Laughter Synthesis:
Combining Seq2seq modeling with Transfer Learning

Noé Tits, Kevin El Haddad, Thierry Dutoit

Numediart Institute, University of Mons

{noe.tits, kevin.elhaddad, thierry.dutoit}@umons.ac.be

Abstract

Despite the growing interest for expressive speech synthesis, synthesis of nonverbal expressions is an under-explored area. In this paper we propose an audio laughter synthesis system based on a sequence-to-sequence TTS synthesis system. We leverage transfer learning by training a deep learning model to learn to generate both speech and laughs from annotations. We evaluate our model with a listening test, comparing its performance to an HMM-based laughter synthesis one and assess that it reaches higher perceived naturalness. Our solution is a first step towards a TTS system that would be able to synthesize speech with a control on amusement level with laughter integration.

Index Terms: Laugh synthesis, speech synthesis, TTS, deep learning, transfer learning

1. Introduction and Motivations

Given the progress in speech technologies and Human-Agent Interactions (HAI), several applications of voice assistants and virtual agents have been developed. These applications are evolving towards breaking the barriers between robot-sounding synthetic sounds to human-like conversations. One of the under-explored domains is the synthesis of nonverbal conversational expressions, particularly laughter. Laughter is an important component of speech and daily interactions. It has been shown to be very frequent cross-cultural expression in conversations, to communicate emotions and to have conversational and social functionalities [1, 2, 3].

Laughter can be expressed in many different ways and is particular to each individual. It is therefore rather difficult to collect naturalistic genuine laughter in a sound clean environment. This is the main reason why resources available for synthesis purposes are rather limited compared to speech.

In this paper we present a deep learning-based laughter synthesis system. This work is part of a larger project aiming at synthesizing laughter alongside speech in HAI systems. An initial system was trained from scratch using the same deep-learning based model described here and only laughter data. The results obtained were not satisfying. This was probably due to the limited amount of data available for training. It thus motivated the use of speech data as well. Indeed, although different, both speech and laughs share common sound characteristics since laughs are sequences of fricatives, vowel-sounds and breathing. So, we leverage the knowledge learned by TTS systems when trained with speech in order to improve laughter synthesis.

The paper is organized as follows: related work is summarized in Section 2; Section 3 presents the datasets involved in this work; Section 4 describes the proposed system for audio laughter synthesis; the procedure of the perceptive evaluation is described in Section 5; the results of the evaluation are presented and discussed in Section 6; finally we conclude and detail our plans for future work in Section 7. Most of the resources used here are accessible [1].

2. Related Work

The techniques studied for laughter synthesis have generally followed those used for speech synthesis. In this work we will do the same for two main reasons: first, the signals share a lot of common characteristics; second, using TTS systems will allow to incorporate our laughter synthesis system into a fully functioning TTS system and thus achieve our ultimate goal of TTS with control over amusement levels.

Speech synthesis methods can be grouped in three main categories: synthesis by concatenation, parametric synthesis and statistical parametric synthesis [4]. Among the few studies on laughter synthesis, the first attempts included techniques like synthesis by diphone concatenation [5], parametric synthesis and by using a mass-spring approach [6]. Then Hidden Markov Models (HMM)-based models were introduced to laughter synthesis due to their wide-use in speech synthesis back then [7].

In our previous work we used an HMM-based approach to synthesize laughter alongside amused speech in [8]. Synthetic laughs were obtained by training HMM systems on laughter data. The amused speech was obtained by adapting HMMs trained with another speaker's neutral speech data and adapting them to a smaller dataset containing speech from the speaker from which the laughs were recorded (due to the limited amount of data available for that speaker).

However, deep learning-based laughter synthesis has been little explored yet. A recent approach of synthesis with wavenet was recently proposed by [9]. Wavenet [10] is an autoregressive CNN synthesizing audio sample by sample from features, typically linguistic features for TTS, or acoustic features for vocoding. In their attempt of application to laughter synthesis, they conditioned the wavenet model on information of inhalation/exhalation sequence and parameters of durations and power contour predicted by an HMM model. This approach is therefore still relying on previous HMM approaches for a part of the information.

Given the breakthrough of sequence-to-sequence (seq2seq) approaches in speech synthesis systems, we propose an adapted approach for audio laughter synthesis. Along with the synthesis quality, the method proposed in this paper also offers control over the specific sound sequences to be generated rather than syllable-level control. This allows the flexibility of choosing the specific sequence of voiced (vowels) and unvoiced sounds to be generated. Given the aforementioned goal of obtaining a fully functioning TTS system generating laughter alongside
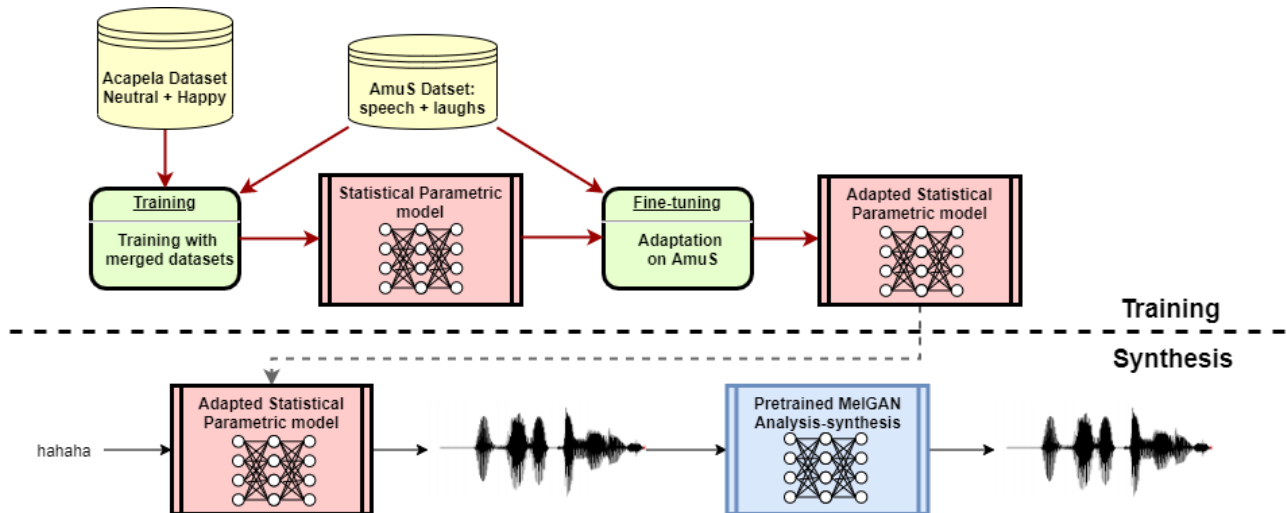
---

[1]https://github.com/numediart/LaughterSynthesis

Figure 1: *Block diagram of the proposed method for model adaptation. The statistical parametric model is based on DCTTS and described in Section 4.1. The text input showed on the figure is symbolic, its detailed format is described in Section 3.*

Table 1: *Quantity of laughs per vowel context.*

Vowel	[a]	[e]	[i]	Total
number of samples	54	33	25	112
duration (sec)	101	63	38	202

speech, another advantage is not only to synthesize naturalistic human-like laughs, but also to do it in a speech context.

This will allow a later integration in a fully functioning speech and laugh synthesis system as planned.

3. Dataset

In order to apply the transfer learning approach described above, the data used are formed of subsets of a proprietary dataset recorded by Acapela and of the AmuS dataset [11].

Acapela's dataset was recorded to build a narrating framework to construct book recordings from transcriptions. It contains phonetically rich sentences uttered by a male actor in US English. The actor was asked to utter a set of the sentences in 8 style classes. For the purpose of this work, only the audio recordings of the neutral style were kept along of the corresponding transcription with a total of 150.50 minutes (3299 utterances) of speech data.

The AmuS dataset contains recordings of amused speech components such as smiled speech, laughs and speech-laughs. For this work, we chose the speaker with the most amount of recorded laughs: SpkB. In this dataset, the purpose of the laughs were to be inserted in speech in order to create an amused effect. This suits well with the goal of generating laughs alongside speech as mentioned above.

In order to record these, the subject was asked to watch stimuli of funny content while sustaining the sound of a vowel, until eventually laughter occurred, naturally interrupting the vowel. This would allow us to collect laughs with transitions from and to vowels. We thus have at our disposal laughs occurring in three vowel contexts: [a], [e], [i] (French IPA symbols). Table 1 breaks down amount of data available.

Isolated laughs are sequences of voiced and unvoiced sounds. In AmuS, the laughs were segmented and each segment was given a label corresponding to a voiced or unvoiced category. These label sequences and the corresponding laughter audio signals were used to train our systems.

4. Seq2seq Audio Laughter synthesis

4.1. System description

Nowadays, one of the major techniques for Text-to-Speech synthesis are deep learning architectures based on the sequence-to-sequence (seq2seq) principle. It consists of an encoder-decoder setup with an interface between the two components called *Attention Mechanism* whose role is to model the alignment between input and output sequences. Well known seq2seq TTS systems are Tacotron [12], Char2wav [13] and DCTTS [14]. In this work we adapted the DCTTS model for audio laughter synthesis using an open implementation available online [2].

In this work, the input sequence is composed of speech phonemes and laughter annotations described in Section 3.

We use festival [15] to extract phones from transcriptions in the Acapela dataset and the speech part of AmuS dataset.

The output sequence is a mel-spectrogram. A second part of DCTTS, trained separately reconstructs a full resolution magnitude spectrogram from the mel-spectrogram to be inverted to a waveform using Griffin-Lim algorithm [16]. For more details about these, see [14].

A first Text-To-Speech and Laughter is trained using the speech from the Acapela dataset due to the quantity of data it provides along with the laughs and smiled speech from the speaker SpkB of the AmuS dataset. This system is then fine-tuned with smiled speech and laughter, both coming from the same AmuS speaker (SpkB). This was done with the perspective of integrating this system into a fully functioning TTS system with control over amusement, as mentioned previously.

This adaptation technique was previously tested on emotional speech and showed promising results in [17], which motivated its use in this work.

---

[2]https://github.com/CSTR-Edinburgh/ophelia

Figure 1 shows a block diagram of the procedure proposed for model adaptation and waveform correction with MelGAN.

4.2. Waveform Correction with MelGAN

To generate the waveform from acoustic features, it has been shown that neural audio synthesizers achieve better quality in terms of naturalness [10, 18]. However it is generally a challenge to design and optimize such models efficiently to reach the expected results described in the literature. They also often lose generalization properties compared to signal processing based vocoders, as they are often speaker dependent.

MelGAN [19] is a recently proposed model that tackled the problems of efficiency and generalization accross speakers. The model is non-autoregressive, fully convolutional and smaller than previous ones.

In this paper we use MelGAN as a waveform corrector. The laughter waveform is first synthesized by the system described in Section 4.1. This waveform contains artifacts due to the Griffin-Lim estimation. Then we apply analysis and synthesis with MelGAN to obtain a corrected laughter waveform that, we show, is perceived as more natural compared to not using MelGAN.

5. Perception Test

To evaluate the obtained results we set-up a perception test using a Mean Opinion Score (MOS) test. We gathered the samples from these different methods:

- Method 1: original laughter samples from AmuS dataset (SpkB) which will serve as a top-line of naturalness for the different methods.

- Method 2: Synthesized laughs based on the HTS system equivalent to the one used in [8]

- Method 3: seq2seq model (seq2seq-GL) described in Section 4.1.

- Method 4: same seq2seq model as method 3 followed by the proposed MelGAN waveform correction (seq2seq-MelGAN)

The MOS test focused on evaluating the naturalness of the synthesized samples of the methods. It was implemented as a web experiment with turkle[3], which is an open-source web server with which one can host a crowdsourcing application locally,

A total of 71 samples were presented to 24 participants in a random order (the participants characteristics are detailed in Table 2)). They were asked to rate each sample in terms of naturalness on a 5-point Likert scale with the following labels: very unnatural (score 1), unnatural (2), fairly natural (3), natural (4) and very natural (5). They could listen to each sample as much times as needed and could stop at any point of the test.

The definition of naturalness for speech and laughter could be interpreted in different ways during evaluation. For example it would have been possible that a participant would rate the acting quality instead of human-likeness of the sound perceived during the listening test. Indeed, "not natural" can be perceived as "fake" or "simulated" instead of "synthetic". But in order to not deviate from previous work and evaluations that used the word "natural", we preferred using it and specify what we mean by it.

---

[3]https://github.com/hltcoe/turkle

Table 2: *Number of participants by gender and age range (in years)*

	Female	Male	Sum
[20,40[	7	13	20
[50,65[	1	3	4
Sum	8	16	24

Table 3: *Number of collected ratings, MOS scores and their standard deviation for each method*

	#ratings	MOS	std
HMM	407	2.64	1.02
seq2seq-GL	431	2.50	1.09
seq2seq-melgan	429	3.28	1.06
original	429	4.10	0.91

That is why, we added an explanation of the meaning of "natural" in the question asked, focusing on the definition of "human-likeness" for "natural".

We also asked some of the participants at the end of the test to comment on what mainly influenced their choices. This mainly serves as our qualitative testing.

6. Results

In this section we present the results obtained from the perception test. These are divided in quantitative and qualitative analyses. The former presents the scores obtained along with interpretations while the latter reports on comments made by the participants on what influenced their choices,

6.1. Quantitative Analysis

A total of 1696 answers were collected. Table 3 gathers the number of ratings for each method, the resulting MOS scores and their standard deviation.

Figure 2 shows the distributions of the ratings as boxplots. Figure 3 shows the percentage of scores chosen for each method.
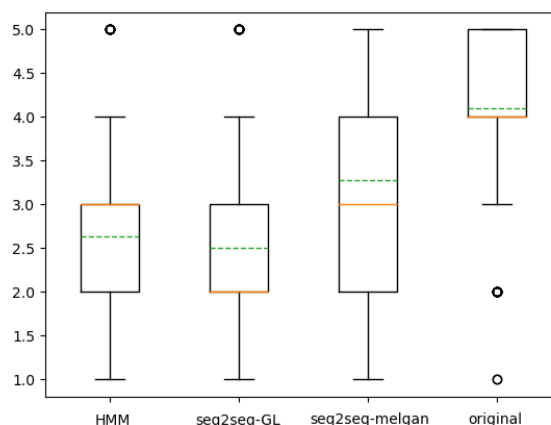

Figure 2: *Boxplots of scores distributions of the different methods. The green lines correspond to the Mean Opinion Scores while the red lines show the median values of the scores.*
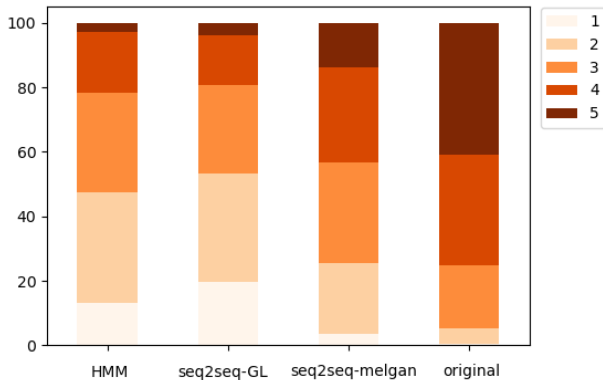
Figure 3: *Score distributions of the different methods. The scale ranges from 1:very unnatural to 5: very natural.*

Original samples of the dataset reach a MOS score that is below 5 as listeners do not rate all original samples as perfect. There is a quite high variance, that is close to one for all methods, which was also the case in [7].

We can clearly see from the results obtained that our seq2seq-MelGAN outperformed both the seq2seq-GL and the HMM-based systems. Both seq2seq-GL and HMM-based system obtained similar MOS. We interpret these results by attributing the loss in MOS obtained, in the latter 2 systems generated laughs, to the distortions found in the seq2seq-GL laughs and the more robotic effect found in the HMMs-obtained laughs. We base our interpretation on previously reported results as detailed in what follows, but also on our personal observations that were not officially tested yet and ones reported by the MOS tests participants as detailed in Section 6.2,

This shows that the seq2seq is efficient at synthesizing laughter, but the question asked in the MOS test to grade naturalness (human-likeness) combined with the distortion generated by the Griffin-Lim algorithm degraded the grades obtained for this test.

Although MOS tests are never executed exactly the same way, it is always interesting to analyze and compare the results obtained here with respect to the results of previous similar experiments found in the literature. This will also partly back our results interpretation.

In [19], the authors compared MelGAN vocoder to the Griffin-Lim algorithm for seq2seq TTS. They reported the Griffin-Lim algorithm to be responsible for a large part of the distortion leading to a loss of 2.95 MOS points compared to original samples. MelGAN, on the other hand, offered a gain of 1.77 points of MOS over the Griffin-Lim algorithm. In this work, the MelGAN waveform correction offered a gain of 0.78 points of MOS. It is important to highlight the fact that training MelGAN directly on generated spectrograms will likely improve our results which was not done in this study but will be part of future work. In [7], the authors compared different variants of HMM-based laughter synthesis. They show that the distortion caused by the vocoder in the copy-synthesis samples is of 0.8 compared to original samples. Their best synthesis solution is 0.6 points below that, and therefore 1.4 below original samples. In this paper, the HMM approach is 1.46 below original samples which is close to their results. In [20], they confirm that HMM-based laughter synthesis have significantly lower quality than copy-synthesis with several vocoders.

6.2. Qualitative Analysis

A part of the participants were asked to comment on what, according to them, influenced their choices in rating the laughs during the test. In this section, we summarise the obtained qualitative results. We report these comments in order to shed a bit more light on the results obtained from the quantitative test. Here is a list of the what some of the participants reported:

- The duration seemed to be another important parameter to consider, as some participants seemed to have based their choice on it. Indeed they reported that some laughs were too long/short to be natural. Also, some participants found short laughs to be ambiguous on a naturalness scale, as if they were focused on finding the fake laughs from the real ones instead of solely focusing grading how natural they perceive the laugh they were listening to.

- Laughs with varying pitch and duration seemed to be perceived as more natural as opposed to laughs with a monotonous prosody. Some laughs were described as "sounding like a repeated sequence" and were perceived as robotic (these would correspond to the HMM generated laughs) whereas laughs containing more randomness were perceived as more natural. Similarly, the ones of which the loudness was fading off (decreasing until the end) were perceived as more natural than ones with a monotonous loudness level and ending abruptly.

Apart from the main parameters influencing degrading decision during the MOS test, we note how important the choice of the question and phrasing is for subjective evaluations such as these.

7. Conclusions and Future Work

In this paper, a new approach of audio laughter synthesis based on seq2seq learning was proposed inspired by the evolution of the TTS field. This system is implemented by leveraging the patterns learned to pass from text to acoustic features in speech, to learn laughter synthesis.

We also use a pretrained MelGAN model as a post waveform corrector allows to remove audio artifacts generated by Griffin-Lim algorithm and thus improve the scores obtained in a MOS test. We believe several modifications could improve the acoustic quality of the synthesis. First end-to-end training could help concerning the accumulation of errors of several blocks: the seq2seq system and the vocoder.

This results in a strong improvement over past methods of audio laughter synthesis (including our own [8]) in terms of naturalness and is promising for later use to build amused speech synthesis systems. The promising results obtained here, allows us to work on incorporating the laughter synthesis system into a fully functioning TTS with control over amusement level. The fact that our laughter synthesis system was developed in a TTS context makes this integration easier.

8. Acknowledgements

9. References

[1] K. Laskowski and S. Burger, "Analysis of the occurrence of laughter in meetings," in *Eighth Annual Conference of the International Speech Communication Association*, 2007.

[2] L. Devillers and L. Vidrascu, "Positive and negative emotional states behind the laughs in spontaneous spoken dialogs," in *Interdisciplinary workshop on the phonetics of laughter*, 2007, p. 37.

[3] M. Soury and L. Devillers, "Smile and laughter in human-machine interaction: a study of engagement." in *LREC*, 2014, pp. 3633–3637.

[4] N. Tits, K. El Haddad, and T. Dutoit, "The Theory behind Controllable Expressive Speech Synthesis: A Cross-Disciplinary Approach," in *Human-Computer Interaction*. IntechOpen, 2019. [Online]. Available: http://dx.doi.org/10.5772/intechopen.89849

[5] E. Lasarcyk and J. Trouvain, "Imitating conversational laughter with an articulatory speech synthesizer," in *Proc. Interdisciplinary Workshop on the Phonetics of Laughter*, 2007.

[6] S. Sundaram and S. Narayanan, "Automatic acoustic synthesis of human-like laughter," *The Journal of the Acoustical Society of America*, vol. 121, no. 1, pp. 527–535, 2007.

[7] J. Urbain, H. Çakmak, and T. Dutoit, "Evaluation of HMM-based laughter synthesis," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 7835–7839.

[8] K. El Haddad, S. Dupont, J. Urbain, and T. Dutoit, "Speech-laughs: An HMM-based Approach for Amused Speech Synthesis," in *Internation Conference on Acoustics, Speech and Signal Processing (ICASSP 2015)*, Brisbane, Australia, 2015, pp. 4939–4943.

[9] H. Mori, T. Nagata, and Y. Arimoto, "Conversational and Social Laughter Synthesis with WaveNet," in *Proc. Interspeech 2019*, 2019, pp. 520–523. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-2131

[10] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu, "WaveNet: A Generative Model for Raw Audio," in *SSW*, 2016.

[11] K. El Haddad, I. Torre, E. Gilmartin, H. Çakmak, S. Dupont, T. Dutoit, and N. Campbell, "Introducing AmuS: The Amused Speech Database," in *Statistical Language and Speech Processing*, N. Camelin, Y. Estève, and C. Martín-Vide, Eds. Cham: Springer International Publishing, 2017, pp. 229–240.

[12] 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: Towards End-to-End Speech Synthesis," in *INTERSPEECH*, 2017.

[13] J. Sotelo, S. Mehri, K. Kumar, J. F. Santos, K. Kastner, A. Courville, and Y. Bengio, "Char2wav: End-to-end speech synthesis," *ICLR2017 workshop submission*, 2017.

[14] H. Tachibana, K. Uenoyama, and S. Aihara, "Efficiently trainable text-to-speech system based on deep convolutional networks with guided attention," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4784–4788.

[15] A. W. Black, P. Taylor, and R. Caley, "The festival speech synthesis system: system documentation," 1997.

[16] D. Griffin and J. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 236–243, 1984.

[17] N. Tits, K. El Haddad, and T. Dutoit, "Exploring Transfer Learning for Low Resource Emotional TTS," in *Intelligent Systems and Applications*, Y. Bi, R. Bhatia, and S. Kapoor, Eds. Cham: Springer International Publishing, 2020, pp. 52–60.

[18] N. Kalchbrenner, E. Elsen, K. Simonyan, S. Noury, N. Casagrande, E. Lockhart, F. Stimberg, A. Oord, S. Dieleman, and K. Kavukcuoglu, "Efficient neural audio synthesis," pp. 2410–2419, 2018.

[19] K. Kumar, R. Kumar, T. de Boissiere, L. Gestin, W. Z. Teoh, J. Sotelo, A. de Brébisson, Y. Bengio, and A. C. Courville, "Mel-gan: Generative adversarial networks for conditional waveform synthesis," in *Advances in Neural Information Processing Systems*, 2019, pp. 14 881–14 892.

[20] B. Bollepalli, J. Urbain, T. Raitio, J. Gustafson, and H. Cakmak, "A comparative evaluation of vocoding techniques for hmm-based laughter synthesis," in *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2014, pp. 255–259.