Automatic Lip-Reading (ALR) with Machine Learning: A Review of Approaches, Challenges, and AV-HuBERT Model Implementation

Kimberly Villegas
Master of Data Science
Rice University
Houston, Tx

I. Introduction

Speech is used everyday to communicate ideas and thoughts, yet most people rarely consider the instinctual nature of understanding visual and acoustic cues for speech perception. Most people, whether consciously or not, leverage visual cues to varying degrees depending on acoustic conditions as well as hearing abilities.

Recent supervised learning models have been able to achieve word-error-rates (WERs) of 1.4-to-2.6% for automated speech recognition (ASR) using solely audio-based input in noisy and quiet environments. Additionally, there has been much research done on the combination of audio- and video-based inputs for speech recognition (AV-ASR) but less so for automatic lip reading (ALR) [note: some academic literature refers to this as visual speech recognition (VSR) as well], which aims to understand what is being spoken based solely on the video data.

For videos that do not have audio available–security cameras, extremely noisy environments, medical patients who cannot speak, or video streams with broken microphones–methods to create transcriptions must rely on the video data alone leveraging ALR. The goal of our project is to build and train a series of models using image stills from videos (data sets below) to capture lip positioning and additionally, deduce the most likely words being said for transcriptions. We will then compare the various models, apply them to real world situations, identify drawbacks/challenges, and recommend avenues for enhancing the efficacy of automated lip reading.

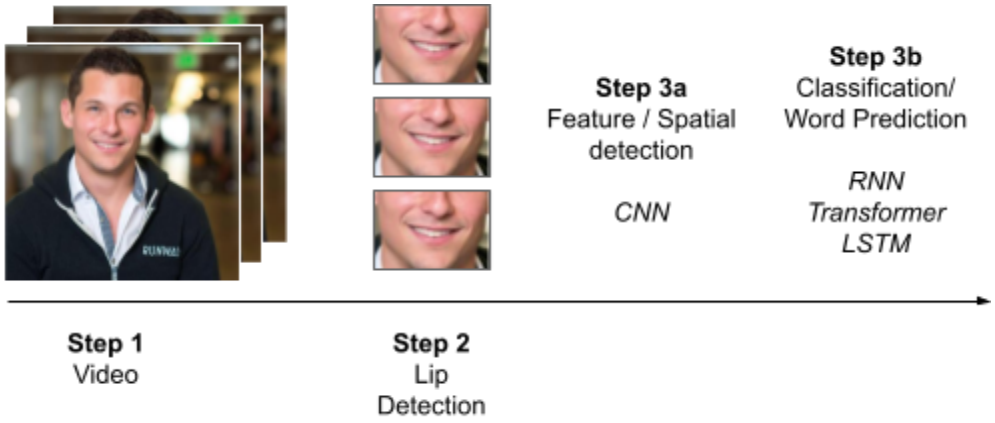


Figure 1. *Typical ALR Process*

II. History of Automatic Lip Reading (ALR) Approaches

Automatic lip reading is a technique used to decipher speech from visual cues, such as lip movements. This technology has been around for several decades and has been used in a variety of applications, including helping people with hearing impairments communicate and improving speech recognition systems. In the 1980s and 1990s significant progress was made in the field. Today, researchers are still working on improving the accuracy of automatic lip-reading technology and expanding its applications. Below is a timeline of significant automatic lip reading advancements:

- **1984:** Audio-Visual Automatic - Speech Recognition (AV-ASR) System - The first significant advancement happened in 1984 when Dr. Eric Petajan "...successfully extracted features from lip movement and combined them

with speech recognition to form an Audio-Visual Automatic Speech Recognition (AV-ASR) system [9]."

- **1994:** Hidden Markov Models - Hao et al. [9] mentions that extracted motion features were used as input features in a hidden markov model in 1994.
- **2007:** SVM - In 2007 an SVM model was used for isolated phrase recognition. [10]

The most recent advancements in ALR involve deep learning methods. Listed below are just a few:

- **2014:** Convolutional Neural Networks (CNN) - CNN models use filters to extract certain features from an image. In 2014, these models were used to extract lip features. [9]
- **2016:** Long Short-Term Memory (LSTM) Model - Huang et al. [10] mentions that in 2016 an LSTM model was used for lip reading and reached 79.6% in word-level lip reading.
- **2017:** Transformer - One of the most recent advancements in lip reading technology is when Google used a Transformer deep learning model for lip reading. This model is beneficial because not only does it speed up training, but it also can increase how well the model performs speech recognition. [10]

The advancements in ALR over the last forty years are remarkable. Thanks to all of these contributions in ALR, people are now able to understand what is going on in videos without having to hear the words. Although technology has come a long way, there are still many obstacles to overcome.

III. Challenges of Automatic Lip Reading (ALR)

And although there have been numerous advancements, many challenges still stand in the way of an optimal ALR system. Some of these challenges focus on data ingestion, such as the frame rate, camera or angle, while others focus on the speaker and the model and itself.

One of the main challenges in ALR systems resides on the visual ambiguities that arise at the word level due to homophenes, i.e characters that are easily confused because they produce the same or very similar lip movements (e.g. [p], [b] and [m]) [11, 21, 13]. Recall that the main objective of speech recognition systems is to understand verbal communication, which is structured in terms of sentences, words and characters, going from larger to smaller speech entities. More precisely, the standard minimum unit in speech processing is not the character, but the phoneme, defined as the minimum distinguishable sound that is able to change the meaning of a word [22].

One of the key challenges is the environment in which the videos are filmed in. Dataset collection is very much a double-edged sword. On the one hand, ALR requires a large enough dataset which has been accurately labeled and transcribed with text. [13] But most of the training data that is in this format is also homogenous. It does not include many of the issues that will arise in real world usage of ALR – such as a jittery camera which can throw off the facial matching algorithm, poor lighting conditions or an appropriate camera distance and quality. Additionally, further demographic distortions occur, leading to inaccurate model performance in real world situations.

A solution for this is the usage of synthetic datasets. Synthetic datasets involve generating artificial data for model training. Doing provides additional training methods with sufficient labeling, which given the relatively small number of ALR datasets in the wild is incredibly important. But additionally, synthetic data can help produce a distribution of speakers which is more accurate for speech recognition and considers differences in face physiology as well as producing utterances for unseen classes [13].

IV. Model Implementation

AV-HuBERT is a machine learning framework that uses audio and visual information to teach a machine to understand speech. It is a self-supervised learning method, meaning that it can improve its performance on its own without the need for human guidance. AV-HuBERT has been shown to produce excellent results in tasks related to speech recognition, such as lip reading and audio-visual speech recognition, on the LRS3 benchmark. This makes it a state-of-the-art technique in the field.

AV-HuBERT is composed of four modules: a feed-forward network (FFN) audio feature extractor, a modified ResNet video feature extractor, a fusion module, and a Transformer backend. The two feature extractors generate frame-level representation for the corresponding stream, which are frame-wise concatenated by the fusion module to form initial audio-visual features. The transformer backend takes these features and produces the contextualized frame-level audio-visual representations. The entire model is optimized to perform masked prediction, where random segments are masked for each stream independently, and the model learns to predict the cluster assignment of the masked frames. Figure 2. The cluster assignments are iteratively refined: it is produced by clustering Mel-frequency cepstral coefficients (MFCC) features in the first iteration, and by clustering previous iteration's AV-HuBERT representations in the subsequent iterations [12]. The model was pre-trained on the LRS3 and the English portion of VoxCeleb2 (VC2) datasets where it learns from both voice and mouth movements. AV-Hubert also has an extensive selection of Fine-Tuned Models for Visual Speech Recognition and Audio-Visual Speech Recognition. They can be found on https://facebookresearch.github.io/av_hubert/.

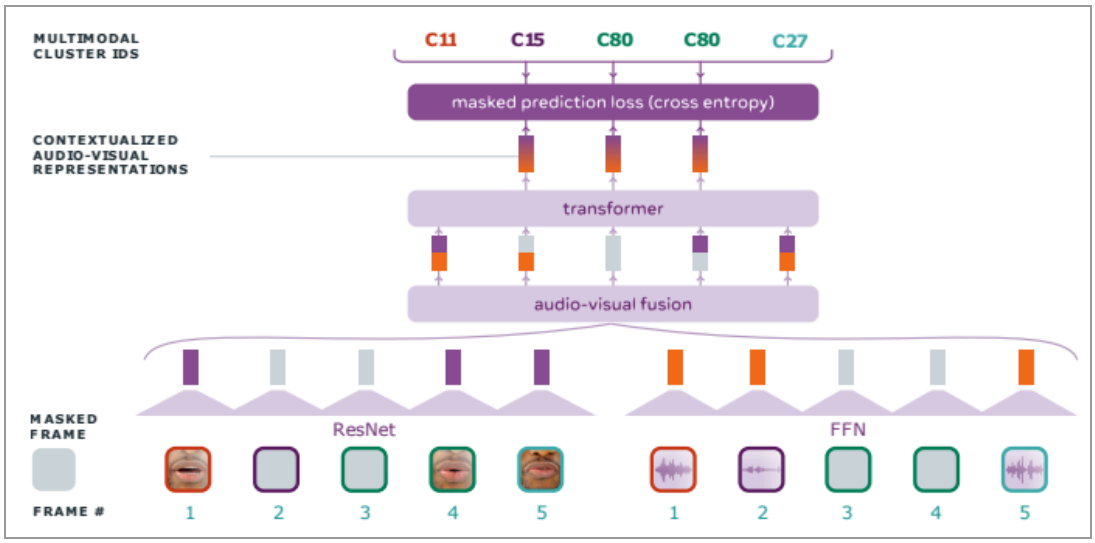


Figure 2. AV-HuBERT Model Schema

We have evaluated AV-HuBERT effectiveness and limitations with a synthetically generated dataset of videos of individuals speaking carefully chosen sentences (Appendix I). These sentences include the following chosen variables:

1. Native vs non-native speakers
2. Angle of taking the video
3. Gender of the person speaking in the video
4. Using words from newer context vs the data that was used for training
5. Sentences with pause, jumbled words, etc.

To evaluate the model, first the error rate between the original sentence spoken in the video and the sentence that was lipread by the system was calculated using the Levenshtein distance. The Levenshtein distance is a measure of the difference between two strings, which is defined as the minimum number of single-character edits (insertions, deletions, or substitutions) required to change one string into the other. This Levenshtein distance has been calculated at the word level, giving us the word error rate, WER, for each sentence, using the formula:

$$\mathrm{WER} = \frac{S+D+I}{N} = \frac{S+D+I}{S+D+C}$$

where, S = Number of Substitutions, D = Number of Deletions, I = Number of Insertions, C = Number of Corrects, N = Number of words in the reference.

The different variables listed above were also recorded for each sentence. An example output of the WER computation is Figure 3.

```
ref = "All our dreams can come true if we have the courage to pursue them"
wer_score = wer(ref.lower(), hypo)
wer_score
```

```
OP      REF     HYP
OK      all     all
OK      our     our
OK      dreams  dreams
OK      can     can
OK      come    come
INS     ****    to
SUB     true    you
OK      if      if
OK      we      we
OK      have    have
OK      the     the
OK      courage courage
OK      to      to
OK      pursue  pursue
OK      them    them
#cor 13
#sub 1
#del 0
#ins 1
{'WER': 0.143,
 'numCor': 13,
 'numSub': 1,
 'numIns': 1,
 'numDel': 0,
 'numCount': 14}
```

Next, the efficiency of the model was calculated by using linear regression and decision tree models.

The effect of variables can be summarized using the figure below:

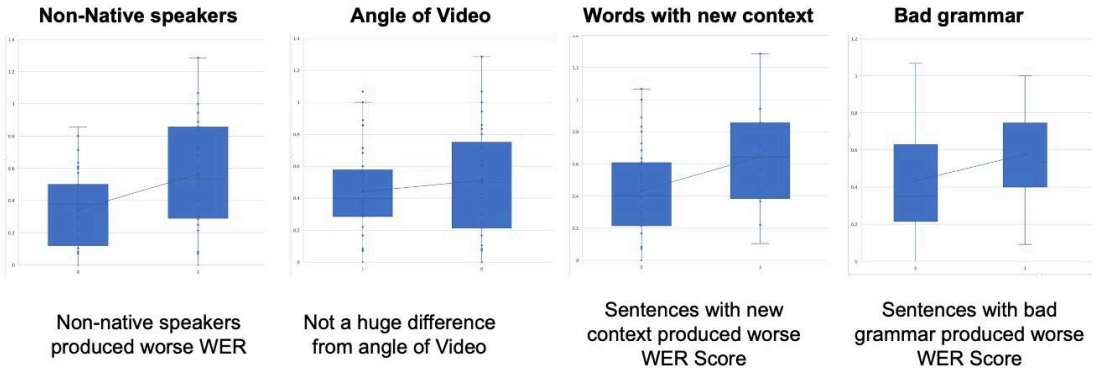


Figure 4. Input variable comparison for AV-HuBERT model

The variable importance, the Decision Tree and the Linear Regression model are described in Appendix II.

The following are the results inferred from them:

- When the speaker is a non-native speaker, the model produces worse results. There is a perceived bias against non-native speakers.
- Angle of taking the video wasn't significant to the model performance
- Gender of the speaker did make a small difference with the model performing better for women speakers
- Sentences with newer context and / or bad grammar got worse results
- The AV-HUBERT model performed surprisingly well, despite the errors. WER was 48.1% and benchmarks are

typically 20-30%

- The model was slow to interpret given the high resolution of videos (typically 1-2.5 minutes using a GPU) making it difficult and / or impractical for real-time transcription

V. Future of ALR

Lip-reading is notoriously difficult, depending as much on context and knowledge of language as it does on visual clues. But researchers are showing that machine learning can be used to discern speech from silent video clips more effectively than professional lip-readers can.

The massive expansion of social and visual media posted online, have enabled researchers to generate far larger datasets, like the Oxford-BBC Lip Reading Sentences 2 (LRS2), which is based on thousands of spoken lines from various BBC programs. LRS3-TED gleaned 150,000 sentences from various TED programs while the LSVSR (Large-Scale Visual Speech Recognition) database, among the largest currently in existence offers 140,000 hours of audio segments with 2,934,899 speech statements and over 127,000 words. And it's not just English: Similar datasets exist for a number of languages such as HIT-AVDB-II, which is based on a set of Chinese poems, or IV2, a French database composed of 300 people saying the same 15 phrases. Similar sets exist too for Russian, Spanish and Czech-language applications.

Today, speech recognition comes in three, which depends on the input source. Visual Speech Recognition- that is, using only visual means to understand what is being conveyed. Conversely, there's Automated Speech Recognition which relies entirely on audio, i.e. "Hey Siri," and Audio-Visual Automatic Speech Recognition, which incorporates both audio and visual cues. The tremendous success of deep learning in these fields has already affected visual speech recognition by shifting the research direction from handcrafted features and Hidden Markov Model (HMM) based models to deep feature extractors and end-to-end deep architectures.

Recently introduced deep learning systems beat human lip-reading experts by a large margin, at least for the constrained vocabulary defined by each dataset. Adversarial attacks have become one of the most actively researched topics in the deep learning space recently. The kind of drawback that these attacks have exposed have made researchers lookout for any anomalies even in reinforcement learning [11]. In addition, the best human lip readers rely on significant amounts of additional information to interpret speech, such as the context of the conversation, the speaker's body movements and a good knowledge of grammar, idioms and common speech. These are factors that computers have yet to get to grips with. Automated lip reading may still be some way off but the early signs are that it is by no means impossible which raises a whole set of other privacy-related issues. For example, it may be that videos of conversations without sound are impossible to interpret now but may be easy to interpret in future with the implantation of one of the above models.

VI. Conclusion

Ultimately, ALR is a fast-developing area of research and development in academia as well as industry. As seen in the chart on the following page, models using video data alone are fast approaching 80-90% accuracy in well-lit rooms with native (and largely-white speakers).

For our final project, our team attempted to implement/fine-tune several models as well as develop models from scratch. We ultimately decided to evaluate the impact of various input video types (evaluation sheet linked here) on the performance of the model. We came to the conclusion that there is quite a bit of bias in the model largely due to the training data set (which is easily confirmed looking at the LSR2 and other data sets which have predominantly white speakers, a common challenge in the industry). We believe that the bias perceived in the model can be improved by training the model with more of the following:

- non-native English speakers
- using synthetically-generated (e.g. diverse speakers with same utterances)

sentences versus video clips taken from other existing videos

Ultimately, this was a challenging project that, while it was feasible to implement, required easier access to data corpuses (some required applications and long wait periods), more compute resources and even pre-developed virtual environments to run PyTorch or other applications, etc. All said, we believe our findings as well as understanding of and the ability to implement the AV-HuBERT led to meaningful conclusions that we are excited to build upon.

Ref.	AI method	Accuracy	Dataset		Task
			Name	Size	
[3]	MLLT + SAT, DNN	48% mean (visemes) 52% mean phonemes.	200 sentences selected from the RM *corpus*.	only the front view vocabulary size of around 1000 words	Word
[4]	VGG-M, 3D Conv. with Early Fusion and Multiple Towers	92.5% at sentence level 88.6% in unseen speakers	Their own dataset	29 speakers 118,166 Utterances Duration 33 h.	Sentences
[8]	Spatiotemporal conv., residual and bidirectional LSTM networks.	83.0% at word level	Videos extracted from BBC TV broadcasts	500-size target-words with 1.28 sec video excerpts	Words
[10]	6-layer Deep Auto-encoder NN (DANN) GMM-HMM and DNN-HMM hybrid	15.4% Compared to shape features 20.4% Compared to ROI features	CUAVE	digits (0 to 9) 36 speakers (19 males and 17 females) 80 isolated digits	Isolated and connected digits
[11]	12-layer CNN with 2 layers of batch normalization	96.5% on training set 52.9% on validation set.	MIRACL-VC1	3000 instances	Word or phrase
[12]	CNN models: AlexNet and Inception V3	Speaker dependent AlexNet 86.6%, Inception V3 64.6% speaker independent AlexNet 37.1% inception-V3.17.6%	Miracl-VC1	15 speakers, 1500 instances	Word
[13]	Pre-trained deep learning architecture VGG Net	94.86% in training, 93.82% in validation and 60% in testing	MIRACL-VC1 dataset with some modifications	15 speakers, 1500 instances	Word
[16]	Deep 3D CNNs, two-stream	84.07%	LRW	number of target words = 500	Word
[15]	CNN + Hahn moments	59.23%, 93.72%, and 90.86% on AV-Letters, OuluVS2 and BBC LRW, respectively.	AV-Letters, OuluVS2 and BBC LRW		Letters, digits or words

References

1. Fernandez-Lopez, Adriana. “Towards Estimating the Upper Bound of Visual-Speech Recognition: The Visual Lip-Reading Feasibility Database.” *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, vol. 12, no. 2017. *iEEEexplore*, https://ieeexplore.ieee.org/abstract/document/7961743.

2. “Lipreading.” *Papers With Code*, https://paperswithcode.com/task/lipreading. Accessed 22 November 2022.

3. “Papers with Code - Distinguishing Homophenes Using Multi-Head Visual-Audio Memory for Lip Reading.” *Papers With Code*, https://paperswithcode.com/paper/distinguishing-homophenes-using-multi-head-1. Accessed 22 November 2022.

4. “Papers with Code - Lip Reading in the Wild Benchmark (Lipreading).” *Papers With Code*, https://paperswithcode.com/sota/lipreading-on-lip-reading-in-the-wild. Accessed 22 November 2022.

5. “Papers with Code - Training Strategies for Improved Lip-reading.” *Papers With Code*, 3 September 2022, https://paperswithcode.com/paper/training-strategies-for-improved-lip-reading. Accessed 22 November 2022.

6. Tarantola, A. “AI is already better at lip reading than we are.” *Engadget*, 29 September 2022, https://www.engadget.com/ai-is-already-better-at-lip-reading-that-we-are-183016968.html? Accessed 22 November 2022.

7. Thangthai, Kwanchiva. “Improving Lip-reading Performance for Robust Audiovisual Speech Recognition using DNNs.” *ISCA Archive*, vol. Sep 11 2015, 2015, p. 127. *ISCA-Speech*, https://www.isca-speech.org/archive_v0/avsp15/papers/av15_127.pdf.

8. Fernandez-Lopez, Adriana and Sukno, Federico. “Survey on Automatic Lip-Reading in the Era of Deep Learning”, https://repositori.upf.edu/bitstream/handle/10230/36119/fernandez_ivc_surv.pdf?sequence=1&isAllowed=y

9. Mingfenf Hao, Mutallip Mamut, Nurbiya Yadikar, Alimjan Aysa, and Kurban Ubul. “A Survey of Research on Lipreading Technology”, https://ieeexplore.ieee.org/ielx7/6287639/8948470/09252931.pdf. Accessed 22 November 2022

10. Hongyang Huang, Chai Song, Nurbiya YadikarJin Ting, Taoling Tian, Kurban UbulChen Hong, Zhang Di, and Danni Gao. “A Survey of Research on Lipreading Technology”, https://ieeexplore.ieee.org/ielx7/6287639/8948470/09252931.pdf.

11. P. Ma, S. Petridis, and M. Pantic, “Detecting adversarial attacks on audiovisual speech recognition,” arXiv.org, 12-Feb-2021. [Online]. Available: https://arxiv.org/abs/1912.08639.

12. B. Shi, A. Mohamed, and W.-N. Hsu, “Learning Lip-Based Audio-Visual Speaker Embeddings with AV-HuBERT,” arXiv:2205.07180 [cs, eess], Jul. 2022, [Online]. Available: https://arxiv.org/abs/2205.07180

13. Oghbaie, M., Sabaghi, A., Hashemifard, K., & Akbari, M. (2021, October 15). Advances and challenges in deep lip reading. arXiv.org. Retrieved December 12, 2022, from https://arxiv.org/abs/2110.07879

Appendix I

The sentences used in this study:

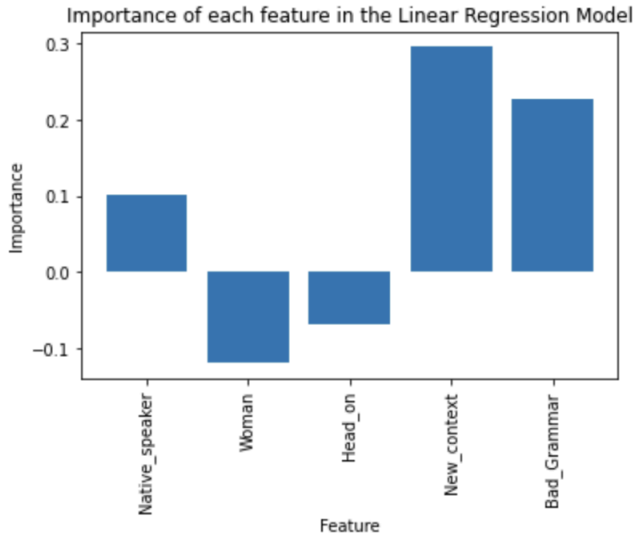
1. All our dreams can come true if we have the courage to pursue them
2. It is the duty of the Madrigal family to use their gifts for the betterment of others around them
3. Without the Black Panther Wakanda will fall
4. Just because something works doesn't mean it can't be improved
5. I make grave mistakes all the time Everything seems to work out
6. We Do Not Follow Maps To Buried Treasure And X Never Ever Marks The Spot
7. Okay here we go Focus speed I am speed One winner 42 losers I eat losers for breakfast
8. With great power comes great responsibility
9. Named must be your fear before banish it you can
10. No Try not Do Or do not There is no try

Appendix II

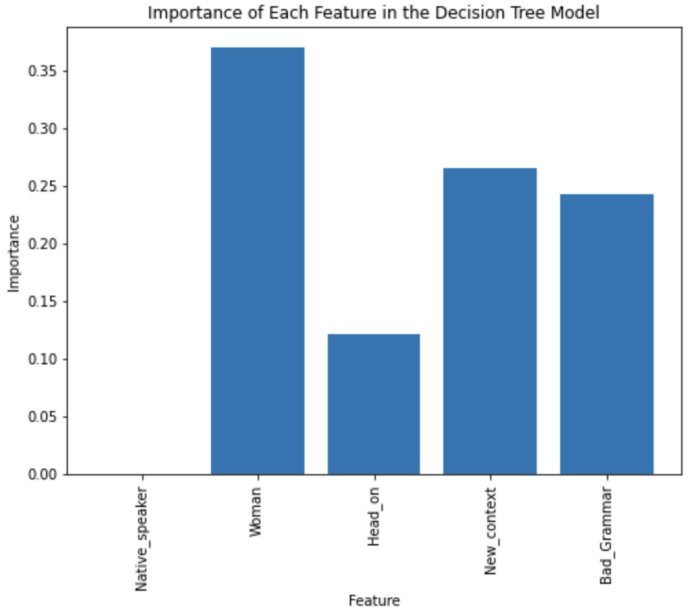
The equation for the linear regression model is :

y = 0.09a - 0.12b - 0.08c + 0.286d + 0.217e where y is the predicted WER

The Variable importance plot for Linear regression:



The variable importance produced by Decision Tree Model :



The tree produced by the Decision Tree Model was :