Unsupervised spoken content mismatch detection for automatic data validation under Indian context for building HCI systems

Thesis submitted in partial fulfilment
of the requirements for the degree of

Master of Science in ***Computer Science and Engineering*** *by Research*

by

Nayan Anand
2021701014
`nayan.anand@research.iiit.ac.in`

International Institute of Information Technology
Hyderabad - 500 032, INDIA
June 2024

International Institute of Information Technology

Hyderabad, India

CERTIFICATE

It is certified that the work contained in this thesis, titled "Unsupervised spoken content mismatch detection for automatic data validation under Indian context for building HCI systems" by Nayan Anand, has been carried out under my supervision and is not submitted elsewhere for a degree.

June 2024

Advisor: Prof. Chiranjeevi Yarra

To my Friends and Family

Acknowledgments

I am immensely grateful to all those who have been part of my research journey over the past few years at IIIT Hyderabad —a journey that has been not only academically enriching but also filled with personal growth and enduring memories.

First and foremost, I extend my deepest gratitude to Professor Chiranjeevi Yarra. His unwavering support and guidance have been instrumental in my achievements and in reaching heights I had previously thought unattainable. Professor Chiranjeevi Yarra has been more than a mentor; he has been a cornerstone in my academic and personal development. His patience in addressing my myriad questions and his willingness to engage deeply with my ideas have profoundly shaped my journey. For this, I am eternally thankful.

A special thank you goes to my amazing friends who have always stood by my side through thick and thin: Aditya, Amruth, Bhoomeendra, Dhaval, Dhruv, Nancy and Prateek. These individuals have made the arduous process of research and my time at IIIT Hyderabad enjoyable. I cannot thank you all enough for filling my life with memories that I will forever cherish.

Furthermore, I would like to extend my sincere gratitude to all my dear labmates who have always been just a call away. I will really miss the fun and frolic time spent with them; may it be regarding academic, co-curricular, extra-curricular, or personal discussions. They have not only played a pivotal role in enriching my research understanding but have also helped me evolve as a person.

I must also express my heartfelt appreciation to my family. Their constant encouragement and emotional support have been my guiding light through uncharted territories of research. Their belief in me has been a constant source of motivation to push forward, even in the toughest times.

To all of you, I owe a debt of gratitude that mere words cannot express. Thank you for being part of this incredible journey. You have all been the real heroes of this academic journey, and I am privileged to have had you all by my side.

Abstract

This thesis explores the critical challenges and provides solutions associated with automatic spoken data validation in the complex multilingual and multicultural context of India, which is crucial for developing efficient human-computer interaction (HCI) systems such as automatic speech recognition (ASR) and Text-to-speech synthesis (TTS). The diversity in linguistic backgrounds and the prevalence of non-native language speakers create unique challenges in speech communication. These challenges are exacerbated by the frequent mismatches between recorded speech and its reference text, referred to as misspoken utterances.

To tackle some of these challenges, this work introduces novel unsupervised techniques for detecting spoken content mismatches. The developed methods leverage state-of-the-art self-supervised speech representation models such as Wav2Vec-2.0 and HuBERT, integrating them with Dynamic Time Warping (DTW) as well as its variants such as Phone level cost maximised DTW approach (Ph-DTW), and Phone level cost maximised weighted DTW approach (Ph-WDTW) along with cross-attention mechanisms. This work develops and tests the techniques on specially curated datasets such as IIITH MM2 Speech-Text and Indic TIMIT, which include a wide variety of phonetic and linguistic features reflective of India's language diversity.

The methodologies proposed are rigorously evaluated for their effectiveness in improving the accuracy and efficiency of spoken data validation in an unsupervised manner. The results demonstrate significant advancements in the automatic detection of mismatches, thereby enhancing the reliability of speech data for training sophisticated HCI systems. By reducing the reliance on labour-intensive manual validation processes, these approaches significantly contribute to the scalability of speech data processing.

Overall, this thesis not only addresses a significant gap in the technological handling of spoken data validation but also sets a foundation for future research and development in speech technology applications within diverse linguistic landscapes. The implications of this work are broad, offering potential

improvements in various data-intensive speech applications such as ASR, TTS, and Computer-aided language learning systems (CALL) to name a few. This would be achieved by ensuring a readily accessible clean training, testing, and validation set for the development of target models for the aforementioned use-cases, thus addressing the reliable data scarcity to a great extent.

Contents

Chapter Page

List of Tables

List of Figures

Chapter 1

Introduction

The non-native signatures are common in the spoken language within societies where there is no common native language used for communication. Particularly in India, despite language diversity, often, English is used as the language of communication in administration, law, education and the workplace. In addition to English, Hindi is also used as the common language to connect the people at work. However, both languages are non-native to the majority of the Indian population. Though both languages are learnt during schooling and often, mainly English, with the help of spoken language training centres, a research report shows that $\sim$ 47% of graduates are not employable due to lack of English language skills. Because of low language proficiency, non-native signatures often exist in ones' spoken language; hence these variations limit human-computer interaction (HCI) due to the errors caused.

In order to achieve better speech-based HCI, it is required to have good quality speech data, which is typically obtained with a reliable validation process. The validation process involves justifying the spoken content in the audio with the text uttered by the speakers while recording. Mostly, this has been done manually using a group of annotators who listen to the audio and justify its spoken content with the text. In case of any word mismatches, the annotators correct the text to match the spoken content. This way, they ensure that the spoken content in the audio always aligns with the corresponding reference text. Similarly, the automatic validation process targets outcomes similar to those of manual validation.

However, in regions with a large language/accent diversity like India, non-native variations limit the validation of speech data. Hence, manual data validation is often considered to obtain reliable speech data. The manual validation is costly and cumbersome, and it limits the scalability of speech data. In the literature, it has been observed that the size of publicly available reliable Indian speakers' speech data is less compared to the speech data of native English and Mandarin speakers.

On the other hand, the need for reliable speech data is growing exponentially due to the current deep learning era for achieving significant performance. In order to cater to this growing demand for speech data, it is required to have automatic speech data validation methods, which could reduce manual intervention and increase scalability.

Spoken data validation in the Indian context is presented with a twofold challenge. Firstly, adapting the standard validation methods to India's linguistic diversity is difficult. This is because these methods were initially designed for less varied languages and accents, making them less effective in the Indian scenario. Hence it become non-trivial to adopt these methods for the Indian context. Secondly, it is complex to identify spoken content mismatches in languages non-native to the speaker. This complexity arises not only from word mismatches but also from accent variabilities, including lexical and acoustic differences, compounded by the speaker's language competency influenced by their native background.

In India, these challenges are amplified due to the country's vast language diversity. With a multitude of regional languages and dialects, each with its own unique phonetic and lexical characteristics, the process of validating speech data requires a highly nuanced approach. Due to this, there arises a pressing need for developing more sophisticated and context-sensitive approaches that can effectively address the complexities introduced by India's linguistic and cultural diversity. This situation calls for new, more tailored methods that can understand and cope with the diverse linguistic landscape of India.

Hence, to address these challenges, this thesis aims to detect spoken content mismatches in a non-native context for automatic data validation in the speech data collection process focusing on HCI applications, including automatic speech recognition (ASR) and text-to-speech synthesis (TTS). The mismatches are under read speech conditions between recorded audio and its corresponding reference text that is used as a prompt for the recording. The proposed work could save the time and cost involved in the annotation and the correction of the spoken content. Also, it would speed up the process of data preparation for training systems like ASR and TTS.

1.1 Motivation

The primary goal of the traditional spoken data validation method is to verify that the words that are spoken in an audio recording correspond precisely to the reference text, i.e. the text intended to be spoken. This is typically accomplished by a group of annotators working together who listen to the audio and manually validate the spoken content. In case of mismatches between the spoken content and reference text, the reference text is edited to ensure that the speech and its corresponding transcription match.

This manual approach, despite being successful, is quite a time and resource-exhaustive process and is thus inefficient. To address this challenge and increase scalability and productivity, there is a need to develop automated systems for spoken data validation. However, developing such systems is a challenging task, especially when the language spoken in the recordings is not the speaker's native language. In many circumstances, the intricacy extends beyond simply recognizing absent or wrong words; it also includes comprehending and interpreting varied dialects, the subtleties of lexical choice (the specific words used), and acoustic nuances (how the word is pronounced). Additionally, the speaker's overall ability in the language has a significant impact on these elements, thus enhancing the complexity even further. In the Indian context, these challenges are intensified even further due to the country's rich linguistic diversity. Each language and dialect brings its unique characteristics, making it harder to create a one-size-fits-all automatic spoken data validation approach.

However, the development of a successful automatic data validation approach would have substantial benefits. It would drastically reduce the time and resources currently needed. Moreover, it would eliminate the potential for human error, which is an inherent risk in any manually-intensive process. For researchers working with diverse Indian languages, this would offer a reliable and efficient way to ensure the accuracy of speech data.

1.1.1 Motivation for Automatic Data Validation

The advent of deep learning era has marked an increase in the data demand for training models to develop various speech applications. This trend is most noticeable in applications like automatic speech recognition (ASR) and text-to-speech (TTS). This increasing demand for training data could be met by the collection of extensive speech corpora. However, data collection in itself is a job half done. The raw recorded corpora cannot be directly used for training the models as they may contain errors that adversely impact the performance of model [1, 2, 3, 4].

In order to utilize the collected corpora for achieving better performance on the models trained, it needs to be refined to ensure good quality. Hence, the recorded corpus needs to be validated prior to the model training.

Data validation refers to cleaning the raw data to remove erroneous records. It has been established that Data validation improves the training data quality and hence the quality of the model trained using the validated data improves as well. The validation, typically done manually, ensures the absence of errors within the recordings while also verifying the congruence between utterances and their corresponding transcripts.

In [5], the authors emphasized the benefit of data validation for end-to-end model training. Furthermore, in [6], the authors observed that adding more training data does not necessarily lead to better performance; instead, it can be achieved with high-quality (validated) training data. This further highlights the need for data validation.

1.1.2 Motivation for Automatic Spoken Data Validation

Manual validation, despite providing us with reliable, high-quality data, is costly and cumbersome [7] and limits the data's scalability. In this process, a group of annotators listen to the audio and justify its spoken content with the text. In case of any word mismatches, they correct the text to match the spoken content, ensuring the speech-text pair used for model training is accurate. However, at times, they may discard a speech utterance if they are not suited to the purpose of the recorded corpora.

Some of the criteria for discarding maybe that the spoken content is either noisy or unintelligible. The discarding includes these cases but is not limited to them. For example, very long, noisy or unintelligible utterances would be discarded from a speech corpora designed for training ASR or TTS models. It should be understood that removing an utterance includes, but is not limited to, these aforementioned conditions.

Hence, In order to meet this growing data demand and overcome the drawbacks of manual validation, it needs to be performed automatically [8, 9], referred to as automatic spoken data validation. This would make the data validation process faster and scalable.

1.2 Current spoken data demand scenarios

In order to better understand the data demand trend and ensure that automatic data validation is indeed required, we examine the dataset size to analyze data demand trend in the ASR domain reported in 204 research papers [1] published in reputed conferences like Interspeech, ICASSP, NEURIPS, ASRU etc over the span of 2016 to 2021. For the analysis, we consider an equal number of papers per year.

[1] https://tinyurl.com/surveypaper1

Ensuring equal distribution of papers per year, the averaged data size (in hours) is shown in Figure 1 considering the following four cases:

1. All ASR types, 2. End-to-End (E2E) ASRs only, 3. DNN-HMM-based ASR only, and 4. GMM-HMM-based ASR only. Considering all ASR types, it is observed that it follows an overall increasing trend with the highest data duration of ∼30000 hours for the year 2021.

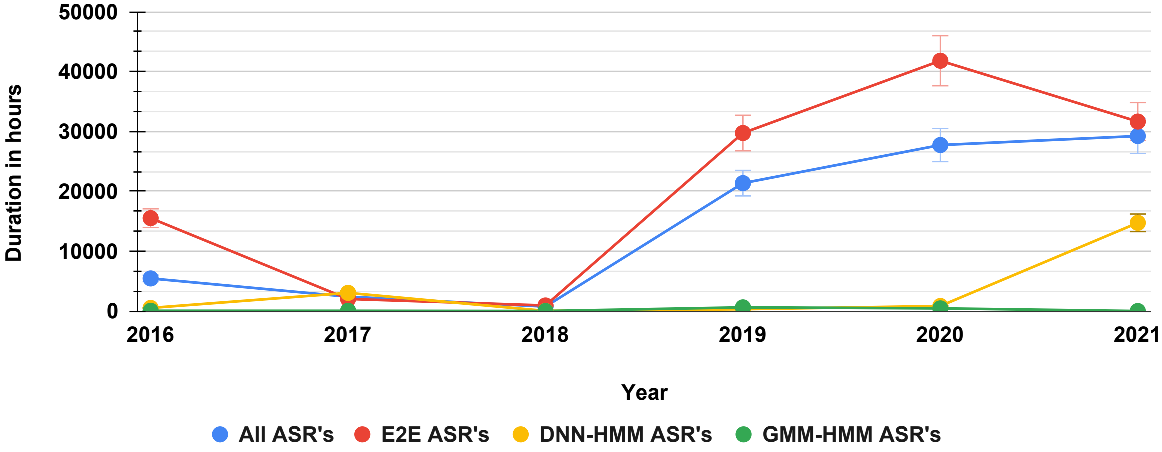


Figure 1.1: Data demand trend for training different types of ASR models from 2016 to 2021.

A comparable pattern is noticeable in the case of E2E ASR, with the highest data size of ∼40,000 hours in the year 2020. Compared to E2E ASRs, lower data is considered for DNN and GMM-based ASRs, but the trend has been increasing over the years. This observation indicates the increasing demand of the data size over the years. Similar observations are found in [10] where the authors stated that a larger amount of training data is required for effective training of neural-based language models as compared to traditional language models. This increasing data demand trend is found to be consistent when the data considered for building ASR by authors of academia and industry, as shown in Figure 2 and Figure 3, respectively. These observations align with the findings in [11]. The authors highlighted that rapid growth in data availability has made data demand. Similar observations were made in [12, 13] where the authors found that the large labelled datasets contributed significantly towards the accelerated success of deep learning-based methods. Similar observations were emphasized in [14] and [15]. The aforementioned observations emphasize the increasing demand for high-quality data. This demand can be achieved quickly, considering automatic data validation methods.

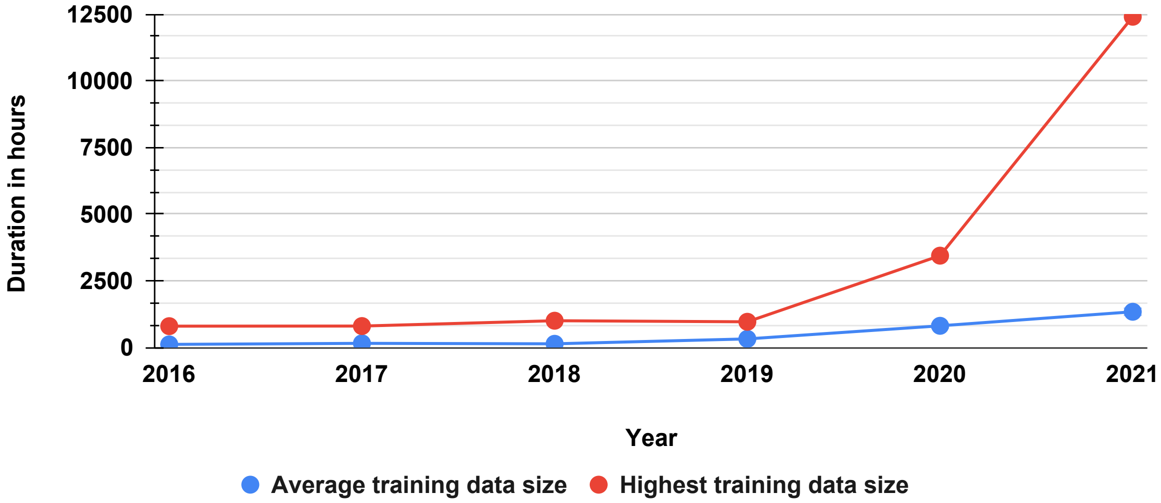


Figure 1.2: *Academia's* data demand trend in training ASR models from 2016 to 2021.

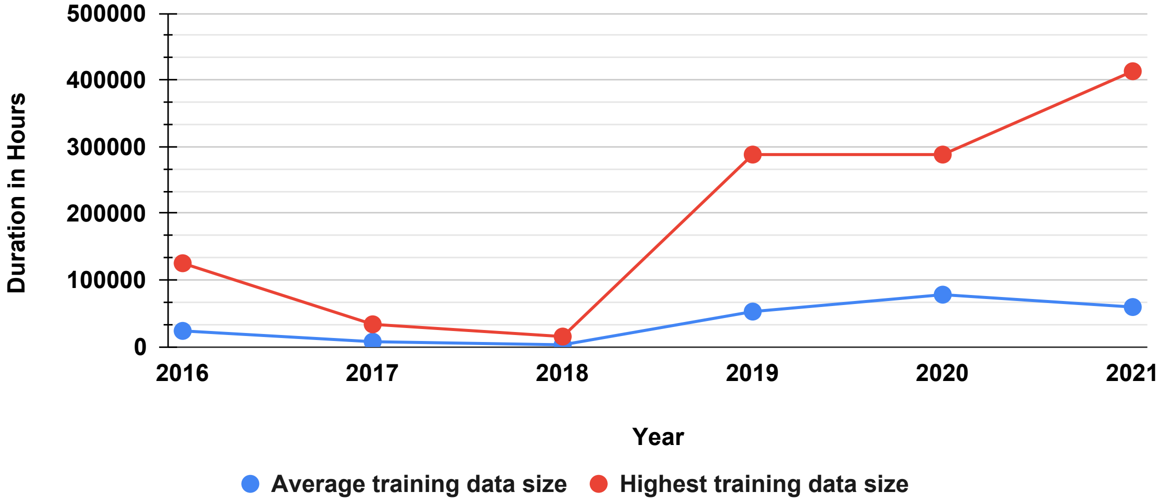


Figure 1.3: *Industry's* data demand trend in training ASR models from 2016 to 2021.

1.3 Thesis outline

This thesis explores innovative solutions to enhance human-computer interaction (HCI) systems through the automatic validation of spoken data in an unsupervised setting, which is crucial in multilingual and multicultural demography like India due to its large language and accent diversity. The thesis is structured into five main chapters, each elaborating on different facets of our approach towards addressing spoken content mismatch detection within HCI systems.

Chapter 1 sets the stage by highlighting the importance and relevance of data validation. This is then extended with the motivation for data validation in Section 1.1, followed by the motivation to automate the data validation in Section 1.1.1. Section 1.1.2 then discusses the need for extending it for spoken data as well, along with highlighting the data demand trend in recent works in Section 1.2, which further advocates the need for automatic spoken data validation. This chapter also includes an overview of the thesis structure in Section 1.3 to guide readers through the subsequent sections.

Chapter 2 presents a detailed discussion of the existing works towards automatic data validation across multiple domains. It critically analyzes existing literature, identifies gaps in current approaches, and positions the thesis as a proposal of new approaches towards achieving a robust automatic spoken data validation under Indian context for building HCI systems.

Chapter 3 motivates the necessity of IIITH MM2 Speech-Text Dataset and describes its creation. It further discusses the proposed joint entropy maximisation approach employed towards stimuli selection followed by the subject section approach in Sections 3.1 and 3.2, respectively. Sections 3.3 and 3.4, then describe the recording setup and protocol followed by the post-processing approach for the recorded samples. Section 3.5 then thoroughly explains the DTW-based baseline developed for detecting spoken content mismatches, followed by a discussion of other potential use cases for this dataset in Section 3.6.

Chapter 4 then describes the details of unsupervised pronunciation assessment analysis using utterance level alignment distance. Section 4.1 describes the importance of unsupervised pronunciation assessment and presents the related works. Section 4.2 explains the dataset used for this analysis, while Section 4.3 covers the methods applied. Sections 4.4 and 4.5 cover the rationale behind choosing Wav2Vec-2.0 and the distance measures used for utterance level alignment distance computation. This is followed by Section 4.6, which details on the approach for computing utterance level alignment distance, and Section 4.7, which explains the classification technique and experiments performed. Section 4.8 describes the experiments conducted. The baseline and final results are then presented in Sections 4.9 and 4.10, respectively.

Chapter 5 then builds upon the baseline presented in Chapter 3 towards spoken content mismatch detection. Section 5.1 delineates the datasets employed for developing unsupervised approaches to detect spoken content mismatches. Subsequently, Section 5.2 elaborates on the methodology utilized. Furthermore, Sections 5.3, 5.4, and 5.5 respectively discuss the rationale for employing self-supervised speech representations, the distance measures used for computing utterance level alignment, and the approaches to detect spoken content mismatches. The chapter proceeds with Section 5.7 and Section 5.8, which respectively examine the experiments conducted and the corresponding results.

Finally, Chapter 6 summarizes the findings and underscores the significant contributions of the thesis towards the development of a more reliable and efficient automatic spoken data validation system in the Indian context.

Chapter 2

Related work regarding automatic data validation

In the literature, there have been some works targeting to automate data validation. Early attempts towards achieving this revolve around anomaly detection as surveyed in the works [16] and [17]. Popular approaches include density estimation [18], one-class SVM-based approach [19], tree-based isolation forest [20] as well as GAN utilization [21].

Several other works have addressed this by attempting to detect out-of-distribution (OOD) samples within the dataset. In [22] the authors the OOD samples by thresholding at the highest softmax score at the output of the neural network. They hypothesize that the overall softmax score of a true positive sample would be higher as compared to a false positive sample. The experiments are conducted across various tasks in the domains of computer vision, natural language processing as well as automatic speech recognition. This approach is seen to be extended in works like [23] and [24] where the authors extend the idea of utilizing OOD samples for automatic data validation. In [23] the authors experiment on image datasets such as CIFAR-10 [25], TinyImageNet [26] and LSUN [27]. They utilize temperature scaling and identify that introducing small perturbations to the input helps enhance the difference between the softmax scores of in-distribution and out-of-distribution samples, thus leading to better data validation. In [24] the authors extend the OOD sample detection by jointly training the classifier responsible for detecting the OOD samples along with a generative neural network designed to produce more reliable OOD samples. This experiment is conducted by using popular image datasets such as CIFAR [25], SVHN [28], ImageNet [29] and LSUN [27].

The idea of OOD detection is further extended in the work [30] where the authors utilize generative ensembles to learn a tractable likelihood approximation of the training distribution and use it to reject OOD samples. Similarly, in [31], the authors make use of multiple semantic label representations to detect OOD samples. In [32], the authors utilize the variational information bottleneck [33] towards achieving the OOD detection. Furthermore, the authors in [34] develop an interesting approach of OOD detection by exposing the models to OOD samples deliberately and then exploring the heuristics for differentiating between the samples that are in-distribution and others that are OOD.

While these OOD detection techniques have proven to be useful, they often utilize labelled data, thus limiting their scalability towards validating large-scale raw datasets for which sufficient reliable labels might not always be available.

Several other works develop rule-based approaches to automate data validation. For example in [35] the authors suggest rule-based automatic data validation techniques for machine learning datasets. In [36], the authors utilize Deequ, an Apache Spark-based library, to automate the data validation. In [37], the authors present a data validation approach via implementing unit testing using the Apache Spark pipeline. However, the challenge with these approaches is that they usually need a lot of domain knowledge to specify explicit rules, constraints, and patterns for data validation.

In [38], pointwise gradients from the model are utilized to obtain the outlier filtering heuristics. These heuristics are then utilized to identify outliers in the given dataset. Similarly, authors in [39] propose a probabilistic model over a dataset. This integrates the integrity constraints of the dataset with external data sources to obtain data cleaning suggestions thus ensuring that the dataset does not contain any OOD values. In [40] the authors target automatically detecting domain value violations on the dataset. Domain value violation refers to cases when attributes take values outside of the permitted domain. For example, say a child is born, so the date of birth cannot be a date from the future followed by [41] where the authors attempt to identify data validation rules from a small subset of clean data. [42] suggests a unique active learning-based approach towards automatically labelling the crowd-sourced dataset by learning from the human-labelled subsets. In [43], the authors propose a value modification-based data validation approach. It utilizes a combination of machine learning and likelihood methods for identifying and modifying OOD records. In [44], data validation is performed by imputing the missing records based on the attribute relationships of the relational database. Several other similar validation approaches via cleaning the data are covered in the survey [45].

While the aforementioned approaches are unique and intuitive, the majority of them are limited to only numerical and categorical attributes, thus limiting their scalability. In addition to this work like [46] discusses a variety of data duplication detection techniques thus wring well for a very niche use-case of data validation. Similarly, in [47], the cleaning of large face datasets is automated by the detection and removal of dissimilar faces tagged to a common identity.

We can also find a lot of work being done towards automatically validating the data in a Machine-learning production pipeline. While this is a good rule-based way to automate dataset validation with known constraints, it might not be an optimal approach for dynamic datasets. In [48] the authors talk about detecting anomalies in time series data. [49] also talks about detecting anomalies in a presented dataset by utilizing dimensionality reduction and statistical hypothesis testing. Similarly in [50] the authors develop an approach involving deducing appropriate data-validation patterns that accurately represent the specific data domain. This technique effectively reduces the occurrence of false positives while enhancing the detection of data quality issues. However, the effectiveness of the developed method is still sub-par to the human validation as well as limited to validating if the newly added data points fall into the original data distribution as well as the datatype check. Furthermore, Google utilizes TensorFlow data validation tool [51] to automate the validation of training and inference samples. Just like Google, LinkedIn has its own data validation platform [52] which generates statistical insights that are then used to adopt datasets for the desired training. Similarly in [53] the authors survey the best practices adopted in industrial machine learning projects towards automatic data validation. In [54] the authors attempt to improve the scores on machine reading comprehension task by utilizing a BERT to improve the overall label quality of the dataset. They leverage the semantic data cleaning over syntactic data cleaning to achieve significant improvements on the TriviaQA [55] dataset.

In the literature, some works perform automatic data validation in the domain of speech as well. In [56], the authors automated the pruning of recordings when the expressive styles differed from the intended. In Librispeech corpus [57], the utterances whose decoding did not match the reference transcripts were discarded. In MUCS dataset [58], Hindi and Marathi utterances are automatically validated using an ASR-based likelihood considering the forced-alignment process. In [59], automatic validation was carried out on emotional recordings at the speaker level using the discriminative classifiers KNN and SVM. Using accuracy as a thresholding criterion, they discarded the recordings.

Though there were works that validated the data automatically, the resultant data from these methods could fail to obtain diversity and reliability of the data. It was reported that the highest F_1-score of only 0.5 from the automatic validation proposed in [10]; thus, the reliability of the method is low. Further, the decoding-based validation criteria in [11] bias to the grammatical structures of highly occurred sentences in the ASR training. Thus, the obtained data might not contain utterances with rare grammatical structures and out-of-vocabulary words. Similarly, the forced-alignment likelihood-based validation also bias to correctly pronounced and highly occurred utterances. In all of these methods, the strategy is proposed only to discard the utterances, but not to facilitate the correction mechanism, which is also part of the manual validation process.

Chapter 3

IIITH MM2 Speech-Text Dataset

In all of the methods discussed in section 2, the strategy is proposed only to discard the utterances, but not to facilitate the correction mechanism, which is also part of manual validation process. Furthermore, in order to comprehensively address the problem of automating the spoken data validation, the development of generalizable automatic data validation methods with robustness towards diverse non-native variations is imperative. For fostering the creation of such methods a corpus with both the correct as well as naturally mismatched utterances for a given set of phonetically rich stimuli set shall be required.

However, to the best of our awareness, no corpora exist to develop robust and scalable automatic data validation. Hence, this motivates us to collect a read speech dataset encompassing phonetic richness in its stimuli set with matched and naturally mismatched utterances.

This motivates us to develop IIITH MM2 Speech-Text corpus for automatic spoken data validation containing the following:

1. Speech data from non-native speakers to include the diversity in the pronunciations.
2. Utterances with naturally occurring spoken errors while reading the text.
3. Manually transcribed text to reflect the spoken errors.
4. Speech data to build the models for automatic correction of the errors.
5. Speech data with word level segmental boundaries.
6. Phonetically rich text stimuli.

We create IIITH MM2 Speech-Text data selecting a subset of 100 stimuli from the total of 2342 TIMIT [60] stimuli to ensure the phonetic richness by proposing a joint entropy maximization approach.

Table 3.1: Examples of Insertion, Deletion and Replacement mismatches from the recorded corpus presented with reference text (RT) and transcription annotated (TSA).

Insertion	**RT:**	**It offered to surrender its right to exclusive trade but asked an indemnity**
	TSA:	**It offered to surrender its right to exclusive trade but asked for an indemnity**
Deletion	**RT:**	**Superior new material for orthodontic work is another result of research**
	TSA:	**Superior new material for orthodontic work is result of research**
Replacement	**RT:**	**In most discussions of this phenomenon the figures are substantially inflated**
	TSA:	**In most discussions of this phenomenon the figures were substantially inflated**

These stimuli are recorded from 50 Indian speakers of Hindi, Marathi, Tamil, Telugu, Bangla, Maithili, Urdu, Gujarati, Malayalam, and Kannada nativities to ensure accent diversity. We segregate the recordings into two sets: 1) the recorded speech matched with reference text, and 2) the recorded speech containing spoken errors, which causes mismatches between speech and the reference text. Further, we perform manual annotation to obtain the text that reflects spoken errors in the mismatched speech.

The collected read speech dataset comprises 5764 utterances with a total duration of ∼7 hours, recorded across 50 Indian speakers. It consists of both matched as well as naturally mismatched utterances with respect to the reference text. While the matched set has 5000 utterances, making its entire duration ∼6 hours, the mismatched set has 764 utterances and has a duration of ∼1 hour. Any utterance whose transcription contains mismatches at the word level is kept in the mismatched set of the dataset, while the other utterances are kept in the matched set. A sentence is said to contain mismatched words if the utterance has an insertion, deletion, or replacement of one or more words that were not present in the original prompt provided to the subject while recording. An example of each of these three cases from the collected dataset is presented in Table 3.1. It consists of reference text and the manually annotated transcription of audio spoken by the subject during the recording process. In the example for insertion error, the word 'an' is extra, which was not present in the corresponding reference text, whereas in the case of deletion error, the word 'another' was present in reference text but was not uttered by the subject and is hence absent in the corresponding annotated transcription. Similarly, for replacement error, the word 'are' was present in the reference text, but while speaking, the subject replaces it with another word 'were'. While these examples are showcased for one word only, any instances where they occur for multiple words are also treated as a mismatched case.

As a preliminary analysis, we build a model for speech-text mismatch detection. The developed model considers self-supervised speech representations from Wav2Vec-2.0 and a DTW [61] distance-based classification approach. The performance in terms of F_1-score is found to be 0.87.

The details of database collection are described in the following five sections: 3.1 discusses the stimuli selection strategy, 3.2 describes the subject selection, followed by 3.3 and 3.4, which elaborate the recording process and postprocessing respectively. 3.5 discusses the preliminary analysis performed on the dataset and 3.6 talks about other potential use cases where this dataset can be utilized.

3.1 Stimuli Selection Strategy

We select 100 unique stimuli from a pool of 2342 TIMIT stimuli [62]. TIMIT was chosen as it was known for ensuring phonetic richness in the data with the choice of stimuli during its data collection. We employ joint entropy maximization approach to ensure that the selected subset also captures adequate phonetic richness. For this, we consider all possible combinations of 100 sentences out of the 2342 TIMIT stimulus. For each combination, phoneme probabilities are computed. Considering these probabilities, an entropy is computed using Equation 3.1 where p_i represents the probability of i^{th} phoneme. The combination with the highest entropy is selected as the stimuli set for the recording. The stimuli selected with this process have a mean word count of 9.09 and a standard deviation of 3.17 words. The phonetic diversity in the selected stimuli set is depicted in Figure 3.1.

$$\text{Entropy}(S) = \sum_{i=1}^{n} -p_i \log_2 p_i \tag{3.1}$$

3.2 Subject Selection

The dataset was collected from 50 subjects who were either undergraduate or postgraduate students at IIIT Hyderabad, India, with a good level of proficiency in the English language. We selected subjects belonging to diverse Indian nativities, including Hindi, Marathi, Tamil, Telugu, Bangla, Maithili, Urdu, Gujarati, Malayalam and Kannada. The primary criteria for subject selection were their English proficiency and ability to articulate clearly. This approach facilitated the selection of participants belonging to the aforementioned nativities thus ensuring a diverse representation Figure 3.2 shows the total number of subjects selected across the nativities, ensuring a gender balance. The subjects' ages ranged from

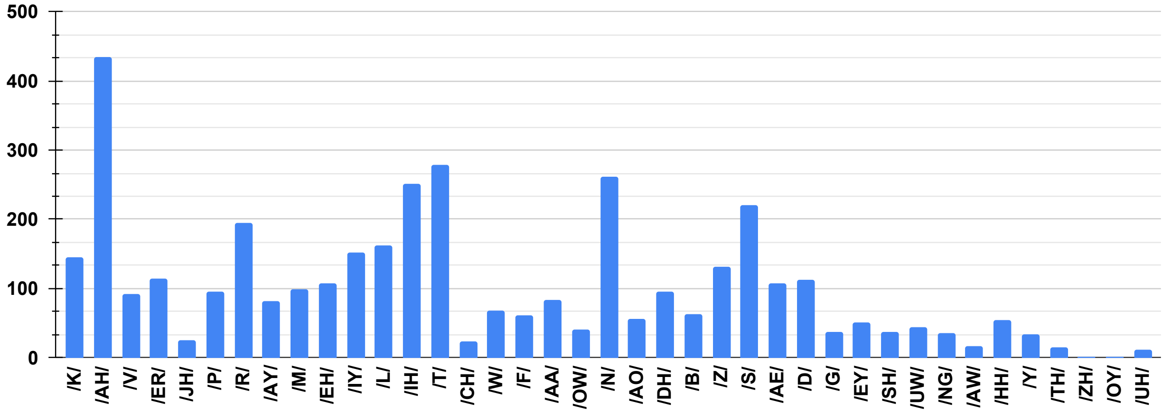


Figure 3.1: Phoneme distribution of the stimuli used for recording the dataset.

18 to 35 years, with an average age of 23.22 years and a standard deviation of 3.56 years. Despite the best efforts to maintain a homogeneous participant set across genders for each language, the gender breakdown does not exactly match due to practical challenges with sourcing the participants; given that the participation was voluntary without any monetary compensation and the speaker selection pool was limited to only IIIT Hyderabad students. Prior to the recording, all participants were confirmed to be in optimal physical and mental health. Written consent was obtained from each subject at the outset of the recording process as per the institute's ethics policies.

3.3 Recording setup and protocol

The dataset was recorded at a sampling rate of 16KHz from all 50 speakers in a noise-free anechoic studio setting. Figure 3.3 illustrates the recording setup used for dataset collection. This setup involves a JBL commercial CSLM20B microphone connected to a Dell Vostro 3020 desktop with a Dell U2412M 24-inch monitor and an Intel i5-13400 processor. This desktop hosts an in-house developed recording tool named 'Collection Module'.

The purpose of designing an in-house recording tool is to automate the user profile creation process for each speaker and store their necessary details. It facilitates speech data recording from speakers by showing them a stimuli prompt that they have to read. Once they record the stimuli, they will have the option to playback their recording, and only if they are satisfied with the recorded audio they can submit it and proceed to the next stimuli. In case the speakers are not satisfied with the recording, they

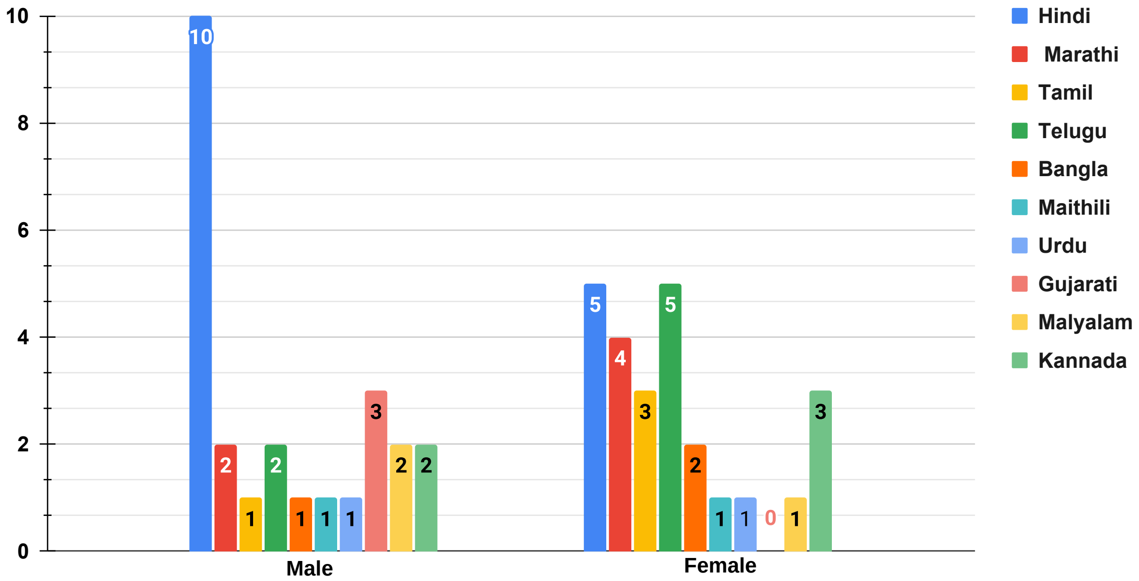


Figure 3.2: Nativities of speakers along with gender breakdown.

will have the option to re-record the same stimuli. This mechanism is enforced to ensure that the data recording is appropriate and doesn't contain any unwanted impurities that adversely impact the results of the experiments for Automatic Spoken data validation.

Furthermore, it organizes all metadata and recordings into dedicated folders for each speaker, thus enhancing the overall structure of the collected dataset. This precautionary function enables effortless continuation of recording from the precise interruption point caused by unforeseen events by simply logging into the user profile.

The collection module, when launched, opens a user signup form where the metadata of the subjects is collected. This metadata collection is done to understand better the speaker diversity associated with the dataset. However, any details about the speakers are which are personal or could be used to identify the subjects are kept confidential and shall not be released in the public domain. Since the metadata form is designed to scroll down while the submit button is stationary, a screenshot of some of the fields from this form is presented in Figure 3.5. All the fields on this page are mandatory. When the user clicks on the submit button all the user details are saved in sqlite server. After a subject successfully registers, they receive a unique username and password followed by a redirection to the login page showcased in Figure 3.4. The subjects use their respective credentials to log in, thereby getting redirected to the recording page, an exemplary instance showcased in Figure 3.6. It displays the sentence stimuli to be

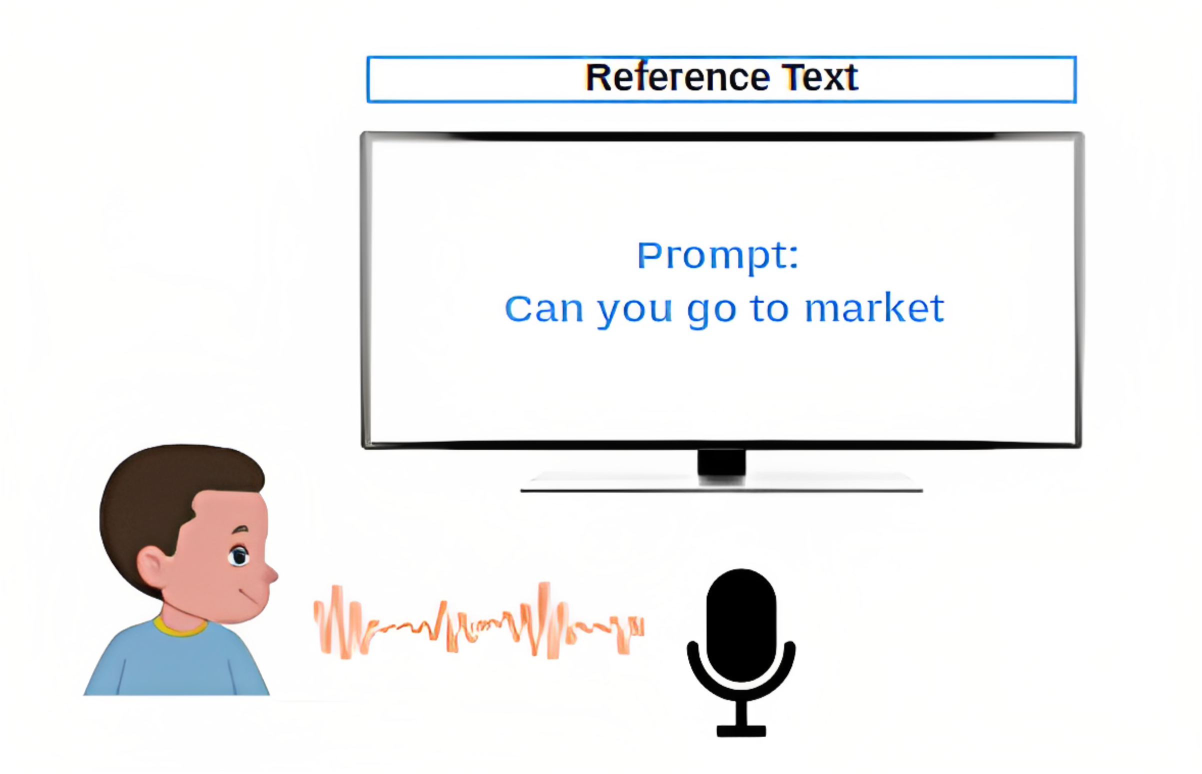


Figure 3.3: Setup used for dataset recording.

spoken by the speaker and consists of four buttons marked by the microphone, stop, playback, and next symbol, respectively.

1. Firstly we have the sentence ID which is nothing but a unique ID mapped to each of the stimuli sentences.

2. In the next line we have the Stimuli corresponding to the ID displayed on top.

3. Next we have four buttons corresponding to the following tasks:

    (a) **Microphone Symbol:** On clicking this button, the recording shall start.

    (b) **Stop Symbol:** On clicking this button, the recording shall stop. This button shall be functional only when the recording has started in the first place.

    (c) **Playback Symbol:** This button allows the user to playback the audio he/she had recorded.

    (d) **Next Symbol:** On clicking this button, the user submits the recorded audio and continues to the next stimuli.

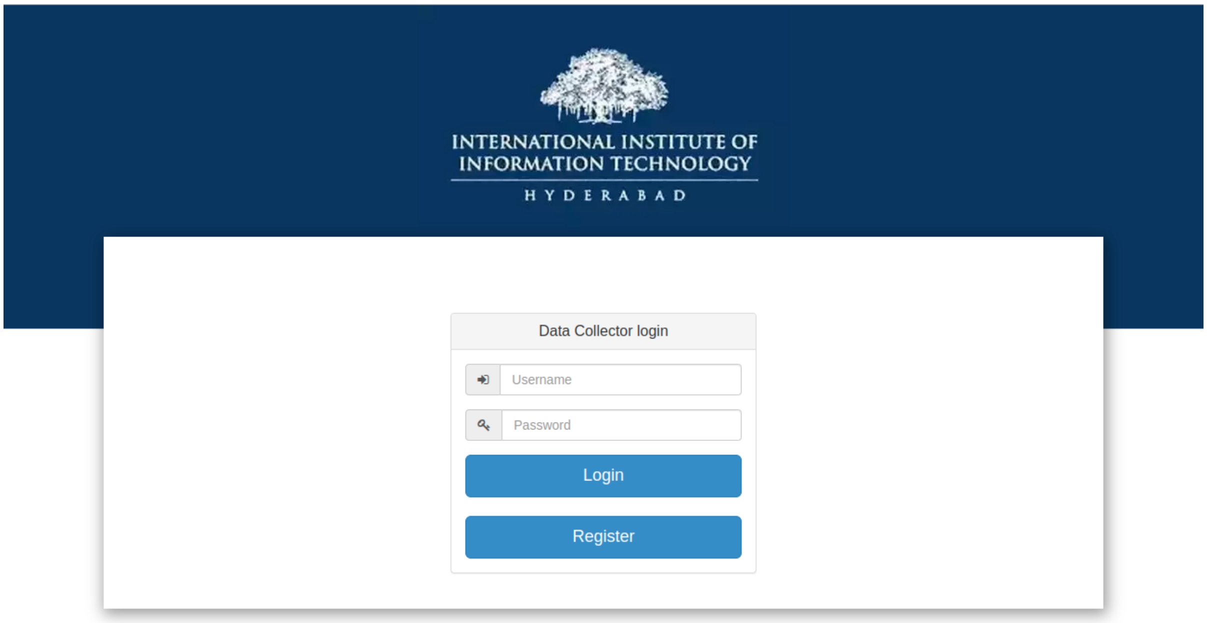
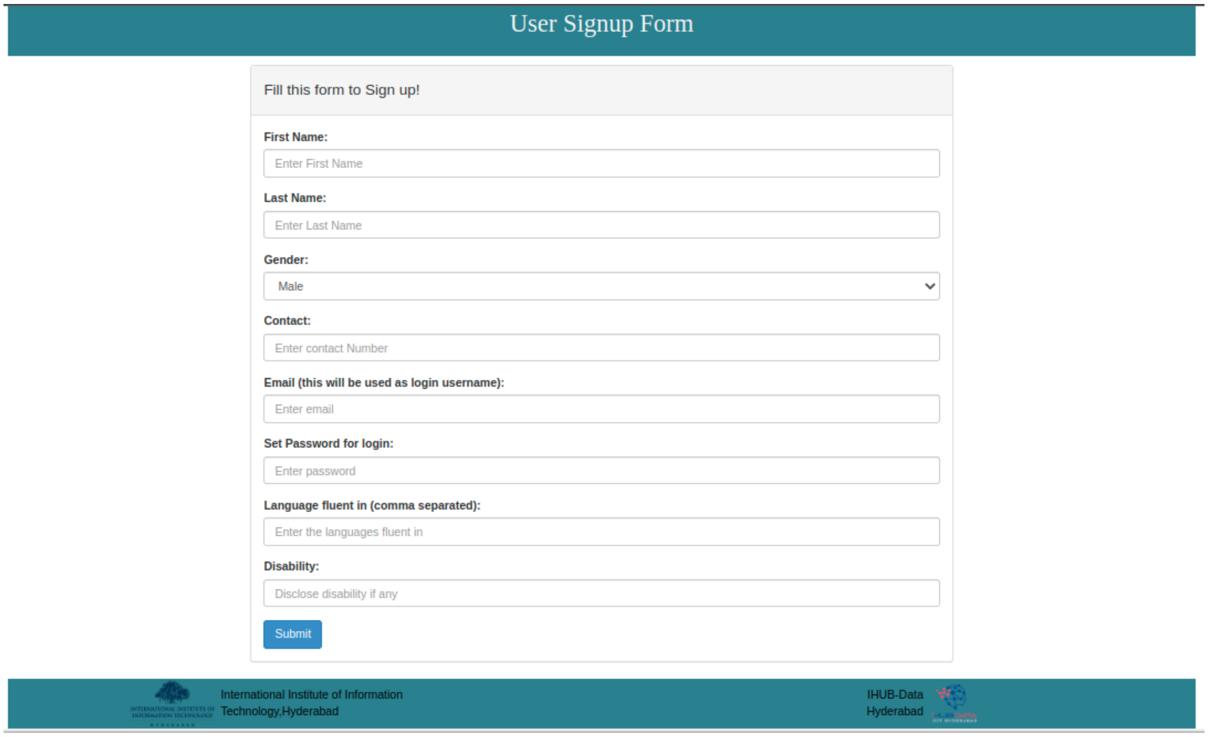


Figure 3.4: The login page in the designed webtool



User Signup Form

Fill this form to Sign up!

First Name:

Enter First Name

Last Name:

Enter Last Name

Gender:

Male

Contact:

Enter contact Number

Email (this will be used as login username):

Enter email

Set Password for login:

Enter password

Language fluent in (comma separated):

Enter the languages fluent in

Disability:

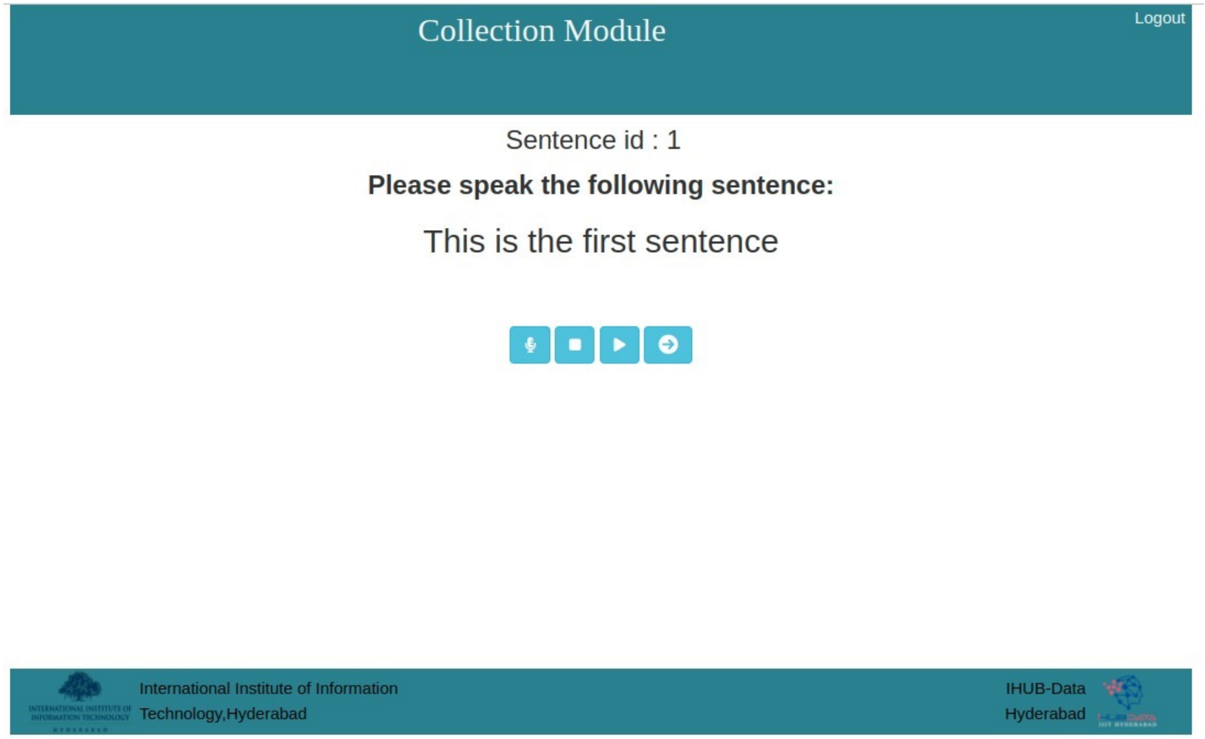
Disclose disability if any

Figure 3.5: A screenshot of the subject signup form.

Before beginning the recording, the speakers were instructed to read each stimulus at their habitual, self-determined rate, labelled ‘Normal speaking rate’. Also, they were suggested to read the stimulus appearing on the computer screen without compromising on the intelligibility of their utterance. The entire recording was carried out under the supervision of an operator, ensuring the verification of all matched and mismatched audio files during the recording process. On clicking the microphone symbol, its background colour turns red, thus marking the beginning of the recording. Once the speaker utters the displayed stimuli, the microphone symbol is clicked again. This turns its background colour back to cyan, thus indicating that the stimulus has been recorded. The playback button is clicked to verify the recorded stimuli in case the operator is doubtful about the spoken utterance. If the recording is satisfactory, the ‘next’ button is pressed, which presents the next stimuli on the screen, and the entire process is repeated. However, if the recording contains naturally mismatched utterances, the stop button is pressed, which stores the recorded sample with a mismatched tag. When the recording is tagged as mismatched or noisy or has a delayed start, the speaker is asked to re-utter the stimulus. In all such cases, the fresh recording is started by clicking the microphone symbol again. This process is repeated until all stimuli are recorded. Thus, recordings were segregated into matched and mismatched sets based on the presence of spoken errors relative to the reference text. The matched set consisted of recordings that accurately followed the given text, whereas the mismatched set included recordings with deviations such as insertions, deletions, or replacements of words.

3.4 Post Processing

After completion of the recording process, the 5000 recordings from the matched set were paired with their corresponding reference stimuli. Simultaneously, a set of 764 utterances are found with mismatched tags, which are considered the mismatched set. This set is manually transcribed and verified to reflect the mismatches between speech and the text made by the subjects. The percentage spread of mismatches along stimuli length is presented in Figure 3.7. From the figure, it is observed that, on average, there is an increase in mismatch percentages as the stimulus length increases. While the highest mismatch percentage for stimulus length below ten words is 15%, the lowest mismatch percentage for stimulus length greater than or equal to ten words is 15.3%. This may be due to the fact that longer stimuli typically contain more phonetic content and complex sentence structures, increasing the likelihood of deviations from the expected speech output. The likelihood of deviation may also increase



Collection Module

Sentence id : 1

Please speak the following sentence:

This is the first sentence

International Institute of Information Technology,Hyderabad

IHUB-Data Hyderabad

Figure 3.6: An exemplary instance of the recording page.

due to the additional cognitive load on the speakers while reading the longer sequences in an impromptu manner. We obtain word-level boundaries for both matched and mismatched sets using an ASR-based forced-alignment process. The considered ASR is GMM-HMM based and is trained on Librispeech [57] corpus using Kaldi [63] toolkit. The reason for manually transcribing the mismatched set and capturing word-level boundaries is to make the dataset suitable for developing automatic mismatch detection and correction.

3.5 Prilimnary Analysis

This section describes the speech mismatch detection of the collected data. We follow the work from [64] and [65] to perform analysis. In [64], DTW is computed between two utterances with cosine distance as a distance measure using perceptual linear prediction (PLP) [66] and frequency domain linear prediction (FDLP) features. Furthermore, [65] implements a similar technique and considers average

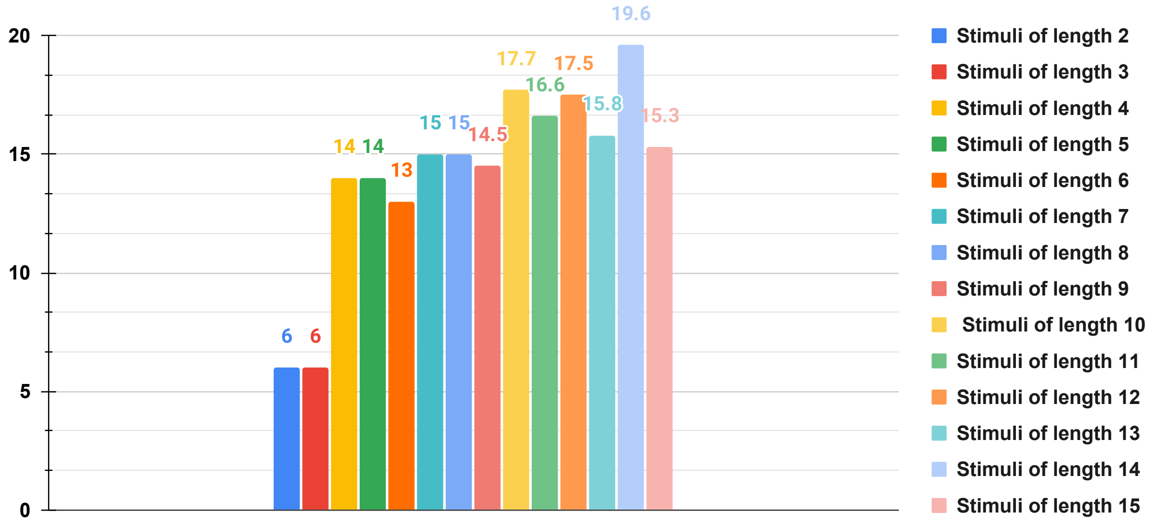


Figure 3.7: Spread of mismatched errors (in percentage) across stimuli length.

precision as well as Precision-Recall Breakeven (PRB) as the thresholding criteria. PRB threshold is the DTW distance at which precision and recall values are equal or closest.

Considering these two works, the DTW is computed with cosine distance (CD) as a distance between reference and test utterances. While PLP and FDLP have played significant roles in speech analysis, we utilize Wav2vec-2.0 [67] based self-supervised speech representations for the computation. The choice for Wav2Vec-2.0-based representations is due to its ability to obtain robust and generalizable representations from raw audio. This is because it captures linguistic [68] as well as semantic and syntactic [69] contents. Hence, these features could be useful for differentiating between correct and natural read speech mismatches. However, detailed experimentation done to justify this choice of feature is presented in Chapter 4.

For computing the DTW-based distance, we select a male and a female speaker's recording from the matched set as reference sets 1 and 2 (RS_1 and RS_2), respectively. This selection is based on the minimal occurrence of mismatched errors during the dataset recording procedure. After selecting reference utterances, the detection is performed separately, considering each reference. We separately compute DTW distance for each utterance from matched and mismatched sets with RS_1 and RS_2. These distances are used to detect the matches and mismatches between the speech and the text. The utterance having a distance below and above the PRB threshold is considered as matched and mismatched, respectively.

Table 3.2: Classification accuracy (Accuracy) (in percentage), Precision, Recall, F_1-score (F_1) at PRB and Area under Precision-Recall curve (AUPR) across distance metric CD for both reference sets (RS_1 & RS_2).

	Accuracy	**Precision**	**Recall**	$\mathbf{F_1}$	**AUPR**
RS_1	**77.731**	**0.871**	**0.870**	**0.870**	**0.927**
RS_2	77.534	0.868	**0.870**	0.869	0.926

The results reported in Table 3.2 present the area under precision-recall curve (AUPR) as well as accuracy, precision [70], recall [71], F_1-score [72] at the threshold where PRB is achieved. These results are presented for the baseline using DTW-based alignment with CD as a distance metric. From the table, it is observed that the obtained results are comparable across both reference sets, thereby advocating the usability of the collected dataset towards detecting read speech mismatches.

3.6 Other Potential use cases

The potential use of the developed IIITH MM2 Speech-Text corpus is for automatic spoken data validation. In addition to this, we believe the data can be used in other applications. As we noticed in the data, the mismatched set includes spoken grammatical errors; thus, the data can be used for detecting spoken grammatical errors automatically to aid in developing computer-assisted language learning systems. It is also to be noted that, unlike the synthetically generated grammatical errors, this data is richer in naturally made spoken grammatical errors. Another application could be automatic speech intelligibility assessment. It is often known that speech intelligibility is affected due to word-level errors. So, intelligibility can be assessed by considering the utterances in matched and mismatched sets as intelligible or not. Further, the data can be extended to collect speech intelligibility ratings for building more accurate assessment models. Moreover, the data can also be used to build the Indian accent recognition models robust to naturally made spoken errors. Generally, the accent recognition data contains either read or spontaneous speech. The read speech data always matches speech and text. On the other hand, spontaneous speech contains unknown spoken errors. The IIITH MM2 Speech-Text could be useful to analyze accent-specific variations introduced by non-native speakers in the grammatical structures based on the errors in the mismatched set.

Chapter 4

Unsupervised pronunciation assessment

For any experiment to be performed successfully, the features selected to represent the speech play an important role. In the preliminary analysis presented in Chapter 3, we go ahead with choosing Wav2vec-2.0 as the feature of choice as it is known to capture linguistic [68] as well as semantic and syntactic [69] contents. However, we test it on the task of unsupervised pronunciation assessment as well to justify that indeed it could be used as a feature for the task of unsupervised spoken content mismatch detection.

We perform analysis considering the unsupervised assessment approach computing DTW-based alignment distance between Wav2Vec-2.0 representations of expert's and learner's speech for assessing all the seven factors. The alignment distance computed from DTW is used as the metric to differentiate the binary class of each factor with a threshold computed from equal error rate (EER) [73] criterion. For the distance computation, we use three distance metrics: 1) mean absolute error (MAE) [74], 2) mean square error (MSE) [75], and 3) cosine distance (CD). For the experimentation, we consider voisTUTOR corpus [76] containing recordings from 16 learners and two experts, in which each speaker spoke a set of 1676 stimuli. Further to analyze the drawback of the recording requirement of speech from an expert, we obtain two sets of speech samples synthesized from the state-of-the-art text-to-speech (TTS) systems for all 1676 stimuli. We consider these two synthesized speech sets as the two experts' speech and perform the proposed unsupervised assessment. Among all the factors, the highest accuracy of 81.24% is obtained for intelligibility factor. When compared with the baseline, the highest relative improvement was found to be 48.28% for pause placement factor. It is also observed that the performance variability among experts is very low for all the factors, which suggests that the synthesized speech can be used in place of recorded speech from human experts to achieve scalable and economical solutions.

The key contributions in the proposed analysis are as under:

1. Assessment of the factors using a common feature unlike the factor-specific features considered in the existing works discussed in section 4.1.

2. Consideration of the current state-of-the-art Wav2Vec-2.0 representations for the assessment.

3. Computation of alignment distance using cosine distance (inspired from cosine similarity) [77] that measures angular dissimilarity between two input features.

4. Analysis of the need for the recording of the paired expert speech, which is costly and time-consuming, to that of the learner's speech.

The details of unsupervised pronunciation assessment analysis using utterance level alignment distance is described in the following sections: Section 4.1 discusses the significance of unsupervised pronunciation assessment analysis along with the related works. 4.2 describes the dataset used to perform unsupervised pronunciation assessment analysis followed by 4.3 discussing the methodology employed. 4.4 and 4.5 then talk about the motivation behind having Wav2Vec-2.0 and the feature of choice as well as the distance measures used for utterance level alignment distance computation. This is followed by the sections 4.6, 4.7 and 4.8 which describe the approach of computing utterance level alignment distance, the classification approach, and the experiments performed. The baseline and the final results produced are then discussed in sections 4.9 and 4.10, respectively.

4.1 Significance and Related works

The rising number of second language (L2) learners led to an increase in the popularity of automated assessment tools, which consider Computer Assisted Language Learning (CALL) [78] techniques. These tools benefit the L2 learners in enhancing their spoken proficiency. Among the different aspects of spoken proficiency, pronunciation plays a critical role and it is affected by various factors. As per [79], the pronunciation quality depends on the following seven factors: Intelligibility, Phoneme quality, Phoneme mispronunciation, Syllable stress, Intonation, Correctness of pause placement, and Mother tongue influence (MTI). Thus, automatically assessing the quality of these factors could help L2 learners in obtaining detailed feedback about their pronunciation. Hence, spoken proficiency can be enhanced with the tools incorporating this detailed feedback.

Various automatic assessment works were done in the literature to assess all seven factors except MTI and Phoneme quality in an unsupervised manner [80, 81, 82, 83, 84, 85, 86]. In most of these works, one of the common techniques used is Dynamic Time Warping (DTW), which calculates similarity between two sequences of different lengths. The similarity level is determined by the alignment distance resulting from DTW [86]. In the context of language learning, the similarity is computed between expert (reference) and learner (testing) utterances with different distance metrics and input features.

In [82], intelligibility factor was assessed using phoneme-based posteriorgrams (PPGs) as input features, and the DTW is computed between expert and learner using the Bhattacharyya distance (BD) metric. Miodonska et al. [83] showed that the assessment of phoneme mispronunciation factor was more effective with DTW than hidden Markov model (HMM) when the DTW is computed using Mel-frequency cepstral coefficients (MFCCs) with euclidean distance metric. In [84], intonation factor was assessed considering MFCC and pitch features for distance computation. Further, in this work, stress factor was also assessed using pitch and energy features. In [85], the pause placement factor was assessed considering MFCCs and the Euclidian distance metric-based DTW. In [86], Euclidian, Bhattacharya distance, and Kullback Leibler divergence metrics are considered for DTW computation for assessing the pronunciation quality considering posteriorgrams as input features.

In all the existing works, the factors were assessed considering heuristically computed factor-specific input features of learners' speech with respect to that of experts' speech. Thus, it requires a pair of speech recordings for each stimulus one from the learner and another from the expert. Also, as these factors influence a common phenomenon, i.e. pronunciation quality, a common input feature for all the seven factors can be analyzed for better modeling. In recent studies, the effectiveness of self-supervised learning (SSL) [87] was showcased in various speech-processing applications [88]. In these studies, it was demonstrated that the contextual representations obtained from these pre-trained SSL models are capable of capturing linguistic information [89], suprasegmental pronunciation, syntactic and semantic text-based features [90]. Among the existing SSL representations, Wav2Vec-2.0 [91] is one of the popular methods and was learned using cosine similarity metric. Recently, in [92] Wav2Vec-2.0 representations were used to predict the speaker proficiency level of L2 English learners. This further justifies our choice of using Wav2Vec-2.0 representations.

4.2 Dataset description

For the experiments conducted towards unsupervised pronunciation assessment analysis, voisTUTOR corpus [76] was considered. This corpus consists of English speech recordings of 1676 unique stimuli obtained from 2 experts and 16 learners. Among the experts, one is a male voice-over artist, referred to as Expert 1, with more than 20 years of experience, while the other is a female voice-over artist, referred to as Expert 2, and a spoken English teacher with more than 25 years of experience. The stimuli were carefully selected from spoken English materials to cover various aspects of pronunciation, encompassing phonological elements such as fricatives, stops, nasals, glides, laterals, consonant clusters, vowels, diphthongs, and semi-vowels. Additionally, for each audio recording of the learner, a set of seven binary ratings (0 or 1) was given by the female expert to evaluate the influence of seven factors on the overall quality.

Furthermore for each of 1676 unique stimuli, Indian and American accented speech samples are synthesized with male voice using Google TTS API. The speech samples from Indian and American TTS are referred to as Expert 3 and Expert 4, respectively. Considering these, in total, four sets of expert speeches are used.

4.3 Methodology

Figure 4.1 shows the block diagram of DTW-based unsupervised assessment approach for assessing all seven factors. It has four steps. Given the stimulus, the first step obtains Wav2Vec-2.0 representations for expert and learner speech samples separately using a pre-trained model. In this process, we discard the silence from the start and end of the speech samples. The second step computes the utterance level alignment distance between expert and learner Wav2Vec-2.0 representations with DTW considering a distance metric. The third step considers EER based threshold criterion to compute threshold (τ) for predicting the binary rating of each factor. The fourth step detects the binary rating of each factor as label 1 or 0 considering τ. It is to be noted that, for a given stimulus, the computation flow in Figure 1 repeats for each combination of expert and distance metrics. Similar to existing works, we hypothesize that the alignment distance computed with Wav2Vec-2.0 has lower values when factor-specific features in learner's speech are similar to expert's speech and vice versa. We hypothesize that all the seven factors' specific features are embedded in Wav2Vec-2.0 representations.

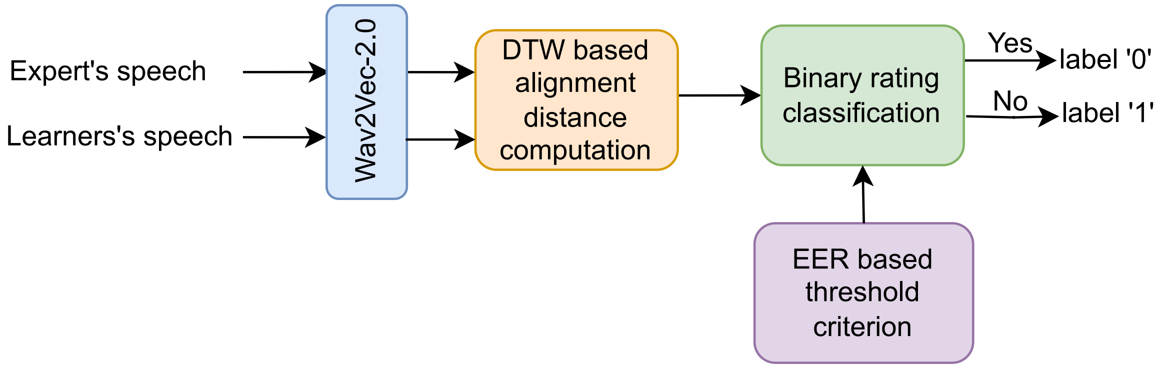

Expert's speech
Learners's speech
Wav2Vec-2.0
DTW based alignment distance computation
Binary rating classification
Yes
label '0'
No
label '1'
EER based threshold criterion

Figure 4.1: Block diagram of DTW-based factors' assessment with Wav2Vec-2.0 representations

4.4 Motivation for using Wav2vec-2.0 representations

Wav2Vec-2.0 [93] is a state-of-the-art model for obtaining representation sequence for an input raw audio considering a self-supervised representation learning framework. These representations have been considered in many end-to-end speech recognition tasks [94, 95]. Self-supervised representation learning approach in Wav2Vec-2.0 takes raw audio data and learns the representations that could disentangle the linguistic [96, 89], pronunciation, syntactic and semantic [90] aspects in the audio. These representations are then used for fine-tuning the downstream tasks including speech recognition [96]. Wav2Vec-2.0 focuses on capturing complex patterns from waveform; introducing non-linearity by choosing activation functions such as GELU [97]. This in turn enhances its generalizability and ensures a robust representation of the waveform.

Due to this effective learning process, Wav2Vec-2.0 representations have been considered in a wide variety of tasks besides speech recognition such as Speaker recognition [98] Speaker adaptation [99], Speaker verification [96], Cross-lingual knowledge transfer [100], Mispronunciation detection [101, 102], Voice activity detection [103], Prosodic boundary detection [104], Emotion identification [105] as well as Non-verbal vocalization detection [106]. Furthermore, it has been widely explored for medical domain tasks such as Stuttering [107], Alzheimer detection [108] as well as developing system for rating children speech with speech sound disorder [109].

We obtain Wav2Vec-2.0 representations for expert and learner speech utterances. Since Wav2Vec-2.0 captures linguistic features, suprasegmental pronunciation, syntactic and semantic text-based features [90], we hypothesize that these representations can be used for assessing the considered seven factors contributing to pronunciation quality.

4.5 Distance metrics

The following distance measures for computing utterance level alignments between expert-learner speech pairs for a given stimulus:

Mean absolute error (MAE): It is a metric used to quantify the average absolute difference between any two vectors r_e and $\tilde{r_l}$ of dimension D. It ranges between [0,∞).

$$\mathrm{MAE}(\mathrm{r_e}, \tilde{\mathrm{r_l}}) = c_{MAE}(e,l) = \frac{1}{D}\sum_{i=1}^{D}\left|r_e^i - \tilde{r_l}^i\right| \tag{4.1}$$

Mean squared error (MSE): It is a metric used to quantify the average squared difference between any two vectors r_e and $\tilde{r_l}$ of dimension D. It is highly prone to outliers as compared to MAE due to the squaring of errors. It ranges between [0,∞).

$$\mathrm{MSE}(\mathrm{r_e}, \tilde{\mathrm{r_l}}) = c_{MSE}(e,l) = \frac{1}{D}\sum_{i=1}^{D}\left(r_e^i - \tilde{r_l}^i\right)^2 \tag{4.2}$$

Cosine Distance (CD): It is a metric that quantifies the angular dissimilarity between any two vectors r_e and $\tilde{r_l}$ of dimension D. It remains unaffected by the vector magnitudes and has a range of [0,2].

$$CD(r_e, \tilde{r_l}) = c_{CD}(e,l) = 1 - \frac{\sum_{i=1}^{D} r_e^i \tilde{r_l}^i}{\sqrt{\sum_{i=1}^{D}(r_e^i)^2}\sqrt{\sum_{i=1}^{D}(\tilde{r_l}^i)^2}} \tag{4.3}$$

4.6 Alignment distance computation

For a given stimulus, the number of frames is different in the expert's and learner's speech. Hence, we cannot obtain a direct one-to-one mapping between their frames in a sequential manner as it would always leave some frames unmatched. Furthermore, there is always the possibility of different phonemes being stretched temporally for an expert-learner speech pair. This gives rise to an uneven distribution of frames across phonemes. To address this issue, DTW [110] algorithm is considered (for which the cost can be a distance metric); which is highly time efficient with a focus on cost optimization. In this work, the distance metrics considered are MAE, MSE and CD. Considering a distance metric, the DTW obtains the best possible frame level alignments between each of the expert-learner Wav2Vec-2.0 representation sequence pairs.

Algorithm 1 Utterance level alignment distance computation

```
Input: $\mathcal{E}$={$r_1, r_2, ..., r_E$} and $\mathcal{L}$={$\tilde{r}_l, \tilde{r}_2, ..., \tilde{r}_L$};
Initialization: $\tilde{C}(e,l) \leftarrow \infty, \forall e,l; 0 \leq e \leq E, 0 \leq l \leq L$, $\tilde{C}(0,0) \leftarrow 0$;
Distance (cost) $\tilde{C}$ matrix updation:
$e \leftarrow 1, l \leftarrow 1$;
while $e \leq E$ do
    while $l \leq L$ do
        Compute $c(e,l)$;
        $\tilde{C}(e,l)$ = $c(e,l)$ + min[$\tilde{C}(e-1,l)$, $\tilde{C}(e,l-1)$, $\tilde{C}(e-1,l-1)$];
        $l \leftarrow l+1$;
    end
    $e \leftarrow e+1$;
end
Output: $C(E,L) = \tilde{C}(E,L)$
```

Let $\mathcal{E}$={$r_e; 1 \leq e \leq E$} and $\mathcal{L}$={$\tilde{r}_l; 1 \leq l \leq L$} are the $D-$dim Wav2Vec-2.0 representation sequences of expert and learner with lengths E and L, respectively. The best alignment between $\mathcal{E}$ and $\mathcal{L}$ sequence pair is computed using DTW as shown in Equation 4. The equation computes accumulated cost ($C(e,l)$) considering representation sequences ($r_1, r_2, \ldots, r_e$ and $\tilde{r}_1, \tilde{r}_2, \ldots, \tilde{r}_l$) till e-th and l-th frames from the respective $\mathcal{E}$ and $\mathcal{L}$. The accumulated cost computation at e and l involves three accumulated costs: $C(e-1,l)$ $C(e,l-1)$ $C(e-1,l-1)$ and one local cost $c(e,l)$, which is the distance between r_e and $\tilde{r}_l$ using one of the Equations 4.1, 4.2 and 4.3. Considering the accumulated and local costs in Equation 4.4, we obtain the utterance level alignment distance (cost), which is equal to $C(E,L)$. It is to be noted that the cost $C(E,L)$ computation is not straightforward forward and it is recursive in nature as $C(E,L)$ involves three accumulated costs from the previous frames and local cost from frames E and L. Further, the accumulated costs from the previous frames depend on downstream accumulated costs, which involve the previous of previous frames and so on. Under this recursive relation of the costs, the utterance level alignment distance (cost) is computed using Algorithm 1.

$$C(e,l) = c(e,l) + \min\left[C(e-1,l), C(e,l-1), C(e-1,l-1)\right] \tag{4.4}$$

4.7 Classification Approach

We consider the alignment distance for detecting factors' binary ratings by computing a threshold (τ) considering EER criterion. The EER is an average of the false acceptance rate (FAR) and false rejection rate (FRR). The equations for FAR, FRR, and EER are provided in Equations 4.5, 4.6 and 4.7 respectively. As per this criterion, the τ is computed as the alignment distance value at which the values of FAR and FRR are equal. Considering τ value, we classify the learner speech samples whose alignment distance is above τ as one label and the below as another label.

$$FAR = \frac{\text{Number of false acceptances}}{\text{Number of identification attempts}} \tag{4.5}$$

$$FRR = \frac{\text{Number of false rejections}}{\text{Number of identification attempts}} \tag{4.6}$$

$$EER = \frac{FAR + FRR}{2} \tag{4.7}$$

4.8 Experiments

The entire learner dataset from the voisTUTOR corpus was used for the experiments. We consider four experts' speech sample sets of 1676 stimuli, out of which two experts are from voisTUTOR corpus and the remaining two experts' speech are synthesized samples. EER [73] criterion was used to determine the threshold (τ) using 5% of the data. The performance of the suggested approach was assessed on the remaining 95% of data, using classification accuracy as the performance indicator with the help of ground truth labels present in the data. The classification was carried out independently for each of the three alignment distances, with τ being determined for each distance and used to categorize alignment distance results. This was repeated considering each expert's speech samples set across all the seven factors considered in the experiments. The alignment distances were calculated independently for all three distance measures in a stimuli-specific manner using the respective speech from each expert.

4.9 Baseline

We compare the performance of the proposed unsupervised assessment approach across all three distance metrics using random selection baseline. This approach corresponds to the random assignment of labels for the entire learners' dataset using a binary distribution designed to replicate the original data distribution in the randomly selected 5% dataset from voisTUTOR corpus. The classification accuracy thus obtained using these labels is compared against results obtained for the proposed unsupervised assessment.

4.10 Results

The results obtained for all classification approaches across all factors considered in the experiments are discussed in the following sub-sections. Sub-section 4.10.1 talks about the overall performance followed by sub-section 4.10.2 that discusses about the phoneme category-specific performance. Lastly sub-section 4.10.3 presents an analysis with illustrative examples.

4.10.1 Overall Performance

Table 4.1 shows the classification accuracies obtained with the proposed analysis using all three distance measures – MAE, MSE, and CD across all four experts along with BL. From the table, it is observed that the accuracies obtained with the proposed analysis with all three distance measures are higher than BL except for intelligibility. This may be due to other factors impacting intelligibility that have not been considered in this study. This limitation highlights the need for developing more nuanced features or algorithms that can capture the subtleties of intelligibility more effectively.

However, the highest accuracy of 81.24% is observed for intelligibility among all the factors with Expert 2 for distance measure CD. The highest relative improvement is found to be 17.48%, 23.05%, 13.21%, 48.28%, 34.20%, and 27.24% for intonation (Expert 1 using MSE), phoneme mispronunciation (Expert 3 using MSE), MTI (Expert 3 using MSE), pause placement (Expert 3 using MAE), phoneme quality (Expert 3 using MAE), and syllable stress (Expert 1 using MSE), respectively compared with BL. This indicates the benefit of the proposed analysis for assessing all seven factors. The results across experts exhibit minimal variation, suggesting that synthesized speech can effectively substitute human expert speech in CALL systems thereby boosting its scalability.

Table 4.1: Classification accuracy (in percentage) obtained with proposed analysis for all the seven factors.

	Expert 1			**Expert 2**			**Expert 3**			**Expert 4**			**Baseline**
Factors	**MAE**	**MSE**	**CD**	**MAE**	**MSE**	**CD**	**MAE**	**MSE**	**CD**	**MAE**	**MSE**	**CD**	**RS**
Intelligibility	80.79	80.36	**81.09**	81.23	80.86	**81.24**	80.86	80.89	**81.20**	80.09	79.92	**80.79**	85.44
Intonation	**59.40**	**59.67**	**58.91**	**58.44**	**58.16**	**58.06**	**59.10**	**59.14**	**59.13**	**59.49**	**59.37**	**59.34**	50.79
Phoneme mispronunciation	**59.50**	**60.18**	**55.69**	**58.50**	**59.22**	**55.03**	**60.08**	**60.63**	**57.41**	**58.95**	**58.69**	**56.72**	49.27
MTI	**57.33**	**57.90**	**53.74**	**56.49**	**57.09**	**52.67**	**57.76**	**58.19**	**55.14**	**56.91**	**56.31**	**54.69**	51.40
Pause Placement	**74.17**	**73.70**	**73.90**	**74.25**	**74.26**	**73.96**	**75.21**	**75.16**	**74.38**	**74.31**	**74.10**	**73.66**	50.72
Phoneme quality	**65.31**	**65.08**	**63.92**	**64.69**	**65.15**	**63.67**	**65.41**	**65.06**	**64.26**	**64.61**	**63.78**	**63.53**	48.74
Syllable stress	**59.90**	**60.58**	**57.34**	**59.16**	**59.10**	**56.50**	**60.20**	**60.33**	**58.56**	**59.25**	**58.97**	**57.89**	47.61

4.10.2 Phoneme category-specific performance

Figure 4.2 illustrates the classification accuracies obtained across different phoneme categories with the proposed analysis with all three distance measures – CD, MAE, and MSE. The highest classification accuracies for all seven factors with respect to each expert are as follows: Intelligibility - 86.25%, 88.20%, 86.67%, and 83.88% are obtained for nasals and semi-vowels with experts 1, 2, 3, and 4, respectively. Intonation - 52.65%, 51.78%, 53.35%, and 57.75% are obtained for consonant clusters with experts 1, 2, 3, and 4, respectively. Pause Placement - 78.07%, 81.08%, 79.67%, and 85.36% are obtained for consonant clusters with experts 1, 2, 3, and 4, respectively. Phoneme mispronunciation - 48.67%, 51.21%, 48.86%, and 58.73% are obtained for semi-vowels and vowels with experts 1, 2, 3, and 4, respectively. Syllable stress - 55.00%, 56.76%, 55.57%, and 54.22% are obtained for semi-vowels with experts 1, 2, 3, and 4, respectively. MTI - 62.61%, 60.14%, 61.00%, and 67.19% are obtained for diphthongs and nasals with experts 1, 2, 3, and 4, respectively. Phoneme quality - 58.79%, 59.15%, 56.43%, and 68.27% are obtained for vowels with experts 1, 2, 3, and 4, respectively. Except for MTI, the results from the synthesized speech are found to be inline with that from human expert speech. This further strengthens the observations made from Table I thereby suggesting that synthesized speech can be interchangeably used with human expert speech in CALL systems.

On a comprehensive analysis of the results from all four experts combined, it is observed that the phoneme categories vowels and semi-vowels exhibit the highest count of achieving the highest classification accuracy. This could be due to the fact that vowels are one of the most critical phoneme categories

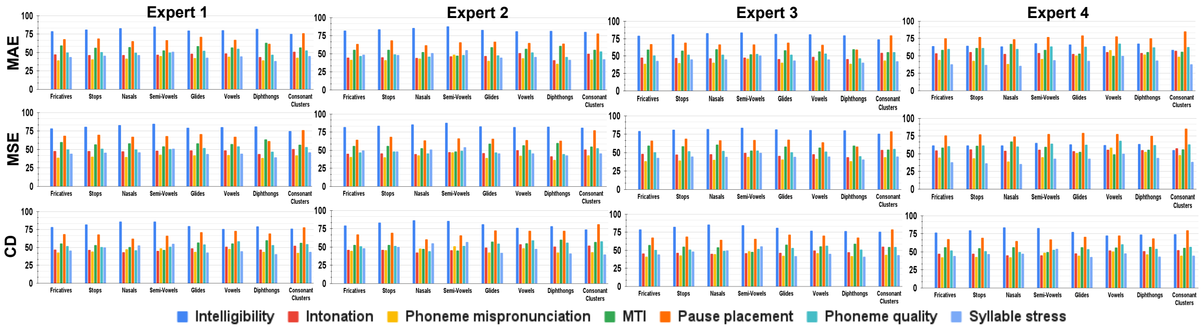


Figure 4.2: Classification accuracy (in percentage) obtained for all the seven factors under the following phoneme categories: Fricatives, Stops, Nasals, Semi Vowels, Glides, Vowels, Diphthongs, and Consonant Clusters.

in determining the pronunciation of a speech utterance as they carry high energy and hence have significant auditory prominence over other phoneme categories. They also usually form the nucleus of syllabic units, thereby leading to a drastic shift in the meaning with even a slight deviation in their pronunciation. Since semi-vowels are phonetically very similar to vowel sounds thereby carrying similar auditory prominence they play an equally important role in determining the pronunciation quality of a speech utterance.

4.10.3 Analysis with illustrative examples

Figure 4.3A and 4.3B shows the distance (cost) matrices resulting from Algorithm 1 between an expert and a learner for the spoken text “They walk” and “Wonderful”, respectively. On the figures, phoneme boundaries of expert and learner are shown with green dotted lines. The intersection of phone boundaries for expert and learner has been highlighted by red circles. The black line indicates the frame locations of expert which are aligned with that of learner. For the learners’ speech considered in Figure 4.3A and 4.3B, all the factors are correctly predicted considering utterance level alignment distance. Particularly, in Figure 4.3A, all the factors’ labels result in a good pronunciation quality i.e. the learner speech sample is intelligible (1), pronunciation quality is good (1), phoneme mispronunciation is absent (0), stress placement is correct (1), proper intonation (1), pause placement is correct (1), and mother tongue influence is absent (0). On the other hand, in Figure 4.3B, all the factors’ labels result a poor pronunciation quality. From Figure 4.3A, it is observed that red circles fall on the black line, however,

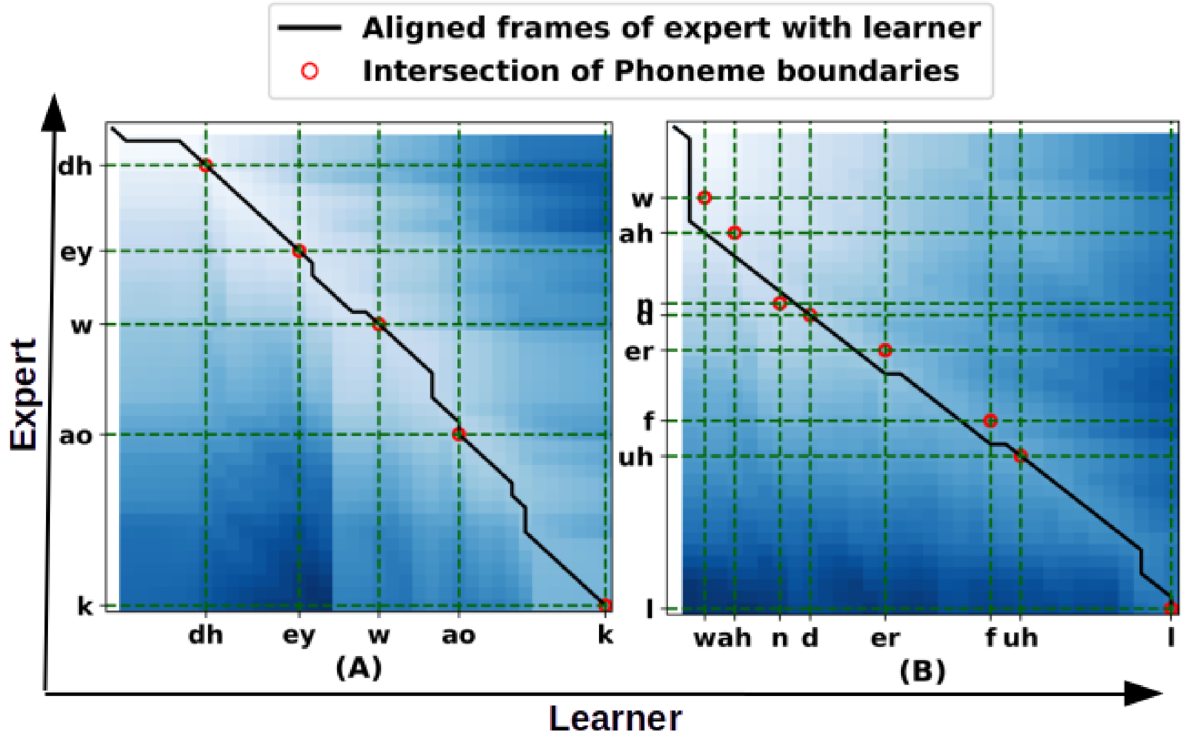


Figure 4.3: Distance matrices between expert & learner for spoken texts 'They walk' (A) and 'Wonderful' (B), for which all the factors' labels correspond to good and poor pronunciation quality, respectively.

not in Figure 4.3B. This indicates that when all the seven factors correspond to good pronunciation quality, phoneme boundaries in learner's speech aligns closely with those in expert's speech resulting the lower alignment distance. On the other hand, when all the seven factors correspond to bad pronunciation quality, phoneme boundaries in learner's speech significantly deviate from those in expert's speech resulting in a higher alignment distance. This suggests the effectiveness of the considered alignment distance.

Chapter 5

Unsupervised spoken content mismatch detection

The results presented in Chapter 4 ensure that Wav2vec-2.0 representations not only are able to capture linguistic [68] as well as semantic and syntactic [69] contents but also perform well for the task of Unsupervised pronunciation assessment analysis. Hence we can utilize its embeddings as intermediate speech representations.

Furthermore, we also consider another model named HuBERT [111] to generate speech embeddings for this task. The rationale behind this choice being that it, just like Wav2vec-2.0 is self-supervised in nature. Furthermore, it have proven to be equally effective in addressing various downstream tasks.

Hence, building upon the baseline proposed in Table 3.2 of Chapter 3; three approaches are developed towards addressing the unsupervised spoken content mismatch detection task namely Phone level cost maximized DTW and Phone level cost maximized Weighted DTW which are an extension of the DTW approach along with a cross-attention-driven approach. The aforementioned approaches utilize Indic TIMIT [112] dataset along with the IIITH MM2 Speech-Text Dataset [113] for the experimentation. In all these approaches the utterance level alignment distance is computed using the Wav2vec-2.0 as well as HuBERT speech representations using mean square error (MSE) [75] as a distance metric. The Precison-Recall breakeven point (PRB) [65] based thresholding criterion is utilized for the binary classification of the target utterances during inference. Among all these approaches the highest classification accuracy of 89.226% is obtained for the cross-attention driven approach while utilising HuBERT speech embedings.

The key contributions in this proposed analysis are as under:

1. Extension of the DTW driven baseline presented in Chapter 3 with Phone level cost maximized DTW and Phone level cost maximized Weighted DTW towards addressing the unsupervised spoken content mismatch detection.

2. Development of a cross-attention-driven approach towards unsupervised spoken content mismatch detection.

3. Consideration of the current state-of-the-art Wav2Vec-2.0 as well as HuBERT representations for the assessment.

4. Computation of alignment distance MSE across all three proposed approaches

The details of the aforementioned approaches are presented in the following sections: Section 5.1 describes the datasets utilized for the development of unsupervised spoken content mismatch detection approaches followed by 5.2 which discusses the methodology employed. 5.3, 5.4 and 5.5 then talk about the motivation behind utilizing the self-supervised speech representations, the distance measures used for utterance level alignment distance computation, and spoken content mismatch detection approaches, respectively. This is followed by Sections 5.7 and 5.8 which discuss the experiments performed and the corresponding results respectively.

5.1 Datasets utilized

For the experiments conducted towards developing Unsupervised spoken content mismatch detection approaches, Indic TIMIT [112] and IIITH MM2 Speech-Text datasets [113] are utilized. The Indic TIMIT dataset is a phonetically rich Indian English speech corpus designed to reflect pronunciation variations specific to Indian speakers. It contains approximately 240 hours of speech recordings from 80 subjects, each speaking 2342 stimuli from the TIMIT [114] corpus. The dataset also includes phoneme transcriptions for a subset of these recordings, manually annotated by linguists to capture the speaker's pronunciation nuances. Accompanying the corpus is the Indic English lexicon, which integrates pronunciation variations typical of Indian speakers, identified through their common pronunciation errors, into an existing native English lexicon.

Subjects for the recordings were selected from major Indian languages spoken by about 90% of the population, grouped into six regions, and influenced by four major language families — Indo-Aryan, Dravidian, Austro-Asiatic, and Tibeto-Burman. The recordings were conducted in a controlled environment with each subject reading aloud the stimuli displayed on a laptop, and were recorded using a Zoom H6 mixer.

Since, Indic TIMIT is recorded in a controlled environment, which ensures superior speech quality with little to no noise present in the recordings. Additionally, it covers native Indian speech and pronunciation variabilities. This makes Indic Timit a suitable dataset for the cross-attention pre-training task.

The IIITH MM2 Speech-Text dataset is an innovative corpus designed for the specific purpose of detecting and correcting mismatches between spoken audio and written text. It stands out as it includes both matched and mismatched speech-text pairs, with a preliminary analysis, yielding an $\mathbf{F_1}$-score of 0.87 using Wav2Vec-2.0 representations and Dynamic Time Warping as showcased in Table 3.2. The dataset comprises 5764 utterances recorded from 50 speakers from diverse Indian nativities, ensuring a representative phonetic richness by selecting 100 stimuli from 2342 available in the TIMIT corpus through a joint entropy maximization method. The details of this approach are described in sub-section 3.1. Recordings were conducted in a controlled anechoic studio environment using professional equipment and a custom-built software tool to facilitate the collection and organization of data. The dataset includes 5000 matched utterances and 764 with identified mismatches, categorized by insertion, deletion, or replacement errors at the word level. Furthermore, recordings from a male and a female speaker from the matched set are chosen as reference sets 1 and 2 (RS_1 and RS_2, respectively). For ease of addressability, all the other speakers apart from RS_1 and RS_2 shall be referred to as target speakers. This selection is made using the criterion of minimal presence of mismatched errors encountered during the recording process of the dataset.

This rich dataset not only provides a unique resource for testing and enhancing speech recognition systems but also includes detailed metadata and transcription verifications. This makes IIITH MM2 Speech-Text dataset ideal for the task of fine-tuning the pre-trained cross-attention model on Indic TIMIT dataset followed by inferencing for the task of spoken content mismatch detection.

Thirty speaker data from the Indic Timit dataset is utilized for the pre-training task of cross-attention followed by fine-tuning done on 30% dataset of the IIITH MM2 Speech-Text dataset. while the subset selection from both datasets is random, a gender ratio of 1:1 is ensured.

5.2 Methodological framework

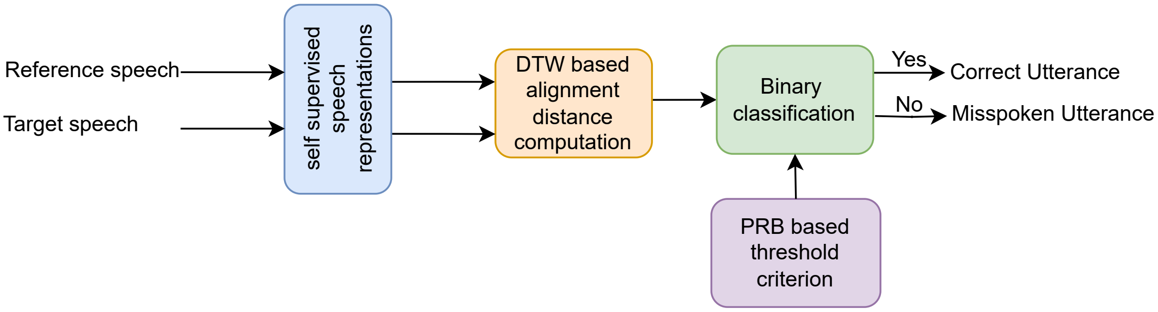

Reference speech
Target speech
self supervised speech representations
DTW based alignment distance computation
Binary classification
Yes
Correct Utterance
No
Misspoken Utterance
PRB based threshold criterion

Figure 5.1: Block diagram of DTW-based unsupervised spoken content mismatch detection using self-supervised speech representations with Wav2Vec-2.0 and HuBERT.

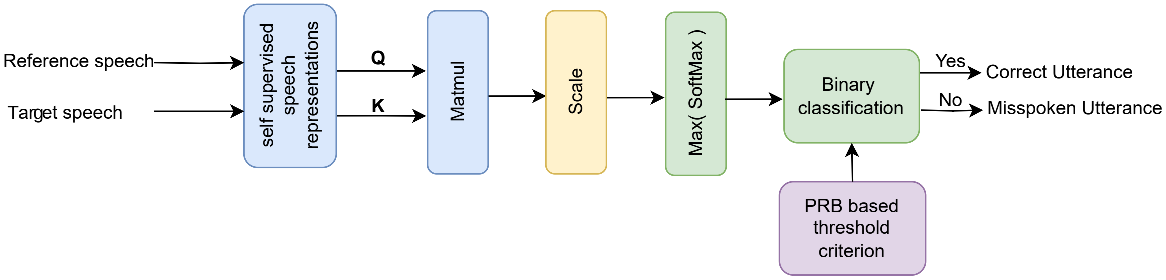


Figure 5.2: Block diagram of cross-attention based unsupervised spoken content mismatch detection using self-supervised speech representations with Wav2Vec-2.0 and HuBERT.

The block diagram for the DTW [115] based unsupervised spoken content mismatch detection is presented in Figure 5.1. It utilizes the IIITH MM2 Speech-Text dataset and has four steps. Firstly, the intermediate speech representations are obtained for both RS_1 and RS_2 along with speech samples from the remaining speakers through a pre-trained Wav2Vec-2.0 and HuBERT model. Subsequently, the alignment distance between these representations is calculated using the proposed Dynamic Time Warping (DTW) variants namely Phone level cost maximized DTW and Phone level cost maximized Weighted DTW, where Mean Square Error (MSE) is chosen as the designated distance metric. Following this, a threshold τ is obtained using the Precision-Recall Breakeven (PRB) as the thresholding criteria. PRB threshold is the DTW distance at which precision and recall values are equal or closest to each other. The fourth step detects the binary rating of each spoken utterance of the target speaker with reference to the corresponding utterances by RS_1 and RS_2 and the determined threshold τ.

For the cross-attention-based approach, the block diagram is presented in Figure 5.2. In this approach, after the intermediate speech representations are obtained for the selected thirty speakers' speeches are obtained a cross-attention model is trained using the same word-level utterances of all the selected speakers iteratively. Once the training is completed, the trained model is further fine-tuned on the 30% dataset selected for training from the IIITH MM2 Speech-Text dataset. After training, the threshold τ is obtained based on the PRB criterion for which the attention scores between same-same word pair and different-different word pairs are treated as two classes. Once the threshold τ is determined, inferencing is performed on the remaining 70 % of the dataset and based on the same-same word cross attention score a binary classification for the word is done. If any target sentence contains at least one word whose cross-attention score falls below the threshold, then that target sentence is said to contain the spoken content mismatch.

This procedure is iteratively applied to each pair of reference and target speakers. The underlying hypothesis being smaller alignment distances and higher attention scores between the self-supervised speech representations indicate a higher similarity in specific speech features between the utterances of reference speakers and the target speakers. Thus the hypothesis is extended to form the assumption that if the utterance level alignment distance exceeds the obtained threshold using the PRB criterion, the utterances in comparison are different and vice versa. Similarly, for the cross-attention scores, the hypothesis is that cross-attention would learn an association between the same and different words, assigning higher attention scores to the same words and lower attention scores to the different words.

This entire flow as showcased in Figure 5.1 and Figure 5.2 is repeated independently for each of the Wav2Vec-2.0 and HuBERT-based representations in combination with both the variants of DTW namely Phone level cost maximized DTW and Phone level cost maximized Weighted DTW as well as the cross-attention approach. The detailed explanations for each of them are present in sub-sections 5.5.1, 5.5.2, and 5.5.3 respectively.

5.3 Motivation for using self-supervised speech represenations

We obtain Wav2Vec-2.0 as well as HuBERT speech representations for performing the experiments discussed in section 5.7. The rationale behind choosing them is that these models have proven to provide robust and generalizable speech representations. Since a detailed discussion regarding the choice of

Wav2Vec-2.0 as an intermediate speech representation is presented in section 4.4, we shall only discuss the rationale behind choosing HuBERT as an intermediate speech representation here.

HuBERT just like Wav2Vec-2.0 is another self-supervised model, which is known for generating robust speech representation. Since it leverages the self-supervised learning framework, its training is independent of the labeled data. Instead, it utilizes large volumes of unlabeled speech to learn the sequence patterns via selective masking and leverages masked language modeling as well as acoustic modeling techniques simultaneously with its uniquely designed loss function which computes cross-entropy loss [116] on both its masked as well as unmasked frames.

The pseudo-labels are generated via clustering which are refined iteratively throughout the training process, drawing inspiration from the DeepCluster method for self-supervised learning of visual features [117]. Since HuBERT focuses on predicting the labels for the masked regions of the speech, it inherently develops a good high-level representation for the unmasked inputs. By predicting the cluster assignments of these masked regions iteratively, it learns to interpret the acoustic properties, enhancing its ability to capture subtle pronunciation details over time. To further improve the label prediction task for the masked regions, HuBERT also learns to capture more complex syntactic and semantic information. This occurs because the model must understand the broader context in which specific sounds and words appear, including the relationships between sequential elements in speech to minimize the prediction error.

Due to this robust its robust speech feature representation it is used in a wide variety of tasks such as end-to-end speech recognition [118] [119], Speech emotion recognition [120] [121] [122], speaker verification [123], spoken language understanding [124], dysarthric speech recognition [125] [126], speech pronunciation assessment [127] to name a few.

5.4 Distance measures utilised

MSE is the distance metric used to compute the alignment distance between the utterances for all the three proposed approaches presented in section 5.5. It measures the average of the squares of the differences between two vectors r_e and $\tilde{r_l}$ bearing D dimensions. It has a range of [0,∞).

$$\mathrm{MSE}(\mathrm{r_e}, \tilde{\mathrm{r_t}}) = c_{MSE}(e,t) = \frac{1}{D}\sum_{i=1}^{D}\left(r_e^i - \tilde{r_t}^i\right)^2 \tag{5.1}$$

5.5 Spoken content mismatch detection approaches

Building upon the DTW-based baseline presented in Chapter 3 we extend it to derive two more variants of DTW namely Phone level cost maximized DTW and Phone level cost maximized Weighted DTW towards unsupervised spoken content mismatch detection. Furthermore, we also propose a cross-attention-based implementation towards achieving the aforementioned goal. All these three approaches are discussed in detail in the sub-sections 5.5.1, 5.5.2, and 5.5.3 respectively.

5.5.1 Phone level cost maximized DTW approach (Ph-DTW):

In the Phone-level cost-maximized DTW approach (Ph-DTW), instead of computing the entire accumulated cost and then normalizing it with the path length as in the case of DTW, we identify the phone-level boundaries and normalize the accumulated cost between the starting and end of the phone with the respective path length covered. Once all the phone-level normalized cost is obtained, we identify all the phonemes belonging to each word. This is again done by obtaining word-level boundaries. The cost per word is then determined as the maximum cost among all the phonemes belonging to the target word. The aforementioned phone and word level boundaries are obtained using the force-alignment process. The ASR [128] utilized for this is trained in-house on Librispeech [57] corpus using Kaldi [63] toolkit.

The idea behind this focuses on making the DTW cost robust to capture even the slightest mismatches between the reference and the target utterance. In a normal DTW setting for a long utterance the overall accumulated cost is normalized with the overall path length at the end of the sequences. Thus the smallest mismatches, say at a single phone somewhere in the entire utterance are quite likely to go unnoticed, as the contribution of a single phone to the entire accumulated cost of DTW sequence for a long speech utterance is negligible. Ph-DTW effectively overcomes this limitation of DTW by normalizing the accumulated cost for each phone with its corresponding path length, thereby highlighting the match quality of each target phone with its corresponding reference phone.

A better understanding of Ph-DTW is showcased in Algorithm 2. Let $\mathcal{R} = \{r_e; 1 \leq e \leq r\}$ and $\mathcal{T} = \{\tilde{r}_t; 1 \leq t \leq T\}$ represent the D-dimensional feature sequences of reference and target utterances, with lengths r and T, respectively. In the Ph-DTW approach, the alignment between these sequences is

Algorithm 2 Phoneme-level Dynamic Time Warping (Ph-DTW) for Utterance Alignment

```
Input: $\mathcal{R} = \{r_1, r_2, ..., r_R\}$ and $\mathcal{T} = \{\tilde{r}_1, \tilde{r}_2, ..., \tilde{r}_T\}$, Phone boundaries $\mathcal{P}$, Word boundaries $\mathcal{W}$;
Initialization: Initialize $\tilde{C}(e, t) \leftarrow \infty$, for all $e, t$; $\tilde{C}(0, 0) \leftarrow 0$;
Distance (cost) $\tilde{C}$ matrix updation:
$e \leftarrow 1, t \leftarrow 1$;
while $e \leq r$ do
    while $t \leq T$ do
        Compute $c(e, t)$;
        $\tilde{C}(e, t)$ = $c(e, t)$ + min[$\tilde{C}(e-1, t)$, $\tilde{C}(e, t-1)$, $\tilde{C}(e-1, t-1)$]; $t \leftarrow t + 1$;
    end
    $e \leftarrow e + 1$;
end
Normalize cost per phone:
for each phone $p$ in $\mathcal{P}$ do
    $S_p, E_p \leftarrow$ start and end indices of phone $p$;
    Normalize $\tilde{C}(E_p, T)$ over the path length from $S_p$ to $E_p$;
end
Compute maximum cost per word:
for each word $w$ in $\mathcal{W}$ do
    Identify phonemes $\mathcal{P}_w$ in word $w$;
    $C_w = \max_{p \in \mathcal{P}_w} \tilde{C}(p)$;
end
Output: Word-level costs $\{C_w\}$ for each word in $\mathcal{W}$
```

Algorithm 3 Phoneme-level Weighted Dynamic Time Warping (Ph-WDTW) for Utterance Alignment

```
Input: $\mathcal{R} = \{r_e; 1 \leq e \leq r\}$ and $\mathcal{T} = \{\tilde{r}_t; 1 \leq t \leq T\}$, Phone boundaries $\mathcal{P}$, Word boundaries $\mathcal{W}$;
Initialization: Initialize $\tilde{C}(e,t) \leftarrow \infty$, for all $e, t$; $\tilde{C}(0,0) \leftarrow 0$;
Distance (cost) $\tilde{C}$ matrix updation:
$e \leftarrow 1, t \leftarrow 1$;
while $e \leq r$ do
    while $t \leq T$ do
        Compute $c(e,t)$;
        $\tilde{C}(e,t)$ = $c(e,t)$ + min[$\tilde{C}(e-1,t)$, $\tilde{C}(e,t-1)$, $\tilde{C}(e-1,t-1) * \sqrt{2}$]; $t \leftarrow t+1$;
    end
    $e \leftarrow e+1$;
end
Normalize cost per phone:
for each phone $p$ in $\mathcal{P}$ do
    $S_p, E_p \leftarrow$ start and end indices of phone $p$;
    Normalize $\tilde{C}(E_p, T)$ over the path length from $S_p$ to $E_p$;
end
Compute maximum cost per word:
for each word $w$ in $\mathcal{W}$ do
    Identify phonemes $\mathcal{P}_w$ in word $w$;
    $C_w = \max_{p \in \mathcal{P}_w} \tilde{C}(p)$;
end
Output: Word-level costs $\{C_w\}$ for each word in $\mathcal{W}$
```

computed using a modified Dynamic Time Warping algorithm. This revised method constructs a cost matrix $\tilde{C}(e,t)$ that accumulates costs up to the e-th frame of $\mathcal{R}$ and the t-th frame of $\mathcal{T}$. The cost at each matrix point includes the local cost $c(e,t)$, which measures the distance between r_e and $\tilde{r}_t$, and is determined using the Equation 5.1. The recursive accumulation at each point considers the minimum of three possible preceding costs: $\tilde{C}(e-1,t)$, $\tilde{C}(e,t-1)$, and $\tilde{C}(e-1,t-1)$.

5.5.2 Phone level cost maximized Weighted DTW approach (Ph-WDTW):

Weighted DTW [129] provides the flexibility to emphasize certain paths based on their relative importance. This leads to better performance in comparison to the vanilla DTW, especially for time-series classification tasks [130] [131]. Extending this idea to Ph-DTW we update the diagonal path weight to $\sqrt{2}$, whereas the non-diagonal path weights remain as 1. The rest operations for Ph-WDTW remain the same as Ph-DTW. A better understanding of Ph-WDTW is showcased in Algorithm 3. While all the steps remain exactly same as Algorithm 2, the recursive accumulation at each point considers the minimum of three possible preceding costs: $\tilde{C}(e-1,t)$, $\tilde{C}(e,t-1)$, and $\tilde{C}(e-1,t-1) * \sqrt{2}$ instead of $\tilde{C}(e-1,t)$, $\tilde{C}(e,t-1)$, and $\tilde{C}(e-1,t-1)$.

5.5.3 Cross attention based approach:

Cross-attention [132] is all about comparing two sequences from different sources and assigning the weights in the range of zero to one that mimics the importance of each token in the target sequence from the reference sequence token acting as the query. This way we get the corresponding scores in the target token referred to as attention scores which implies the importance of the target tokens for the query token. For the computation of cross-attention scores, each reference token acts as the query iteratively. This way we get a matrix of dimension [D_{ref} x D_{target}] where D_{ref} and D_{target} are the total token counts in the reference and the target sequences. Since the attention scores are computed using Equation 5.2, where Q, K and V represent the ”queries”, ”keys”, and ”values”, respectively, which are all inputs into the cross-attention mechanism. In the context of cross-attention specifically, Q comes from one set of data called the reference set and K as well as V comes from a different dataset called the target set. The dot product of Q and K^T is scaled down by $\sqrt{d_k}$ where d_k is the dimensionality of the keys and queries. This scaling helps in stabilizing the gradients during training, as it prevents the softmax function [133] from having extremely small gradients when the dot products are large.

Drawing inspiration from this, we implement the cross-attention architecture with a small modification to compute only the attention map as showcased in Equation 5.3. We only compute the dot product of Q and K^T and normalize it by dividing with $\sqrt{d_k}$ followed by applying the softmax function. Here Q and K represent the intermediate representations of word sequences in the reference and the target set and d_k represents their dimensionalities. These intermediate representations are obtained from Wav2Vec-2.0 and HuBERT. We then compute the MSE loss between the obtained attention map and an identity matrix $I_{D_r X D_t}$ where D_r and D_t is the length of word sequences in any given reference and corresponding target sequence. This is represented in Equation 5.4 where $A_{i,j}$ is the element at the i^{th} row and j^{th} column in the attention map A. $I_{i,j}$ is the element at the i^{th} row and j^{th} column in the identity matrix $I_{D_r \times D_t}$ which equals 1 if $i = j$ and 0 otherwise.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V \tag{5.2}$$

$$\text{Attention map}(Q, K) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) \tag{5.3}$$

$$CrossAttentionLoss_{MSE} = \frac{1}{D_r \cdot D_t} \sum_{i=1}^{D_r} \sum_{j=1}^{D_t} (A_{i,j} - I_{i,j})^2 \tag{5.4}$$

5.6 Classification Approach

We consider the final word level score for Ph-DTW, Ph-WDTW as well as for cross-attention-based approach. Each sentence gives us an array of scores, where the length of the array corresponds to the word count per sentence. We compare the word level scores iteratively to the PRB-based obtained threshold τ for deriving a word level binary label marking the correctly spoken word with 1 and the misspoken word with 0. If the score is less than or equal to τ we say the the target word has been spoken correctly else the target word is misspoken. This rating is then extended to a sentence level wherein if the sentence contains at least one misspoken word it too is labeled as misspoken.

5.7 Experiments

The entire IIITH MM2 Speech-Text dataset was utilized for the experiments for implementing the Ph-DTW and Ph-WDTW approaches. Using the intermediate representations of the correct set of IIITH

MM2 Speech-Text dataset, word-level scores were obtained for all combinations of the same stimulus spoken across all the speakers. Hence for all such combinations a matrix of dimension [$W_{S_{a_i}}$ X $W_{S_{b_i}}$], where $W_{S_{a_i}}$ and $W_{S_{b_i}}$ are the words spoken by speaker S_a and S_b for the i^{th} stimulus of the dataset. All the diagonal values in all such matrices obtained represent the scores between same - same word pairs, whereas all the non-diagonal values represent the scores between different - different word pairs across all speaker combinations. Using these two set of scores, their respective distributions are obtained namely $Dist_{corr}$ and $Dist_{incorr}$. The PRB-based thresholding τ is then obtained based on $Dist_{corr}$ and $Dist_{incorr}$.

However, for the cross-attention-based approach, the selected subset of thirty speakers from Indic TIMIT, as well as the IIITH MM2 Speech-Text dataset, is utilized. The pre-training of the cross-attention model followed by fine-tuning on the selected subsets of Indic TIMIT and IIITH MM2 Speech-Text dataset is performed respectively. Then, the next step is to obtain the threshold τ which is identical to the approach followed for Ph-DTW and Ph-WDTW methods.

Once the threshold τ is obtained, the next step is to obtain word-level scores across all stimuli between the reference speakers(RS_1 and RS_2) and the other speakers, taking into account all the respective speaker utterances in the correct as well as misspoken set and assigning binary label for the same. The performance is then assessed across these obtained scores with classification accuracy being the performance indicator. This approach is repeated for both Wav2vec-2.0 as well as HuBERT-based representations for all three approaches.

5.8 Results

The results of the experiments performed for all three proposed approaches are presented in Table 5.1. It compares the performance of the baseline which is a DTW-based method with the three proposed approaches namely Ph-DTW, Ph-WDTW, and CAA. Performance metrics include Accuracy, Precision, Recall, and F1-score, leveraging speech embeddings from Wav2vec-2.0 and HubERT models. The evaluation is performed under the thresholding criterion of PRB.

The baseline DTW approach achieves an F_1-score of 0.927 with Wav2vec-2.0-based speech embeddings. The Ph-DTW and Ph-WDTW methods show substantial improvements in Recall, especially when employing HubERT embeddings where the highest recall is reported to be 0.992 for the Ph-DTW approach. The CAA approach demonstrates the highest Accuracy of 89.226% and the best F_1-score

Table 5.1: Classification accuracy (Accuracy) (in percentage), Precision, Recall, F_1-score (F_1) at PRB for the Baseline (DTW based approach) as well as all the three proposed approaches Ph-DTW, Ph-WDTW and CAA.

	Embeddings	**Accuracy**	**Precision**	**Recall**	**F$_1$**
Baseline (DTW)	Wav2vec-2.0	77.731	0.871	0.870	0.927
Ph-DTW	Wav2vec-2.0	86.780	0.883	0.970	0.927
	HubERT	86.849	0.873	**0.992**	0.929
Ph-WDTW	Wav2vec-2.0	86.832	0.871	0.990	0.929
	HubERT	86.884	0.868	0.990	0.929
CAA	Wav2vec-2.0	88.913	**0.967**	0.900	0.933
	HubERT	**89.226**	0.959	0.914	**0.936**

of 0.936 with HubERT embeddings, reflecting an overall superior performance in terms of precision and recall balance. These results suggest that the proposed approaches outsmart the presented baseline. These findings demonstrate that the proposed approaches especially CAA significantly advance the performance for the unsupervised spoken content mismatch detection task over the conventional DTW approach.

Chapter 6

Conclusion

This thesis consolidates a set of approaches towards developing solutions in automatic spoken data validation for non-native English speakers in the Indian context, which is crucial for enhancing HCI systems. This work has unfolded across multiple layers of analysis and experimentation, culminating in the approaches that significantly improve the process of automatic spoken data validation.

The first chapter of this thesis introduces the problem statement and motivates the necessity of developing unsupervised spoken data validation approaches. Chapter two then introduces the IIITH MM2 Speech-Text dataset, a unique corpus that includes both matched and naturally misspoken read speech utterances from a diverse group of Indian speakers. The dataset is not only robust due to its incorporation of various Indian nativities but also versatile, facilitating the development and evaluation of algorithms designed for automatic mismatch detection and correction. Experimental baselines established using DTW and Wav2vec-2.0 representations have shown promising results, setting a strong foundation for further enhancements and expansion of the dataset.

In chapter three, the focus shifts to the pronunciation quality assessment of second language learners using an unsupervised approach that utilizes DTW between expert and learner speech representations from Wav2Vec-2.0. This is done to highlight that indeed the choice of self-supervised features is a good choice for speech representation. The analysis highlighted the effectiveness of synthesized speech as a viable alternative to human expert speech in CALL systems, marking a significant advancement in the accessibility and scalability of pronunciation training tools. Although the approach showed substantial improvements in most assessed factors, further research is required to enhance intelligibility assessments, suggesting a direction for future studies.

Chapter four addresses the automatic validation of speech data with the help of enhanced DTW approaches namely Ph-DTW and Ph-WDTW as well as a cross-attention mechanism-based approach.

These approaches are implemented on speech representations obtained from self-supervised models such as Wav2Vec-2.0 and HuBERT. The results from applying these methods on the Indic TIMIT and IIITH MM2 Speech-Text datasets showcase the significance and reliability of these automated approaches over traditional manual validation in terms of accuracy, precision, and recall.

Overall, this thesis acts as the stepping stone towards reliable automatic spoken data validation. The showcased approaches effectively handle the challenges posed by India's linguistic diversity, enhancing the potential of HCI systems to work efficiently across speech from diverse speaker backgrounds. Moving forward, the continued exploration and refinement of these validation techniques will undoubtedly contribute to the development of more robust and inclusive speech-based applications.

Chapter 7

Publications

[1] N. Anand, M. Sirigiraju, and C. Yarra, “Iiith mm2 speech-text: A preliminary data for auto- matic spoken data validation with matched and mismatched speech-text content,” in 2023 26th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Stan- dardisation of Speech Databases and Assessment Techniques (O-COCOSDA). IEEE, 2023, pp. 1–6.

[2] N. Anand, M. Sirigiraju, and C. Yarra,, “Unsupervised pronunciation assessment analysis using utterance level alignment distance with self-supervised representations,” in 2023 IEEE 20th India Council International Conference (INDICON). IEEE, 2023, pp. 409–414.

[3] N. Anand, M. Sirigiraju, and C. Yarra, “Unsupervised spoken content mismatch detection,” in preparation for ICASSP 2025, IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). IEEE, 2025, submission deadline August 2024.

Bibliography

[1] E. Caveness, P. S. GC, Z. Peng, N. Polyzotis, S. Roy, and M. Zinkevich, “Tensorflow data validation: Data analysis and validation in continuous ml pipelines,” in *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*, 2020, pp. 2793–2796. 3

[2] C. Xie, J. Gao, and C. Tao, “Big data validation case study,” in *2017 IEEE third international conference on big data computing service and applications (BigDataService)*. IEEE, 2017, pp. 281–286. 3

[3] L. E. Lwakatare, E. Rånge, I. Crnkovic, and J. Bosch, “On the experiences of adopting automated data validation in an industrial machine learning project,” in *2021 IEEE/ACM 43rd International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)*. IEEE, 2021, pp. 248–257. 3

[4] T. Johnson and T. Dasu, “T3: Data quality and data cleaning: An overview,” 2003. 3

[5] E. Breck, N. Polyzotis, S. Roy, S. Whang, and M. Zinkevich, “Data validation for machine learning.” in *MLSys*, 2019. 4

[6] X. Zhu, C. Vondrick, C. C. Fowlkes, and D. Ramanan, “Do we need more training data?” *International Journal of Computer Vision*, vol. 119, no. 1, pp. 76–92, 2016. 4

[7] T. Manjunath, R. S. Hegadi, and H. Mohan, “Automated data validation for data migration security,” *International Journal of Computer Applications*, vol. 30, no. 6, pp. 41–46, 2011. 4

[8] F. Biessmann, J. Golebiowski, T. Rukat, D. Lange, and P. Schmidt, “Automated data validation in machine learning systems,” 2021. 4

[9] S. Savanur and K. Shreedhara, “Automated data validation for data warehouse testing,” in *2016 International Conference on Electrical, Electronics, Communication, Computer and Optimization Techniques (ICEECCOT)*. IEEE, 2016, pp. 223–226. 4

[10] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, *et al.*, “A survey of large language models,” *arXiv preprint arXiv:2303.18223*, 2023. 5

[11] M. I. Jordan and T. M. Mitchell, “Machine learning: Trends, perspectives, and prospects,” *Science*, vol. 349, no. 6245, pp. 255–260, 2015. 5

[12] J. Schmidhuber, “Deep learning in neural networks: An overview,” *Neural networks*, vol. 61, pp. 85–117, 2015. 5

[13] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016. 5

[14] L. Liu, W. Ouyang, X. Wang, P. Fieguth, J. Chen, X. Liu, and M. Pietikäinen, “Deep learning for generic object detection: A survey,” *International journal of computer vision*, vol. 128, pp. 261–318, 2020. 5

[15] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai, *et al.*, “Recent advances in convolutional neural networks,” *Pattern recognition*, vol. 77, pp. 354–377, 2018. 5

[16] V. Chandola, A. Banerjee, and V. Kumar, “Anomaly detection: A survey,” *ACM computing surveys (CSUR)*, vol. 41, no. 3, pp. 1–58, 2009. 9

[17] M. Markou and S. Singh, “Novelty detection: a review—part 1: statistical approaches,” *Signal processing*, vol. 83, no. 12, pp. 2481–2497, 2003. 9

[18] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, “Lof: identifying density-based local outliers,” in *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, 2000, pp. 93–104. 9

[19] B. Schölkopf, R. C. Williamson, A. Smola, J. Shawe-Taylor, and J. Platt, “Support vector method for novelty detection,” *Advances in neural information processing systems*, vol. 12, 1999. 9

[20] F. T. Liu, K. M. Ting, and Z.-H. Zhou, “Isolation forest,” in *2008 eighth ieee international conference on data mining*. IEEE, 2008, pp. 413–422. 9

[21] T. Schlegl, P. Seeböck, S. M. Waldstein, U. Schmidt-Erfurth, and G. Langs, “Unsupervised anomaly detection with generative adversarial networks to guide marker discovery,” in *International conference on information processing in medical imaging*. Springer, 2017, pp. 146–157. 9

[22] D. Hendrycks and K. Gimpel, “A baseline for detecting misclassified and out-of-distribution examples in neural networks,” *arXiv preprint arXiv:1610.02136*, 2016. 9

[23] S. Liang, Y. Li, and R. Srikant, “Enhancing the reliability of out-of-distribution image detection in neural networks,” *arXiv preprint arXiv:1706.02690*, 2017. 9

[24] K. Lee, H. Lee, K. Lee, and J. Shin, “Training confidence-calibrated classifiers for detecting out-of-distribution samples,” *arXiv preprint arXiv:1711.09325*, 2017. 9

[25] A. Krizhevsky, G. Hinton, *et al.*, “Learning multiple layers of features from tiny images,” 2009. 9

[26] A. Torralba, R. Fergus, and W. T. Freeman, “80 million tiny images: A large data set for nonparametric object and scene recognition,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 30, no. 11, pp. 1958–1970, 2008. 9

[27] F. Yu, A. Seff, Y. Zhang, S. Song, T. Funkhouser, and J. Xiao, “Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop,” *arXiv preprint arXiv:1506.03365*, 2015. 9

[28] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, A. Y. Ng, *et al.*, “Reading digits in natural images with unsupervised feature learning,” in *NIPS workshop on deep learning and unsupervised feature learning*, vol. 2011, no. 5. Granada, Spain, 2011, p. 7. 9

[29] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255. 9

[30] H. Choi, E. Jang, and A. A. Alemi, “Waic, but why? generative ensembles for robust anomaly detection,” *arXiv preprint arXiv:1810.01392*, 2018. 9

[31] G. Shalev, Y. Adi, and J. Keshet, “Out-of-distribution detection using multiple semantic label representations,” *Advances in Neural Information Processing Systems*, vol. 31, 2018. 9

[32] A. A. Alemi, I. Fischer, and J. V. Dillon, “Uncertainty in the variational information bottleneck,” *arXiv preprint arXiv:1807.00906*, 2018. 9

[33] A. A. Alemi, I. Fischer, J. V. Dillon, and K. Murphy, “Deep variational information bottleneck,” *arXiv preprint arXiv:1612.00410*, 2016. 9

[34] D. Hendrycks, M. Mazeika, and T. Dietterich, “Deep anomaly detection with outlier exposure,” *arXiv preprint arXiv:1812.04606*, 2018. 10

[35] N. Hynes, D. Sculley, and M. Terry, “The data linter: Lightweight, automated sanity checking for ml data sets,” in *NIPS MLSys Workshop*, vol. 1, no. 5, 2017. 10

[36] S. Schelter, F. Biessmann, D. Lange, T. Rukat, P. Schmidt, S. Seufert, P. Brunelle, and A. Taptunov, “Unit testing data with deequ,” in *Proceedings of the 2019 International Conference on Management of Data*, 2019, pp. 1993–1996. 10

[37] S. Schelter, D. Lange, P. Schmidt, M. Celikel, F. Biessmann, and A. Grafberger, “Automating large-scale data quality verification,” *Proceedings of the VLDB Endowment*, vol. 11, no. 12, pp. 1781–1794, 2018. 10

[38] S. Krishnan, M. J. Franklin, K. Goldberg, J. Wang, and E. Wu, “Activeclean: An interactive data cleaning framework for modern machine learning,” in *Proceedings of the 2016 International Conference on Management of Data*, 2016, pp. 2117–2120. 10

[39] T. Rekatsinas, X. Chu, I. F. Ilyas, and C. Ré, “Holoclean: Holistic data repairs with probabilistic inference,” *arXiv preprint arXiv:1702.00820*, 2017. 10

[40] S. Krishnan, M. J. Franklin, K. Goldberg, and E. Wu, “Boostclean: Automated error detection and repair for machine learning,” *arXiv preprint arXiv:1711.01299*, 2017. 10

[41] C. Gokhale, S. Das, A. Doan, J. F. Naughton, N. Rampalli, J. Shavlik, and X. Zhu, “Corleone: Hands-off crowdsourcing for entity matching,” in *Proceedings of the 2014 ACM SIGMOD international conference on Management of data*, 2014, pp. 601–612. 10

[42] B. Mozafari, P. Sarkar, M. Franklin, M. Jordan, and S. Madden, “Scaling up crowd-sourcing to very large datasets: a case for active learning,” *Proceedings of the VLDB Endowment*, vol. 8, no. 2, pp. 125–136, 2014. 10

[43] M. Yakout, L. Berti-Équille, and A. K. Elmagarmid, “Don’t be scared: use scalable automatic repairing with maximal likelihood and bounded changes,” in *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*, 2013, pp. 553–564. 10

[44] C. Mayfield, J. Neville, and S. Prabhakar, “Eracer: a database approach for statistical inference and data cleaning,” in *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*, 2010, pp. 75–86. 10

[45] H. Müller and J. C. Freytag, *Problems, methods, and challenges in comprehensive data cleansing*. Professoren des Inst. Für Informatik, 2005. 10

[46] A. K. Elmagarmid, P. G. Ipeirotis, and V. S. Verykios, “Duplicate record detection: A survey,” *IEEE Transactions on knowledge and data engineering*, vol. 19, no. 1, pp. 1–16, 2006. 10

[47] H.-W. Ng and S. Winkler, “A data-driven approach to cleaning large face datasets,” in *2014 IEEE international conference on image processing (ICIP)*. IEEE, 2014, pp. 343–347. 10

[48] S. Schelter, D. Lange, P. Schmidt, M. Celikel, F. Biessmann, and A. Grafberger, “Automating large-scale data quality verification,” *Proceedings of the VLDB Endowment*, vol. 11, no. 12, pp. 1781–1794, 2018. 11

[49] S. Rabanser, S. Günnemann, and Z. Lipton, “Failing loudly: An empirical study of methods for detecting dataset shift,” *Advances in Neural Information Processing Systems*, vol. 32, 2019. 11

[50] N. Polyzotis, M. Zinkevich, S. Roy, E. Breck, and S. Whang, “Data validation for machine learning,” *Proceedings of machine learning and systems*, vol. 1, pp. 334–347, 2019. 11

[51] E. Breck, N. Polyzotis, S. Roy, S. Whang, and M. Zinkevich, “Data validation for machine learning.” in *MLSys*, 2019. 11

[52] A. Swami, S. Vasudevan, and J. Huyn, “Data sentinel: A declarative production-scale data validation platform,” in *2020 IEEE 36th International Conference on Data Engineering (ICDE)*. IEEE, 2020, pp. 1579–1590. 11

[53] L. E. Lwakatare, E. Rånge, I. Crnkovic, and J. Bosch, “On the experiences of adopting automated data validation in an industrial machine learning project,” in *2021 IEEE/ACM 43rd International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)*. IEEE, 2021, pp. 248–257. 11

[54] Y. Yang, S. Kang, and J. Seo, “Improved machine reading comprehension using data validation for weakly labeled data,” *IEEE Access*, vol. 8, pp. 5667–5677, 2020. 11

[55] M. Joshi, E. Choi, D. S. Weld, and L. Zettlemoyer, “Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension,” *arXiv preprint arXiv:1705.03551*, 2017. 11

[56] I. Iriondo, S. Planet, J.-C. Socoró, E. Martínez, F. Alías, and C. Monzo, “Automatic refinement of an expressive speech corpus assembling subjective perception and automatic classification,” *Speech Communication*, vol. 51, no. 9, pp. 744–758, 2009. 11

[57] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: an asr corpus based on public domain audio books,” in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210. 11, 21, 42

[58] A. Diwan, R. Vaideeswaran, S. Shah, A. Singh, S. Raghavan, S. Khare, V. Unni, S. Vyas, A. Rajpuria, C. Yarra, *et al.*, “Multilingual and code-switching asr challenges for low resource indian languages,” *arXiv preprint arXiv:2104.00235*, 2021. 11

[59] I. Păvăloi and E. Muscă, “Experimental study in development of speech corpus for emotion recognition with data validation,” in *2015 International Symposium on Signals, Circuits and Systems (ISSCS)*. IEEE, 2015, pp. 1–4. 11

[60] J. S. Garofolo, “Timit acoustic phonetic continuous speech corpus,” *Linguistic Data Consortium, 1993*, 1993. 13

[61] H. Sakoe and S. Chiba, “Dynamic programming algorithm optimization for spoken word recognition,” *IEEE transactions on acoustics, speech, and signal processing*, vol. 26, no. 1, pp. 43–49, 1978. 15

[62] V. Zue, S. Seneff, and J. Glass, “Speech database development at mit: Timit and beyond,” *Speech communication*, vol. 9, no. 4, pp. 351–356, 1990. 15

[63] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. CONF. IEEE Signal Processing Society, 2011. 21, 42

[64] K. Levin, K. Henry, A. Jansen, and K. Livescu, "Fixed-dimensional acoustic embeddings of variable-length segments in low-resource settings," in *2013 IEEE workshop on automatic speech recognition and understanding*. IEEE, 2013, pp. 410–415. 21

[65] M. A. Carlin, S. Thomas, A. Jansen, and H. Hermansky, "Rapid evaluation of speech representations for spoken term discovery," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011. 21, 36

[66] H. Hermansky, "Perceptual linear predictive (plp) analysis of speech," *the Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990. 21

[67] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020. 22

[68] F. Javanmardi, S. Tirronen, M. Kodali, S. R. Kadiri, and P. Alku, "Wav2vec-based detection and severity level classification of dysarthria from speech," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5. 22, 24, 36

[69] J. Shah, Y. K. Singla, C. Chen, and R. R. Shah, "What all do audio transformer models hear? probing acoustic representations for language delivery and its structure," *arXiv preprint arXiv:2101.00387*, 2021. 22, 24, 36

[70] M. Buckland and F. Gey, "The relationship between recall and precision," *Journal of the American society for information science*, vol. 45, no. 1, pp. 12–19, 1994. 23

[71] J. Davis and M. Goadrich, "The relationship between precision-recall and roc curves," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 233–240. 23

[72] C. Goutte and E. Gaussier, “A probabilistic interpretation of precision, recall and f-score, with implication for evaluation,” in *European conference on information retrieval*. Springer, 2005, pp. 345–359. 23

[73] J.-M. Cheng and H.-C. Wang, “A method of estimating the equal error rate for automatic speaker verification,” in *2004 International Symposium on Chinese Spoken Language Processing*. IEEE, 2004, pp. 285–288. 24, 31

[74] J. Qi, J. Du, S. M. Siniscalchi, X. Ma, and C.-H. Lee, “On mean absolute error for deep neural network based vector-to-vector regression,” *IEEE Signal Processing Letters*, vol. 27, pp. 1485–1489, 2020. 24

[75] Z. Wang and A. C. Bovik, “Mean squared error: Love it or leave it? a new look at signal fidelity measures,” *IEEE signal processing magazine*, vol. 26, no. 1, pp. 98–117, 2009. 24, 36

[76] C. Yarra, A. Srinivasan, C. Srinivasa, R. Aggarwal, and P. K. Ghosh, “voistutor corpus: A speech corpus of indian l2 english learners for pronunciation assessment,” in *2019 22nd Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*. IEEE, 2019, pp. 1–6. 24, 26

[77] M. Senoussaoui, P. Kenny, T. Stafylakis, and P. Dumouchel, “A study of the cosine distance-based mean shift for telephone speech diarization,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 1, pp. 217–227, 2013. 25

[78] M. H. Lim and V. Aryadoust, “A scientometric review of research trends in computer-assisted language learning (1977–2020),” *Computer Assisted Language Learning*, vol. 35, no. 9, pp. 2675–2700, 2022. 25

[79] V. Ramanarayanan, P. L. Lange, K. Evanini, H. R. Molloy, and D. Suendermann-Oeft, “Human and automated scoring of fluency, pronunciation and intonation during human-machine spoken dialog interactions.” in *INTERSPEECH*, 2017, pp. 1711–1715. 25

[80] M. Yang, K. Hirschi, S. D. Looney, O. Kang, and J. H. Hansen, “Improving mispronunciation detection with wav2vec2-based momentum pseudo-labeling for accentedness and intelligibility assessment,” *arXiv preprint arXiv:2203.15937*, 2022. 25

[81] J. Tepperman and S. Narayanan, “Automatic syllable stress detection using prosodic features for pronunciation evaluation of language learners,” in *Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, vol. 1. IEEE, 2005, pp. I–937. 25

[82] C. Zhu, R. Hakoda, D. Saito, N. Minematsu, N. Nakanishi, and T. Nishimura, “Multi-granularity annotation of instantaneous intelligibility of learners' utterances based on shadowing techniques,” in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2021, pp. 1071–1078. 25, 26

[83] Z. Miodonska, M. D. Bugdol, and M. Krecichwost, “Dynamic time warping in phoneme modeling for fast pronunciation error detection,” *Computers in Biology and Medicine*, vol. 69, pp. 277–285, 2016. 25, 26

[84] J. P. Arias, N. B. Yoma, and H. Vivanco, “Automatic intonation assessment for computer aided language learning,” *Speech communication*, vol. 52, no. 3, pp. 254–267, 2010. 25, 26

[85] A. Lee and J. Glass, “A comparison-based approach to mispronunciation detection,” in *2012 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2012, pp. 382–387. 25, 26

[86] J. Yue, F. Shiozawa, S. Toyama, Y. Yamauchi, K. Ito, D. Saito, and N. Minematsu, “Automatic scoring of shadowing speech based on dnn posteriors and their dtw.” in *INTERSPEECH*, 2017, pp. 1422–1426. 25, 26

[87] S. Liu, A. Mallol-Ragolta, E. Parada-Cabaleiro, K. Qian, X. Jing, A. Kathan, B. Hu, and B. W. Schuller, “Audio self-supervised learning: A survey,” *Patterns*, vol. 3, no. 12, 2022. 26

[88] S.-w. Yang, P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G.-T. Lin, *et al.*, “Superb: Speech processing universal performance benchmark,” *arXiv preprint arXiv:2105.01051*, 2021. 26

[89] F. Javanmardi, S. Tirronen, M. Kodali, S. R. Kadiri, and P. Alku, “Wav2vec-based detection and severity level classification of dysarthria from speech,” in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5. 26, 28

[90] J. Shah, Y. K. Singla, C. Chen, and R. R. Shah, “What all do audio transformer models hear? probing acoustic representations for language delivery and its structure,” *arXiv preprint arXiv:2101.00387*, 2021. 26, 28

[91] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020. 26

[92] S. Bannò and M. Matassoni, “Proficiency assessment of l2 spoken english using wav2vec 2.0,” in *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2023, pp. 1088–1095. 26

[93] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020. 28

[94] C. Yi, J. Wang, N. Cheng, S. Zhou, and B. Xu, “Applying wav2vec2. 0 to speech recognition in various low-resource languages,” *arXiv preprint arXiv:2012.12121*, 2020. 28

[95] R. Jain, A. Barcovschi, M. Yiwere, D. Bigioi, P. Corcoran, and H. Cucu, “A wav2vec2-based experimental study on self-supervised learning methods to improve child speech recognition.” *IEEE Access*, 2023. 28

[96] Z. Fan, M. Li, S. Zhou, and B. Xu, “Exploring wav2vec 2.0 on speaker verification and language identification,” *arXiv preprint arXiv:2012.06185*, 2020. 28

[97] D. Hendrycks and K. Gimpel, “Gaussian error linear units (gelus),” *arXiv preprint arXiv:1606.08415*, 2016. 28

[98] N. Vaessen and D. A. Van Leeuwen, “Fine-tuning wav2vec2 for speaker recognition,” in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7967–7971. 28

[99] M. K. Baskar, T. Herzig, D. Nguyen, M. Diez, T. Polzehl, L. Burget, J. Černockỳ, *et al.*, “Speaker adaptation for wav2vec2 based dysarthric asr,” *arXiv preprint arXiv:2204.00770*, 2022. 28

[100] C. Yi, J. Wang, N. Cheng, S. Zhou, and B. Xu, “Transfer ability of monolingual wav2vec2. 0 for low-resource speech recognition,” in *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2021, pp. 1–6. 28

[101] M. Yang, K. Hirschi, S. D. Looney, O. Kang, and J. H. Hansen, “Improving mispronunciation detection with wav2vec2-based momentum pseudo-labeling for accentedness and intelligibility assessment,” *arXiv preprint arXiv:2203.15937*, 2022. 28

[102] X. Xu, Y. Kang, S. Cao, B. Lin, and L. Ma, “Explore wav2vec 2.0 for mispronunciation detection.” in *Interspeech*, 2021, pp. 4428–4432. 28

[103] M. Kunešová and Z. Zajíc, “Multitask detection of speaker changes, overlapping speech and voice activity using wav2vec 2.0,” in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5. 28

[104] M. Kunešová and M. Řezáčková, “Detection of prosodic boundaries in speech using wav2vec 2.0,” in *Text, Speech, and Dialogue: 25th International Conference, TSD 2022, Brno, Czech Republic, September 6–9, 2022, Proceedings*. Springer, 2022, pp. 377–388. 28

[105] Y. Wang, A. Boumadane, and A. Heba, “A fine-tuned wav2vec 2.0/hubert benchmark for speech emotion recognition, speaker verification and spoken language understanding,” *arXiv preprint arXiv:2111.02735*, 2021. 28

[106] P. Tzirakis, A. Baird, J. Brooks, C. Gagne, L. Kim, M. Opara, C. Gregory, J. Metrick, G. Boseck, V. Tiruvadi, *et al.*, “Large-scale nonverbal vocalization detection using transformers,” in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5. 28

[107] S. Sheikh, M. Sahidullah, F. Hirsch, and S. Ouni, “Introducing ecapa-tdnn and wav2vec2. 0 embeddings to stuttering detection,” in *Submitted to Interspeech 2022*, 2022. 28

[108] L. Gauder, L. Pepino, L. Ferrer, and P. Riera, “Alzheimer disease recognition using speech-based embeddings from pre-trained models.” in *Interspeech*, 2021, pp. 3795–3799. 28

[109] Y. Getman, R. Al-Ghezi, K. Voskoboinik, T. Grósz, M. Kurimo, G. Salvi, T. Svendsen, and S. Strömbergsson, “Wav2vec2-based speech rating system for children with speech sound disorder,” in *Interspeech*, 2022. 28

[110] H. Sakoe and S. Chiba, “Dynamic programming algorithm optimization for spoken word recognition,” *IEEE transactions on acoustics, speech, and signal processing*, vol. 26, no. 1, pp. 43–49, 1978. 29

[111] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, “Hubert: Self-supervised speech representation learning by masked prediction of hidden units,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021. 36

[112] C. Yarra, R. Aggarwal, A. Rajpal, and P. K. Ghosh, “Indic timit and indic english lexicon: A speech database of indian speakers using timit stimuli and a lexicon from their mispronunciations,” in *2019 22nd Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*. IEEE, 2019, pp. 1–6. 36, 37

[113] N. Anand, M. Sirigiraju, and C. Yarra, “Iiith mm2 speech-text: A preliminary data for automatic spoken data validation with matched and mismatched speech-text content,” in *2023 26th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, 2023, pp. 1–6. 36, 37

[114] J. S. Garofolo, “Timit acoustic phonetic continuous speech corpus,” *Linguistic Data Consortium, 1993*, 1993. 37

[115] P. Senin, “Dynamic time warping algorithm review,” *Information and Computer Science Department University of Hawaii at Manoa Honolulu, USA*, vol. 855, no. 1-23, p. 40, 2008. 39

[116] A. Mao, M. Mohri, and Y. Zhong, “Cross-entropy loss functions: Theoretical analysis and applications,” in *International Conference on Machine Learning*. PMLR, 2023, pp. 23 803–23 828. 41

[117] M. Caron, P. Bojanowski, A. Joulin, and M. Douze, “Deep clustering for unsupervised learning of visual features,” in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 132–149. 41

[118] X. Chang, T. Maekaku, P. Guo, J. Shi, Y.-J. Lu, A. S. Subramanian, T. Wang, S.-w. Yang, Y. Tsao, H.-y. Lee, and S. Watanabe, “An exploration of self-supervised pretrained representations for

end-to-end speech recognition," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2021, pp. 228–235. 41

[119] J. W. Yoon, B. J. Woo, and N. S. Kim, "Hubert-ee: Early exiting hubert for efficient speech recognition," *arXiv preprint arXiv:2204.06328*, 2022. 41

[120] M. A. Pastor, D. Ribas, A. Ortega, A. Miguel, and E. Lleida, "Cross-corpus speech emotion recognition with hubert self-supervised representation," in *IberSPEECH 2022*. ISCA, 2022, pp. 76–80. 41

[121] E. Morais, R. Hoory, W. Zhu, I. Gat, M. Damasceno, and H. Aronowitz, "Speech emotion recognition using self-supervised features," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6922–6926. 41

[122] I.-H. Chu, Z. Chen, X. Yu, M. Han, J. Xiao, and P. Chang, "Self-supervised cross-modal pretraining for speech emotion recognition and sentiment analysis," in *Findings of the Association for Computational Linguistics: EMNLP 2022*, Y. Goldberg, Z. Kozareva, and Y. Zhang, Eds. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 5105–5114. [Online]. Available: https://aclanthology.org/2022.findings-emnlp.375 41

[123] J. Lin, M. Ge, W. Wang, H. Li, and M. Feng, "Selective hubert: Self-supervised pre-training for target speaker in clean and mixture speech," *IEEE Signal Processing Letters*, pp. 1–5, 2024. 41

[124] Y. Wang, A. Boumadane, and A. Heba, "A fine-tuned wav2vec 2.0/hubert benchmark for speech emotion recognition, speaker verification and spoken language understanding," *arXiv preprint arXiv:2111.02735*, 2021. 41

[125] C. Yu, X. Su, and Z. Qian, "Multi-stage audio-visual fusion for dysarthric speech recognition with pre-trained models," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 31, pp. 1912–1921, 2023. 41

[126] A. Hernandez, P. A. Pérez-Toro, E. Nöth, J. R. Orozco-Arroyave, A. Maier, and S. H. Yang, "Cross-lingual self-supervised speech representations for improved dysarthric speech recognition," *arXiv preprint arXiv:2204.01670*, 2022. 41

[127] E. Kim, J.-J. Jeon, H. Seo, and H. Kim, "Automatic pronunciation assessment using self-supervised speech representation learning," *arXiv preprint arXiv:2204.03863*, 2022. 41

[128] R. Esposito and M. Hendriks, “Literature review of modelling approaches for asr in concrete: a new perspective,” *European Journal of Environmental and Civil Engineering*, vol. 23, no. 11, pp. 1311–1331, 2019. 42

[129] H. Li, “Time works well: Dynamic time warping based on time weighting for time series data mining,” *Information Sciences*, vol. 547, pp. 592–608, 2021. 45

[130] Y.-S. Jeong, M. K. Jeong, and O. A. Omitaomu, “Weighted dynamic time warping for time series classification,” *Pattern recognition*, vol. 44, no. 9, pp. 2231–2240, 2011. 45

[131] D. Toshniwal and R. C. Joshi, “Similarity search in time series data using time weighted slopes,” *Informatica*, vol. 29, no. 1, 2005. 45

[132] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” 2017. [Online]. Available: https://arxiv.org/pdf/1706.03762.pdf 45

[133] D. Dwivedi, A. Ganguly, and V. Haragopal, “6 - contrast between simple and complex classification algorithms,” in *Statistical Modeling in Machine Learning*, T. Goswami and G. Sinha, Eds. Academic Press, 2023, pp. 93–110. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780323917766000166 45