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

Carl Robinson, *Nicolas Obin*

IRCAM, CNRS, Sorbonne Université
Paris, France

ABSTRACT

Voice interfaces are becoming wildly popular and driving demand for more advanced speech synthesis 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. Experiment participants indicated the emotions they perceived in converted and original voice samples. It was discovered that, while highly phrase-dependent, conditioning the model on the position of the syllable in the phrase significantly improved recognition rates. This paper successfully demonstrates the use of sequence-to-sequence models to produce convincing voice emotion transformations that rival those of previous studies.

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

1. INTRODUCTION

1.1. Speech, Emotion and Conversion

The more sophisticated our technology becomes, the greater the need for natural, intuitive interfaces. Voice-enabled emotion-aware interfaces 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. Real-world applications include conversational interface design, generating dialogue for film and video-games, real-time voice transformation, and sound design. 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, 2, 3], and when speech is interpreted, the prosodic component is given priority over the verbal component [4]. Variations in the vocal pitch, or fundamental frequency (F0), are known as intonation. Intonation 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.

1.2. Related Works

Modelling F0 contours is challenging due to their discontinuous nature; the F0 is only defined for the voiced phonemes, not the whole phrase. F0 contours can be modelled in a number of ways: 1) Linearly, as a sequence of raw F0 values defined at each time step. This can include quantised F0 values rather than interpolated continuous valued contours [5]. 2) As a stylisation over linguistic units. Discrete cosine transform (DCT) coefficients are used at the syllable level to model the temporal correlation between syllables [6, 7]. Continuous wavelet transform (CWT) decompositions of F0 and energy contours into wavelets describe prosody patterns in different temporal scales [8]. Phonetically-aware sparse representations of have also been created using CWT [9]. 3) Using multi-linear/super-postionnal modelling [10, 11], for example in text-to-speech (TTS) [12, 13].

Recurrent neural networks (RNN) and long-short-term memory (LSTM) cell models effectively exploit the temporal dependencies in audio data, and have modelled timbre and prosody [14], and also pitch transformation [15, 16]. They produce more natural sounding voice synthesis conversion at lower-latency than HMM models [17, 18]. By modelling the spectral (frequency) and the temporal (pitch contour) features together in a 2D time-frequency LSTM (TFLSTM), a structured output layer (SOL) can then capture dependencies between the two [19]. By including multi-tier links and feedback loops at the frame, syllable and phoneme levels, the segmental and suprasegmental correlations between F0 contours and the unvoiced regions can be preserved [5]. A standard RNN can be extended to make both the hidden state and the output values recurrent [18]. Separate LSTMs can be used for predicting phone durations and the other acoustic features [20].

Sequence-to-sequence (seq2seq) transcoder models use a pair of multi-layered RNN networks to map an input sequence to an output sequence via a fixed size vector [21]. An LSTM-based sequence-to-sequence text-to-speech (STT) model has approached WaveNet in terms of quality at low computational cost and latency [22, 23]. At the time of writing, no paper that demonstrated a sequence-to-sequence being applied to the task of voice transformation could be found. In the above research, the durations of the F0 contours are often normalised (lost) by stylisation methods such as the DCT and CWT, so instead the durations are modelled and processed separately to the F0. Additionally, the existing F0 contour transformation models are generally conditioned on the linguistic context, not on the source neutral speech signal to be converted. In contrast, sequence-to-sequence models simultaneously represent both the pitch values and the duration of the contours, so no information is lost. Furthermore, the output contour can be easily conditioned on both the source input signal and its linguistic context, presenting the opportunity for improved voice transformations.

2. PROPOSED MODEL

The proposed model for the F0 conversion is a transcoder that maps between two sequences of pitch values; 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 was built, based on a sequence-to sequence neural network architecture. 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.

2.1. Sequence-to-Sequence Architecture

2.1.1. Introduction to the Basic 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 trans-coder is similar, except that the objective is no longer to approximate the input vector, but another data vector. Basic auto-encoders/trans-coders do not process sequences, and the input and output data must be the same length. The seq2seq encoder-decoder architecture overcomes these limitations [21]. 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,

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

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],

$$h_t = f(x_t, h_{t-1}) \quad (2)$$

$$c = g(h_1, \ldots, h_{T_x}) = h_{T_x} \quad (3)$$

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 $\mathbf{y}$,

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

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.

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

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

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 (sample) 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'.

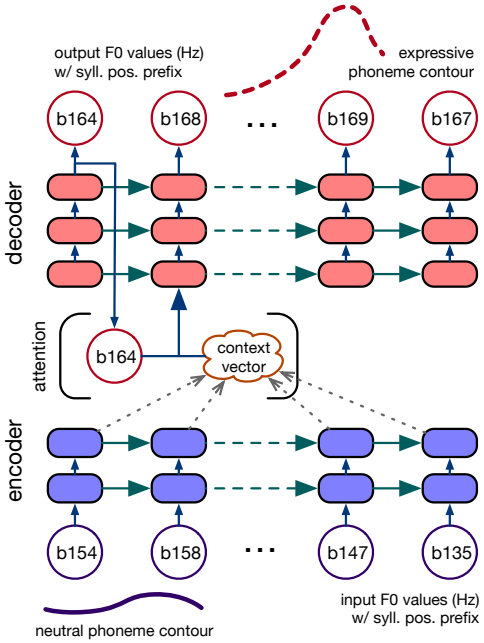

Fig. 1. Encoder-decoder sequence-to-sequence architecture with attention mechanism.

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 defacto standard today [25, 26]. F0 contours are relatively long sequences for a seq2seq model, and using attention ensures the information from the start of the sequence is sufficiently exploited. Unlike the basic seq2seq architecture, our proposed model only used the attention mechanism [27] to link the encoder to the decoder (Figure 1).

2.1.3. Our Loss and Optimisation Functions

Cross entropy loss (log loss) indicates the distance between the learned and observed distributions. It was used as have a softmax activation in 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). Cross entropy loss is defined in Equation 6.

$$H(p, q) = -\sum_i p_i \log(q_i) = -y \log \hat{y} - (1-y) \log(1-\hat{y}) \quad (6)$$

where: $p \in \{y, 1-y\}$ the set of true labels (0 or 1), $q \in \{\hat{y}, 1-\hat{y}\}$ the set of predicted labels, y is a true label, $\hat{y}$ is a predicted label probability, and H is the mean of the cross entropies of all observations in the sample i.e. the cost. Minimising the cross entropy over the target embedding of F0 integers maximises the log

likelihood of the predictions. The ADAM optimiser was used for this, with an initial learning rate of 0.001.

2.2. Signal Conversion Process: Neutral to Expressive

During inference, the trained seq-to-seq model is used to predict the expressive F0 contours for all voiced phonemes in a phrase. 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. These transformed contours were applied to the neutral speech signal using SuperVP [28]:

Time stretching the source speech signal: Each corresponding section of the neutral speech signal was time-stretched to fit the new contour length.

Interpolation across unvoiced sections: Linear interpolation was performed over the silence and unvoiced sections to eliminate gaps.

Harmonicity threshold: Harmonicity (signal to noise ratio) of the neutral file was 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

3.1. Speech Emotion Database

Audio data for training: To generate samples for the experiment, the model was trained on a portion of a parallel audio database of studio recordings, Olivia2006, developed at IRCAM. Recordings were raw 16-bit WAVs at 48000 KHz. This portion comprised: 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' one (i00). This provided 40 recorded utterances of neutral speech ($10 \times 4 \times 1$), for source data, and 200 of expressive speech ($10 \times 4 \times 5$), for target data.

F0 extraction, interpolation and alignment: The fundamental frequency (F0) values were extracted using AudioSculpt, providing continuous sequences of real valued contours for the voiced sections of speech, interspersed by sections of silence corresponding to the unvoiced sections (noise). To facilitate model training, we interpolated between the voiced sections to create an unbroken F0 signal contour. Using IrcamAlign [29] and the transcribed text, syllable and phoneme alignments were calculated, and used to split the phrase level F0 contour into a set of smaller syllable/phoneme level contours, for use as source and target training input.

3.2. Model Setups

Architecture used for training: Our encoder contained two layers of 128 bidirectional LSTM cells [17, 20], to capture the full F0 contours of each phoneme (Figure 1). Dropout on the input values was set at 0.8. 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 rounding and source/target pairing: As the seq2seq model used an embedding layer, the real-valued F0 values were rounded 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. The training set of source/target F0 contour pairs used the neutral intensity i00 files as the source, and various combinations of intensities (i01-i03, i03-i05, and i01-i05) as the target.

Voiced phonemes only: Splitting a phrase at the syllable level would train the model on F0 contours that included the unvoiced parts of a syllable, introducing noise. Therefore the data was split by phoneme, and only the *voiced* phonemes were used for training.

Matching phoneme position in the same phrase: To preserve the context of each phoneme, only phoneme contours from the same position in the same phrase were matched as source/target pairs.

With and without conditioning on syllable position in the phrase: To add further context to the model, the F0 values for a phoneme were tagged with a prefix that indicated whether they belonged to the first, last or other syllable in 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 enhancement were trained for comparison in the experiment.

Train, validation and test split: The complete dataset of source/target pairs of phoneme contours was first split by phrase, to ensure that the phrases used for training were not used for testing. Of the 10 phrases available, 6 were used for training, and 4 for testing i.e. a 60% training data / 40% test data split. The training data was then split again, 85% for the training set / 15% for the validation set. This resulted in around 1100 phoneme contour pairs used for training, and 170 used for validation.

Conversion of neutral audio files: After training and inference of the models to be evaluated, the F0 contours were applied to the neutral speech signal using a harmonicity threshold of 0.7. The files were normalised to the same volume level, and converted to 160 Kbps MP3 format.

3.3. Experiment Methodology

Selection of converted audio files: 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.

Survey evaluation of emotion conversions: An online survey tool with audio playback facilities was used, allowing participants to hear a voice sample, then select an emotion from four options (happiness, sadness, anger, fear). Additionally, they judged how natural the voice sounded, and select a corresponding voice quality (bad, poor, fair, good, excellent). The survey asked participants to identify the emotion for 20 files, selected at random for each participant from the pool of 96 files. Participants also indicated their level of French (native, non-native, cannot speak French), and the listening equipment used (headphones, earphones, speakers). Participants included both fellow researchers and the general public.

3.4. Results and Discussion

The survey was completed by 87 participants, providing 1734 responses. Here we compare the recognition rates of the original emotion samples (Table 3.4) with those of the proposed model, both with and without linguistic conditioning (Table 2). Firstly, the recognition rates obtained for the original samples used as control highlight the difficulty of this task. For instance, joy and sadness are strongly

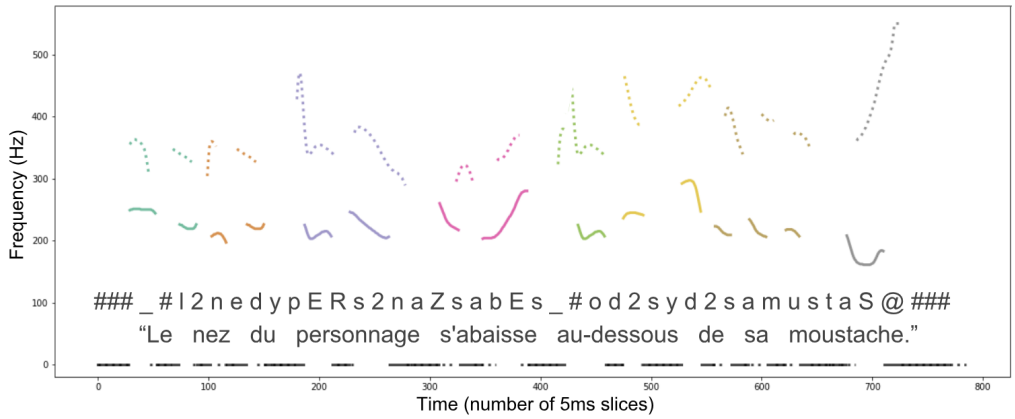

Fig. 2. Original and transformed voiced and unvoiced phoneme F0 contours across an entire phrase.

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 percentage points over the non-conditional model. Responses improved by 13.9 points for fear, and by 14.5 points 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 points 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 [6]. This is an encouraging result as, unlike the model proposed by [6], 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. Finally, the naturalness of the speech was judged to be between good and excellent for the original speech samples (4.22 in mean, with a standard deviation of 0.84), while the converted samples were judged to be between fair and good (3.40 in mean, with a standard deviation of around 1). This slight drop in naturalness, commonly reported in speech transformation, is encouraging for the quality of the conversion, and may partly explain the greater difficulty to correctly recognise an emotion from degraded speech.

		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).

4. CONCLUSION

In this paper, we presented a supervised method for voice emotion conversion based on a sequence-to-sequence architecture. Experimental results showed that despite some of the original emotions being ambiguous to participants, notably sadness and fear, the model was able to extract pertinent features for each emotion, and generate F0 contours that can convert the emotion in neutral utterances effectively. The addition of syllable position conditioning was an important innovation that helped improve results for all emotions, especially anger. Further research will use a larger parallel database of multiple speakers and genders to improve generalisability, focus on modelling inter-syllable correlations, and introduce new features such as voice amplitude, syllable stress, duration of silences and unvoiced syllables, and linguistic information. Project code and audio samples are available online [1].

[1] https://github.com/carl-robinson/voice-emotion-seq2seq

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] J. Ohala, "Ethological theory and the expression of emotion in the voice," in *International Conference on Spoken Language Processing*, Philadelphia, USA, 1996, pp. 1812–1815.

[3] J. A. Bachorowski, "Vocal expression and perception of emotion," *Current Directions in Psychological Science*, vol. 8, pp. 53–57, 1999.

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

[5] 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.

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

[7] S. Ronanki, G. E. Henter, Z. Wu, and S. King, "A template-based approach for speech synthesis intonation generation using LSTMs," in *Interspeech*, 2016, p. 2463–2467.

[8] 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.

[9] B. Sisman, H. Li, and K. C. Tan, "Transformation of prosody in voice conversion," in *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, Dec. 2017, pp. 1537–1546.

[10] B. Holm, "SFC : un modèle de superposition de contours multiparamétriques pour la génération automatique de la prosodie - apprentissage automatique et application l'énonciation de formules mathématiques," PhD. Thesis, Institut de la Communication Parlée, Grenoble, 2003.

[11] B. Gerazov, G. Bailly, O. Mohammed, and P. N. Garner, "A variational prosody model for the decomposition and synthesis of speech prosody," in *Speech Prosody*, Poznan, Poland, 2018.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

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

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

[18] 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.

[19] R. Li, Z. Wu, Y. Ning, L. Sun, H. Meng, and L. Cai, "Spectro-Temporal modelling with Time-Frequency LSTM and structured output layer for voice conversion," in *Interspeech 2017*, 2017.

[20] 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.

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

[22] 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," Sep. 2016.

[23] 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*, 2017.

[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] Thang Luong, Eugene Brevdo,, "Neural machine translation (seq2seq) tutorial — TensorFlow," https://www.tensorflow.org/tutorials/seq2seq, accessed: 2018-3-23.

[26] D. Bahdanau, K. Cho, and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate," *ArXiv e-prints*, Sep. 2014.

[27] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *CoRR*, vol. abs/1409.0473, 2014. [Online]. Available: http://arxiv.org/abs/1409.0473

[28] "Supervp (super phase vocoder), a library and executable for analysis, synthesis, and transformation of sounds." [Online]. Available: http://anasynth.ircam.fr/home/english/software/supervp

[29] 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.