SEQUENCE-TO-SEQUENCE MODELLING OF F0 FOR SPEECH EMOTION CONVERSION

Carl Robinson, *Nicolas Obin*, *Axel Roebel*

IRCAM, CNRS, Sorbonne Université
Paris, France

ABSTRACT

Voice interfaces are becoming wildly popular and driving demand for more advanced speech synthesis and voice transformation systems. Current text-to-speech methods produce realistic sounding voices, but they lack the emotional expressivity that listeners expect, given the context of the interaction and the phrase being spoken. Emotional voice conversion is a research domain concerned with generating expressive speech from neutral synthesised speech or natural human voice. This research investigated the effectiveness of using a sequence-to-sequence (seq2seq) encoder-decoder based model to transform the intonation of a human voice from neutral to expressive speech, with some preliminary introduction of linguistic conditioning. A subjective experiment conducted on the task of speech emotion recognition by listeners successfully demonstrated the effectiveness of the proposed sequence-to-sequence models to produce convincing voice emotion transformations. In particular, conditioning the model on the position of the syllable in the phrase significantly improved recognition rates.

Index Terms: speech emotion conversion, intonation, sequence-to-sequence models

1. INTRODUCTION

1.1. Speech, Emotion and Conversion

The sound of a human voice is changed as a consequence of the somatic (bodily) effects of emotional responses. Once simply impulsive expressions, emotions have now evolved into an essential component of human communication. Emotion is conveyed by speech prosody (pitch, intensity, speech rate, voice quality) [1], and when speech is interpreted, the prosodic component is given priority over the verbal component [2]. In particular, the intonation, describing the variations of the vocal pitch, or fundamental frequency (F0) is a key aspect of speech emotion that takes place over different time domains, from local contours over the syllables, to global contours over an entire phrase. Consequently, speech emotion conversion involves learning the transfer function between the continuous, variable-length F0 sequences of natural speech and those of expressive speech. The more sophisticated our technology becomes, the greater the need for natural, intuitive interfaces. In this context, voice-enabled emotion-aware interfaces can create very little friction for users and encourage acceptance, leading to greater use, engagement, and higher perceived utility of products. By modelling and transforming speech emotion, we can improve the naturalness of text-to-speech synthesis, and manipulate recordings of human speech, with applications to voice assistants, conversational agents, and sound design.

1.2. Related Works

Modelling speech intonation and associated F0 contours is a challenging task that has been faced in the past decades for a variety of speech applications: from text-to-speech, voice identity conversion, and speech emotion conversion among others. The representation of such F0 variations is a challenging task for at least two main reasons: first, the F0 sequence corresponding to a speech signal is discontinuous by nature: F0 values are only over speech segments that are voiced, and undefined otherwise; second, the F0 varies over multiple time scales associated with pre-defined linguistic units (e.g., syllable, phrase) or with latent units. Accordingly, a number of models have been proposed to model F0 variations : 1) Basically, as a linear sequence of F0 values defined at each time step, either from discontinuous raw F0 values or from continuous interpolated F0 values over voiced instants. 2) As a parametric stylization of the defined F0 values over linguistic units, based on the decomposition of the F0 values over a set of slow time-varying functions, pre-defined as the Discrete cosine transform (DCT) [3] or learned from speech datasets [4]. 3) Using multi-scale modelling, from multi-linear models[5] to more complex models such as the continuous wavelet transform (CWT) decomposition of F0 variations over multiple time scales [6]. These representations have been largely designed and exploited for generative modelling tasks, such as text-to-speech synthesis (TTS) and voice conversion [7, 8, 9, 10, 11, 12].

Over the recent years, recurrent neural networks (RNN) and long-short-term memory (LSTM) cell architectures have been developed to effectively exploit the temporal dependencies in audio data. In particular LSTM-RNN networks have been used to model timbre and prosody variations over time, improving the speech quality over models [13] in text-to-speech synthesis [14, 15] and voice conversion [16, 17]. By including multi-tier links and feedback loops at the frame, phoneme, and syllable levels, the segmental and suprasegmental structure of the F0 contours can be efficiently modelled and preserved during synthesis [18]. A standard RNN can be extended to make both the hidden state and the output values recurrent [15], and separate LSTMs can be used for predicting phone durations and the other acoustic features [19]. Sequence-to-sequence (seq2seq) transcoder models use an encoder/decoder architecture of multi-layered RNN networks to map a variable-length input sequence onto a variable-length output sequence via a fixed size vector [20]. Sequence-to-sequence models have recently been applied with success for text-to-speech synthesis [21, 22], approaching WaveNet [23] in terms of quality at lower computational cost and latency. At the time of writing, no paper that demonstrated a sequence-to-sequence being applied to the task of speech emotion conversion could be found.

This paper presents a seq2seq modelling of F0 for speech emotion conversion [†]. In the above research on voice transformation, the proposed models learns to predict the F0 values corresponding

[†]Project code and audio samples are available online at: `https://github.com/carl-robinson/voice-emotion-seq2seq`

to the target emotion without considering the actual F0 values of the neutral speech. Moreover, the F0 values and the durations of the F0 contours are modelled and processed separately during learning and prediction. For instance, the duration of the F0 contours are often normalised over time by stylisation methods such as the DCT and CWT, so instead the durations are modelled and processed separately to the F0. In contrast, the proposed sequence-to-sequence models simultaneously represent both the pitch values and the duration of the contours, so pitch and time information are jointly modelled. Furthermore, the output contour can be easily conditioned on both the source input signal and its linguistic context, presenting the opportunity for improved speech emotion conversion.

2. PROPOSED MODEL

The proposed model for the F0 conversion is a transcoder that maps between a source sequence for neutral speech and a target sequence for a single emotion (i.e., anger, sadness, joy, or fear). A voice transformation system capable of converting neutral speech to emotive speech is then constructed, based on the sequence-to sequence neural network architecture (seq2seq). The conversion process involves three main steps: 1) an extraction of the F0 contours from the neutral source speech signal; 2) a transformation of those contours using the seq2seq model; 3) an application of the transformed F0 contours back into the neutral source speech signal, which produces a new speech signal containing the desired expressive form of the utterance. The remaining of this section introduces the core seq2seq architecture used and the application of the seq2seq network predictions to convert the original speech signal.

2.1. Sequence-to-Sequence Architecture

2.1.1. Encoder-Decoder Architecture

An auto-encoder is a variant of a neural network in which the output is the approximation of the input data. It is composed of an encoder module that learns a latent lower-dimension representation of the data, and a decoder that reconstructs the observed data from this latent code. A transcoder is similar, except that the objective is no longer to approximate the input vector, but another data vector. Basic auto-encoders/transcoders do not process sequences, and the input and output data must be the same length. The seq2seq encoder-decoder architecture overcomes these limitations [20].

For speech emotion conversion, the input and output vectors are variable length sequences of pitch values calculated on each syllable/phoneme, the fundamental building blocks of speech prosody. Let $\mathbf{x} = [x_1, \ldots, x_{T_x}]$ the source F0 sequence corresponding to neutral speech, and $\mathbf{y} = [y_1, \ldots, y_{T_y}]$ the target F0 sequence corresponding to emotional speech. The seq2seq model is trained to predict the target sequence conditionally to the source sequence,

$$p(\mathbf{y}|\mathbf{x}) = \prod_{t=1}^{T_y} p(y_t|\mathbf{y}_{<t}, \mathbf{x}) \tag{1}$$

To do so, an encoder RNN is first used to map the variable length input $\mathbf{x}$ into a fixed length context vector $\mathbf{c}$ [24],

$$h_t = f(x_t, h_{t-1}) \tag{2}$$
$$c = g(h_1, \ldots, h_{T_x}) = h_{T_x} \tag{3}$$

where: f is the recurrent function, for instance a RNN-LSTM.

Then, a decoder RNN is used to map the fixed length code $\mathbf{c}$ to the target sequence $\mathbf{y}$,

$$p(\mathbf{y}|\mathbf{x}) = \prod_{t=1}^{T_y} p(y_t|\mathbf{y}_{<t}, \mathbf{c}) \tag{4}$$

The decoder also uses teacher forcing, in that it predicts the next element of a target sequence by taking into account the element it predicted on the previous step.

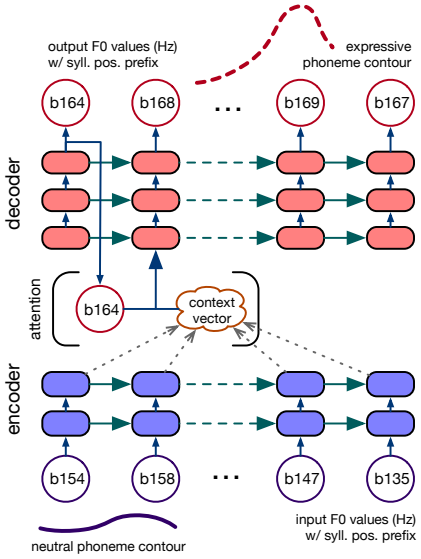

Fig. 1. Encoder-decoder sequence-to-sequence architecture with attention mechanism.

During inference, the predicted target sequence is obtained by maximising the conditional probability:

$$\hat{\mathbf{y}}_{[1:T_y]} = \underset{\mathbf{y}_{[1:\mathbf{T_y}]}}{\operatorname{argmax}}\ p(\mathbf{y}_{[\mathbf{1}:\mathbf{T_y}]}|\mathbf{c}) \tag{5}$$

The prediction starts by supplying the decoder with a target sequence of length 1 (the start of sequence character, 'SOS'), and the context vector. It then calculates prediction probabilities for all the possible values in the embedding, and uses the argmax to select the next value (i.e. the first F0 value in the converted output sequence). It then feeds the output value back into the decoder in order to generate the next value in the sequence, and repeats this process until the decoder generates an end of sequence character 'EOS'.

2.1.2. Attention Mechanism

Attention mechanisms augment the performance of the basic seq2seq model by allowing the implicit alignment of input and output sequences, and are the standard today in seq2seq architectures [25, 26], with some recent applications to speech processing [27]. Basically, the attention mechanisms allows to perform an internal alignment between the source and the target F0 sequences, and to identify the most important past F0 values to be used for conversion at the current time of the target sequence. An illustration of the seq2seq architecture with the attention mechanism for the conversion of F0 contours is provided in Figure 1.

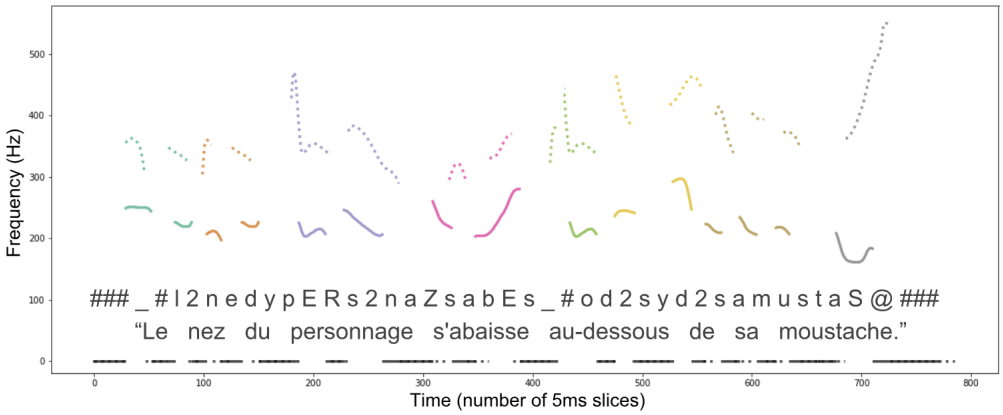

Fig. 2. Original and transformed voiced and unvoiced phoneme F0 contours across an entire phrase.

2.1.3. Implementation Details

The seq2seq model was learned separately on the F0 contours corresponding to the vowel of each syllable, and additional linguistic conditioning was optionally added to indicate the position of the syllable within the phrase. The following specifications were used for the implementation of the proposed seq2seq model:

Architecture: Our encoder contained two layers of 128 bidirectional LSTM cells, to capture the full F0 contours of each phoneme (Figure 1). Dropout on the input values was set at 0.5. Residual connections were not used as these conflicted with the attention mechanism, essentially bypassing the non-linearity functions and passing the raw input values directly to the attention mechanism. Our decoder was the same except that it had three layers, and residual connections were used. LSTM peepholes degraded results significantly, so were not used.

F0 embedding: As the seq2seq model used an embedding layer, the real-valued F0 values were quantized beforehand to produce a finite range of discrete integers. This was restricted to the range 50 to 550 Hz, to encompass the vocal ranges of both males and females.

Optimization: Cross entropy, which indicates the distance between the learned and observed distributions, was used as the loss function for learning. A softmax activation layer was used as the output layer of our neural network. Our proposed model outputs a probability for each element in the target embedding (each a unique F0 integer value, restricted to the range 50 to 550 Hz). The ADAM optimiser was used for this, with an initial learning rate of 0.001.

2.1.4. Linguistic Conditioning

To add further context to the model, a preliminary conditioning of the network on the linguistic context was investigated. In particular, the position of the syllable within a phrase is known as one of the most important factor of variations of the syllable F0 contour, especially for the initial and last syllables of the phrase. To do so, the F0 values for a syllable were tagged with a prefix that indicates whether they belong to the first, last or other positions within the phrase. This caused the model to treat the three types of contour separately, and allowed it to generate the important inflections often found at the start and the end of a phrase. Models with and without this conditioning were also trained for comparison in the experiment.

2.2. Neutral to Expressive Conversion

The trained seq2seq model is then used to predict the expressive F0 contours for all the vowels of a speech utterance in the desired expressivity. In Figure 2, the solid lines are the original neutral contours, while the dotted lines are the transformed expressive contours (in this case, joy). Like-phonemes have the same colour, while black lines are unvoiced phonemes or periods of silence. The predicted contours are applied to the neutral speech signal using the superVP phase vocoder, by time-stretching the neutral speech to fit with the predicted F0 contour lengths, and linearly interpolating the predicted F0 contours in the unvoiced and silence segments, and applying the F0 conversion only on the voiced frames if the warped neutral speech. The harmonicity (harmonic to noise ratio) of the neutral speech is used to determine which parts of the interpolated F0 contour to use for conversion. Using a conservative value of 0.7 produced the least audible distortion.

3. EXPERIMENT

An experiment was conducted to evaluate and compare the proposed seq2seq architectures, consisting in a speech emotion recognition task by human listeners of original acted speech utterances and converted neutral speech utterances.

3.1. Speech Emotion Database

The emotional speech database used for this study is the same as the one previously used and described in [11]. The database comprises one female French actor speaking 10 different utterances with 4 emotions (joy, fear, sadness, anger), each acted with 5 levels of intensity (i01-i05), plus a neutral version (i00). This provided 40 recorded utterances of neutral speech for source data for each intensity, and a total 200 expressive speech utterances for all intensities ($10 \times 4 \times 5$). Recordings were stored in 16-bit uncompressed audio files at 48000 KHz, and corresponding prompts were stored in plain text files with UTF-8 encoding. The fundamental frequency (F0) values were estimated using SuperVP, and then linearly interpolated between voiced regions. Phoneme, syllable, and phrase alignment were created by using the ircamAlign library [28] from the speech files and the text prompts. Finally, the F0 contours were created for each vowel of the speech utterance, and linguistic context was calculated from the alignement by calculating the position of the syllable within the corresponding phrase.

		Perceived			
		Joy	Sadness	Anger	Fear
Original Acted	Joy	**91.3%**	2.4%	6.3%	0.0%
	Sadness	4.4%	**85.7%**	0.0%	33.3%
	Anger	2.2%	2.4%	**70.3%**	7.8%
	Fear	2.1%	9.5%	23.4%	**58.9%**
	Total	100%	100%	100%	100%

Table 1. Confusion matrix of participant responses for the original emotion samples, as performed by the actor.

		Perceived			
		Joy	Sadness	Anger	Fear
Converted *with* Cond.	Joy	**74.8%**	13.0%	9.2%	12.9%
	Sadness	17.6%	**40.2%**	18.5%	22.5%
	Anger	5.9%	13.6%	**64.9%**	14.5%
	Fear	1.7%	33.2%	7.4%	**50.1%**
	Total	100%	100%	100%	100%
Converted *w/o* Cond.	Joy	**67.8%**	20.9%	30.7%	17.0%
	Sadness	19.4%	**25.7%**	7.7%	25.5%
	Anger	9.6%	25.5%	**42.4%**	21.3%
	Fear	3.2%	27.9%	19.2%	**36.2%**
	Total	100%	100%	100%	100%

Table 2. Confusion matrices of participant responses for emotion conversion with linguistic conditioning (top), and without (bottom).

3.2. Model Setups

A seq2seq model was learned for each emotion separately, from the source and target F0 contours coming from the parallel utterances of the speech emotion database. Since the speech emotion database is parallel, each F0 syllable contour of a neutral utterance can be systematically paired with a its corresponding F0 contour in another expressivity. For learning, the source sequence was chosen as the F0 contour corresponding to the neutral version of the syllable and the target sequence as the F0 contour corresponding to the expressive version of the syllable. The complete dataset of source/target pairs of F0 contours was first split by utterances, to ensure that the utterances used for training were not used for testing. Of the 10 utterances available, 6 were used for training and 4 for testing. The training data was then split again, 85% for training and 15% for validation. This resulted in around 1100 syllable F0 contour pairs used for training, and 170 used for validation.

3.3. Experiment Methodology

Three sets of 32 files were manually selected for evaluation: 32 from the syllable-position conditioned model; 32 from the non-conditioned model; and 32 from the original non-converted expressive samples provided by the actor (for use as a control). Each set comprised 8 samples from each of the 4 emotions. Additionally, each emotion set contained 2 examples for each of the 4 test phrases, to reduce the bias from the wording of the samples influencing the participant's choice. A total of 96 files were available for evaluation. The experiment consisted in a on-line speech emotion recognition task by human listeners. The survey asked participants to identify the emotion for 20 files, selected at random for each participant from the pool of 96 files. For each audio file, the participant were asked to select one emotion among four possible (happiness, sadness, anger, fear). Participants were encouraged to use headphones and to do the experiment in quiet listening conditions.

3.4. Results and Discussion

The survey was completed by 87 participants, providing 1734 responses. Table 3.2 presents recognition rates obtained with the original acted speech utterances, and Table 2 those obtained from the conversion obtained by the proposed model with and without linguistic conditioning.

Firstly, the recognition rates obtained for the original acted samples used as control highlight the difficulty of this task. For instance, joy and sadness are strongly recognised by participants (91.3% and 85.7%) whereas anger and fear are much more ambiguous (70.3% and 58.9%). Fear is very often confused with anger which may be due to the actor's performance, or to a general ambiguity that exists between emotions expressed by speech. Secondly, the recognition rates obtained for the converted speech are fairly good and consistent with the ones reported for the acted emotions. In particular, converted joy, anger, and fear were consistently recognised (respectively, 74.8%, 64.9%, and 50.1%), at rates comparable to their original expressions. Converted sadness is the exception, being much less well recognised than its original version, and often perceived as fear.

The model with linguistic conditioning performed considerably better across all four emotions. Participants correctly identified joy 74.8% of the time, an increase of 7.0% over the non-conditional model. Responses improved by 13.9% for fear, and by 14.5% for sadness. In the non-conditioned model results, sadness was indistinguishable from the other three emotions, whereas in the conditioned model results, while still often confused for fear, the majority of sadness responses were correct. The results for anger improved the most, increasing by 22.5% to a value of 64.9%. By comparison, the results obtained by the proposed seq2seq model are only slightly worse than those obtained on the same speech emotion database by [11]. This is an encouraging result as, unlike the model proposed by [11], the proposed approach explicitly modelled the durations of the syllables, and did not use forced alignment of the predicted durations to the original ones for the experiment.

4. CONCLUSION

In this paper, we presented a sequence-to-sequence architecture for speech emotion conversion based on F0 conversion learned from parallel databases. Experimental results showed can generate F0 contours that can convert the emotion of neutral utterances effectively. The addition of syllable position conditioning helped to improve the quality of the conversion for all emotions, especially anger. The sequence-to-sequence architecture will be next applied to multiple speakers in order to improve the generalisation ability of the network and its adaptation to a specific speaker's strategy. Also, further research will focus on extending the sequence-to-sequence architecture to allow a parametric modelling and conversion of emotional speech (F0, duration, intensity, voice quality). Finally, more advanced linguistic embedding strategies will be explored.

5. REFERENCES

[1] K. R. Scherer, R. Banse, H. G. Wallbott, and T. Goldbeck, "Vocal cues in emotion encoding and decoding," *Motivation and Emotion*, vol. 15, p. 123–148, 1991.

[2] P. Taylor, *Text-to-Speech Synthesis*. Cambridge University Press, 2009.

[3] J. Teutenberg, C. Watson, and P. Riddle, "Modelling and Synthesising F0 contours with the Discrete Cosine Transform," in *International Conference on Acoustics, Speech, and Signal Processing*, Las Vegas, U.S.A, 2008, pp. 3973–3976.

[4] N. Obin and J. Belião, "Sparse coding of pitch contours with deep auto-encoders," in *International Conference on Speech Prosody*, 2018, pp. 799–803.

[5] G. Branislav, G. Bailly, O. Mohammed, Y. Xu, and P. N. Garner, "A variational prosody model for the decomposition and synthesis of speech prosody," in *ArXiv e-prints*, 2018. [Online]. Available: arxiv.org/abs/1806.08685

[6] Z. Luo, J. Chen, T. Takiguchi, and Y. Ariki, "Emotional voice conversion using neural networks with arbitrary scales F0 based on wavelet transform," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2017, no. 1, 2017.

[7] A. Black, H. Zen, and K. Tokuda, "Statistical parametric speech synthesis," in *International Conference on Audio, Speech, and Signal Processing*, 2007, pp. 1229–1232.

[8] J. Latorre and M. Akamine, "Multilevel parametric-base F0 model for speech synthesis," in *Interspeech*, Brisbane, Australia, 2008, pp. 2274–2277.

[9] N. Obin, A. Lacheret, and X. Rodet, "Stylization and Trajectory Modelling of Short and Long Term Speech Prosody Variations," in *Interspeech*, Florence, Italy, 2011, pp. 2029–2032.

[10] N. Obin, "MeLos: Analysis and Modelling of Speech Prosody and Speaking Style," PhD. Thesis, IRCAM - UPMC, 2011.

[11] C. Veaux and X. Rodet, "Intonation conversion from neutral to expressive speech," in *Interspeech*, Jan. 2011, pp. 2765–2768.

[12] X. Yin, M. Lei, Y. Qian, F. K. Soong, L. He, Z.-H. Ling, and L.-R. Dai, "Modeling F0 trajectories in hierarchically structured deep neural networks," *Speech Communication*, vol. 76, pp. 82–92, 2016.

[13] H. Zen, "Statistical parametric speech synthesis: from HMM to LSTM-RNN," 2015.

[14] Y. Fan, Y. Qian, F.-L. Xie, and F. K. Soong, "TTS synthesis with bidirectional LSTM based recurrent neural networks," *Interspeech*, 2014.

[15] H. Zen and H. Sak, "Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.

[16] H. Ming, D. Huang, L. Xie, J. Wu, M. Dong, and H. Li, "Deep bidirectional LSTM modeling of timbre and prosody for emotional voice conversion," in *Interspeech 2016*, 2016.

[17] R. Li, Z. Wu, H. Meng, and L. Cai, "DBLSTM-based multi-task learning for pitch transformation in voice conversion," in *2016 10th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, 2016.

[18] X. Wang, S. Takaki, and J. Yamagishi, "An RNN-Based quantized F0 model with Multi-Tier feedback links for Text-to-Speech synthesis," in *Interspeech 2017*, 2017.

[19] S. Ronanki, G. E. Henter, Z. Wu, and S. King, "A Template-Based approach for speech synthesis intonation generation using LSTMs," in *Interspeech 2016*, 2016.

[20] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in Neural Information Processing Systems (NIPS)*, 2014.

[21] 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*, 2017. [Online]. Available: http://arxiv.org/abs/1703.10135

[22] V. Wan, Y. Agiomyrgiannakis, H. Silen, and J. Vít, "Google's Next-Generation Real-Time Unit-Selection synthesizer using Sequence-to-Sequence LSTM-Based autoencoders," in *Interspeech*, 2017.

[23] 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," 2016.

[24] K. Cho, B. Merrienboer, C. Gulcehre, F. Bougares, H.Schwenk, and Y. Bengio., "Learning phrase representations using rnn encoder-decoder for statistical machine translation," in *Empirical Methods in Natural Language Processing (EMNLP)*, 2014.

[25] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *International Conference on Learning Representations (ICLR)*, 2014.

[26] C. Raffel, M.-T. Luong, P. J. Liu, R. J. Weiss, and D. Eck, "Online and linear-time attention by enforcing monotonic alignments," in *International Conference on Machine Learning (ICML)*, 2017.

[27] H. Sak, M. Shannon, K. Rao, and F. Beaufays, "Recurrent neural aligner: An encoder-decoder neural network model for sequence to sequence mapping," in *Interspeech*, 2017, pp. 1298–1302.

[28] P. Lanchantin, A. Morris, X. Rodet, and C. Veaux, "Automatic Phoneme Segmentation with Relaxed Textual Constraints," in *International Conference on Language Resources and Evaluation*, Marrakech, Morroco, 2008, pp. 2403–2407.