulation). By giving information of the future behavior of the predicted sequence, there are

10.21437/IberSPEECH.2018-7

no discontinuities and artifacts can be reduced. This is translated into better quality speech as rated by human listeners.

The proposals were initially conceived to improve the speech obtained with a deep generative network able to model multiple speakers with the same structure. Nevertheless, the speaker-dependent normalization (see section 2.2) could be used as a new pre-processing technique in a variety of problems, and the look ahead approach (section 2.3) can be generalized to time series modeling. Current state-of-the art TTS models like WaveNet [1], Tacotron [8] or VQ-VAE [9] already model several speakers with a unique model, but do not apply speaker-dependent normalizations, which is shown to deteriorate results. The look ahead proposal is not either mentioned, but it outperformed our baseline model and could be applied to other time series modeling system.

In the next section, first the baseline system is presented. It consist of SampleRNN [4], extended to generate speech conditioned to acoustic features and speaker identity. Then, the speaker dependent normalization and the look ahead are introduced. In section 3, the experimental setup is described. Section 4 presents the experimental results that show how both proposals outperform the baseline system.

2. Multi-speaker Network

2.1. Baseline

The proposed neural vocoders are based on SampleRNN, an unconditional end-to-end neural audio generation model [4] that consists of two recurrent modules running at different clock rates that aim to model the short and long term dependencies of speech signals, and one module with auto-regressive multilayer perceptrons (MLPs) that processes speech sample by sample. The authors of SampleRNN reported that gated recurrent unit (GRU) [10] cells worked slightly better than long short-term memory (LSTM) ones, hence this is the recurrent architecture adopted for this work. The three tier architecture provides flexibility in allocating the amount of computational resources for modeling different levels of abstraction and results very efficient in memory during training. The final output of SampleRNN model is the probability of the current sample value conditioned on all the previous values of the sequence that can be expressed following the chain rule of probability as stated in equation (1). This follows a Multinoulli distribution, which could be unintuitive due to the naturalness of speech signals, which are real-valued, but achieves better results as it does not assume any distribution shape of the data and thus can more easily model arbitrary distributions. In this work, speech samples are quantized with 8 bits, having therefore 256 possible values. Differing from the linear quantization proposed in SampleRNN, we apply a $\mu-$law companding transformation [11] before classifying into the 256 possible classes to flatten the Laplacian-like distribution of the speech signals.

$$P(\boldsymbol{X}) = \prod_{t=1}^{T} P(x_t | x_1, \ldots, x_{t-1}) \tag{1}$$

In order to generate speech coherent spoken contents, the model was conditioned like in [3] with acoustic features obtained with Ahocoder [12], a high-quality harmonics-plus-noise vocoder that predicts a set of features that can characterize speech signals. The adapted model with its conditioning inputs is depicted in figure 3. This changes the previous equation as a new dependence factor is included, thus new formulation follows equation (2), where $\boldsymbol{l}_t \in \mathbb{R}^{43}$ stands for a 43-dimensional acoustic vector corresponding to the analysis window of the current sample x_t.

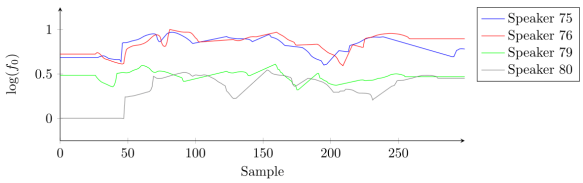

Figure 1: *Classical speaker-independent normalization*

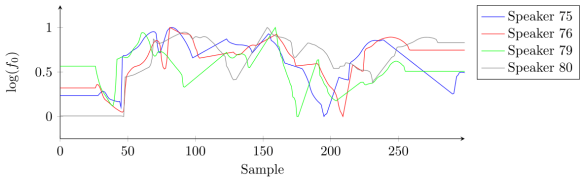

Figure 2: *Proposed speaker-dependent normalization*

$$P(\boldsymbol{X}|\boldsymbol{L}) = \prod_{t=1}^{T} P(x_t | x_1, \ldots, x_{t-1}, \boldsymbol{l}_t) \tag{2}$$

Differing from the original SampleRNN model [4] and apart from the previously mentioned addition of the acoustic conditioners that allow to synthesize coherent speech, the authors also incorporated the blocks in the left of figure 3. These aim to differentiate among all the speakers of the database by means of embedding an identifier which is also used to condition the model jointly with the aforementioned Ahocoder features. Hence $\boldsymbol{l}_t$ is augmented to include the speaker identity by concatenating the embedding to the acoustic features, resulting in a vector $\hat{\boldsymbol{l}}_t \in \mathbb{R}^{49}$.

2.2. Speaker-dependent feature normalization

Features fed to a neural network are often previously normalized to control the magnitude of both the activations and gradients in training. With the hypothesis of having speaker-dependent features, an independent normalization for each of the speakers was proposed to isolate the speech features from the source. Maximum and minimum values for each of the parameters were found within the training partition so it could happen that some features of the train or validation partitions overpass the bounds. The chosen normalization function was a simple feature scaling that follow equation (3), which bound each of the features from 0 to 1. This approach could be also applied with other normalization functions like the $z-$score, i.e. statistical normalization. This last option was not tested before the writing of this paper due to the low improvement in results of this only modification (see table 1). Nevertheless, as it can be seen in the same table, this approach outperforms the other models when combined with the look ahead approach (explained in section 2.3). Therefore, a statistical normalization could also be tested in future work.

$$\hat{\boldsymbol{x}} = \frac{\boldsymbol{x} - \boldsymbol{x}_{\min}}{\boldsymbol{x}_{\max} - \boldsymbol{x}_{\min}} \tag{3}$$

This proposal aims to give importance to the speaker identity to ideally allow voice conversion without the need of a complex mapping of features. Inspiration came from the behavior of the pitch for every speaker, which is depicted for both speaker-independent and speaker-dependent normalizations in figures 1 and 2 respectively. These plots illustrate the evolution of the logarithmic fundamental frequency for four different speakers including two males and two females that read the exact same text and thus are very similar once normalized following a speaker-dependent approach. Note that there is some time shifting due to different duration of phonemes and pauses but the signal is yet very similar.

After the classical speaker-independent normalization (figure 1), it is very easy to distinguish between females (75, 76) and males (79, 80). This means that it would be impossible to perform voice conversion because the network doesn't need the speaker identifier for being this information implicit in the features. This is why this redundancy is translated into the futility of this input observed when trying to change the speaker identity at will. The behavior of the pitch once normalized by speaker is very similar if the intonation is comparable. Nevertheless, the other features that are fed to the network (see next section) resulted in very similar normalizations for both speaker-independent and speaker-dependent approaches.

2.3. Look ahead

In the modeling of non-real-time sequences such as the generation of speech in a TTS system, the features that will be fed to the network are known beforehand. This means that, in contrast with a possible phone call where both ends are talking at real time, the features that will condition the sequence at future time steps are always known and thus can be used to better model the generated signal.

With this idea in mind, the causality that speech synthesizers inherited from the vocoders used in decoding is questioned and both the current and future windows of features are fed to the network. This results in a larger model because the number of features is duplicated at each time step but also achieves better quality without the need of more features.

Note that the look ahead approach modifies the architecture because the upper right 1D-convolution block doubles its input size (the original value of 43 is crossed out in the figure and replaced by 86 to accept both the features of the current and future frames).

3. Experimental setup

In this section we characterize the experimental conditions to evaluate the previous approaches. First we describe the speech data used to estimate the models. Then, the acoustic parameters, architecture and learning hyperparameters are outlined. Finally, the methodology used to evaluate the system is described.

3.1. Dataset

The speech dataset used in the experiments is formed by six Spanish voices from the TC-STAR project [13], where half of them are males and the other half are females. The database was unbalanced with one of the female speakers barely having a quarter of speech recording time compared to the others. Notwithstanding some works like [14] recommend balancing the data per user so that all speakers have approximately the same amount of samples to train, we choose to use all the available data per speaker to avoid restricting all of them to only 14 minutes of speech instead of an hour. The total duration of the whole dataset including the six speakers amounts to 5.25 hours, which we divide into 80% for training, 10% for validation and 10% for testing.

3.2. Feature Design and Hyper-parameters

The acoustic parameters are extracted with Ahocoder in frames of length $15\,ms$ shifted every $5\,ms$, obtaining 40 Mel-frequency cepstral coefficients, the maximum voiced frequency (fv), the logarithmic F0 value and the voiced/unvoiced flag (uv). To tackle the discontinuity in the $\log$F0 statistics in unvoiced signals, the extracted pitch is post-processed with a log-linear interpolation for the unvoiced segments following previous strategies [14].

All these features are thus scaled following either the proposed speaker-dependent or the more classical speaker-independent normalization. The normalized features are then rearranged to match the speech samples dimensions used in training and the speaker embedding is added as an independent input to the system, as mentioned earlier.

The learning strategy was to train each of the models derived from the previous proposals with mini-batch stochastic gradient descent (SGD) using a mini-batch size of 128 and minimizing the negative log-likelihood (NLL). The chosen optimizer is the adaptive moment estimation (ADAM) [15] for its effectiveness in many problems and ease of use. It is an SGD algorithm with adaptive learning rate, having an initial value of 10^{-4}, which we enhance for our task with an external rate controller known as scheduler. This had two milestones at epochs 15 and 35. In each of these milestones, the learning rate is scaled down by a factor 0.1, which counterattacks the sudden changes in the loss curve that shows up at first epochs. Weight normalization [16] is also used in the 1D-convolutional layers to speed-up the convergence of the model.

3.3. Subjective evaluation

As this is a generative task that involves synthesized nuances in the speech that are difficult to evaluate with any objective metric, a mean opinion score (MOS) test is conducted. The MOS is a rating of the naturalness of the speech signal with an integer scale ranging from 1 to 5. The meaning of each scale value is translated as Excellent (5), Good (4), Fair (3), Poor (2), and Bad (1).

In total 4 systems are evaluated combining both proposed improvements with all possibilities: (1) speaker dependent normalization and look-ahead (Spk-D + LA); (2) speaker independent normalization and look-ahead (Spk-Ind + LA); (3) speaker dependent normalization (Spk-D); and (4) only speaker independent normalization (Spk-Ind). Hence to perform the test 25 subjects were asked to rate each of the 4 proposed systems under a set of 8 test utterances, one per modelled speaker (4 males and 4 females). In total 32 systems were prompted to be rated per listener, and they could listen the different systems as many times as required to compare and rate them. For each sentence, the transcription of the audio was provided to ease the listening, and the audios of each of the different systems synthesizing the same sentence, were disposed side by side to compare, having a random order per utterance (i.e. the system identity was hidden and mixed among the different utterances).

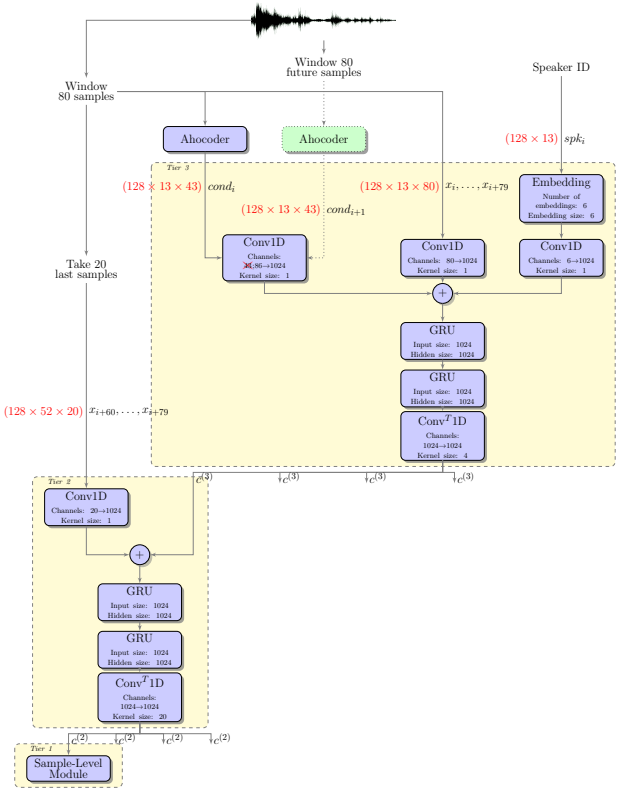

Figure 3: *Conditioned SampleRNN with multi-speaker adaptation*

Normalization: Look ahead:	Spk-D No	Spk-Ind No	Spk-D Yes	Spk-Ind Yes
Female:	3.3	3.3	**3.8**	3.6
Male:	3.6	3.6	**4.0**	3.8
Total:	3.5	3.5	**3.9**	3.8

Table 1: *Table with MOS results comparing the* 4 *proposed systems: (Spk-D + LA), (Spk-Ind + LA), (Spk-D) and (Spk-Ind). For each score the higher the better.*

4. Results

The test results are shown in table 1, obtained by averaging the evaluation of the 25 listeners for every system. Some of the comments written by the subjects that took the test highlighted the difference in quality between males and females. This is the reason why table 1 does a separation between genders that shows that best results are indeed obtained with male voices. This is attributed to the unbalanced speech database, which contained one female speaker with only 25% of recordings compared to all other speakers. Seemingly, this affected the modeling of female voices, which were noisier. Overall we can observe that best scores are achieved when we combine both proposals with the (Spk-D + LA) system.

The samples belonging to the best system combining both proposals can be heard in the GitHub of the main author [1], as well as the code that is publicly available.

[1] https://github.com/Barbany/Multi-speaker-Neural-Vocoder

5. Conclusions

Speaker-dependent normalization achieve substantially better results in naturalness for the generation of speech modeled with balanced datasets when combined with the look ahead approach. Regarding the second proposal, it outperforms the previously achieved scores and reaches state-of-the-art results when combined with the speaker-dependent normalization.

Speaker-dependent normalization was not enough for voice conversion purposes, so more complex architectures were proposed in [2]. Nevertheless, human listeners preferred the speech modeled with speaker-dependent normalization and, given the similarity of the features normalized for each speaker, a better quantization could be applied for coding applications or for deployment of neural networks with limited resources.

Whilst the speaker-dependent normalization itself does not seem to improve the results obtained with the classical speaker-independent feature scaling, when combined with the look ahead approach, it achieves the best score with the balanced dataset. In summary, with the combination of the two proposals, a state of the art MOS score has been achieved for a multi-speaker speech synthesis system. Both of these approaches were novelties introduced in this work and results show that they can be potentially beneficial to other TTS systems as well as for a bunch of other applications involving features from different sources and modelling of non-real-time sequences.

6. Acknowledgements

This research was supported by the project TEC2015-69266-P (MINECO/FEDER, UE).

7. References

[1] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WaveNet: A Generative Model for Raw Audio," ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, pp. 1–15, 2016. [Online]. Available: http://arxiv.org/abs/1609.03499

[2] O. Barbany Mayor, "Multi-Speaker Neural Vocoder," Bachelor's thesis, Universitat Politècnica de Catalunya, 2018.

[3] A. Bonafonte, S. Pascual, and G. Dorca, "Spanish statistical parametric speech synthesis using a neural vocoder," Proc. Interspeech, pp. 1998–2001, 2018.

[4] S. Mehri, K. Kumar, I. Gulrajani, R. Kumar, S. Jain, J. Sotelo, A. Courville, and Y. Bengio, "SampleRNN: An Unconditional End-to-End Neural Audio Generation Model," ICLR, pp. 1–11, 2017. [Online]. Available: http://arxiv.org/abs/1612.07837

[5] Y. Fan, Y. Qian, F. K. Soong, and L. He, "Unsupervised speaker adaptation for DNN-based TTS synthesis," ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, pp. 4475–4479, 2015.

[6] S. Pascual and A. Bonafonte, "Multi-output RNN-LSTM for multiple speaker speech synthesis and adaptation," in 24th European Signal Processing Conference, EUSIPCO 2016, Budapest, Hungary, August 29 - September 2, 2016, 2016, pp. 2325–2329. [Online]. Available: https://doi.org/10.1109/EUSIPCO.2016.7760664

[7] T. Toda, L. H. Chen, D. Saito, F. Villavicencio, M. Wester, Z. Wu, and J. Yamagishi, "The voice conversion challenge 2016," Proc. Interspeech, vol. 08-12-Sept, pp. 1632–1636, 2016.

[8] Y. Wang, R. J. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. V. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, "Tacotron: A fully end-to-end text-to-speech synthesis model," CoRR, vol. abs/1703.10135, 2017. [Online]. Available: http://arxiv.org/abs/1703.10135

[9] A. van den Oord, O. Vinyals, and K. Kavukcuoglu, "Neural Discrete Representation Learning," in NIPS, 2017. [Online]. Available: http://arxiv.org/abs/1711.00937

[10] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation," CoRR, 2014. [Online]. Available: http://arxiv.org/abs/1406.1078

[11] ITU-T. Recommendation G. 711, "Pulse Code Modulation (PCM) of voice frequencies," 1988.

[12] D. Erro, I. Sainz, E. Navas, and I. Hernaez, "Harmonics plus Noise Model based Vocoder for Statistical Parametric Speech Synthesis," IEEE Journal on Selected Topics in Signal Processing, vol. 8, no. 2, pp. 184–194, 2014.

[13] A. Bonafonte, H. Höge, I. Kiss, A. Moreno, U. Ziegenhain, H. V. D. Heuvel, H. Hain, X. S. Wang, and M. N. Garcia, "TC-STAR : Specifications of Language Resources and Evaluation for Speech Synthesis," Proceedings of the Language Resources and Evaluation Conference LREC06, pp. 311–314, 2006.

[14] S. Pascual de la Puente, "Deep learning applied to speech synthesis," Master's thesis, Universitat Politècnica de Catalunya, 2016.

[15] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," CoRR, vol. abs/1412.6980, 2014. [Online]. Available: http://arxiv.org/abs/1412.6980

[16] T. Salimans and D. P. Kingma, "Weight normalization: A simple reparameterization to accelerate training of deep neural networks," CoRR, vol. abs/1602.07868, 2016. [Online]. Available: http://arxiv.org/abs/1602.07868