Platform Agnostic Emotional Speech Synthesis With Mel to Mel Translation

Laxmaan Balaji
Dept. of Computer Science and Engineering
Pennsylvania State University
laxmaanb@psu.edu

Mike Burnham
Dept. of Political Science
Pennsylvania State University
mlb6496@psu.edu

***Abstract*—**
Embedding emotion into speech synthesis is a challenging problem with broad applications from enhancing human-computer interactions to sound editing. In recent years speech synthesis has made significant strides by leveraging mel spectrograms for generating raw audio. Currently, the primary methods of synthesizing emotional speech involve embedding emotion during the spectrogram generation process or altering audio after it has already been generated. Here we test an approach of synthesizing emotion via picture translation on mel spectrograms. To do so, we use the conditional adversarial network pix2pix to transform emotionally neutral spectrograms to a spectrogram correlating with a specific emotion, and then generate audio from the spectrogram with Goggle's wavenet. This approach has the benefit of being platform agnostic in that it can be applied to a variety of speech generating platforms. We conclude that while platform compatibility may be a challenge for this approach, it shows significant promise.

***Index Terms*—Speech synthesis, emotions, affective computing, image translation, text to speech**

I. Introduction

Emotional speech synthesis aims to generate natural sounding speech embedded with specific emotions. Creating emotional speech has broad applications. Current widely used speech based human-computer interfaces such as Apple's Siri or Goggle's assistant, for example, speak in largely neutral tones. Adding appropriate emotional depth to their language could enhance user experience. Likewise, the ability to edit the emotional content of speech could have wide applicability to those editing sound for movies, music, or radio. However, this remains a difficult task because emotions alter every aspect of speech such as volume, tempo, and pitch. Additionally, while there are multiple points in the speech synthesis pipeline at which audio can be embedded, there are drawbacks to each in terms of audio quality, computational resources, and flexibility.

In this paper we build on on a method previously proposed by Choi et al. by constructing and testing an emotional speech synthesis system that works by converting neutral mel spectrograms to emotional mel spectrograms [1]. This approach inserts itself into the middle of the typical speech synthesis pipeline. It has the benefit of working for a variety of tasks and is platform agnostic. Our iteration of this process relies on the image translation executed by the pix2pix generative adversarial network [2]. We assess the feasibility of this methodology in terms of its ability to synthesize emotions that are identifiable, natural sounding, and to do so in a computationally efficient manner. We show it is promising in all three of these areas. The drawback, however, is that it does not yet naturally integrate with end-to-end speech synthesis pipelines and thus can present compatibility challenges.

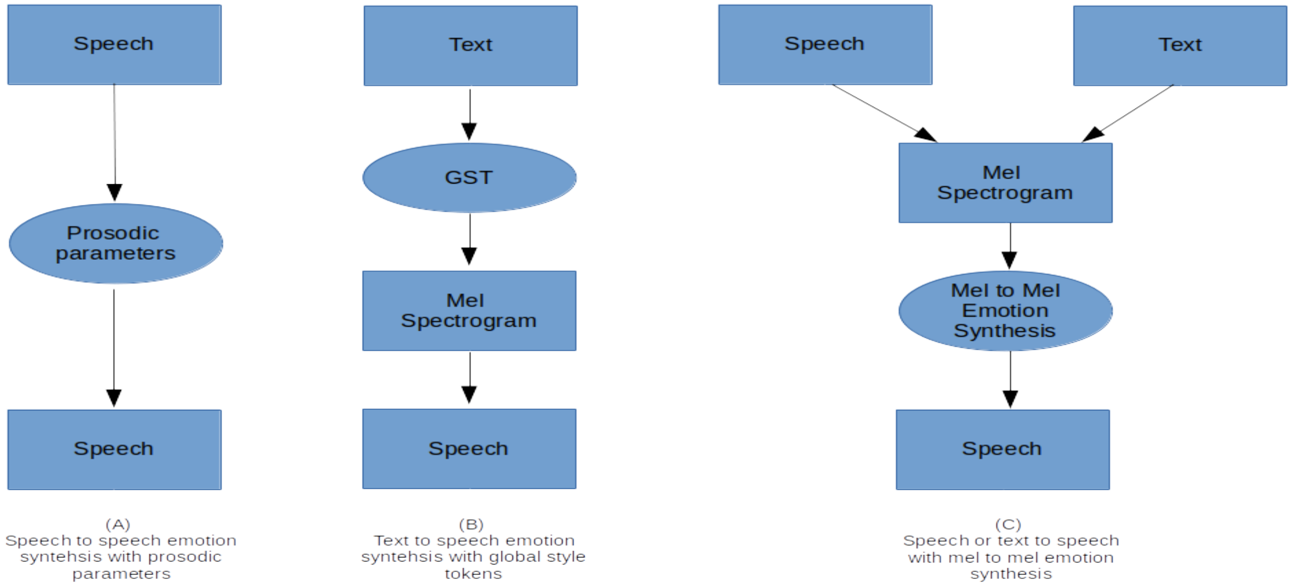
The rest of the paper is organized as follows. Section 2 will briefly summarize the current state of emotional speech synthesis including the motivations for this work and the current state-of-the-art techniques. Section 3 outlines our data sources and technical approach. Section 4 presents the results of our experiments. Finally in section 5 we discuss implications and future research.

II. Related Work

A. Emotion Synthesis

Speech synthesis takes two forms. Text-to-speech programs are what generally first come to mind in the field of speech synthesis and are the more well researched form. They accept text as user input and synthesize a variety of natural sounding speech typically for use in human-computer interactions. The current state-of-the-art in this field is Google's Tacotron 2 and Wavenet [3], [4]. The second form is speech-to-speech synthesis, which translates previously generated speech to another voice or style. This is often used for real time translation purposes or simulating another person's voice such as with the MelGAN-VC program [5].

In recent years, both types of speech synthesis made strides by leveraging mel spectrograms – a visual representation of sound that is scaled for the human perception of hearing. Spectrograms provide an ideal representation of audio for the purpose of speech synthesis. They are more efficient in terms of space than audio, can be visualized as an image or decomposed to a numeric array, and can generate audio via a direct mathematical conversion or neural network. They are also well suited to other non-synthesis related tasks such as classification, and can be easily manipulated. For these reasons, in addition to the fact that both text-to-speech and speech-to-speech synthesis utilize them, we argue that the mel spectrogram is an ideal point at which to interject emotion. Mel spectrogram based emotion synthesis includes all of the aforementioned benefits and can be integrated into a wide variety of pipelines.



(A)
Speech to speech emotion syntehsis with prosodic parameters

(B)
Text to speech emotion syntehsis with global style tokens

(C)
Speech or text to speech with mel to mel emotion synthesis

Fig. 1. Emotional speech synthesis current models (A and B) vs proposed model (C)

Current methods of synthesizing emotional speech primarily focus on either the beginning or the end of the synthesis pipeline rather than on the mel spectrogram. Early techniques used global prosodic parameters, which are considered to be "universal or near universal cues for emotion" [6]. These parameters are voice attributes such as range, tempo, or loudness that are correlated with specific emotions. While these parameters do elucidate much about the nature of emotion in speech, it is difficult to achieve natural sounding speech through post-hoc manipulation of them. More recent attempts to synthesize emotion in speech comes by defining the emotions to be generated a-priori. One such example is Google's global style tokens which is a bank of embeddings trained and integrated into tacotron that can generate speech in a number of different styles [7]. Um et al. later iterated on this approach by tuning style tokens to specific emotions [8]. While this approach has generated some of the most authentically sounding results, it requires deep integration into an existing platform, vast resources, and does not work for speech-to-speech synthesis.

Our proposed method capitalizes the the advantages of the mel spectrogram and uses mel to mel translation in order to synthesize emotion. Figure 1 illustrates a basic overview of current emotional synthesis pipelines as well as the proposed pipeline. Introducing emotion at the mel spectrogram stage is both a benefit in that is platform agnostic, and a challenge in that it lacks of integration with other systems and potentially suffers from compatibility challenges. Because the approach necessarily relies on other systems for generating audio it is worth giving a brief overview of current speech synthesis.

B. Text to Mel with Tacotron

Tacotron [3] is an end-to-end generative model for text-to-speech-synthesis. It is based on a seq2seq architecture incorporating the attention mechanism as shown in Figure 2. The Tacotron model can be trained completely from scratch with <text, audio> pairs as input, and outputs the raw spectrogram. Tacotron 2 [9] also uses a seq2seq model in which a sequence of letters is mapped to a sequence of features encoding the audio. These features are represented by an 80-dimensional audio spectrogram. They help to capture word pronunciation, along with specific details like speed and volume. A model similar to that of a WaveNet is used to convert these features into a waveform. Some of the limitations of Tacotron 2 is that the model is unable to pronounce some complicated words, sometimes generating random noises in an attempt to do so. The model also cannot be controlled to generate speech with emotion.

C. Mel to Speech with WaveNet

WaveNet [4] is an example of parametric speech synthesis model. This means that all the information required for generation is stored as model parameters and that the speech's characteristics are determined through model inputs. However, this kind of model often produces speech that sounds unnatural. The WaveNet model was one of the first to overcome this limitation. WaveNet is a Convolutional Neural Network, as seen in 3 which has different dilation factors for each of its convolutional layers, allowing the receptive field of each layer to increase exponentially with its depth. It takes as input waveforms which represent raw audio recording from speaking humans during training. The network is sampled after training, and at each step the network computes a probability distribution. From this, a value is fed back to the input and the next step is predicted. The WaveNet model was tested by humans using Mean Opinion Scores (MOS). It achieved

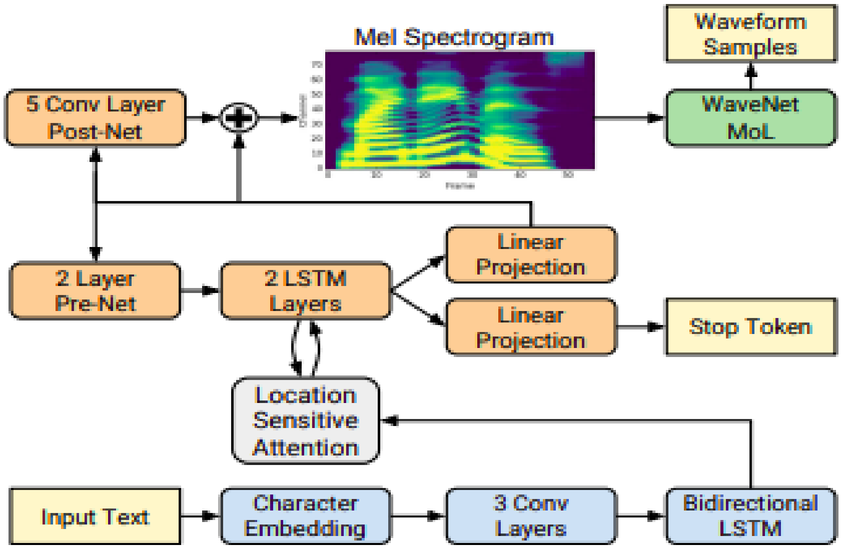


Fig. 2. Tacotron 2 System Architecture

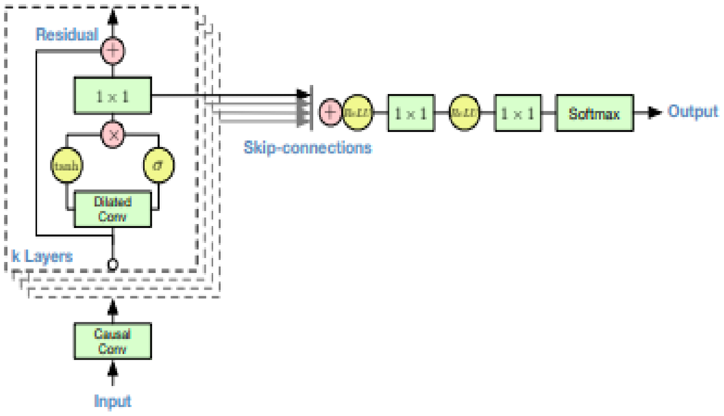


Fig. 3. WaveNet Architecture

naturalness above 4.0 (on a 5 scale), making it really close to actual human speech.

D. Mel to Speech with WaveGlow

WaveGlow [10] shows that auto-regressive methods are not necessary for speech synthesis. It is a flow-based generative network, as shown in 4 that generates audio by sampling from a zero mean spherical Gaussian distribution. These samples are then transformed into the desired distribution by passing them through a sequence of layers. The distribution of audio samples is modelled conditioned on a mel-spectrogram. The main idea of WaveGlow was to incorporate the benefits of both the WaveNet [4] and the Glow [11] model, thus constraining the network to be invertible. This provides quick and efficient speech synthesis, of a very high-quality. One of the main advantages of WaveGlow is that it was implemented using one network and one cost function which makes the model stable and simple to train.

III. Approach

A. Proposed Process

For our approach to emotion synthesis via mel to mel conversion, we use established image translation techniques on paired mel spectrograms. Image translation refers to a process by which an algorithm receives an image as input, and outputs some variation of that image. The output can be a slight variation on the source material such as a change in color scheme or a higher resolution version of the image, or it can be a radical transformation such as creating photo-realistic images from sketches or converting a photo to a painting in the style of a specific artist. In our case, we input a spectrogram of

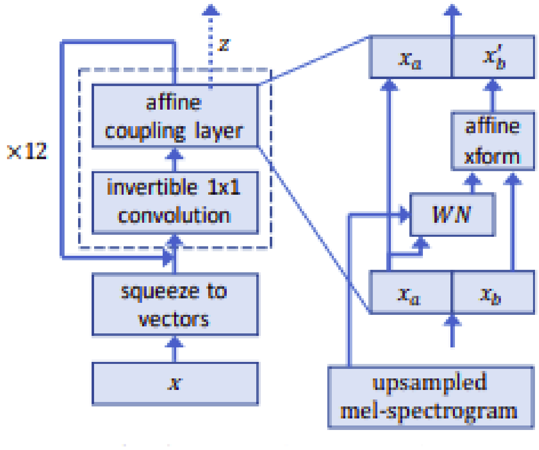


Fig. 4. WaveGlow Architecture

a phrase spoken in a neutral tone and generated spectrograms of a specific emotion.

To accomplished this, we relied on the pix2pix conditional adversarial network [2]. This technique was introduced by Isola et al. In 2018 and is an ideal platform for our task for a couple of reasons. First, pix2pix has shown to be particularly adapt at identifying and translating the edges of a photograph. This is especially important for mel spectrograms. Frequencies within the spectrogram are represented by waves visualized in the spectrogram and blurring the edges between waves can lead to a dramatic loss in quality. Second, pix2pix has already been successfully applied to mel spectrograms for other purposes. A variations of this framework called pix2pixHD [12] has been used by Sheng and Pavlovskiy [13] to successfully increase the resolution and thus audio quality of mel spectrograms, and by Michelsanti and Tan to reduce background noise in spectrograms [14].

Our proposed training process is shown in figure 5. We use speech to speech synthesis in our experiment because training the model requires spectrograms that contain the same words and vary only in emotional content. It is impossible to label the same text to two different emotions. Thus, mel to mel translation must first be trained on speech based spectrograms and can then be generalized to text based spectrograms given enough training data. Due to the architecture of pix2pix a new model is trained for each emotion.

B. Data and Pre-processing

To train our model we used the Toronto Emotional Speech Set (TESS) [15]. Due to the relatively small size of currently available English language emotional speech data sets, we had to narrow the scope of our model. While we argue that our results show a widely generalizable model is possible, currently available data does not support such functionality. Accordingly, the TESS data set was chosen because of its limited scope as well as its high quality audio and relatively high number of samples. The data consists of two female speakers that each provide 200 samples for seven emotions as well as neutral emotions. Spoken phrases are all some variation of “Say the word X” where X is one of 200 different words. The TESS data set also provided the highest quality audio samples with the clearest frequencies and the most consistent length, a significant factor in generating good image translation results.

Preprocessing of the data and mel spectrogram generation was done with python’s LibROSA library [16]. The initial step is to decompose the audio via the Fourier transform. This breaks the sample into discrete frequencies which can then be converted to the mel scale. At this point the audio is an array of numbers that can be manipulated, transformed back in to audio, or saved as either a spectrogram or numeric array. Eighty mels were chosen as those were the number of mels necessary for Wavenet compatibility. Because the image translation algorithm requires pictures of consistent size we first standardize the length of each audio sample. Each sample was fit to an 80x128 array that represented the loudest window of the sample. For the majority of samples this window represented the entire audio clip minus any leading or trailing white noise. For those that exceeded this length the audio was truncated at the end. For samples that were shorter than this window, silence was added to the end of the sample. To further augment the data set and improve the sample size for training, spectrograms were inverted and rotated. This allowed us to effectively double the size of our training set.

C. Image Generation

The source images in all runs are the *neutral* mel spectrograms and the target images in each run are the mel spectrograms of a particular emotion. As in [2], we optimize the objective given by:

$$G* = \arg\min_G \max_D \mathcal{L}_{cGAN}(G,D) + \lambda\mathcal{L}_{L1}(G)$$

where $\mathcal{L}_{cGAN}(G,D)$ is the conditional GAN objective and is given by:

$$\mathcal{L}_{cGAN}(G,D) = \mathbb{E}_{x,y}[\log D(x,y)]{+}\mathbb{E}_{x,z}[\log(1{-}D(x,G(x,z))]$$

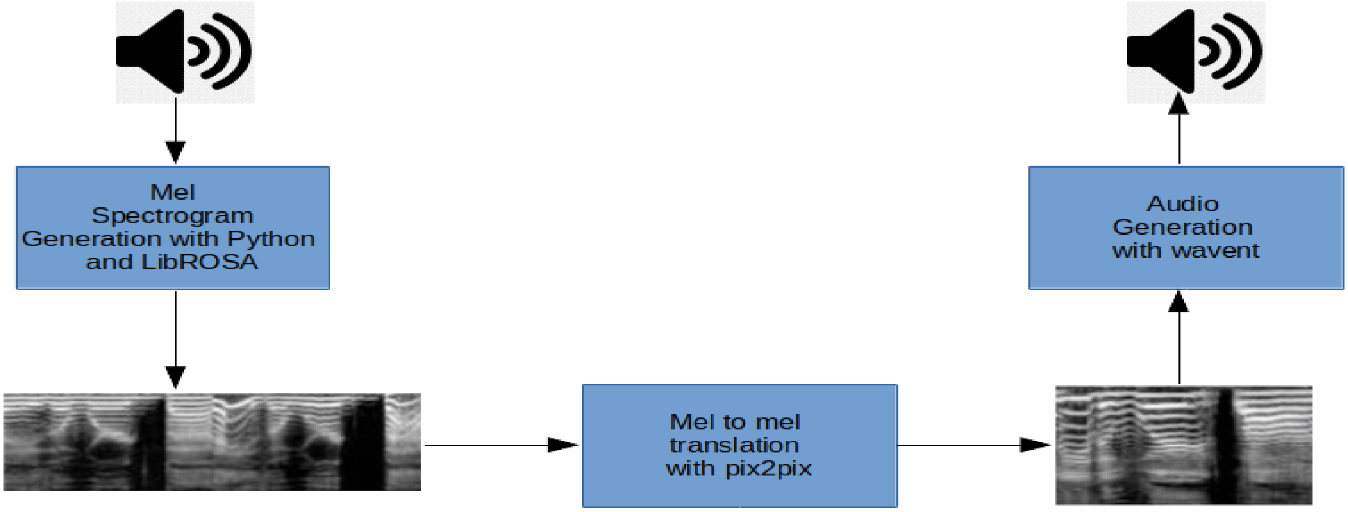

Mel Spectrogram Generation with Python and LibROSA
Mel to mel translation with pix2pix
Audio Generation with wavent

Fig. 5. proposed emotional speech synthesis model training process

where $\mathcal{L}_{L1}(G)$ is the L1 loss which minimizes image blurring and is given by:

$$\mathcal{L}_{L1}(G) = \mathbb{E}_{x,y,z}[||y - G(x,z)||_1]$$

To evaluate results, image translation does not leverage typical loss metrics as its primary indicator of success. Rather, human opinion is the primary metric by which results are judged. We used a similar approach in evaluating mel to mel conversion. Generated spectrograms were saved at periodic intervals and assessed for how identifiable frequencies were in the image, as well the depth of silences. High quality spectrogrms show frequencies as discrete waves in the image and silence as deep blacks. In addition, generated samples were decoded and transformed in to audio with Librosa as an intermittent test of sound quality. We found the best results were achieved by training for around 600 epochs with a batch size of one. At this point the model began to converge and intermittent testing produced audio quality comparable to the true spectrograms.

A rudimentary metric (inferior to human opinion) that we used in conjunction with human testing is the Peak Signal-to-Noise Ratio (PSNR) given by :

$$PSNR = 20\log_{10}\frac{f_{\max}}{\sqrt{MSE}}$$

where $f_{\max}$ corresponds to maximum pixel intensity in the source image and MSE is the *pixel-wise* mean squared error between the source and target images. We use this as one dimension in our analysis of the images reconstructed per emotion.

To train the model we used the parameters shown in table I These parameters were chosen based on a combination of what produced the best results while operating within the constraints imposed by wavenet. Currently wavenet's architecture can only produce speech from spectrograms with certain parameters. Given the resources to reconfigure and retrain wavenet, further optimization may be possible.

Hyparameter	value
pre-process	crop
crop size	80
netG	resnet$_6blocks$
Lambda$_l$1	100

TABLE I
PIX2PIX IMAGE TRANSLATION PARAMETERS

Hyparameter	value
wav max value	32768.0
sampling rate	22050
filter length	1024
hop length	256
window length	1024
mel channels	80
min mel frequency	0.0
max mel frequency	11025

TABLE II
WAVEGLOW/WAVENET AUDIO PARAMETERS

D. Audio Generation

In order to generate audio, we used NVIDIA's implementation of WaveGlow in Pytorch [17] and the Wavenet implementation in [18] with the hyperparameters as listed in Table II. The output images from the Mel Spectrogram Synthesis stage are PNG files with values between $[0, 255]$. In order to convert back to the mel scale, the values are linearly interpolated to the range $[-12, -1]$ corresponding to the log scale values. The resultant array is fed as the input to WaveGlow/Wavenet.

We found that Waveglow [17] was faster than Wavenet [18] by a factor of 10 and the final generated audio came from Waveglow.

Emotion	MSE	PSNR
Anger	100.32	28.12
Disgust	100.07	28.14
Happy	**97.80**	**28.23**

TABLE III
MEAN PSNR AND MSE FOR THE SYNTHESIZED MELS BY EMOTION

Emotion, sample #	Mean	Std. Dev
Anger 1	3.75	0.60
Anger 2	3.81	0.72
Disgust 1	3.36	1.19
Disgust 2	3.18	0.94
Happy 1	**4.09**	0.67
Happy 2	3.55	0.78

TABLE IV
MEAN SCORE PER SAMPLE

IV. RESULTS

A. Spectrogram Generation

Examples of generated spectrograms as training progresses is shown in figure 6. As can be seen, spectrograms generated early in training show blurred frequencies and artifacts in the images. At around epoch 200 it becomes difficult to notice differences with the naked eye, although some differences were still noticeable in audio testing. At around 600 epochs, audio differences between generated and target spectrograms were near indistinguishable. Figure 7 shows examples of generated spectrograms with their neutral and emotional counterparts.

The results for mean PSNR and MSE for all images in the test set across 3 emotions are shown in Table III. We see that PSNR is highest for the *Happy* emotion and also the MSE is lowest for the same emotion. The metrics then say that generated samples for *Disgust* are second best, followed by samples for *Anger*.

B. Speech Generation

As the quality of generated samples of audio is contingent on both the input mel spectrogram and the implementation of the vocoder, we have obtained separate scores for the synthesized audio. As in Choi et. al. [1], we obtain the Mean Opinion Scores (MOS) from a sample of 14 responses, we recorded MOS for samples from the 3 synthesized emotions. A total of 6 samples were presented, with two happy, two disgusted, and two angry samples. Each sample was presented with its corresponding synthesized neutral audio. All words spoken in the samples were unique. The ratings went from *terrible* with a score of 1 to *excellent* with a score of 5.

Table IV shows the mean and standard deviation of scores for each sample. We see that sample *Happy-1* received the highest scores. Table V consolidates the results of Table IV into the 3 emotions. We see that *Happy* is scored highest, followed by *Anger* and then *Disgust*.

While the PSNR and MOS scores agree for *Happy*, we see that although *Disgust* has higher PSNR, *Anger* has a higher MOS. Since PSNR is only a guiding metric and is not always in line with human opinion, this behavior is expected.

Emotion #	Mean	Std. Dev
Anger	3.78	0.66
Disgust	3.27	1.05
Happy	**3.81**	0.77

TABLE V
MEAN OPINION SCORE (MOS) BY EMOTION

V. DISCUSSION

We consider these results encouraging. By way of comparison, Professionally recorded natural speech achieves a mean opinion score of 4.58 [9]. Google's own in-house implementation of Wavenet with no emotional content achieves a score of 4.53 [9]. While our results are noticeably lower, the vast majority of loss in audio quality is due to a lack of integration with Wavenet rather than loss in emotion synthesis. Achieving intelligible emotional speech on such a limited data set without organic Wavenet implementation provides significant evidence that this approach can be fruitful.

Analysis of the generated samples yields a few insights:

- The vocoders are conditioned on a sampling rate and audio parameters that are highly coupled with the Text-To-Mel system (Tacotron2 in our case). This leads to an incompatibility with datasets that are not sampled at that specific rate necessitating the need for retraining on a large audio corpus at the same sampling rate as the datasets. Coupled with the relative paucity of large emotional audio corpora, the process of end-to-end conditioning is prohibitive in training time.
- A key flaw that the evaluators for the audio samples identified for the *disgust* samples was that the audio was abruptly cut off. This is an inherent flaw of Pix2Pix and other image based GANs as the input and output sizes are fixed. Mel spectrograms are fixed along the mel axis with a constant number of mels but they can extend along the time axis. Emotions such as disgust involve speech that is more "dragged out" across timeframes. The rate of speech is also a factor in emotion expression. Any emotional mapping that incorporates an elongation along the time axis is poorly handled.
- Despite the above shortcomings, this work shows that it is possible to effectively map emotions to mel spectrograms and condition a GAN to transform audio from one emotion's domain to another. The approaches detailed in section VI-A elaborate on how to scale and refine the procedure.

VI. FUTURE WORK

A. Future Work

This section details on methods to scale the approach outlined in this process to more general use cases of arbitrary and longer audio samples.

- Generalizability of this method requires a larger data set. Currently there is a dearth of high quality emotional speech data sets. Some possible solutions may include

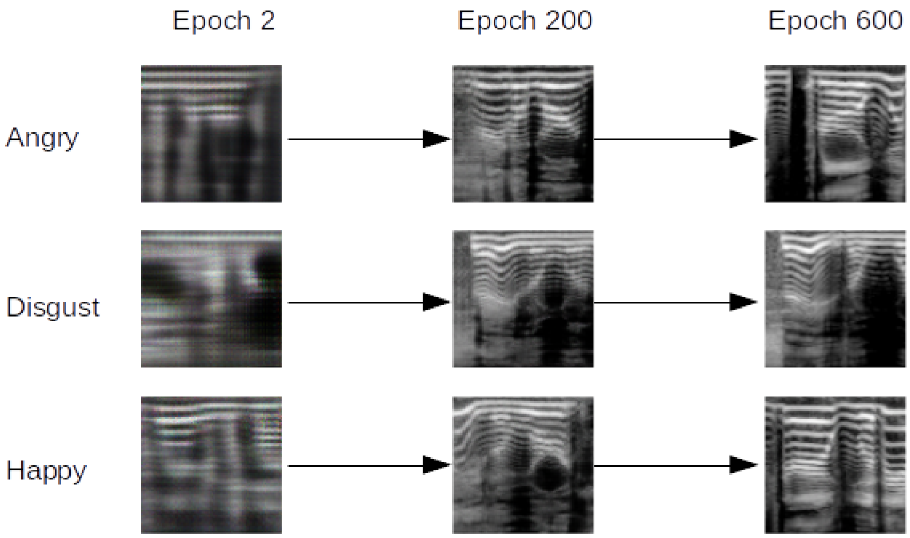


Fig. 6. Gradual convergence of generated spectrograms during training from blurred to clear frequency bands

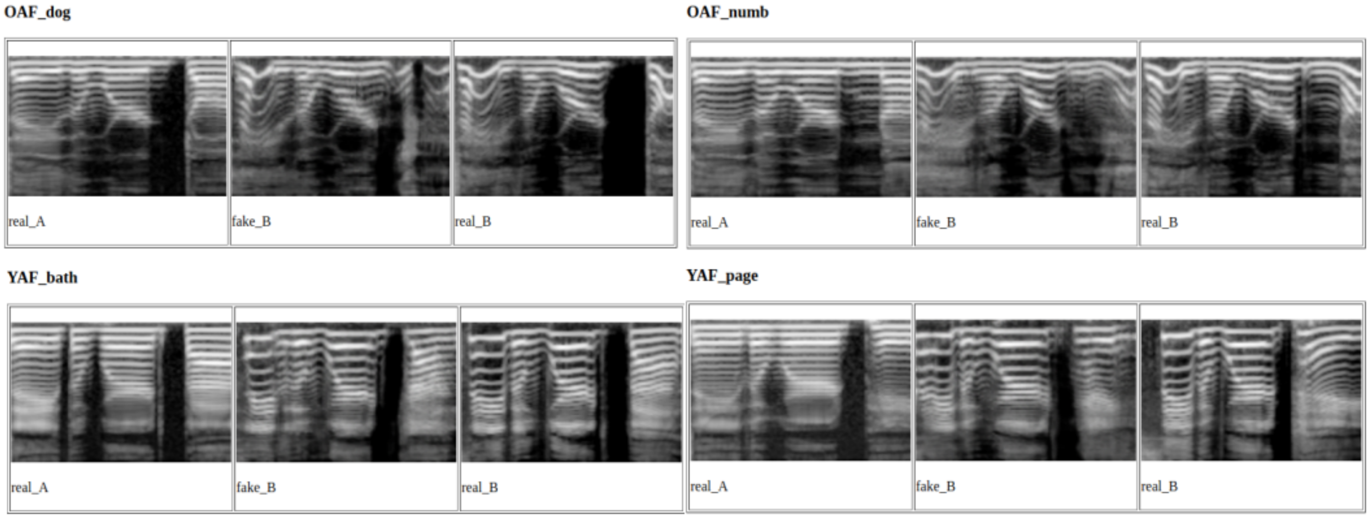


Fig. 7. Examples of final predictions vs actual neutral and actual emotional spectrograms. Generated spectrograms are in the middle while original neutral are on the left and original emotional are on the right.

stripping audio from emotional film data sets, or upscaling data sets with low quality audio via the aforementioned pix2pixHD.

- At the core, a mel spectrogram is an array of numbers (in the range $[-12, 1]$). The conversion into an image format and reconstuction back to a numeric array in this range is possibly lossy. One approach would be to **condition Pix2pix directly on the mel array**.
- Since Convolutional layers restrict the input size, we can replace Pix2pix with **RNN Encoder-Decoder systems**. This casts the problem into a similar domain as that of Neural Machine Translation. This addresses the problem of abrupt truncations. In this manner it is possible to scale the model to be able to process arbitrarily large sequences.
- Even with an Encoder-Decoder system, there is still a *pairwise correspondence between emotions*. **For** N **emotions, this means** $\binom{N}{2}$ **models constructed, which grows quadratically in** N. To solve this problem, a **Variational Auto Encoder coupled with RNN** can be used to generate sequences with the emotions embedded in a latent space. This allows for a large number of emotions (subtle and intense) represented by their coordinates on the *Valence-Arousal plane* to be incorporated.
- pix2pix currently utilizes one-to-one neutral to single emotion translation, necessitating a new model be trained for each emotion. Neutral to many emotion translation may be possible with a different image translation model such as cycleGAN [19].

References

[1] H. Choi, S. Park, J. Park, and M. Hahn, "Emotional speech synthesis for multi-speaker emotional dataset using wavenet vocoder," in *2019 IEEE International Conference on Consumer Electronics (ICCE)*. IEEE, 2019, pp. 1–2.

[2] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, 2017.

[3] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio *et al.*, "Tacotron: Towards end-to-end speech synthesis," *arXiv preprint arXiv:1703.10135*, 2017.

[4] A. v. d. 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," *arXiv preprint arXiv:1609.03499*, 2016.

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

[6] M. Schröder, "Emotional speech synthesis: A review," in *Seventh European Conference on Speech Communication and Technology*, 2001.

[7] Y. Wang, D. Stanton, Y. Zhang, R. Skerry-Ryan, E. Battenberg, J. Shor, Y. Xiao, F. Ren, Y. Jia, and R. A. Saurous, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," *arXiv preprint arXiv:1803.09017*, 2018.

[8] S.-Y. Um, S. Oh, K. Byun, I. Jang, C. Ahn, and H.-G. Kang, "Emotional speech synthesis with rich and granularized control," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7254–7258.

[9] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan *et al.*, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4779–4783.

[10] R. Prenger, R. Valle, and B. Catanzaro, "Waveglow: A flow-based generative network for speech synthesis," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 3617–3621.

[11] D. P. Kingma and P. Dhariwal, "Glow: Generative flow with invertible 1x1 convolutions," in *Advances in Neural Information Processing Systems*, 2018, pp. 10 215–10 224.

[12] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional gans," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[13] L. Sheng and E. N. Pavlovskiy, "Reducing over-smoothness in speech synthesis using generative adversarial networks," in *2019 International Multi-Conference on Engineering, Computer and Information Sciences (SIBIRCON)*. IEEE, 2019, pp. 0972–0974.

[14] D. Michelsanti and Z.-H. Tan, "Conditional generative adversarial networks for speech enhancement and noise-robust speaker verification," 09 2017.

[15] K. Dupuis and M. K. Pichora-Fuller, *Toronto emotional speech set (TESS)*. University of Toronto, Psychology Department, 2010.

[16] B. McFee, V. Lostanlen, M. McVicar, A. Metsai, S. Balke, C. Thomé, C. Raffel, A. Malek, D. Lee, F. Zalkow, K. Lee, O. Nieto, J. Mason, D. Ellis, R. Yamamoto, S. Seyfarth, E. Battenberg, , R. Bittner, K. Choi, J. Moore, Z. Wei, S. Hidaka, nullmightybofo, P. Friesch, F.-R. Stöter, D. Hereñú, T. Kim, M. Vollrath, and A. Weiss, "librosa/librosa: 0.7.2," Jan. 2020. [Online]. Available: https://doi.org/10.5281/zenodo.3606573

[17] R. V. Ryan Prenger and B. Catanzaro, "Waveglow: a flow-based generative network for speech synthesis," https://github.com/NVIDIA/waveglow, 2018.

[18] R. Yamamoto, M. Andrews, M. Petrochuk, W. Hy, cbrom, O. Vishnepolski, M. Cooper, K. Chen, and A. Pielikis, "r9y9/wavenet_vocoder: v0.1.1 release," Oct. 2018. [Online]. Available: https://doi.org/10.5281/zenodo.1472609

[19] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.