CS230 Winter 2018

Lily Chen {chlily,
Laura Shen lashen,
Meng Tang mengtang}
@stanford.edu

Accent Classification and Neural Accent Transfer of English Speech

INTRODUCTION

About 75% of the 1.5 billion English speakers in the world are English as a Foreign Language (EFL) speakers.[1] Many EFL speakers cannot effectively communicate due to heavily accented speech.

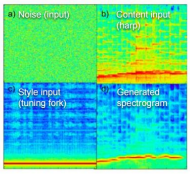
We aimed to apply neural style transfer to make accented English sound like native American English.

We first trained a convolutional neural network (CNN) on a binary classification task ("native" vs "non-native"). Classification performance on Indian accents vs native US/Canadian accents achieved 99.2% accuracy on training data.

The trained CNN network was then used in neural style transfer, with a non-native accent clip as the "content" using activations from earlier layers and a US accent clip as the "style" using activations from later layers. The CNN classifier was able to classify the generated spectrograms as "native" despite the generated audio sounding like white noise.

RELATED WORK

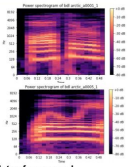
Neural style transfer,[2] often applied to images, combines "content" and "style" inputs to generate a new image. It has been applied to spectrograms of audio, e.g. to transform the timbre and bandwidth of musical instruments.[3]



Accent transfer using MFCC spectral features[4] and neural style transfer applied to voice transformation[5] have also been shown, both with limited success.

DATA

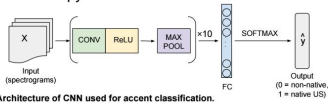
We used speech samples from CMU_Arctic,[6] specifically Indian accents (labeled 0) vs. native US and Canadian accents (labeled 1). Wav files were processed into spectrograms of non-overlapping 500 ms clips via STFT, yielding a total of 92k spectrograms.



Plots of processed spectrograms.

ACCENT CLASSIFICATION

Our accent classifier used a convolutional neural network (CNN) architecture with 10 convolutional layers followed by a fully connected layer, with a cross-entropy cost function.



Architecture of CNN used for accent classification.

NEURAL ACCENT TRANSFER

We applied the neural style transfer algorithm developed by Gatys et al.,[2] using the parameters learned during the training of our classifier CNN. The activations of the early layers represent the structure (content) of the spectrograms, while the deeper layers represent the style (accent).

To generate a spectrogram G of non-native speech transferred to a native accent, we minimized:[7]

$$J(G) = \alpha J_{content}(C,G) + \beta J_{style}(S,G)$$

where:

$$J_{content}(C,G) = \frac{1}{4 \times n_H \times n_W \times n_C} \sum_{\text{all entries}} (a^{(C)} - a^{(G)})^2$$

$$J^{[l]}_{style}(S,G) = \frac{1}{4 \times {n_C}^2 \times (n_H \times n_W)^2} \sum_{i=1}^{n_C} \sum_{j=1}^{n_C} (G^{(S)}_{ij} - G^{(G)}_{ij})^2$$

An Indian accented clip was used as the content input C, and a native US accented clip was used as the style input S. We initialized G as C with added white noise. The generated audio was reconstructed from the generated spectrogram G by using an inverse short-time Fourier transform (ISTFT).

RESULTS

Accent Classification Network

Our training set consisted of 82k examples, and our test and dev sets consisted of 5k examples each.

Performance improved when the CNN was deepened from 2, to 5, then finally to 10 convolutional layers.

After 20 training epochs, classifier performance was very good:

- Accuracy on train set: 99.2%
- Accuracy on test set: 97.8%
- Precision on test set: 98.1%
- Recall on test set: 98.3%

		Predicted:	
	True:	Non-N	Native
	Non-N	1945	59
	Native	51	2981

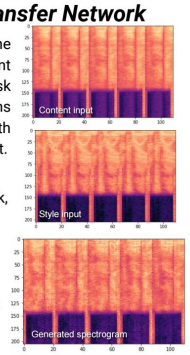
Confusion matrix on test set.

Training the model beyond 20 epochs was found to produce little or no improvement.

Neural Accent Transfer Network

Initial attempts using the CNN trained on our accent classification task generated spectrograms that appeared as noise with no visible shape or content.

We switched to using a pretrained VGG-19 network, generating spectrograms that were classified as "native" by our accent classifier. However, the audio sounds garbled with no discernible speech content.



DISCUSSION

- Classifier performance was very good, consistent with our subjective evaluation that humans can perform nearly perfectly on the classification task.
- Capturing the "content" and "style" of our data was not as straightforward as it often is with images.
- The classifier performed nearly as well with significantly fewer layers (e.g. 5 instead of 10), so perhaps later layers did not provide additional feature information, thus hampering style transfer.
- The generated spectrograms were classified as "native" by our original classifier, indicating that style transfer took place.
- Generated audio sounds garbled with no discernible speech sounds, perhaps partially due to the difficulty of reconstructing audio from spectrograms since phase information was lost.

FUTURE WORK

- Expand classifier and style transfer to multiple non-native accents, not just Indian accents.
- Experiment with other loss function terms to more accurately capture content and style features.
- Experiment with other NN architectures.
- Use a Generative Adversarial Network (GAN) to generate the accent-transferred clips.
- Use more sophisticated methods of recovering speech from spectrograms, e.g. phase recovery using the Griffin-Lim algorithm.

REFERENCES

[1] "How And Why Did English Supplant French As The World's Lingua Franca?" www.babbel.com/en/magazine/how-and-why-did-english-supplant-french-as-the-world-s-lingua-franca

[2] L. A. Gatys, A. S. Ecker, M. Bethge. "A Neural Algorithm of Artistic Style." arxiv.org/abs/1508.06576

[3] P. Verma, J. O. Smith. "Neural Style Transfer for Audio Spectograms." arxiv.org/abs/1801.01589

[4] A. Bearman, K. Josund, G. Fiore. "Accent Conversion Using Artificial Neural Networks." web.stanford.edu /class/cs224s/reports/Amy_Bearman.pdf

[5] D. Foote, D. Yang, M. Rohaninejad. "Do Androids Dream of Electric Beats?" audiostyletransfer.wordpress.com/

[6] Carnegie Mellon University. "CMU_ARCTIC", http://www.festvox.org/cmu_arctic/

[7] Coursera, Convolutional Neural Networks. "Deep Learning & Art: Neural Style Transfer." www.coursera.org/learn/convolutional-neural-networks

Acknowledgements:
Many thanks to our project mentor, Suraj Heereguppe, for invaluable advice and guidance!