Efficient Deep Learning Network Implementation for NLP based Voice Modulation using Mel Frequency Cepstral Coefficient (MFCC)

Ishita Aggarwal[1] Maneesh K Singh[2] Dr. Sandeep Sharma[3]
[1,2,3]Department of Electronics & Communication Engineering
[1,2,3]Dehradun Institute of Technology

Abstract— Neural networks and deep learning currently provide the best solutions to many problems in image recognition, speech recognition, and natural language processing. In this Paper we propose a simple technique train the neural network for speech modulation. The data sets used in the training have used the technique of Mel frequency cepstral coefficient (MFCC) to extract features from speech and map the differences in them between different speakers to generate a modulation vector. We aim that the host voice when modulated into target voice use this network to learn these modulation from large-scale unlabeled data. The network would modulate the speech without the modulation vector after a certain amount of time students.

Key words: Deep Neural Nets , Natural Language Processing , MFCC

I. INTRODUCTION

A. Deep Learning:

Deep learning (deep machine learning, or deep structured learning, or hierarchical learning, or sometimes DL) is a branch of machine learning based on a set of algorithms that attempt to model high-level abstractions in data by using model architectures, with complex structures or otherwise, composed of multiple non-linear transformations.

Deep learning is part of a broader family of machine learning methods based on learning representations of data. An observation (e.g., an image) can be represented in many ways such as a vector of intensity values per pixel, or in a more abstract way as a set of edges, regions of particular shape, etc... Some representations make it easier to learn tasks (e.g., face recognition or facial expression recognition) from examples. One of the promises of deep learning is replacing handcrafted features with efficient algorithms for unsupervised or semi-supervised feature learning and hierarchical feature extraction. [6]

Research in this area attempts to make better representations and create models to learn these representations from large-scale unlabeled data. Some of the representations are inspired by advances in neuroscience and are loosely based on interpretation of information processing and communication patterns in a nervous system, such as neural coding which attempts to define a relationship between the stimulus and the neuronal responses and the relationship among the electrical activity of the neurons in the brain.

Various deep learning architectures such as deep neural networks, convolutional deep neural networks, deep belief networks and recurrent neural networks have been applied to fields like computer vision, automatic speech recognition, natural language processing, audio recognition and bioinformatics where they have been shown to produce state-of-the-art results on various tasks.

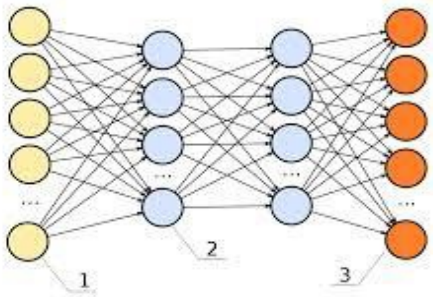

Fig. 1: Deep Learning

B. NLP:

Natural language processing (NLP) is a field of computer science, artificial intelligence, and computational linguistics concerned with the interactions between computers and human (natural) languages. As such, NLP is related to the area of human–computer interaction. Many challenges in NLP involve natural language understanding, that is, enabling computers to derive meaning from human or natural language input, and others involve natural language generation. Modern NLP algorithms are based on machine learning, especially statistical machine learning. The paradigm of machine learning is different from that of most prior attempts at language processing. Prior implementations of language-processing tasks typically involved the direct hand coding of large sets of rules. The machine-learning paradigm calls instead for using general learning algorithms — often, although not always, grounded in statistical inference — to automatically learn such rules through the analysis of large corpora of typical real-world examples. A corpus (plural, "corpora") is a set of documents (or

sometimes, individual sentences) that have been hand-annotated with the correct values to be learned. Many different classes of machine learning algorithms have been applied to NLP tasks. These algorithms take as input a large set of "features" that are generated from the input data. Some of the earliest-used algorithms, such as decision trees, produced systems of hard if-then rules similar to the systems of hand-written rules that were then common. Increasingly, however, research has focused on statistical models, which make soft, probabilistic decisions based on attaching real-valued weights to each input feature. Such models have the advantage that they can express the relative certainty of many different possible answers rather than only one, producing more reliable results when such a model is included as a component of a larger system.

C. MFCC:

In sound processing, the mel-frequency cepstrum (MFC) is a representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency. Mel-frequency cepstral coefficients (MFCCs) are coefficients that collectively make up an MFC. They are derived from a type of cepstral representation of the audio clip (a nonlinear "spectrum-of-a-spectrum"). The difference between the cepstrum and the mel-frequency cepstrum is that in the MFC, the frequency bands are equally spaced on the mel scale, which approximates the human auditory system's response more closely than the linearly-spaced frequency bands used in the normal cepstrum. This frequency warping can allow for better representation of sound, for example, in audio compression. MFCCs are commonly used as features in speech recognition systems, such as the systems which can automatically recognize numbers spoken into a telephone. They are also common in speaker recognition, which is the task of recognizing people from their voices.

MFCCs are also increasingly finding uses in music information retrieval applications such as genre classification, audio similarity measures, etc. MFCC values are not very robust in the presence of additive noise, and so it is common to normalize their values in speech recognition systems to lessen the influence of noise. Some researchers propose modifications to the basic MFCC algorithm to improve robustness, such as by raising the log-mel-amplitudes to a suitable power (around 2 or 3) before taking the DCT, which reduces the influence of low-energy components.

II. ANALYSIS OF MAPPING AND DATASET

The paper NLP based voice modulation using Mel frequency cepstral coefficient (MFCC) by Harsh Joshi and Harshita Agarwal discusses the mapping of speech. The overview of contents of mapping contains:-

A. PRAAT:

PRAAT (the Dutch word for "talk" or "speak") is a free scientific computer software package for the analysis of speech in phonetics. It was designed, and continues to be developed, by Paul Boersma and David Weenink of the University of Amsterdam

B. NLP Classifier:

Using the CMU pronunciation dictionary we create a list of 500 words which when used as a transcript for profile modelling of speaker, then cover maximum possible variation of phonemes. To do this we parse the dictionary set in our scripts and generate a lexical tree structure. Then modelling the knapsack algorithm on the sparse tree we form our list of final dictation transcript. The process is straight forward and can be used repeatedly after updating the parent dictionary.

C. Sample Isolation and Segmentation

The approach derives its influence; however when the classifier text is known then the reverse mapping of segmented phonemes saves further processing required in audio to phone text alignment, thereby making it a verbose model. 2.4. Dictionary Preparation In this stage we map the corresponding MFCC segments of audio signal to the phoneme classification list prepared for a linear time hash preparation as the dictionary is to be indexed.

III. TYPES OF FEATURES

In our experiment, we employ several different sets of features extracted from the audio content of the music, and compare them to each other. Specifically, we use the MARSYAS, Chroma, Rhythm Patterns, Statistical Spectrum Descriptors, and Rhythm Histograms audio feature sets

A. Temporal Features

The temporal features (time domain features), which are simple to extract and have easy physical interpretation, like: the energy of signal, zero crossing rate, maximum amplitude, minimum energy, etc.

B. Spectral Features

The spectral features (frequency based features), which are obtained by converting the time based signal into the frequency domain using the Fourier Transform, like: fundamental frequency, frequency components, spectral centroid, spectral flux, spectral density, spectral roll-off, etc. These features can be used to identify the notes, pitch, rhythm, and melody. 4. Convolution Network The modular design enables us to simply pass the perceptual map and audio data sets of the speakers into the convolution network and get the parse audio in output stream. Hardware bottleneck are overcome by the use of CUDA as offered by Theano and Torch framework. Both are compatible with the python which was primarily used for the development purposes in the prototype stage. 4.1 Classification In this segment we use the map to make a cluster that can be used for training process. Further, to reduce the error in the network we simply use 4.1.1 K-means clustering K-means clustering is an approach for vector quantization. In particular, given a set of n vectors, k-means clustering groups them into k

clusters (i.e., subsets) in such a way that each vector belongs to the cluster with the closest mean. The problem is computationally NP-hard, and suboptimal greedy algorithms have been developed for k-means clustering. In feature learning, k-means clustering can be used to group an unlabeled set of inputs into k clusters, and then use the centroids of these clusters to produce features. These features can be produced in several ways. The simplest way is to add k binary features to each sample, where each feature j has value one if the jth centroid learned by k-means is the closest to the sample under consideration. It is also possible to use the distances to the clusters as features, perhaps after transforming them through a radial basis function (a technique that has used to train RBF networks). Coates and Ng note that certain variants of k-means behave similarly to sparse coding algorithms. In a comparative evaluation of unsupervised feature learning methods, Coates, Lee and Ng found that k-means clustering with an appropriate transformation outperforms the more recently invented auto-encoders and RBMs on an image classification task. K-means has also been shown to improve performance in the domain of NLP, specifically fornamed-entity recognition; there, it competes with Brown clustering, as well as with distributed word representations (also known as neural word embedding). Here, we make use of the map here so the features of the target are matched with the host voice. For this purpose we used the python / MATLAB for training our network. The major problem is that we did not consider the RNN here. To minimize the error in the network we use the edit distance method or difference in the value of weights of the neurons and direction as the novel technique to improve the output of the system. And thus, the K-means clustering is used here to cluster different features in map and give the preferred output of target voice.

IV. CONCLUSIONS

Through this project we have explored the latest speech modulation methods which can be replaced regarding this method and a solution to the problem we faced is overcame by the above proposed model convolution network that is the heavy biasing towards one person's accent.

REFERENCES

[1] Graves, Alex. "Sequence transduction with recurrent neural networks." arXiv preprint arXiv: 1211.3711 (2012).
[2] Graves, Alex, A-R. Mohamed and Geoffrey Hinton. "Speech recognition with deep recurrent neural networks." Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on. IEEE, 2013.
[3] G. E. Hinton, et al. "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups". Signal Processing Magazine, IEEE, 29(6):8297, 2012.
[4] Deng, Li, et al. "Recent advances in deep learning for speech research at Microsoft." ICASSP, 2013.
[5] A. Krizhevsky, et al., "Image Net Classification with Deep Convolutional Neural Networks." NIPS, 2012.
[6] A. K. Halberstadt, "Heterogeneous acoustic measurements and multiple classifiers for speech recognition." Ph.D. Dissertation at MIT, 1998.
[7] Hannun, Awni, et al. "DeepSpeech: Scaling up end-to-end speech recognition." arXiv preprint arXiv: 1412.5567 (2014).
[8] Graves, Alex, and Jrgen Schmidhuber. "Framewise phoneme classification with bidirectional LSTM and other neural network architectures." Neural Networks 18.5 (2005): 602-610.
[9] Graves, Alex, et al. "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks." Proceedings of the 23rd international conference on Machine learning. ACM, 2006.
[10] Graves, Alex, Navdeep Jaitly, and A-R. Mohamed. "Hybrid speech recognition with deep bidirectional LSTM." Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on. IEEE, 2013.
[11] Bourlard, H., and N. Morgan. "Connectionist speech recognition: a hybrid approach. 1994."
[12] Jia, Yangqing, et al. "Caffe: Convolutional architecture for fast feature embedding." Proceedings of the ACM International Conference on Multimedia. ACM, 2014.
[13] Bengio, Yoshua. "Markovian Models for Sequential Data." (1996).
[14] O. Abdel-Hamid, et al., "Applying CNN Concepts to Hybrid NN-HMM model for speech recognition." IEEE ICASSP, 2012.
[15] T. N. Sainath, et al., "Deep Convolutional Neural Networks for LVCSR." IEEE ICASSP, 2013.
[16] Maas, et al. "Lexicon-Free Conversational Speech Recognition with Neural Networks." NAACL, 2015.
[17] Graves, Alex, and Navdeep Jaitly. "Towards end-to-end speech recognition with recurrent neural networks." Proceedings of the 31st International Conference on Machine Learning (ICML- 14). 2014
[18] Sutskever, Ilya, et al. "On the importance of initialization and momentum in deep learning." Proceedings of the 30th international conference on machine learning (ICML-13). 2013.