SPEECH TO SPEECH TRANSLATION

Prof. ShashidharHalligerimath, Pooja Ghodke, RichaliDesurkar, Supriya Kulkarni, Vandana Patel
Department of Computer Science and Engineering
KLE Dr. M. S. Sheshgiri College of Engineering and Technology
Belagavi, India

(E-mail: supriyamk41@gmail.com)

Abstract—In today's world, one important challenge is the need for effective ways of communication. This has led to the rise of machine translation. Research has shown that we can build systems that translate speech from one language to another. Speech-to-speech translation technology represents a technology which automatically translates one language to another language in order to enable communication between two parties with different native tongues [2]. The existing systems make use of an intermediate text representationwhich leads to the loss of important characteristics of the voice. The concepts of neural network and deep learning can be used to solve the above problem [1]. Thus we can translate audio input directly into spoken words without the need of text representation by making use of machine learning algorithms.

Keywords—*speech translation; machine learning; neural networks; svm*

I. INTRODUCTION

In the age of increasing globalization, there is a need for effective ways of communication. Much of human communication being spoken, the problem of spoken language translation must be addressed. One of the possible solution for this could be the human translators, however human translators being short in number, this is not an efficient way and also it expensive and slow. This has led to the rise of machine translation. Research has shown that we can build systems that translate speech from one language to another.

Speech translation was first demonstrated as a concept in 1983, by NEC Corporation at ITU Telecom World (Telecom'83) C-Star-2 consortium demonstrated speech-to-speech translation of 5 languages i.e. English, Japanese, Italian, Korean, and German in 1999. In 2015, a speech translator for 23 languages, Blabber Messenger was developed.An MIT Enterprise Technology Review has listed “Universal Translation” as one of top ten technologies in an article “10 Emerging Technologies That Will Change Your World”.

Research and development in the speech to speech translation field has been developing gradually. Examples of software systems which perform speech to speech translation are:Jibbigo: An application for mobile devices in which the users speak a sentence and then the app translates the sentence and outputs the translationthrough the speakers. It performs a two way translation. This system depends on an intermediate text representation. Google Translate: It is a translation system from Google which has a built-in functionality to take input from microphone and then output the translated audio through speaker. It can translate multiple forms of text and media, including text, speech, images, sites or real-time video from one language to another. It supports over 100 languages at various levels. Skype translator: It is a speech translation system developed by Skype. It is built on deep neural networks. It uses an acoustic model for mapping the source voice to a translated voice which is similar to source voice.

In the existing systems,a speech recognition system identifies the words spoken and transcode them into text followed by a text-to-text translation model which produces a text representation of the spoken words. The speech synthesizer then converts this text into speech. However when we transcode from the speech signal into a text representation we tend to lose important characteristics of the voice such as loudness,sharpness,frequency,emotions etc. that is needed to get a good dynamic voice. This problem can be overcome by applying the concepts of neural network and deep learning.By applying these concepts we have come up with an efficient tool called speech-to-speech translator. This tool takes speech from one language as its input and produces a speech of desired language as its output.It enables communication between speakers of different languages and is thus of tremendous value for humankind in terms of science,cross-cultural exchange and global business.

II. BACKGROUND

Arthur Samuel defined machine learning as – “Field of study that gives computers the capability to learn without being explicitly programmed”**.** The main aim is to allow the computers learn automatically without human intervention or assistance and adjust actions accordingly. The process involves feeding data and training the machines by building machine learning models using the data and different algorithms. The choice of algorithms depends on what type of data we have and what kind of task we are trying to automate. Machine learning enables analysis of large quantities of data. Though it gives faster and accurate results, it requires additional time and resources to train it.

A. Neural Networks

Neural networks are artificial information processing systems that are based on biological nervous system. It is consist of a large number of highly interconnected processing elements (neurons) which work together to solve specific problems. Learning in biological systems involves adjustments to the synaptic connections that exist between the neurons. The first artificial neuron was produced in 1943 by the Warren McCulloch and Walter Pits. Neural networks, like humans learn to perform tasks through various datasets and examples. The system generates identifying characteristics from this data with a pre-programmed understanding of these datasets. Neural networks process information in a similar way the human brain does. A neural network involves neurons, connections, weights, biases, propagation function, and a learning rule. Neurons will receive an input from predecessor neurons Connections consist of connections, weights and biaseswhich rules how neuron *i*transfers output to neuron *j*. Propagation computes the input and outputs the output and sums the predecessor neurons function with the weight. The learning rule modifies the weights and thresholds of the variables in the network. Neural networks are used in fields such as pattern recognition or data classification.

B. *Supervised Machine Learning*

Supervised machine learning is a learning in which we teach or train the machine by providing both input and desired output data. It provides the learning algorithms with known quantities to support future judgments to the learning systems. The input and output data are labelled to provide a learning basis for future data processing. . After that, the machine is provided with new set of data so that supervised learning algorithm analyses the training data and produces a correct output from labeled data. Supervised learning classified into two categories of algorithms: Classification: A classification problem is when the output variable is a category such as “yes” or “no". Regression: A regression problem is when the output variable is a real value, such as “dollars” or “weight”.

Supervised learning consists of input variables (x) and an output variable (Y) and an algorithm to learn the mapping function from the input to the output.

$Y = f(X)$

The goal is to approximate the mapping function so that the output variables (Y) for new input data (x) can be predicted.

C. *SVM*

Support Vector Machines are supervised learning methods based on the concept of decision planes that define decision boundaries used for classification and regression. A decision plane is one that separates between a set of objects having different class memberships. Support Vector Machines are used for tasks in which lines distinguishing objects of different class are drawn. These lines are called hyperplanes. Studies have shown that the SVM (support vector machines) generally are capable of delivering higher performance in terms of classification accuracy than the other data classification algorithms [3].

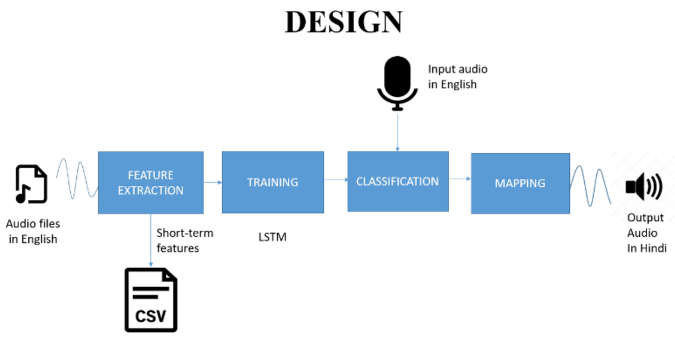
D. *MFCC*

The most popular feature extraction technique is the Mel Frequency Cepstral Coefficients called MFCC as it is less complex in implementation. They were introduced by Davis and Murmelstein in the 1980's.

Steps involved in MFCC are Pre-emphasis, Framing, Windowing, FFT, Mel filter bank, computing DCT [4].

III. METHODOLOGY

It involves the following steps:



A. *Dataset collection*

The dataset contains 1,350 audio files [5]. The training data comprises of digits from 1 to 9 in English. Translation is done from English to Hindi.

B. *Feature Extraction*

There are two stages in the audio feature extraction methodology:

- Short-term feature extraction: The input signal is split into short-term windows (frames) and a number of features for each frame are computed. This process leads to a sequence of short-term feature vectors for the whole signal.
- Mid-term feature extraction: The signal is represented by statistics on the extracted short-term feature sequences described above. A number of statistics (e.g. mean and standard deviation) over each short-term feature sequence are extracted.

The total number of short-term features implemented is 34.

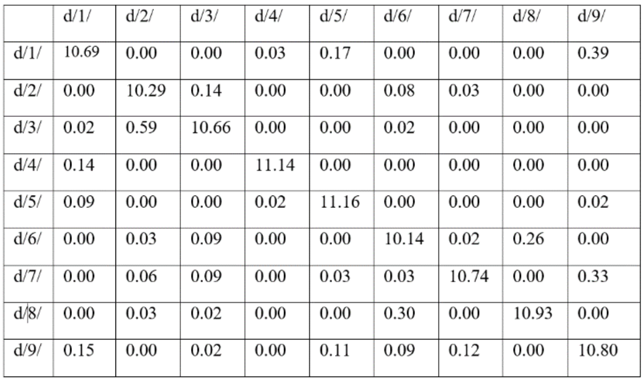
Some of the features extracted are: Zero Crossing Rate,Energy, Entropy of Energy, Spectral Centroid etc.

C. *Training and Classification*

A segment classification functionality is used in order to train and use supervised models that classify an unknown audio segment to a set of predefined classes (e.g. music and speech).

IV. RESULT

A. Confusion Matrix



	d/1/	d/2/	d/3/	d/4/	d/5/	d/6/	d/7/	d/8/	d/9/
d/1/	10.69	0.00	0.00	0.03	0.17	0.00	0.00	0.00	0.39
d/2/	0.00	10.29	0.14	0.00	0.00	0.08	0.03	0.00	0.00
d/3/	0.02	0.59	10.66	0.00	0.00	0.02	0.00	0.00	0.00
d/4/	0.14	0.00	0.00	11.14	0.00	0.00	0.00	0.00	0.00
d/5/	0.09	0.00	0.00	0.02	11.16	0.00	0.00	0.00	0.02
d/6/	0.00	0.03	0.09	0.00	0.00	10.14	0.02	0.26	0.00
d/7/	0.00	0.06	0.09	0.00	0.03	0.03	10.74	0.00	0.33
d/8/	0.00	0.03	0.02	0.00	0.00	0.30	0.00	10.93	0.00
d/9/	0.15	0.00	0.02	0.00	0.11	0.09	0.12	0.00	10.80

Columns represent the set of audio files that were predicted to be each label. Here, the first column represents thefiles that were predicted to be “one”, the second that were predicted to be “two”, the third "three", and so on.

Rows represent audio files by their correct truth labels. The first row represents the files that were “one”, the second that were “two”, the third "three", and so on.

Therefore, this matrix gives a summary of the performance of the network.

REFERENCES

1) *Fredrik Bredmar, Speech-to-speech translation using deep learning, 2017*

2) *J.Poornakala , A.Maheshwari,Automatic Speech-Speech Translation System from English to Tamil Language, 2016*

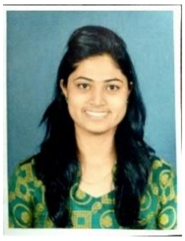
3) *DATA CLASSIFICATION USING SUPPORT VECTOR MACHINE, DURGESH K. SRIVASTAVA, LEKHA BHAMBHU*

4) *Parwinder Pal Singh, Pushpa Rani, An Approach to Extract Feature using MFCC,2014*

5) https://github.com/Jakobovski/free-spoken-digit-dataset

ABOUT AUTHORS

ShashidharHalligerimath, working as an assistant professor in CSE department at KLE Dr.MSSCET. Trained students on machine learning using python and pursuing Ph.D in the machine learning domain.



Pooja Ghodke
Pursuing 8th Semester
B. E(Computer Science and Engineering)
KLE Dr. M.S. Sheshgiri College of
Engineering and Technology, Belagavi

RichaliDesurkar
Pursuing 8th Semester
B. E(Computer Science and Engineering)
KLE Dr. M.S. Sheshgiri College of
Engineering and Technology, Belagavi

Supriya Kulkarni
Pursuing 8th Semester
B. E(Computer Science and Engineering)
KLE Dr. M.S. Sheshgiri College of
Engineering and Technology, Belagavi

Vandana Patel
Pursuing 8th Semester
B. E(Computer Science and Engineering)
KLE Dr. M.S. Sheshgiri College of
Engineering and Technology, Belagavi