BTP-I report

Generating aligned transcripts for broadcast news audio

by

Mitali Meratwal
190070033

under the guidance of

Prof. Preeti Rao

Department of Electrical Engineering
Indian Institute of Technology Bombay

1 Introduction

Automatic speech recognition (ASR) systems are trained to generate text for the input speech. The training data for the ASR systems comprises of audio files with their corresponding word-level transcription. The availability of large datasets has aided the shift from statistical HMM-based speech recognition to deep learning based models. The size and quality of the training data significantly impacts the accuracy of such methods. While the audio files can largely vary in length, considering GPU limitations and the need of contextual information, most ASR models require longer audio files to be segmented.

Given a long duration audio file (eg. 10 minutes) and the corresponding text transcript, our goal is to (1) segment the audio into smaller chunks containing only speech, (2) rejection of poor-quality audio segments, and (3) get the text corresponding only to the audio segment from the complete transcription. The chunks containing only jingles, a lot of background noise, and a poor signal-to-noise ratio (SNR) are removed. In this work, we explore preprocessing stages of an ASR toolkit Vakyansh[1] offering some of these solutions. Particularly, we are concerned with the evaluation of the techniques used by them on broadcast news audio, identifying their systematic errors, and testing other available solutions.

Our work can be summarized as (1) studying the characteristics of the dataset from News On AIR broadcasts[2] (2) We test WebRTC used in Vakyansh for voice activity detection (VAD) and do a grid search on other parameters like frame duration and ”aggressiveness”. We report changes in VAD output and quality. (3) We address the issues of WebRTC like high false positives, by using another VAD technique called Silero VAD. We report the best values of tunable parameters of Silero VAD that result in good speech-only segments for News On Air broadcasts. (4) Default values of upper and lower thresholds on SNR in Vakyansh cause about 50% of audio chunks to be rejected. Using histogram analysis on SNR values of audio chunks, we find values of these thresholds that mostly reject only noisy segments. (5) Speech-text alignment by instead aligning text generated from the speech segment and original transcription is proposed. Using python's SequenceMatcher to find similarity scores between transcribed text and ngrams in a moving window on each tokenized sentence gave better results than all other methods.

2 Background and motivation

Cleaning and preprocessing of data, especially for low-resource languages where data is already limited, is crucial for the training of ASR models. Our work focuses on broadcast news audio in Marathi and Gujarati. On a sample of News On Air audio bulletins, we observed that the news audios can have background music, multiple speakers, advertisements, field reporters, audience interaction, etc. However, the transcript contains text corresponding to the speech of only the main speaker. Sometimes there is also a mismatch between the speech and transcript. We typically do not want to feed such training data to ASR systems and remove audio segments not belonging to the main speaker or those not having the corresponding transcript. So simply performing forced alignment between entire audio and text is not valid for such datasets.

An approach to addressing these problems is to first section the audio into smaller segments containing only speech. Voice activity detection techniques can help in breaking audio into only speech chunks and reject non-speech parts of the audio. However, singing is also considered speech and so depending on the input dataset, additional processing steps like speaker clustering might be needed to further reject such outlier audio segments. The audio chunks containing a lot of background noise and disturbances can be filtered out by thresholding their SNR. The threshold needs to be set carefully to prevent the loss of data that is actually clean and otherwise would not have been removed. With Vakyansh's WADA SNR, we found around 50% of the data getting rejected, which motivates us to look for other SNR techniques and thresholding appropriately. Ultimately we want an aligned text for only the speech in the audio chunk and exclude those that are not transcribed. Vakynash does not clearly mention the details of how exactly the speech-text alignment is performed after all the preprocessing steps. So we leverage machine learning models already trained on large datasets with reasonable performance, such as whisper, to get the text for the audio chunks. The generated text and actual transcript can then be aligned using string similarity methods.

3 Dataset

All the experiments in this work have been performed on arbitrarily selected audio samples from News On AIR for Marathi and Gujarati. Each audio file is roughly around 10 minutes long. Most of the audio recordings started and ended with music, followed by a short announcement of the ensuing news by the main speaker. For audio in Gujarati, the speech starts only after the music ends, but for Marathi, the music seemed to slowly die down after the speaker began speaking. We definitely want to reject only music parts but also those segments where music is more prominent than news. The news audio also contained an advertisement section spoken by 2 different speakers (one male and one female) lasting for about 1 minute. For Marathi, we found the speaker summarizing the main news points at the end. However, transcription for this part was found to be completely missing. We also came across instances where the main audio is in Gujarati, but around starting 4 minutes of 14-minute audio contained news in English. Some audio occasionally included interviewing the audience, speech clippings of PM Modi in Hindi, news by field reporters etc. Again there is an absence of text transcription for these parts, and should not be passed as training data. These characteristics of the dataset highlight why we can't just use forced alignment between the entire audio and transcript and need aligned text for only the relevant speech parts.

4 Vakyansh pipeline

Vakyansh[1] is an ASR toolkit that offers data collection, data validation, data processing and filtering, and model training pipeline for building applications leveraging speech recognition. We wish to test the performance of pre-processing stages of Vakyansh on our dataset. Fig. 1 shows the flow of data through three key methods. First, the long audio file is passed through WebRTC to perform VAD and break into shorter segments of speech chunks. This is to ensure words are not broken at the boundaries of clipped audio. Next, for every audio segment generated from WebRTC VAD, segments that mostly contain background noise or music are removed by finding if the SNR (based on waveform amplitude distribution analysis[3]) lies outside the threshold range. Since there is no mapping between the audio and speaker label, good-quality chunks of the unique main speaker can be obtained by speaker clustering on embeddings of clean chunks obtained from the previous step. The top k SNR segments for each speaker are finally selected. We evaluate the results for each method and better alternatives in the next sections.

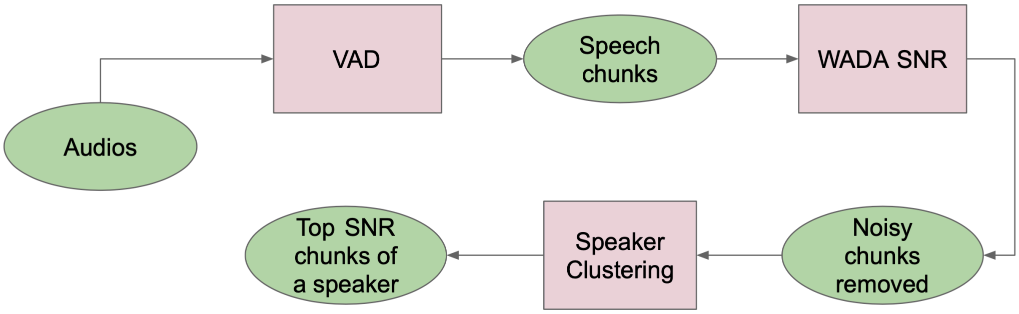


Figure 1: Vakyansh pre-processing stages

5 Voice Activity Detection Methods

5.1 WebRTC

The aim of VAD is to detect the presence or absence of human speech in a signal. The audio is segmented into frames of a given frame duration. For every frame, a VAD decision is found, 0 for non-speech and 1 for speech, based on the likelihood ratio test. In this method, the probability for both speech and background noise is calculated using GMMs, and a hypothesis test is performed to decide which is more probable. The input features to GMMs are log_{10}(energy in frequency band) with length equal to the number of frequency bands i.e. 6. A circular buffer of these frames is maintained with a maximum buffer size configured by the padding duration parameter. So for a padding duration of 300ms and frame duration of 30ms, a moving window of 10 frames (30ms each) is maintained. Speech is considered to have started when at least 9/10 (>90%) frames have VAD decision as speech and the frames currently in buffer, and incoming frames are concatenated subsequently. Similarly, the segment ends if at least 9/10 (>90%) frames in the buffer are non-speech. Another parameter that affects WebRTC's VAD decision is "aggressive mode" (AM). A Higher AM value results in a higher probability of a frame being speech when its VAD decision is 1.

Case	Frame duration (ms)	Aggressive mode	Avg. chunk duration (s)
I	10	2	5
II	20	2	10
III	30	0	26
IV	30	2	15
V	30	3	7

Table 1: Effect of different WebRTC parameters on VAD output

The default value of frame duration, padding duration, and aggressiveness level chosen in Vakyansh is 30ms, 300ms, and 2, respectively. Table 1 shows how WebRTC performs for different frame duration and aggressiveness levels. In case I, when the frame duration is small, the average duration of the output segments becomes too small. This affects the speaker clustering stage, where all the segments are either labeled as noise or grouped in a single cluster as the audio segment length is too short to differentiate between speaker embeddings. Even changing the AM does not produce a significant difference in VAD output. On increasing the frame duration in case II, continuous phrases get broken into separate chunks, making speech-text alignment challenging. Case III onwards, we fix frame duration at 30ms and increase AM. AM = 0 produces longer chunks, and audio does not get broken even at considerably long silences in the segment. Some segments were as long as 2 minutes. While AM = 1 produced a desirable length of audio segments, the audio was not always broken at news boundaries. For a very aggressive level of 3, speaker clustering again fails even for different cluster sizes. Cluster size decides the minimum number of points required to form a cluster.

However, in all cases, we observed that the 1st and last segments containing only music were also considered speech sounds. Similarly, the music in advertisements also passed the speech test. Even performing SNR thresholding did not completely filter out such chunks. Thus, due to the high false positives, chunks with only silences, and observations discussed above, we look for alternate VAD methods.

5.2 Silero VAD

Silero[4] provides enterprise-grade speech-to-text and text-to-speech models. Their open-source Silero VAD is based on multi-head attention neural network. The long audio input to the model is broken into lengths of a given window size. Short-time Fourier transform of each window serves as an input feature to the model that returns the speech probability for each window. The windows are appended till the speech probability lies above the threshold parameter and include additional minimum silence duration frames before separating the segments. Final audio segments with lengths less than the minimum speech duration are discarded by the API. For our audio samples, we found a speech probability threshold of 0.6 to successfully reject most of the music containing chunks. Further, performing SNR thresholding removed all, if any, music chunks that were produced after Silero VAD. Best results were obtained using a default window size of 1536 samples, a minimum silence duration of 150 ms, and a minimum speech duration of 250 ms. The average chunk duration is about 7 s which is close to the typical length of a sentence. On increasing speech threshold probability, the chunks became shorter, with words getting broken at boundaries, while a low threshold value failed to reject music windows. Decreasing minimum speech duration increased false positives. So we stick to the default values for other parameters.

6 WADA SNR

In order to obtain only clean speech audio segments after VAD, Vakynash applies thresholding on the SNR of audio chunks calculated using waveform distribution analysis (WADA). WADA-SNR assumes clean speech can be modeled using gamma distribution and additive noise as Gaussian distribution. WADA-SNR gives better performance than the standard NIST STNR even in the presence of other types of noise that may not be gaussian, like background noise or an interfering speaker. The probability density function of clean speech can be described as :

$$f_x(x|\beta_x) = \frac{\beta_x}{2\Gamma(\alpha_x)}(\beta_x|x|^{\alpha_x-1})exp(-\beta_x|x|) \tag{1}$$

where α_x is chosen to be 0.4 and some arbitrary value for β_x as it is just normalizing the density function and does not affect SNR estimation. Chunks with WADA-SNR less than 20 dB and greater than 60 dB are rejected in Vakyansh. From Fig. 2, we observe that with this threshold range, around 50% of the sample data was lost after WebRTC VAD due to many clean chunks having low SNR values being rejected.

To analyze differences between Silero and WebRTC post SNR thresholding, we manually labeled regions of audio as main speaker, music, advertisement, etc., for some audio post VAD. Using WebRTC, some noisy or music chunks produced high SNR values, passing the threshold test. For example, in Fig. 4, the bottom row corresponds to the audio after WebRTC VAD, and the last music chunk predicted as speech by WebRTC has high SNR values. Clearly, for WebRTC, a threshold of even 15 dB fails to remove the last chunk containing only music apart from chunks where background noise is present when the main speaker speaks. But in the case of Silero VAD, the top row in Fig. 4, note that the last music chunk is not passed in the pipeline, and the chunks with background music have low SNR value that will be removed with 15 dB SNR threshold. Fig 5 shows SNR results for Gujarati audio that has part of a clip in Hindi by another speaker. WADA-SNR's output depends on the quality of VAD also, as we see for Silero, the SNR values are low when compared to other parts of the audio but high in the case of WebRTC. Rejecting non-main speakers at SNR thresholding stage itself can improve the results of the speaker clustering phase also.

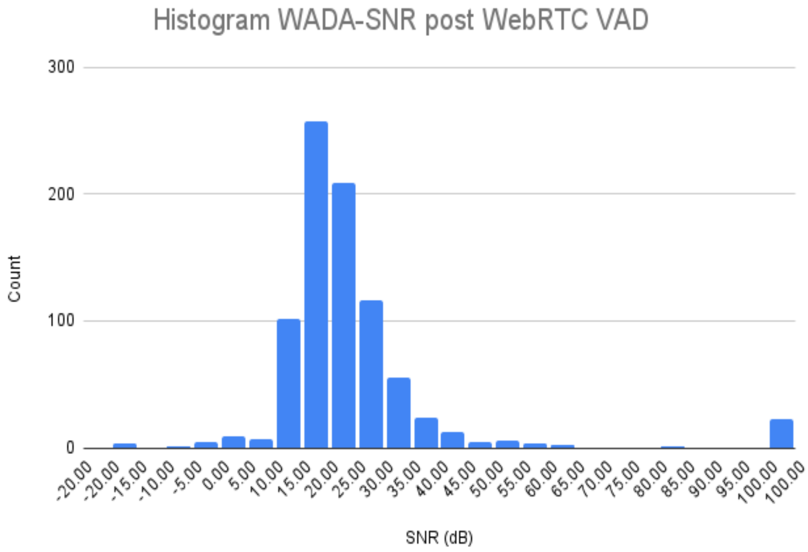


Figure 2: WADA-SNR histogram on 20 sample audio files after WebRTC VAD

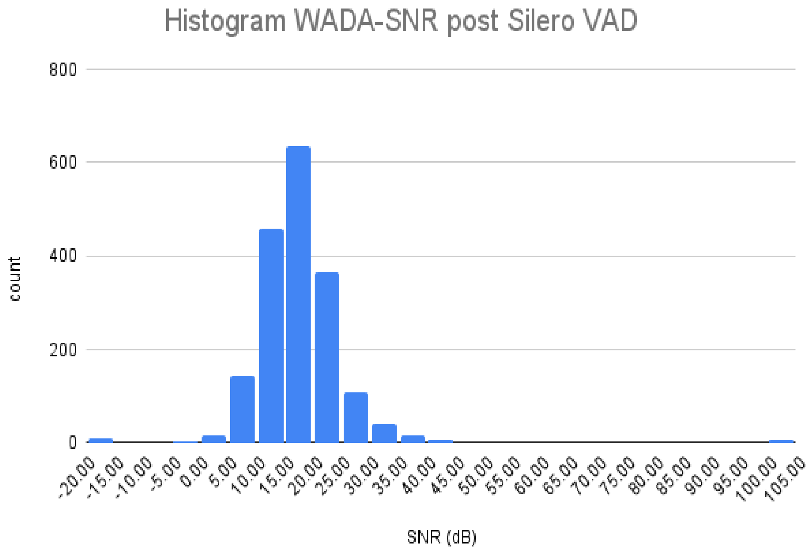
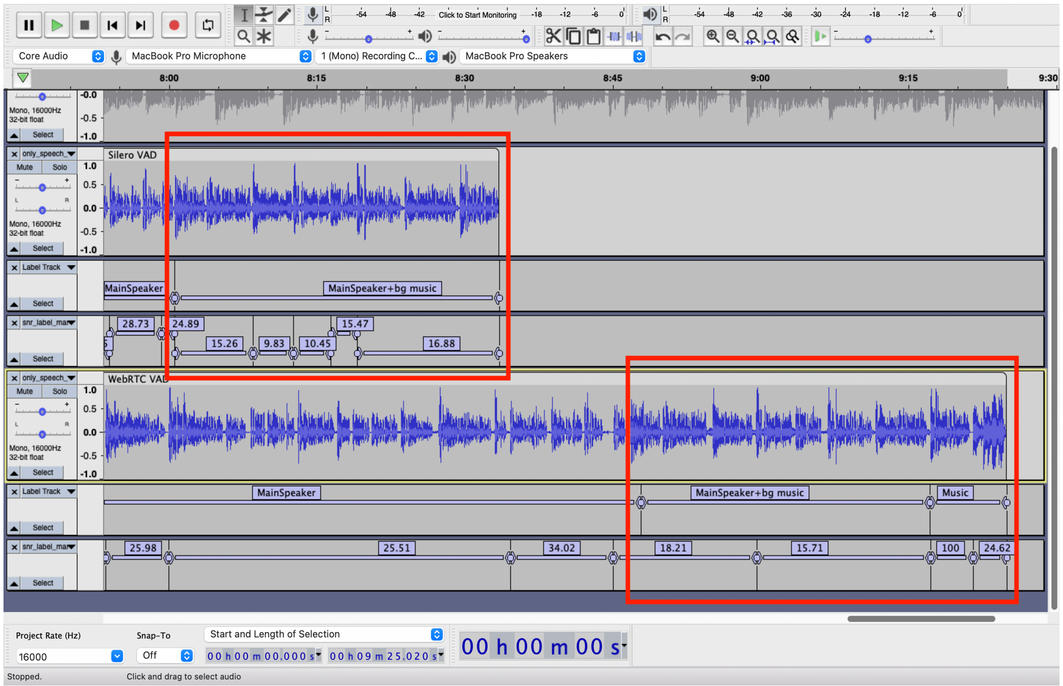


Figure 3: WADA-SNR histogram on 20 sample audio files after Silero VAD



Project Rate (Hz): 16000

Snap-To: Off

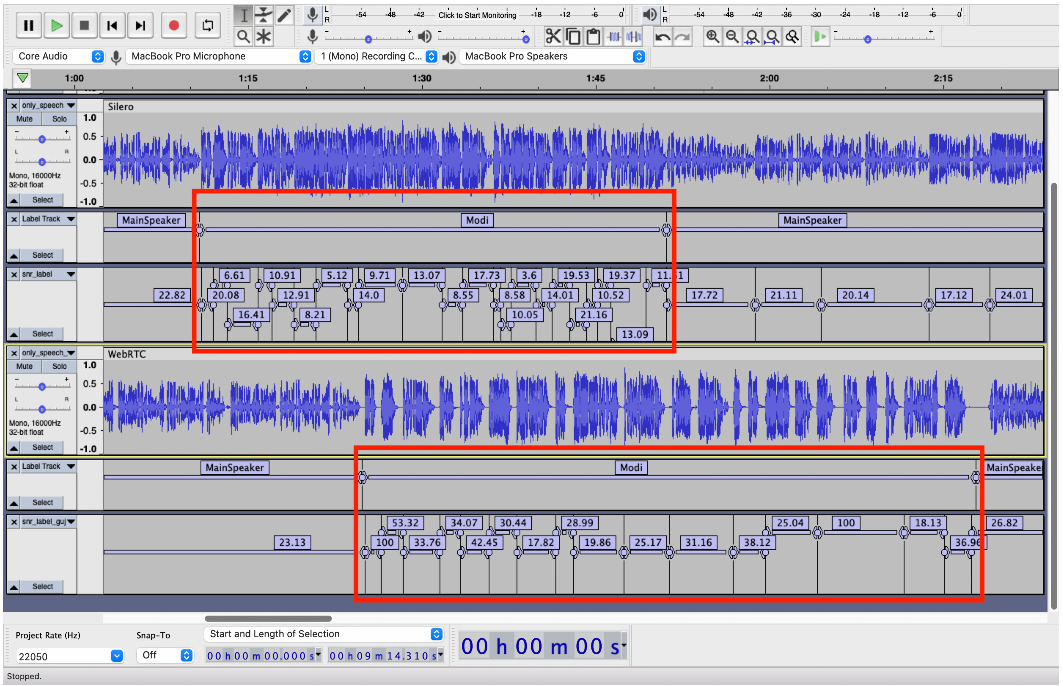
Start and Length of Selection: 00 h 00 m 00.000 s 00 h 09 m 25.020 s

00 h 00 m 00 s

Stopped.

Click and drag to select audio

Figure 4: Comparing Silero and WebRTC VAD after SNR estimation for Marathi audio



Project Rate (Hz): 22050

Snap-To: Off

Start and Length of Selection: 00 h 00 m 00.000 s 00 h 09 m 14.310 s

00 h 00 m 00 s

Stopped.

Figure 5: Comparing Silero and WebRTC VAD after SNR estimation for Gujarati audio

7 Speech-text alignment

After performing voice activity detection and SNR thresholding, the next stage is to get the text for the corresponding segmented audio from the entire transcript. In Vakyansh, the transcript generated from the existing STT model is directly used as labels for the filtered audio chunks. This can affect the model performance as the STT model itself may not be very accurate and have many grammatical errors. Moreover, good open-source STT implementation may not exist for low-resource languages. For News on Air bulletins, they just suggest forced alignment on unfiltered audio. We propose speech text alignment by first generating transcripts for individual pre-processed chunks using whisper[5]. The transcript is sentence tokenized using spaCy NLP library. The substring in the original sentence tokens with which generated transcript provides best matches is then used as the final text label for the audio segment.

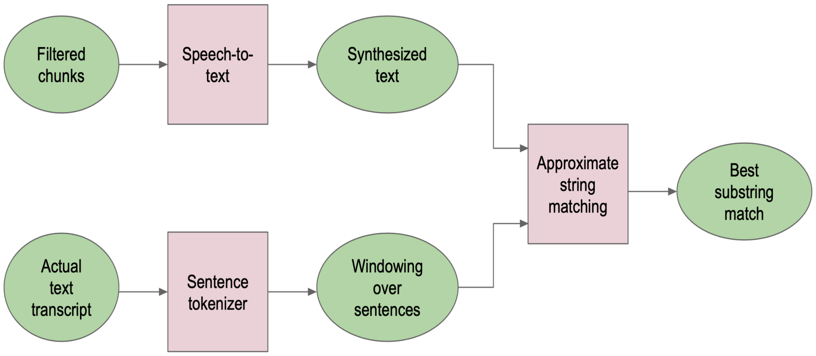


Figure 6: Speech-text alignment stages

7.1 Speech to text using whisper

Whisper is a multi-task speech processing model that is trained on a large, diverse dataset. It can perform multilingual speech recognition, speech translation, language detection, etc. We use whisper to transcribe Marathi and Gujarati audio segments. There are 5 different models available with different performance and memory requirements. The default small model provides reasonably good transcription. In fact, it gave better results than the larger medium model for Marathi. It confuses similar phones and is not very accurate in vowel diacritics. For our purpose, as long as the STT model does not produce completely wrong or garbage words, the transcription can be used for string matching. But using whisper as is without any changes in parameters resulted in repeated gibberish text for some audio segments. This is a failure mode of whisper, where the model tries to predict the next word while trying to transcribe the audio itself. As suggested by the authors, lowering the value default value of the compression ratio threshold parameter from 2.4 to 1.8 improved the results. Further, we were able to completely mitigate the looping by lowering the no_speech threshold to 0.4, as after VAD, we do not expect non-speech segments. Lastly, the parameter condition_on_previous_text had to be disabled as otherwise, the previous output of the model is used as a prompt for the next window. Setting this parameter off does not decrease the accuracy of transcription as our segments are 9 s, much less than 30 s window length.

Transcription using whisper	Actual text
बेंकान जा दिर्गखालिन कर्यगाडा जोखिम तारनेक्र ता शाश्वत लबाचा उपाएजुना ना चालना द्ऱा अवश्यक अस्लय ज़ा या अह्वालात मतला है	बँकांच्या दीर्घकालीन कार्यकाळात जोखीम टाळण्याकरता शाश्वत लाभाच्या उपाययोजनांना चालना देणं आवश्यक असल्याचं या अहवालात म्हटलं आहे
भारद्या बैंकेंग उद्ध्योगाची स्तितिट सानली होती	भारतीय बँकिंग उद्योगाची स्थिती चांगली होती
त्यास्म्रुते प्रित्ते त्यात्त भार्ताचा विका सात प्रदेशास्त्त भार्त्ये समुदायेचा योगदानाची दखालगेना से त्यार्वर्षी नुु जानेवरेला प्रावासी भार्ते दिव साज्रा किलाजातो	त्या च्या स्मृतीप्रित्यर्थ भारताच्या विकासात परदेशस्थ भारतीय समुदायाच्या योगदानाची दखल घेण्यासाठी दरवर्षी 9 जानेवारीला प्रवासी भारतीय दिवस साजरा केला जातो

Table 2: Comparison of transcript generated by whisper with original text for Marathi audio

7.2 Approximate string matching

After getting the synthesized transcript for each audio segment, we need to find which portion of text in the original transcript it best matches with. The generated transcript may contain words not present in the true transcript and can also be discontinuous due to SNR thresholding. But we know after Silero VAD, audio segments will not contain speech beyond a sentence due to natural pauses humans take between sentences. Therefore, instead of finding a match in the entire transcript, we can search for substring match in one sentence at a time. Since there can be a large number of errors in the generated transcript, doing a character-wise comparison is not useful. Doing an approximate match of a smaller string in a larger string has been widely used in spell checking, protein alignment, spam filtering, etc. We explored edit distanced based and matching subsequence lengths based techniques in this direction.

7.2.1 Edit distance based

Consider a pattern string $P = p_1 p_2 p_3 \cdots p_m$ and a text string $T = t_1 t_2 \cdots t_n$. Algorithms under this category try to find the substring $T_{jj'} = t_{j'} \cdots t_j$ in T that has the minimum edit distance with P. Python's regex search allows to specify the maximum number of errors (substitution + insertion + deletion) allowed to find the substring in a longer string. Needleman-Wunsch is another technique used to align protein or nucleotide sequences. It uses dynamic programming and finds an optimal global alignment. If the characters at an index match, then a score of 1 is assigned, and -1 if there is a mismatch or matches to a gap (insertion or deletion). In the end, it chooses the alignment having the maximum score. However, none of these produced desirable results as spaces are ignored, the synthesized transcript can have a single word broken into multiple, and with so many vowel diacritics in Devanagari as compared to English, makes the character level edit distance-based matching poor. This motivated us to look for approaches based on the similarity between strings that ignores character-level alignment.

7.2.2 Subsequence match based

Python's difflib module provides SequenceMatcher class to find all the contiguous matching subsequences. It finds the similarity between two strings as:

$$R = \frac{2 * M}{T} \tag{2}$$

where M is the number of matches and T is the total number of elements in both strings. It also allows ignoring user-specified chars while matching. Thus, spaces and punctuation can be ignored for better results. Let n be the number of words in a synthesized transcript for an audio segment. Then, for every sentence in the original transcript, we further tokenize it into words and apply a sliding window considering n-1, n, n+1 words at a time. The window across all sentences that results in the maximum similarity score is estimated as the approximated aligned text for the audio.

Table 3 displays results of transcription obtained using whisper and the best substring match found in the entire transcript with their similarity scores. The alignment produced is fairly accurate except for some cases. In a few instances, additional one or two words are included that are not actually present in the audio chunk. Even though we included n-1 words in our window size, the similarity score comes higher when these extra words are included because the total number of substring matches also increases. The last row is of an audio segment that does have the corresponding text in the original transcript. Thus, as expected, the similarity score is very less. We can apply a threshold on the similarity scores. This will ensure any misalignments or speech with no text labels are not passed as training data.

8 Conclusion and future work

In this work, we evaluated the pre-processing stages of the Vakyansh pipeline on the News on Air dataset. We further filled gaps in the performance by using other publicly available solutions. Using WebRTC for VAD, we could reject most non-speech audio segments and reduce the amount of clean data loss post-SNR thresholding. Text for the filtered audio segments is aligned using a similarity score between the strings.

The final speech-text alignment is still not 100% accurate, and by designing another cost metric to further align texts after using SeqeunceMatcher, results can be improved in the future. The alignment was performed in the text domain. It will be worth exploring sentence similarity between embeddings that ignore spelling errors and look only for semantic similarity. We did not explore speaker clustering in this work and can be taken in future. We believe it will further improve quality of training data and results post alignment by rejecting speech segments of non main speakers.

References

[1] H. Singh Chadha, A. Gupta, P. Shah, N. Chhimwal, A. Dhuriya, R. Gaur, and V. Raghavan, "Vakyansh: ASR Toolkit for Low Resource Indic languages," *arXiv e-prints*, p. arXiv:2203.16512, Mar. 2022.

Transcription using whisper	Aligned text	Similarity score
इंदेशातल्या बैंका प्रशासक्या दूष्त्या सक्शम होत अस्लया	देशातल्या बँका प्रशासकीयदृष्टया सक्षम होत असल्या तरी	0.6
मात्र 2013-14 नत्र याद माल्मत्त, गुणुवत्त, अन नफ्याचा दुष्टी कोना तुन गस्रन सुरुजाली, असर्या अह्वालात मत्ला है.	मात्र २०१३-२०१४ नंतर यात मालमत्ता, गुणवत्ता आणि नफ्याच्या दृष्टिकोनातून घसरण सुरू झाली असं या अहवालात म्हटलं आहे	0.71
सवरीं गोल्डबाश्टी दुस्री मालिका ये तेस उमर पसुन विक्री साटी खूली होनारसुन फि विक्री शुक्रवार परेंत शुरूर अनार है	सॉव्हरिन गोल्ड बॉड्सची दुसरी मालिका येत्या सोमवारपासून विक्रीसाठी साठी खुली होणार असून ही विक्री शुक्रवारपर्यंत सुरु राहणार आहे.	0.71
तरी समाजिक धुश्टी को तों अदिक सुदारना अवश्यक अज्लेचा	तरी सामाजिक दृष्टिकोनातून अधिक सुधारणा आवश्यक असल्याचं भारतीय रिझर्व बँकेच्या	0.63
किंद्रिया प्रुत्फी विद्यान अनी अवकाष, तेस बरवर विद्यान तन्त्यान विबागात्से राज्जा वन्त्री जी तिंद्रसी हा, आज दों दिव सांचा पुने दवरे अरहेद	केंद्रीय पृथ्वी विज्ञान आणि अवकाश त्याचबरोबर विज्ञान तंत्रज्ञान विभागाचे राज्यमंत्री जितेंद्र सिंह आज दोन दिवसांच्या पुणे दौऱ्यावर आहेत.	0.55
राजा सानात पाली इता रस्ते अबगाता चाथ जन ठार, तर चोभी सुन अदिखषक्मी, अनी जिम्भाभे वरोडवा तुसर्या एक दिश्या क्रिकेट सामने नाने प्रिक चिंकों भरताचाक शेटर अर्ख्रान चानिरने, या वरोडवराच अकष्मनी मुमभाई केंद	राजस्थानच्या पाली जिल्ह्यात काल रात्री झालेल्या एका भीषण रस्ते अपघातात किमान ७ जणांचा मृत्यू झाला असून २४ हुन अधिक लोक जखमी झाले आहेत.	0.25

Table 3: Comparison of transcript generated by whisper with estimated aligned text for Marathi sample audio

[2] P. Bharati, “News on all india radio,” https://newsonair.gov.in/RNU-NSD-Audio-Archive-Search.aspx, 2022.

[3] C. Kim and R. M. Stern, “Robust signal-to-noise ratio estimation based on waveform amplitude distribution analysis,” in *Interspeech*, 2008.

[4] S. Team, “Silero vad: pre-trained enterprise-grade voice activity detector (vad), number detector and language classifier,” https://github.com/snakers4/silero-vad, 2021.

[5] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, “Robust speech recognition via large-scale weak supervision,” *OpenAI Blog*, 2022.