INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

A Multi-Speaker Emotion Morphing Model Using Highway Networks and Maximum Likelihood Objective

Ravi Shankar, Jacob Sager, Archana Venkataraman

Department of Electrical and Computer Engineering, Johns Hopkins University
rshanka3@jhu.edu, jsager@jhu.edu and archana.venkataraman@jhu.edu

Abstract

We introduce a new model for emotion conversion in speech based on highway neural networks. Our model uses the contextual pitch, energy and spectral information of a source emotional utterance to predict the framewise fundamental frequency and signal intensity under a target emotion. We also incorporate a latent gender representation to promote cross-speaker generalizability. Our neural network is trained to maximize the error log-likelihood under an assumed Laplacian distribution. We validate our model on the VESUS repository collected at Johns Hopkins University, which contains parallel emotional utterances from 10 actors across 5 emotional classes. The proposed algorithm outperforms three state-of-the-art baselines in terms of the mean absolute error and correlation between the predicted and target values. We evaluate the quality of our emotion manipulations via crowd-sourcing. Finally, we apply our emotion morphing model to utterances generated by Wavenet to demonstrate our unique ability to inject emotion into synthetic speech.

Index Terms: Emotion morphing, expressive speech synthesis, highway network, maximum likelihood estimation, wavenet

1. Introduction

Emotion is a hallmark attribute of human speech, with the ability to convey speaker intent [1], mood, and temperament [2, 3]. While humans are exceptionally good at encoding and decoding emotional states, the same cannot be said of automated platforms. For example, modern day speech synthesis can now mimic human intonation but, synthesizing emotions remain an open challenge. One reason for the limited progress is that emotions are inherently subjective and involve a complex set of signal characteristics. Another reason is the lack of sufficient training data to learn different emotional speech representations.

In this paper, we circumvent the data limitations by learning a multi-speaker model that transforms a neutral utterance to one of the three target emotions. We focus on modifying two prosodic features namely, pitch and signal energy [4]. Pitch, being a correlate of the fundamental frequency, controls the intonation. In general, pitch tends to rise for anger and happiness, and it tends to fall for sadness and fear. Energy, on the other hand is a correlate of the intensity and controls the fluctuations in loudness profile. Typically, the loudness is higher when speaker is excited and is lower when in sad emotional state. However, beyond these general trends, the actual relationship between pitch/energy and emotion is highly complex and is governed by both local and global speech properties. Our strategy is to learn a mapping function for these two prosody features from a neutral state to an emotional state by factoring in both the segmental and supra-segmental nature of speech.

Several previous works have explored the problem of emotion morphing. For example, the work of [5] explicitly models the fundamental frequency (F0) contour using a linear model, a Gaussian mixture model (GMM), and a classification-regression tree (CART). In contrast, the work of [6] develops an independent transformation model for pitch, duration and spectrum. A GMM model constrained by global variance [7] was introduced by [8]. This framework estimates the joint distribution of the source and target spectral and prosody features. Another strategy relies on dictionary learning and sparse recovery to estimate the emotional transfer function. For example, the work of [9] uses parallel exemplars aligned using dynamic time warping [10], a greedy optimization based sequence alignment procedure, to create a source and a target dictionary. Going one step further, the work of [9] estimates a sparse encoding using an active set Newton method based non-negative matrix factorization (NMF) [11]. The sparse encoding is used only to estimate the contextual spectrum envelope, whereas the fundamental frequency is directly copied from the corresponding frame of target dictionary. A more recent approach in emotion conversion is the application of bi-directional long-short term memory networks (Bi-LSTM) [12]. LSTMs are particularly suited for time series data, such as speech. Simultanous conversion of both spectral and prosody features is carried out in [13]. In this method, the F0 and energy contour are parameterized using 10 scales of continuous wavelet transform [14]. An approximate reconstruction of converted F0 and energy values synthesizes the final speech signal using the STRAIGHT [15] module.

Unlike prior work, our approach converts the prosodic features without any explicit parameterization. We rely on a highway network architecture which is faster to train than the Bi-LSTM and more robust on small datasets. Our highway network input consist of the smoothed spectrogram averaged within the standard Mel-frequency bands, along with the F0 values in a 360 ms context window, and a novel gender embedding. The highway network uses a likelihood based loss function to predict the framewise pitch and energy for the target emotion. We do not change the spectrum of the signal itself to maintain speaker identity. Our model is trained from scratch using the VESUS emotion dataset collected at Johns Hopkins [16]. We perform both objective and subjective evaluation to compare the results of our proposed model with three state-of-the-art baseline methods. Finally, we apply the emotion morphing model to synthetic utterances generated by Google Wavenet [17].

2. Highway Network for Emotion Warping

We use the STRAIGHT vocoder [15] to extract the F0 and energy contours. During training, we align the source and target emotional utterance using dynamic time warping [10]. This process allows us to learn a framewise transformation for pitch and energy values. However, it is not applied to any of the test utterances during conversion. We incorporate a novel latent representation for gender to improve the generalizability of our model across multiple speakers. Finally, we again use STRAIGHT to re-synthesize the modified utterances.

http://dx.doi.org/10.21437/Interspeech.2019-2512

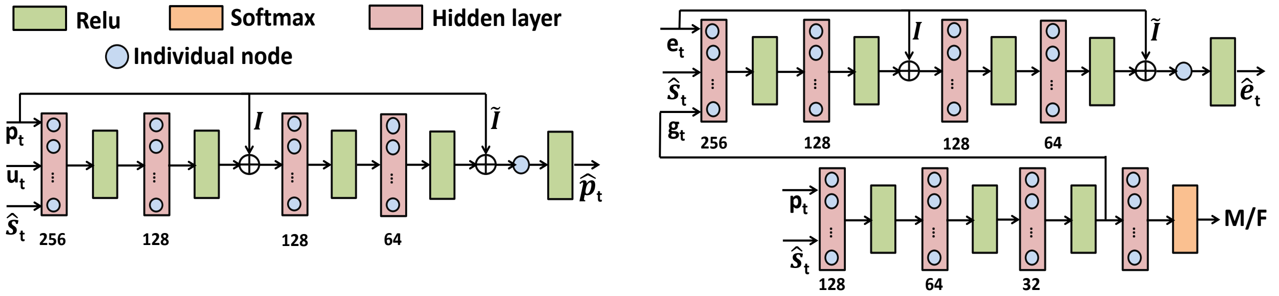

Figure 1: *(Left) shows the highway network architecture for pitch prediction and (right) shows the model used for prediction of energy. The gender embedding* $\mathbf{g_t}$ *is obtained from the smaller network trained for gender classification on the same dataset.*

2.1. Feature Extraction

As described above, our model predicts the framewise pitch and energy variations from source to target emotion. The features used for pitch prediction include a compressed form of smoothed spectral envelope, the utterance-level normalized fundamental frequency, and un-normalized pitch values with a context of 180 ms on both sides of the frame. We use only the voiced portion of each utterance to extract the pitch normalization parameters. The reason behind using a long contextual window for fundamental frequency is to account for both local and global properties. In other words, prosody is affected by both segmental (phonetic level) and supra-segmental (syllable or word level) characteristics of an utterance. A context of 360 ms ensures that the pitch information for mapping function is provided over on average two syllables. All features are extracted using a window of size 10 ms and a 10 ms stride.

To reduce the dimensionality of the input space, we compress the spectral envelope using the standard Mel frequency filterbanks. Namely, we first compute a 1,024 point FFT for each frame, resulting in a 513 dimensional magnitude spectrum $S \in \mathbb{R}^{513\times1}$ (frequency range 0 to π). We then use the normalized Mel filterbank matrix to obtain a 128-dimensional input representation. The filterbank matrix preserves the shape of the spectrum while accelerating the training of our deep highway network. Reducing the size to below 128 dimensions leads to noticeable loss in the shape of spectral envelope.

The utterance-normalized pitch (zero mean, unit variance), $\mathbf{u_t}$ allows us to capture the extreme values of target distribution. Conceptually, this feature acts as a flag forcing the neural network to sample from the tails of output distribution.

The computation of energy for each frame t is done by squaring and summing the short-time spectrum $S \in \mathbb{R}^{513\times1}$:

$$\mathbf{e_t} = \sqrt{\sum_{k=1}^{513} S_{k,t}^2} \quad t \in 1, 2, ...T \tag{1}$$

where T is the total number of frames extracted from an utterance. Similar to the pitch contour, a context of 360 ms is used for energy as well. The VESUS repository contains parallel emotional utterances across ten different speakers. We obtain a framewise correspondence between source and target prosody features using DTW for training the neural network.

2.2. Highway Network Architecture

We employ a highway neural network with one input layer, four hidden layers and one output layer along with multiple skip connections [18]. Fig. 1 shows the schematic diagram of the highway network architectures used for predicting pitch and energy. The input spectral features $\mathbf{\hat{s}_t}$ are normalized to mean zero and unit variance while the pitch contours $\mathbf{p_t}$ are fed in without any normalization. The output of highway network is given by:

$$\begin{aligned}\mathbf{\hat{p}_t} = \phi[W_{45} \times (\phi[W_{34} \times (\phi[W_{23} \\ \times \phi[W_{12} \times \phi[W_{01} \times \{\mathbf{\hat{s}_t}, \mathbf{u_t}, \mathbf{p_t}\} + b_1] \\ + b_2] \oplus \mathbf{Ip_t}) + b_3] + b_4] \oplus \mathbf{\tilde{I}p_t}) + b_5]\end{aligned} \tag{2}$$

The variables W_{ij} denote the weights going from layer i to layer j, and ϕ is the Relu non-linearity [19] applied at the output of each hidden node. The terms $\mathbf{Ip_t}$ and $\mathbf{\tilde{I}p_t}$ represent the skip connections to the output of the second and fourth hidden layer, respectively. While $\mathbf{I}$ is the identity matrix, $\mathbf{\tilde{I}}$ denotes just the three central rows of the identity matrix $\mathbf{I}$ which provides a short pitch context of 30 ms to the neural network before the final output. As designed, the highway network learns a perturbation on top of the input pitch values conditioned on the source spectrum and pitch contour. The skip connections add the correct bias, i.e, source pitch, back into the signal to better match the ground truth target. Closer to the input layer, the full 360 ms contextual pitch information is provided to extract important features from the contour, but as we go deeper, a shorter context proves to be sufficient. We use $\mathbf{\hat{p}_t}$ to denote the pitch predicted for the input source frame at time t. A log transformation of the pitch tends to collapse its dynamic range and makes the predictions to saturate at the mean of the training samples.

A similar architecture is used for the energy prediction. We replace the contextual pitch contour $\mathbf{p_t}$ by contextual energy contour $\mathbf{e_t}$. Unlike pitch, which inherently carries gender information, predicting energy requires an auxiliary gender input, as illustrated in Fig. 1. Here, we train a relatively shallow neural network having three hidden layers with the smoothed spectrum and pitch contour as input. The output of the final hidden layer is used as a latent embedding for the gender $\mathbf{g_t}$ to predict energy at time index t. Denoting the input by $\{\mathbf{\hat{s}_t}, \mathbf{e_t}\} \oplus \mathbf{g_t}$, the predicted energy is:

$$\begin{aligned}\mathbf{\hat{e}_t} = \phi[W_{45} \times (\phi[W_{34} \times (\phi[W_{23} \\ \times \phi[W_{12} \times \phi[W_{01} \times \{\mathbf{\hat{s}_t}, \mathbf{e_t}\} \oplus \mathbf{g_t} + b_1] \\ + b_2] \oplus \mathbf{Ie_t}) + b_3] + b_4] \oplus \mathbf{\tilde{I}e_t}) + b_5]\end{aligned} \tag{3}$$

During training we use a dropout [20] rate of 0.3 and batch normalization [21] after every hidden layer and before the skip connections with identity map are concatenated. These implementation details help us to improve the generalization capability of our highway network. We use the Adam optimizer [22] with a fixed learning rate of 0.01 and mini-batches of size 500.

2.3. Maximum Likelihood Objective

Since the dynamic range of pitch is very high, the standard l_2 loss is not appropriate because of its sensitivity towards penalizing extreme values in the difference. In contrast, mean absolute error (i.e., l_1 penalty) allows the highway network to evenly focus on the less extreme values of pitch (such as around 200 Hz which occur more frequently in the data). We train the highway networks by maximizing the likelihood of the error for each training sample in a mini-batch [23]. In particular, we assume that the error function defined by $\mathscr{E}_n = y_n - \hat{y}_n$, where y_n is the true value and $\hat{y}_n$ is the model estimate, is drawn from a Laplacian distribution with mean 0 and variance b:

$$\mathscr{E}_n \sim \frac{1}{2b} \exp\left\{-\frac{\|y_n - \hat{y}_n\|_1}{b}\right\} \tag{4}$$

The parameters of highway network, denoted by θ get updated via standard backpropagation algorithm. From here, the variance of the error distribution b is updated after every epoch of the highway network update in a maximum likelihood framework similar to the expectation maximization (EM) algorithm:

$$\hat{b} = \frac{1}{N} \sum_{n=1}^{N} \|y_n - \hat{y}_n\|_1 \tag{5}$$

The algorithm for training the model parameters and estimating the Laplacian variance alternates between the following steps:

- Update θ to minimize $\sum_{n=1}^{N} \|y_n - \hat{y}_n\|_1$ while b fixed.
- Update b using Eq. (5) while θ is fixed.

At a high level, our maximum likelihood strategy acts as a learning rate scheduler by re-scaling the step size by variance in each epoch. In practice, this approach improves the correlation observed between the ground truth and predicted pitch/energy beyond the standard minimum absolute error objective.

2.4. Reconstruction

In the reconstruction stage, the predicted pitch and energy values over the input frames are smoothed using a mean filter to ensure the continuity in pitch and energy contour. While the pitch is directly used for synthesis, the energy values are implicitly used by re-scaling the spectrum using the equation:

$$\hat{S}_t = S_t \times \frac{\hat{\mathbf{e}}_\mathbf{t}}{\mathbf{e}_\mathbf{t}} \quad for \;\; t = 1, 2, ...T \tag{6}$$

Here, $\mathbf{e}_\mathbf{t}$ is the original energy value of frame t while $\hat{\mathbf{e}}_\mathbf{t}$ is the predicted energy value. The aperiodicity component of the STRAIGHT vocoder is copied directly from the source speech.

3. Experiments and Results

We carry out both the quantitative and qualitative evaluations to compare our performance with the current state-of-the-art techniques for emotion and prosody conversion in speech.

3.1. Dataset and Experimental Setup

Our training and evaluation relies on the VESUS emotional dataset collected at Johns Hopkins University [16]. VESUS contains a set of parallel emotional utterances spoken by a mix of amateur and professional actors. The original database has 2500 utterances for each of the five emotional classes: happiness, anger, sadness, fear and neutral. The repository also contains an emotion perception rating for each utterance provided by ten Amazon Mechanical Turk (AMT) raters.

For the proposed model, we use only those utterances from VESUS repository which are agreed upon by more than 50% of the AMT raters. We also omit the fear category from our experiments because of its high confusion with sad and neutral emotions. The total numbers in our experiment are:

- For **Neutral to Angry**: 1534 utterances for training, 72 for validation and, 61 for testing.
- For **Neutral to Happy**: 790 utterances for training, 43 for validation and, 43 for testing.
- For **Neutral to Sad**: 1449 for training, 75 for validation and, 63 for testing.

Objective evaluation includes the mean absolute error and the Pearsons correlation coefficient measure between the predicted value of pitch and energy and their ground truth counterparts. For subjective evaluation, we ask raters on AMT to classify each of the converted test samples for perceived emotion. Our designed survey asks AMT workers to listen to two speech files. One of them is the baseline neutral speech and the other one is the speech converted into some target emotion. The order of neutral and emotional speech is randomized to weed out any non-diligent raters or bots. After they finish listening, we ask them to classify the emotion in both audio files. We find this type of bias correction using source (neutral) speech to be important because emotion perception is highly dependent on the knowledge about speaker articulation and speaking style.

3.2. Baseline methods

We compare our proposed model with three state-of-the-art baseline methods. The first baseline fits a Gaussian mixture model (GMM) [8] to the joint distribution of the source and target STRAIGHT cepstral features and fundamental frequency. We further incorporate the Global variance constraint proposed by [7] to improve the GMM based conversion model.

The second baseline uses the sparse Non-Negative Matrix Factorization (NMF) method developed in [9]. Here, two parallel dictionaries of STRAIGHT spectrum are constructed from the training dataset. An active Newton set based NMF estimates the sparse coding of input spectral features over the source dictionary. This encoding is then used to construct the converted spectrum and fundamental frequency from the target dictionary.

The third baseline is the Bi-LSTM model [13] which is pre trained for voice conversion using the CMU-ARCTIC corpus [24] and then fine-tuned for emotion conversion on the VESUS database. This method simultaneously converts both spectral and prosodic (pitch, energy) features. The prosodic features are parameterized by continuous wavelet transform [14] coefficients. The intention behind such parameterization is to consider both short-term and long-term pitch and energy trajectories by using multiple scales for the wavelet transform.

3.3. Experimental Results

Table 1 reports the quantitative performance of all four methods. Like the baseline algorithms, we train separate models for each target emotion category. Note that the proposed model outperforms all the baselines by a significant margin. The global variance based GMM model is the second best algorithm for emotion conversion on the VESUS dataset. The high performance of GMM compared to the Bi-LSTM can be attributed to its simplicity, which makes it less prone to overfitting, and the large number of speakers in the VESUS repository. Our results also suggest that the procedure used in [13] of fine-tuning an Bi-LSTM model can not achieve the good performance for emotion

Table 1: *MAE and Pearson's Correlation measures for pitch and energy across target emotions using universal model.*

Alg.	MAE($\hat{\mathrm{p}}_\mathrm{t}$)	Cor($\hat{\mathrm{p}}_\mathrm{t}$)	MAE($\hat{\mathrm{e}}_\mathrm{t}$)	Cor($\hat{\mathrm{e}}_\mathrm{t}$)
		Neutral-to-Angry		
GMM	44.3	0.54	4.24	0.57
NMF	94.2	0.22	4.2	0.22
Bi-LSTM	57.4	0.34	5.77	0.56
Proposed	**39.6**	**0.64**	**1.9**	**0.6**
		Neutral-to-Sad		
GMM	29.1	0.8	5.87	0.53
NMF	65.3	0.4	7.9	0.32
Bi-LSTM	29.6	0.78	5.23	0.5
Proposed	**22.2**	**0.83**	**3.4**	**0.67**
		Neutral-to-Happy		
GMM	53.8	0.51	4.24	0.53
NMF	106.7	0.25	6.5	0.23
Bi-LSTM	67.6	0.48	4.8	0.52
Proposed	**49.8**	**0.54**	**2.5**	**0.68**

conversion. Further, the assumption that local optima for emotion conversion should be close to the voice conversion solution on the error surface may not necessarily be true. NMF based sparse recovery and reconstruction performs the worst among all four models. This result is expected because there is no explicit constraint on the estimation of sparse coding. Specifically, there are multiple acoustic units that have very similar spectral envelopes and hence the algorithm also does not guarantee a smooth transition going from one frame to another.

In contrast to the baselines, our proposed pitch and energy prediction model is more robust because it focuses on learning a single, highly relevant transformation, rather than attempting to modify the entire spectrum. In addition, the highway network architecture allows us to learn a perturbation model that translates easily across multiple speakers. From a technical standpoint, it also facilitates for a smooth flow of gradients during the backpropagation [25]. Further, the EM type update of variance and weights of highway network in each iteration has a scaling effect on the mini-batch loss. This indirectly adjusts the learning rate during training, thereby helping the network converge to a better local optima than the Bi-LSTM model.

We evaluate the subjective quality of our emotion conversion using AMT. Empirically, we found the reconstructed speech from the GMM and NMF models to be highly distorted and unintelligible. Therefore, we only obtain crowd-sourced ratings for our highway network and the Bi-LSTM model. We crowd-source the same utterances spoken by same speakers for the highway network and Bi-LSTM model to get a uniform comparison between the two. Fig. 2 (top) shows the emotion classification accuracy on the testing utterances. Compared to the baseline model, our proposed model has higher classification accuracy across all three emotions. Further, the classification for neutral-to-sad is best followed by neutral-to-angry and then neutral-to-happy. This result is in line with the objective measures for pitch prediction (see Table 1). The Bi-LSTM model performs poorly because it fails to capture important prosody variations that contribute to emotion perception.

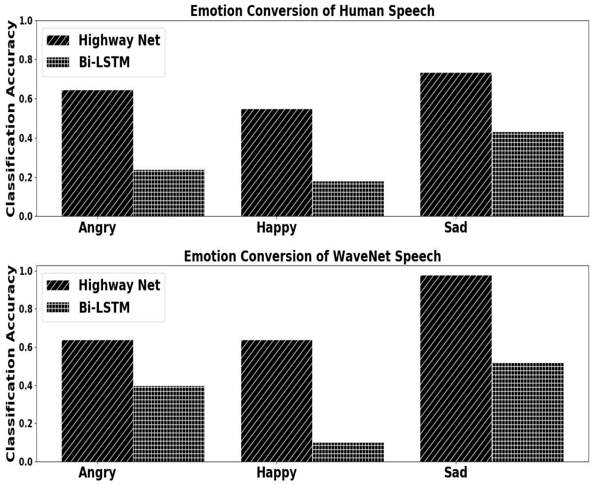

Figure 2: *Emotion Classification accuracy for human (top) and Wavenet's speech (bottom) obtained via crowd-sourcing.*

The final experiment examines our ability to inject emotional cues into synthetic speech generated by Google Wavenet [17]. We use the text-to-speech API provided by Google to generate same utterances as spoken in the VESUS repository. The utterances are generated for a female American English speaker. For this experiment, we fine tune the highway network by picking a speaker from the VESUS dataset who has the most expressive emotional utterances. We use 220/120/220 samples for fine tuning the neutral to angry/happy/sad model, respectively. The fine tuning procedure runs for only 50 epochs starting from the mixed speaker model weights. The crowd sourcing setup is unchanged from the previous case. Fig. 2 (bottom) shows the emotion classification result on speech generated by Wavenet. We see that speech modified by the highway network is clearly perceived as emotional, in contrast to the Bi-LSTM. This first-of-its-kind demonstration shows that our model is highly adaptable to new (and even synthetic) speakers with minimal training data for fine tuning. In contrast, the Bi-LSTM does worse for the same setup due to its complex architecture and attempt to modify the entire spectral range.

In summary, our quantitative and qualitative results together show the markedly improved performance for our proposed model over three competing baselines. Our results also suggest that modifying just the pitch and energy contours is sufficient for emotion conversion. Finally, our experiment on Wavenet demonstrates that we can infuse emotions into a synthetic speech by fine tuning our cross-speaker model.

4. Conclusions

We have demonstrated the first multi-speaker emotion conversion model based on modifying pitch and energy. Our novel highway network based prosody prediction model has the lowest mean absolute error and highest correlation with the ground truth values when trained and tested on the VESUS emotional dataset. We trained our highway network in an alternating fashion by maximizing the error likelihood. A Laplacian assumption on the residual distribution in each mini-batch was made and was motivated by the data itself. Our algorithm outperformed the state-of-the-art methods for emotion conversion on subjective listening tasks by significant margins thereby proving the effectiveness of our procedure. Finally, we showed that our model is capable of injecting emotion into vocoder output which has not been done before in the literature.

5. References

[1] T. Johnstone and K. Scherer, "Vocal communication of emotion," *Handbook of Emotions,*, 01 2000.

[2] D. Schacter, D. T. Gilbert, and D. M. Wegner, *Psychology (2nd Edition)*. New York: Worth, 2011.

[3] J. A. Russell, J.-A. Bachorowski, and J.-M. Fernandez-Dols, "Facial and vocal expressions of emotion," *Annual Review of Psychology*, vol. 54, pp. 329–349, 11 2003.

[4] R. W. Frick, "Communicating emotion. the role of prosodic features," *Psychological Bulletin*, vol. 97, pp. 412–429, 05 1985.

[5] Y. Kang, J. Tao, and B. Xu, "Applying pitch target model to convert f0 contour for expressive mandarin speech synthesis," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, vol. 1, 06 2006, pp. I – I.

[6] Z. Inanoglu and S. Young, "A system for transforming the emotion in speech: Combining data-driven conversion techniques for prosody and voice quality," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 1, 01 2007, pp. 490–493.

[7] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2222–2235, Nov 2007.

[8] R. Aihara, R. Takashima, T. Takiguchi, and Y. Ariki, "Gmm-based emotional voice conversion using spectrum and prosody features," *American Journal of Signal Processing*, vol. 2, pp. 134–138, 12 2012.

[9] R. Aihara, R. Ueda, T. Takiguchi, and Y. Ariki, "Exemplar-based emotional voice conversion using non-negative matrix factorization," in *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2014 Asia-Pacific*, Dec 2014, pp. 1–7.

[10] *Dynamic Time Warping (DTW)*. Dordrecht: Springer Netherlands, 2008, pp. 570–570.

[11] T. Virtanen, B. Raj, J. Gemmeke, and H. Van hamme, "Active-set newton algorithm for non-negative sparse coding of audio," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 05 2014, pp. 3092–3096.

[12] M. Schuster and K. Paliwal, "Bidirectional recurrent neural networks," *Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, nov 1997.

[13] H. Ming, D.-Y. Huang, L. Xie, J. Wu, M. Dong, and H. Li, "Deep bidirectional lstm modeling of timbre and prosody for emotional voice conversion," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 09 2016, pp. 2453–2457.

[14] M. Sam Ribeiro and R. A. J. Clark, "A multi-level representation of f0 using the continuous wavelet transform and the discrete cosine transform," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 04 2015, pp. 4909–4913.

[15] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveign, "Restructuring speech representations using a pitch-adaptive timefrequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, pp. 187–207, 04 1999.

[16] J. Sager, R. Shankar, J. Reinhold, and A. Venkatarman, "Vesus: A crowd-annotated database to study emotion production and perception in spoken english," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2019.

[17] 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," *CoRR*, vol. abs/1609.03499, 2016. [Online]. Available: http://arxiv.org/abs/1609.03499

[18] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Highway networks," *CoRR*, vol. abs/1505.00387, 2015. [Online]. Available: http://arxiv.org/abs/1505.00387

[19] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ser. ICML'10. USA: Omnipress, 2010, pp. 807–814.

[20] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, 2014. [Online]. Available: http://jmlr.org/papers/v15/srivastava14a.html

[21] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *CoRR*, vol. abs/1502.03167, 2015. [Online]. Available: http://arxiv.org/abs/1502.03167

[22] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2015.

[23] Y. Wang, J. Du, L.-R. Dai, and C.-H. Lee, "A maximum likelihood approach to deep neural network based nonlinear spectral mapping for single-channel speech separation," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 08 2017, pp. 1178–1182.

[24] J. Kominek and A. W Black, "The cmu arctic speech databases," *SSW5-2004*, 01 2004.

[25] S. S. Du, J. D. Lee, H. Li, L. Wang, and X. Zhai, "Gradient descent finds global minima of deep neural networks," *CoRR*, vol. abs/1811.03804, 2018.