American Journal of Computer Science and Technology
2021; 4(3): 75-82
http://www.sciencepublishinggroup.com/j/ajcst
doi: 10.11648/j.ajcst.20210403.13
ISSN: 2640-0111 (Print); ISSN: 2640-012X (Online)

Activation Maximization with a Prior in Speech Data

Sho Inoue, Tad Gonsalves

Department of Information & Communication Sciences, Sophia University, Tokyo, Japan

Email address:
s-inoue-tgz@eagle.sophia.ac.jp (S. Inoue), t-gonsal@sophia.ac.jp (T. Gonsalves)

To cite this article:
Adane Fekadu Wogu, Shanshan Zhao, Hazel B Nichols, Jianwen Cai. Activation Maximization with a Prior in Speech Data. *American Journal of Computer Science and Technology*. Vol. 4, No. 3, 2021, pp. 75-82. doi: 10.11648/j.ajcst.20210403.13

Received: July 30, 2021; **Accepted:** August 13, 2021; **Published:** August 31, 2021

Abstract: Recently, more and more studies regarding neural networks have been done. However, the learning process of neural networks is often elusive to human beings, which leads to the advent of feature visualization techniques. Activation Maximization (AM) is one of the feature visualization techniques, originally designed for image data. In AM, the input data is optimized to find the data that activates the selected neuron. In this paper, the emotion recognizer's output is selected as the neuron, and the latent code of a generator (of Generative Adversarial Networks) is optimized instead of the input raw data. The aim of this study is to apply AM to different representations of audio data (waveform-based data and mel-spectrogram-based data) and different model structures (CNN, WaveNet, LSTM), and to find out the most suitable condition for AM in audio domain data. Additionally, we have also tried to visualize the essential features of being a certain class for emotion classification in speech data, using 2 datasets: the Toronto emotional speech set (TESS) and the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). The mel-spectrogram-based models were found to be superior to the others, showing the distinctive features of selected emotions. More specifically, the CNN-mel-spectrogram-based model was the best in both qualitative and quantitative (FID score) results. Moreover, as demonstrated in this study, AM can also be employed as an output enhancer for generative models.

Keywords: Deep Learning, Signal Processing, Feature Visualization, Activation Maximization, GAN

1. Introduction

Neural Networks are predominant for processing various tasks such as object detection, speech recognition, emotion detection, and so on. However, their internal processing is, in general, not understandable for human beings. To understand how the models tackle the problems, some visualization techniques such as feature visualizations are proposed. In this paper, we have discussed the applications of Activation Maximization(AM) [1], which is one of the feature visualization techniques.

In AM, the input data is optimized to the data that activates the selected neuron in the network in order to understand how the models make a decision. The reference neurons can be the filter of layers, the pre-trained classifier's output, and so on. There are quite a few applications of AM such as class-based activation maximization [2] and activation maximization with a prior (generator) [3]. Both of these are relevant to our experimentation. In the former application, the input data is gradually altered to improve the output of the learned classifier to observe the reason for being in a certain class. Therefore it is called class-based Activation Maximization. In the latter concept, they optimize the noise of the generator in the generative models such as Generative Adversarial Networks (GAN) [4] which is employed as a prior.

In this paper, we have applied AM to audio data in different conditions, which are defined by the form of data and the classifier's structure and make a comparison among different conditions. There was no application of Activation Maximization in raw audio with a vocoder, which is modeled after WaveGlow [5]. A generator in GAN is also employed as a prior for AM to increase the stability. It is trained with two audio emotion datasets, namely the Toronto emotional speech set (TESS) [6] and the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [7]. The purpose of the experiment is two-fold: (1) To find out the features in a speech that are important for the classifier to judge the class by optimizing the input audio data into the data which activates the score of the classification. (2) To examine the different conditions and find out the one suitable for AM. In addition,

from a different perspective, AM acts as an output enhancer. In other words, AM strengthens the output of the generator to have a distinctive feature of a specific emotion. This could be used as any kind of generator that has the noise vector as input. The implementation and the results of the audio are available at: https://github.com/shinshoji01/AM_with_GAN_for_melspectrogram.

2. Related Work

2.1. GAN

Generative Adversarial Networks (GAN) [4] is one of the deep generative models and is often compared with Variational Auto-Encoder (VAE) [8]. VAE tries to reconstruct exactly the same data as the input, whereas GAN learns to generate data that is similar to the input data distribution (not necessarily the same data). GAN is basically composed of two models: Generator (G) and Discriminator (D). G generates data that is able to deceive D and D tries to judge whether the input data is real or fake generated by G. GAN optimizes the function below:

$$\min_{G}\max_{D} V(D,G) = \mathbb{E}_{x\sim p_{data}(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1-D(G(z))] \ ) \quad (1)$$

Where the first term indicates the judgement of the real data and the second shows that of generated data. In addition, x and z represent the input data and the noise vector, respectively. The noise vector z was optimized in our experiment. GAN is generally considered as having an unstable training process, so we employ 2 techniques to stabilize: UnrolledGAN [9] and Consistency Regularization [10]

2.2. Activation Maximization

Activation Maximization (AM) is first proposed to visualize the way and the reason why deep neural networks work [1]. In general, in deep neural networks, the model's weights (parameters) are optimized to have a better result via backpropagation. Instead, the input data tries to be optimized to the data that maximizes the activation of selected neurons in AM. These include the filter of layers, the classification output [2], and so on. AM is initially volatile and cannot find the optimal data. The data tends to end up with an incomprehensive result that excessively activate the selected neuron. To tackle this problem, some regularization techniques were invented, including frequency penalization [11], dataset examples [12], transformation robustness [13], and the employment of the learned prior [3]. In our case, the noise of the generator in GAN, which is employed as a prior, is optimized to activate the output of the classifier to observe how the model determines the classes.

2.3. Tacotron2 and WaveGlow

There was a breakthrough in signal data generation when WaveNet [14] was proposed. It was first invented as the generative model of raw audio data. Previously, the reconstruction of audio data from mel-spectrogram was considered a difficult problem due to the simplicity of the mel-spectrogram. However, it was solved by the vocoder modeled after WaveNet, which was employed to achieve a whole speech generation system with Tacotron2 [15]. Tacotron2 was developed from the former version named Tacotron [16], which makes use of either linear-scale spectrogram or mel-scale spectrogram as the intermediary and reconstruct speech data from them. These models had a great impact on the development of speech synthesis, especially in Tacotron2 that was found to be able to generate neutral speech with high fidelity. This system has improved its performance with the change of vocoder, one of which was WaveGlow [5]. This model borrows ideas from Glow [17] and WaveNet to generate high-quality speech from mel-spectrogram with a flow-based network. Additionally, thanks to the flow-based network, it enables us to generate speech much faster.

3. Experiment

3.1. Overview

In this experiment, we are going to apply the class-based Activation Maximization (AM) to the audio domain. As shown in Figure 1(a), a generator in Generative Adversarial Networks [4] was used as a prior for stable training process, which is detailed in Sec.3.5. The experiment includes the effectiveness of using different forms of data and the model (classifier) structures. As for the form of audio data, 2 types of audio features are employed, which are raw audio and mel-spectrogram. Three different models are implemented: CNN-based, WaveNet-based, and LSTM-based models.

3.2. Mel-Spectrogram-based Activation Maximization

In mel-spectrogram-based AM, the classifier's input is a mel-spectrogram and the generator's output is also a mel-spectrogram. Therefore, the generator and the classifier are directly connected to each other as shown in Figure 1(b) and the input of the generator (noise vector) is optimized to make the output more emotional.

3.3. Waveform-based Activation Maximization

As for waveform-based AM, since the classifier uses the waveform data instead of mel-spectrogram as the input, the model, which converts the mel-spectrogram into the audio, is required and a pre-trained WaveGlow is capable of this task. Thus, the flowchart becomes as shown in Figure 1(c).

3.4. Classifier

As class-based AM, the classifier plays an important role and it could strongly affect the validity of the results. Therefore, we have compared 3 kinds of model structures, which can be considered as CNN-based, WaveNet-based, and LSTM-based models.

CNN-based model: CNN stands for Convolutional Neural Network. CNN-based model just has a simple structure which consists of some convolutional blocks with a max-pooling layer as compression and a ReLU function as an activation function, and a fully-connected layer after a global average pooling layer [18]. Moreover, a batch normalization [19] is installed between a max-pooling layer and a ReLU function.

WaveNet-based model: WaveNet [14] is first proposed as a deep generative model for raw audio and it was one of the largest steps for raw audio generation. The biggest takeaway of this model is its massive receptive field due to the employment of various dilated convolutional layers. So, in this model, we have built a classifier based on the structure of WaveNet to allow the model to have a large receptive field.

LSTM-based model: When humans understand sequential data such as audio and sentences, they usually consider the former part of data. This feature was first imitated in Recurrent Neural Networks (RNN) and currently it has a lot of applications. One of them is called Long Short TermMemory networks (LSTM) [20] and it solved the problem of RNN not being able to 'memorize' the data in the distant past. In this model, LSTM is employed followed by some stacked-convolutional blocks.

Hyper-parameter tuning is done in all models and the optimal and detailed structures can be found at https://github.com/shinshoji01/AM_with_GAN_for_melspectrogram.

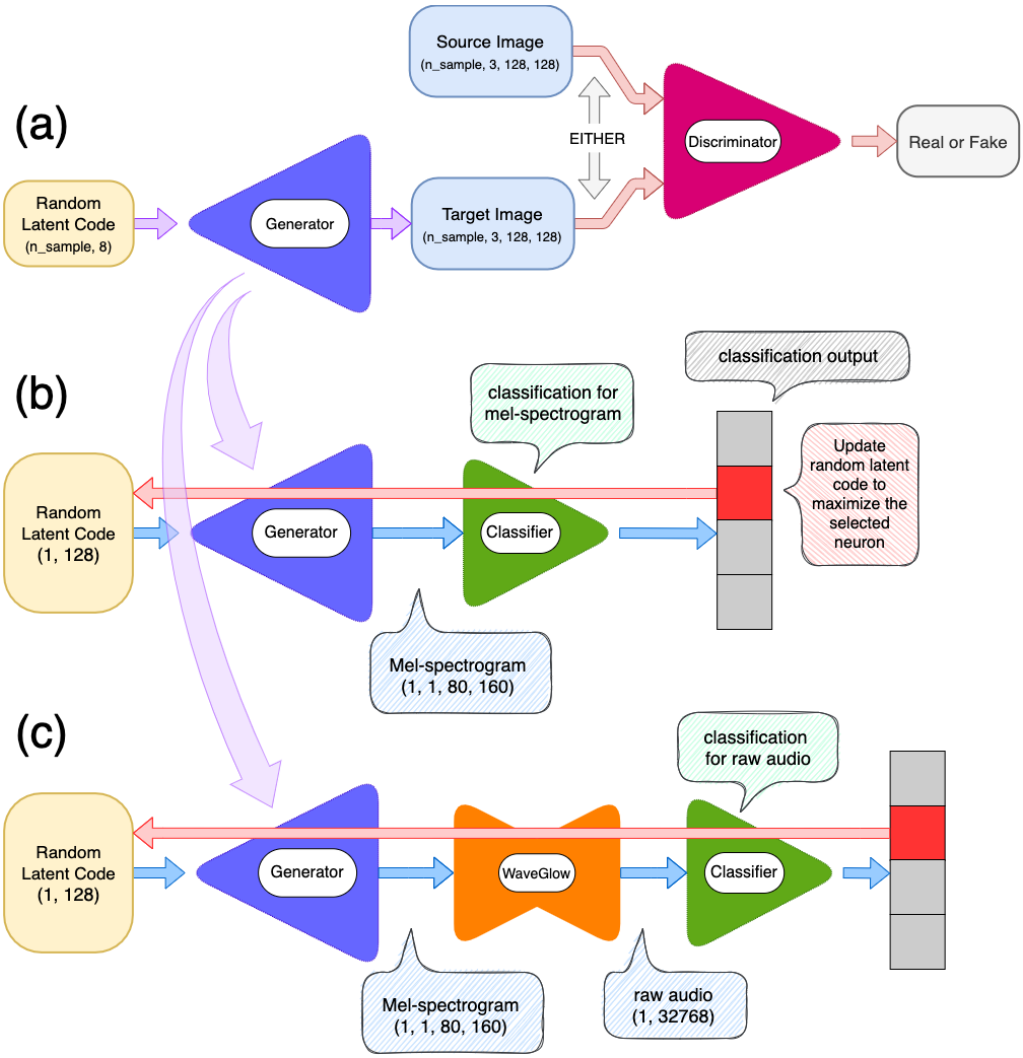

Figure 1. *(a) Flowchart of GAN (b) Mel-spectrogram-based AM (c) Audio-based AM.*

3.5. GAN

The main part of this model is to employ a generator as a prior to stabilize the result. For this reason, we also make use of GAN. GAN requires 2 models to train; the generator and the discriminator, and both structures are modeled after DCGAN [21]. More specifically, the pooling layer is eliminated, the compression process is done by setting the stride of the convolutional layer as 2, and the transposed convolution layer is employed for the expansion process. In addition, to make the generator learn semantic representation and to stabilize the training process, consistency regularization [10] and UnrolledGAN [9] are adopted, respectively. As for the training condition, the hyper parameters $\lambda_{consistency\_regularization}$, $\lambda_{k\_for\_UnrolledGAN}$, and the learning rate were 0.1, 5, 0.01, respectively. The loss function is MSE Loss as employed in LSGAN [22]. In addition, Adam optimizer [23] and learning rate scheduler which reduces the learning rate exponentially were employed. As for these parameters, $\beta_1 = 0.5$ and $\beta_2 = 0.999$ were set for the optimizer and $\gamma = 0.99$ for the scheduler.

3.6. Condition in Activation Maximization

In general, original activation maximization is not quite as stable to work as the other machine learning models. Some regularization techniques were invented to tackle this problem, which is detailed in Sec.2.2. In this paper, we have made use of 3 kinds of regularization techniques, including L2 regularization, preventing undesired output, and a learned prior. Eventually, the loss will be:

$$\text{loss} = -desired + \lambda_{undesired} * undesired + \lambda_{L2} * \|noise\_vector\|^2 \tag{2}$$

Where, $desired$ indicates the score of selected classes before softmax function as mentioned here [24] and similarly, $undesired$ is the score of *not* selected classes before softmax function. The $noise\_vector$ is the noise vector that is inserted into the generator. The hyper parameters; $\lambda_{undesired}$, λ_{L2}, and learning rate were 1, 100, 0.01, respectively. In addition, Adam optimizer [23] and learning rate scheduler which reduces the learning rate exponentially were employed. As for these parameters, $\beta_1 = 0.9$ and $\beta_2 = 0.999$ were set for the optimizer and $\gamma = 0.99$ for the scheduler.

3.7. Quantitative Evaluation

To avoid the analysis being too subjective, we have taken an analysis based on the FID score [25]. FID score indicates the similarity of 2 sets of representations. In our case, we have computed the FID score between the samples in the dataset and the same amount of optimized samples of AM (80 samples), so, for instance, the "angry" audio in the TESS dataset and the samples which are optimized to be more "angry" sound. However, the inception v3 model [26], which is used for the computation of FID, is originally trained with ImageNet [27], so, we could not assure that FID is appropriate for evaluation in mel-spectrogram. Therefore, in addition to the model pretrained with ImageNet, we compute FID score with the model pretrained with the sound datasets used in this experiment. The learning rate was 0.01 and the loss function was Cross Entropy Loss. Similar to Activation Maximization, Adam optimizer [23] and learning rate scheduler which reduces the learning rate exponentially were employed. As for these parameters, $\beta_1 = 0.9$ and $\beta_2 = 0.999$ were set for the optimizer and $\gamma = 0.99$ for the scheduler.

4. Dataset and Preprocessing

2 datasets were used in this experiment: The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [7] and Toronto emotional speech set (TESS) [6]. For preprocessing, some techniques are employed for training, such as silent removal and mel-spectrogram conversion. The RAVDESS dataset includes speech and song audio whose sampling rate is 48kHz and audio depth is 16 bit. 2 statements are spoken by 24 actors (12 female and 12 male) in 8 emotions such as "neutral", "happy", and so on. TESS is also an audio dataset labeled with 7 emotions such as "fear", "disgust" and so on. The sampling rate is 24414Hz. The sentence is spoken by 2 female speakers and every audio is prefaced with "Say the word" followed by a certain word. Since they have different sampling rates, it is downsampled to 22050 Hz and its silent section is removed. As the data length is different from each other, it is adjusted by random-zero-padding and random-cropping. In addition, the audio labeled with "calm" is integrated into the label "neutral" due to the similarity. Eventually, the dataset is utilized in the following condition.

1. sampling rate: 22050 Hz
2. no silent section
3. emotions: "neutral", "happy", "sad", "angry"
4. audio depth: 16 bit

Concerning mel-spectrogram conversion, we have imitated the parameter used in Tacotron2 [15]. To sum up, the parameters are:

1. frame size (STFT): 50 ms
2. frame hop (STFT): 12.5 ms
3. mel scale: 80 channels

5. Results

In this experiment, we tried the class-based Activation Maximization in (AM) audio domain. Since we are not able to post any audio data in this paper, we have uploaded the result with the soundtrack on GitHub (https://github.com/shinshoji01/AM_with_GAN_for_melspectrogram). In this paper, the results are represented in the form of a mel-spectrogram computed with the parameters

in training. We analyzed the result in both qualitative and quantitative ways. The classifier structures for mel-spectrogram-based AM were CNN-based, LSTM-based, and WaveNet-based. Similarly, those for waveform-based AM were LSTM-based and WaveNet-based.

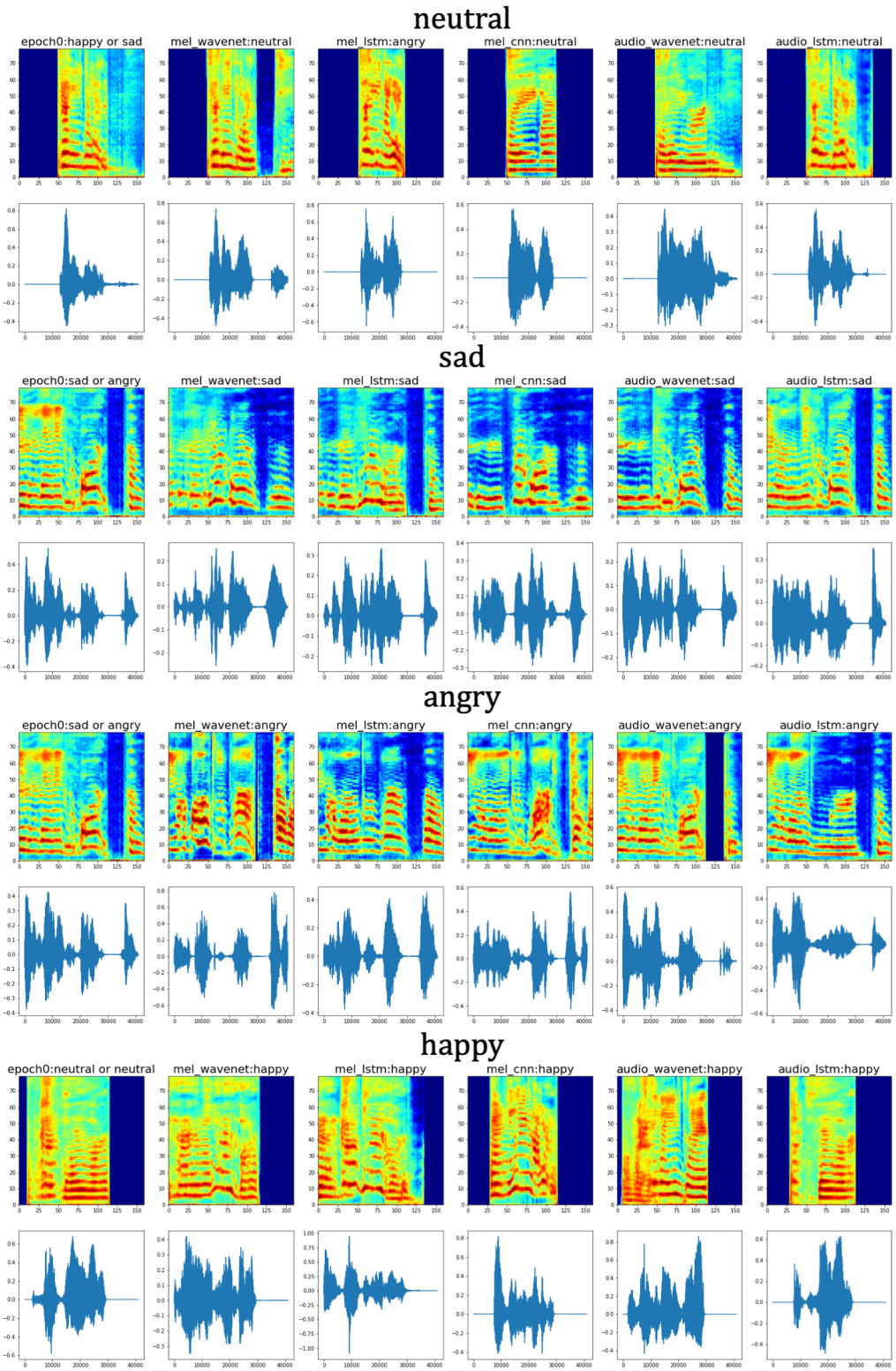

Figure 2. *Qualitative Result: initial (the left column) and optimized (the rest columns) mel-spectrograms (waveform).*

5.1. Qualitative Analysis

Activation Maximization is an algorithm to explain the thought process of the model by optimizing the vector to activate the selected neuron. Therefore, it is important to go over (in this case, and listen to) the result carefully and individually.

We posted the distinctive result in Figure 2. Primarily, it shows the initial result where AM is not applied in the first column, followed by the results optimized with the mel-WaveNet-based, mel-LSTM-based, mel-CNN-based, audio-WaveNet-based, and audio-LSTM-based models. The bold texts indicate the emotion which AM aims at and the emotions just above each mel-spectrogram represent the final predicted emotions.

As for the result of neutral emotion, it was not classified as "neutral" in the first place. However, some models have greatly succeeded in optimization such as mel-CNN and audio-WaveNet models. The intonation became monotonous. In general, the neutral sound had horizontal lines in the mel-spectrogram and it seems to lead to plain sound. As far as sad emotion is concerned, mel-spectrogram based models had better results. The ups and downs of the intonation have diminished and the strong accent has transferred from the beginning to the middle as shown in the waveform. The intonation became weaker in the last part. On the other hand, in the angry emotion, the gap between the highs and lows of the intonation was intense and it changed on and off quickly. Mel-spectrogram-based models were superior to the audio-based models. We can find unambiguous results in the happy emotion. Most models have changed the mel-spectrogram to reshape it to have a mountainous line in the middle of a speech. When we listen to the sounds, we can hear the drastic surge of the intonation in the middle.

To sum up, the mel-spectrogram-based models had better results than the audio-based models due to simplicity. The more simple the input data is, the more stable the training process can be. Mel-spectrogram-based models had good results because, at least in emotion, we can barely find the differences only by observing the mel-spectrogram. Presumably, there is an optimal result within the whole sets in audio-based models, but there is no efficient way to tune the hyper-parameters in the feature visualization field. Therefore, we should prioritize stability or simplicity rather than the possibility of getting a perfect result.

Table 1. *FID scores: ImageNet.*

			mel-spectrogram		audio	
Emotion	**Data set**	**wavenet**	**lstm**	**cnn**	**wavenet**	**lstm**
neutral	ravdess	169.842	158.846	143.283*	151.482	164.045
	tess	158.807	156.328	145.729*	149.674	154.831
happy	ravdess	180.494	173.904	167.680	160.182	159.255*
	tess	182.499	182.426	164.719*	166.192	166.248
sad	ravdess	144.581	134.962	140.281	130.331	129.905*
	tess	164.213	180.635	158.353*	173.494	160.990
angry	ravdess	171.898	175.071	164.053	165.599	161.147*
	tess	181.111	182.707	153.969*	168.848	163.930

The value with * means the highest score among the models.

Table 2. *FID scores: TESS and RAVDESS.*

			mel-spectrogram		audio	
Emotion	**Data set**	**wavenet**	**lstm**	**cnn**	**wavenet**	**lstm**
neutral	ravdess	511.088	338.269	306.979*	482.245	786.82
	tess	261.918	281.961	175.165*	196.721	333.123
happy	ravdess	390.420	385.269	254.387*	340.724	406.724
	tess	568.427	587.502	399.811*	522.355	618.015
sad	ravdess	96.635	63.138	38.862*	155.651	213.153
	tess	110.931	84.246	47.317*	207.254	260.014
angry	ravdess	167.203	327.067	70.736*	437.072	482.265
	tess	192.875	730.336	100.696*	856.200	945.929

The value with * means the highest score among the models.

5.2. Quantitative Analysis

The method to evaluate the result quantitatively is mentioned in Sec. 3.7. The average of the 80 FID scores in each emotion and classifier is illustrated below (Table 1 and Table 2 for the models pretrained with ImageNet and the sound datasets, respectively). In the original FID score, obviously, mel-CNN-based models were better than the other models in most cases, especially in neutral emotion. Nonetheless, the audio-based model had a comparatively good result in the emotions other than "neutral" and it might be because of the way to show the emotion. In other words, these emotions are often expressed using the accent in the sentence, which is elusive in mel-spectrogram. When it comes to the feature extractor pretrained with the sound datasets (Table 2), mel-spectrogram-CNN model was superior to the others in all cases and other mel-spectrogram based models were better than audio-based models in most cases.

6. Conclusion

We have demonstrated the application of Activation Maximization (AM) specifically class-based AM with a prior (a generator) in audio data. The classifier, which is essential for AM, varied in the structure and representation of data. Overall, the mel-spectrogram-based models, especially CNN model, are superior to the audio-based models in both qualitative as well as quantitative results. In the qualitatives result, we could find out the reason why and how the emotion recognizer classified the emotions. Take the emotion "angry", for example; the intonation of the generated audio was changed drastically, indicating that the speech with an angry emotion tends to have that feature. In the quantitative result, we employed FID score with not only the model pretrained with ImageNet (the original model), but also with the model pretrained with sound datasets (TESS and RAVDESS). Both the evaluation methods proved that the CNN-mel-spectrogram-based model was better than any other model, especially with the sound datasets. To sum up, AM has altered the output of the generator by activating the desired score. However, seeing from another point of view, AM can be regarded as the enhancer of the generator. The future plan is to conduct research on the following related topics:

1. AM while fixing the selected feature, such as emotion conversion without changing the text information.
2. Using sophisticated generator as priors and use AM as an output enhancer.

References

[1] Dumitru Erhan, Y. Bengio, Aaron Courville, and Pascal Vincent. Visualizing higher-layer features of a deep network. *Technical Report, UniveristÃ© de MontrÃ©al*, 01 2009.

[2] K. Simonyan, A. Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *CoRR*, abs/1312.6034, 2014.

[3] Anh Nguyen, Alexey Dosovitskiy, Jason Yosinski, Thomas Brox, and Jeff Clune. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks, 2016.

[4] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014.

[5] Ryan Prenger, Rafael Valle, and Bryan Catanzaro. Waveglow: A flow-based generative network for speech synthesis, 2018.

[6] M. Kathleen Pichora-Fuller and Kate Dupuis. Toronto emotional speech set (TESS), 2020.

[7] Steven R. Livingstone and Frank A. Russo. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), April 2018. Funding Information Natural Sciences and Engineering Research Council of Canada: 2012-341583 Hear the world research chair in music and emotional speech from Phonak.

[8] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2014.

[9] Luke Metz, Ben Poole, David Pfau, and Jascha Sohl-Dickstein. Unrolled generative adversarial networks, 2017.

[10] Han Zhang, Zizhao Zhang, Augustus Odena, and Honglak Lee. Consistency regularization for generative adversarial networks. In *International Conference on Learning Representations*, 2020.

[11] Aravindh Mahendran and Andrea Vedaldi. Understanding deep image representations by inverting them, 2014.

[12] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks, 2014.

[13] Alexander Mordvintsev, Christopher Olah, and Mike Tyka. Inceptionism: Going deeper into neural networks, 2015.

[14] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio, 2016.

[15] Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, RJ Skerry-Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis, and Yonghui Wu. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions, 2018.

[16] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc Le, Yannis Agiomyrgiannakis, Rob Clark, and Rif A. Saurous. Tacotron: Towards end-to-end speech synthesis, 2017.

[17] Diederik P. Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions, 2018.

[18] Min Lin, Qiang Chen, and Shuicheng Yan. Network in network, 2014.

[19] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 448–456, Lille, France, 07–09 Jul 2015. PMLR.

[20] Keiron O'Shea and Ryan Nash. An introduction to convolutional neural networks. *ArXiv e-prints*, 11 2015.

[21] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks, 2016.

[22] Xudong Mao, Qing Li, Haoran Xie, Raymond Y. K. Lau, and Zhen Wang. Multi-class generative adversarial networks with the L2 loss function. *CoRR*, abs/1611.04076, 2016.

[23] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

[24] Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature visualization. *Distill*, 2017. https://distill.pub/2017/feature-visualization.

[25] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium, 2018.

[26] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision, 2015.

[27] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.