

## **Speech-to-Text Analysis**

Krishna Sai Biradar, Sasidhar Guthi, Suresh Ganta

Data Science Department, University of Maryland Baltimore County

DATA 690: Natural Language Processing

Dr. Tony Diana

Fall 2022

## **Abstract**

This paper focuses on speech-to-text conversion and how this would be useful for users. It discusses how we have used transformers and their different pre-trained models for speech-to-text conversion. Acquiring the text summarization and named entity recognition and how this would help the user. Finding the sentiment behind the text we have transcribed from the audio input. This goes on to discuss how we have created a webpage for the users to upload files and obtain the above outputs. This also discusses the translation of the audio files so that the text analysis is irrespective of the language spoken.

## **Introduction**

Speech-to-text analysis transforms audio input into text that can be analyzed. There are many instances where there is a difference between what is said and understood. Analysing a person's speech tells us how the person is reacting to a situation or the emotion behind a statement. While people would be talking about how they are feeling about a situation we would be able to measure sentiment and emotions. We also found that many people are wondering about the message behind an audio file. As we transcribed the audio file from speech to text, we were able to identify the important topics, and we were also able to summarise the transcribed text into a short description of the audio file. After determining the named entities and summarized text, we assessed the speaker's sentiment in the audio file. Our text analysis provided the summarized text of the audio file with the main topics highlighted and how much percentage of the audio was positive and negative. For audio or speech inputs in different languages, we transcribed them first and then translated them to achieve the output. We developed a webpage where we would be able to upload the audio file of our choice to generate a text summary, determine the named entities of the audio file, and measure the sentiment of the audio file's speaker.

## Literature Review

In the past speech-to-text conversion is to stenotype a protocol or to record the event and convert it into text. Nowadays we can provide real-time-speech-to-text conversion. Automatic speech recognition can recognize and write more than 90 per cent of a long series of spoken words (Wagner,2005). One of the main reasons for selecting this ASR (Automatic Speech Recognition) is for people who cannot hear or understand the English language. Some of the models involved in the speech-to-text conversion used The Mel-Frequency Cepstral Coefficient (MFCC) feature extraction methodology and the Minimum Distance Classifier, Support Vector Machine (SVM) techniques for voice classification, which were introduced in a multilingual speech-to-text conversion system reported by Ghadage and Shelke (2016). Wan (2018) proposed a method for summarizing English texts using association semantic criteria. The author claimed that the new extraction strategy provided superior convergence and accuracy performance. As the most widely used method for topic-based text categorization, LDA (Latent Dirichlet Allocation) is a powerful tool for analysing large amounts of data. A new similarity calculation approach proposes an enhancement of the same. Biswas (2018) reviewed methods of text summarization with a special emphasis on methods that pruned out unnecessary words.

Recently transformers surpassed other neural models, including convolutional and recurrent neural networks, in performance for tasks in natural language interpretation and synthesis, making it the dominating architecture for NLP (Vaswani, 2017). This explains why we selected transformers over the other models. The architecture of our Transformers was influenced by the groundbreaking tensor2tensor library (Vaswani, 2018) and BERT (Devlin, 2018), both from Google Research. Based on Zhao (2022), the wav2vec model, took raw audio as the first input to the feature encoder to obtain latent speech recognitions, before feeding it to the transformer. The model was the Hidden unit BERT (Hubert). It is a BERT-like pre-training approach based on an offline clustering step to generate noisy labels first. The model architecture is nearly the same as wav2vec2.0, while the training task is masked prediction rather than contrastive learning. Contrastive learning creates positive pairs which use back-translation to generate another view of the original English data (Fang et al. 2020).

## Methodology

### Data and Data Preparation

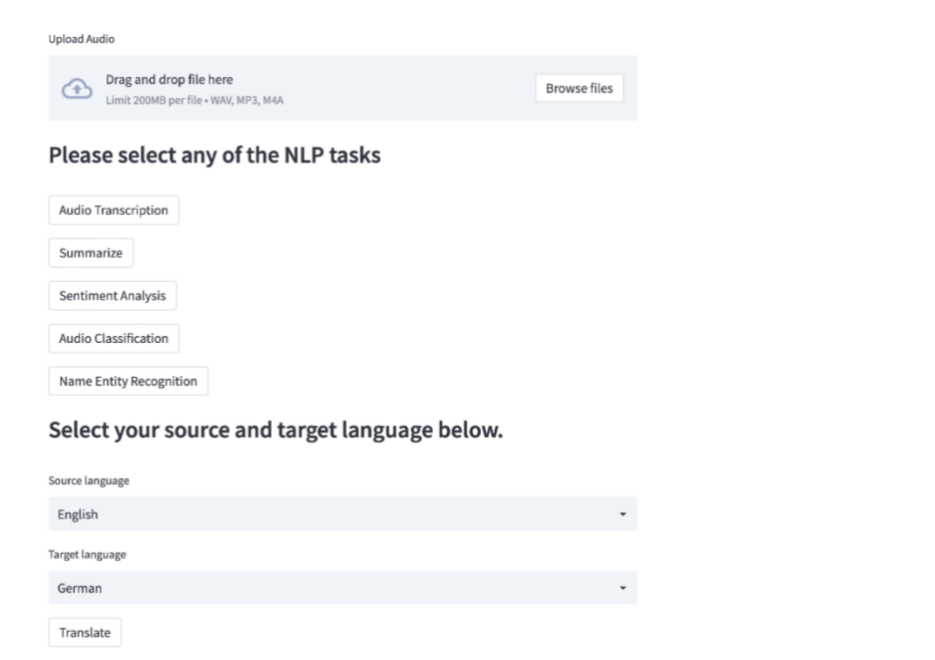
We collected data from various Internet sources in the form of audio files. Each audio file size is less than 1MB, and the audio file is in .wav format.

The audio files consist of interviews between two speakers, commentaries of different sports which contain crowd noise, speeches with noise and low pitch, speeches without noise etc. We have considered different audio files to test the performance of the model under critical conditions to ensure goodness of fit.

### Models and Results

#### *Streamlit*

To allow user interactions we developed a streamlit application where users would be able to upload their audio files with minimum wait time before, receiving the desired output of their audio file (summarization, named entities, sentiment polarity). Fig 1 shows the streamlit application.



The screenshot displays the Streamlit application interface. At the top, there is an 'Upload Audio' section with a light blue box containing a cloud icon, the text 'Drag and drop file here', a limit note 'Limit 200MB per file • WAV, MP3, M4A', and a 'Browse files' button. Below this, a heading 'Please select any of the NLP tasks' is followed by five buttons: 'Audio Transcription', 'Summarize', 'Sentiment Analysis', 'Audio Classification', and 'Name Entity Recognition'. The next section, 'Select your source and target language below.', contains two dropdown menus. The 'Source language' dropdown is set to 'English', and the 'Target language' dropdown is set to 'German'. A 'Translate' button is located at the bottom of this section.

Fig 1 Streamlit Application Interface.

## Audio Transcription

HuBERT is one of the pre-trained models for audio transcription. As Hsu (2021) mentioned, this model outperformed other pre-trained models in the area of literature. It is based on word2vec 2.0 and contrastive learning.

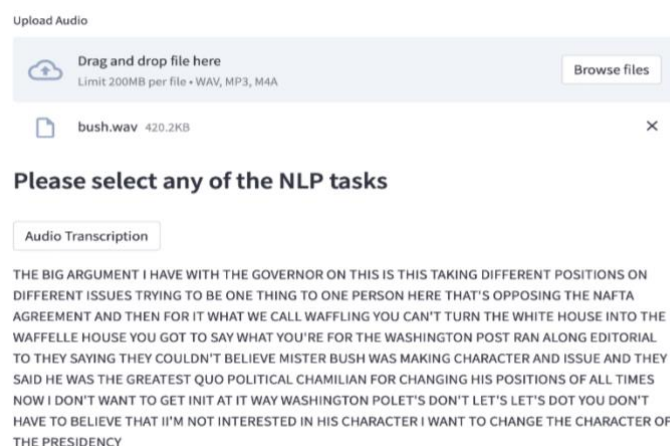


Fig 2 Transcribed audio.

After uploading the file using the webpage, we generated output as the audio transcription, at the click of a button. The audio transcription in fig 2 utilized the HuBERT model. We will be using this transcribed audio for the rest of our text analysis.

Whisper is an NLP model built by OpenAI. This model was developed by Radford (2022) trained for long-form transcriptions. We have used it to provide speech-to-text conversion for large audio files.

## Summarization

BART represented one of our models for text summarization. BART as Lewis M (2019) mentioned did better than the BERT model which was previously considered the best model for text Summarization. BART model outputs are highly abstractive with phrases copied from the input that is generally factually accurate and integrate supporting evidence.

Upload Audio

Drag and drop file here  
Limit 200MB per file • WAV, MP3, M4A

Browse files

bush.wav 420.2KB

Please select any of the NLP tasks

Audio Transcription

Summarize

The White House can't turn the White House into the WAFFELLE House, says President George W. Bush . President Bush has been criticized for changing his positions on issues such as NAFTA and NAFTA . The president says he wants to change the role of the president of the presidency .

Sentiment Analysis

Name Entity Recognition

Select your source and target language below.

Source language

English

Target language

German

Fig 3 Summarized Text

Fig 3 shows the summarized text of the transcribed audio file as the result of the BART model.

### ***Sentiment Analysis***

To maximize our computer resources, we have used DistilBERT see Sanh (2019). It is a smaller general-purpose language representation model. This model is a pre-trained version of BERT, 40 per cent smaller, and 60 per cent faster, that retains 97 per cent of the language understanding capabilities.

Upload Audio

Drag and drop file here  
Limit 200MB per file • WAV, MP3, M4A

Browse files

bush.wav 420.2KB

Please select any of the NLP tasks

Audio Transcription

Summarize

Sentiment Analysis

[[{"label": "NEGATIVE", "score": 0.9966676831245422}]]

Name Entity Recognition

Select your source and target language below.

Source language

English

Target language

German

Fig 4 Sentiment Polarity

Sentiment analysis of the transcribed text is performed by the DistilBERT model, as you can see in Fig 4 the audio file uploaded is 99.6 per cent negative.

### ***Named Entity Recognition***

Spacy has been showing a significantly more accurate recognition system, it not only gives out responses in the standard form of Names, Places, and Organizations but also adds layers of detail to outputs.

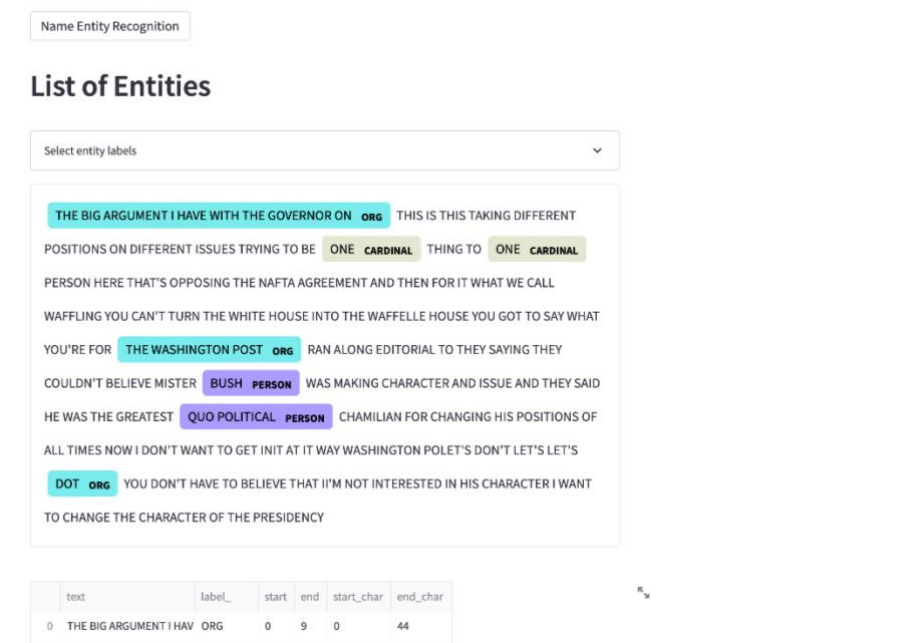


Fig 5 Name Entity Recognition

Named entities were recognized for the audio file uploaded as Spacy took the transcribed text as input. The output of the named entities can be seen in Fig 5.

### ***Translation***

We have used one of the models in Spacy called en\_core\_web\_sm for English-speaker text translation. Fig .6 illustrates the translated text.

Summarize

sentiment-analysis

Name

Select your source and target language below.

Source language  
English

Target language  
German

Translate

**Translated Text**

DER GROSSE ARGUMENT, DEN ICH MIT DER GOUVERNREGIERUNG ÜBER DIESE VERFÜGBARKEIT IST, DIE VERFÜGBARKEIT DER VERTRAG DER NAFTA AN DER JEDER PERSON IST, DIE DEN VERTRAG OPPOSIIERT, UND DIE DAMIT W

Fig 6 Translated Text.

We added languages such as German and French. We will be able to add different languages based on the user's preference.

### **Audio Classification**

We used the Speech processing Universal PERformance Benchmark (SUPERB) model for audio classification. This Pretrained model measured the emotions shown in the audio file by categorizing them into happy, sad, and angry, among others see Yang S (2021). Fig 7 shows the emotion categorization.

Drag and drop file here  
Limit 200MB per file • WAV, MP3, M4A

Browse files

bush.wav 420.2KB

Please select any of the NLP tasks

Audio Transcription

Summarize

Sentiment Analysis

Audio Classification

Name Entity Recognition

[{'score': 0.9869334697723389, 'label': 'hap'}, {'score': 0.010098624974489212, 'label': 'sad'}, {'score': 0.002216066466444063, 'label': 'neu'}, {'score': 0.0007517611375078559, 'label': 'ang'}]

Fig 7 Emotion Categorization.



### ***Challenges Faced and future improvements***

Here are some of the challenges that we experienced:

- The processing power of the local machines was not enough, which required the use of Google Colab Pro, even though we were not able to process more than 1MB of the audio file.
- As Whisper AI dealt with large data transcriptions, we could not observe the model's full potential as we were only restricted to smaller audio files.
- Users would like to upload files with different formats, but since mp3 file quality was not sufficient enough for the models to process speech to the text we were confined to using .wav format files.

### **Conclusion**

In this project, we developed a streamlit application, in which users would be able to upload the audio file.

- This audio file was being transcribed into text, which would be useful for many users facing difficulty in understanding the English language.
- We provided a summarized text output of this audio file for users who wanted to get a short description of the audio files.
- A named entity output of this audio file allowed users to identify the important points mentioned in the audio input, which would help with keywords, places, products, etc.
- We translated the text of the audio file so that users from a different language would be able to understand the audio file in their language.
- Finally, we made it possible for the users to assess the emotional content of the audio file.

## References

- Wagner, S. (2005, May). Intralingual Speech-to-text conversion in real-time: Challenges and Opportunities. In *Challenges of Multidimensional Translation Conference Proceedings*.
- Y. H. Ghadage and S. D. Shelke, "Speech to text conversion for multilingual languages," 2016 International Conference on Communication and Signal Processing (ICCSP), Melmaruvathur, pp. 0236-0240, 2016.
- L. Wan, "Extraction Algorithm of English Text Summarization for English Teaching," 2018 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS), Xiamen, China, 2018, pp. 307-310.
- S. Biswas, R. Rautray, R. Dash and R. Dash, "Text Summarization: A Review," 2018 2nd International Conference on Data Science and Business Analytics (ICDSBA), Changsha, China, pp. 231-235, 2018.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I Guyon, U V Luxburg, S Bengio, H Wallach, R Fergus, S Vishwanathan, and R Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.
- Ashish Vaswani, Samy Bengio, Eugene Brevdo, Francois Chollet, Aidan N. Gomez, Stephan Gouws, Llion Jones, Łukasz Kaiser, Nal Kalchbrenner, Niki Parmar, Ryan Sepassi, Noam Shazeer, and Jakob Uszkoreit. 2018. Tensor2tensor for neural machine translation. CoRR, abs/1803.07416
- Zhao, J., & Zhang, W. Q. (2022). Improving Automatic Speech Recognition Performance for Low-Resource Languages With Self-Supervised Models. *IEEE Journal of Selected Topics in Signal Processing*, 16(6), 1227-1241.
- Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

- Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., ... & Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.
- Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). Robust speech recognition via large-scale weak supervision. *OpenAI Blog*.
- Yang, S. W., Chi, P. H., Chuang, Y. S., Lai, C. I. J., Lakhotia, K., Lin, Y. Y., ... & Lee, H. Y. (2021). Superb: Speech processing universal performance benchmark. *arXiv preprint arXiv:2105.01051*.
- Hsu, W. N., Bolte, B., Tsai, Y. H. H., Lakhotia, K., Salakhutdinov, R., & Mohamed, A. (2021). Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29, 3451-3460.
- Fang, H.; Wang, S.; Zhou, M.; Ding, J.; and Xie, P. 2020. CERT: Contrastive Self-supervised Learning for Language Understanding. *arXiv preprint arXiv:2005.12766*, 1–16.