Identifying Gendered Vocal Features with WaveNet

Stephanie Huang
Harvey Mudd College
sthuang@hmc.edu

Saheli Patel
Harvey Mudd College
sapatel@hmc.edu

Abstract

This project investigates whether deep-learning neural networks can detect meaningful vocal features, specifically pitch and resonance, contributing to gender perception through vocal gender-recognition tasks. This is done using WaveNet, a deep-learning model for raw audio generation. Experiments show that the network captures features resembling resonance while pitch is not so discernably captured. Experiments also show that the network is sensitive to linguistic differences which correlate with qualitative differences in pitch.

1 Introduction and related work

Convolutional neural networks are known for their ability to learn and pick up "features" from data [1]. These networks were initially built for image recognition tasks but were later adapted for auditory tasks. The recent WaveNet model uses this for raw audio generation but its architecture can also be extended to audio recognition tasks. While most applications of WaveNet utilize its generational prowess, this project explores WaveNet's recognition ability; we want to see if neural networks like WaveNet are capable of identifying features that are meaningful to humans. Specifically, we want to see how neural networks can recognize androgenized voices and if they do so by picking up features that are significant to humans (i.e. pitch and resonance). This may help deconstruct how convolution-based neural networks like WaveNet function and determine if they are sensitive to meaningful vocal patterns contributing to gender perception.

2 Datasets

The WaveNet model was trained on the Mozilla Common Voice dataset. This dataset consists of 1-10 second audio voice files across multiple languages. Languages included Arabic, German, English, Spanish, French, Italian, Japanese, Rwandan, Swahili, Thai, Turkish, and Chinese. Files are labeled with speaker sex, age, and language. The voices were volunteered into the dataset.

3 Methods

3.1 WaveNet

We used the Wavenet architecture as outlined in "Wavenet: A generative model for raw audio" [2]. While the model was designed with voice synthesis in mind, we have modified the model to output binary predictions. Our model outputs a series of predictions across the voice sample. To make a prediction, we then average the results into one number with a decision threshold of .5. A prediction above .5 is a prediction that a voice is female and a prediction below .5 is a prediction that the voice is male.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

3.2 Training

The model was trained on a mixture of Arabic, Begnali, Welsh, German, English, Spanish, Basque, Frnech, Italian, Japanese, Russian, Kinyarwanda, Swahili, Thai, and Turkish. The training set was chosen from the premade training set provided by Mozilla. Additionally, we excluded voices from teens and pre-adolescents from the training sets. Each language was represented equally in the dataset with male and female voices being balanced across languages. This did mean that we needed to drop samples from the dataset so that we could balance out representation of languages and speaker sex. The model was fed 1 second clips from each sample present in the dataset, the clips were randomly chosen from the sample each epoch. We used Binary Cross Entropy as the loss function as the model was performing binary classification.

3.3 Feature Extraction

The supposed features were extracted using the filters of the channels in the last convolutional layer. To gauge what these channels were capturing, we fed the trained model five batches of 32 voices sampled from the following languages, which we'll dub as the "Language Set": Arabic, German, English, Spanish, French, Italian, Japanese, Rwandan, Swahili, Thai, Turkish, and Chinese. The results would be merged into one batch of 160 voices (batch size was capped at around 40 due to data constraints on the server). The filters of the twenty strongest positively-weighted channels and the twenty strongest negatively-weighted channels were then analyzed to see if they were capturing any meaningful features (we'll call these features/channels "significant" features from now on).

4 Results and experiments

4.1 Training and Validation

The model achieved a training loss of .3648. For validation we validated the model over several languages separately. We shall present the results for validation over English, Spanish, Arabic and Polish.

The accuracy over English was 82.6%, Spanish had 89.2% accuracy, Arabic had 91.9% accuracy, and Polish had 88.2% accuracy. Polish was not in the training set, yet the model performed well over the dataset. As shown by Figure 1 the model tends to classify male and female to curves that we would expect, where there is a peak for male near 0 and a peak for female near 1 and the curve tends to taper off as it goes in the other direction. Note that for English, the model is biased towards classifying speakers as male however in Arabic, the model is biased towards classifying the speaker as female.

4.2 Classifications

Across our multi-language set, the average proportion of female predictions was around 0.49-0.53. So there seems to be an even split between male and female classifications. We will refer back to this in our "Linguistic Comparisons" section.

4.3 Features

4.3.1 Pitch

Pitch is perhaps one of the first features that comes to mind when thinking about vocal differences. To determine if the learned features were coded for pitch, we tried to determine if there was a strong correlation between pitch and activation for each significant feature. This was done with two experiments.

In the first experiment, we compared areas of high activation with the areas of high pitch for each significant feature. We first used numpy's built-in Fourier analysis to determine the relative frequencies (often a suitable proxy for pitch [3]) of each batch file across time. A frequency heatmap was then constructed, as shown in figure 3. This was then compared to the heatmaps of the 40 most significant features.

It's worth mentioning that we sorted the batch files based on the model's prediction of how "feminine" they were. This would allow us to visualize any patterns that may be meaningful since heatmaps for

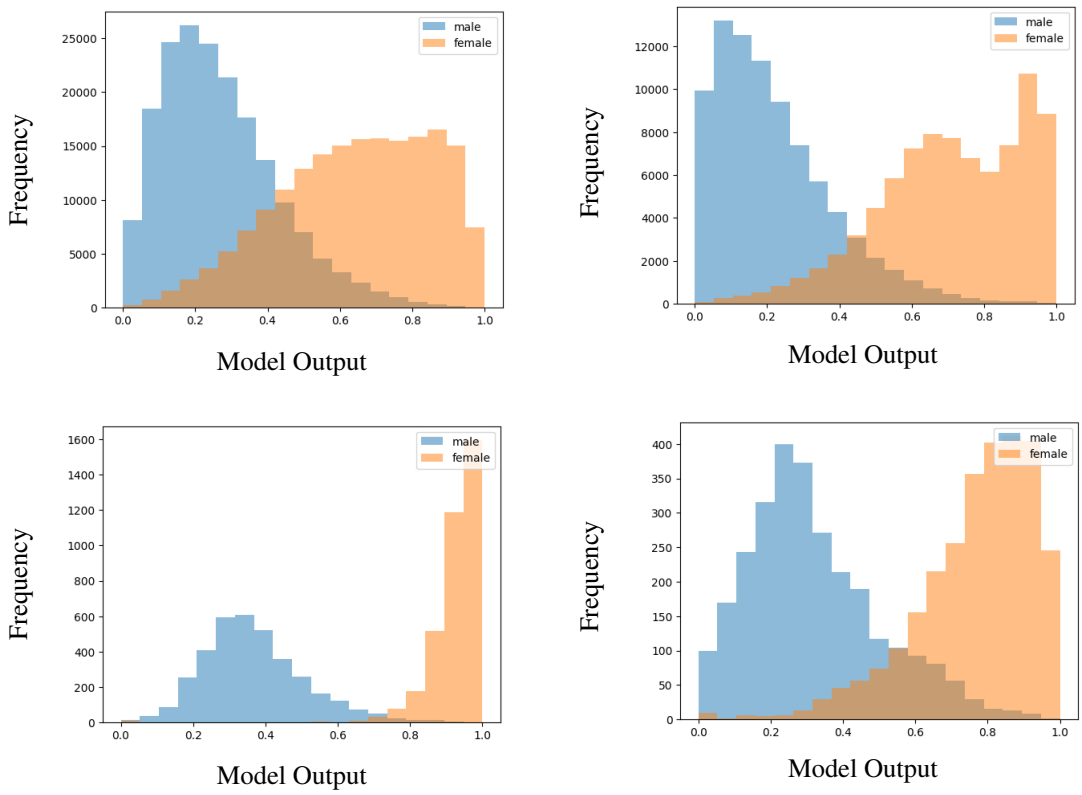

Figure 1: A histogram of the model's outputs that represents how it categorizes various speakers. The color represents the true classification of the speaker. Top Left: English, Top Right: Spanish, Bottom Left: Arabic, Bottom Right: Polish

features that strongly correlate with gender perception may appear as strong gradients, as features 132 and 173 exemplify in figure 4.

While the frequency heatmap exhibited patterns that other channel heatmaps had (higher activations for more female voices for example), we weren't able to find one that was a close match. Thus, we can only conclude that some features may be correlated with pitch but weren't able to conclude that any features explicitly coded for pitch or frequency.

In the second experiment, we compared *average* frequency of each batch file with its average level activation for each feature. Interestingly, we found that average pitch appeared to correlate positively with perceived femininity, as shown in figure 5, but did not appear to correlate with any of the significant features, exemplified in figure 6.

From these experiments, we were unable to conclude that pitch was adequately captured by the significant features in our WaveNet model.

4.3.2 Formants

We took each filter and found the subsequences that caused each filter to have the greatest activation value and then took the subsequences and performed Fourier analysis on the sequence. As a result, in some filters we got results that indicated that the filters seemed to be activated by patterns that are consistent with formants. Formants are patterns in spectrograms that appear to be lines that are produced by vowel sounds. Formants are important in the detection of resonance in the voice. Thus the presence of filters that detect formants indicate that these filters may be attempting to detect resonance in the voice. It is important to note that the presence of formants in these clips is not consistent and so its not entirely clear what the filter is detecting, but it appears that formants may be one of the feautes that are being detected.

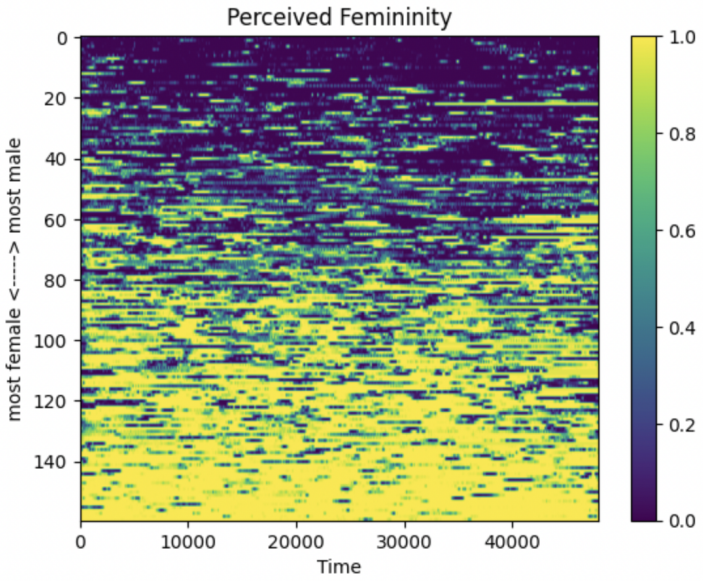

Figure 2: Perceived Femininity for multi-language batch files. Files are sorted from most male (least female) to most female

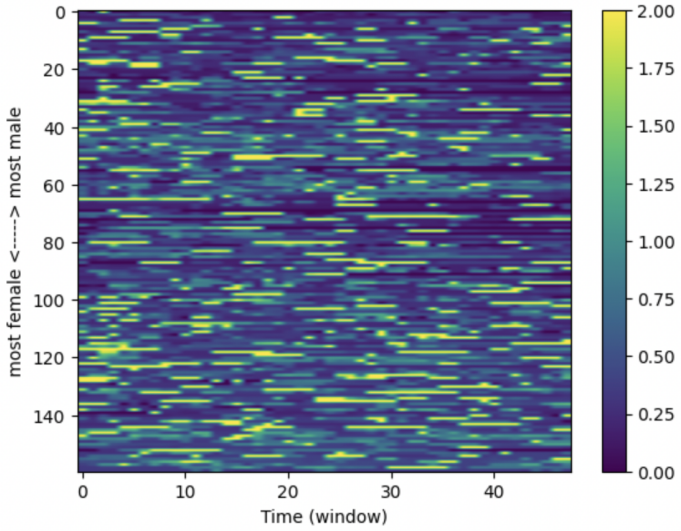

Figure 3: Frequency heatmap across audio files, sorted from most female to most male

4.4 Linguistic Comparison

There's a big possibility that the learned features aren't just physiological but also linguistic. To test this, we fed the trained model five batches of 32 for each language in our Language Set and looked at the average perceived femininity relative to other languages. We found that some languages were perceived to be more feminine than others, as exhibited in

Possible causal factors we looked into were pitch and we found that these rankings roughly correlated with pitch, as shown in figure 9. Because we weren't able to find specific features that closely matched with pitch, it's possible that pitch may not be a direct causal factor and it just happened to correlate with some of the other features the model was detecting. However, this does show evidence of qualitative differences between languages that may contribute to gender perception.

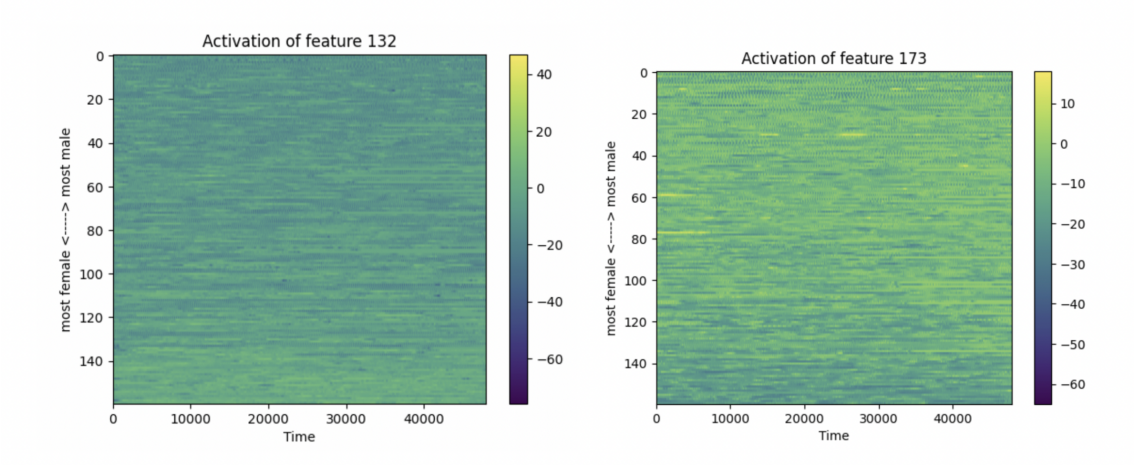

Figure 4: Activation heatmaps for features 132 and 173, both significant features. Feature 132 is a positively-weighed feature while 173 is negatively-weighted.

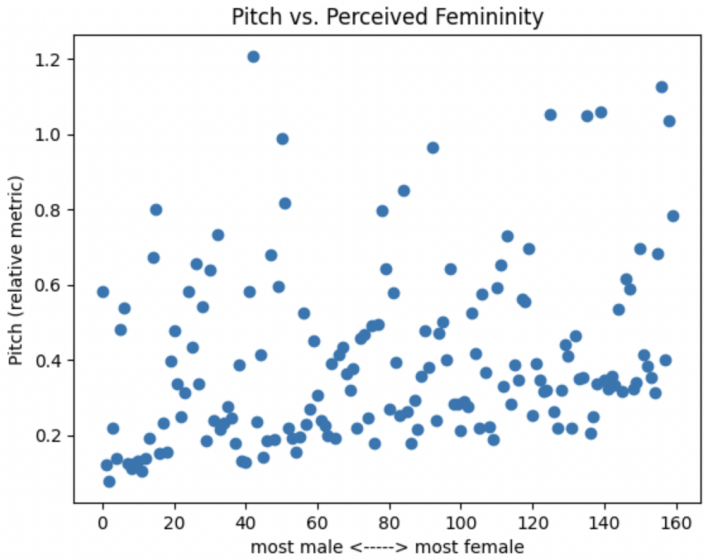

Figure 5: Weak positive correlation between pitch and perceived femininity

5 Conclusion and future work

From the results of the model's classification, we found that the model could generalize speaker sex classifications beyond the languages it was trained on. Thus implying that cross-cultural speaker sex classification is feasible. We also found that the model's classifications were weakly correlated with pitch, thus implying that pitch may be possibly playing some role in the model's decision. This is further supported by the fact that languages with higher average pitch tended to be classified as more female. Because we weren't able to find a filter that closely matched our proxy of pitch, it is still an open question as to whether this is specifically a feature that the model picks up or something that just happens to correlate with other features. We also found that the model had trained specific filters to detect formants and thus detect resonance in the voice. It is not yet clear what the model is doing with these filters.

We would recommend that others use our results to understand that while there are some human-understandable ways to understand how the model works, the model seems to work in ways that are

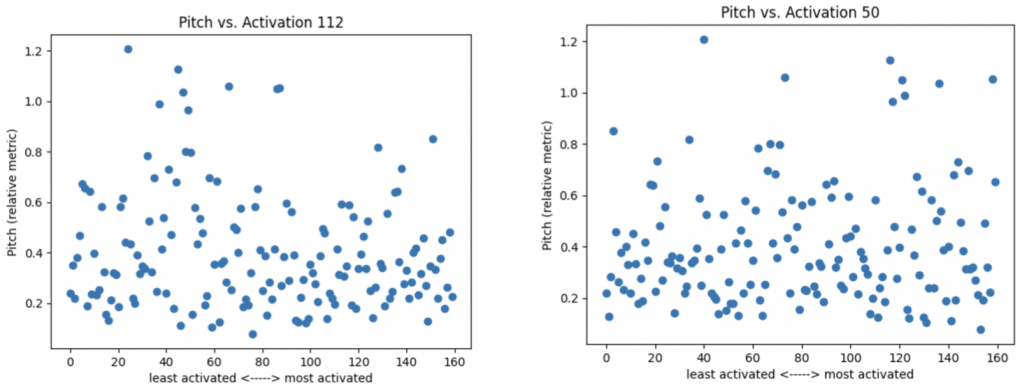

Figure 6: Features 112 and 50 are significant features but show little to no correlation with pitch. Graphs for other significant features appear the same

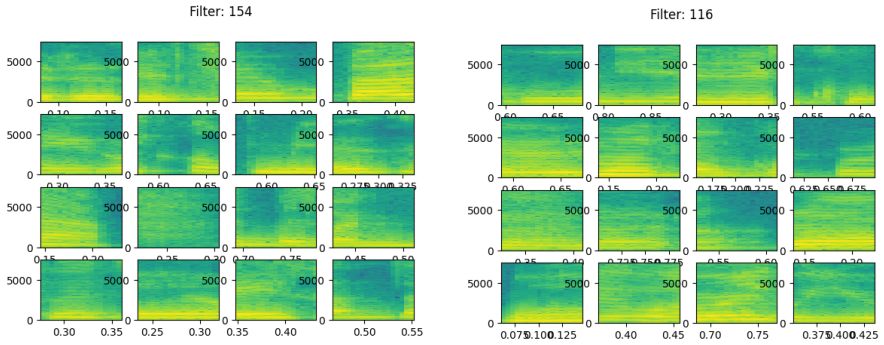

Figure 7: Filters 154 and 116 exhibit patterns that in the sequences that activate them. They appear to be activated by the presence of formants. However there are also clips that do not appear to exhibit formants that activate these features as well.

not quite understandable to a person. Others can also use our techniques to understand how their models are making decisions or to gleem some insight into what may be going on.

This project could be extended in the future by training a larger model in order to get better performance over the data and seeing if that correlates to the model having more human recognizable features. Additionally the model could be presented with adversarial cases in order to understand what is going on.

6 Broader impacts

The model itself does not hold much value or have much impact. However, the methods we use to examine the model may be used to examine other models to determine how they may be working. People may attempt to use the model to try to detect people's gender through voice, however this is ill-advised there will always be pathological cases that occur that will confound the model. Thus it is important that this model be not made public as there is no productive use of this model. The model itself used significant resources to train as the dataset it needed was large and the model itself was large. It required multiple days of gpu compute time in order to train. It used a significant amount of energy to train and validate.

7 Code

`https://github.com/tzarii/CS152-Final-Project`

Language	Prop. of Female predictions
Thai	0.651
Arabic	0.632
Turkish	0.590
Swahili	0.571
Japanese	0.545
All	0.514
French	0.501
Rwandan	0.499
Italian	0.482
Spanish	0.468
German	0.452
English	0.439
Chinese	0.418

Figure 8: Proportion of Female Predictions per language. Some languages were perceived to be more feminine than others.

Language	Prop. of Female Predictions	Avg. Pitch (relative metric)
Thai	0.651	0.497
Arabic	0.632	0.460
Turkish	0.590	0.585
Swahili	0.571	0.438
Japanese	0.545	0.439
All	0.514	0.402
French	0.501	0.338
Rwandan	0.499	0.388
Italian	0.482	0.451
Spanish	0.468	0.366
German	0.452	0.316
English	0.439	0.364
Chinese	0.418	0.304

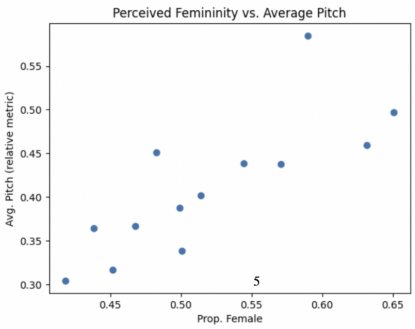

Figure 9: Perceiced Femininity vs. Average Pitch across different languages seem to exhibit another loosely positive correlation

References

[1] Jogin, M., Mohana, Madhulika, M. S., Divya, G. D. Meghana, R. K. & Apoorva, S. (2018) Feature Extraction using Convolution Neural Networks (CNN) and Deep Learning. *2018 3rd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)*, Bangalore, India, pp. 2319-2323.

[2] Oord, A.V.D., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A. & Kavukcuoglu, K. (2016) Wavenet: A generative model for raw audio. arXiv preprint arXiv:1609.03499.

[3] Pernet, C.R. & Berlin, P. (2012) The role of pitch and timbre in voice gender categorization. *Frontiers in Psychology* **3**