Voice-Changing Detection with Convolutional Neural Network

ChunTung Zhuang
Basis International School of Guangzhou
z2750418749@gmail.com

Abstract

Voice-changing is a voice transformation technique that directly modifies a voice's pitch, tone, and timbre. Like other types of voice transformation, voice-changing is threatening our society from an economic, social, and legal perspective; hence in this study, we tackle this problem by adopting a convolutional neural network. As far as we know, our proposed work is the first solution to distinguish voice-changed voices and authentic voices. In addition, we also conducted experiments and gained significant insights that can help facilitate further research into voice-changing and audio-related machine learning. Foremost, we find the architecture with four convolutional layers to be the most effective. Second, we find a strong and positive correlation between the training data size and the performance in terms of accuracy and stability. Finally, we discovered that different training languages yield very different performances, with Chinese far outweighing English as a training language. We were inspired by the findings of different training languages, which motivated us to conduct a supplementary experiment: we concluded that tone is one factor that contributes to making a training language better in spotting voice-changed voices.

Keywords: *voice-changing, Convolutional Neural Network*

Contents

1 Introduction

Voice-changing techniques impose danger on many areas. Concerning telemarketing fraud. Eric Cole, the founder of Secure Anchor, claimed that he worked with 17 companies that have lost an average of 175,000$ apiece from voice cloning scams. In one case, the hackers gained access to a firm's IT systems [1]. When legal services collect evidence in the court, voice-changing becomes a potential distraction, increasing the difficulty of finding the criminal. In some countries, like the USA, Law enforcement agencies were already aware of the threat imposed by voice transformation technologies [2].

The primary goal of this paper is to propose a voice-changing detecting neural network architecture and investigate audio classification neural networks. In Sec. 2.1, we introduce the mechanism of voice-changing for a better understanding of the problem we intend to address. We chose convolutional neural network as the solution for this problem because it was proven to be successful in dealing with other audio tasks, and we provided a survey of the neural network's application specifically in audio in Sec. 2.3. Moving on to Sec. 3, we illustrate in detail the pre-training, architecture and training of our neural network. We tested different settings in the experiments in Sec. 4, including various neural network convolutional layers in the architecture, various data quantities in the training process, and different training languages. In the language experiment section, we extended our research further into finding the correlation between the tones (in Chinese), a change in pitch to contrast the meaning of the same syllables, and the performance of convolutional neural network.

2 Related Works

This section will first introduce mechanisms of voice transformation using a voice-changer to understand our research better. Then, we will survey the developments of neural networks in general. Finally, we will investigate the more relevant neural networks' progress in the field of audio application.

2.1 Mechanism of a voice-changer

A voice's essential components (pitch, timbre, and formant) can be modified through a voice-changer. In essence, our goal is to identify such modifications.

- **Pitch** adjusts the volatility rate of the input sound without affecting other attributes. A low-pitch sound corresponded to a bass singer. Conversely, a high-pitch sound created a sharper output like the cries of a baby.
- **Timbre** can be best understood as the tone of the sound, which defines the sound. Hence, even if two voices' pitch and volume are the same, they still differ to our ears because of timbre. In the voice-changer, timbre is a digitized value that becomes mellow and dark like the voice of an older man if the value is low and becomes bright and thin like the voice of a little girl if the value is high. Furthermore, the voice-changer we evaluated in this experiment provided more refined tuning, including HiS and LoS for high and low extension, Ls for low adjustment, and Sm for smoothness.

[1] https://fortune.com/2021/05/04/voice-cloning-fraud-ai-deepfakes-phone-scams/

[2] https://www.forbes.com/sites/forbestechcouncil/2021/05/10/analyzing-the-rise-of-deepfake-voice-technology/?sh=33149ecc6915

- **Formant** is the resonance frequency of the voice. Modifying formant can change the sense of naturalness for output.

2.2 Neural Network in general

The history of the neural network can be dated back to 1943 when McCulloch & Walter Pitts used electrical circuits to implement a simple neural network they proposed. About 45 years later, Stanford researchers Bernard Widrow and Marcian Hoff applied neural networks to a real-world problem to eliminate echos on phone lines for the first time [3]. Currently, neural networks' wide-range applications, including but not limited to classification, prediction, optimization, and association, are forever redefining our way of life. On the one hand, vehicle detection and pedestrian detection contribute to the emergence of autonomous driving, which is estimated to save 1.35 million lives per year [4]. On the other hand, fraud classifying neural networks were able to distinguish phishing emails that cause harm both economically and intellectually. In this paper, we build a neural network similar to fraud emails detector. We proposed a neural network detector that classifies voice-changed voices and natural voices.

2.3 Neural Network Applications in audio

To the best of our knowledge, we are the first to propose using convolutional neural networks to identify voice-changed voices. To better understand our contribution as an audio classification neural network, we examine other audio neural networks research, including noise-canceling, the transformation between voice and text, and detection of other types of voice transformation techniques.

2.3.1 Noise Canceling

Noise-canceling, referring to the removal of the noisy signals from the audio without causing significant distortion, is already applied to many areas, especially in our headphones and cell phones. Notice that, in many cases, a neural network is the backbone for smart noise suppression. For example, in RNNoise [1], deep learning was incorporated into classic signal processing to achieve noise canceling. In particular, RNNoise used a Gated Recurrent Unit technique, which contained two extra gates than common recurrent units that allow the network to memorize desirable information for a more extended period while also solving the vanishing gradient problem.

2.3.2 Speech and Text

- "Reading" the text written down that can be hardly distinguished from a natural human speaking can widely benefit different groups of people, such as those with visual impairments or people who are too busy. Google DeepMind's WaveNet [2] is one of many efforts that are reaching this goal. WaveNet is not the first one to approach Text to Speech function using a neural network. However, unlike precedents that use Recurrent Neural networks, WaveNet adopted Convolutional Neural Network with dilated causal convolutional layers, allowing the convolution to operate more coarsely, significantly increasing the receptive field.

[3] https://cs.stanford.edu/people/eroberts/courses/soco/projects/neural-networks/History/history1.html
[4] https://www.asirt.org/safe-travel/road-safety-facts/

- Relentless efforts have been poured into generating texts based on human voices, and we surveyed two state-of-art research: Fully Convolutional Speech Recognition [3]and ContextNet [4]. Fully Convolutional Speech Recognition introduces an end-to-end fully convolutional method to speech-to-text. The whole training process can be divided into training the raw waveform input to predict letters and using beam-searching decoding with a convolutional language model to predict sentences. ContexNet made significant improvements over existing convolutional neural networks for automatic speech recognition by introducing the squeeze and excitation layer, which successfully enhances the global context of the model.

2.3.3 DeepFake Detection and Voice Conversion

DeepFake includes all types of synthetic media, and with the growing AI technology that makes DeepFake hyper-realistic, many people become aware of the damage it might cause (i.e., impersonating voices for fraud or faking celebrities' pornography). Consequently, research on spotting DeepFake emerged. In terms of identifying DeepFake audio, which is highly relevant to our research direction, we surveyed DeepSonar in particular [5] which uses a binary-classifier to monitor neuron behaviors of a DNN-based Speaker Recognition model. Although researchers were already trying to detect AI-synthesized speech as early as in 2019 [6], DeepSonar not only acquired outstanding performance in classifying DeepFake, but it also examined the model's effect on finding voice-converted voices, including re-sampling, speed, and pitches. Our work can be seen as an alternative approach to extend this research by considering pitch, timbre, and formant changes using a voice-changer. The following chart explains the differences in various voice transformations [7]including voice conversion, DeepFake, and voice-changing.

	Voice Converting	DeepFake	**Voice-changing**
Description	Converting the source sound into another sound with a target speaker in mind.	Often DNN and GAN based, involving directly mapping linguistic features to acoustic feature representation	Directly modifying the components of a voice
Modification	Resampling, Speed, Pitch	Various	Formant, Pitch, Timbre

3 Our Method

To distinguish changed voices from authentic voices, we trained a convolutional neural network after pre-training or pre-processing the data. In this section, we discuss the implementation details and pipeline of our proposed network in the order of pre-training, architecture and training.

3.1 Pre-training Setup: Audio Processing (Pytorch-based)

- **Cutting duration** is a function to ensure that all training samples are shaped and consistent by cutting the signals through slicing if the actual number of samples is greater than the number of samples expected. This is important because the data sets we used are based on real-world audio files with various duration, and we want to avoid duration being the factor learned by the neural network.

- **Right Padding** is the inverse of the method cutting duration, which appends new items on the right side of the arrays of the signals if the number of samples is less than expected. We calculated the number of missing samples and applied it to the PyTorch defined function for padding.
- **Mixing down** the audio resolved the problems of having unwanted stereo or multiple channels. As the same suggested, the solution is to mix down all the channels into a mono channel unless the audio file is already a mono channel.
- **Resampling** ensured the coherency of the audio dimension by adjusting various sample rates into targeted sample rates.
- **Mel_Spectrogram** is the ideal audio feature for the neural network to learn and classify different audio files because it is an effective representation of perceptually relevant amplitude and frequency and the relationship between time and frequency. By applying the Transforms function defined by TorchAudio, we can extract Mel_Spectrogram from the original wav files.

3.2 Architecture of Neural Network

The primary motivation for us to choose a convolutional neural network to tackle this challenge is its ability to detect and classify without the need for human supervision and feature extraction, and the strength in identifying patterns, so we can let the machine decide ”what to learn.” In each convolutional neural network layer, a kernel of a set of weights and corresponding inputs will be applied to fulfill the linear operation. In this case, our input is a mel-spectrogram. Initially, we are using three layers to start with our architecture. Throughout the course of this paper, we examine other numbers of layers, including four layers, five layers, and six layers. Our architecture using four layers outperformed its counterparts, gaining 86.5% test set accuracy in the best training trial. Hence, in this section, we will present our architecture with four layers.

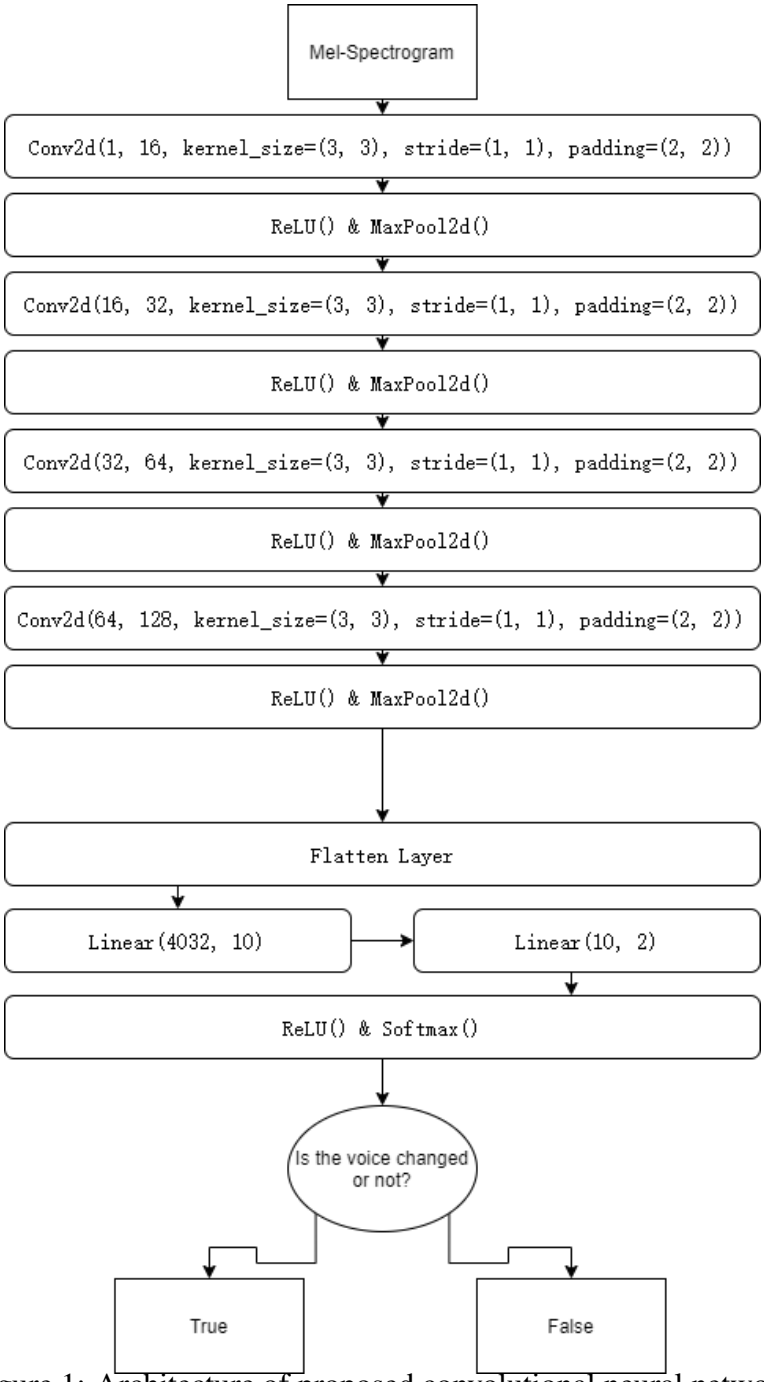

Figure 1: Architecture of proposed convolutional neural network

We used four sequential convolutional layers with ReLU, the activation function, and MaxPool2D that extract sharp and smooth features. Then we have a flatten layer to flatten the 2D input into 1D because our desired output is a binary classification task to determine whether the voice is changed or not. Following on, we contained two linear layers, which refer to a single layer of neuron, that learns an average rate of correlation between the input of 4032 with the output of 2. Finally, we applied the activation function ReLU again and the Softmax layer that assigns possibilities for each class label that enables binary classification.

3.3 Training

Finally, we started training our model. Initially, we began with 20 epochs. Soon, we realized 20 epochs were insufficient or under-fitting, so we gradually increased the epochs by five at a time, and we observed the training loss to continuously increase until 50 epochs when we reached 99.8% training set accuracy. Hence, we use 50 epochs in the training process to avoid under-fitting or over-fitting. Other training settings include a batch size of 128 and a learning rate of 0.001. For the optimizer that contributes to reducing the loss, we select Adam because we discovered that alternatives such as SGD do not produce satisfactory results. For the loss function that calculates the loss, we choose CrossEntrophyLoss because we are dealing with a binary classification task with a large decision boundary.

4 Experiment

In this section, we first explained data preparation. Then we walked through our experiment to evaluate the performance of our voice-changer detection neural network. We examined the effects of different convolutional layers and different quantities of data size. We compared the performance of different languages (English and Chinese) in this model.

The data sets in this experiment consist of original voices and voice-changed voices. Initially, we downloaded original voice clips from **Mozilla Common Voice data sets** [8], a well-established initiative focused on common voices of human speakers from different ages and regions in multiple languages. We trust the validity of this data set because two separate users were in place to validate the voice clips, which ensures the data to be natural and high-quality. We downloaded both Chinese and English for experiments from Mozilla Common Voice data sets, which became our original voices data set.

Then we acquired the voice-changing data set by passing a section of the original voice clips into a voice-changing software called **AV voice-changer Software Diamond** [5], a mature commercial product that followed the voice-changing mechanism discussed in Sec 2.1. In particular, there are pre-settings available in the software, including but not limited to "women aging 40", "women aging 20", and "men aging 20". These pre-settings offered realistic voice-changing from one sex to another that can hardly be distinguished. To clarify, we only use those pre-settings that can match the Mozilla Common Voice data sets. For example, we used the voice-changing pre-settings of "women aging 20", but not "female babies," because only the former appeared in the data sets. We believe this is an essential safeguard to prevent neural networks from memorizing unwanted patterns such as a particular age group that will only present in either voice-changed voices or natural voices, ensuring we evaluate the model correctly.

Table 1: Training and Testing data sets for the experiment

	#Clips	En&Ch
Training	1000, 800, 600, 400, 200	Both
Testing	200	Both

To further ensure fairness and to avoid random guessing of the neural network, we evenly distribute the data (in both training and testing data sets) into male and female original voices, male change into female voices, and female changed into male voices. For instance, in the 200 Chinese voice clips data set

[5] https://www.voicechangerdiamond.com/

we prepared, 100 voice clips were voice-changed. Among these 100 processed voice clips, 50 of them are male voices into female voices. For the simplicity of this work, we only considered cross-gender voice-changing, but we also recognize the importance of other types of voice-changing.

4.1 Network Size

We evaluated different network sizes, specifically testing convolutional layers of three to six layers, to find the optimal performance for our proposed neural network. We ran the 1000 Chinese training data sets for each architecture with a different number of layers and then tested them against the 200 Chinese testing data sets. We repeated this process for four trials, and we presented our data with box plots.

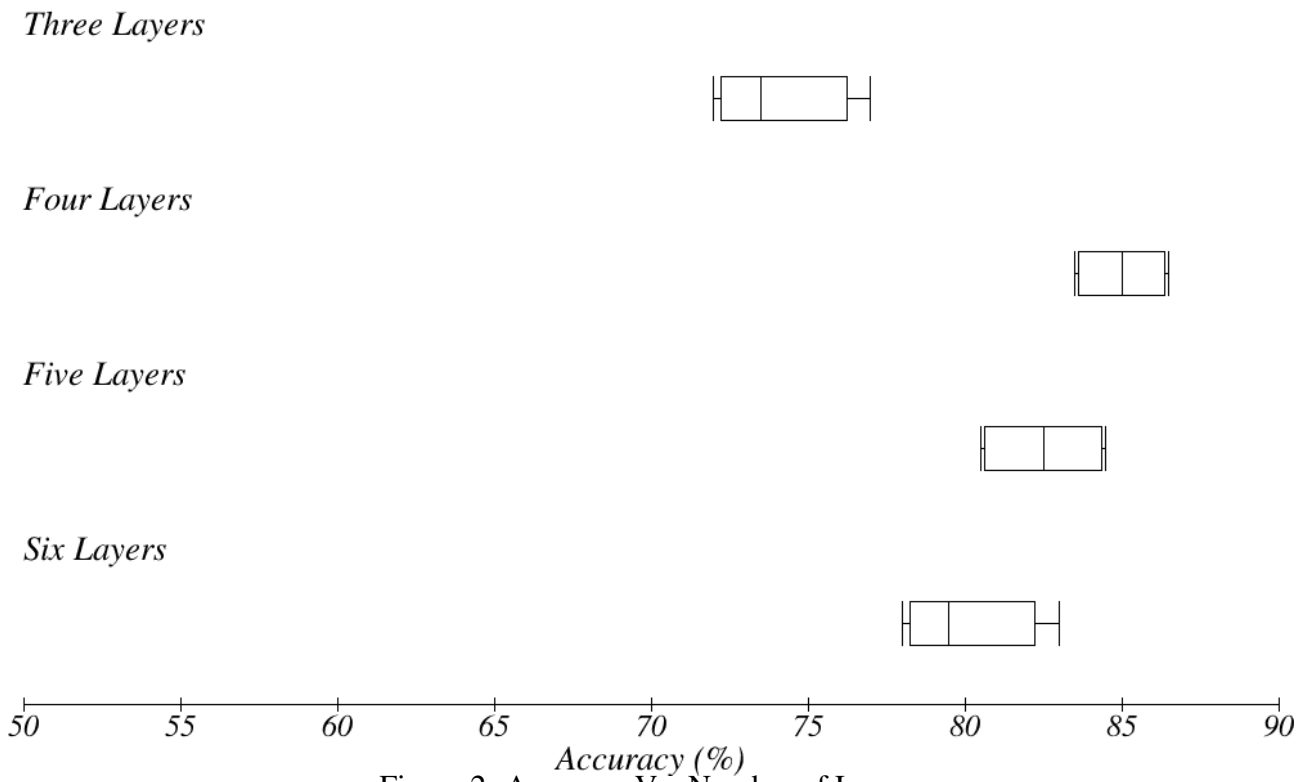

Figure 2: Accuracy Vs. Number of Layers

We found that the application of four layers reaches the optimal accuracy by a significant degree. We also discovered a pattern of less spread out data for four layers, followed by five layers (three layers and six layers are equally distributed, but six layers gained higher accuracy). This scenario can explain that four layers account for more stable performance and are more suited for this research purpose. Hence, we use four-layers architecture for the rest of our experiment.

4.2 Data Size

We wanted to determine whether data size affects the validation accuracy, so we compared different data sizes by running five separate trials on each Chinese training set with quantity of 200, 400, 600, 800 and 1000. Then we used box plots to present the four trials we ran on each quantity of data set. Initially, we

also tried to expand the data size to provide a further analysis by adopting data augmentation, but we discovered the training loss of the new data-augmented data set was unexpectedly high. We assumed this high training loss might be caused by speed interruption or inverse sound that deviated significantly from the rest of the data. Consequently, the results presented didn't include the trials with data augmentation.

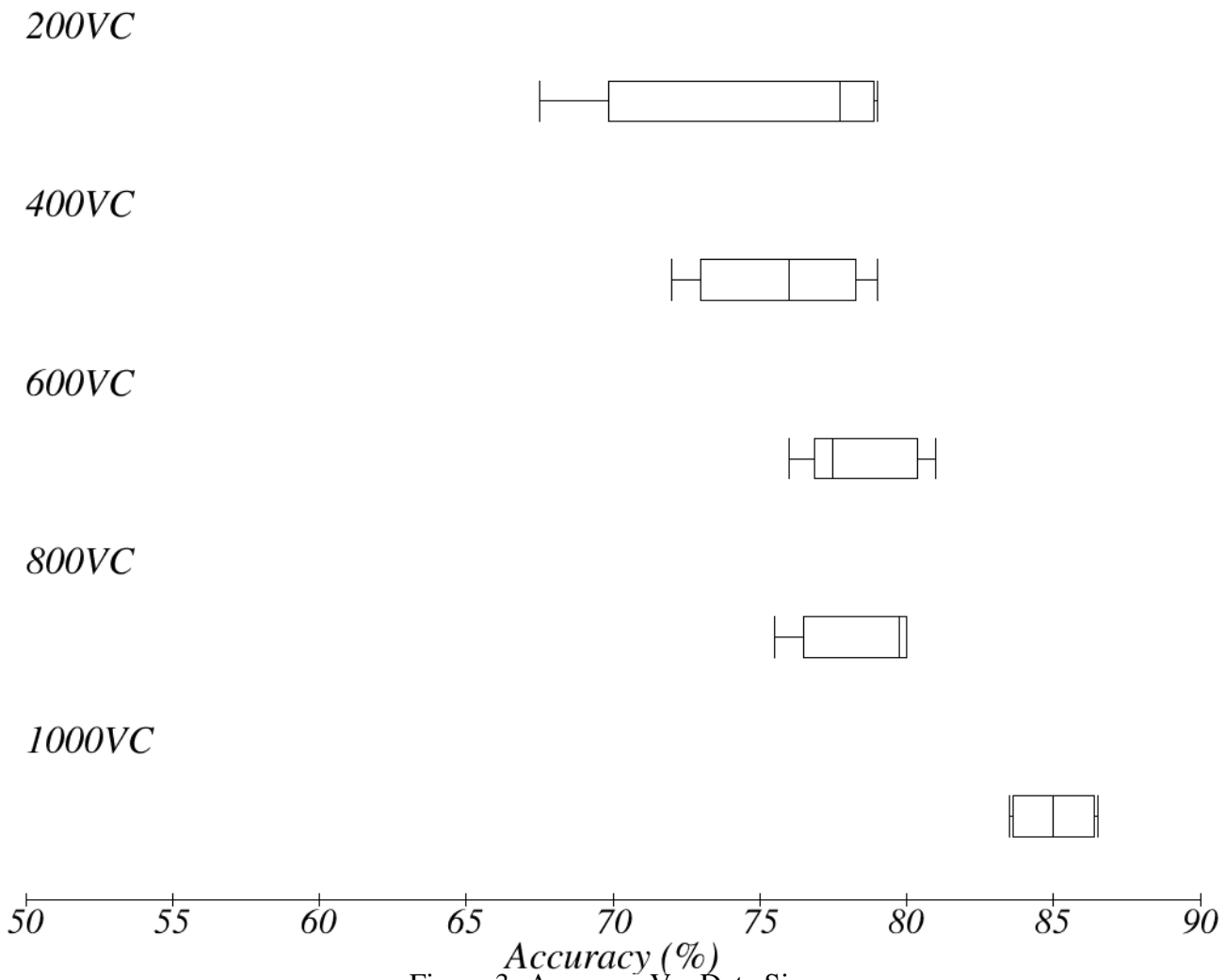

Figure 3: Accuracy Vs. Data Size

We found a growing trend of accuracy with respect to the quantity of training data. We acquired the best performance from the 1000 training data set, reaching 86.5% accuracy in the best trial and 85% on average, which is a considerable improvement compared to the 200 training data set with only 77% accuracy in the best trial and 75.5% in average. Furthermore, we also observed the results of the trials became more constant and less spread out as the data size increased, which indicated the training became more stable as data size expanded. Based on this result, we utilize 1000 clip data sets for the following experiments.

4.3 Language

We investigated whether different languages are a determining factor in the performance of the neural network. To do so, we fed 1000 Chinese voice clips to the neural network in four separate trials and average the validation accuracy. Then we repeated the same process for 1000 English voice clips. We presented our results in box plots.

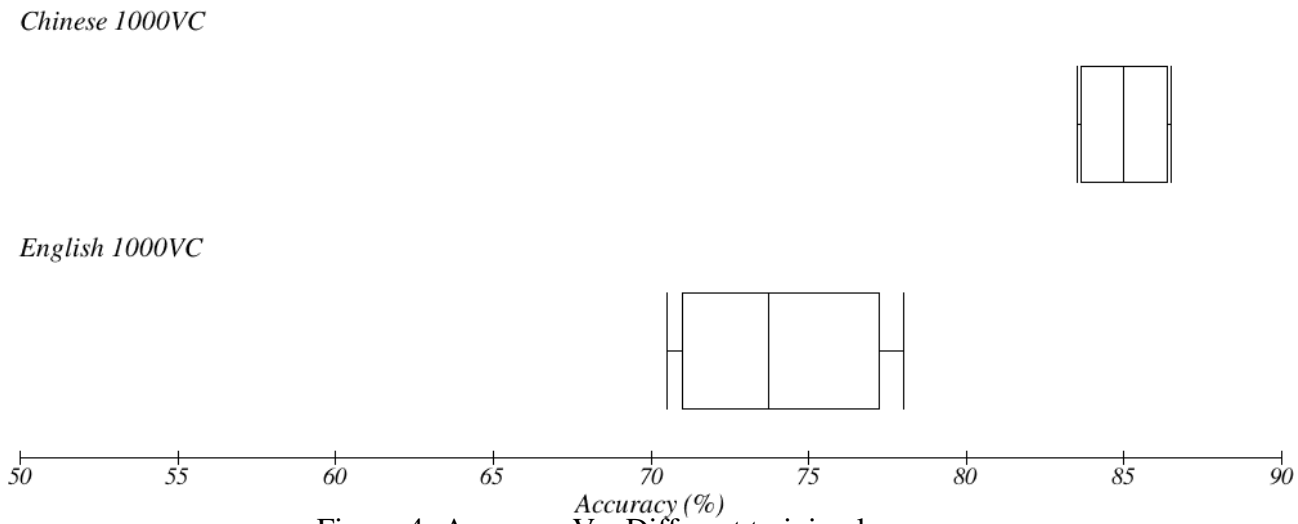

Figure 4: Accuracy Vs. Different training languages

We discovered that languages do make a difference in the neural network's ability to classify, and Chinese is a better indicator of whether the voices were changed or not. This result shares similarity in the experiment of DeepSonar [5], in which the accuracy of identifying Chinese AI-synthetic voice is much higher than English AI-synthetic voice. We believe this is because Chinese is a language with a more complex pronunciation system, with more features for the neural network to capture, like **tone** [9], [10], which refers to the changing pitch to contrast the meaning of the same syllables (For a better understanding of tone, please refer to figure 4). In this case, Chinese is a tonal language with four different tones, while English is a non-tonal language. Hence, we want to investigate whether tone is responsible for such difference, and we recognize the importance of the following experiment contributing to further understandings of machine learning audio tasks.

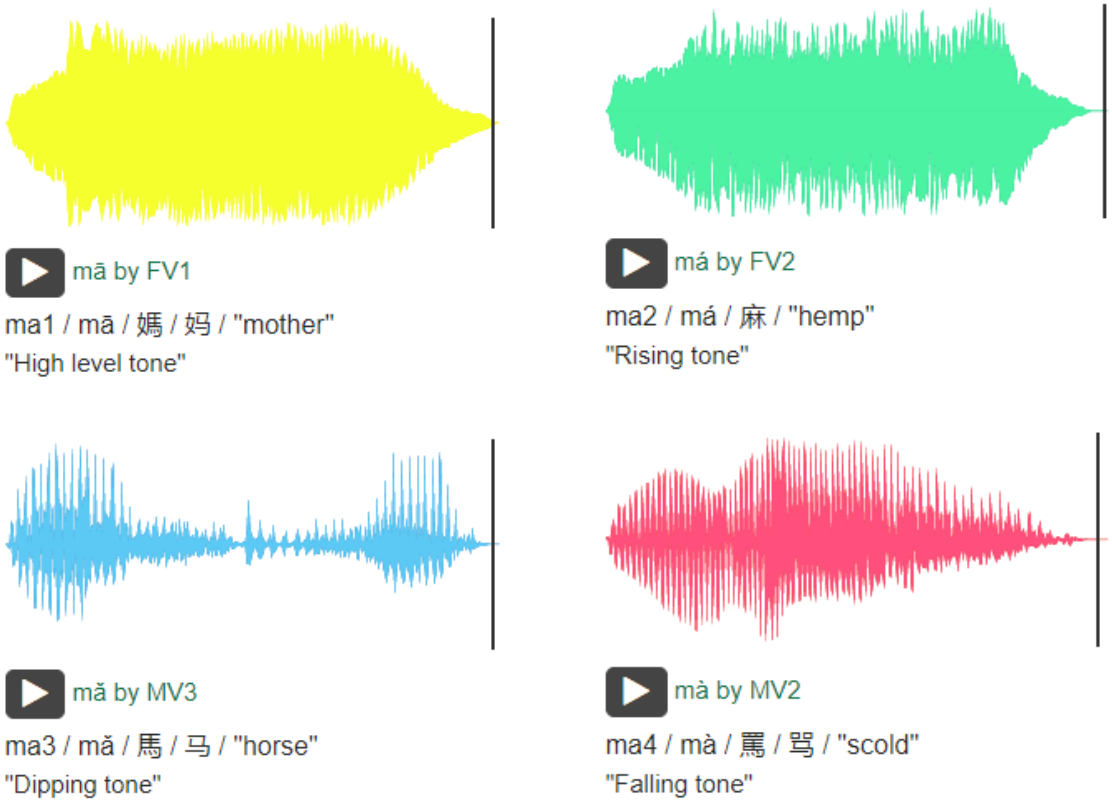

Figure 5: Four tones of Chinese, images provided by Michigan State University, https://tone.lib.msu.edu/

4.4 Supplementary Study - Chinese Tone's impact on machine learning

We evaluated whether tone, an important feature of many languages, impacts the machine learning model's performance and whether it accounts for the higher accuracy of Chinese compared with English. We primarily discuss the corresponding data sets and results of this experiment in sequence.

4.4.1 Data

Thanks to Michigan State University, we accessed the **Tone Perfect: Multimodal Database for Mandarin Chinese** data set for this experiment. We value this data set not only because it is the 2019 Open Scholarship Award and the 2018 Esperanto "Access to Language Education" Award winner, and it is widely embraced by many outstanding projects[6], but also because it is in line with our research direction. We quote from the official website that introduces the dataset: *"The Tone Perfect collection includes the full catalog of monosyllabic sounds in Mandarin Chinese (410 in total) in all four tones (410 x 4 = 1,640). Spoken by six native Mandarin speakers (three female and three male), the collection is comprised of 9,860 audio files (6 sets of 1,640)."* [11]

We then prepared four subsets of data from Tone Perfect.

[6]https://tone.lib.msu.edu/related

	#Clips	voice-changed Clips	#Tones
TrAT	1920	960	4
TeAT	240	120	4
TrOT	1920	960	1
TeOT	240	120	1

- **Training with All Tones (TrAT)** consists of 1920 voice clips with equal distribution in speakers (three male and three female) and tones (four types). Similar to the voice-changing technique we used in other parts of the experiment, we conducted cross-gender voice-changing on half of the monosyllabic sounds.
- **Testing with All Tones (TeAT)** consists of 240 voice clips with equal distribution in speakers (three male and three female) and tones (four types). Similar to the voice-changing technique we used in other parts of the experiment, we conducted cross-gender voice-changing on half of the monosyllabic sounds.
- **Training with One Tone (TrOT)** consists of 1920 voice clips with equal distribution in speakers (three male and three female), but it contains only the high-level tone or the first tone. Similar to the voice-changing technique we used in other parts of the experiment, we conducted cross-gender voice-changing on half of the monosyllabic sounds.
- **Testing with One Tone (TeOT)** consists of 240 voice clips with equal distribution in speakers (three male and three female), but it contains only the high-level tone or the first tone. Similar to the voice-changing technique we used in other parts of the experiment, we conducted cross-gender voice-changing on half of the monosyllabic sounds.

Please note that these monosyllabic sound clips are around or less than 1s, significantly lower than the sample size we choose, and it might not be the best data set for learning real-world voices, which usually is in sentences. However, we believe this will not interfere with our goal to determine whether tone impacts machine learning performance. In particular, we addressed this difference in desired sample size and actual sample size by using right padding (in Sec. 3.1) that will automatically adjust all voice clips to have the same sample size by appending new items on the right side of the arrays of the signals.

4.4.2 Result

We trained TrAT and TrOT with the same architecture of four convolutional layers, and then we evaluated the differences on the testing dataset TeAT and TeOT. We gathered the result from four trials, and we presented them with box plots. From the graph, we can see TrAT outperforming TrOT significantly, which also applies to the average of TrAT and TrOT: **67.81% Vs. 58.13%**.

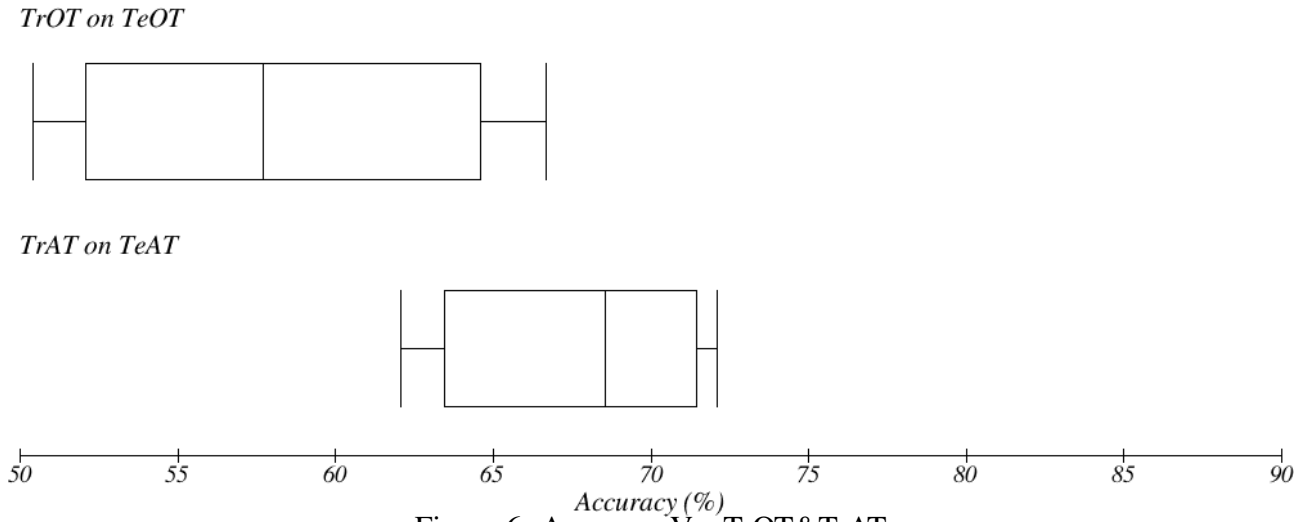

Figure 6: Accuracy Vs. TrOT&TrAT

Based on the results, we concluded that tone played an important role in machine learning to distinguish between voice-changed and natural voices. One explanation we come up with is that tone might be a complex variable to preserve during voice-changing, so the machine can easily capture this pattern, which can also be proven as the distribution of accuracy in the four trials in TrAT is more concentrated than that of TrOT. However, this conclusion doesn't mean machines can't replicate tone or other features in a sound; we believe other voice transformation techniques such as DeepFake and voice conversion are better at maintaining the tone of a voice because voice-changing is simply using pre-defined parameters to adjust digitized sounds regardless of the features of the source speakers.

5 Conclusion

We were motivated to research voice-changer because we saw an urgent need to prevent this tool from harming our society, from telemarketing fraud to distracting legal services. Hence, this paper proposed a convolutional neural network architecture to distinguish voice-changed voices from authentic voices and conducted experiments to draw several significant conclusions about the audio criteria of machine learning. Initially, we demonstrate the importance of having such a detector by referring to real-world problems like telemarketing fraud, hindering effective legal service, and online streaming fraud. Then, we conducted a throughout survey that revolved around voice-changing and machine learning, including the mechanism of voice-changing, the neural network in general, and how neural network was applied to audio-related tasks in specific. Subsequently, we clarify the specific implementation of our proposed architecture by discussing the pipeline: pre-training, architecture, and training. Consequently, we seized several significant conclusions through experiments in three fields. Firstly, we evaluated different network sizes' performance, and we discovered that the architecture with four layers acquired the best performance. Secondly, we investigated the relationship between training data size and the neural networks' performance, and we noticed a pattern indicating the accuracy in the validation increases as the data size increases. The results also illustrate a more stable output as data size expands. Lastly, we contrasted the neural networks' accuracy between English and Chinese as the training language, and we found that Chinese produces better machine learning performance, which inspired us to investigate

the reason behinds it. Hence, we conducted a supplementary experiment, and our hypothesis that tone played an important role in machine learning model's classification performance was confirmed. We believe our work will contribute not only to the audio criteria of neural network but also to our society as a whole because of this novel approach to address threats and challenges imposed by voice-changing.

6 Future Prospect

Admittedly, although we might be the first to tackle this challenge, we are superficial, and we firmly believe more could be done to investigate the topic of voice-changing and corresponding detection. We want to point to two limitations of the paper. Firstly, while many software and programs are available for voice-changing, we only examined and used one. Consequently, it might not be sufficient in generalizing our research on every voice-changing software. However, we believe the convolutional neural network architecture proposed, and the unique insights we learned from researching this voice-changer can be extended on other voice-changers because of their similar mechanisms and performance, which is worth further research. Secondly, we only evaluated cross-gender voice-changing, which means our study cannot be directly applied to other types of voice-changing like age changing (i.e., ten years old boy changed to 90 years old grandpa).

References

[1] Constantine C Doumanidis, Christina Anagnostou, Evangelia-Sofia Arvaniti, et al. "RNNoise-Ex: Hybrid Speech Enhancement System based on RNN and Spectral Features". In: arXiv preprint arXiv:2105.1181 (2021).

[2] Aaron van den Oord, Sander Dieleman, Heiga Zen, et al. "Wavenet: A generative model for raw audio". In: arXiv preprint arXiv:1609.03499 (2016).

[3] Neil Zeghidour, Qiantong Xu, Vitaliy Liptchinsky, et al. "Fully convolutional speech recognition". In: arXiv preprint arXiv:1812.06864 (2018).

[4] Wei Han, Zhengdong Zhang, Yu Zhang, et al. "Contextnet: Improving convolutional neural networks for automatic speech recognition with global context". In: arXiv preprint arXiv:2005.03191 (2020).

[5] Run Wang, Felix Juefei-Xu, Yihao Huang, et al. "Deepsonar: Towards effective and robust detection of ai-synthesized fake voices". In: Proceedings of the 28th ACM International Conference on Multimedia. 2020, pp. 1207–1216.

[6] Ehab A AlBadawy, Siwei Lyu, and Hany Farid. "Detecting AI-Synthesized Speech Using Bispectral Analysis." In: CVPR Workshops. 2019, pp. 104–109.

[7] Seyed Hamidreza Mohammadi and Alexander Kain. "An overview of voice conversion systems". In: Speech Communication 88 (2017), pp. 65–82.

[8] Rosana Ardila, Megan Branson, Kelly Davis, et al. "Common voice: A massively-multilingual speech corpus". In: arXiv preprint arXiv:1912.06670 (2019).

[9] Katherine M Crosswhite and Joyce McDonough. "Comparison of intonation patterns in Mandarin and English for a particular speaker". In: ().

[10] Gang Peng et al. “Temporal and tonal aspects of Chinese syllables: A corpus-based comparative study of Mandarin and Cantonese”. In: Journal of Chinese Linguistics 34.1 (2006), p. 134.

[11] Catherine Ryu, Mandarin Tone Perception amp; Production Team, and Michigan State University Libraries. Tone perfect: Multimodal database for Mandarin chinese. URL: https://tone.lib.msu.edu/.