Investigating Speaker Diarization on Code-Switching Speech Executive Summary

Bronston Ashford z5146619

Justification for Thesis

Code-switching (CS) speech, the act of alternating between two or more languages in or between sentences, is a prevalent phenomenon in most multilingual societies [1, 2]. As the world becomes increasingly interconnected, speech technologies have an increasing demand to interface with these non-standard forms of speech. Among these technologies, speaker diarization (SD) systems, which identify 'who spoke when' in an audio clip, play a significant part.

SD systems serve as an important component in the pipeline of various speech technologies, such as speech recognition or meeting transcription. Since the quality of upstream SD directly impacts the performance of the parent technologies, audio recordings containing CS speech may be at a disadvantage compared to monolingual speech [2, 20]. Therefore, mitigating errors introduced by CS in audio would contribute to better accessibility and understanding in multilingual societies.

However, developing robust SD systems that can effectively handle CS speech requires an understanding of how much and to what extent CS speech impacts these systems. Yet, due to a notable lack of research in this domain, the question remains unexplored. Furthermore, contributing to the lack of research is the absence of freely available SD datasets containing CS speech. The quality of research on diarizing CS speech is intrinsically tied to the quality of available datasets, the generation of which can be challenging and time-consuming. Therefore, the addition of a free CS dataset to the field stands to facilitate future research and contribute to the development of this domain.

This thesis intends to contribute by initiating an investigation into the impact of CS speech on SD systems with a specific language pair. It aims to offer initial insights into how much SD systems are influenced by CS speech and to identify which aspects of CS speech present the greatest challenges. In doing so, a new CS dataset fit for diarization will be developed and made freely accessible.

Objectives

The primary objectives of this thesis are to curate a dataset for CS and to study the impact of CS on modern SD systems. To achieve these objectives, the MIAMI corpus [3], a resource not initially designed for speech processing, will be adapted. This adaptation requires a conversion from the original *Codes for the Human Analysis of Transcripts* (CHAT) format to *Rich Transcription Time Marked* (RTTM) format for SD evaluation. Since no software provides this conversion, a custom program will need to be made using python. Additionally, CHAT transcriptions frequently present CS speech within single timestamp segments. For simplified analysis, these timestamps will be manually edited to ensure each segment contains only monolingual speech. The final dataset will be categorised into sections of English, Spanish, and English-Spanish code-switched speech.

The SD system will first be tested on the dataset's monolingual files to establish a baseline, followed by a comparative analysis of the error rate on the CS section. A requirement for conducting these tests is some foundational machine learning theory. Subsequently, a software tool will be developed to compute CS characteristics within the dataset. These metrics will then be correlated with diarization errors to understand the influence of CS on SD performance.

Literature Review

SD systems primarily come in two key architectures: a modular approach or an end-to-end strategy. The modular approach segments the diarization task, initially extracting embeddings, followed by clustering to assign speaker labels. Conversely, the end-to-end strategy combines these steps into one process [2]. Because of their ability to handle varying numbers of speakers, modular design tends to outperform in diarization tasks, including the gold-standard benchmark for diarization performance, the DIHARD challenges [4]. Consequently, a survey of available systems reveals a tendency towards this architecture. In particular the Pyannote diarization system. The competitive results achieved by this open-source system's pre-trained model on the demanding DIHARD benchmark [5] reflect contemporary SD capabilities, making it a viable choice for this thesis.

The performance of SD systems on CS speech is a field that is largely under-researched. The only literature found directly addressing this issue is that of the inaugural 2023 DISPLACE challenge. The paper introducing the challenge implies that SD systems perform poorly during CS speech, highlighting the need for further research. Their aim is to draw attention to this issue by developing a CS dataset and establishing a benchmark [6]. Despite this, some evidence suggests that SD systems might not be significantly affected by CS speech. Investigations into the MFCC characteristics during a speaker's language switch shows a consistent formant structure [7], which are often used by SD systems for input embedding. This could imply that the issue is less severe than anticipated. However, due to the limited research in this area, the proficiency of SD systems in CS speech remains unclear.

A significant barrier to exploring diarization in CS speech is the lack of CS datasets. As the vast majority of datasets curated for speech processing tasks are strictly monolingual, the number of CS datasets is severely limited. Moreover, many of the existing datasets are not publicly available [1, 9]. To effectively analyse SD system performance on CS data, it's important to understand the nature of the CS in the dataset. Metrics such as the I-index, for quantifying language switching frequency, and burstiness, for irregularities, play an important role in this process [8].

Preliminary Work

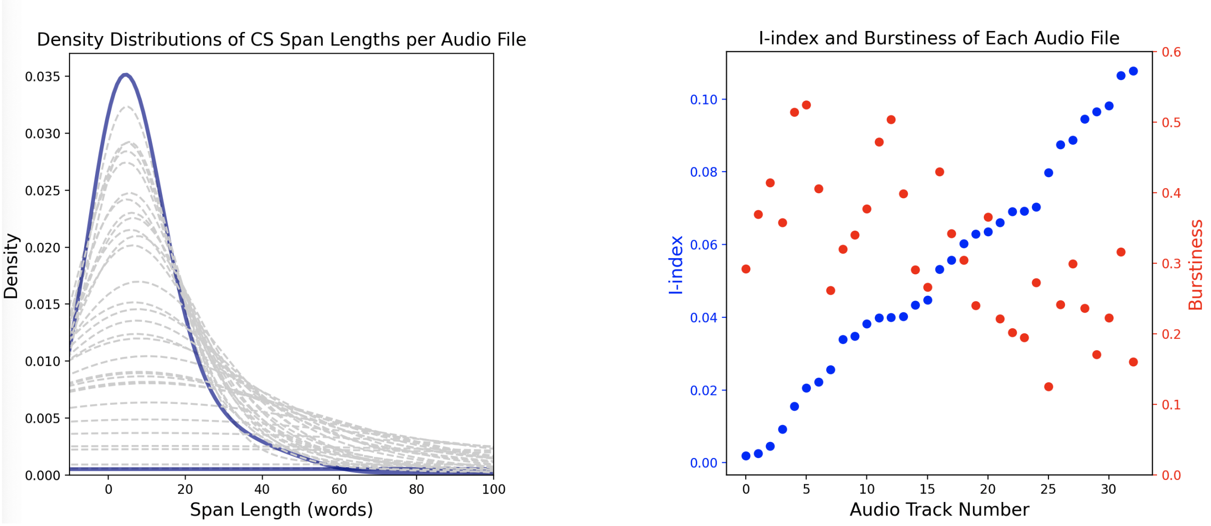
For the task of developing a new dataset suitable for testing and evaluating SD systems, the *Corpus: MIAMI*, a CS corpus originally designed for linguistic studies were identified to form the basis of the new dataset. For this dataset to be used for the purpose of the thesis, it needed to contain both English-only and Spanish-only audio tracks. This would enable the establishment of a diarization baseline. In addition, speaker labels and timestamps were required for the creation of RTTM files, and timestamps marking language change points were necessary to carry out CS metrics assessments. The MIAMI corpus contained the relevant information but in a CHAT file format. As there is no software that can make the conversion from CHAT to RTTM a program had to be designed to automate this.

In the creation of this software, the `Annotate` data structure from the *Pyannote Metrics* Python library was used as the foundational structure. This choice was driven by its alignment with RTTM file formats and its built-in methods for manipulating audio file annotations. This structure was extended to handle language labels along with speaker labels, which required a deeper understanding of Python class structures. The software was designed to perform several tasks: extract timestamp information and speaker labels from CHAT transcriptions, identify missing timestamp segments, split multilingual segments into monolingual ones, and output RTTM files. However, manual correction of modified timestamps from splitting multilingual segments is necessary, creating a significant bottleneck in the process.

Despite the manual corrections of timestamps being incomplete, analysis on the CS characteristics can still be performed at the word level. A separate Python program was made to measure the words between language switch points, as seen in figure 1a. The distributions in the figure give an indication to how long, or short utterances per language tend to be per audio transcription. This information helps in the selection of audio files to convert into monolingual tracks. Tracks with flatter peaks generally consist of longer utterances in a particular language, which eases their conversion to monolingual tracks. Approximately 2 hours of Spanish and 3.5 hours of English content have been marked for this conversion process, leaving 12 hours of CS speech.

In addition, the CS metrics i-index and burstiness were calculated for each audio track, as seen in figure 1b. The i-index of the corpus has a large variation in language switching frequency distributed evenly from the highest being 50 times more frequent than the lowest audio track. In terms of burstiness, the speech patterns display a uniform distribution of values from low to moderate, reflecting varying degrees of cluster-based language switching behaviour within the tracks.

The Pyannote diarization system was set up on the Katana servers, and a trial was conducted using a fully converted audio segment from the MIAMI set. This trial run resulted in a 16% diarization error rate when evaluated using the Pyannote Metrics library. Subsequent work focused on gaining a deeper understanding of Pyannote Metrics' inner workings, which aided in the development of tools for in-depth analysis of diarization output. A Python program was created to isolate the errors in the SD system's output, enabling an examination of the language being used and the speaker during the occurrence of these errors. This tool will be useful in exploring the impact of language switching on SD systems.



(a) Span length distributions per transcription.

(b) I-index and Burstiness for each audio track.

Figure 1: Characteristics of the dataset.

Plans

Building upon the tools developed in thesis B, the conversion of the MIAMI corpus into a dataset suitable for SD can now commence. The immediate next step is to carry out the manual timestamp corrections required to complete the conversion. A 20-minute segment took approximately 1 hour to convert. Therefore, the conversion of a 16-hour dataset is projected to take around 50 hours. This work will be undertaken during the term break, prior to Thesis C. Once the dataset is ready, the diarization will be performed on the monolingual audio tracks to establish a baseline, then applied to the CS dataset to understand the impacts of CS on diarization.

The diarization error rates can then be compared with the i-index and burstiness of the tracks to identify potential correlations between language switching frequency and error rates. Lastly, a specially developed tool will be used to examine the precise points of error in the audio to determine whether errors predominantly occur during a language switch. The findings and any conclusions will be documented and presented in the final poster and report.

Meeting Log

Bibliography

[1] S. Sitaram, K. R. Chandu, S. K. Rallabandi, and A. W. Black. A Survey of Code-switched Speech and Language Processing. [Online]. Available: http://arxiv.org/abs/1904.00784

[2] T. J. Park, N. Kanda, D. Dimitriadis, K. J. Han, S. Watanabe, and S. Narayanan. A Review of Speaker Diarization: Recent Advances with Deep Learning. [Online]. Available: http://arxiv.org/abs/2101.09624

[3] M. Deuchar, "The Miami Corpus: Documentation File." [Online]. Available: http://bangortalk.org.uk/speakers.php?c=miami

[4] H.-S. Heo, J.-w. Jung, Y. Kwon, Y. J. Kim, J. Huh, J. S. Chung, and B.-J. Lee, "NAVER CLOVA SUBMISSION TO THE THIRD DIHARD CHALLENGE."

[5] Pyannote/speaker-diarization · Hugging Face. [Online]. Available: https://huggingface.co/pyannote/speaker-diarization

[6] S. Baghel, S. Ramoji, Sidharth, R. H, P. Singh, S. Jain, P. R. Chowdhuri, K. Kulkarni, S. Padhi, D. Vijayasenan, and S. Ganapathy. DISPLACE Challenge: DIarization of SPeaker and LAnguage in Conversational Environments. [Online]. Available: http://arxiv.org/abs/2303.00830

[7] J. Mishra and S. R. Mahadeva Prasanna, "Challenges in spoken language diarization in code-switched scenario," in *2023 National Conference on Communications (NCC)*, pp. 1–6.

[8] G. Guzmán, J. Ricard, J. Serigos, B. E. Bullock, and A. J. Toribio, "Metrics for Modeling Code-Switching Across Corpora," in *Interspeech 2017*. ISCA, pp. 67–71. [Online]. Available: https://www.isca-speech.org/archive/interspeech_2017/guzman17_interspeech.html