Progress Report 1
Capstone Project, JPMC 1 Team 1
Columbia University, Data Science Institute

Analyzing & Interpreting Wav2Vec & HuBERT Models for Keyword Spotting, Speaker Verification, Spoken Language Identification and Speech Emotion Recognition

Team Members:
Angad Nandwani - an3077
Eubin Park - ep3048
Mridul Gupta - mg4364
Siddhant Pravin Mahurkar - sm5129
Siddhant Rajeev Kumar - sk4975

Mentors:
Akshat Gupta, JP Morgan
Prof. Sining Chen

Table of Contents

Problem Definition

Recently there have been great improvements in our ability to process text owing to breakthroughs such as Word2Vec. It follows that the next challenge to tackle is our ability to process speech using recently published models such as Wav2Vec and HuBERT. In order to do this, we have focused our project on the four major tasks in the domain of audio, speech, language, and speaker recognition. These tasks are Speech Emotion Recognition, Language Identification, Keyword Spotting, and Speaker Verification.

For this project, we have restricted the scope to only pre-trained speech models - Wav2Vec & HuBERT. The objective of this project is to understand and study how well these models are suited for each of the aforementioned tasks. This would require model probing to understand how these models understand and work for different tasks, domains, and languages. Below is the descriptions for each of these tasks:

1. **Speech Emotion Recognition** - Speech Emotion Recognition (SER) is a fundamental task to predict the emotion label from speech data. Studies of automatic emotion recognition systems aim to create efficient, real-time methods of detecting the emotions of mobile phone users, call center operators and customers, car drivers, pilots, and many other human-machine communication users.

2. **Language Identification** - In natural language processing, language identification or language guessing is the problem of determining which natural language the given speech is in.

3. **Keyword Spotting -** To detect when a single word is spoken, from a set of ten or fewer target words, with as few false positives as possible from background noise or unrelated speech. With the emergence of Natural Language Processing (NLP), keyword extraction has evolved into being effective as well as efficient.

4. **Speaker Verification -** Speaker verification is verifying the identity of a person from characteristics of the voice (acoustic features of speech). Recognizing the speaker can simplify the task of translating speech in systems that have been trained on specific voices or it can be used to verify the identity of a speaker as part of security.

Introduction to Wav2Vec & HuBERT Models

1. Wav2Vec

Wav2Vec [1] is a speech encoder model released by the Facebook AI team in 2019. Currently, there have been 2 versions that have been released: Wav2Vec & Wav2Vec 2.0. Wav2Vec uses only a Convolutional Neural Network (CNN) while Wav2vec2.0 also uses a transformer architecture, enabling state of the art performance for many speech tasks such as Automatic Speech Recognition or Emotion Recognition. Wav2Vec essentially tries to predict the future of an audio sequence. To do so, the model is first pre-trained on a large network of unlabeled data to learn useful contextual representations of the audio sequence. These representations can then be used for those downstream tasks for which we do not have enough labeled data such as Language Identification, Speech Emotion Recognition etc.

However, the problem with this approach is that the high dimensionality of audio data makes modeling the waveform directly very challenging. To counter this issue, Wav2Vec first reduces the dimensionality of speech data by first encoding it into a latent space, and then predicting the future in this latent space. Therefore, we can see that Wav2vec comprises two distinct networks: **the encoder network & the context network**. Both are CNNs, but with different settings.

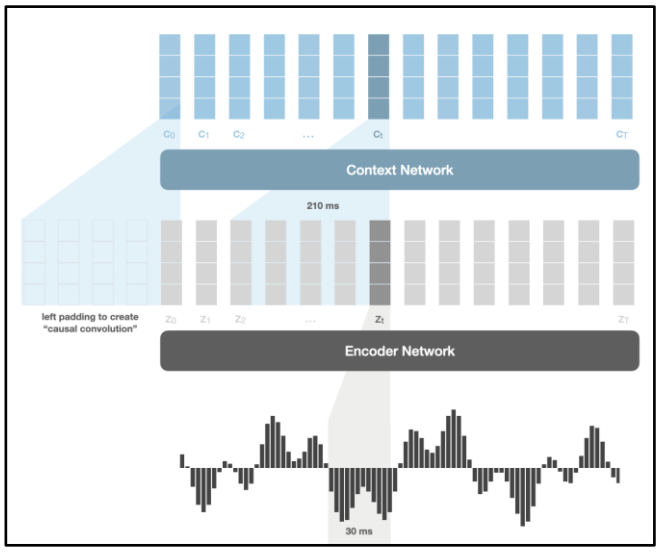

Fig 1 - The Encoder & Context Network of Wav2Vec (Boigne 2021) [1]

<u>Encoder Network</u>:
Given raw audio samples $x_i \in X$, we apply the encoder network $f : X \rightarrow Z$ parameterized as a five-layer convolutional network (kernel sizes (10, 8, 4, 4, 4) and strides (5, 4, 2, 2, 2)). The output of the encoder is a low frequency feature representation $z_i \in Z$ which encodes about 30 ms of 16 kHz

of audio. The striding results in representations z_i every 10ms. The output is a 512-dimensional feature vector z_t.

Context Network:

The output of the encoder network (i.e., multiple latent representations) are then fed into the context network which has a nine-layer CNN with kernel size 3 and stride 1. The objective is to aggregate information over a longer time frame to model higher-order information. This network outputs contextual representations c_t that are used to predict future audio samples.

An important detail is the causal nature of these convolutional networks; Wav2Vec should not refer to the future when predicting future samples. Hence, the convolutional layers are structured in such a way that each output at time t never attends to positions after t. In practice, this is done through left-padding the input (as shown in Fig. 1). These contextual representations can then be used for downstream tasks such as speech recognition.

2. Wav2vec 2.0

Wav2Vec 2.0 [2] brings the famous transformer-based neural network BERT architecture, which is currently the state-of-the-art for multiple natural language processing downstream tasks, to the speech processing domain. It has an architecture deriving from the encoder block of Transformer and shares the masked language prediction objective from BERT, tweaking it for speech related tasks. The architecture of Wav2Vec 2.0 comprises four major components, namely the feature encoder, the contextual network, the quantization module and the contrastive loss function module.

The model is trained in two phases. The first phase is in a self-supervised mode, which is done using unlabeled data and it aims to achieve the best speech representation possible. You can think about that in a similar way as you think of word embeddings. The second phase of training is supervised fine-tuning, during which labeled data is used to teach the model to predict particular words or phonemes.

3. HuBERT

HuBERT [3] is a new approach for learning self-supervised speech representations for modeling rich lexical and non-lexical information in audio. HuBERT also learns both acoustic and language models from continuous inputs. To achieve this, Hubert uses an offline k-means clustering step and learns the structure of spoken input by predicting the right cluster for masked audio segments. HuBERT draws inspiration from Facebook AI's DeepCluster method for self-supervised visual learning and Google's Bidirectional Encoder Representations from Transformers, or BERT to represent the sequential structure of speech.

The first step is to extract the hidden units (pseudo-targets) from the raw waveform of the audio. The K-means algorithm is used to assign each segment of audio (25 ms) into one of K clusters. Each identified cluster will then become a hidden unit, and all audio frames assigned to this cluster will be assigned with this unit label. Each hidden unit is then mapped to its corresponding embedding vector that can be used during the second step to make predictions.

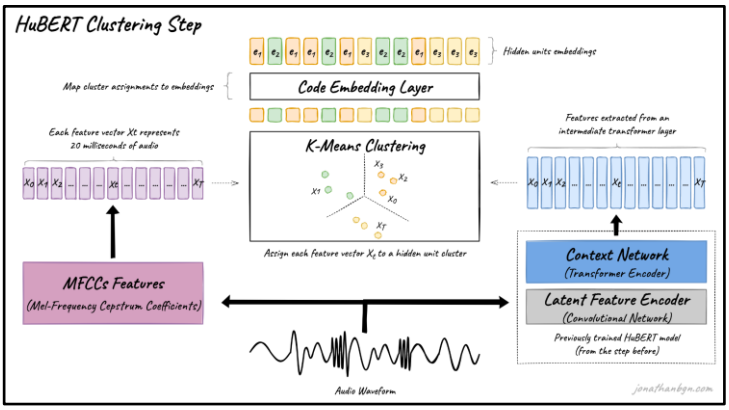

Figure 2. Clustering Step of HuBERT (Boigne 2021) [3]

The most important decision for clustering is into which features to transform the waveform for clustering. Mel-Frequency Cepstral Coefficients (MFCCs) are used for the first clustering step, as these features have been shown to be relatively efficient for speech processing. However, for subsequent clustering steps, representations from an intermediate layer of the HuBERT transformer encoder (from the previous iteration) are re-used.

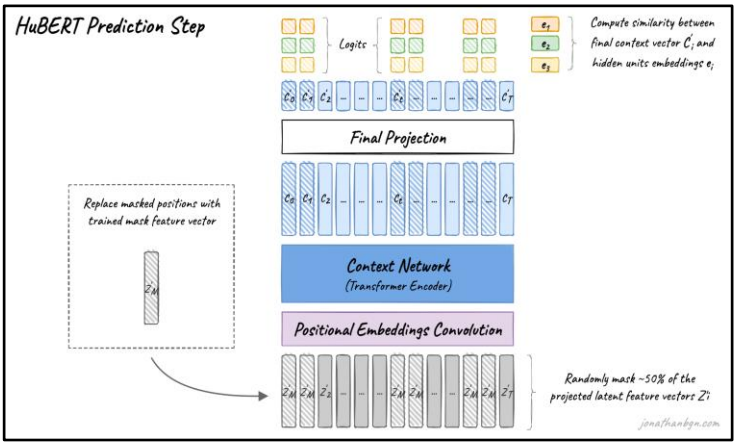

Figure 3. Prediction Step of HuBERT (Boigne 2021) [3]

The second step is the same as for the original BERT: training with the masked language modeling objective. Around 50% of transformer encoder input features are masked, and the model is asked to predict the targets for these positions. For this, the cosine similarity is computed between the transformer outputs (projected to a lower dimension) and each hidden unit embedding from all

possible hidden units to give prediction logits. The cross-entropy loss is then used to penalize wrong predictions. The loss is only applied to the masked positions as it has been shown to perform better when using noisy labels.

Literature Review

Over the last few years, Speech Recognition has gained importance and researchers have been trying to test different types of models over the popular known datasets. Our project, to evaluate Wave2Vec and Hubert models for the 4 different tasks (Speech Emotion Recognition, Language Identification, Keyword Spotting and Speaker Verification), needs a baseline for comparison and for the same, we did a literature review of existing models and their results. This section below highlights those research techniques and datasets over which results were obtained.

1. Wav2Vec 2.0 for Speaker Verification and Language Identification [4]

<table>
<tr><th>Dataset</th><th>Description</th></tr>
<tr><td>VoxCeleb1</td><td>100,000+ utterances from 1,251 celebrities.
<table>
<tr><th></th><th>Train</th><th>Validation</th><th>Test</th></tr>
<tr><td>Samples</td><td>1,43,642</td><td>5,000</td><td>4,874</td></tr>
<tr><td>Unique Speaker</td><td>1211</td><td>1145</td><td>40</td></tr>
</table></td></tr>
<tr><td rowspan="3">AP17-LR</td><td>10 different languages (Mandarin, Cantonese, Indonesian, Japanese, Russian, Korean, Vietnamese, Kazakh, Tibetan and Uyghur)</td></tr>
<tr><td>The duration of training data for each language is about 10 hours with the speech sampled at 16 kHz</td></tr>
<tr><td>The Test Set contains 2 subsets, 1 second audio and full-size audio. Each of them had 17,964 samples.</td></tr>
</table>

Table 1 - Description to dataset used for Speaker Verification and Language Identification

Results:

- For speaker verification, this paper obtained a new state-of-the-art result, where they achieved 3.61% Equal Error Rate (EER) on the VoxCeleb1 dataset.

- For language identification, they obtain an EER of 3.47% on full size condition and 12.02% on 1 second condition for the AP17-LR dataset.

2. Wav2Vec2.0 for Speech Emotion Recognition [5]

In this paper, they proposed a transfer learning method, where features extracted from pre-trained wav2vec 2.0 are modeled using simple neural networks.

<table>
<tr><th>Dataset</th><th>Description</th></tr>
<tr><td>RAVDESS
(Ryerson Audio-Visual Database of Emotional Speech and Song)</td><td>Features 24 different actors (12 males and 12 females) enacting 2 statements: “Kids are talking by the door” and “Dogs are sitting by the door.” with 8 different emotions: happy, sad, angry, fearful, surprise, disgust, calm, and neutral.

<table><tr><th></th><th>Train</th><th>Test</th><th>Validation</th></tr><tr><td>Actors Used</td><td>20 actors</td><td>20-22</td><td>22-24</td></tr></table>
These emotions are expressed in two different intensities: normal and strong</td></tr>
<tr><td rowspan="3">IEMOCAP
(Emotional Dyadic Motion Capture)</td><td>12 hours, improvised by 10 speakers.</td></tr>
<tr><td>5 sessions were composed, consisting 1 actor and 1 actress in each. 5-fold cross-validation was used for training & validation, leaving one session out for each of the folds.</td></tr>
<tr><td>4 emotional classes: anger, happiness, sadness and neutral,</td></tr>
</table>

Table 2 - Description to dataset used for Speech Emotion Recognition

Results:

Model	IEMOCAP (Average recall %)	RAVDESS (Average recall %)
Wav2vec2-PT (base-model pre-trained in LibriSpeech)	67.2 ± 0.7	84.3 ± 1.7
Wav2vec2-FT (fine-tuned using 960 hours of LibriSpeech2)	63.8 ± 0.3	68.7 ± 0.9

Table 3 -Model Results for Speech Emotion Recognition for Wav2Vec Models

3. Wav2KWS for Keyword Spotting [6]

This paper uses an encoder pretrained with large scale speech corpus as the backbone network and then design a transfer network for keyword spotting. It also demonstrates that the English speech corpus can be used for keyword spotting in other languages. This proposed network outperformed the state-of-the-art deep neural network for google speech command dataset.

<table>
<tr><th>Dataset</th><th>Description</th></tr>
<tr><td>Google Speech Command Dataset V2</td><td>Dataset V1 comprises 10 commands, namely, yes, no, up, down, left, right, on, off, stop, and go. Along with 10 commands from V1, V2 adds 10 commands corresponding to the digits from zero to nine, thus totaling 22 commands
<table>
<tr><th></th><th>Training</th><th>Validation</th><th>Test</th></tr>
<tr><td>Samples</td><td>36,923</td><td>4,445</td><td>490</td></tr>
</table>
</td></tr>
</table>

Table 4 -Description to dataset used for Keyword Spotting

Results:

- The proposed Wav2KWS model provides a 0.8% higher accuracy than the state-of-the-art DenseNet-BiLSTM model.
- Although the Wav2KWS model is more computationally expensive, it can produce a robust high performance.

Model	Accuracy %
LSTM	93.7
ATT-RNN	94.5
DenseNet - BiLSTM	96.6
Wav2KWS	97.8

Table 5 - Model Results for Keyword Spotting for Wav2KWS

4. Hubert benchmark for Speech Emotion, Speaker Verification [7]

Wav2Vec 2.0 and HuBERT have been making revolutionary progress in Speech Recognition systems. This paper along with Wav2Vec 2.0 results, also brings to light the 4 types of Hubert's model, which are described below:

Results:

Model	Speaker Verification (Equal Error Rate - EER) on VoxCeleb1	Speech Emotion with Speaker Dependent setting - weighted accuracy on IEMOCAP
EF-hbt-base	2.84	76.53
EF-hbt-large	2.86	78.52
PF-hbt-base	3.13	76.60
PF-hbt-large	3.21	79.58

EF- Entirely Fine Tuned, PF - Partially Tuned, Hbt - Hubert Model
base - 12 transformers blocks and 768 embedding dimensions, pre-trained on 960 hours of LibriSpeech Data
large - 24 transformer blocks and 1024 embedding dimensions, pre-trained on 60K hours Libri-Light data.

Table 6 - Model Results for Speech Emotion & Speaker Verification from HuBERT Model

Introduction of Datasets

1. **CREMA-D** [8]

The Crowd-sourced Emotional Multimodal Actors Dataset (CREMA-D) is an audio-visual dataset consisting of a total of 7,442 clips. The data was collected from 91 actors, each of whom were recorded expressing one of the 'universal' emotions: happy, sad, anger, fear, disgust, and neutral. Each emotion was displayed at one of the four emotion levels: low, medium, high, and unspecified. The clips were labeled through crowd-sourcing with 2,443 raters. For this project, we will be using the audio clips from this dataset as input for the *Speech Emotion Recognition* task.

2. **IEMOCAP** [9]

The Interactive Emotional Dyadic Motion Capture (IEMOCAP) database is a dataset recorded at the SAIL Lab at USC. It contains audio, visual, and motion capture data of 10 actors in 5 dyadic sessions. Each session contains 3 scripted scenarios and 7 or 8 improvised scenarios. The total corpus contains approximately 12 hours of data, which has been segmented into 10,039 utterances. Each utterance has a label of one of the following emotions: neutral, happiness, sadness, anger, surprise, fear, disgust, frustration, excited, and other. For this project, we will be using the IEMOCAP dataset for the *Speech Emotion Recognition* task.

3. **Speech Commands** [10]

The Speech Commands dataset is a dataset collected by the Google Brain team, often used for deep learning tasks concerning keyword or command detection. It contains more than 65,000 utterances, each utterance being 1 second long and containing a single spoken word from a list of 35 keywords. This dataset will be used for the *Keyword Spotting* task of our project.

Preliminary Results: Baseline Models

1. Speech Emotion Recognition

We Fine-tuned the Hugging face implementation of Facebook hubert-base-ls960 model on crema-d [8] by unfreezing last 4 layers including the classification head and its biases and weights. The number of training, validation and test samples are 6027, 335 and 335. For training, we ran 5 epochs each with training batch size of 8, validation batch size of 8, gradient accumulation steps as 4, loss function as cross entropy loss and learning rate 3x10-5. We achieved an accuracy of 0.64 on the test data.

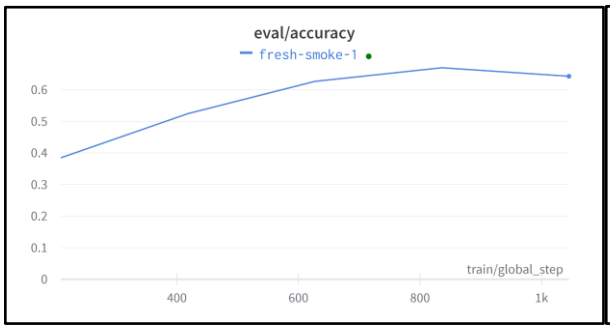

Figure 4. Evaluation Steps Vs. Accuracy

Figure 5. Training Steps Vs. Loss Value

2. Keyword Spotting

We fine-tuned the Hugging face implementation of Facebook wav2vec2-base model on Speech commands dataset [10]. The number of training, validation and test samples are 26000, 2000 and 2000. For training, we ran 3 epochs each with training batch size of 8, validation batch size of 8, gradient accumulation steps as 4, loss function as cross entropy loss and learning rate 3x10-5. We achieved an accuracy of 0.84 on the test data.

3. Next Steps

- Over the course of upcoming weeks, we aim to implement Wav2Vec 2.0 and HuBERT models for Language Identification on VoxLingua107 [11] and Speaker Verification on VoxCeleb [12] dataset.
- For keyword spotting due to computational constraints, we were able to fine-tune the model for a fewer batch size and fewer epochs with reduced data size. We plan to train the models on higher computational resources and improve upon the baseline results obtained currently.
- Further, building upon the models developed we will move towards probing Wav2Vec and HuBERT models for the different tasks and assess how specific transformer layers are learning to perform the tasks.

VoxLingua107 is a dataset popularly used for spoken language recognition. The data was automatically collected by scraping YouTube videos found using search phrases specific to certain languages. In total, the corpus contains around 6,628 hours of data, with 107 languages represented. On average, there are approximately 62 hours of data per language.

VoxCeleb is a large-scale speaker identification dataset comprising over 100,000 utterances for 1,251 celebrities. The data was retrieved using YouTube videos containing interview clips and then running these clips through CNNs for speaker verification and facial recognition. The chosen celebrities represent diverse backgrounds in order to account for variations in speech.

References

[1] Boigne, Jonathan. “The Illustrated wav2vec.” *Jonathan Bgn*, 29 June 2021, https://jonathanbgn.com/2021/06/29/illustrated-wav2vec.html.

[2] Boigne, Jonathan. “An Illustrated Tour of wav2vec 2.0.” *Jonathan Bgn*, 30 Sept. 2021, https://jonathanbgn.com/2021/09/30/illustrated-wav2vec-2.html.

[3] Boigne, Jonathan. “Hubert: How to Apply Bert to Speech, Visually Explained.” *Jonathan Bgn*, 30 Oct. 2021, https://jonathanbgn.com/2021/10/30/hubert-visually-explained.html.

[4] Fan, Zhiyun, Meng Li, Shiyu Zhou, and Bo Xu. "Exploring wav2vec 2.0 on speaker verification and language identification." *arXiv preprint arXiv:2012.06185* (2020).

[5] Chen, Li-Wei, and Alexander Rudnicky. "Exploring Wav2vec 2.0 fine-tuning for improved speech emotion recognition." *arXiv preprint arXiv:2110.06309* (2021)

[6] Seo, Deokjin, Heung-Seon Oh, and Yuchul Jung. "Wav2kws: Transfer learning from speech representations for keyword spotting." *IEEE Access* 9 (2021): 80682-80691.

[7] Wang, Yingzhi, Abdelmoumene Boumadane, and Abdelwahab Heba. "A fine-tuned wav2vec 2.0/hubert benchmark for speech emotion recognition, speaker verification and spoken language understanding." *arXiv preprint arXiv:2111.02735* (2021).

[8] Cao H, Cooper DG, Keutmann MK, Gur RC, Nenkova A, Verma R. CREMA-D: Crowd-sourced Emotional Multimodal Actors Dataset. IEEE Trans Affect Comput. 2014 Oct-Dec;5(4):377-390. doi: 10.1109/TAFFC.2014.2336244. PMID: 25653738; PMCID: PMC4313618.

[9] Busso, Carlos, et al. “IEMOCAP: Interactive Emotional Dyadic Motion Capture Database.” *Language Resources and Evaluation*, vol. 42, no. 4, 2008, pp. 335–359., https://doi.org/10.1007/s10579-008-9076-6.

[10] Warden, Pete. “Launching the Speech Commands Dataset.” *Google Research*, https://ai.googleblog.com/2017/08/launching-speech-commands-dataset.html.

[11] Valk, Jorgen, and Tanel Alumae. “Voxlingua107: A Dataset for Spoken Language Recognition.” *2021 IEEE Spoken Language Technology Workshop (SLT)*, 2021, https://doi.org/10.1109/slt48900.2021.9383459.

[12] Nagrani, Arsha, et al. “Voxceleb: A Large-Scale Speaker Identification Dataset.” *Interspeech 2017*, 2017, https://doi.org/10.21437/interspeech.2017-950.

Appendix

Member Contributions:

- Angad Nandwani: Literature and Datasets Review
- Eubin Park: Literature and Datasets Review
- Mridul Gupta: Problem Definition, Wav2Vec Explanation
- Siddhant Pravin Mahurkar: Preliminary Results: Baseline Models
- Siddhant Rajeev Kumar: Wav2Vec 2.0, HuBERT Explanations

IEMOCAP Dataset Distribution

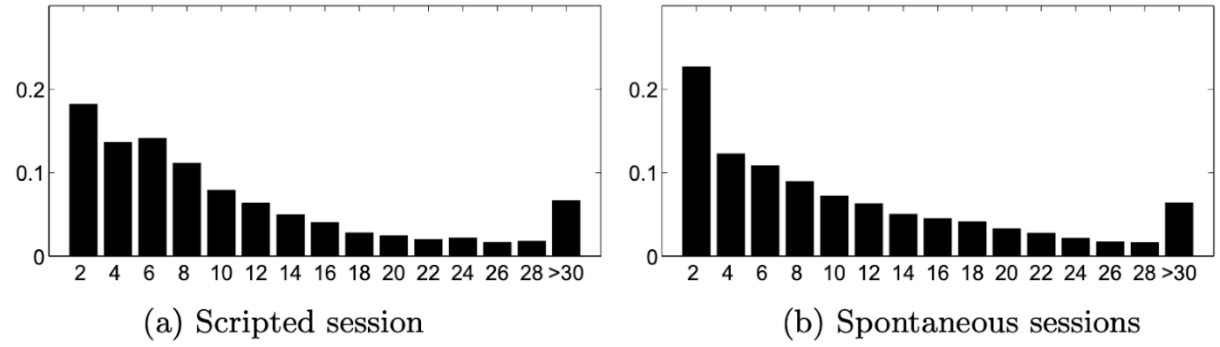

Figure 6. Distribution of the number of words per utterance (as a percentage) for the scripted (a) and improvised (b) sessions (Busso 2008) [9]

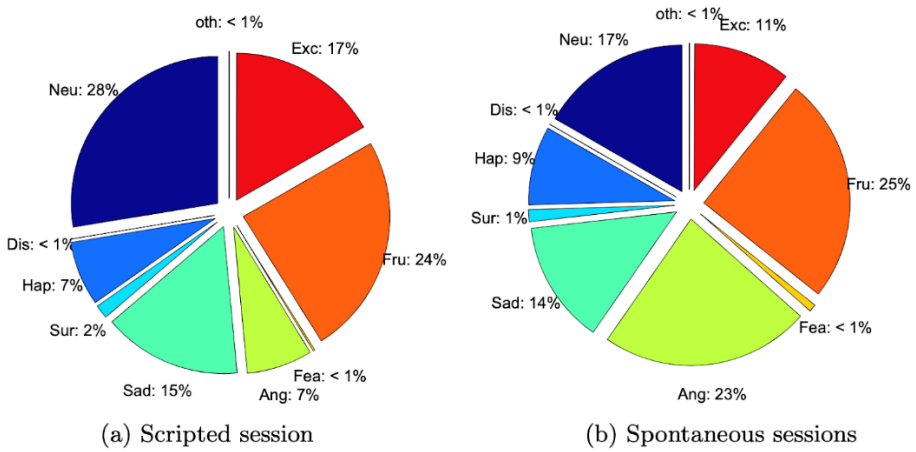

Figure 7. Distribution of the data over emotional categories for scripted (a) sessions and improvised (b) sessions (Busso 2008) [9]

Speech Commands Dataset Distribution

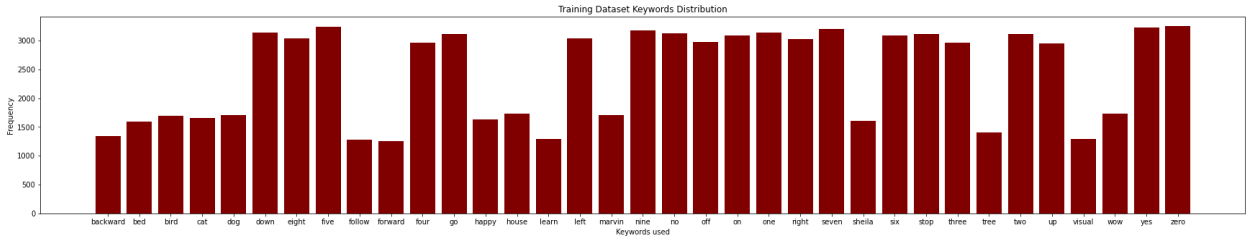

Figure 8. Training Dataset Labels distribution

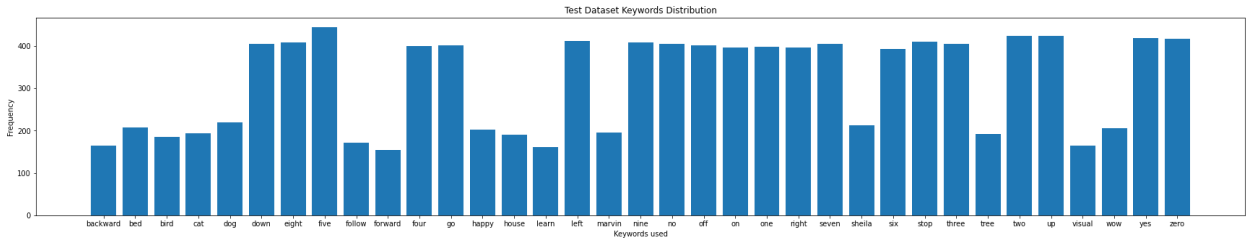

Figure 9. Test Dataset Labels distribution

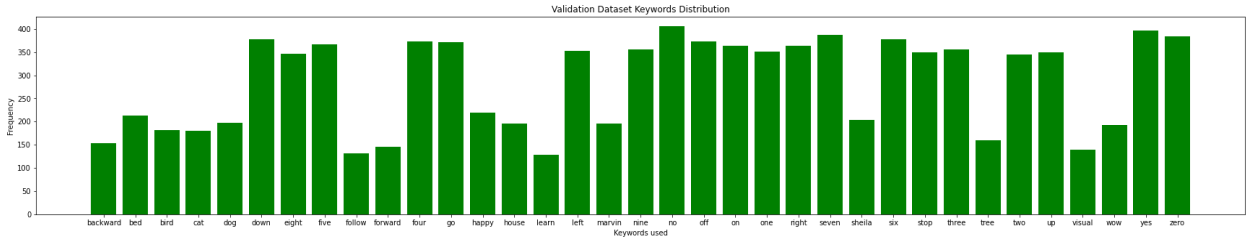

Figure 10. Validation Dataset Labels distribution