# Deep learning for voice faker

CHANG,YU-JEN[1], CHIANG,YUNG-CHU[2], CHEN,CHIH-CHENG[3], *student, NYCU*
[1]Course of Introduction to Artificial Intelligence , Group 22 , School of Electrical and Computer Engineering

**In view of the significant increase in the application of artificial intelligence in various fields in recent years, understanding the basic concepts of deep learning and the implementation of programs has also become an important learning goal. Among the application fields of deep learning, the most important ones are Image Recognition and Natural Language Processing. Therefore, this study intends to carry out a simple program implementation for the latter field. By learning and reading multiple textbooks on the Internet, we finally use Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) to implement a simple speech recognition model. Word identification is also carried out through parameter tuning design and experiments, in order to develop a high-accuracy identification model.**

***Index Terms*—Convolutional Neural Networks, Classification, Deep learning, Deep voice, Fake voice, Feature map, Imitation,LaTeX.**

## I. Introduction

THIS project is interesting because deepfake is popular, and we can combine them together. Ideally , we can fake someone totally. When it comes to voice faker, it downs on us that we often use the APP for instant speech recognition and text conversion in class or while listening to speeches. We are interested in the recognition method of audio files such as human voice, so we would like to take the opportunity of the implementation of this topic to understand the logic more, and carry out some words.

First of all, we would like to construct simple identification model. This research will select a public speech data set—Common voice Dataset. To implement, we use the wavfile of Pytorch's torchaudio package through the data preprocessing process to visualize the frequency of the audio file into a picture, and then train it through a deep learning model , so that it can recognize some simple words. And we hope to improve the accuracy by adjusting the parameters. We expected to input the sound from one person and output the sound from another person but with the same semantics. If everything is good, we can input a string and choose which tones we want to use and make AI to fake our voices. It can help the lazy guys and the speaking handicapped to have a better life.

January 10, 2022

## II. Related work

### A. Audio to text technique

Speech recognition is a technology which a computer converts the speaker's pronunciation into text by comparing acoustic features. In the 1980s, research in this field was initiated by the laboratory of the Massachusetts Institute of Technology, but due to the low recognition rate, it has not been able to be applied for commercial purposes. It was not until 2012 that scientists replaced the traditional Gaussian distribution calculation with the deep neural network (DNN) calculation method, which greatly improved the recognition rate, and gradually attracted the attention of large international enterprises.

Manuscript received December 1, 2012; revised August 26, 2015. Corresponding author: M. Shell (email: http://www.michaelshell.org/contact.html).

### B. Applying Deep Learning Models to Speech Recognition

The main process of using deep network to achieve automatic speech recognition (ASR) is: input speech fragments (Spectrogram, MFCCs, etc.), convert the original language into acoustic features, and then go through the judgment and probability distribution of the neural network, and finally output the corresponding text content. The two neural networks used in this study are Convolutional Neural Network (CNN) and Recurrent Neural Networks (RNN) Long Short-Term Memory (LSTM). CNN is a convolutional neural network consisting of a convolutional layer, a fully connected layer, and a pooling layer. With the operation of the backpropagation algorithm, it can use the two-dimensional structure of the input data to extract features and properly converge and learn. All in all, it has been proved that CNN has excellent performance in image and speech recognition. RNN is a neural network with an active data memory called LSTM, which can be used for a series of data to guess what will happen next. Its output is not only related to the current input and the weight of the network, but also related to the input of the previous network . Moreover, it is often used to process time series data. It has been widely used in natural language understanding ,such as speech-to-text, translation, generation of handwritten text, image recognition and other fields.

### C. Text-To-Speech

Text-To-Speech is a technique that can let your devices read text to you. And is also an important part in dialogue robots.Text-To-Speech is a kind of speech sythesis, which can convert the text generated by the computer itself or input from the outside into fluent spoken language output that can be understood. It can be used not only to help people with visual or reading impairments,but also convenient for those who are lazy to talk. A typical speech synthesis technology component contains two parts.One is Natural Language Processing

(NLP) also known as linguistic formalisms , inference engines or logical inferences , which is responsible for generating intonation and rhythm. It can transfer the text input from others or generated by the computer into the narrow phonetic transcription. The output phones prosody from NLP on the left are sent to the digital signal processing module (DSP) on the right to convert the received symbols into natural sounds.

## III. Method design

### A. Our project

In the method design part, since we haven't been learning audio-related artificial intelligence topics before, we feel that it is a bit difficult to directly convert text to voice. Therefore, we implement classification and voice-to-text. The classification database mainly uses Urbansound8k, and the model mainly uses to converted the sound into MelSpectrom and the noise have also been removed. All in all, the classification trained model can achieve 70% accuracy. The voice-to-text part uses Libasound train-100 as the database. The model uses Greedy decoding and CTCloss and is modified with reference to deepspeech3. It is limited by the number of trains and our available resources, so the accuracy of each letter is The degree is about 70%, however, a sentence is composed of many letters, so it is easy to get a text that is not suitable for human reading.


### B. The state-of-art benchmark of our project datasets

#### 1) Urbansound8k — Classification

Accuracy from 70% to 79% and the highest model in : Justin Salamon and Juan Pablo Bello , Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound , 2016 ,Classification model accuracy 79%

#### 2) Libasound train-100 — Audio-to-text

Accuracy from 80% to 84% in Statista , 2020, Global speech-to-text transcript error rating.

### C. Our procedure

Step 1 Problem design (1/2) read four deep voice technical articles — know What is sound and how it is digitized. What problems is audio deep learning solving in our daily live. What are Spectrograms and why they are all-important.

Step 2 Problem design (2/2) read two related research papers — understand Why Mel Spectrograms perform better and how to generate them. The classification model we have built in Colab has about 70% accuracy rate.

Step 3 Method and experiment design (1/3) sound recognition — Enhance Spectrograms features for optimal performance by hyper-parameter tuning and data augmentation. Using MFCC (Mel Frequency Cepstral Coefficients) instead of Mel spectrogram.

Step 4 Method and experiment design (2/3) sound classification — Speech-to-Text algorithm and architecture, using CTC (Connectionist temporal classification) Loss and Decoding for aligning sequences.

Step 5 Method and experiment design (3/3) sound to text — Difficult Beam search Algorithm: commonly used by Speech-to-Text and NLP applications to enhance

## IV. Experiment result and analysis

Our Colab link : https://reurl.cc/MbbLbv

### A. Experiment result

From our colab, we've implement the classification part and work in a accuracy around 70% as average. The difficult to use this classification is that we don't add another classification to means the sounds from none of the city sounds, like human sounds. Thus, when we implement our speech to text part, we cannot use the classification part to improve the accuracy and loss. However, our final result still can reach around 70% accuracy for a single alphabet. This accuracy is low because the data processing not include the MelSpectrom and the other normalised processing. Therefore, if we try to test this model with different accent, we will find that the results are very terrible and not readable.

### B. Experiment analysis

To sum up, our project has many problems needed to alleviate, but we still can use it in some specific occasions. From this project, we can learn that normalisation is necessary part to level up the accuracy and enlarge the ability to accept more different test data.

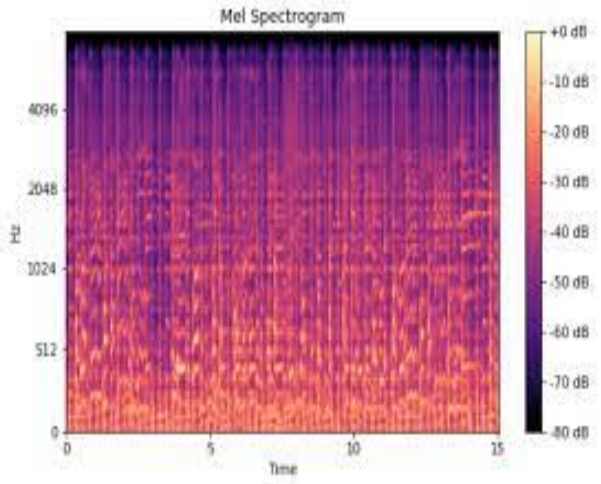
## V. Conclusion and Future Work

We refer to a paper as our future work on [Neural Voice Cloning with a few samples], which aims to study a technique for training Text-to-speech models with few samples, also known as voice cloning. The whole idea is very simple, that is, how to train a speech synthesis model with very few samples.The article uses the method of pretrained combined with finetone, and as a mature generator, deep voice3 has a trained multi-speaker model, we can use it directly, treat it as a black box, we only need to input text embedding and speaker embedding, that is to say, we only need a qualified timbre input, match the appropriate corresponding text, and throw it in to complete.Next is the part of voice cloning.,two methods are used, speaker adaptation and Speaker encoding. After various evaluations, both methods can achieve the task

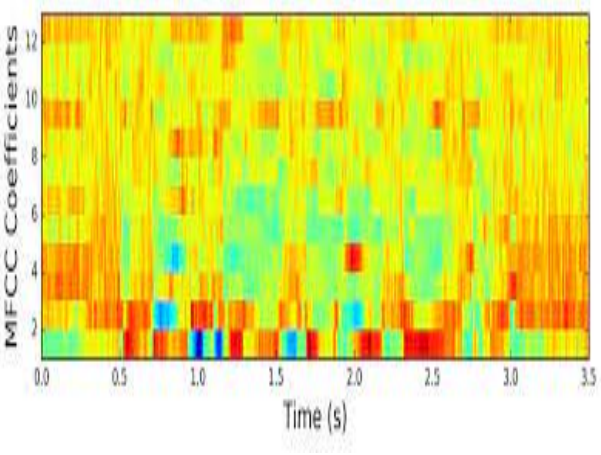
with a small amount of data and achieve good results. To sum up, for the Text-to-speech task, the generation network of a single speaker requires nearly 20 hours of training data, but when voice cloning an unseen new person speaks, it only takes a few minutes or a few seconds. We feel that this is very valuable and can be used as our future work, and we hope that our future research in this area can also move in this direction.

## APPENDIX A
## NOMENCLATURE

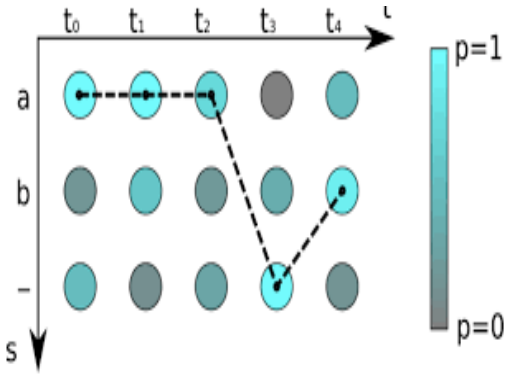
1. Mel Spectrograms




2. Mel Frequency Cepstral Coefficients




3. Connectionist temporal classification




4. Beam search Algorithm

**Algorithm1** Beam Search

**Input:** beam size $B$ ,input $x$ ,Parameters
**Output:** Approx. $B$ -best summaries
$\pi[0] \leftarrow \{\varepsilon\}$
$S = V \ if \ abstractive \ else \ \{x_i \mid \forall i\}$
**for** $i = 0 to N-1$ **do**
- Generate Hypotheses

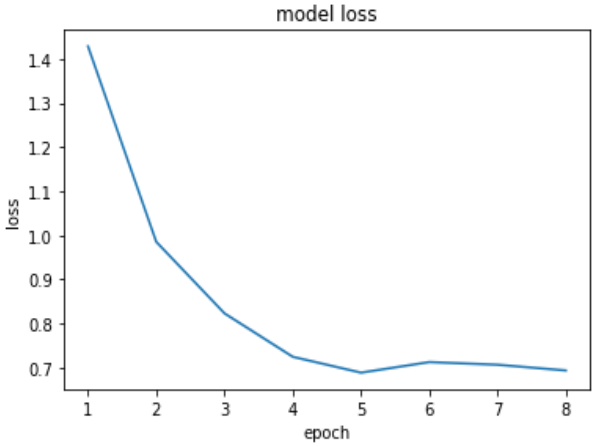
$N \leftarrow \{[y, y_{i+1}] \mid y \in \pi[i], y_{i+1} \in S$

- Hypotheses Recombination

$$H \leftarrow \begin{cases} y \in N \mid s(y,x) > s(y',x) \\ \forall y' \in N \ s.t. \ y_c = \end{cases}$$

Filter B-Max

5. Common voice dataset
https://commonvoice.mozilla.org/en/datasets
6. Our model performance




## ACKNOWLEDGMENT

The authors would like to thank Professor Wang for the impressive and wonderful courses during this semester. This paper is for the final project of Introduction to Artificial Intelligence and it's our pleasure to join the class. Thank a lot!

## REFERENCES

[1] Jitong Chen,Kainan Peng,and Wei Ping, *Neural Voice Cloning with a Few Samples*,in NIPS, 2018.
[2] Dora M.,Yohanna,Diego,Gonzalo, *Deep4SNet: deep learning for fake speech classification*, Sciencedirect Volume 184,ISSN 0957-4174,2021
[3] Lucioles, Sophia,, Antipolis, Cedex, *AFE characteristics proposed by the European Telecommunications Standards Institute*, ETSI Secretariat RES/STQ-00044, 2003
[4] Ketan Doshi, *Audio Deep Learning Made Simple: Sound Classification, Step-by-Step*,An end-to-end example and architecture for audio deep learning's foundational application scenario, 2021
[5] Jyh-Shing Roger Jang, *Audio Signal Processing and Recognition*,available at the links for on-line ressources at the author's homepage at http://mirlab.org/jang/books/audiosignalprocessing/., 2005