

# Fairness of Automatic Speech Recognition in Cleft Lip and Palate Speech

Susmita Bhattacharjee<sup>a,\*</sup>, Jagabandhu Mishra<sup>b,\*</sup>, H.S. Shekhawat<sup>a</sup>, S. R. Mahadeva Prasanna<sup>c</sup>

<sup>a</sup>*Department of Electronics and Electrical Engineering, Indian Institute of Technology Guwahati, Guwahati, 781039, Assam, India*

<sup>b</sup>*School of Computing, University of Eastern Finland, Joensuu, Finland*

<sup>c</sup>*Department of Electrical Engineering, Indian Institute of Information Technology Dharwad (IIIT Dharwad), Dharwad, India*

---

## Abstract

Speech produced by individuals with cleft lip and palate (CLP) is often highly nasalized and breathy due to structural anomalies, causing shifts in formant structure that affect automatic speech recognition (ASR) performance and fairness. This study hypothesizes that publicly available ASR systems exhibit reduced fairness for CLP speech and confirms this through experiments. Despite formant disruptions, mild and moderate CLP speech retains some spectro-temporal alignment with normal speech, motivating augmentation strategies to enhance fairness. The study systematically explores augmenting CLP speech with normal speech across severity levels and evaluates its impact on ASR fairness. Three ASR models—GMM-HMM, Whisper, and XLS-R—were tested on AIISH and NMCPC datasets. Results indicate that training with normal speech and testing on mixed data improves word error rate (WER). Notably, WER decreased from 22.64% to 18.76% (GMM-HMM, AIISH) and 28.45% to 18.89% (Whisper, NMCPC). The superior performance of GMM-HMM on AIISH may be due to its suitability for Kannada children's speech, a challenge for foundation models like XLS-R and Whisper. To assess fairness, a fairness score was introduced, revealing improvements of 17.89% (AIISH) and 47.50% (NMCPC) with augmentation.

**Keywords:** Automatic speech recognition, XLS-R, Whisper, KALDI, Fairness objective score

---

## 1. Introduction

CLP is a congenital abnormality of the craniofacial region [1, 2, 3]. Globally, the incident of CLP was recorded to be 1,92,708 in 2019 [4]. The abnormality in the craniofacial region causes defects in the produced speech [1, 2, 3, 5]. Mostly, due to the opening between the lip and nasal cavity, the produced speech is breathy and highly nasalized [6], [7]. Further, the opening between the oral and nasal cavity shifts the speech resonances [1], [8]. Hence, the produced speech is not like normal speech. However, to avoid the digital divide, paying attention and making the speech technologies fair for CLP speech is necessary.

There exist several attempts in the literature using CLP speech, mostly to perform classification and intelligibility enhancement. The work in [9], utilizes wav2vec2 embeddings and uses them with machine learning classifiers and in [10], transformer classifier to perform normal and CLP classification. [11] reports the intelligibility of the speech can be improved by compensating the hypernasality. The work in [12], proposed a technique using glottal activity region to compute the hypernasality score, whereas in [13] deep learning approaches are used to estimate hypernasality. In [14], the work proposed a procedure to detect the stop consonant and trill sounds and then replace them to enhance the speech perception. The paper [15] used a MaskCycleGAN approach to enhance the CLP speech and show the improvement in the phone decoding performance, using the GMM-HMM. Further, several works also attempt to enhance the intelligibility of CLP speech using the generative end-to-end deep learning frameworks [16, 17]. However, as per the authors' knowledge, no work exists in the literature, that studies the fairness of ASR in CLP speech.

This study first examines the performance gap in ASR between normal and CLP speech by evaluating them using the state-of-the-art Google ASR API. To further analyze the similarities and differences between normal and

---

\*Corresponding authors: Susmita Bhattacharjee and Jagabandhu Mishra.

Email addresses: [sbhattacharjee@iitg.ac.in](mailto:sbhattacharjee@iitg.ac.in) (Susmita Bhattacharjee), [jagabandhu.mishra@uef.fi](mailto:jagabandhu.mishra@uef.fi) (Jagabandhu Mishra), [h.s.shekhwat@iitg.ac.in](mailto:h.s.shekhwat@iitg.ac.in) (H.S. Shekhawat), [prasanna@iitdh.ac.in](mailto:prasanna@iitdh.ac.in) (S. R. Mahadeva Prasanna)

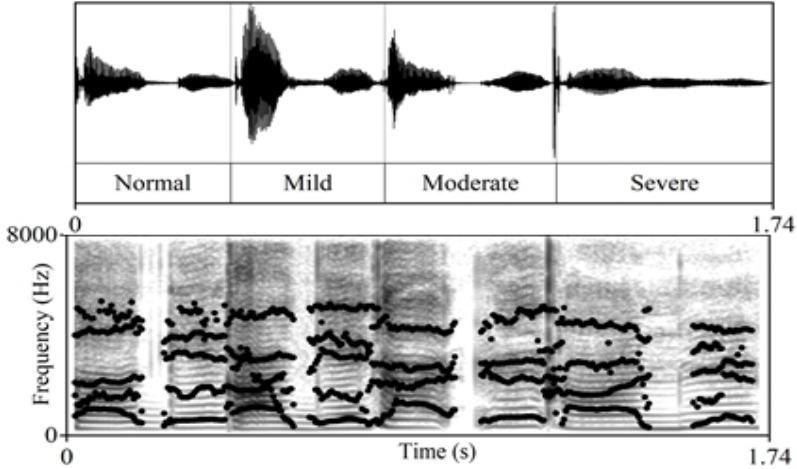


Figure 1: Speech signal produced by Normal, mild, moderate and severe subjects, and their corresponding spectrogram and formant contour, while producing the sound *"kage"*.

CLP speech, we conduct cross-training and testing between the two speech types. Given the available data, we evaluate ASR performance using both the traditional GMM-HMM approach and foundation models, namely cross-lingual self-supervised speech representation (XLS-R) and Web-scale Supervised Pretraining for Speech Recognition (Whisper). XLS-R [18] leverages self-supervised learning, making it highly effective for low-resource, cross-lingual, and multilingual scenarios. In contrast, Whisper [19] is trained using supervised learning, enabling it to generalize across multiple speech tasks with minimal fine-tuning while maintaining strong robustness to noisy data.

Figure 1 illustrates the speech signal, spectrogram, and formant contours for speech produced by normal, mild, moderate, and severely affected CLP subjects. The figure indicates a progressive shift in formant locations from normal to severe, likely due to the increasing degree of oral-nasal tract opening. However, while the shape of the formant contours remains largely intact in normal and mild speech, it gradually deteriorates in moderate and severe cases. This suggests that ASR performance may decline due to formant frequency shifts.

However, the preservation of formant contour shapes, particularly in mild cases, motivates further investigation into strategies for developing a fair ASR system. Inspired by this, we train the ASR system by augmenting normal speech with various combinations of CLP speech across severity levels. The resulting performance is then analyzed with a focus on system fairness. The main contributions of this work are summarized as follows:

1. This study evaluates the fairness of publicly available ASR systems for CLP speech and confirms its adverse impact on recognition performance.
2. An augmentation strategy combining CLP and normal speech is proposed to enhance ASR fairness across varying severity levels.
3. A fairness score is introduced to quantify and assess improvements in ASR performance for CLP speech.

## 2. Related studies on CLP speech

Research on cleft lip and palate (CLP) speech has primarily focused on hypernasality detection, intelligibility assessment, and speech enhancement. Hypernasality, a key characteristic of CLP speech, has been extensively studied using various signal processing techniques. Early works such as [20] and [21] employed Zero Time Windowing (ZTW) and Homomorphic processing-based cepstral features for hypernasality detection, with support vector machines (SVMs) used for classification. Other studies, including [22], analyzed formant frequency shifts to measure vowel space area reduction in CLP children. Additionally, [23] proposed a posterior probability-based approach for estimating hypernasality scores, overcoming limitations of traditional perceptual evaluation methods. Deep learning-based methods, such as the attention-based BLSTM models [24], have further improved automatic hypernasality detection, reducing dependence on specialized clinical expertise. Studies such as [25] and [26] introduced pitch-adaptive features and constant-Q cepstral coefficients (CQCC) to enhance classification accuracy, while [27] explored

sinusoidal model-based features for robust hypernasality detection. More recent works, such as [28], investigated advanced spectral and articulatory-based features for improved classification.

Intelligibility assessment studies have employed a variety of methodologies to estimate speech intelligibility, including Gaussian posteriograms [13], dynamic time warping (DTW) [29], and regression models [30]. These approaches have demonstrated strong correlations with human perceptual ratings, underscoring their effectiveness for objective intelligibility assessment. For instance, studies like [31] have analyzed spectral moments of fricatives in CLP speech, while [32] leveraged epoch-synchronous features to detect nasalized voiced stops. In the realm of speech enhancement, significant efforts have been made to improve intelligibility using both signal processing and deep learning techniques. Early works, such as [33], focused on modifying phoneme transitions, whereas [34] proposed enhancements for hypernasal speech by altering vocal tract system characteristics. Misarticulated fricative and stop consonants were addressed in [35] and [36] through spectral transformations. More recent advancements have utilized deep learning-based methods, such as CycleGAN in [37], which significantly improved WER in ASR evaluations for CLP speech. Additionally, [38] demonstrated enhanced CLP speech recognition performance using data augmentation techniques like vocal tract length perturbation and reverberation. And here are some recent works such as [39] which does the classification between CLP and healthy voices with latent representations from the lower and middle encoder layers of a pre-trained wav2vec 2.0 system, reached an accuracy of 100% and [40] which focussed on improving hypernasality with ASR in CLP.

A summary of these works is provided in Table 1. However, despite these advancements, none of the existing studies address the fairness of ASR systems for CLP speech. In this work, we aim to bridge this gap by evaluating the fairness of ASR systems for CLP speech and exploring the application of data augmentation tailored to the severity levels of CLP speech.

### 3. Database setup

This section provides a brief description of the database used in this study. There are relatively very few datasets available on CLP speech [45, 46, 47]. This work used two datasets to perform the experiments. Initially, the analysis is performed with All India Institute of Speech and Hearing (AIISH) [48], and then, for generalization purposes, some experiments are repeated with the New Mexico Cleft Palate Centre (NMCPC) dataset [45]. The former is in the Kannada language, whereas the latter is in English. Both datasets are randomly divided into training and testing partitions with 80 : 20 ratio. After that, the training partition is further partitioned to 80 : 20 to form the training and development set.

**AIISH dataset:** The dataset is collected from the All India Institute of Speech and Hearing, India. It consists of 60 speakers with 31 normal and 29 CLP speakers. Out of them, 19 and 12 are normal female and male speakers respectively, and 9 and 20 are CLP female and male speakers, respectively. The dataset is collected with 19 unique utterances. Each sentence has a maximum of 3 words. The participants are native Kannada speakers, who are within the age group 7 – 12 years. They did not have any other congenital syndromes like hearing impairment. The detailed statistics of the dataset are given in Table 2. Out of a total of 2,726 utterances, the dataset is partitioned as follows: training (1,731 utterances – 1,106 normal, 625 CLP), development (429 utterances – 278 normal, 151 CLP), and evaluation (566 utterances – 357 normal, 209 CLP).

**NMCPC dataset:** The dataset is collected at the New Mexico Cleft Palate Centre. It has a total of 65 speakers and consists of speech utterances from 41 CLP speakers (22 male and 19 females) and 24 normal speakers (20 male and 4 females). The dataset consists of 76 unique utterances. Each sentence has a maximum of 5 words. The age group of the speakers is 9 – 13 years. The CLP speakers were classified into mild, moderate and severe. The detailed statistics of the dataset are given in Table 2 [45]. Out of a total of 1,463 utterances, the dataset is divided into training (929 utterances – 649 normal, 280 CLP), development (235 utterances – 165 normal, 70 CLP), and evaluation (299 utterances – 89 normal, 210 CLP).

### 4. Fairness of publicly available Google application programming interface (API) ASR in CLP speech

This section discusses the performance of publicly available Google API ASR and its fairness when used with CLP speech.

Table 1: Summary of work done so far in CLP

Sl.	Ref, Year	Task	Dataset	Method
1	[20] , 2016	Hypernasality Detection	AIISH	Based on presence/absence of extra nasal peak in low, high and voiced consonants in the HNGD spectrum, the severity rating of the hypernasal speech can be decided for /a/ and /i/ vowels and voice consonants /b/, /d/ or /g/
2	[33] , 2018	Intelligibility improvement	AIISH	This study enhances intelligibility of /s/ substituted by a glottal stop by inserting sustained portions and modifying transitions using 2D-DCT projections onto normal speech SVD vectors.
3	[14] , 2018	Misarticulated trills analysis	AIISH	Acoustic analysis of misarticulated trills in CLP children using glottal and vocal tract features shows significant differences from normal speech. A DTW-based system using trill-specific features outperforms MFCCs in detecting misarticulations.
4	[41] , 2018	Glottal Activity Errors Detection in Stops	AIISH	The proposed algorithm detects glottal activity errors (GAE) in stop consonant production of CLP speakers using low-frequency voiced consonant evidence from zero-frequency filtering (ZFFS) and band-pass filtering (BPFS)
5	[23] , 2018	Hypernasality estimation	AIISH	Motivated by the functionality of nasometer, a posterior probability-based approach is proposed here which estimates hypernasality scores using MFCCs from glottal regions. DNN outperforms GMM and also nasometer.
6	[12] , 2018	Intelligibility assessment	AIISH	This study analyzes CLP speech intelligibility using glottal landmarks (g LMs) and acoustic features, showing that Mel-2DDCT-based GMMs outperform MFCCs by better capturing abrupt transitions and correlating with perceptual ratings.
7	[13] , 2018	Intelligibility assessment	AIISH	This study uses GP-based speech representation and DTW distance to assess CLP child speech intelligibility, showing that pitch-normalized Mel-2D-DCT features best correlate with SLP perceptual ratings, outperforming MFCCs and LP-2D-DCT features.
8	[29] , 2018	Intelligibility Assessment	AIISH	This study proposes an SSM-based unsupervised framework for estimating CLP children's speech intelligibility, showing that GP-based SSMs outperform MFCCs and DTW in correlating with perceptual ratings and discriminating intelligibility groups.
9	[21] , 2018	Hypernasality Detection & assessment	AIISH	This study introduces the HNGDF cepstral feature for hypernasality detection, showing superior accuracy over EDM features, with further improvement when combined with MFCCs, making it promising for hypernasality severity analysis.
10	[25] , 2018	Hypernasality Detection	AIISH	This study proposes Pitch-Adaptive MFCC (PAMFCC) for hypernasality detection, improving low-frequency nasality cue capture and achieving higher classification accuracy than MFCCs by mitigating pitch harmonics effects in CLP speech.
11	[31] , 2019	Study of voiceless sibilant fricatives	AIISH	This work analyzes NAE-affected voiceless sibilant fricatives in Kannada, showing spectral deviations due to VPD. An SVM classifier using spectral moments and peak ERBN-number achieves high accuracy in detecting NAE-distorted fricatives.
12	[32] , 2019	Segmentation & detection of nasalized voiced stops	AIISH	This study proposes an automatic segmentation and detection algorithm for nasalized voiced stops in CP speech, using glottal activity and spectral features to enhance SVM-based classification, outperforming HMM-based segmentation and MFCCs.
13	[30] , 2019	Composite measure of speech intelligibility	AIISH	This study proposes a composite intelligibility measure for CLP speech using articulation and hypernasality features, with SVR achieving the best prediction of PCC scores using wM2DDCT, wMFCC, and gMFCC features.
14	[28] , 2019	Hypernasality Detection	AIISH	This study detects hypernasality using VTC, PSR, and SMAC features, capturing spectral distortions in vowels. SVM classification with combined features outperforms baselines for both detection and severity classification.
15	[26] , 2019	Hypernasality severity Detection	AIISH	This study detects hypernasality severity in /i/ and /u/ vowels using CQCC features, which capture nasal formant variations more effectively than MFCCs, improving accuracy but with challenges in mild case classification.
16	[27] , 2020	Hypernasality Detection	AIISH	This study proposes NHA, HAR, and PHF features for hypernasality detection using a sinusoidal speech model, with SVM classification showing that their combination outperforms individual features and baseline methods.
17	[34] , 2020	Enhancement of Vowels	AIISH	This study explores hypernasal speech enhancement using XLP residual modification, vocal tract system modification, and their combination, with evaluations showing the combined approach most effectively reduces nasalization.
18	[42] , 2020	Detection of misarticulated stops	AIISH	This study segments CV transitions in CLP speech using VOPs and SPF-based 2D-DCT features, with SVM classification outperforming STFT-based 2D-DCT, MFCCs, and HMM models in detecting misarticulated stops.
19	[35] , 2021	Modification of fricatives	AIISH	This study modifies misarticulated /s/ in CLP speech using spectral adjustments and synthesized insertions, improving spectral similarity and intelligibility, though MOS scores remain below normal.
20	[36] , 2021	Misarticulated stops Enhancement	AIISH	This study enhances CLP speech intelligibility by modifying misarticulated stops using NMF-based spectral transformation, improving detection and perceptual similarity, though MOS ratings suggest room for further quality enhancement.
21	[43] , 2021	Hypernasality Assessment	Americleft, NMCPC	This study proposes OHM, a DNN-based hypernasality assessment metric trained on healthy speech, achieving high correlation with expert ratings and sensitivity to mild hypernasality, performing comparably to clinicians.
22	[37] , 2021	Intelligibility Enhancement	AIISH	This study uses CycleGAN to enhance CLP children's speech intelligibility, with ASR and subjective evaluations confirming improvements, benefiting speech-controlled device usability and therapy outcomes.
23	[38] , 2021	Data augmentation for improving CLP ASR	AIISH	This study explores data augmentation for CLP speech recognition, with CycleGAN, VTLP, and reverberation showing the best improvements, significantly reducing phone error rates.
24	[44] , 2021	Data augmentation based on Frequency Warping	ATR Japanese speech	This paper proposes frequency warping for data augmentation in CLP speech ASR, enhancing robustness to formant fluctuations and improving accuracy when combined with self-supervised learning, outperforming SpecAugment.
25	[15] , 2022	Intelligibility Enhancement	AIISH	This paper proposes MaskCycleGAN for data augmentation in CLP speech ASR and obtained better WER outperforming CycleGAN.
26	[40] , 2022	Hypernasality Estimation	CNH-CLP, NMCPC	This study improves hypernasality estimation by fine-tuning a pre-trained ASR encoder, leveraging larger ASR datasets and text labels for better feature extraction, achieving superior performance on cleft palate datasets.
27	[39] , 2023	Classification of CLP & Normal	Erlangen-CLP	Classification between CLP and healthy voices with latent representations from the lower and middle encoder layers of a pre-trained wav2vec 2.0 system, reached an accuracy of 100%.
27	[10] , 2023	Classification of CLP & Normal using transformers	AIISH, NMCPC	This study fine-tunes pretrained transformer models on CLP speech, showing superior classification performance, with DistilHuBERT achieving near 100% accuracy.
29	Ours, 2025	Fairness of ASR in CLP Speech	AIISH, NMCPC	This study examines ASR fairness for CLP speech, evaluating GMM-HMM, Whisper, and XLS-R systems and exploring the impact of augmenting CLP speech with varying severity levels of normal speech.

Table 2: Description of AIISH and NMCPC [45] datasets

Dataset	Type of audio Data	Severity of Speaker	Subjects	Utterances	Total (Normal / CLP)
AIISH	Normal	-	31	1741	1741 / 985
	CLP	Mild	14	473	
		Moderate	11	379	
		Severe	4	133	
NMCPC	Normal	-	24	439	439 / 1024
	CLP	Mild	11	385	
		Moderate	14	324	
		Severe	16	315	

#### 4.1. Google API ASR performance in CLP speech

We use publicly available English and Kannada ASR from Google <sup>1</sup> to evaluate the performance of the NMCPC and AIISH datasets, respectively. We use the evaluation set of both datasets to evaluate performance. The performance obtained in terms of WER is tabulated in Table 3 [45].

After decoding, the performance in terms of word error rate (WER) is evaluated by comparing them with the ground truth text transcription. The performance obtained in terms of WER for AIISH dataset is 93% and 98.94% in the normal and CLP test sets, respectively, and for NMCPC dataset 30.21% and 74.27% in the normal and CLP test sets, respectively, for the same test set utterances. The performance degradation from 93% to 98.94% for AIISH and 30.21% to 74.27% for NMCPC justifies the claim that the fairness of the ASR system is compromised in CLP speech. These findings emphasize the significant performance gap in ASR systems when processing CLP speech, highlighting concerns about fairness and accessibility. The substantial increase in WER for CLP utterances suggests that current ASR models struggle with disordered speech, necessitating targeted improvements in acoustic modeling and adaptation techniques. Furthermore, it should be noted that the relatively high WER for normal speech specifically for AIISH can be attributed to the challenges of child speech recognition. The literature also suggests that ASR performance decreases with child speech [49]. This highlights the need for child-inclusive ASR training.

Table 3: Fairness of ASR available publicly through Google API,  $W_C$ ,  $W_N$  are the obtained WER from CLP and normal test utterances and fairness score, respectively for AIISH & NMCPC datasets.

Dataset	$W_N \downarrow$	$W_C \downarrow$
AIISH	93.00	98.94
NMCPC	30.21	74.27

#### 4.2. Fairness as a metric

Inspired by the work [50, 51], we use fairness metrics to observe the degree of fairness of the ASR system. The Fairness Score (FS) is defined as a weighted combination of the negative average error rate and the error disparity between two groups. The average error rate is the mean of the error rates across the two groups, and error disparity is the absolute difference in error rates between the two groups. The same is defined in the Equations below,

The fairness score (FS) is computed as follows:

$$FS = -\alpha \cdot \text{Average Error Rate} - \beta \cdot \text{Error Disparity}; \alpha, \beta \geq 0$$

where:

1. **Average Error Rate:**

---

<sup>1</sup><https://cloud.google.com/speech-to-text>

$$\text{Average Error Rate} = \frac{\text{Error}(G_1) + \text{Error}(G_2)}{2}$$

Here,  $\text{Error}(G_1)$  and  $\text{Error}(G_2)$  are the error rates for the two groups, here the group  $G_1$  indicates Normal and  $G_2$  refers to CLP.

## 2. Error Disparity:

$$\text{Error Disparity} = |\text{Error}(G_1) - \text{Error}(G_2)|$$

3.  $\alpha$  and  $\beta$  are weights to balance the importance of overall error minimization and fairness (disparity minimization).

The  $-\alpha$  and  $-\beta$  indicate that higher fairness leads to a higher  $FS$ . The range of  $FS$  is  $-\infty \leq FS \leq 0$ . The value of  $FS$  closer to zero signifies **better fairness** of the system, and vice versa. That is, the value of  $FS$  represents the **degree of fairness** of the system. From the given equation, the average error rate is influenced by  $\text{error}(G_1)$  which is the WER for Normal speech and  $\text{error}(G_2)$  which is WER for CLP speech. If either has a high value, the overall average error rate increases. A lower average error rate indicates a better overall performance. The error disparity is determined by the absolute difference between  $\text{error}(G_1)$  and  $\text{error}(G_2)$ . A large disparity indicates a significant difference in ASR performance between normal and CLP speech, meaning the system is less fair. A smaller disparity suggests a more balanced ASR performance between both groups and indicates a more fair system. WER values for normal and CLP speech directly affect both the average error rate and the error disparity, affecting the overall fairness score ( $FS$ ).

The fairness score and WER obtained using the Google API are tabulated in Table 4. The scores obtained in the AIISH and NMCPC evaluation sets are  $-50.95$  and  $-48.15$ , respectively. Also, a good system has both a low average error rate and error disparity, resulting in a fairness score close to zero. Fairness scores farther from zero indicate greater unfairness toward CLP speech. The maximum degradation in normal speech is mainly due to the noisy environment. The CLP speech is mainly spoken by adult doctors. The state of the art WER for ASR is  $5 - 10\%$ . It was noted that the main error in the normal case was substitution error whereas for CLP, the main error was insertion error. In AIISH, CLP eval set Google API was not able to provide the text output for 11 utterances out of 209 text files, NMCPC CLP eval set, Google API was not able to provide the text output for 26 utterances out of 210 text files. The WER of the existing files was obtained as  $98.94\%$  for AIISH and  $74.27\%$  for NMCPC with  $\alpha, \beta = 0.5$ .

Table 4: Fairness of ASR available publicly through Google API,  $W_C$ ,  $W_N$ , and  $FS$  are the obtained WER from CLP and normal test utterances and fairness score, respectively for AIISH & NMCPC datasets.

Dataset	$W_N \downarrow$	$W_C \downarrow$	$FS \downarrow$
AIISH	93.00	98.94	-50.95
NMCPC	30.21	74.27	-48.15

## 5. Proposed approach for CLP ASR

In this work, we used ASR systems that included GMM-HMM, Whisper, and XLS-R models. A comparison study was initially performed to understand the distortions in CLP speech. The aim is to observe the distortions and, accordingly, propose a strategy to improve fairness. Figure 2(a-h), shows the speech signal and the corresponding spectrogram of normal, mild, moderate, and severe speech utterances, respectively. Variations in amplitude and duration suggest differences in speech articulation and phonation among the different severity levels. The visual observation suggests that, compared to normal, the spectral resonance pattern is less distorted in mild and the distortion increases gradually from mild to moderate and severe. To better understand the same, a distance-based study is performed considering three utterances from each category (i.e., normal, mild, moderate, and severe). Figure 2(e-h), present the frequency content over time for each speech signal. The yellow bands indicate voiced regions, and the differences in the spectrograms reflect spectral deviations due to increasing severity in nasalization or articulation distortions. The more severe cases may exhibit spectral smearing or reduced harmonics due to increased hypernasality or reduced phonatory control. First, from the speech signals, the spectrogram is obtained considering 20 milliseconds as framesize and 10 milliseconds as frameshift. Then, using an energy-based voice activity detector (VAD), only voiced

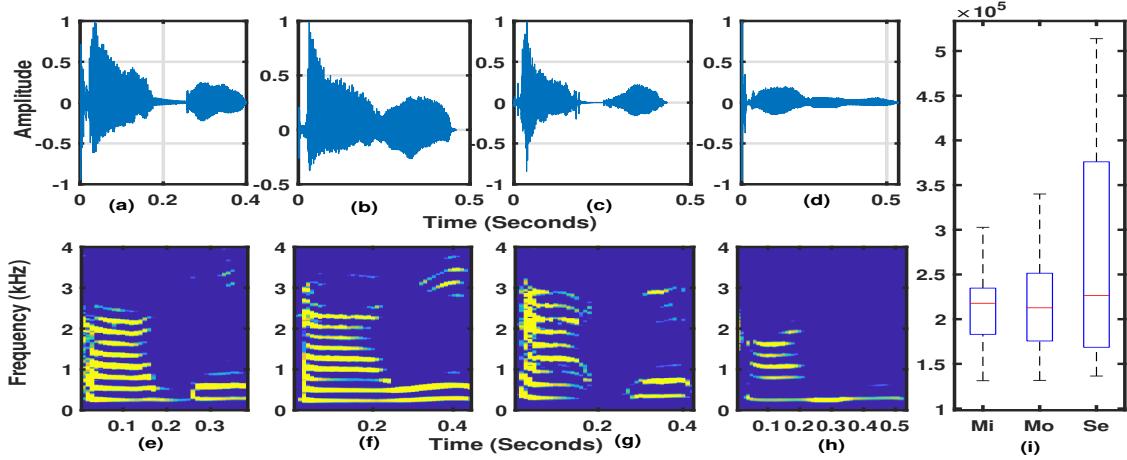


Figure 2: (a-h) Speech signal produced by Normal, mild, moderate, and severe subjects, and their corresponding spectrogram, and (i) showing the DTW distance distribution between the voiced spectrogram of normal to mild (Mi), moderate (Mo), and severe (Se), respectively, while producing the sound *"kage"*.

frames are detected. The reason only voiced frames are detected is because they contain harmonics and formants, which are crucial for analyzing nasalization, articulation distortions and phonatory control. Also, hypernasality affects voiced sounds primarily making them more informative than unvoiced segments. By using energy-based VAD, non-speech regions and background noise are removed, for a cleaner analysis. The VAD considers 6% of the average energy as the threshold to decide on the voice/unvoice frame. After that, using voiced spectrogram frames of each normal utterance, the dynamic time warping (DTW) [52] distance is calculated with the utterances of the mild, moderate and severe categories. Figure 2(i) shows the distance distribution obtained. The figure suggests that the spectral distance gradually increases from mild to moderate and severe. This justifies the claim that distortion gradually increases from mild to moderate and severe. The mild category (Mi) has the smallest spread and range of values, indicating more consistent or tightly grouped values. The moderate category (Mo) shows a higher spread, and the severe (Se) category shows a wide distribution, with higher variability, suggesting that severe category has the highest variability, indicating greater differences in speech patterns. Hence DTW results further validate our hypothesis, allowing us to proceed with the experiments. The DTW optimal alignment technique formula is represented as below

$$DTW(i, j) = d(i, j) + \min \begin{cases} DTW(i - 1, j) \\ DTW(i, j - 1) \\ DTW(i - 1, j - 1) \end{cases}$$

- $DTW(i, j)$  represents the accumulated cost at point  $(i, j)$ .
- $d(i, j)$  is the distance between the elements of two sequences at indices  $i$  and  $j$ .
- The recursion considers the minimum cost among:
  1. Insertion:  $DTW(i - 1, j)$
  2. Deletion:  $DTW(i, j - 1)$
  3. Match:  $DTW(i - 1, j - 1)$

The preservation of formant contour shapes particularly in mild cases as seen in 1, motivates further investigation into strategies for developing a fair ASR system. Building on this motivation, we analyze the impact of data augmentation. We train the ASR system by augmenting normal speech with various combinations of CLP speech across different severity levels. The performance resulting from this cross-testing, along with WER analysis by the foundation models, is then evaluated with a focus on system fairness.

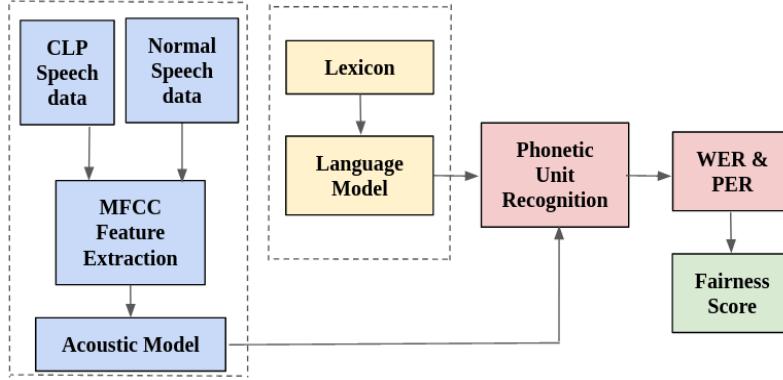


Figure 3: Fairness in ASR computation pipeline .

### 5.1. Description of the used ASR models

The block diagram in Figure 3 shows the pipeline of obtaining Fairness score from raw speech data. The input speech data (CLP and normal speech) goes through feature extraction (MFCC), followed by model evaluation (acoustic model, lexicon, and language model). After phonetic unit recognition, the performance is assessed through WER and PER. Finally, the fairness of the model is evaluated using the fairness score given in section 4.2. It illustrates the overall structure of a system that performs both speech recognition and fairness assessment, providing a framework for understanding how different components of the ASR pipeline work together to improve both recognition accuracy and fairness in speech processing systems.

A self supervised Cross-lingual Representation Learning for Speech Recognition (XLS-R) is a multilingual speech recognition model which is built on wav2vec 2.0. It is trained over 100 languages which includes Kannada as one of the Indian language. The robust self-supervised learning framework of Wav2Vec2 is combined with a multilingual training strategy in XLS-R to create a versatile ASR model that can handle multiple languages. The essential actions consist of using contrastive loss for self-supervised pretraining on multilingual data and adjusting the pretrained model using CTC loss on particular ASR tasks.

XLS-R's ability to learn language-agnostic speech representations makes it particularly effective for cross-lingual and multilingual tasks because it can generalize across languages. This allows it to be more efficient in recognizing speech, even in languages with limited resources, and without the need for language-specific models. Wav2Vec2's multilingual capability is extended by XLS-R through the use of multilingual training data. Learning language-neutral speech representations that are cross-linguistically generalizable is the main objective. The following actions are taken to accomplish this:

**Multilingual Pretraining:** The model can learn universal speech features in the pretraining stage by using speech data from multiple languages. Despite training on different languages, XLS-R applies a single contrastive loss function across all of them. The contrastive loss helps the model learn language-agnostic speech representations by pulling together similar speech segments (positive pairs) and pushing apart dissimilar ones (negative pairs). Instead of training separate models for different languages, the same loss function is applied uniformly, allowing the model to learn shared features across languages. This approach enables cross-lingual transfer, meaning that learning from high-resource languages can improve performance on low-resource languages.

**Language-Agnostic Representations:** The model learns common speech features that are applicable to various languages by training on a variety of languages. The model does not rely on language-specific features. Instead, it captures universal acoustic and phonetic patterns that are common across different languages. This allows the model to generalize well across languages, even for those it has not seen much during training.

#### 5.1.1. Gaussian Mixture Model-Hidden-Markov model (GMM-HMM)

The GMM-HMM [53] combination uses HMMs to model temporal dynamics and GMMs to model the probability distribution of acoustic features. In Kaldi, the emission probabilities of HMM states are modeled using the GMM. Every HMM state has a corresponding GMM, where the emission probability is a combination of various Gaussian

distributions [54, 55, 56]. An HMM consists of  $N$  states, which often stand for various phonemes or speech-related sub-phonetic components.

### 5.1.2. XLS-R and Whisper

#### Self-supervised cross-lingual speech representation learning (XLS-R)

1. **Feature Extractor** XLS-R [18] uses multilingual training data to expand Wav2Vec2.0’s ability to handle various languages. It has a self supervised approach but trains on cross-lingual data using multilingual datasets trained on 436k hours of speech containing over 128 languages. There are overlapping frames created from the raw audio input ( $y$ ). A convolutional neural network (CNN), ( $c(y)$ ) processes each frame in order to generate latent speech representations ( $z$ ).

$$z = \text{CNN}(y) \quad (1)$$

- **Contextual Transformer:**

Context representations ( $c$ ) are produced by passing the latent representations ( $z$ ) through a Transformer network ( $T(z)$ ) in order to extract long-range dependencies.

$$c = T(z) \quad (2)$$

For training the model, feature encoder representations are discretized with a quantization module to represent the targets in the self-supervised learning objective.

- **Contrastive Loss:**

The context representations  $c$  are used to predict the true latent representation  $z$  for masked time steps  $t$ , denoted by  $c_t$ . It predicts the masked frames based on the unmasked context.

$$\mathcal{L}_{\text{cont}} = -\log \frac{\exp(\text{sim}(\mathbf{c}_t, \mathbf{z}_{\text{true}}))}{\sum_{z \in Z} \exp(\text{sim}(\mathbf{c}_t, z))} \quad (3)$$

where  $Z$  is the set of latent representations. Applying this loss across multiple languages encourages the model to learn representations that perform well in a range of linguistic contexts.  $\text{sim}$  is the cosine similarity [57] measure which is a measure of similarity between two non-zero vectors defined in an inner product space.

- **Diversity Loss**

$$\mathcal{L}_{\text{div}} = \frac{1}{C} \sum_{c=1}^C \left( \sum_{k=1}^K \mathbf{p}_{c,k} \log \mathbf{p}_{c,k} \right) \quad (4)$$

where  $C$  is the number of codebooks,  $K$  is the size of each codebook and  $p_k$  is the probability of selecting the code  $k$ .

## 2. Cross-Lingual training Stage

The XLS-R model is trained on 128 languages and generalizes well across different linguistic contexts. It creates a shared representation space across languages using the multilingual data and thereby assures robustness to language-specific variations. After pretraining, XLS-R can be fine-tuned on language specific tasks like LID, ASR.

#### Web-scale Supervised Pretraining for Speech Recognition (Whisper)

The Whisper [19] model for ASR is based on transformer architecture which uses a supervised approach for training and trained on 99 languages for 680k hours of data. It allows efficient processing and transcription of multilingual speech with high accuracy. It includes spectral analysis for preprocessing followed by self-attention mechanisms in the encoder-decoder structure, and finally probabilistic decoding techniques during inference. The encoder  $E$  extracts latent representations,  $z$  from the spectrogram input,  $y$ . It is then passed through an attention-based mechanism in order to capture long-range dependencies across time [19] represented as:

$$z = \mathbf{E}(y) \quad (5)$$

The latent representations are then passed to the decoder,  $D$  for generating target sequence,  $y_T$ . The attention mechanism application allows selecting tokens. The target sequence consists of this tokenized text.

$$y_t = \mathbf{D}(z, y_{1:t-1}) \quad (6)$$

A scaled dot-product attention mechanism is employed between the query  $Q$ , key  $K$ , and value  $V$  matrices to compute attention scores which is calculated as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{D_k}}\right)V \quad (7)$$

$D_k$  represents the dimensionality of the key vectors.

The cross-entropy loss between actual tokens  $y_t$  and predicted tokens  $\hat{y}_t$  is given as:

$$L = \sum_{t=1}^T y_t \log(\hat{y}_t) \quad (8)$$

Besides, to provide information about the positions of the input frames Positional encodings are also used in Whisper. Finally, greedy decoding or beam search is applied by the model for inference purpose to get the most probable sequence of output tokens,  $Y$  given as:

$$Y = \arg \max_Y P(Y | y) \quad (9)$$

### 5.2. Fairness of the proposed ASR approach

We utilized the XLS-R and Whisper models to evaluate the performance of our ASR system. These models were selected for their robustness in handling diverse speech patterns and we applied these models to the datasets with varying degrees of severity. After training the system with augmented data, we computed fairness by analyzing how well the models performed across different speech types, focusing on any biases or disparities in error rates between normal and CLP speech across severity levels. This analysis helps ensure that the ASR system delivers equitable performance for all speech types, without favoring any particular category or group.

## 6. Experimental Setup and Results

To ensure a structured evaluation, we designed our experiments in a two-step process. First, we trained and tested the GMM-HMM model using conventional feature extraction methods. Next, we applied transformer-based ASR models (Whisper, XLS-R) to the same dataset. This approach allowed us to compare performance across different architectures.

For fairness evaluation, we cross-tested models by augmenting normal speech with various levels of CLP speech and analyzing their word error rate (WER) and phoneme error rate (PER) across different severity categories. The XLS-R model was fine-tuned using pre-trained weights from Hugging Face, while GMM-HMM relied on handcrafted MFCC features. The Whisper model employed in this work has 244 million parameters and the XLS-R has 300 million parameters.

To further assess robustness, we also experimented with data augmentation, incorporating mild and moderate CLP speech into the training process. This provided insights into how model biases were affected by exposure to different speech patterns.

### 6.1. Experimental setup

The NMCPC and AIISH datasets are used with GMM-HMM, Whisper and XLS-R for ASR task. Kaldi is a Conventional pipeline-based method utilizing language modeling, acoustic modeling (GMM-HMM), and explicit feature extraction (MFCC, for example). The architecture includes GMM-HMM and DNN-based models. From each utterance, the MFCC features are extracted using a 10 msec frameshift with 20 msec window length. The considered number of senones and Gaussians are 50 and 500, respectively. The speech signals are processed in frames of 20 ms with a shift of 10 ms and characterized by 1024 fast-fourier transform (FFT) points. After training the decoding is done using the Viterbi algorithm. Wav2Vec 2.0 (XLS-R) is a Self-supervised learning from unprocessed audio waveforms. Pre-trained on large, unlabeled datasets, fine-tuned on labeled data; CNN is used for feature extraction and transformer for context. Large unlabeled datasets are needed for pre-training, while less labeled data is required for fine-tuning. It is highly adaptable and flexible. It makes multilingual model training easier; less domain expertise is required to access powerful pre-trained models. It is PyTorch-implemented and emphasizes contemporary deep learning workflows. In brief Kaldi is adaptable and a conventional ASR toolkit; requires significant labeled data, high customization, and expertise. XLS-R, on the contrary, is a cutting-edge, comprehensive deep learning method that makes use of self-supervised learning to minimize the need for labeled data and streamline multilingual ASR developed on Hugging Face which is an open-source machine learning platform<sup>2</sup>.

#### 6.1.1. Experimental setup for ASR models

In GMM-HMM and XLS-R the transcriptions were formulated by keeping speaker id and utterances together. The training partition in all the experiments is used to train the wav2vec2 feature extractor for XLS-R and GMM-HMM takes 39 dimensional MFCC+ $\Delta$  +  $\Delta\Delta$  feature vectors as input. Wav2Vec 2.0's feature extractor functions as the model's first stage, converting unprocessed audio input into latent speech representations that identify specific waveform patterns. A sequence of CNN layers make up this stage mainly [58, 59]. The metrics used for estimating the performance are WER and phoneme error rate (PER).

#### 6.1.2. Setup for fairness ( $\alpha, \beta$ )

The fairness analysis of the ASR models was conducted using various combinations of alpha ( $\alpha$ ) and beta ( $\beta$ ) values from the fairness scores equation to evaluate the performance of the system across three combinations of weights of Average Error Rate and Error Disparity namely ( $\alpha = 0.1, \beta = 0.9$ ), ( $\alpha = 0.5, \beta = 0.5$ ) and ( $\alpha = 0.9, \beta = 0.1$ ). The analysis was done on the GMM-HMM and Whisper models trained and tested on AIISH and NMCPC datasets. The first case has higher weight on average error rate ( $\alpha = 0.9, \beta = 0.1$ ), i.e., when overall accuracy is the priority is relevant if the application demands high accuracy across all groups, even at the cost of fairness. Secondly, if data imbalance exists and one group (e.g., normal) has significantly more samples than the other, focusing on reducing total error may help improve generalization. In the second case where a higher weight on error disparity ( $\alpha = 0.1, \beta = 0.9$ ) is applicable when fairness is the primary concern in applications where equal treatment of groups is crucial such as speech disorder classification where CLP should not be disproportionately misclassified or in ethical requirements to ensure equal performance across groups where minimizing error disparity becomes more critical. Finally the balanced case ( $\alpha = 0.5, \beta = 0.5$ ) is useful when both accuracy and fairness are equally important, ensuring a trade-off between minimizing overall errors and maintaining similar performance across groups. At times, a system may perform better, but the performance may be skewed to one class. But we need the performance to be good and fair. On the other hand, some cases require a better WER instead of fairness and, depending on the requirements, different ranges of  $\alpha$  and  $\beta$  can be considered. For the sake of comparison, two ends have been considered as 0.1 and 0.9.

### 6.2. Experimental results

This section presents the performance analysis of three different ASR models GMM-HMM, XLS-R, and Whisper on CLP and normal speech. The results are evaluated using WER and PER, and further analyzed for fairness across different severity levels. The initial cross-testing experiments indicate that WER and PER increase as speech severity

---

<sup>2</sup><https://huggingface.co>

Table 5: Performance of criss-cross analysis for **AIISH**. Total: total CLP test set.

Model	Test→		Normal	CLP			
	Train↓			Mild	Moderate	Severe	Total
<b>GMM-HMM</b>	<b>Normal</b>	WER	2.39	22	22.50	84.17	42.89
		PER	34.6	56.49	58.08	77.53	64.03
	<b>CLP</b>	WER	2.66	11	16.67	81.67	36.44
		PER	48.87	55.56	58.08	75.07	62.90
<b>XLS-R</b>	<b>Normal</b>	WER	7.53	27	30.41	113.33	56.91
	<b>CLP</b>	WER	24.73	25.33	26.25	105.83	52.47
<b>WHISPER</b>	<b>Normal</b>	WER	2.21	35	34.58	90	53.19
	<b>CLP</b>	WER	4.25	28.66	40	98.33	55.66

Table 6: Performance of criss-cross analysis for **NMCPC**. Total: total CLP test set.

Model	Test→		Normal	CLP			
	Train↓			Mild	Moderate	Severe	Total
<b>GMM-HMM</b>	<b>Normal</b>	WER	19.03	46.44	64.84	87.19	66.15
		PER	65.17	80.24	80.3	90.02	83.52
	<b>CLP</b>	WER	19.03	40.68	39.45	81.4	53.84
		PER	90.6	88.96	86.62	88.69	88.09
<b>XLS-R</b>	<b>Normal</b>	WER	32.93	50.16	79.68	95.86	75.23
	<b>CLP</b>	WER	47.73	44.74	55.07	69.83	56.54
<b>WHISPER</b>	<b>Normal</b>	WER	7.03	9.31	22	57.56	29.62
	<b>CLP</b>	WER	22.01	28.62	17.6	62.18	36.13

worsens. When trained and tested on normal speech, the models achieve low WER values, but performance degrades significantly when tested on CLP speech, particularly in the severe category. Augmenting training data with CLP speech from different severity levels improves fairness but does not fully close the performance gap. The fairness analysis shows that Whisper consistently performs better than XLS-R and GMM-HMM, particularly for the NMCPC dataset, whereas GMM-HMM is more suitable for AIISH due to its better performance on child speech.

The following subsections provide a detailed breakdown of ASR performance for each model.

### 6.2.1. GMM-HMM

Initially, the ASR is done using GMM-HMM and using the MFCC features trained with NMCPC and AIISH datasets in Kaldi toolkit. Then in the similar way XLS-R based transformer model is applied for the same task. The obtained performance of the initial crisis cross-experiments is tabulated in Table 5 and Table 6.

The performance in terms of WER and PER shows the least value when training and testing happening using the normal category. When trained using normal and testing using normal and CLP, the WER of Normal, CLP, the pooled WER and FS are 2.39%, 22%, 22.50%, and 84.17%, respectively for AIISH. Similarly, for NMCPC WER was obtained as 19.03%, 46.44%, 64.84%, 87.19%. This shows that the WER increases due to degradation in speech signal introduced by the severity of CLP. Further, by training, the ASR with CLP, and testing with normal, mild, moderate, and severe provides the WER of 2.66%, 11%, 16.67%, and 81.67%, respectively for AIISH and 19.03%, 40.68%, 39.45%, 81.40%, for NMCPC.

For AIISH data, the performance obtained in terms of WER and PER for GMM-HMM in CLP is 42.89% and 64.03%, respectively for AIISH and 66.15% and 83.52% for NMCPC. The experimental results are tabulated in Table 7 and Table 8. From the table, it can be seen that FS of CLP is better than normal. It indicates that the system trained with only normal data is unfair for CLP compared to the system trained with CLP data which is fair for CLP test set. The result of the system trained with Normal+Mild+Moderate+Severe gives FS score as -25.92 indicating is system is better than the other methods.

Table 7: GMM-HMM, XLS-R and Whisper WER performance comparison of Augmented data and Fairness ratio of **AIISH** dataset.

Train↓	Test→											
	GMM-HMM				XLS-R				Whisper			
	Normal	CLP	Pooled WER	FS	Normal	CLP	Pooled WER	FS	Normal	CLP	Pooled WER	FS
<b>Normal</b>	2.39	42.89	22.64	-31.57	7.53	56.91	32.22	-40.80	2.21	53.19	27.70	-39.34
<b>CLP</b>	2.66	11.00	16.67	-26.67	24.73	52.47	38.60	-33.17	7	44.77	25.88	-31.83
<b>Mild+Normal</b>	2.48	39.66	21.07	-29.13	20.21	64.82	42.51	-43.56	9.57	58.66	34.11	-37.71
<b>Mild+Moderate+Normal</b>	2.48	36.75	19.61	-26.94	24.11	62.33	43.22	-40.72	6.91	49.80	28.35	-35.62
<b>Mild+Moderate+Severe+Normal</b>	2.22	35.30	18.76	-25.92	10.37	51.30	30.83	-35.88	8.59	42.19	25.88	-29.49

Table 8: GMM-HMM, XLS-R and Whisper WER performance comparison of Augmented data and Fairness ratio of **NMCPC** dataset

Train↓	Test→											
	GMM-HMM				XLS-R				Whisper			
	Normal	CLP	Pooled WER	FS	Normal	CLP	Pooled WER	FS	Normal	CLP	Pooled WER	FS
<b>Normal</b>	19.03	66.15	42.59	-44.86	32.93	75.23	54.08	-48.19	7.33	49.57	28.45	-35.34
<b>CLP</b>	19.03	53.84	36.43	-35.62	47.73	56.54	52.13	-30.47	23.85	25.05	24.45	-12.82
<b>Mild+Normal</b>	19.03	41.98	30.50	-26.72	24.00	49.07	39.03	-29.55	10.09	36.55	23.32	-24.89
<b>Mild+Moderate+Normal</b>	19.34	39.17	29.25	-24.54	30.51	45.89	38.20	-26.79	9.78	28	18.89	-18.55
<b>Mild+Moderate+Severe+Normal</b>	22.05	39.89	30.97	-24.40	33.83	51.90	42.86	-30.47	14.67	34.99	24.83	-22.58

### 6.2.2. XLS-R and Whisper

Similarly, for XLS-R and Whisper, after training with Normal train set, the test performances on CLP WER is 56.91% and 53.19%, respectively. As evidenced by the study, the performance of transformer provides an improvement over the best performance achieved on GoogleAPI which is 74.27% for NMCPC. As AIISH dataset is child speech and Kannada language so the foundation models were not performing well for it and traditional models are giving better results whereas for NMCPC being in English language, foundation models are giving better results. This justifies the claim that, transformers are best suitable for predicting the speech to text from low resource pathological speech data, the performance of the fairness score also justifies the same.

Interestingly, though the ASR is trained using CLP, testing with normal does not have much performance degradation. For AIISH and NMCPC datasets respectively, as seen from Table 7. The WER performance is (2.39% and 19.03%) with the normal training scenario in GMM-HMM. It is 7.53% to 32.93% in XLS-R and 2.21% to 7.33% in Whisper for AIISH and NMCPC respectively. This shows that, though the CLP has utterances shifted formant locations, the ASR can model them and provide competitive performance while testing with normal. This may be due to the somewhat intact nature of the formant contour shape in normal and CLP speech. Further, it is also observed that, even with using CLP in training the performance in severe is very high in both GMM-HMM and the foundation models as well. This shows that severe utterances have some random distortion and are difficult to learn through ASR. To avoid the confusion that the claim of capturing formant contour shape and observed performance is due to the impact of the language model, the performance is evaluated without using LM. The performances in terms of PER also show a similar trend to WER. Hence justifies the claim. The nature of ASR to capture the formant contour shape, and the stability in performance in the normal test set when trained with CLP, motivates to augment CLP utterances to the training set. This may help in improving the fairness of the system. Motivated by the facts, an augmentation study is performed and obtained performance along with the degree of unfairness or the fairness score tabulated in Table 9.

From the tables, it is also observed that when the system is trained using normal, the fairness score improves gradually after augmenting with severities from mild to mild+moderate and mild+moderate+severe (i.e. increased from -29.13 in normal+mild to -25.92 in normal+mild+moderate+severe) in GMM-HMM, from -43.56 to -35.88 in the case of XLS-R and from -37.71 to -29.49 in the case of Whisper respectively for AIISH data as seen from Table 7. Thus the augmentation of the mild, moderate, and mild, moderate, and severe with the normal training set is able the further improve the fairness of the system. The performance is also evaluated in the NMCPC dataset. It is observed from the table that the observed performance shows a similar trend to AIISH. Further, the performance

Table 9: alpha and beta for GMM-HMM **AIISH** and **NMCPC** dataset.

Train↓	Test→									
	GMM-HMM (AIISH)					Whisper (NMCPC)				
	CLP	Pooled WER	$\alpha = 0.5, \beta = 0.5$	$\alpha = 0.1, \beta = 0.9$	$\alpha = 0.9, \beta = 0.1$	CLP	Pooled WER	$\alpha = 0.5, \beta = 0.5$	$\alpha = 0.1, \beta = 0.9$	$\alpha = 0.9, \beta = 0.1$
<b>Normal</b>	42.89	22.64	-31.57	-38.71	-24.42	49.57	28.45	-35.34	-40.86	-29.83
<b>CLP</b>	11.00	16.67	-26.67	-32.36	-20.97	25.05	24.45	-12.82	-3.52	-22.12
<b>Mild+Normal</b>	39.66	21.07	-29.13	-35.57	-22.68	36.55	23.32	-24.89	-26.14	-23.63
<b>Mild+Moderate+Normal</b>	36.75	19.61	-26.94	-32.80	-21.08	28	18.89	<b>-18.55</b>	<b>-18.28</b>	<b>-18.82</b>
<b>Mild+Moderate+Severe+Normal</b>	35.30	18.76	<b>-25.92</b>	<b>-31.65</b>	<b>-20.19</b>	34.99	24.83	-22.58	-20.77	-24.38

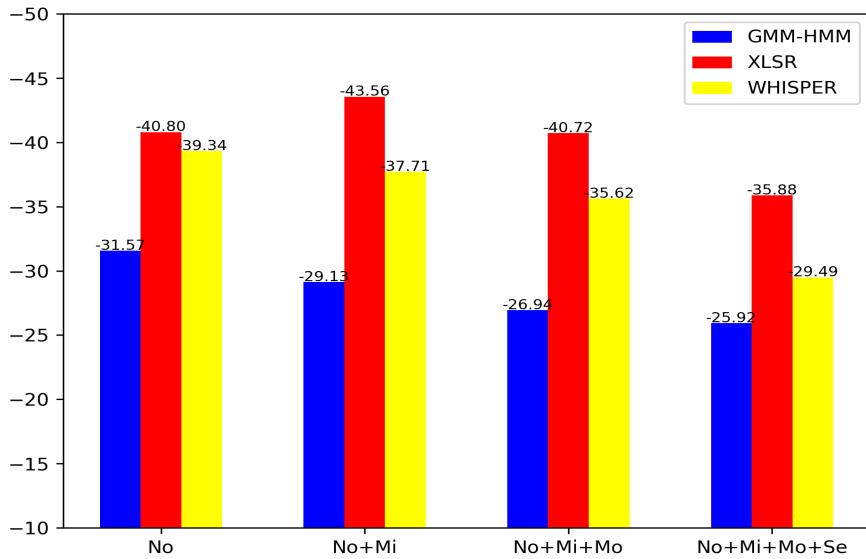


Figure 4: Bar graph showing FS for AIISH dataset. No represents Normal, Mi represents Mild, Mo represents Moderate and Se represents Severe.

of ASR in over CLP test for NMCPC as seen from Table 8, improved by observing the obtained improvement in FS from  $-26.72$  to  $-24.40$ , in the case of GMM-HMM from  $-29.55$  to  $-26.79$  in the case of XLS-R and from  $-24.89$  to  $-18.55$  in the case of Whisper respectively. In some cases, mild+moderate performed better and in other cases, mild+moderate+severe provided better FS. Figure 4 and 5 present bar graphs illustrating the FS values for different models (GMM-HMM, XLS-R, and WHISPER) on the AIISH and NMCPC datasets, respectively. The x-axis represents different subsets of the dataset and refers to speech impairment. The y-axis indicates the FS values, with higher values or in other words, the values closer to zero, generally reflecting better model performance. Across both datasets, the XLS-R model (red) consistently shows the highest (least favorable) FS values, whereas GMM-HMM (blue) and WHISPER (yellow) exhibit higher values, indicating better performance. WHISPER tends to achieve the lowest FS values in several cases, particularly in the NMCPC dataset, suggesting that it may handle pathological speech more effectively than the other models. The trend in both figures indicates that as the severity of speech impairment increases, the difference in FS values between models becomes more pronounced. In the Table 9, the first case shows higher weight on average error rate ( $\alpha = 0.9, \beta = 0.1$ ), the second case has a higher weight on error disparity ( $\alpha = 0.1, \beta = 0.9$ ) and finally, the balanced case ( $\alpha = 0.5, \beta = 0.5$ ) is useful when both accuracy and fairness are equally important, ensuring a trade-off between minimizing overall errors and maintaining similar performance across groups. The bold values show which augmentation is favorable for both cases. For NMCPC, Mild+Moderate+Normal system provides better WER as compared to other systems. But for the AIISH dataset, Mild+Moderate+Severe+Normal system gives the best WER which performed better than the CLP system as well. Depending on the user's specifications, different operating systems can be designed accordingly as per the best performance observation of FS.

This shows the augmentation of CLP data helps in improving the normal speech ASR performance as well along

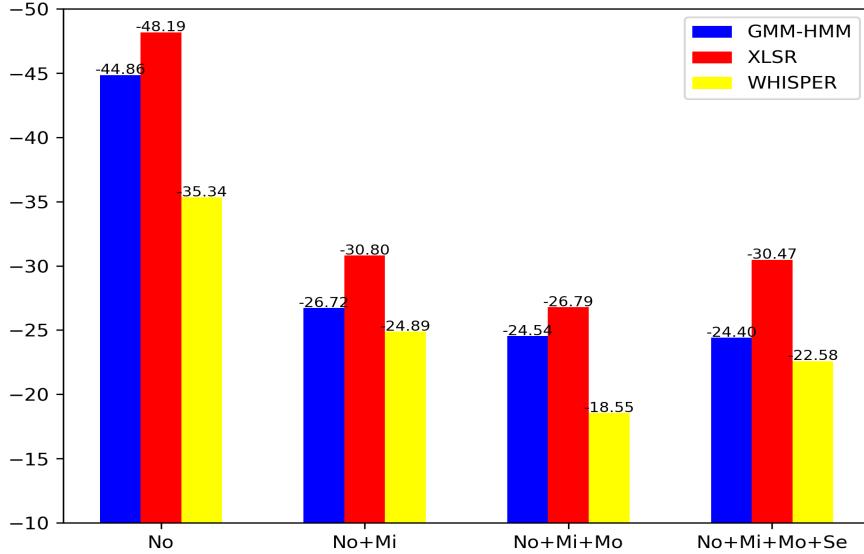


Figure 5: Bar graph showing FS for NMCPC dataset. No represents Normal, Mi represents Mild, Mo represents Moderate and Se represents Severe.

with the CLP speech. Finally, the study concludes the fairness and the performance of the ASR system can be improved by augmenting the utterances from different severity levels during ASR training. The fairness score across all the results depict that there is still a wide gap in speech recognition of CLP speech data. The criss cross results have bridged the gap a little and a detailed results shown in the experiments show that augmentation would help further in the improvement of fairness score.

## 7. Discussion

Initially, to understand the effect of training testing, the normal and CLP speech is used to train and evaluate the GMM-HMM system in a crisscross manner. The motivation is to observe the ASR performance in different combinations, specifically the performance on evaluating with normal utterances, when the system is trained with CLP utterances. Now, the hypothesis is, that the degree of unfairness (*FS*) might increase by testing the ASR using subjects from mild to moderate and severe. Alternatively, as the distortion in the mild utterances is less, hence the augmentation of the same with normal during training might improve the system's fairness, without much hampering the performance of normal ASR. To address the same, various combinations of augmentation strategies are planned and their fairness is analyzed.

This study uses the GMM-HMM system to train the ASR using the AIISH dataset. Keeping the fairness analysis into consideration, the system is initially trained with normal speech and then evaluated in normal and CLP speech, and the fairness is analyzed. After that, as per the hypothesis, the separate GMM-HMM system is trained by augmenting normal with mild, then normal with moderate, and normal with severe, respectively. Further, the GMM-HMM system is also trained by augmenting the normal with mild and moderate, and normal with mild, moderate and severe utterances, respectively. After that, all the trained systems are evaluated using normal, mild, moderate, and severe utterances, and the fairness of the systems is analyzed. Finally, for the generalization of the obtained observations, the same set of experiments is repeated with the NMCPC dataset.

## 8. Conclusion

The work proposes a fairness metric to calculate the fairness of CLP speech. The criss-cross experiment shows that though the training done in CLP by augmenting mild, moderate, and severe cases improves the performance of

inferencing. Training with several sets of CLP severity helps to further draw a conclusion about how the ASR system will perform if augmented with pathological speech. XLS-R and Whisper-based cross-lingual ASR performs better as compared to GMM-HMM based KALDI models with a decrement in WER. This shows that transformer based models outperform for ASR tasks for low resource pathological speech in English language. It has also been observed that the speech recognition rate is highest for mild CLP and least for severe CLP. Therefore, our future attempts will try to develop a better framework, that can provide enhanced speech recognition system whose performance should be independent less affected by the variations of CLP patient's severity. The augmentation study confirms that in the future, the fairness of the ASR system can be improved by augmenting the utterances from different severity levels during ASR training.

## References

- [1] D. J. Zajac, L. Vallino, Evaluation and Management of Cleft Lip and Palate: A Developmental Perspective, 2017. doi:10.1109/ICME.2016.7552917.
- [2] A. Lohmander, M. Olsson, Methodology for perceptual assessment of speech in patients with cleft palate: A critical review of the literature, *The Cleft Palate-Craniofacial Journal* 41 (2004) 64 – 70.
- [3] J. Stengelhofen, Cleft palate: The nature and remediation of communication problems, Churchill Livingstone (1993).
- [4] D. Wang, B. Zhang, Q. Zhang, Y. Wu, Global, regional and national burden of orofacial clefts from 1990 to 2019: an analysis of the global burden of disease study 2019, *Annals of Medicine* 55 (2023). doi:10.1080/07853890.2023.2215540.
- [5] K. M. V. Lierde, S. Claeys, M. D. Bodt, P. V. Cauwenberge, Vocal quality characteristics in children with cleft palate: a multiparameter approach., *Journal of voice : official journal of the Voice Foundation* 18 3 (2004) 354–62.
- [6] D. J. Zajac, C. Plante, A. Lloyd, K. L. Haley, Reliability and validity of a computer-mediated, single-word intelligibility test: Preliminary findings for children with repaired cleft lip and palate, *The Cleft Palate-Craniofacial Journal* 48 (2011) 538 – 549.
- [7] T. L. Whitehill, C. H. F. Chau, Single-word intelligibility in speakers with repaired cleft palate, *Clinical Linguistics & Phonetics* 18 (2004) 341 – 355.
- [8] A. W. Kummer, Cleft Palate and Craniofacial Anomalies: Effects on Speech and Resonance, 2007.
- [9] I. Baumann, D. Wagner, F. Braun, S. P. Bayerl, E. Noth, K. Riedhammer, T. Bocklet, Influence of utterance and speaker characteristics on the classification of children with cleft lip and palate, *INTERSPEECH* 2023 (2022).
- [10] S. Bhattacharjee, H. S. Shekhawat, S. R. M. Prasanna, Classification of cleft lip and palate speech using fine-tuned transformer pretrained models, in: B. J. Choi, D. Singh, U. S. Tiwary, W.-Y. Chung (Eds.), *Intelligent Human Computer Interaction*, Springer Nature Switzerland, Cham, 2024, pp. 55–61.
- [11] S. Kalita, G. K S, P. Mariswamy, S. Prasanna, S. Dandapat, Objective assessment of cleft lip and palate speech intelligibility using articulation and hypernasality measures, *The Journal of the Acoustical Society of America* 146 (2019) 1164–1175. doi:10.1121/1.5121310.
- [12] S. Kalita, S. R. M. Prasanna, S. Dandapat, Importance of glottis landmarks for the assessment of cleft lip and palate speech intelligibility., *The Journal of the Acoustical Society of America* 144 5 (2018) 2656.
- [13] S. Kalita, S. R. M. Prasanna, S. Dandapat, Intelligibility assessment of cleft lip and palate speech using gaussian posteriograms based on joint spectro-temporal features., *The Journal of the Acoustical Society of America* 144 4 (2018) 2413.
- [14] C. M. Vikram, S. Macha, S. Kalita, S. R. M. Prasanna, Acoustic analysis of misarticulated trills in cleft lip and palate children., *The Journal of the Acoustical Society of America* 143 6 (2018) EL474.
- [15] S. Bhattacharjee, R. Sinha, Sensitivity analysis of maskcyclegan based voice conversion for enhancing cleft lip and palate speech recognition, 2022, pp. 1–5. doi:10.1109/SPCOM55316.2022.9840769.
- [16] X. Wang, S. Yang, M. Tang, H. Yin, H. Huang, L. He, Hypernasalinitynet: Deep recurrent neural network for automatic hypernasality detection, *International Journal of Medical Informatics* 129 (2019) 1–12. doi:<https://doi.org/10.1016/j.ijmedinf.2019.05.023>.
- [17] J. H. Ha, H. Lee, S. M. Kwon, H. Joo, G. Lin, D. Y. Kim, S. Kim, J. Y. Hwang, J. H. Chung, H. J. Kong, Deep learning-based diagnostic system for velopharyngeal insufficiency based on videofluoroscopy in patients with repaired cleft palates, *Journal of Craniofacial Surgery* 129 (2023).
- [18] A. Babu, C. Wang, A. Tjandra, K. Lakhota, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. M. Pino, A. Baevski, A. Conneau, M. Auli, Xls-r: Self-supervised cross-lingual speech representation learning at scale, *ArXiv abs/2111.09296* (2021).
- [19] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, I. Sutskever, Robust speech recognition via large-scale weak supervision, in: *Proceedings of the 40th International Conference on Machine Learning, ICML'23, JMLR.org*, 2023.
- [20] A. K. Dubey, S. R. M. Prasanna, S. Dandapat, Zero time windowing based severity analysis of hypernasal speech, 2016 IEEE Region 10 Conference (TENCON) (2016) 970–974.
- [21] A. K. Dubey, S. R. M. Prasanna, S. Dandapat, Hypernasality detection using zero time windowing, 2018 International Conference on Signal Processing and Communications (SPCOM) (2018) 105–109.
- [22] K. Nikitha, S. Kalita, M. VikramC., M. Pushpavathi, S. R. M. Prasanna, Hypernasality severity analysis in cleft lip and palate speech using vowel space area, in: *Interspeech*, 2017.
- [23] M. VikramC., A. Tripathi, S. Kalita, S. R. M. Prasanna, Estimation of hypernasality scores from cleft lip and palate speech, in: *Interspeech*, 2018.
- [24] V. C. Mathad, N. J. Scherer, K. Chapman, J. M. Liss, V. Berisha, An attention model for hypernasality prediction in children with cleft palate, *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2021) 7248–7252.
- [25] A. K. Dubey, S. R. M. Prasanna, S. Dandapat, Pitch-adaptive front-end feature for hypernasality detection, in: *Interspeech*, 2018.
- [26] A. K. Dubey, S. R. M. Prasanna, S. Dandapat, Hypernasality severity detection using constant q cepstral coefficients, in: *Interspeech*, 2019.

- [27] A. K. Dubey, S. R. M. Prasanna, S. Dandapat, Sinusoidal model-based hypernasality detection in cleft palate speech using cvcv sequence, *Speech Commun.* 124 (2020) 1–12.
- [28] A. K. Dubey, S. R. M. Prasanna, S. Dandapat, Detection and assessment of hypernasality in repaired cleft palate speech using vocal tract and residual features., *The Journal of the Acoustical Society of America* 146 6 (2019) 4211.
- [29] S. Kalita, S. R. M. Prasanna, S. Dandapat, Self-similarity matrix based intelligibility assessment of cleft lip and palate speech, in: *Interspeech*, 2018.
- [30] S. Kalita, K. S. Girish, P. M., S. R. M. Prasanna, S. Dandapat, Objective assessment of cleft lip and palate speech intelligibility using articulation and hypernasality measures., *The Journal of the Acoustical Society of America* 146 2 (2019) 1164.
- [31] S. Kalita, P. N. Sudro, S. R. M. Prasanna, S. Dandapat, Nasal air emission in sibilant fricatives of cleft lip and palate speech, in: *Interspeech*, 2019.
- [32] C. M. Vikram, N. Adiga, S. R. M. Prasanna, Detection of nasalized voiced stops in cleft palate speech using epoch-synchronous features, *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 27 (2019) 1189–1200.
- [33] P. N. Sudro, S. Kalita, S. R. M. Prasanna, Processing transition regions of glottal stop substituted /s/ for intelligibility enhancement of cleft palate speech, in: *Interspeech*, 2018.
- [34] P. N. Sudro, S. M. Prasanna, Enhancement of cleft palate speech using temporal and spectral processing, *Speech Communication* 123 (2020) 70–82.
- [35] P. N. Sudro, S. M. Prasanna, Modification of misarticulated fricative/s/in cleft lip and palate speech, *Biomedical Signal Processing and Control* 67 (2021) 102088.
- [36] P. N. Sudro, C. M. Vikram, S. R. M. Prasanna, Event-based transformation of misarticulated stops in cleft lip and palate speech, *Circuits, Systems, and Signal Processing* 40 (2021) 4064 – 4088.
- [37] P. N. Sudro, R. K. Das, R. Sinha, S. R. M. Prasanna, Enhancing the intelligibility of cleft lip and palate speech using cycle-consistent adversarial networks, *2021 IEEE Spoken Language Technology Workshop (SLT)* (2021) 720–727.
- [38] P. N. Sudro, R. K. Das, R. Sinha, S. R. Mahadeva Prasanna, Significance of data augmentation for improving cleft lip and palate speech recognition, in: *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2021, pp. 484–490.
- [39] I. Baumann, D. Wagner, F. Braun, S. P. Bayerl, E. Nöth, K. Riedhammer, T. Bocklet, Influence of utterance and speaker characteristics on the classification of children with cleft lip and palate, in: *Interspeech*, 2022.
- [40] K. Song, T. Wan, B. Wang, H. Jiang, L. K. Qiu, J. Xu, L. ping Jiang, Q. Lou, Y. Yang, D. Li, X. Wang, L. Qiu, Improving hypernasality estimation with automatic speech recognition in cleft palate speech, in: *Interspeech*, 2022.
- [41] M. VikramC., S. R. M. Prasanna, A. K. Abraham, M. Pushpavathi, S. GirishK., Detection of glottal activity errors in production of stop consonants in children with cleft lip and palate, in: *Interspeech*, 2018.
- [42] V. C. Mathad, S. R. M. Prasanna, Vowel onset point based screening of misarticulated stops in cleft lip and palate speech, *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 28 (2020) 450–460.
- [43] V. C. Mathad, N. J. Scherer, K. Chapman, J. M. Liss, V. Berisha, A deep learning algorithm for objective assessment of hypernasality in children with cleft palate, *IEEE Transactions on Biomedical Engineering* 68 (2020) 2986–2996.
- [44] K. Fujiwara, R. Takashima, C. Sugiyama, N. Tanaka, K. Nohara, K. Nozaki, T. Takiguchi, Data augmentation based on frequency warping for recognition of cleft palate speech, in: *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2021, pp. 471–476.
- [45] M. H. Javid, K. Gurugubelli, A. K. Vuppala, Single frequency filter bank based long-term average spectra for hypernasality detection and assessment in cleft lip and palate speech, in: *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6754–6758. doi:10.1109/ICASSP40776.2020.9054684.
- [46] T. Bocklet, A. K. Maier, K. Riedhammer, U. Eysholdt, E. Nöth, Erlangen-clp: A large annotated corpus of speech from children with cleft lip and palate, in: *International Conference on Language Resources and Evaluation*, 2014.
- [47] A. Eshky, M. S. Ribeiro, J. Cleland, K. Richmond, Z. Roxburgh, J. M. Scobbie, A. A. Wrench, Ultrasuite: A repository of ultrasound and acoustic data from child speech therapy sessions, *ArXiv abs/1907.00835* (2018).
- [48] K. Nikitha, S. Kalita, C. Vikram, M. Pushpavathi, S. M. Prasanna, Hypernasality severity analysis in cleft lip and palate speech using vowel space area., in: *Interspeech*, 2017, pp. 1829–1833.
- [49] M. Russell, S. D'Arcy, Challenges for computer recognition of children's speech, in: *Proc. Speech and Language Technology in Education (SLaTE 2007)*, 2007, pp. 108–111. doi:10.21437/SLaTE.2007-26.
- [50] J. J. Howard, E. J. Laird, Y. B. Sirotin, R. E. Rubin, J. L. Tipton, A. R. Vemury, Evaluating proposed fairness models for face recognition algorithms, in: *ICPR Workshops*, 2022.
- [51] A. Liang, J. Lu, X. Mu, Algorithm design: Fairness and accuracy\*, 2022.
- [52] H. Sakoe, Dynamic programming algorithm optimization for spoken word recognition, *IEEE Transactions on Acoustics, Speech, and Signal Processing* 26 (1978) 159–165.
- [53] D. A. Reynolds, T. F. Quatieri, R. B. Dunn, Speaker verification using adapted gaussian mixture models, *Digital Signal Processing* 10 (2000) 19–41.
- [54] M. J. F. Gales, S. J. Young, The application of hidden markov models in speech recognition, *Found. Trends Signal Process.* 1 (2007) 195–304.
- [55] L. R. Rabiner, A tutorial on hidden markov models and selected applications in speech recognition, *Proc. IEEE* 77 (1989) 257–286.
- [56] L. R. Rabiner, A tutorial on hidden markov models and selected applications in speech recognition, *Proc. IEEE* 77 (1989) 257–286.
- [57] G. Salton, M. J. McGill, *Introduction to Modern Information Retrieval*, McGraw-Hill, 1983.
- [58] A. Conneau, A. Baevski, R. Collobert, A. rahman Mohamed, M. Auli, Unsupervised cross-lingual representation learning for speech recognition, *ArXiv abs/2006.13979* (2020).
- [59] A. Baevski, H. Zhou, A. rahman Mohamed, M. Auli, wav2vec 2.0: A framework for self-supervised learning of speech representations, *ArXiv abs/2006.11477* (2020).