Journal of Future Artificial Intelligence and Technologies

E-ISSN: 3048-3719

*(Research/Review) Article*

# Enhancing Audio Topic Segmentation with Unsupervised Learning Techniques: A Comparative Study with Traditional Approaches

**Aarav Minocha [1] and Renu Balyan [2]**

[1] William A. Shine Great Neck South High School; e-mail : aarav.minocha@gmail.com
[2] State University of New York, Old Westbury; e-mail : balyanr@oldwestbury.edu

**Abstract:** Topic identification and audio segmentation are critical in applications such as content analysis, automatic transcription, and information retrieval. It's especially helpful in educational settings where large lectures need to be segmented into specific topics for easier content organization and processing. Traditional methods that rely on transcribed speech data are often impractical, due to extensive manual intervention and the need for large datasets. To address these challenges we propose a novel approach integrating Unsupervised Term Discovery (UTD) with audio features such as the intensity, pitch, waveform, and Mel-Frequency Cepstral Coefficients (MFCC). UTD identifies and aligns phoneme-like units to the audio without relying on transcriptions and Dynamic Time Warping (DTW) normalizes and aligns the multimodal features to the audio. The combined semantic features are then segmented with K-means clustering, creating segments of audio. This approach effectively segments audio based on thematic content, improving the accuracy for topic identification when compared to conventional Automatic Speech Recognition (ASR) systems, as shown by reduced Word Error Rates (WER) and Character Error Rates (CER) with english and spanish audio. This method offers a more practical and efficient solution for analyzing and segmenting audio content by combining phonetic-like data with audio features.

**Keywords:** Topic Identification; Phonemes; Audio Segmentation; Speech Processing



## 1. Introduction

In recent years there has been an increasing use of artificial intelligence (AI) techniques for audio analysis and processing. AI has been used for audio processing such as word/phoneme transcription, keyword identification, and audio segmentation [1]. However AI often requires a large amount of labeled training data in the audio domain such as manual transcriptions [2]. While unlabeled training data is easy to collect, manually labeled training data oftentimes is tedious, time consuming and requires a lot of effort that is not feasible to achieve for large datasets [2]. We seek to address the non-availability of labelled data for supervised learning within audio processing by utilizing basic semantic processes of sound which do not require explicit labeling.

One common way that researchers get around the inaccessibility of labeled data is by generating automatic word transcriptions through ASR machine learning models such as Whisper, Kaldi, Vosk, and more [3]. ASR packages such as these take audio as input, process the audio signal, and output the transcripts (written form of words as spoken) from the input audio. While these tools/packages are useful but are often inaccurate, limited in the scope of actions they can perform, and unable to process large files over 25MB [3]. In response to these limitations we utilize features such as phonemes and sound/audio characteristics to get a more informative and accurate analysis of the audio.

Additionally ASR-based methods fall short when capturing more nuanced aspects of speech beyond simple word transcriptions. In many cases, these systems overlook subtle variations in tone, emphasis, and inflection which provide context to spoken word. Phonetic content of audio provides more information regarding the content of the audio and tone of the speaker. This is a more informative measure to extract from audio when compared to only words, which are unable to convey the intent and inflection of the speaker [4]. Utilizing phonemes can help machine learning models learn data more efficiently and more accurately as it makes it easier to understand the

relationship between words, by understanding the phonetic pronunciation of each word, and therefore can discern topics from audio with greater accuracy [5].

Characteristics/properties of audio/speech such as intensity, pitch, waveform, and Mel Frequency Cepstral Coefficients (MFCCs) seek to explain different aspects of sound [6]. Intensity measures the amplitude of sound waves, representing the "loudness" of the audio. Pitch represents the frequency of sound waves within the audio. Waveform allows for a visual of the audio's temporal structure and energy distribution. Finally, MFCC analyzes the short term power structure of the audio, based on a linear cosine transform of a log power spectrum on a nonlinear mel frequency scale, allowing for a measure of the timbre and textual characteristics of the audio.

This paper aims to perform audio topic identification by training ASR machine learning models on semantic properties of speech including the phonetic and characteristics of the audio. Topic identification identifies major focus areas of the audio to be able to more accurately discern the content and entities within the audio. This task is especially important in educational settings where students may need to rewatch recorded videos of lectures that are often multiple hours long in order to find an explanation for a concept they are struggling with [7]. Topic Identification would allow a student to sort through educational content, automatically categorized by topics, and find relevant materials quickly [8]. The goal of the study was to identify the most prevalent topics from English and Spanish audios extracted from an existing set of recorded videos in the health domain. This study forms a part of a bigger NSF-funded project that is developing a culturally sensitive health intelligent tutoring system (ITS) for the Hispanic population. In order to achieve the said goal, some of the research questions (RQ) that were answered in this study are:

RQ1: Can an unsupervised machine learning model accurately extract phonetic transcriptions from audio and align them temporally to audio?

RQ2: Does the inclusion of audio features with phonetic content allow for accurate topic identification?

## 2. Related Work

Traditional ASR systems like Whisper and Kaldi face challenges in handling long audio files and languages other than English. In contrast, recent research in unsupervised learning techniques has sought to overcome these limitations by leveraging phonetic and audio feature analysis for more effective topic segmentation without the need for transcriptions.

Latent Dirichlet Allocation (LDA) [9], has been utilized in research for topic modeling in textual data. LDA uncovers hidden thematic structures by representing documents as mixtures of topics, with each topic being a distribution over words. In recent years, LDA has been adapted for use in audio data by converting transcribed speech or extracted feature representations (e.g., word embeddings of phonemes) into a form amenable to topic modeling. Examples of this are in [10] and [11], which demonstrate converting audio signals into bag-of-words representations. [10] applies LDA to phonetic embeddings from audio segments, while [11] transforms acoustic features into textual tokens for effective audio topic clustering. .

Dynamic Time Warping (DTW) addresses the inherent variability in spoken language by measuring the similarity between two sequences that may vary in speed or timing, making it effective for aligning segments of audio that differ in speaking rate, tone, or pronunciation. A study demonstrated the effectiveness of DTW in aligning MFCC feature vectors extracted from speech signals with varying speaking rates and background noise [12]. By aligning MFCC with audio temporally, the study was able to achieve high word recognition accuracy by compensating for differences in speech timing.

Additional work in the literature has explored various techniques for audio topic identification. For instance, several studies have employed deep neural network architectures such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) to learn representations directly from raw audio features for effective topic segmentation [13]. Another paper proposed an end-to-end framework that integrates attention mechanisms with feature extraction to selectively focus on key audio segments for improved topic clustering [14].

In contrast to these approaches, our work integrates unsupervised term discovery (UTD) with both phonetic transcriptions and multiple audio features, including intensity, pitch, waveform, and MFCC, and aligns them using DTW before clustering with K-means. This framework minimizes the reliance

on extensive labeled data and captures a richer set of acoustic and linguistic cues. An integrated unsupervised approach for topic identification in audio such as this has not been widely explored, potentially offering improved performance, especially for long-format audio files, while relying on only inherent audio characteristics.

# 3. Methodology

## 3.1. Data Collection

11 English and 11 Spanish mp4 videos ranging from 16 seconds to 2 minutes and 33 seconds on cancer survivorship recorded by a doctor were used as the initial data for the study. Audio was extracted from these videos and stored as mp3 files. These extracted audio (mp3 files) were 32-bit dual channel.

## 3.2. Audio Preprocessing

The audio extracted from the videos were 32-bit dual channel mp3, while many ASR systems and packages require 16-bit mono-channel wav files [15]. Therefore these mp3 files were converted to wav format to make it compatible with these tools and easy to process. Any extraneous noise was also removed using audio intensity and pitch normalization using z-score normalization. This also allows for feature comparison across different audio files by bringing them to a common scale, with a mean of 0 and a standard deviation of 1. In addition, the files were dual channel as they were meant to be played on systems with left and right audio output channels as these were stereo wav files. For this reason, the audio files were converted to mono-channel wav files, and the sampling rate was changed to 16-bit.

## 3.3. Audio Feature Extraction

Four audio features including the waveform, intensity, pitch, and Mel Frequency Cepstral Coefficients (MFCC) were extracted from the audio to represent the inherent features of the audio data. First, each audio file was processed using the Librosa library to load the audio and process the signal's time series and sampling rate. The *librosa* python package is used to extract waveform data from the audio, representing the amplitude of the audio signal over time, and the silent intervals within the audio to identify when the speaker is not speaking [16]. The *parselmouth* python package is used to extract the intensity (dB) of the audio, measuring the energy of the sound in the audio [17]. These values were then normalized by excluding any non-positive values, which are often caused by noise or errors while reading the audio file. *Parselmouth* was also used to extract the pitch (Hz) of the audio, measuring the frequency of the audio, which is normalized by getting rid of zero values and abnormal values which could potentially be due to any background noise. Finally, MFCC was extracted also using *parselmouth*, providing information about the vocal tract of the audio and compressing it into a small number of coefficients, showing spectral characteristics and overall shape of the spectral envelope. MFCC contains about 10-20 spectral characteristics making it useful for machine learning and audio manipulation tasks such as Montreal Forced Aligner (MFA) [18].

## 3.4. Manual Word Transcriptions

Human transcriptions were created by a fluent English and Spanish speaker and the transcriptions were also verified and validated by another speaker fluent in both the languages. The expert word transcriptions were utilized to analyze the accuracy of the automatic phonetic transcriptions by matching the phonetic pronunciation of each word with the automatic phonetic transcription output. These transcriptions were generated using the exact words spoken in each video/audio. The descriptives (number of words and number of sentences) for the transcriptions generated from these videos/audios by the human expert are shown in Figure 1.

The data descriptives shown in Figure 1 for each transcription were obtained using spaCy, an open-source NLP python library that extracts different linguistic characteristics from text [19]. SpaCy utilizes tokenization to analyze the number of words in a text. SpaCy obtains the number of words from the transcripts by breaking down the transcript into individual tokens, where each token is an individual word, punctuation, or space. SpaCy analyzes the number of sentences in a text using token-based rules as well as a statistical model trained on annotated corpora, which allows it to segment the text into sentences and obtain the number of sentences from the transcripts.

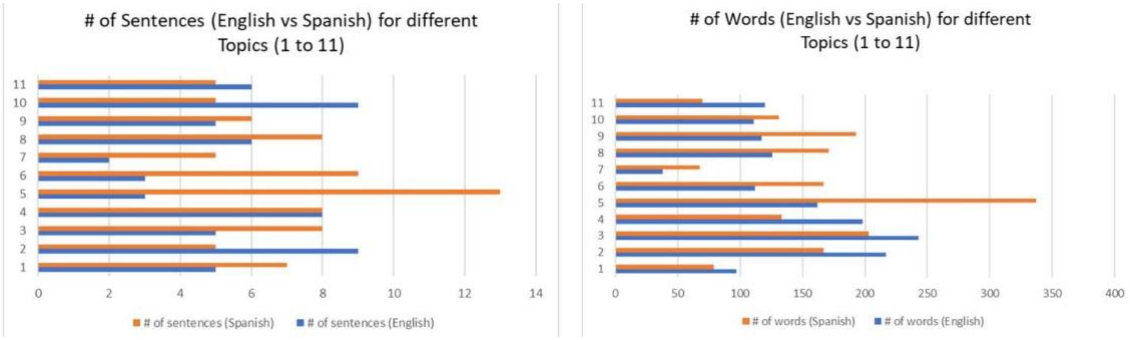


**Figure 1:** Data Descriptives for the English and Spanish Expert/Reference Transcriptions

There were differences between the data descriptives for English and Spanish transcriptions, even though the videos were on the same topics. Some of these differences result due to varying lengths of videos for the two languages, which leads to different numbers of sentences. In addition, the other differences are caused due to linguistic differences between the two languages. For example, a word in English may not have a single word equivalent in Spanish but is represented by multiple words or vice versa, which results in a difference between the number of words in the two language transcripts.

### 3.5. Automatic Phonetic Transcriptions

Automatic Phonetic Transcriptions were obtained and temporally aligned to their corresponding audio recordings using MFA. MFA has a pre-trained Acoustic Model which is trained on a dataset consisting of audio recordings and their corresponding phonetic transcriptions [20]. This model allows the MFA to recognize distinct phonemes from the audio. To create the phonetic transcriptions MFA extracts the MFCC of the audio signal, which is fed into the acoustic model so it can predict the phonemes of the audio. This allows it to create phonetic transcriptions for each audio file without requiring the manual word transcription.

The MFA then performs temporal alignment where it takes each individual phoneme and aligns it with its corresponding audio segment. It aligns the two by incrementally adjusting the timing of the phonemes until it matches the speech patterns of the audio precisely. The output of this process is a TextGrid file in Praat software consisting of the phonetic transcriptions and their timestamps in the audio [15]. The manual transcription was then added to the TextGrid file of each audio file to visualize each word and its corresponding phonetic pronunciation and when the occurred within the audio. An example output of this process is shown in Figure 4.

### 3.6. Audio-Feature Alignment

Even though both features (attributes of the audio and the phonetic transcriptions) were aligned to the same audio timeline, they varied due to the differing extraction methods and nature of their data, requiring it to normalize the data. The two semantic feature outputs were concatenated to form one feature vector using Dynamic Time Warping (DTW) [21]. DTW creates a cost matrix where each element (I, J) represents the cost of aligning the ith frame of audio features with the jth phoneme. The cost matrix is a grid where the rows represent the frames of audio features and the columns represent the phonetic segments. The cost of aligning each feature vector separately with the phonetic transcription, results in multiple cost matrices corresponding to each feature vector. The distance between each frame of the individual feature vectors and the phonetic transcription was calculated using the Euclidean Distance. This operation is performed independently for each cost matrix to find the optimal alignment path for each feature vector. The warping path, is the optimal alignment between the two sequences, is identified to specify which frames of the audio features correspond to which phonetic segments. The two features are then concatenated onto the same TextGrid file for visualization and UTD training as shown later in Figure 6 in the Results section.

### 3.7. Unsupervised Term Discovery (UTD)

UTD is a process that identifies recurring patterns in audio without relying on labeled data. It allows extraction of meaningful speech segments without requiring any data aside from the audio. After audio features and phonetic transcriptions were aligned LDA extracted topics from the concatenated files [8]. The synchronized data obtained from DTW was converted into a corpus of documents,

each representing a segment of aligned features and phonetic transcriptions. LDA typically requires word data to identify topics so we converted our phonetic transcripts into word transcripts utilizing clustering and word-phonetic pronunciation dictionaries.

First, K-means clustering categorized similar segments of the audio [22]. These segments contain groups of audio features (phonemes and corresponding speech/audio attributes) representing words. Then these groups of phonemes were converted to text using word pronunciation dictionaries. *CMU_Dict* was utilized for our English dictionary and a dictionary obtained from MFA was utilized for Spanish.

LDA performed topic identification through a multi-step process involving data loading, preprocessing, word tokenization, dictionary and corpus creation, and finally topic modeling [8]. LDA first loads the data from the directory containing the word transcriptions and reads each file into a dictionary where keys are file names and values are the content within the file. Stopwords are obtained and removed from the dictionary using the NLTK library, eliminating uninformative words for the analysis. Then the spaCy library is used to lemmatize the text to reduce words to their root forms, leaving behind only the specific parts of speech (nouns, verbs, adjectives, and adverbs). The lemmatized text is tokenized into individual words using the Gensim package, removing punctuation and converting text to lowercase [23]. Finally, a corpus is created from the modified dictionary where each document is represented as a bag-of-words where each word is replaced by its own unique id and frequency count to measure how many times it occurs in the audio. The final output is the most prevalent topics and their corresponding words.

### 3.8. Evaluation

To analyze if the automatic word clustering produced accurate word transcriptions the Word Error Rate (WER) and Character Error Rate (CER) were calculated and compared to other automatic word transcription packages (Whisper) to see if the performance improved (Figures 7 and 8 in the Results section). Whisper is a general-purpose, multitasking speech recognition model by OpenAI, trained on a large dataset of diverse audio that can perform several tasks including language identification, speech translation, and multilingual speech recognition [24]. Whisper models are trained on 680,000 hours of labeled audio and the corresponding transcripts collected from the internet. This training data constitutes 65% (i.e., 438,000 hours) of English audio and the matching English transcripts; roughly 18% (i.e., 125,000 hours) represents X→ English translation data, and the remaining 17% (i.e., 117,000 hours) represents non-English audio and the corresponding transcript, covering 96 other languages, including Spanish. As it scales well, the model was trained using an encoder-decoder transformer [25]. The overall methodology is shown in Figure 2.

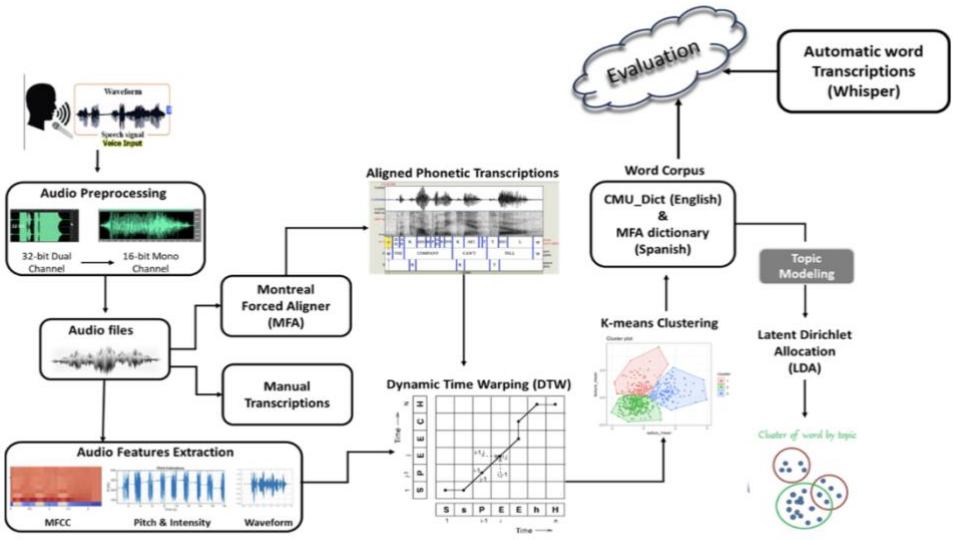

**Figure 2:** The Overall Workflow Methodology

# 4. Results

## 4.1. RQ1: Can an unsupervised machine learning model accurately extract phonetic transcriptions from audio and align them temporally to audio?

### 4.1.1 Automatic Phonetic Transcription

Automatic phonetic transcriptions for all 22 audio files were generated using MFA. The output was transcriptions consisting of the phonetic pronunciation of all words identified and used by MFA. Figure 3 shows an example of the phonetic pronunciation transcription produced by MFA alongside a corresponding manual word transcription.

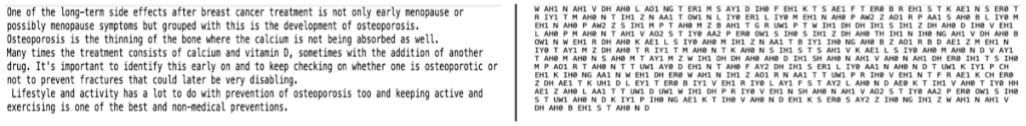


**Figure 3:** Manual word transcription for Osteoporosis discussed in the audio/video (left). Automatic phonetic transcription for the same topic using the MFA (right).

The numbers after the vowel syllables indicate the stress of the syllable. 0 indicates that the syllable is unstressed, 1 indicates primary stress, meaning the syllable is emphasized in pronunciation, and 2 indicates secondary stress, meaning the syllable isn't the primary one being emphasized but is still emphasized more than unstressed syllables.

### 4.1.2 Audio-Phonetic Alignment

Phonetic transcriptions were aligned with their corresponding audio temporally using MFA. This resulted in a TextGrid displaying phonemes and when they occurred in the audio. Manual transcriptions were also added to the TextGrid to visualize when the phonemes occurred compared to spoken word. Figure 4 shows an example of a TextGrid loaded with phonemes and manual word transcriptions.

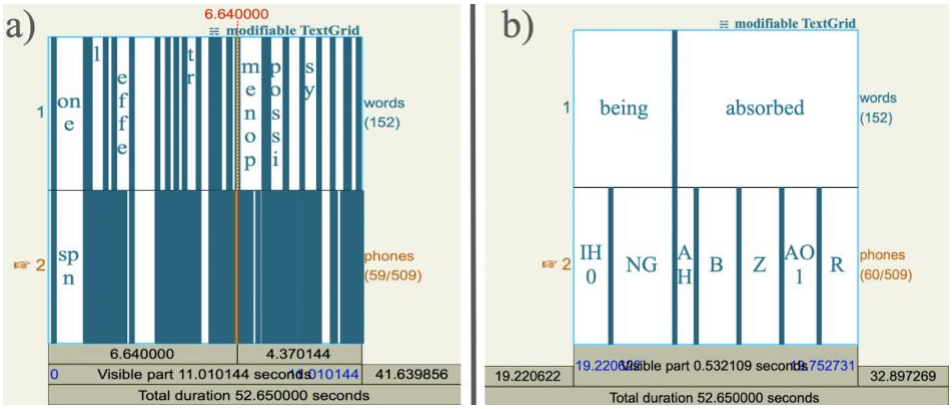


**Figure 4:** TextGrid audio for Osteoporosis loaded into Praat software showing the words and corresponding phonemes temporally aligned with the audio using MFA to show when the word or phoneme was spoken. a) multiple words and corresponding phonetic amounts b) zoomed in on two words to see the specific phonetic pronunciations

## 4.2. RQ2: Does the inclusion of audio features with phonetic content allow for accurate topic identification?

### 4.2.1 Audio Feature Extraction

The semantic properties of the audio were extracted using Librosa and Parselmouth python libraries. The features extracted from the audio were the waveform, intensity, pitch, and MFCC. These audio

features displayed information about the audio such as the speaker's infliction and volume as well as the frequency and amplitude. These semantic features were extracted for all audio files and used along with the phonetic transcriptions to create automatic transcriptions. This was done by clustering similar feature points, consisting of the audio features and phonetic transcriptions, in an audio file through K-means clustering of all data points and creating a final transcription through phonetic-word dictionaries. Figure 5 shows values for the waveform, intensity, pitch, and MFCC extracted from a file that was plotted using MatLab.

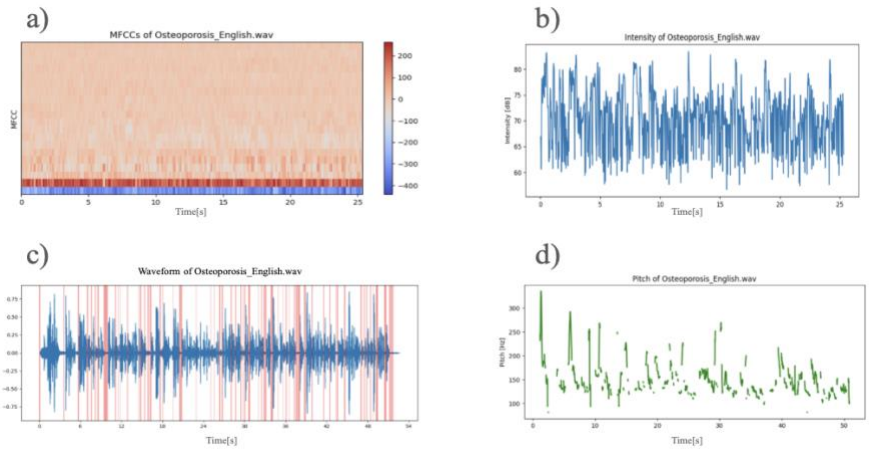


**Figure 5:** Audio feature graphs for the Osteoporosis audio. a) MFCC of the audio signal. The colors represent the amplitude of the cepstral coefficients at different times and frequencies b) Intensity of the audio signal c) Waveform of the audio signal with pauses overlaid onto the graph. d) Pitch of the audio signal

Pauses are overlaid on the waveform graph to show when the speaker was not speaking. Areas of pauses are susceptible to background noise corrupting phonetic extraction. The pitch graph has numerous gaps instead of being a continuous function. This is because the pitch graph had erroneous values which were too high or too low which were removed as these corresponded to pauses, which potentially caused the abnormal values.

### 4.2.2 Feature Alignment

DTW was used to align the audio features and phonetic transcriptions. First, a cost matrix was created which represented the cost of a frame of an audio feature with a phoneme. A cost matrix was generated for all 4 audio features (Intensity, Pitch, Waveform, and MFCC). The distance between the phonetic transcriptions and feature vectors was calculated to find the optimal alignment between the two modalities of data. Figure 6 shows an example TextGrid file with all features aligned alongside each other to normalize data.

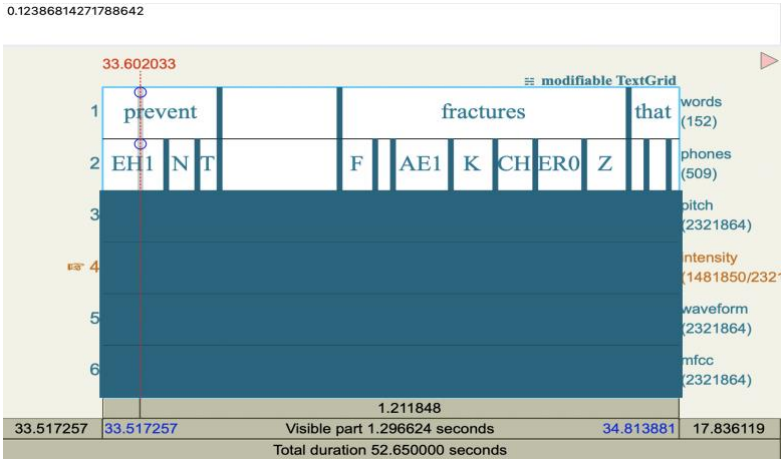

**Figure 6:** TexGrid for Osteoporosis after DTW to combine the phonetic features and audio features (Intensity, Pitch, Waveform, and MFCC). The intensity measurement tier is selected (number 4 highlighted in orange) and the number on the top represents the intensity measurement for the audio at that time stamp.

### 4.2.3 ASR Evaluation

K-means clustering segmented the audio into various clusters, with each segment/cluster consisting of audio with phonetic transcriptions and audio features. These segments were then utilized to create word transcriptions as the phonemes within these segments were converted to words using word-phoneme dictionaries. This allowed us to analyze the accuracy of segmenting the audio based on multiple feature types and utilizing the phonemes within the segments to form word transcriptions. These word transcriptions were compared with ASR systems which are able to take in raw audio and output word transcriptions [3]. Both automatic transcriptions, proposed approach and ASR approach, were compared to the manual word transcriptions to measure the accuracy of the two methods.

ASR systems are typically evaluated on certain metrics to measure their usefulness and effectiveness. The purpose of evaluating ASR systems is to simulate human judgment of the performance of the systems in order to measure their usefulness and assess the remaining difficulties especially when comparing systems; the standard metric of ASR evaluation is the Word Error Rate (WER), which is defined as the proportion of word errors to words processed [3]. The WER is based on how much the output (typically a string of words) called the Hypothesis, returned by the ASR system differs from a reference transcription generated by a human expert. The WER is computed using equation (1)

$$WER = \frac{S+D+I}{S+D+C} = \frac{S+D+I}{N} \quad (1)$$

Where I = number of insertions, D = number of deletions, S = number of substitutions, C = number of correct words and N = number of words in the reference.

The Python Jiwer package was used to automatically calculate the WER. The measures are computed with the use of the minimum edit distance between one or more reference and hypothesis sentences. Although WER is the most popular and commonly used metric to evaluate ASR, it has certain drawbacks [26]. Therefore, the Character Error Rate (CER) was also calculated. The CER value indicates the percentage of characters that were incorrectly predicted [27]. A lower CER value indicates a better performance of the ASR system with a CER of 0 being a perfect score. The CER is computed using equation (2):

$$CER = \frac{S+D+I}{S+D+C} = \frac{S+D+I}{N} \quad (2)$$

Where S = the number of substitutions, D = the number of deletions, I = the number of insertions, C = the number of correct characters, and N = the number of characters in the reference.

## Transcriptions Error Rates

Figure 7 shows the evaluation metric results for English Transcription performed by Whisper using the 'medium-en' model. The evaluation metrics include the percentages returned by the JiWER package for the WER and CER for automatic transcriptions and the proposed approach. The average WER for English transcriptions is 22.93%, and CER is 4.56% for the automatic Whisper generated transcriptions. Whereas the WER and CER for our proposed approach are slightly less and hence better, and are 21.51%, and 4.26% respectively. The average percentages for the Whisper generated transcriptions for Spanish were WER (24.45%) and CER (8.37%). On the other hand the error rate percentages for our proposed approach for Spanish (Figure 8) were better and lower than the Whisper transcription such that WER was 21.73% and the CER was 5.98%. Overall it was found that the error rates for Spanish were slightly higher as compared to English transcription as the existing ASRs have been trained on a large English corpora and much smaller corpora for other languages.

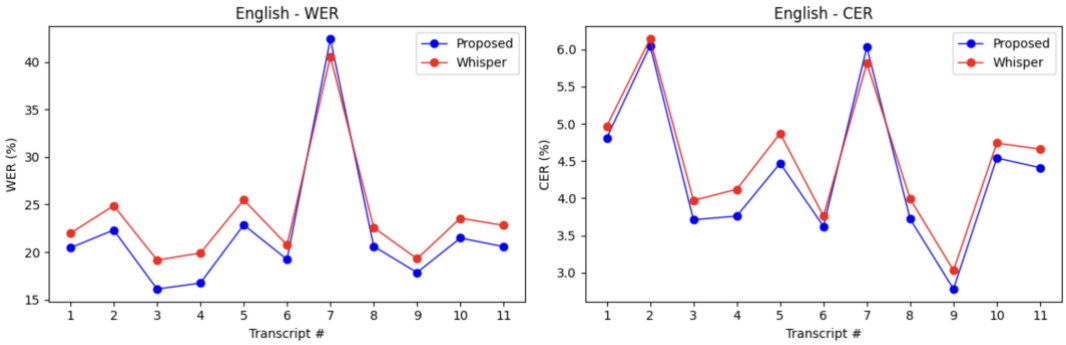


**Figure 7:** Graphical representations comparing the WER and CER between our proposed approach and Whisper for English audio.

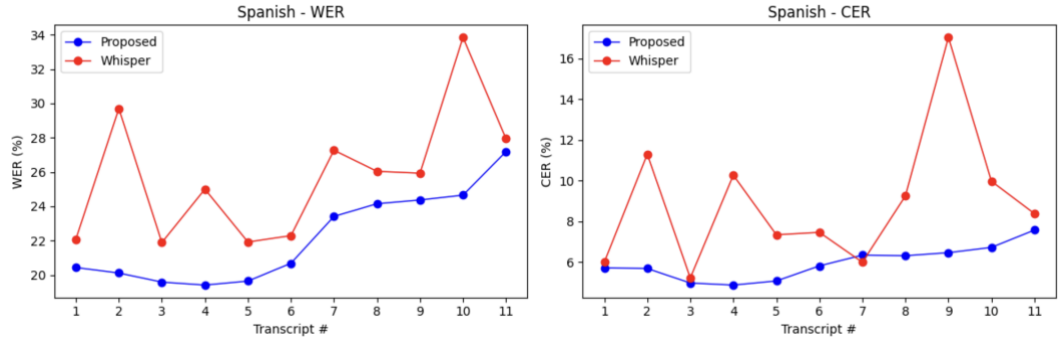


**Figure 8:** Graphical representations comparing the WER and CER between our proposed approach and Whisper for Spanish audio.

We observed that our proposed semantic approach outperforms the Whisper model in all the transcription's accuracy metrics except for WER and CER in transcript 7 for English and CER only in transcript 7 for Spanish. This is most likely because transcript 7 is a significantly shorter audio file than the others (0:17 and 0:33 respectively) and Whisper has high performance for short audio files. Whisper performed similarly to our approach on shorter audio files as our proposed approach was more inaccurate towards the beginning of each audio file, as it didn't pick up on many of the phonemes. Longer audios had more accurate phonetic transcriptions as they had accurate phonetic transcriptions for a larger percentage of their runtime. Our proposed semantic approach greatly improved the WER and CER for Spanish audios as Whisper is largely inaccurate for non-English audios as it tries to directly extract word transcriptions from audio [28]. In addition to this the Spanish audio files were longer on average, which made our approach perform better.

### 4.2.5 Unsupervised Term Discovery

Automatic word transcriptions obtained using our proposed approach were fed into LDA for term/topic identification. The LDA returns the most prevalent topics in the audio and their associated words. Figure 9 shows the topics identified for all audio files formatted with pyLDAvis to visually represent topics, their words, and any overlap between topics [29]. The pyLDAvis outputs not only show the number of topics (8 for Spanish and 10 for English audios) but also reveal the relationships between topics through overlapping clusters of keywords. Many topics appear to capture major themes related to health survivorship and treatment, indicating that the model is successfully isolating key content areas from the audio. However, some topics exhibit significant overlap, suggesting that certain themes are interrelated or that the boundaries between topics are not entirely distinct. In some cases, the presence of less informative or generic terms points to challenges in pre-processing, such as incomplete stopword removal or limitations in the phoneme-to-word conversion step. While the approach effectively segments and identifies dominant themes, further refinement in feature integration and lexical normalization could enhance topic specificity and reduce redundancy.

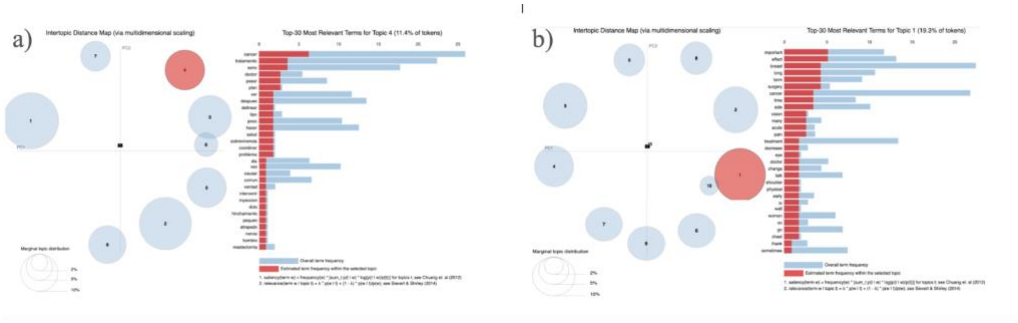


**Figure 9:** LDA outputs graphically visualized using pyLDAvis for a) spanish audios with 8 topics and b) English audios with 10 topics. The relevant terms associated with each topic are shown on the right with overlapping topic clusters representing that the topic has overlapping relevant terms with another.

## 5. Discussions & Conclusions

In this study we developed a novel framework to identify topics within audio without the need for labeled data. To analyze the effectiveness of our framework we had two research questions discussed in the Introduction Section.

### 5.1 RQ1: Can machine learning accurately extract phonetic transcriptions from audio and align them temporally to audio?

Several machine learning models capable of extracting phonemes from audio were explored, including PocketSphinx, Kaldi, DeepSpeech, and MFA. PocketSphinx, Kaldi, and MFA are able to extract phonemes without corresponding word transcriptions, meaning they were viable for unsupervised learning. MFA was able to align extracted phonemes with audio, allowing for insight into when phonemes occurred in the audio. MFA was used for phoneme extraction and alignment as training the model for both processes would allow for a more in depth analysis of the data. MFA was able to accurately extract phonemes and align them with audio in a TextGrid file. This allowed for further data integration since the audio feature data was aligned into the same file for data normalization. The methodology using MFA successfully captured phonetic details and synchronized them with audio features, even though performance on the initial segments of audio files was somewhat less accurate due to blending with background noise. We did not find an alternative approach that outperformed our proposed method across all conditions. The integration of additional noise-handling strategies could further improve the results. Overall, our experiments confirm that our approach is effective for the intended task, with some scope for optimization in handling very short or noisy audio segments.

### 5.2 RQ2: Does the inclusion of audio features with phonetic content allow for accurate topic identification?

Audio Features (Waveform, Pitch, Intensity, and MFCC) were extracted alongside phonetic transcriptions. Further integration of the two data was done by individually aligning each to the audio, allowing for a more detailed analysis of the feature in the context of the audio. In addition to this, the two modalities of data (numerical and textual) were combined into a TextGrid file using DTW, resulting in one file containing both modalities of data for each audio. After these features were combined, K-means clustering was performed on each file, resulting in segments of the audio. These segments were then converted to word transcriptions through translation with word-phonetic dictionaries.

This approach proved to be much more accurate at extracting word transcriptions compared to traditional approaches (Average WER improvement of 10.7% for English transcripts and 10.40% for Spanish transcripts). This was due to the integration of two data modalities from the audio, allowing for a more informative analysis of the audio to extract word transcriptions. To extract topics from the audio the word transcriptions were fed into a LDA model, resulting in the extraction of topics and their most associated words from all audio files. This approach was able to filter out unnecessary words and create more informative topics as it was performed on more accurate word transcriptions.

### 5.3 Implications for Educational Settings

Automatically identifying major topics within audio opens up the potential for a more seamless navigation of educational environments, where students oftentimes have to watch through long lessons to find topics they are struggling with. The issue with most traditional topic identification systems is that they rely heavily on automatic word transcriptions, which are not only impersonal but can also be highly inaccurate, especially in noisy or poor-quality recordings. These systems often fail to capture the nuances of speech, such as tone and intent, which are crucial for understanding the context and relevance of the spoken content [30]. Additionally, if educators want to utilize a supervised system for a more personalized topic identification suited specifically to the type of course, they would need to provide extensive datasets of manually labeled transcriptions to train the models, which are not readily available to most educators and institutions due to the significant time and cost involved in creating these datasets [31]. Our approach leverages unsupervised learning techniques to bypass the need for these labeled transcriptions, utilizing phonetic and audio feature analysis to accurately identify topics within the audio. This method not only enhances the accuracy and personalization of topic identification but also is able to capture parts of speech such as tone, emphasis, and inflection, making our approach much more effective for topic identification within educational settings.

### 5.4. Limitations

One limitation we encountered was that MFA had difficulty transcribing phonemes for the beginning of a sentence i.e., the first one or two words of the audio. While reviewing the audio we found that this was because the audio would cut in, making the first word blend in with the background, potentially causing it to be labeled as background noise by the MFA model. This limitation led to shorter audios having higher WER and CER than the audios that were longer.

Another limitation we encountered was that pauses in the audio, where the speaker is silent or not speaking, led to very abnormal pitch values. This could have been caused by parselmouth plotting background noise for those moments instead of the speaker, who was silent. To overcome this limitation these erroneous values were omitted, however, this could have impacted pitch analysis as numerous spots had gaps in place of the abnormal values.

The final limitation we encountered was that there was no metric which could accurately measure the accuracy of the UTD as term identification is subjective. There is no objective way to extract topics from audio due to inherent bias. One speaker may believe a speaker is emphasizing one topic while another believes another topic is being emphasized.

**Funding:** This research was funded by the National Science Foundation (NSF), grant numbers 2219587 and 2318636.

**Data Availability Statement:** The videos from which the audio data was extracted and transcriptions generated can not be shared publicly due to health videos created by a medical professional as a part of another grant. However, the code created to perform the experiments in this study will be shared on request to perform analysis on their own datasets and will be made available on Github as the project moves into its final stages.

**Conflicts of Interest:** The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

## References

[1] Kheddar, H., Himeur, Y., Al-Maadeed, S., Amira, A., & Bensaali, F. (2023). Deep transfer learning for automatic speech recognition: Towards better generalization. *Knowledge-based Systems*, *277*, 110851. https://doi.org/10.1016/j.knosys.2023.110851

[2] Kamper, H., Jansen, A., & Goldwater, S. (2017). A segmental framework for fully-unsupervised large-vocabulary speech recognition. *Computer Speech & Language*, *46*, 154–174. https://doi.org/10.1016/j.csl.2017.04.008

[3] Ma, H., Peng, Z., Shao, M., Li, J., & Liu, J. (2023, December 13). *Extending Whisper with prompt tuning to target-speaker ASR*. arXiv.org. doi: https://doi.org/10.48550/arXiv.2312.08079

[4] Xue, W., Van Hout, R., Cucchiarini, C., & Strik, H. (2023). Assessing speech intelligibility of pathological speech in sentences and word lists: The contribution of phoneme-level measures. *Journal of Communication Disorders*, *102*, 106301. https://doi.org/10.1016/j.jcomdis.2023.106301

[5] Muzammel, M., Salam, H., Hoffmann, Y., Chetouani, M., & Othmani, A. (2020). AudVowelConsNet: A phoneme-level based deep CNN architecture for clinical depression diagnosis. *Machine Learning With Applications*, *2*, 100005. https://doi.org/10.1016/j.mlwa.2020.100005

[6] Vasconcelos, D., Nunes, N. J., Förster, A., & Gomes, J. P. (2024). Optimal 2D audio features estimation for a lightweight application in mosquitoes species: Ecoacoustics detection and classification purposes. *Computers in Biology and Medicine*, *168*, 107787. https://doi.org/10.1016/j.compbiomed.2023.107787

[7] Joglekar, A., Espejo, I. L., Hansen, J. H. L., & Aalborg Universitet. (2024). Fearless Steps APOLLO: Identifying Conversational Mission-Critical Topics in NASA Apollo Missions Audio based on keyword spotting. In *2024 NASA Human Research Program IWS*. doi: https://doi.org/10.21437/Interspeech.2018-1942

[8] H. Meinedo and J. Neto (2023). Audio segmentation, classification and clustering in a broadcast news task. 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. (ICASSP '03)., Hong Kong, China, 2003, pp. II-5, doi: 10.1109/ICASSP.2003.1202280.

[9] Blei, David & Ng, Andrew & Jordan, Michael. (2001). Latent Dirichlet Allocation. The Journal of Machine Learning Research. 3. 601-608. https://www.researchgate.net/publication/221620547_Latent_Dirichlet_Allocation

[10] T. -H. Le, P. Gilberton and N. Q. K. Duong, "Discriminate Natural versus Loudspeaker Emitted Speech," *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, 2019, pp. 501-505, doi: 10.1109/ICASSP.2019.8683227.

[11] Su-Youn Yoon, Klaus Zechner, Combining human and automated scores for the improved assessment of non-native speech, Speech Communication, Volume 93, 2017, Pages 43-52, ISSN 0167-6393, doi: 10.1016/j.specom.2017.08.001.

[12] Bhadragiri Jagan Mohan and Ramesh Babu N., "Speech recognition using MFCC and DTW," *2014 International Conference on Advances in Electrical Engineering (ICAEE)*, Vellore, India, 2014, pp. 1-4, doi: 10.1109/ICAEE.2014.6838564.

[13] Eleni Tsalera, Andreas Papadakis, Maria Samarakou, and Ioannis Voyiatzis. 2023. CNN-based Segmentation and Classification of Sound Streams under realistic conditions. In Proceedings of the 26th Pan-Hellenic Conference on Informatics (PCI '22). Association for Computing Machinery, New York, NY, USA, 373–378. doi: 10.1145/3575879.3576020

[14] Bahdanau, D., Cho, K., & Bengio, Y. (2014). *Neural Machine Translation by Jointly Learning to Align and Translate.* doi: 10.48550/1409.0473

[15] K. R. Lekshmi and E. Sherly, "An ASR System for Malayalam Short Stories using Deep Neural Network in KALDI," *2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)*, Coimbatore, India, 2021, pp. 972-979, doi: 10.1109/ICAIS50930.2021.9395945.

[16] McFee, B., Raffel, C., Liang, D., Ellis, D. P. W., McVicar, M., Battenberg, E., & Nieto, O. (2015). *librosa: Audio and Music Signal Analysis in Python.* Proceedings of the 14th Python in Science Conference, 18-24. doi: 10.25080/Majora-7b98e3ed-003

[17] Jadoul, Y., Thompson, B., & De Boer, B. (2018). Introducing Parselmouth: A Python interface to Praat. *Journal of Phonetics*, *71*, 1–15. https://doi.org/10.1016/j.wocn.2018.07.001

[18] Muda, L., Begam, M., & Elamvazuthi, I. (2010, March 22). *Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques*. doi: https://doi.org/10.48550/arXiv.1003.4083

[19] F. N. A. Al Omran and C. Treude, "Choosing an NLP Library for Analyzing Software Documentation: A Systematic Literature Review and a Series of Experiments," *2017 IEEE/ACM*

*14th International Conference on Mining Software Repositories (MSR)*, Buenos Aires, Argentina, 2017, pp. 187-197, doi: 10.1109/MSR.2017.42.

[20] McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. *INTERSPEECH 2017*. https://doi.org/10.21437/interspeech.2017-1386

[21] Boersma, Paul & Weenink, David. (2007). PRAAT: Doing phonetics by computer (Version 5.3.51). https://www.researchgate.net/publication/259810776_PRAAT_Doing_phonetics_by_computer_Version_5351

[22] Permanasari, Y., Harahap, E. H., & Ali, E. P. (2019). Speech recognition using Dynamic Time Warping (DTW). *Journal of Physics. Conference Series*, *1366*(1), 012091. https://doi.org/10.1088/1742-6596/1366/1/012091

[23] Kamper, Herman & Livescu, Karen & Goldwater, Sharon. (2017). An embedded segmental K-means model for unsupervised segmentation and clustering of speech. 719-726. doi: https://doi.org/10.48550/arXiv.1703.08135

[24] Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). Robust Speech Recognition via Large-Scale Weak Supervision. doi: https://doi.org/10.48550/arXiv.2212.04356

[25] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30. doi: https://doi.org/10.48550/arXiv.1706.03762

[26] *Automatic text summarization using Gensim Word2VEC and K-Means clustering algorithm*. (2020, June 5). IEEE Conference Publication. doi: 10.1109/TENSYMP50017.2020.9230670.

[27] Piotr Szyma´nski, Piotr ´Zelasko, Mikolaj Morzy, Adrian Szymczak, Marzena ´Zyła-Hoppe, Joanna Banaszczak, Lukasz Augustyniak, Jan Mizgajski, and Yishay Carmiel. 2020. WER we are and WER we think we are. In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 3290–3295, Online. Association for Computational Linguistics. doi: https://doi.org/10.48550/arXiv.2010.03432

[28] TouchMetrics, https://torchmetrics.readthedocs.io

[29] Mangsor, N. S. M. N., Nasir, S. a. M., Abdul-Rahman, S., & Ismail, Z. (2023). Identifying topic modeling technique in evaluating textual datasets. In *Lecture notes on data engineering and communications technologies* (pp. 507–521). doi: https://doi.org/10.1007/978-981-99-0741-0_36

[30] Dua, M., Aggarwal, R. K., & Biswas, M. (2021). Speaker recognition using noise robust features and LSTM-RNN. *Progress in Advanced Computing and Intelligent Engineering*. doi: https://doi.org/10.1007/978-981-33-4299-6_2

[31] Li, D., Gao, Y., Zhu, C., Wang, Q., & Wang, R. (2023). Improving speech recognition performance in noisy environments by enhancing lip reading accuracy. *Sensors*. doi: https://doi.org/10.3390/s23042053