Digitalization of speech therapy with AI-based personalized feedback

Eugenia Rykova

University of Eastern Finland
University of Applied Sciences TH Wildau
Catholic University Eichstätt-Ingolstadt
eugenryk@uef.fi

Abstract

This paper introduces an ongoing PhD thesis carried out in the framework of a research project, in which an application for speech and language therapy support of German speakers with aphasia is developed. Work completed in the selection and implementation of ASR solutions, and creating a semantic analysis pipeline is described, and challenges and future work perspectives are discussed.

Index Terms: automatic speech recognition, aphasic speech, speech and language therapy, digital health.

1. Introduction

Aphasia, literally translated from (Ancient) Greek as „speechlessness“ [1], is an acquired language disorder due to a focal brain injury. It affects some or all language modalities, which makes communication difficult and decreases the quality of life. High intensity and duration of speech and language therapy (SLT) bring certain benefits to communication improvements [2-3]. However, not all people with aphasia (PWA) have access to sufficient SLT (e.g., due to lack of specialists or geographical remoteness). Research shows the efficiency of supplementing in-person therapy with independent usage of digital therapy solutions [4].

Various researchers have explored the possibilities of automatic assessment of speech produced by PWA (see, for example, [5-6]). Naming-oriented semantic exercises have been automated with the help of automatic speech recognition (ASR) for Portuguese [7], English [8-9], and German [10-11]. The latter, however, are not in active use yet. When the answer is only rated as correct/incorrect, with no further analysis of users' errors, the reported acceptance/rejection accuracies on PWA's speech range from 75% [8] to 89.5% [9].

The current project [12] focuses on developing a mobile application for German-speaking PWA that will provide personalized detailed feedback in naming and other exercises. In the present PhD research, ASR and further text processing solutions are used for multilevel feedback: phonemic/phonetic, semantic, and grammatical. To build the corresponding pipeline, the following questions need to be addressed.

1. Which existing (open-source) ASR solutions are suitable for the task-specific speech of German-speaking PWA?

2. How can selected ASR solutions be improved and/or adapted for the purposes of SLT?

3. How can a combination of selected ASR solutions and existing tools for semantic and grammatical analysis serve for speech production errors analysis?

4. What are patients' and therapists' attitudes to the proposed digital solution?

2. ASR solutions

2.1. Model selection

Evaluation of the ASR systems consisted of several steps. First, the suitability of more than 50 open-source ASR solutions was assessed with the help of several speech recordings from different corpora, including PWA's speech [13-14]. Based on the ranking of error rates, 13 models were selected for further evaluation. In the absence of necessary data from PWA, test material from other corpora with atypical speech (presenting abnormalities similar to PWA's speech) was considered for further evaluation, namely speech of adult cochlear implant (CI) users and normal-hearing speakers as a counterpart [15], and speech produced under alcohol intoxication the same speakers under no intoxication as a counterpart [16]. Additionally, two small datasets with aphasic speech were used. AvEv recordings (39 single words) were obtained from four PWA who had taken part in an avatar evaluation experiment was used [17]. UniSt recordings (79 single words) had been made during Aachen Aphasia Test (AAT) [18] sessions and were obtained on request from Stuttgart University Institute for Natural Language Processing [19].

Finally, four open-source ASR models were selected for the backend of the app [20-23] based on character error rates, the number of empty outputs, and the number of precisely recognized words. Three of these models [20-22] are to a certain extent independent from pronunciation and language models and are suitable for phoneme-level pronunciation analysis, while the fourth model [23] gives only existing orthographic forms as output, which is more suitable for subsequent semantic and grammatical error analysis. All four models are to a greater or lesser extent robust to speaker gender and age. The experiments suggest that for better single-word recognition the audio samples should be not too short and pronounced neither too slowly nor too fast (i.e. intentionally speeded up).

2.2. Post-hoc implementation of non-standard phonetic features

Although the selected ASR models present a possibility for fine-tuning, the project lacks adequate data for model (re)training. Thus, it was decided to research the possibility of applying the knowledge about non-standard phonetic features post hoc to ASR output. The methodology combines generating alternative pronunciations based on non-standard patterns [24] and using alternatives for evaluation [25].

First, the orthographic ASR output form is phonemized using automatic grapheme-to-phoneme conversion (g2p) [26]. Then, the phonetic transcription is subject to modifications based on the non-standard phonetic features. The first set of features considers aphasic speech: syllabification with greater pauses between syllables, which causes recognition of syllables as separate words; and slow and careful speech production, which causes vowel prolongation. The next set comprises relevant dialect features selected from the Thuringia-Upper Saxon dialect group [27-29] due to the geography of the project and the data available for the experiments. The modified transcription is then compared to the target transcription, and the error rate (ER) threshold is applied. If the ER is lower than the threshold, the error is considered phonemic/phonetic, and semantic otherwise (see example in Figure 1).

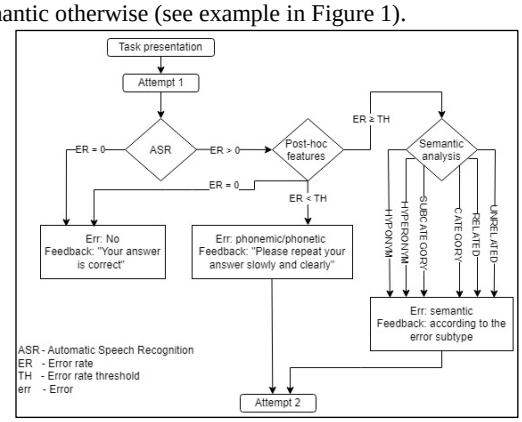


Figure 1: *Error analysis pipeline – attempt 1.*

The proposed method was tested on the 412 single-word recordings made during AAT sessions and obtained on request from the University of Leipzig Medical Center. It has proven to work: after the implementation of each feature set, ASR error rates decrease significantly, general acceptance/rejection accuracy improves, and the accuracy of error attribution also increases.

2.3. Challenges and future work

The selected four models present a possibility to be fine-tuned: to PWA speech or speech of a particular user in a customized version, and to single-word recognition task rather than continuous speech recognition. This requires a certain amount of corresponding data. Obtaining the data from PWA, possible anonymization of such data, and further application for model fine-tuning are seen as the following steps. The complete solution should be tested "live" to take into account the corresponding audio quality and processing time.

3. Semantic and grammatical analysis

3.1. Semantic analysis pipeline

If the answer of the speaker does not pass the ER threshold (i.e. is not recognized as correct or containing phonemic/phonetic errors only), it is subject to further analysis. In particular, it must be compared to the target in terms of their semantic relationship and distance. The current semantic analysis is built upon GermaNet – a semantic network for the German language [30]. It consists of two parts: semi-automatic enrollment of the exercise item into the system and the analysis of a semantic error. The latter includes but should not be limited to the recognition of hyponymy/hypernymy and belonging to the same semantic (sub)category. If the answer is not recognized as an existing word (i.e. contains both semantic and phonemic/phonetic errors), a search for close orthographic matches is performed, and the match that is semantically the closest to the target is subject to the relationship analysis described above.

Current work is concentrated on including further types of semantic relationships in analysis, for example, synonymy, antonymy, and meronymy/holonymy. Close orthographic matches search is extended with close phonemic matches search, using automatic g2p. On the other hand, the search is thought to be limited to the members of the target lexical and semantic categories, while the final assumption is based on both semantic and orthographic/phonemic distance to the target.

3.2. Challenges and future work

The current pipeline, or GermaNet in general, has certain limitations. First, it is mostly suitable for the words of the same lexical category (except for causative relationship, pertainyms, and participles), so that the relationship between "to eat" and "food" would not be recognized. Second, GermaNet takes only lemmas as input, which makes it necessary to implement an additional step with a lemmatizer. Further limitations can arise from a mismatch of the semantic categories in typical SLT tasks or a broader common understanding of language and GermaNet. Exploring other semantic networks (e.g., BabelNet [31]), adding grammatical analysis, elaborating more intuitive semantic categories, and joint implementation with selected ASR solutions will be addressed next.

4. Users' evaluation

The automatic error analysis process includes the following components: ASR, post-hoc phonetic features implementation (if applicable), phonemic/phonetic error analysis (if applicable), semantic and grammatical error analysis (if applicable), and issuing corresponding feedback.

Based on this general pipeline, digital exercises are to be designed and implemented in an app and then tested in SLT practice. A questionnaire is then designed to collect therapists' and patients' opinions on the exercises. Such parameters as, for example, plausibility, clarity, and user-friendliness should be assessed. The answers will be evaluated both quantitatively and qualitatively, and compared between the subsets. Based on the feedback, necessary changes or suggestions can be made to improve the solution (cf. [32]).

5. Limitations

The greatest limitation of the current work is the lack of relevant data. ASR solutions were mostly tested with other atypical speech and to much less extension with aphasic speech. Furthermore, the analyzed data mentioned in this paper are not suitable for ASR model (re)training or adaptation.

On the other hand, few semantic errors are present in the data, and the examples to test the semantic analysis pipeline have to be constructed artificially. The current basis for semantic analysis, GermaNet [30], presents certain limitations on its own, described above in the corresponding section.

The current project is a regional one and therefore is focused on the German language, in particular on the dialects of the Thuringian-Upper Saxon group. However, general principles and pipelines elaborated as the result of the present research can be scaled to other dialects and languages.

6. References

[1] J. Ryalls, "Where does the term "aphasia" come from?" *Brain and Language*, 21, pp. 358-363, 1984.

[2] S. K. Bhogal, R. Teasell, and M. Speechley, "Intensity of aphasia therapy, impact on recovery," *Stroke*, 34, pp. 987-993, 2003.

[3] M. C. Brady, H. Kelly, J. Godwin, and P. Enderby, "Speech and language therapy for aphasia following stroke (Review)," *Cochrane Database of Systematic Reviews*, 2016(6), CD000425, 2016.

[4] M. Braley, J. S. Pierce, S. Saxena, E. D. Oliveira, L. Taraboanta, V. Anantha, . . . S. Kiran, "A virtual, randomized, control trial of a digital therapeutic for speech, language, and cognitive intervention in post-stroke persons with aphasia," *Frontiers in Neurology*, 12, 626780, 2021.

[5] A. Adikari, N. Hernandez, D. Alahakoon, M. L. Rose, and J. E. Pierce, "From concept to practice: a scoping review of the application of AI to aphasia diagnosis and management," *Disability and Rehabilitation*, pp. 1-10, 2023.

[6] G. Pottinger and Á. Kearns, "Big data and artificial intelligence in post-stroke aphasia: A mapping review," *Advances in Communication and Swallowing*, vol. Pre-press, no. Pre-press, pp. 1-15, 2024.

[7] A. Pompili, A. Abad, I. Trancoso, J. Fonseca, I. P. Martins, G. Leal, and L. Farrajota, "An on-line system for remote treatment of aphasia," *Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies*, pp. 1-10, 2011.

[8] K.J. Ballard, N. M. Etter, S. Shen, P. Monroe, and C. Tien Tan C., "Feasibility of automatic speech recognition for providing feedback during tablet-based treatment for apraxia of speech plus aphasia," *American Journal of Speech-Language Pathology*, 28, pp. 818–834, 2019.

[9] D. S. Barbera, M. Huckvale, V. Fleming, E. Upton, H. Coley-Fisher, C. Doogan, I. Shaw, W. Latham, A. P. Leff, and J. Crinion, "NUVA: A Naming Utterance Verifier for Aphasia Treatment," *Computer Speech & Language*, 69, 101221, 2021.

[10] J. Heide, J. Netzebandt, S. Ahrens, J. Brüsch, T. Saalfrank, and D. Schmitz-Antonischki, "Improving lexical retrieval with LingoTalk: an app-based, self-administered treatment for clients with aphasia," *Frontiers in Communication*, 8:1210193, 2023.

[11] Y. Lin, P. Klumpp, J. Pfab, A. Abdelioua, D. Gebray, and M. Späth, "Eine automatische Sprachbewertung für die neolexon Aphasie-App mithilfe Künstlicher Intelligenz [Automatic language assessment with artificial intelligence. for the neolexon aphasia app]," *Poster session presentation at Sprachtherapie aktuell: Forschung - Wissen – Transfer 9(1): XXXIV. Workshop Klinische Linguistik e2022-11*, April 2022.

[12] TDG - TRANSLATIONSREGION FÜR DIGITALE GESUNDHEITSVERSORGUNG [Translational Region for Digital Healthcare], "*AphaDIGITAL: Entwicklung einer digitalen, dezentralen sprachtherapeutischen Versorgung [Development of digital, decentralized speech therapy solutions"*. Accessed: January 25, 2024. Available: https://inno-tdg.de/projekte/aphadigital/

[13] B. MacWhinney, D. Fromm, M. Forbes, and A. Holland, "AphasiaBank: Methods for Studying Discourse," *Aphasiology*, *25*(11), pp. 1286–1307, 2011.

[14] Rhein-Zeitung, Germany. *Am Anfang war das Wort: Zu Besuch bei einem Aphasiker [In the beginning was the word: Visiting a person with aphasia]*. (Oct. 26, 2017). Accessed: May 16, 2022. [Online Video]. Available: https://www.youtube.com/watch?v=Z1ZglYMSx1Y.

[15] V. Neumeyer, "Phonetische Untersuchungender Artikulation von CI-Trägern [Phonetic studies of CI users' articulation] [Master's thesis]," Master Thesis, Ludwig-Maximilians-Universität, München, Germany, 2011.

[16] F. Schiel, C. Heinrich, S. Barfüsser, and T. Gilg, "ALC — Alcohol Language Corpus," *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, pp. 1641-1645, 2008.

[17] E. Zeuner, J. Pietschmann, S. Voigt-Zimmermann, E. Rykova, and M. Walther, "aphaDIGITAL - Avatar-gestützte digitale Aphasietherapie: Evaluation [aphaDIGITAL: Avatar-supported digital aphasia therapy - evaluation study]," presented as poster at DGSS Annual Conference 2022, Stimme und Geschlecht im Wandel' – Implikationen für Theorie und Praxis in der Sprechwissenschaft und Phonetik ["Voice and Gender in Transition" – Implications for Theory and Practice in Speech Science and Phonetics], Jena, Germany, September 23-25, 2022. Available at https://www.researchgate.net/publication/371510421_aphaDIGITAL_-Avatar-gestutzte_digitale_Aphasietherapie_Evaluation.

[18] W. Huber, *Aachener aphasie test (AAT) [Aachen Aphasia Test]*. Verlag für Psychologie Hogrefe, Göttingen, Zürich, 1993.

[19] Universität Stuttgart. „Sprache und Gehirn: Ein neurolinguistisches Tutorial [Language and brain: a neurolinguistics tutorial]". Accessed: June 17, 2023. [Online]. Available: https://www2.ims.uni-stuttgart.de/sgtutorial/index.html.

[20] M. Fleck, "*Wav2vec2-large-xls-r-300m-german-with-lm"*. Accessed: September 12, 2022. [Online]. Available: https://huggingface.co/mfleck/wav2vec2-large-xls-r-300m-german-with-lm.

[21] J. Grosman, "*Fine-tuned XLSR-53 large model for speech recognition in German"*. Accessed: September 12, 2022. [Online]. Available: https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-german.

[22] O. Guhr, "*wav2vec2-large-xlsr-53-german-cv9"*. Accessed: September 12, 2022. [Online]. Available: https://huggingface.co/oliverguhr/wav2vec2-large-xlsr-53-german-cv9.

[23] NVIDIA, "*NVIDIA Conformer-Transducer Large (de)"*. Accessed: September 12, 2022. [Online]. Available: https://huggingface.co/nvidia/stt_de_conformer_transducer_large.

[24] A. Masmoudi, M. E. Khmekhem, Y. Est`eve, L. H. Belguith, and N. Habash, "A corpus and phonetic dictionary for Tunisian Arabic speech recognition," in *LREC*, 2014, pp. 306-310.

[25] A. Ali, P. Nakov, P. Bell, and S. Renals, "WERd: Using social text spelling variants for evaluating dialectal speech recognition," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, IEEE, 2017, pp. 141–148.

[26] M. Bernard and H. Titeux, "Phonemizer: Text to phones transcription for multiple languages in Python," *Journal of Open Source Software*, vol. 6, no. 68, p. 3958, 2021.

[27] U. Wallraff, "Ausgewählte phonetische Analysen zur Umgangssprache der Stadt Halle an der Saale [Selected phonetic analyses of the colloquial language of the city of Halle an der Saale]," Doctoral Dissertation, Martin-Luther-Universität Halle-Wittenberg, Halle, Germany, 2007.

[28] M. J. Rocholl, *Ostmitteldeutsch – eine moderne Regionalsprache? Eine Untersuchung zu Konstanz und Wandel im thüringisch-obersächsischen Sprachraum [East-Central German – a modern regional language? An investigation into constancy and change in the Thuringian-Upper Saxon language area]*. Hildesheim, Zürich, New York: OLMS, 2015.

[29] B. Siebenhaar, private communication, Jan. 2024.

[30] B. Hamp, and H. Feldweg, "GermaNet - a Lexical-Semantic Net for German," *Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, 1997.

[31] R. Navigli, and S. P. Ponzetto, "BabelNet: Building a Very Large Multilingual Semantic Network," in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 2010, pp. 216–225.

[32] A. Pompili, A. Abad, I. Trancoso, J. Fonseca, and I. P. Martins, "Evaluation and Extensions and of an Automatic Speech Therapy Platform," in *Computational Processing of the Portuguese Language. PROPOR 2020. Lecture Notes in Computer Science*, P. Quaresma, R. Vieira, S. Aluísio, H. Moniz, F. Batista, and T.-Gonçalves, Eds., 12037, Springer Cham, 2020, pp. 43–52.

Algorithmic methods towards increased fairness in automatic speech processing

Felix Herron[1,2]

[1]Paris Dauphine University - PSL, France
[2]University of Grenoble Alps, France
felix.herron@dauphine.eu

1. Motivation

As automatic speech processing systems have grown more ubiquitous and relied upon, there has been significant complementary study examining their biases. This study tends to come through the lens of *statistical parity*, i.e. whether all subpopulations are treated equally [1]. As is noted in [2]: "no dialect is inherently more or less intelligible"; in other words, there is no natural hierarchy in merit of access to speech processing systems. Hence, if our speech processing systems treat some speaker groups better than others, then they are suboptimal.

Most study of fairness/bias in speech processing has come through the lens of automatic speech recognition (ASR), which will likewise be the focus of this paper. ASR lends itself intuitively to fairness analysis, given the relatively unambiguous nature of the ASR task, transcription. Indeed, by measuring the difference in transcription error rate between various speaker groups, we are measuring the most significant manifestation of bias in ASR systems. Thus, reducing performance difference in ASR is a societally motivated goal, and one whose improvement can be relatively objectively measured.

2. Related work

There has been significant study in measuring fairness and bias in ASR models[1]. All studies that on bias in ASR that I have reviewed measure fairness through the lens of statistical parity in model performance as measured by word error rate (WER). Given two speaker groups, *A* and *B*, studies investigate whether ASR models have statistically significantly lower average WER for *A* than for *B*. The choice of speaker groups tends to follow relatively objectively measurable demographic attributes, such as age, gender, native language, or dialect. The definition of these speaker groups is thus *semantically motivated*, in that it is in the interest of societal equality for them to be treated equally by ASR systems. This framework of statistical parity is useful for its simplicity; however, it leaves for room for improvement, as I will discuss further (in Section 4).

2.1. Bias identification in ASR models

Unfortunately, there are many *semantically motivated* speaker group pairs which are treated unequally by ASR. I will highlight several here; however, note that this list is far from exhaustive.

- Gender bias in ASR has been studied extensively in various languages [3, 2, 4, 5], though the results are varied - some studies find men favored, others find women favored, and still others find no statistically significant discrepancy. One consistent observation is that women are often under-represented in corpora, though the performance impact of this relative paucity is likewise varied and might depend on the corpus [6, 7, 8].
- ASR models have been shown to be biased against non-native speakers of several languages, with multiple studies demonstrating this for English [9, 10], Mandarin [8], and Dutch [11, 12, 8], on the order of 100% worse [11]. ASR performance does improve with improving language ability [8].
- Different dialects are not all transcribed equally well by ASR systems, a result replicated over multiple studies in several languages, such as English [13, 14, 5, 15, 16, 17], Portuguese [18], and Dutch [12]. Despite each dialect being equivalently valid, some are much better recognized by ASR systems, likely due to greater representation in training corpora.

2.2. Studies in bias reduction

There are two main strategies for creating fairer ASR models. The first strategy is to train on datasets with better calibrated speaker group balance. Models that see more data from a specific speaker group generally perform better on it [19, 8] (though not always [7]); thus, in order to improve the performance of a speaker group, increasing its abundance in the training corpus is a simple way to achieve this. In the extreme case, models trained exclusively on single-speaker-group corpora (such as only women, or only speakers of AAVE dialect) can perform better on that target speaker group than models that have seen more varied data [16], though not always [7]. Meanwhile, many ASR corpora (such as CommonVoice [20] and SWITCHBOARD [21]) are unbalanced in some fundamental characteristics, such as age, gender, or accent of speaker; creating fairer mainstream training corpora would likely result in fairer models. Furthermore, intentional dataset curation and data quality are known to be of undervalued importance in machine learning in general, thus the strategy of smarter data curation ought not be overlooked [22].

The second strategy for creating fairer ASR models is to implement training strategies that produce fairer models for a fixed dataset. There are several common approaches to this end:

1. Data augmentation techniques such as **Voice Conversion** (VC) [11] involve synthesizing voices from under-represented speaker groups to supplement unbalanced corpora. These techniques are motivated by two observations: first, greater training corpus representation tends to lead to better treatment by ASR; and second: collecting highly representative corpora is a tedious and expensive process.
2. Another approach to making fairer models is to intentionally

[1]For the purposes of this article, as a reflection of terminology used in the literature, I consider those terms to be exact complements - i.e., a model that exhibits *bias against* one speaker group is not behaving *fairly towards* that speaker group, and vice versa. See Section 4 for further discussion of this.

blind them to differences between speaker groups. **Domain adversarial training** (DAT) trains the speech model to fool a discriminator that attempts to identify the speaker group associated with each utterance [23]. This too has shown promising results for a small number of speaker groups.

3. There has been work in **Domain Enhancing Training** (DET), an application of multi-task learning (MTL), to make models more aware of the speaker group of the utterance they are attempting to transcribe. For example, forcing the model to learn a transcription objective *and* dialect classifier in parallel has proven fruitful [24, 25, 26]. DET is essentially the opposite of DAT, as rather than teaching a model to be domain-agnostic, we are teaching the model to "lean into" domain idiosyncrasies.
4. Finally, there are models which integrate separately calculated **speaker group embeddings** into the ASR pipeline [27, 28]. Speaker group embeddings are concatenated (or summed), either to input features prior to feature extraction, or the output of feature extraction and prior to the downstream layers (in a hybrid setting). Speaker group embeddings can help the model contextualize speech based on the speaker identity.

There has been limited study comparing these bias reduction methods amongst each other. Some studies have shown DAT to be more effective than DET, and VC more effective than DAT, though further study is necessary. However, many of these methods can be used in parallel, where they tend to complement each other [29, 23, 11, 30].

While each of these methods has shown promise, they all suffer from the weakness of needing to explicitly define underrepresented speaker groups. This requires more overhead in (future) corpus creation, and potentially discriminates against some speaker groups which are not explicitly targeted.

3. Research objectives

3.1. Non-enumerative fairness

The first main research objective of my thesis research is to develop unsupervised **speaker group embedding** discovery methods, as an improvement upon method (4). This is motivated by the goal of zero-shot speaker group adaptation, which for existing enumerative methods is not well achieved. [28, 27]. Furthermore, speaker groups which are not learned using or defined by explicit labels reduces the need for metadata-labeled training data. Unsupervised speaker group embeddings will also be able to target multiple speaker group dimensions simultaneously, such as age and accent, for example.

There has been some scholarship aimed at non-enumerative fairness enforcement, such as [31], which performs automatic cohort discovery based on predicted ASR failure. The method I am currently developing aims to use clustering methods, such as k-means or Latent Dirichlet Allocation (LDA, as inspired by [32]), to define and map each speaker to a continuous speaker group embedding space. Preliminary results of this work show that topic distributions generated using LDA on discretized audio segments contain information corresponding to some speaker attributes, such as gender, age, and accent. Table 1 shows macro F1 scores of linear regressions predicting three different speaker group features based on unsupervised LDA-based speaker embeddings. The speaker embeddings are trained on a sample of English CommonVoice data and for increasing number of topic k, and tested on unseen utterances from the same corpus. Note that increasing k increases utility in predicting the three features. Furthermore, the fact that we do not achieve perfect classification accuracy is not a problem; such analysis simply serves to determine whether any relevant information is contained within these unsupervised embeddings.

Table 1: *LDA embeddings contain speaker group information*

	$k=2$	3	10	20	35	50	100
age	0.28	0.37	0.39	0.41	0.41	0.41	0.41
gender	0.65	0.84	0.88	0.91	0.90	0.90	0.91
accent	0.05	0.12	0.13	0.15	0.18	0.20	0.20

3.2. Interpretability/Explainability of deep speech processing models

The better we understand the inner workings of our models, the better we can steer them towards fair behavior. I will build on works like [33, 34], which identify which layers are primarily responsible for modelling acoustic or linguistic features, as well as works like [27] which attempt to distinguish between different accents in various embedding spaces. Furthermore, I will apply work from causal mediation analysis to speech networks, to determine which layers or neurons in different networks are responsible for which low-level aspect of speech processing, as has been done for text-based networks [35].

There is ample opportunity for scholarship in better understanding the mechanisms behind which fairness enforcement algorithms actually work. Furthermore, with a more nuanced understanding of the functioning of each layer (or neuron), I will be able to more precisely consider new learning objectives or architectures to promote fairness.

4. Challenges

4.1. Insufficient metadata

In order to test for statistical parity in ASR performance, one requires datasets that are labeled with precise speaker group metadata. While some datasets contain useful metadata [36, 20], these are the exception, rather than the rule; furthermore, existing metadata is often insufficiently precise or incomplete. In general, speaker group metadata cannot (and/or may not, for privacy reasons [37]) be reverse-engineered. Furthermore, collecting a dataset (with speaker group metadata) is a long and challenging process, further complicated by the fact that precisely which metadata I deem "useful" might change over time.

4.2. Simplicity of performance difference measurements

Simple statistics like statistical parity are straightforward to implement, easy to understand, and provide a useful rough picture of which speaker groups are treated better than others. They are therefore widely used in the ASR bias literature. However, they are not consummate measures of a model's treatment of specific speaker groups. Not all transcription errors are created equal, and to have a deeper understanding of model fairness, it is important to study these in more intricate detail. There has been limited work to this end; for example, authors in [15] examine how ASR systems handle grammatical differences in the AAVE dialect of English, which informs the analysis of transcription errors the system makes. As I continue my work on fairness, I hope to develop more nuanced fairness metrics founded on semantic in addition to syntactic transcription quality, such as described in [38]. This will hopefully guide further fairness enforcement efforts more precisely towards speaker groups which are more lacking equal treatment.

5. References

[1] S. Verma and J. Rubin, "Fairness definitions explained," in *Proceedings of the International Workshop on Software Fairness*, ser. FairWare '18. New York, NY, USA: Association for Computing Machinery, May 2018, pp. 1–7.

[2] R. Tatman and C. Kasten, "Effects of Talker Dialect, Gender & Race on Accuracy of Bing Speech and YouTube Automatic Captions," in *Interspeech 2017*. ISCA, Aug. 2017, pp. 934–938.

[3] M. Garnerin, S. Rossato, and L. Besacier, "Gender Representation in French Broadcast Corpora and Its Impact on ASR Performance," Aug. 2019.

[4] M. Abushariah and M. Sawalha, "The effects of speakers' gender, age, and region on overall performance of Arabic automatic speech recognition systems using the phonetically rich and balanced Modern Standard Arabic speech corpus," Jan. 2013.

[5] R. Tatman, "Gender and Dialect Bias in YouTube's Automatic Captions," in *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, D. Hovy, S. Spruit, M. Mitchell, E. M. Bender, M. Strube, and H. Wallach, Eds. Valencia, Spain: Association for Computational Linguistics, Apr. 2017, pp. 53–59.

[6] M. K. Ngueajio and G. Washington, "Hey ASR System! Why Aren't You More Inclusive?" in *HCI International 2022 – Late Breaking Papers: Interacting with eXtended Reality and Artificial Intelligence*, ser. Lecture Notes in Computer Science, J. Y. C. Chen, G. Fragomeni, H. Degen, and S. Ntoa, Eds. Cham: Springer Nature Switzerland, 2022, pp. 421–440.

[7] M. Garnerin, S. Rossato, and L. Besacier, "Investigating the Impact of Gender Representation in ASR Training Data: A Case Study on Librispeech," in *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*, M. Costa-jussa, H. Gonen, C. Hardmeier, and K. Webster, Eds. Online: Association for Computational Linguistics, Aug. 2021, pp. 86–92.

[8] S. Feng, B. M. Halpern, O. Kudina, and O. Scharenborg, "Towards inclusive automatic speech recognition," *Computer Speech & Language*, vol. 84, p. 101567, Mar. 2024.

[9] Y. Wu, D. Rough, A. Bleakley, J. Edwards, O. Cooney, P. R. Doyle, L. Clark, and B. R. Cowan, "See What I'm Saying? Comparing Intelligent Personal Assistant Use for Native and Non-Native Language Speakers," in *22nd International Conference on Human-Computer Interaction with Mobile Devices and Services*, ser. MobileHCI '20. New York, NY, USA: Association for Computing Machinery, Oct. 2020, pp. 1–9.

[10] S. Hollands, D. Blackburn, and H. Christensen, "Evaluating the Performance of State-of-the-Art ASR Systems on Non-Native English using Corpora with Extensive Language Background Variation," in *Proc. Interspeech 2022*, 2022, pp. 3958–3962.

[11] Y. Zhang, Y. Zhang, B. Halpern, T. Patel, and O. Scharenborg, "Mitigating bias against non-native accents," in *Proc. Interspeech 2022*, 2022, pp. 3168–3172.

[12] S. Feng, O. Kudina, B. M. Halpern, and O. Scharenborg, "Quantifying Bias in Automatic Speech Recognition," Apr. 2021.

[13] A. B. Wassink, C. Gansen, and I. Bartholomew, "Uneven success: Automatic speech recognition and ethnicity-related dialects," *Speech Communication*, vol. 140, pp. 50–70, May 2022.

[14] G. I. Winata, S. Cahyawijaya, Z. Liu, Z. Lin, A. Madotto, P. Xu, and P. Fung, "Learning Fast Adaptation on Cross-Accented Speech Recognition," Mar. 2020.

[15] J. L. Martin and K. Tang, "Understanding Racial Disparities in Automatic Speech Recognition: The Case of Habitual "be"," in *Interspeech 2020*. ISCA, Oct. 2020, pp. 626–630.

[16] R. Dorn, "Dialect-Specific Models for Automatic Speech Recognition of African American Vernacular English," in *Proceedings of the Student Research Workshop Associated with RANLP 2019*, V. Kovatchev, I. Temnikova, B. Šandrih, and I. Nikolova, Eds. Varna, Bulgaria: INCOMA Ltd., Sep. 2019, pp. 16–20.

[17] M. Masson, "Identifying non-native English speech patterns in ASR systems," *International Speech Communication Association Doctoral Consortium*, 2023.

[18] L. Lima, V. Furtado, E. Furtado, and V. Almeida, "Empirical Analysis of Bias in Voice-based Personal Assistants," in *Companion Proceedings of The 2019 World Wide Web Conference*. San Francisco USA: ACM, May 2019, pp. 533–538.

[19] H.-J. Na and J.-S. Park, "Accented Speech Recognition Based on End-to-End Domain Adversarial Training of Neural Networks," *Applied Sciences*, vol. 11, p. 8412, Sep. 2021.

[20] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common Voice: A Massively-Multilingual Speech Corpus," Mar. 2020.

[21] J. Godfrey, E. Holliman, and J. McDaniel, "SWITCHBOARD: Telephone speech corpus for research and development," in *[Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, Mar. 1992, pp. 517–520 vol.1.

[22] M. Marion, A. Üstün, L. Pozzobon, A. Wang, M. Fadaee, and S. Hooker, "When Less is More: Investigating Data Pruning for Pretraining LLMs at Scale," Sep. 2023.

[23] S. Sun, C.-F. Yeh, M.-Y. Hwang, M. Ostendorf, and L. Xie, "Domain Adversarial Training for Accented Speech Recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2018, pp. 4854–4858.

[24] W. Zhou, H. Wu, J. Xu, M. Zeineldeen, C. Lüscher, R. Schlüter, and H. Ney, "Enhancing and Adversarial: Improve ASR with Speaker Labels," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Jun. 2023, pp. 1–5.

[25] J. Zhang, Y. Peng, P. Van Tung, H. Xu, H. Huang, and E. S. Chng, "E2E-based Multi-task Learning Approach to Joint Speech and Accent Recognition," Jun. 2021.

[26] T. Viglino, P. Motlicek, and M. Cernak, "End-to-End Accented Speech Recognition," in *Interspeech 2019*. ISCA, Sep. 2019, pp. 2140–2144.

[27] A. Jain, M. Upreti, and P. Jyothi, "Improved Accented Speech Recognition Using Accent Embeddings and Multi-task Learning," in *Interspeech 2018*. ISCA, Sep. 2018, pp. 2454–2458.

[28] J. Li, V. Manohar, P. Chitkara, A. Tjandra, M. Picheny, F. Zhang, X. Zhang, and Y. Saraf, "Accent-Robust Automatic Speech Recognition Using Supervised and Unsupervised Wav2vec Embeddings," Oct. 2021.

[29] T. Tanaka, R. Masumura, H. Sato, M. Ihori, K. Matsuura, T. Ashihara, and T. Moriya, "Domain Adversarial Self-Supervised Speech Representation Learning for Improving Unknown Domain Downstream Tasks," in *Interspeech 2022*. ISCA, Sep. 2022, pp. 1066–1070.

[30] N. Das, S. Bodapati, M. Sunkara, S. Srinivasan, and D. H. Chau, "Best of Both Worlds: Robust Accented Speech Recognition with Adversarial Transfer Learning," Mar. 2021.

[31] P. Dheram, M. Ramakrishnan, A. Raju, I.-F. Chen, B. King, K. Powell, M. Saboowala, K. Shetty, and A. Stolcke, "Toward Fairness in Speech Recognition: Discovery and mitigation of performance disparities," in *Interspeech 2022*, Sep. 2022, pp. 1268–1272.

[32] T. Maekaku, J. Shi, X. Chang, Y. Fujita, and S. Watanabe, "HuBERTopic: Enhancing Semantic Representation of HuBERT through Self-supervision Utilizing Topic Model," Oct. 2023.

[33] A. Pasad, J.-C. Chou, and K. Livescu, "Layer-wise Analysis of a Self-supervised Speech Representation Model," Dec. 2022.

[34] A. Pasad, B. Shi, and K. Livescu, "Comparative layer-wise analysis of self-supervised speech models," Mar. 2023.

[35] K. Meng, D. Bau, A. Andonian, and Y. Belinkov, "Locating and Editing Factual Associations in GPT," 2022.

[36] C. Sekkat, F. Leroy, S. Mdhaffar, B. P. Smith, Y. Estève, J. Dureau, and A. Coucke, "Sonos Voice Control Bias Assessment Dataset:

A Methodology for Demographic Bias Assessment in Voice Assistants," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue, Eds. Torino, Italia: ELRA and ICCL, May 2024, pp. 15 056–15 075.

[37] "Loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés," Jan. 1978.

[38] T. Bañeras-Roux, M. Rouvier, J. Wottawa, and R. Dufour, "Un paradigme pour l'interprétation des métriques et pour mesurer la gravité des erreurs de reconnaissance automatique de la parole," in *35èmes Journées d'Études sur la Parole (JEP 2024) 31ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2024) 26ème Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL 2024)*. ATALA & AFPC, Jul. 2024, p. 191.

Multilingual Speech Translation System for code-mixed Indian Languages

Atanu Mandal

Jadavpur University
Kolkata, INDIA
atanumandal0491@gmail.com

Abstract

Recent advancements in Artificial Intelligence have unveiled promising prospects for multidisciplinary research, indicating a significant evolution in the field. As research continues, our focus is directed toward Multilingual Speech Translation in the Indian context. India is the most populous nation having rich diverse cultures. India stands out with people speaking more than 100 languages, however, none holds the status of National Language. With this linguistically diverse and intricate landscape, our research aims to offer innovative solutions to the prevalent communication challenges. The foundation of our idea is built on 3 pivotal Research Problems – *language identification*, *speech recognition*, and *machine translation*, each encompassing intricate and nuanced Research Questions (RQs). Within the scope of this paper, we present the RQs that are currently under investigation and give a brief outline of our strategies and methods for tackling these RQs.

Index Terms: Language Identification, Speech Recognition, Machine Translation, Indian Languages, Code-Mixed

1. Introduction

India, a nation known for its linguistic diversity and status as the most populous country globally, exhibits a rich cultural heritage across various domains such as history, language, cuisine, and more. The linguistic landscape of India is characterized by the prevalence of multiple languages, with data from the 2011 census[1] indicating that 26% of the population is bilingual and 7% is trilingual, underscoring the country's diverse linguistic legacy. To safeguard this heritage, the Constitution of India[2] has designated English and Hindi as the official languages, while conferring Scheduled Language status upon 22 languages which encompass a range of language families.

Our work motivation stems from addressing the challenge of communication barriers among individuals, where the mode of communication relies on either English or Hindi. A major problem related to solving the challenges is the amalgamation of languages. It is imperative to note that due to the multilingual nature of Indians, English or Hindi often gets blended with native languages. This phenomenon is termed as Code-Mixing and code-mixed utterances pose a significant challenge to effective communication among the populace, particularly in rural regions. To mitigate these communication hurdles, our proposition involves developing a Speech Translation system, comprising a sequence of Language Identification (LI), Automatic Speech Recognition (ASR), and Machine Translation (MT) systems, tailored specifically for Indic languages.

[1]https://censusindia.gov.in/census.website/data/census-tables
[2]https://www.mea.gov.in/Images/pdf1/Part17.pdf

2. Research Problems

This section outlines the tasks that are at the center of our attention. Figure 1 represents the schematic diagram of the proposed Multilingual Speech Translation System for code-mixed Indian Languages.

2.1. Language Identification of Speech Utterances

RQ1: Sentence Level: How to identify the language of utterances?
RQ2: Word Level: How to identify the language of each word from a multilingual Speech?

2.2. Speech Recognition

RQ3: Monolingual: Given an audio file of monolingual speech, how to transcribe the speech?
RQ4: Multilingual Code-Mixed: Given an audio file of an unknown language, how to transcribe the speech with each word into its respective script?

2.3. Machine Translation

RQ5: Many-to-Many Translation: How to translate text from any source language $X_{i\in n}$ to any target language $Y_{j\in n}$ using a single MT system, where there are n languages involved? Additionally, the system should be able to deal with code-mixed text input.

3. Methodologies

RQ1 and **RQ2:** The process of Language Identification (LI) holds significant importance as a preliminary step in the domain of ASR, focusing on recognizing a spoken utterance. Present-day systems capable of handling speech in multiple languages necessitate users to specify the language beforehand. The crucial role of LI emerges in situations where ASR systems struggle to comprehend spoken languages in multilingual contexts, especially for diverse linguistic landscapes such as India.

For **RQ1**, we proposed a Convolutional Recurrent Neural Network (CRNN) designed to process the Mel-frequency Cepstral Coefficients (MFCCs) features of utterances [1]. Our investigation involved comparing the CRNN framework with the CRNN with the Attention framework. Additionally, our research evaluated the resilience towards various languages and closed language groups, achieving high scores for accuracy. The analysis in Table 1 presents an overview of performance across different datasets and linguistic settings. Our findings [1] also highlight the framework's robustness against noise interference and its potential for adapting to new languages. Notably, the CRNN with the Attention framework showcased a

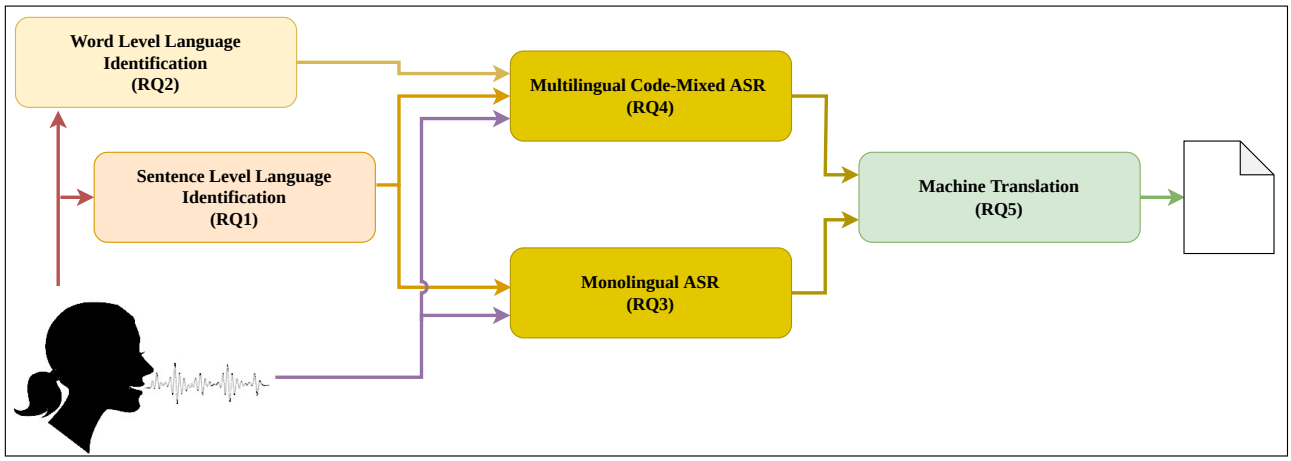


Figure 1: *Schematic Diagram of Multilingual Speech Translation System for code-mixed Indian Languages.*

comparable performance to the CRNN framework, but the Attention mechanism, despite incurring additional computational complexity, did not consistently outperform the CRNN framework.

For **RQ2**, we plan to use the transformer framework [2] that mirrors the concept employed in MT, where the encoder component of the framework is responsible for processing the source sentence, while the decoder handles the target sentence. In this context, we initialize the encoder module with speech features and the decoder module with the sequences of the words' language. The choice of speech features such as MFCCs, Filter Banks, or Wav2Vec2 [3] may vary depending on the characteristics of the audio environment.

Framework		CRNN	CRNN with Attention
Indian Dataset		**0.987**	**0.987**
Close Language Cluster	Cluster 1	0.974	**0.980**
	Cluster 2	**0.999**	**0.999**
	Cluster 3	**1**	0.999
European Language Dataset	No Noise	**0.967**	0.966
	White Noise	**0.912**	0.888

Table 1: *A comprehensive performance analysis*

RQ3 and **RQ4:** The proposed ASR addresses 2 distinct research scenerios. The prevailing state-of-the-art (SOTA) frameworks for ASR predominantly rely on monolingual data. Recent advancements such as OpenAI's Whisper and AI4Bharat's IndicVoices have set a performance benchmark for Indic Languages. Notably, ASR proposed by OpenAI and AI4Bharat necessitates prior language declaration from humans. Our proposed research strategy aims to declare the sentence-level identified language from the output of **RQ1** and word-level identified language from the outputs of **RQ2** to the framework without human intervention. We propose to use a sequence-to-sequence transformer framework with a double encoder and single decoder. The framework will be trained by replacing the decoder $< BOS >$ token with the identified $< Lang\ Tag >$ token from **RQ1**. One encoder of the proposed framework will be the sequence of language tags and the other will be sequences of speech features. Especially for **RQ4**, it is important to understand the acoustic cues when a person changes the language in utterance making communication more difficult. To mitigate the challenge we provide the sequence of language tokens identified from **RQ2** which will help in post-processing. The final output of **RQ4** will be text transcription with each word language tagged. For example, $W_1\#L_{i\in n}\ W_2\#L_{i\in n}\ \cdots\ W_k\#L_{i\in n}$ will be the output of **RQ4**, where there are n languages involved, W are the words and k is the length of the sequence.

RQ5: For MT, the majority of researchers still rely on the Transformer-based MT Systems where a single source language and single target language are involved. Our proposed MT system will translate from any source language $X_{i\in n}$ to any target language $Y_{j\in n}$ using a single system, where there are n languages involved. We plan to use the transformer-based framework, where $< BOS >$ will be replaced by respective $< Lang\ Tag >$ for both the encoder and decoder. Source $< Lang\ Tag >$ with word-level language tags will be used for code-mixed cases. Our intuition behind the idea is that initializing with a $< Lang\ Tag >$ and word-level language tag will help the framework to learn the semantics and syntax, respectively, of specific languages better. The system will act accordingly for code-mixed sentences as MT input by capturing the trigger point of language switching. This can be achieved by introducing word-level language tags with each word in the system. Overall, the objective of the proposed system is to produce SOTA models in Multilingual Code-Mixed Machine Translation.

4. Conclusion

In this paper, we presented our ideas for Multilingual Speech Translation in Indian Languages where multiple languages with multiple dialects are present. Furthermore, we shared our ideas on approaches to solving communication problems. There is further scope for research and improvement in existing systems due to low resources and dialect variations in Indian Languages. In the future, we would like to create a singleton Multilingual Speech Translation system without cascade frameworks for Indian Languages.

5. References

[1] A. Mandal, S. Pal, I. Dutta, M. Bhattacharya, and S. K. Naskar, "Is attention always needed? a case study on language identification from speech," *Natural Language Processing*, p. 1–27, 2024.

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[3] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," 2020.

Fairness in Speech Technologies with an Emphasis on Low-Resource Domains

Hend ElGhazaly

University of Sheffield, UK

helghazaly1@sheffield.ac.uk

Abstract

Speech technologies have become increasingly efficient and prevalent in various domains. However, there is growing concern that these technologies may perpetuate biases and unfairness in outcomes, particularly against people from marginalized communities. "Bias" refers to a performance gap that disproportionately disadvantages certain groups or individuals associated with protected attributes. For example, an Automatic Speech Recognition (ASR) system would be considered biased when there is a significant difference in its performance with different genders. Previous research has proved that the performance of speech technologies is influenced by various factors, such as the speaker's gender or race. Yet, there is a lack of research on practical solutions to mitigate these biases and promote fairness in speech technologies. My project aims to address the current limitations in existing research and investigate novel methods to improve the fairness of speech technologies. The project will also address the challenges in working with speech data from people with different language backgrounds and health conditions. In particular, Arabic and cognitive assessment tools will be used as case studies to assess the effectiveness of the proposed bias mitigation strategies in sparse data settings. This work has the potential to inform the development of inclusive speech technologies in low-resource domains.

Index Terms: Automatic speech recognition, gender bias, data augmentation, pitch manipulation, fairness

1. Motivation and research questions

Standard frameworks and effective methods to address notions of fairness in speech technologies are still not well established. While previous research has proposed various bias mitigation methods, they are not fully applicable to speech systems. Bias in speech technologies may arise from the unequal distribution of demographic groups in training sets [1], in transcripts' corrections [2, 3], or when using a system designed and trained in one context for a different purpose and target users [4, 5, 6, 7]. In speech research, studies on bias have included more emphasis on speakers' demographics [1, 8]. The most common types of demographic bias relate to gender [9, 7], age [7], race [10], accents and language variants [11, 4, 7, 12]. As there are articulation differences, such as in accents or speaking style, imbalanced datasets can lead to a mismatch between the speaker and the trained acoustic model [3]. Consequently, the speech models do not work equally well with all groups of people.

In low-resource domains, sparse datasets represent another challenge that has not been addressed yet. This is the case in healthcare speech and language technologies applications, where datasets are often small and not in the public domain because of ethical concerns. Predominantly, one of the biggest challenges in healthcare speech technologies is the scarcity of speech datasets particularly related to speech disorders [13]. In healthcare systems, biased systems lead to diagnostic inaccuracies, health disparities, and unequal access to treatments or services [14]. For example, biases in ASR systems that transcribe patients' discussions can lead to misinterpretations [5].

On the other hand, much work has shown the discrepancies in speech systems' performance and quantified the bias with the African American language [5, 11], Dutch [4, 3], but not with other low-resourced languages. One of the least studied in regards to bias is the Arabic language despite being one of the top world languages [15]. With limited and unbalanced speech datasets, it is crucial to recognize and address these biases to foster the development of fairer speech technologies.

In addition, most research studies focus on representation bias in data while many other bias types across the machine learning pipeline still need solutions. The interplay of various biases should also be evaluated rather than looking at one bias at a time. Few studies have addressed the bias against other demographics such as nationality, and disability and focused mainly on gender. Speech-specific attributes like dialect, voice timbre, and prosodic attributes have not been considered. The project thus aims to address two main research questions:

- **RQ1:** How can bias against speakers' demographics be mitigated with minimal performance degradation in speech technologies?
- **RQ2:** How effective are specific bias identification and mitigation strategies in improving outcomes within low-resource domains, such as Arabic and healthcare speech systems?

By addressing these questions, this project aims to contribute valuable insights to inform specific methods to combat biases in speech technologies leading to fair speech systems.

2. Results so far

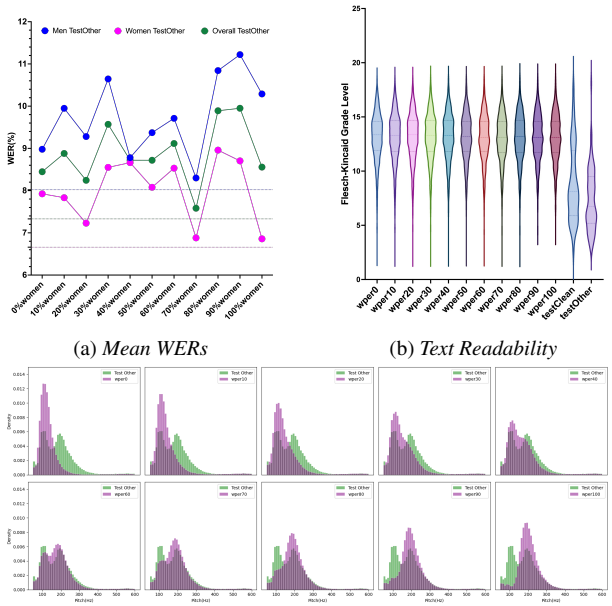
Several studies showed a correlation between the gender distribution in the training set and the performance of speech models, however different studies have come to different conclusions [3, 9, 16, 17]. The contradictory results suggest that other factors, besides the speakers' gender and demographics, might affect the performance. Therefore, I initially conducted a series of experiments to investigate the following hypothesis:

Hypothesis: The gender distribution within the training set, alongside other factors such as text difficulty, semantic similarity, and out-of-domain test sets, significantly impact the performance and fairness of ASR systems.

Artificial bias was induced by creating subsets from the original LibriSpeech [18] training set with varying percentages of women's audio files [1]. Each training subset was used to fine-tune Whisper's small model [19]. The Word Error Rate (WER) was used to compare each speaker's performance in the test set across the different models. The bias was calculated as the difference between the women's and men's WERs [3]. The re-

[1] Acknowledging that the gender spectrum is more diverse, the focus on these two genders is driven by their representation in the dataset used in this investigation.

sults revealed an irregular pattern between the gender ratios in training data and ASR performance as shown in Figure1a, challenging the notion that adjusting these ratios alone can enhance the ASR's performance. The imbalanced gender distribution also did not consistently correlate with improved recognition of the over-represented gender. The findings further suggested that speakers' pitch (F0) variability in the training set significantly affects ASR performance, emphasising the importance of a holistic approach to dataset composition. The text analysis showed that the difficulty and semantic similarity levels between the training and test sets were similar in all subsets and, therefore, were not contributing factors towards the differences observed in the performance.



(a) *Mean WERs* (b) *Text Readability*

(c) *Pitch distributions*

Figure 1: *Illustrations of the ASR performance, text analysis and pitch distributions in the evaluation of the Test Other set*

Based on the initial analysis, the worst-performing models in terms of lowest accuracy and biggest bias size were trained on data consisting of 90% and 100% women audio files (Figure1a). The two seemingly prominent factors among my investigated factors (i.e., gender representation, pitch distribution, text readability and semantic similarity) were gender and pitch distributions. I then attempted to reduce the gender bias by augmenting data while considering these two factors; gender and pitch. I employed data augmentation through pitch shifting with three selection strategies to increase the initial training set size:

1. Gender balanced, to achieve a balanced gender distribution;
2. Pitch distribution balanced, to match a target normal distribution of pitches, regardless of gender; and
3. Random selection, without considering gender or pitch.

In each scenario, I trained the model with the initial 90% women dataset plus the augmented data and evaluated the performance on the TestOther set. The results showed that gender-balanced and random augmentations yielded the best overall WER and reductions in bias. Random augmentation provided the lowest WER for women but did not reduce bias as effectively as gender-balanced or pitch-based augmentations. Pitch-based augmentation achieved the lowest bias size. However, the improvements were not significant, suggesting that these augmentation strategies may not be enough to achieve fair results in speech recognition tasks. To confirm whether the observed patterns are linked to the speech mode or domain, I utilized EdAcc [20] as an alternative domain evaluation set. The results illustrated that while certain augmentation methods may improve fairness, they can also negatively impact overall performance.

3. Current and future research directions

The initial experiments revealed that adjusting the gender ratios and using data augmentation did not significantly improve the model's performance and fairness. Therefore, the next step is to explore advanced pre-processing and in-processing techniques, adjusting the training algorithm itself. My goal is to propose a new bias mitigation method tailored specifically for speech models and data, aimed at reducing the performance gap between different groups. As new datasets have recently become available with more labelled speaker demographics, I will also investigate other protected attributes beyond gender, such as age, ethnicity, and education level. Understanding the interplay between these factors will help identify which biases are most harmful. Subsequently, I will conduct case studies on Arabic and healthcare speech systems to evaluate the proposed bias mitigation strategies in important data-sparse settings. It is anticipated that methods primarily tested on English-speaking, healthy users may not perform as well in low-resource domains. Therefore, I will address the challenge of working with data from these domains and develop novel methods to ensure that current speech technologies can effectively accommodate individuals with diverse backgrounds and health conditions.

4. Challenges

My research has several challenges, particularly due to limited data and resources. The scarcity of data is especially pronounced in low-resource domains, i.e., Arabic and health-related datasets, which further complicates the research process. Unlike image and text data, speech data presents unique characteristics and complexities, rendering the existing de-biasing methods used in these other domains inapplicable. This necessitates the development of novel approaches tailored specifically for speech data to mitigate biases effectively in speech systems. Furthermore, working with speech data presents relatively distinct challenges due to its unique characteristics. Speech data is continuous and temporal, involving sound waves that require complex preprocessing steps such as noise reduction and feature extraction. Speech involves various acoustic-phonetic features, including prosody (e.g. pitch (F0), loudness), and phonetic content (e.g. MFCCs, phoneme duration). These features make speech data more complex to transform or augment. This complexity differs from image data, which is spatially organized and represented as pixel grids, and text data, which consists of sequences of words or characters. The preprocessing of images and text often involves resizing and tokenization, respectively, which are different from the procedures required for speech. Mitigating bias in speech models is particularly challenging due to the nature of speech data and the influence of factors like dialects and acoustic environments. Unlike bias mitigation methods in images and text, with speech, we must account for diverse speaker demographics and varying recording conditions. Additionally, evaluation metrics for speech differ from those used for images and text, necessitating innovative approaches for effective de-biasing in speech technologies.

5. References

[1] J. Meyer, L. Rauchenstein, J. D. Eisenberg, and N. Howell, "Artie bias corpus: An open dataset for detecting demographic bias in speech applications," in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 2020, pp. 6462–6468.

[2] C. Cucchiarini, O. van Herwijnen, F. Smits *et al.*, "Jasmin-cgn: Extension of the spoken dutch corpus with speech of elderly people, children and non-natives in the human-machine interaction modality," in *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, 2006.

[3] S. Feng, O. Kudina, B. M. Halpern, and O. Scharenborg, "Quantifying bias in automatic speech recognition," *arXiv preprint arXiv:2103.15122*, 2021.

[4] Y. Zhang, Y. Zhang, B. M. Halpern, T. Patel, and O. Scharenborg, "Mitigating bias against non-native accents," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2022, 2022, pp. 3168–3172.

[5] J. L. Martin and K. E. Wright, "Bias in automatic speech recognition: The case of african american language," *Applied Linguistics*, 2022.

[6] Y. Wu, D. Rough, A. Bleakley, J. Edwards, O. Cooney, P. R. Doyle, L. Clark, and B. R. Cowan, "See what i'm saying? comparing intelligent personal assistant use for native and non-native language speakers," in *22nd international conference on human-computer interaction with mobile devices and services*, 2020, pp. 1–9.

[7] S. Feng, B. M. Halpern, O. Kudina, and O. Scharenborg, "Towards inclusive automatic speech recognition," *Computer Speech & Language*, vol. 84, p. 101567, 2024.

[8] G. Fenu, M. Marras, G. Medda, G. Meloni *et al.*, "Fair voice biometrics: Impact of demographic imbalance on group fairness in speaker recognition," in *Interspeech*. International Speech Communication Association, 2021, pp. 1892–1896.

[9] M. Garnerin, S. Rossato, and L. Besacier, "Investigating the impact of gender representation in asr training data: A case study on librispeech," in *3rd Workshop on Gender Bias in Natural Language Processing*. Association for Computational Linguistics, 2021, pp. 86–92.

[10] A. B. Wassink, C. Gansen, and I. Bartholomew, "Uneven success: automatic speech recognition and ethnicity-related dialects," *Speech Communication*, vol. 140, pp. 50–70, 2022.

[11] A. Koenecke, A. Nam, E. Lake, J. Nudell, M. Quartey, Z. Mengesha, C. Toups, J. R. Rickford, D. Jurafsky, and S. Goel, "Racial disparities in automated speech recognition," *Proceedings of the National Academy of Sciences*, vol. 117, no. 14, pp. 7684–7689, 2020.

[12] M. A.-A. M. Abushariah, R. N. Ainon, R. Zainuddin, M. Elshafei, and O. O. Khalifa, "Arabic speaker-independent continuous automatic speech recognition based on a phonetically rich and balanced speech corpus." *Int. Arab J. Inf. Technol.*, vol. 9, no. 1, pp. 84–93, 2012.

[13] S. Latif, J. Qadir, A. Qayyum, M. Usama, and S. Younis, "Speech technology for healthcare: Opportunities, challenges, and state of the art," *IEEE Reviews in Biomedical Engineering*, vol. 14, pp. 342–356, 2020.

[14] T. Panch, H. Mattie, and R. Atun, "Artificial intelligence and algorithmic bias: implications for health systems," *Journal of global health*, vol. 9, no. 2, 2019.

[15] C. Hazirbas, Y. Bang, T. Yu, P. Assar, B. Porgali, V. Albiero, S. Hermanek, J. Pan, E. McReynolds, M. Bogen, P. Fung, and C. C. Ferrer, "Casual conversations v2: Designing a large consent-driven dataset to measure algorithmic bias and robustness," 2022.

[16] Y. Meng, Y.-H. Chou, A. T. Liu, and H.-y. Lee, "Don't speak too fast: The impact of data bias on self-supervised speech models," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 3258–3262.

[17] L. Maison and Y. Estève, "Some voices are too common: Building fair speech recognition systems using the common voice dataset," *arXiv preprint arXiv:2306.03773*, 2023.

[18] 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.

[19] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning*. PMLR, 2023, pp. 28 492–28 518.

[20] R. Sanabria, N. Bogoychev, N. Markl, A. Carmantini, O. Klejch, and P. Bell, "The edinburgh international accents of english corpus: Towards the democratization of english asr," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

Articulatory Gesture-Constrained Transfer Learning for Cross-Corpus Speech Emotion Recognition

Shreya G. Upadhyay

National Tsing Hua University, Taiwan

shreya@gapp.nthu.edu.tw

Abstract

Cross-corpus Speech Emotion Recognition (SER) is crucial for various everyday applications. Existing studies in cross-corpus emotion transfer tasks typically focus on constraining acoustic features to adapt features, domains, or labels across corpora. However, acoustic features exhibit high variability and instability due to factors such as speaker differences, domain variations, and recording conditions. This study adopts a contrastive approach by using emotion-specific articulatory movements as the foundational units for analysis. By taking a step down from acoustics to more stable articulatory gestures, we aim to improve emotion transfer in SER tasks. Our experiments and analyses show interesting insights into the commonality of these articulatory gestures, demonstrating their potential as reliable constraints for emotion transfer.

Index Terms: speech emotion recognition, articulatory movements, cross-corpus, transfer learning

1. Motivation

Developing robust Speech Emotion Recognition (SER) strategies is crucial for applications in healthcare, security, education, and entertainment [1]. Current approaches in cross-corpus SER models often address domain discrepancies through transfer learning, semi-supervised learning, and few-shot learning [2,3]. Additional techniques include optimizing distance metrics (e.g., Wasserstein [4]), adversarial training to mitigate domain memorization [5], and generating synthetic data using GANs [6]. Some studies also explore anchoring methods to facilitate emotion transfer across corpora [7]. However, these methods primarily rely on acoustic information as the predominant conditioning factor, which is prone to variability and may not efficiently regulate emotion transfer across domains. Articulatory movements represent fundamental units of speech production, governing physical speech movements and exhibiting less variability than acoustic features influenced by external factors. This study focuses on articulatory movements, which exhibit greater consistency and stability across diverse speakers and environments. By identifying common articulatory gestures associated with specific emotions across corpora and using them as constraints for emotion transfer, our approach aims to enhance the learning of emotional cues in cross-corpus settings.

2. Key Research Question

The primary objective is to determine if similar articulatory gestures exist across different corpora that can be effectively utilized for specific emotions. This exploration involves analyzing articulatory patterns across corpora to identify common gestures. Once these common articulatory gestures are identified, the next challenge is to develop a method that uses them as constraints to improve unsupervised cross-corpus SER. This involves developing models that leverage this prior knowledge to guide learning, thereby aligning the emotional modulation between two corpora, and improving the accuracy of cross-corpus emotion recognition. This research aims to establish emotion-specific articulatory gestures as a reliable feature to enhance the generalizability of SER systems.

3. Methodology

3.1. Multi-Modal Affective Corpora

While the MRI and EMA systems can record articulatory movements, such data are challenging to obtain. Numerous multimodal corpora include visual information. We utilize these visual datasets to analyze articulatory gestures. For this study, we require corpora that include both speech and visual modalities. We select the CREMA-D [8] and MSP-IMPROV [9] datasets due to their diverse settings and multimodal nature. CREMA-D comprises audiovisual recordings of actors performing emotions in controlled conditions, providing both facial expressions and vocal cues. MSP-IMPROV features improvised emotional speech, capturing spontaneous emotional expressions through both audio and visual data. Here, in examining articulatory movements, our focus is on 12 mouth landmark points extracted from visual data.

3.2. Articulatory Movement Analyses

Inspired by the phonetic anchoring work in the literature [7], we focus on vowel articulation in this study. We extract all the facial landmarks and pre-process them by aligning and normalizing the data so that the landmarks from both corpora are in the same space. We conduct two experiments to identify gesture commonalities: articulatory movement transition analysis and emotion modulation similarity.

The first experiment, articulatory movement transition analysis, tracks how mouth shape evolves over time during speech utterances, specifically focusing on vowel articulation. We meticulously track the movement of mouth landmarks across frames, calculating the Euclidean distance of each landmark from its baseline position (the coordinates of the first frame). We assume the first frame represents the beginning of pronouncing the phoneme, serving as the neutral state. By summing these distances for each timestamp and subsequently computing the average change and standard deviation across multiple samples, we gained insights into the temporal dynamics of mouth shape variation. This experiment reveals that some gestures are common to certain emotions across the two corpora. For example, Figure 1 shows the plot of this experiment for the *Happy* emotion across different vowel articulations. We can observe that

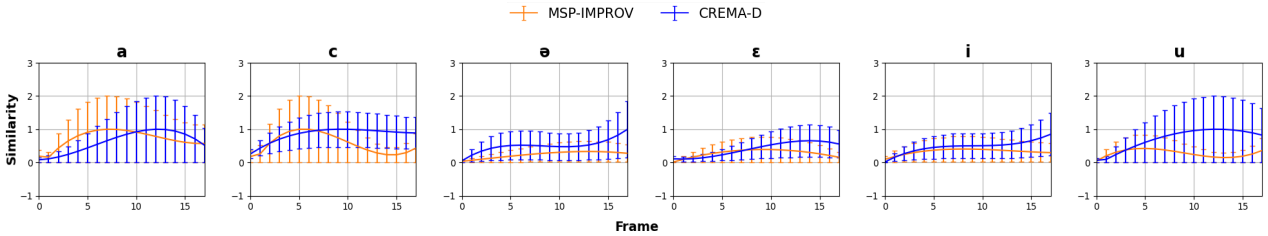


Figure 1: *Mouth landmark transition analysis over frames for different phonemes in Happy emotion, highlighting mean and standard deviation across both corpora.*

Table 1: *Emotion modulation similarity (cosine similarity) over CREMA-D and MSP-IMPROV corpora.*

	Angry		Happy		Sad	
	Mean	±STD	Mean	±STD	Mean	±STD
Overall	0.73	0.5	0.77	0.6	0.73	0.6
a	0.72	0.4	**0.77**	0.3	0.67	0.4
c	**0.81**	0.6	0.51	0.4	0.62	0.4
ə	0.74	0.3	0.68	0.3	**0.73**	0.5
ɛ	0.73	0.4	0.75	0.5	0.71	0.6
i	0.74	0.3	**0.83**	0.4	0.71	0.7
u	**0.78**	0.5	0.71	0.2	**0.74**	0.4

the vowels /i/, /ɛ/, and /a/ exhibit more consistent transition patterns across frames.

For a deeper insight, our second experiment, emotion modulation similarity, aims to uncover commonalities in emotion expression across different corpora, particularly concerning the *Neutral* emotion. In this experiment, we conduct a comparative analysis by subtracting *Neutral* emotion samples from those depicting specific emotional states. We meticulously analyze all 12 landmark points, including their x and y coordinates, as well as time-series data, to capture the full spectrum of facial expression dynamics. Using cosine similarity metrics, we estimate the degree of modulation similarity among sets of emotional samples within each corpus. Table 1 presents the similarity scores alongside their mean and standard deviation. Our findings indicate that while overall sentence modulations exhibit minimal variation, different vowel articulations show varying levels of emotion modulation similarity across specific emotional states. For instance, during the *Happy* emotion, vowels /a/ and /i/ show higher similarity compared to other vowels.

Our experiments have revealed insights that suggest integrating articulatory gestures for reliable emotion transfer in SER tasks. Articulatory gesture anchoring could provide a stable foundation for enhancing cross-corpus emotion recognition.

3.3. Initial Cross-Corpus SER Results

We have shown the initial performance of cross-corpus SER with our proposed Articulatory Gesture constraint (AGC) idea, along with the baseline models for comparison. Table 2 presents the performance across both corpora. Our proposed model (*AGC*), which incorporates the loss with feature space reduction, constrained on articulatory gestures pre-knowledge, shows competitive performance (achieving a UAR of 54.87% for CREMA→IMPRO and 62.11% for IMPRO→CREMA). Due to spatial limitations, we cannot elaborate further on the modeling details. However, Table 2 indicates that our model *AGC* does not yet outperform the *Vowel-Anch* model. This could be attributed to several factors, such as the specific dataset characteristics, the complexity of the emotion recognition task, or the effectiveness of the articulatory gesture constraints compared to vowel-based anchoring. Therefore, we are currently exploring what could be the possible reason behind this performance difference and also exploring some alternative techniques to leverage the articulatory gesture in the SER task.

Table 2: *Cross-corpus SER performances (in UAR %) for 4-category SER task, tested with CREMA-D as the source and MSP-IMPROV as the target (CREMA→IMPRO) and vice-versa (IMPRO→CREMA).*

	CREMA→IMPRO	IMPRO→CREMA
Few-shot [3]	51.08	60.27
Ensemble [10]	52.41	61.95
Vowel-Anch [7]	55.33	63.18
AGC	**54.87**	**62.11**

4. Key and Potential Challenges

Leveraging common articulatory gestures in cross-corpus SER presents significant challenges. Firstly, accurately segmenting facial landmark movements based on phonetics is complicated by variability in speech patterns and emotions. Improving segmentation accuracy necessitates exploring clustering methods to categorize articulatory gestures based on physical characteristics rather than solely relying on phonetic boundaries. Secondly, capturing articulatory gestures using facial landmarks across more than one dimension and time-series data encounters reliability issues influenced by lighting variations, facial poses, and the complexity of facial movements.

5. Plans for the future and thesis roadmap

Moving forward, our research aims to address these challenges effectively. We plan to investigate advanced clustering techniques to enhance the segmentation of articulatory gestures in speech data. Additionally, we will explore adaptive learning functions to improve the robustness of capturing articulatory movements across varying conditions. Our thesis roadmap includes developing and evaluating novel methods that leverage stable common articulatory gestures as a foundation for enhancing cross-corpus SER.

6. Expected Contributions

By focusing on common articulatory gestures, which are more stable and less affected by variability compared to acoustic features, we aim to enhance the generalization of SER systems across different corpora and domains. Our approach intends to provide a more reliable foundation for emotion transfer tasks, thereby improving the accuracy and effectiveness of SER models in recognizing emotions consistently across different contexts. This contribution is pivotal for advancing the applicability and reliability of emotion recognition technologies in various practical domains.

7. References

[1] C.-C. Lee, K. Sridhar, J.-L. Li, W.-C. Lin, B.-H. Su, and C. Busso, "Deep representation learning for affective speech signal analysis and processing: Preventing unwanted signal disparities," *IEEE Signal Processing Magazine*, vol. 38, no. 6, pp. 22–38, 2021.

[2] S. Parthasarathy and C. Busso, "Semi-supervised speech emotion recognition with ladder networks," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 28, pp. 2697–2709, 2020.

[3] Y. Ahn, S. J. Lee, and J. W. Shin, "Cross-corpus speech emotion recognition based on few-shot learning and domain adaptation," *IEEE Signal Processing Letters*, vol. 28, pp. 1190–1194, 2021.

[4] J. Gideon, M. McInnis, and E. Mower Provost, "Improving cross-corpus speech emotion recognition with adversarial discriminative domain generalization (ADDoG)," *IEEE Transactions on Affective Computing*, vol. 12, no. 4, pp. 1055–1068, October-December 2021.

[5] M. Abdelwahab and C. Busso, "Domain adversarial for acoustic emotion recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 12, pp. 2423–2435, December 2018.

[6] B.-H. Su and C.-C. Lee, "A conditional cycle emotion gan for cross corpus speech emotion recognition," in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 351–357.

[7] S. G. Upadhyay, L. Martinez-Lucas, B.-H. Su, W.-C. Lin, W.-S. Chien, Y.-T. Wu, W. Katz, C. Busso, and C.-C. Lee, "Phonetic anchor-based transfer learning to facilitate unsupervised cross-lingual speech emotion recognition," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[8] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "Crema-d: Crowd-sourced emotional multimodal actors dataset," *IEEE transactions on affective computing*, vol. 5, no. 4, pp. 377–390, 2014.

[9] C. Busso, S. Parthasarathy, A. Burmania, M. AbdelWahab, N. Sadoughi, and E. M. Provost, "Msp-improv: An acted corpus of dyadic interactions to study emotion perception," *IEEE Transactions on Affective Computing*, vol. 8, no. 1, pp. 67–80, 2016.

[10] W. Zehra, A. R. Javed, Z. Jalil, H. U. Khan, and T. R. Gadekallu, "Cross corpus multi-lingual speech emotion recognition using ensemble learning," *Complex & Intelligent Systems*, vol. 7, no. 4, pp. 1845–1854, 2021.

Enhancing ASR Systems for Low-Resource Indian Languages

Rishabh Kumar

IIT Bombay, Mumbai, India

krrishabh@cse.iitb.ac.in

1. Research Question & Motivation

Recent advancements in deep learning have significantly enhanced the performance of automatic speech recognition (ASR) systems for several languages [1]. However, these advancements are on the larger dataset, which is not true for low-resource languages. This challenge is particularly pronounced in India, where there are 1652 native languages, and 22 are scheduled (official) languages. Owing to its diversity, most of the Indian languages are low-resource languages compared to other widely spoken global languages. To advance and innovate in ASR for Indian languages, we chose the Sanskrit language, which influences 19 out of 22 scheduled languages. Sanskrit serves as a crucial starting point due to Sanskrit's rich historical and linguistic background emphasizes these complexities [2, 3].

2. Methodology

To advance in ASR for the Sanskrit language, we built **Vāksañcayaḥ**[1], our Sanskrit ASR corpus, with 78 hours of speech data spoken by 27 unique speakers (Sanskrit scholars) [4]. It consists of 46K sentences and a vocabulary size of 91K. The training dataset consists of 56 hours of audio data with 12 speakers (34,000 utterances), whereas the valid and test dataset consists of 3 speakers each (6,004 utterances). The dataset was prepared while accounting for several of the challenges, like Sanskrit's rich cultural heritage, its linguistic characteristics, and the limited availability of resources in both speech and text.

To further enhance the efficiency of creating and refining Sanskrit ASR datasets, we developed **VAgyojaka**, an open-source post-editing and annotation tool for automatic speech recognition (ASR) [5]. VAgyojaka is designed to reduce the human effort required to correct ASR results. It adopts a dictionary-based lookup method to highlight incorrect words in the ASR transcript and provide suggestions by generating the closest valid words. Our tool reduces post-editing effort and time by one-third compared to traditional editing methods. Additionally, VAgyojaka serves as an End-to-End Provenance tool for Spoken Translation, facilitating post-editing and annotation for ASR, machine translation (MT), and text-to-speech (TTS) tasks, ultimately aiming to streamline the correction process across these applications, as shown in Figure 1.

Using the Vāksañcayaḥ dataset, we propose a novel large-vocabulary ASR system for Sanskrit, the first of its kind. Our design choices, influenced by Sanskrit's phonemic orthography, include three encoding schemes for language model tokens: word-based encoding, byte pair encoding (BPE), and a

[1]This speech corpus can be accessed from www.cse.iitb.ac.in/~asr

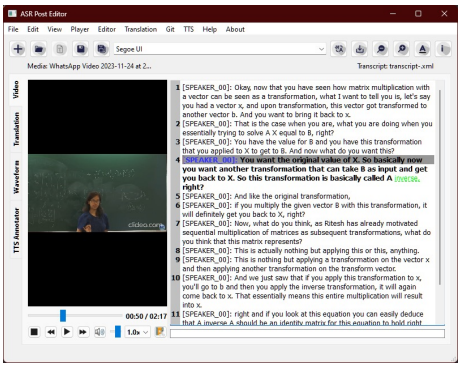


Figure 1: *Screenshot of the ASR Post-Editing tool*

new vowel split encoding inspired by syllabic structure theories. We use the Sanskrit Library Phonetic (SLP1) encoding scheme to address redundancy in Unicode representations, which preserves phonemic orthography. We focus on two graphemic representations: native script and SLP1. We train a Time Delay Neural Network (TDNN) [6] based acoustic model and a subword-based language model, forming a hybrid ASR system. Additionally, we extend our findings to develop ASR systems for Telugu and Gujarati, incorporating SLP1 adaptations for these languages [4].

Traditional word-based ASR models struggle with out-of-vocabulary (OOV) words and rare words in Sanskrit due to its complex morpho-syntactic regularities and sandhi [7]. Subword language models (LMs) help to some extent but still face limitations in handling long-range dependencies and semantic context, as shown in Figure 2.

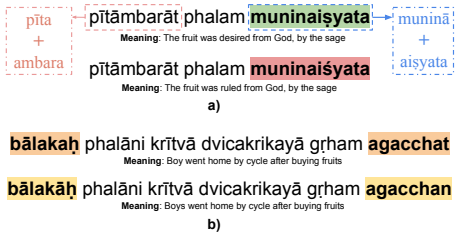


Figure 2: *Examples of Linguistic Challenges: a) Semantic Context (Split for compound and sandhied tokens are shown in boxes) and b) Long Range Dependencies in Sanskrit Text.*

To tackle these challenges, we propose an approach that combines a subword-based language model with a post-processing module enriched with morpho-syntactic information. We used the same TDNN based acoustic model and a

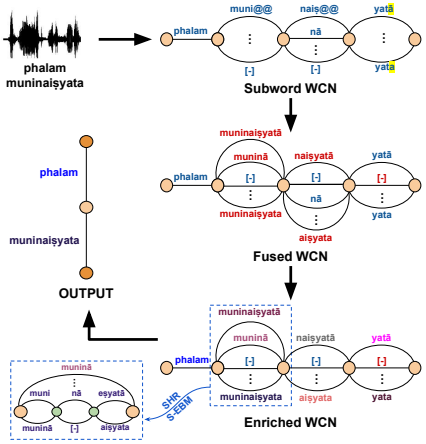


Figure 3: *Different phases of search space of the proposed model for sample audio. A) The audio input is passed to the state-of-the-art subword ASR model. B) The n-best lattice is converted to the* ***Subword WCN****. C) From the subword WCN, we formulate a new* ***Fused WCN*** *where we combine the subword entries deterministically and form fused entries. D) By using a lexicon-driven shallow parser, we construct the* ***Enriched WCN*** *from the fused WCN by incorporating the information from the shallow parser. E) Final selection proceeds based on EBM scores in the enriched WCN.*

subword-based language model, forming a hybrid ASR system. This system converts the subword-based ASR lattice into an ASR word confusion network (WCN) [8], which is refined using a lexicon-driven shallow parser [9]. The parser filters invalid combinations, lists morphological interpretations, segments compound words, and handles sandhi. Using an energy-based model (EBM) framework, we recalibrate scores in the WCN to improve accuracy [10], as shown in Figure 3.

Building on this foundation, we introduce the first large-scale speech-text dataset in Hindi using our tool, VAgyojaka, to capture human-human spoken conversations. We have conducted extensive experiments using several state-of-the-art models, including wav2vec [1, 11], whisper [12, 13], and large language models like ChatGPT [14]. The experimental results reveal significant potential for improvement, particularly within the context of Hindi ASR. This underscores the need for further advancements and innovations in modeling techniques to better handle the linguistic complexities and variability inherent in Hindi and other low-resource languages.

3. Result

Figure 4 shows the performance of ASR systems for Sanskrit, Telugu, and Gujarati languages using different combinations of scripts and language model units. We observe that the use of SLP1 as a graphemic representation scheme performs best for all three languages [4]. The training datasets consist of 56 hours for Sanskrit, 36 hours for Telugu, and 33 hours for Gujarati.

Table 1 compares the word error rate (WER) for all the ASR systems. We find that the subword-based system, Vak-BPE outperforms the word-based system, Vak-Word, by a considerable margin with a decrement of 19.03 WER. Moreover, we found that the search space enriched with linguistic information exceeds the state-of-the-art Vak-BPE system. Among all the search space enrichment approaches, Morph-WCN-EBM

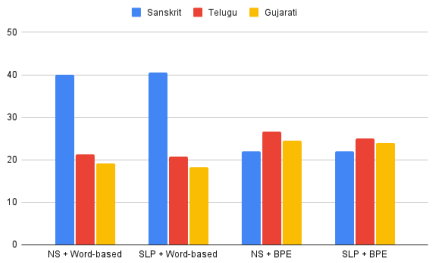


Figure 4: *Word Error Rate (WER) comparison across different language models and scripts. (NS refers to Native Script, Word-based and BPE indicates different language model unit)*

Method	DEV	TEST
Vak-Word	35.68	42.52
Vak-BPE	18.62	23.49
Morph-NBest-EBM	17.80	22.41
Morph-WCN-morphLM	16.18	20.26
Morph-WCN-EBM	**14.15**	**16.31**

Table 1: *WER for ASR systems using different methods*

performs the best with a WER of 16.31 [10].

Strategy	Vak-BPE	Morph-NBest-EBM	Morph-WCN-EBM
Compound Analysis	43.33	44.01	55.67
Syncretism	46.25	47.28	59.47
Homonymy	51.54	52.57	64.79

Table 2: *Results for linguistic analysis for the three different ASR configurations. F-score is used as the metric*

Table 2 shows the results of the morphological analysis for all three systems. Our proposed Morph-WCN-EBM system provides state-of-the-art results in identifying the correct morphological tags, stems, and word forms with an F-score of 75.63. Since Sanskrit is a fusional language where morpheme encodes multiple grammatical categories, it is a tough challenge because Sanskrit has around 1,635 tags. Our system Morph-WCN-EBM significantly improves the identification of compound words with correct components and can resolve obscurity due to syncretism and homonymy [10].

4. Contributions & Future Work

In this study, we developed a large-vocabulary ASR system for Sanskrit using the Vāksañcayaḥ corpus and the VAgyojaka post-editing tool, achieving significant improvements in Word Error Rate (WER) with SLP1 encoding and subword-based language models. Our methods also showed promise for Telugu and Gujarati, suggesting broader applicability for low-resource Indian languages. The findings underscore the potential of combining subword tokenization strategies with search space enrichment through the incorporation of morphological and lexical information in enhancing ASR systems. Future work should be focused on creating a large-scale dataset benchmark and findings to provide new and unique insights into LLM-enhanced ASR.

5. References

[1] 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.

[2] P. Goyal, G. Huet, A. Kulkarni, P. Scharf, and R. Bunker, "A distributed platform for sanskrit processing," in *Proceedings of COLING 2012*, 2012, pp. 1011–1028.

[3] A. Krishna, P. Satuluri, and P. Goyal, "A dataset for sanskrit word segmentation," in *Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, 2017, pp. 105–114.

[4] D. Adiga, R. Kumar, A. Krishna, P. Jyothi, G. Ramakrishnan, and P. Goyal, "Automatic speech recognition in sanskrit: A new speech corpus and modelling insights," *arXiv preprint arXiv:2106.05852*, 2021.

[5] R. Kumar, D. Adiga, M. Kothyari, J. Dalal, G. Ramakrishnan, and P. Jyothi, "Vagyojaka: An annotating and post-editing tool for automatic speech recognition." in *INTERSPEECH*, 2022, pp. 857–858.

[6] A. H. Waibel, T. Hanazawa, G. E. Hinton, K. Shikano, and K. J. Lang, "Phoneme recognition using time-delay neural networks," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 37, pp. 328–339, 1989. [Online]. Available: https://api.semanticscholar.org/CorpusID:9563026

[7] A. Krishna, B. Santra, A. Gupta, P. Satuluri, and P. Goyal, "A graph-based framework for structured prediction tasks in sanskrit," *Computational Linguistics*, vol. 46, no. 4, pp. 785–845, 2021.

[8] G. Tür, J. H. Wright, A. L. Gorin, G. Riccardi, and D. Z. Hakkani-Tür, "Improving spoken language understanding using word confusion networks," in *Interspeech*, 2002. [Online]. Available: https://api.semanticscholar.org/CorpusID:5928650

[9] G. Huet, "A functional toolkit for morphological and phonological processing, application to a sanskrit tagger," *Journal of Functional Programming*, vol. 15, no. 4, pp. 573–614, 2005.

[10] R. Kumar, D. Adiga, R. Ranjan, A. Krishna, G. Ramakrishnan, P. Goyal, and P. Jyothi, "Linguistically informed post-processing for asr error correction in sanskrit." in *INTERSPEECH*, 2022, pp. 2293–2297.

[11] V. S. Lodagala, S. Ghosh, and S. Umesh, "Ccc-wav2vec 2.0: Clustering aided cross contrastive self-supervised learning of speech representations," in *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2023, pp. 1–8.

[12] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International conference on machine learning*. PMLR, 2023, pp. 28 492–28 518.

[13] K. S. Bhogale, S. Sundaresan, A. Raman, T. Javed, M. M. Khapra, and P. Kumar, "Vistaar: Diverse benchmarks and training sets for indian language asr," *arXiv preprint arXiv:2305.15386*, 2023.

[14] Y. Liu, T. Han, S. Ma, J. Zhang, Y. Yang, J. Tian, H. He, A. Li, M. He, Z. Liu *et al.*, "Summary of chatgpt-related research and perspective towards the future of large language models," *Meta-Radiology*, p. 100017, 2023.

Towards Fair Self-Supervised Learning for Speech

Laura Cristina Alonzo Canul[1]

[1]Laboratoire d'Informatique de Grenoble

laura.alonzo-canul@univ-grenoble-alpes.fr

1. Introduction

Similar to text-based and image-based applications powered by Artificial Intelligence (AI), recent years have seen speech-based technologies become ubiquitous, too; from the comfort of our own homes, we live in continuous interaction with smart devices that grant us access to speech-based services. Much like in text and image processing too, speech processing has benefited from the rise of self-supervised learning (SSL), a paradigm that seeks to extract general and robust representations of data without supervision. Speech systems that rely on this paradigm have systematically seen performance gains across tasks and domains thanks to an ever growing variety of network architectures, network sizes and a few large data resources. However, just as self-supervised representations learned from speech have been shown to be meaningful [1], they have also been shown to be harmful, exhibiting a wide range socioeconomic and demographic biases learned from unfiltered data when applied to end-user applications [2, 3, 4, 5].

Mostly characterized by subpar system performance, system biases against certain demographic groups have been identified for speech tasks such as speech and speaker recognition but remain to be investigated for many speech downstream tasks and especially for models learned through self-supervision. In this doctoral project, we leverage existing knowledge about these biases to figure out efficient ways to mitigate them and reduce performance disparities for demographic groups in disadvantage. We look into achieving fairness from representational point of view rather than from a purely data-centric perspective; i.e., we focus on ways we can learn better representations from relevant features rather than learning better models from even larger and higher quality datasets.

2. Related work

Recent works in speech processing have highlighted important performance disparities across demographic groups, especially for those based on ethnicity, dialect and gender. For example, in terms of dialect and race, authors in [4] studied the performance and psychological impact of a state of the art ASR system on a set of native English speakers, all of whom are of African-American ethnicity, and found out these speakers experience word error rates up to two times higher than White standard American English speakers. A different study made in a similar vein also found out ASR systems to be two to four times less capable of correctly inferring instances of habitual "be", an important morpho-syntactic feature of African American English, than instances of non habitual "be", it's counterpart in the standard version of the language [3]. In a third different study targeting dialect and gender, authors in [2], evaluated Youtube's automatic caption ASR system on five different regional dialects of English, observing overall lower performance for female speakers and speakers of highly accented variants (such as that of Scottish English) but finding the effect of gender not to be equal across dialects. Apart from these findings, the presence of socio-economic and socio-demographic biases in speech processing can be tracked down to much more than just end-system performance. As shown by [6] in their study of the VoxCeleb Speaker Recognition challenge, biases exist and can stem from every stage of a speech processing pipeline. In the work presented in this abstract, we focus on bias mitigation at the fine-tuning stage of an automatic speech recognition pipeline, and which is based on an extensive analysis of acoustic features that are similar across demographic groups. We further describe our methodology in Section 4.

3. Motivation and Research Questions

Correcting biases in order to create fairer speech models is a task that requires the conception and exploration of mitigation strategies at different levels of the development and that necessitates of more advanced techniques than simply balancing out training corpora according to textual demographic labels or over-representing groups in disadvantage, both of which have been shown to be insufficient to achieve fairness [7, 8]. So given such premise, we ask ourselves the following questions: if the information provided by demographic labels is not representative enough of the group itself to be helpful in de-biasing speech systems, *can we leverage acoustic feature information learned by self-supervised representations to build better training groups that will improve the performance of demographic groups in disadvantage?* And if so, *do self-supervised representations offer any advantage over features extracted directly from audio for building such groups?* We give answer to these questions using the methodology described below.

4. Methodology

4.1. Building feature-based groups

In the first step of our proposed mitigation pipeline, we focus on building acoustically similar training groups for fine-tuning. The main idea in this step is to build groups of data that share similar speech features, regardless of their demographic label. We take the data from the English portion of the CommonVoice 16.1 dataset [9] for this purpose, which contains human annotations for three demographic categories: accent (16 sub-categories), age (10 sub-categories) and gender (4 sub-categories). As part of this step too, we define a set of acoustic features to include in our study. We take into account that different features might have a different impact on the final performance of the system, as some might be more relevant than

Figure 1: *Our pipeline to build acoustically similar data groups from self-supervised representations for ASR.*

others depending on the downstream task at hand. With that in mind, we select a variety of features that have been shown to be relevant for speech recognition: the Log Power Spectrum (LPS), Mel-frequency Ceptral Coefficients (MFCC) and a set of features commonly referred to as "Prosody" (PROS), that include the interpolated logarithm of the fundamental frequency, voiced and unvoiced probability, zero-crossing rate and energy. We train a feature classifier based on small feed-forward network for each of these features, given the self-supervised representations of the data in CommonVoice (we select Wav2vec 2.0 [10] as our base SSL model). Once trained, the embedded representations are fed on a simple K-Means classifier to build the feature-based training groups based on the clusters.

4.2. Evaluating self-supervised ASR performance on the feature-based groups

The second step in the pipeline is using the newly discovered training groups to fine-tune the chosen encoder on ASR using standard CTC loss. We then assess the impact of using each of the features on the downstream performance of each demographic. We also assess the added utility of self-supervised representations over using the standard features extracted directly from audio to build the training groups and compare downstream performances.

5. Discussion and Future Work

Once all results on the baseline are available, a natural extension of this work will be to apply the same pipeline to explore the utility of different self-supervised representations from different state of the art models such as vq-wav2vec [11], HuBERT [12] and WavLM [13], in reducing performance disparities across demographic groups. Another possible continuation to this work could see the evaluation of our best performing model (i.e., the model achieving the *fairest* performance across demographics) on specific benchmarks targeting the same features of study (accented speech, prosody-related tasks). The next step in this doctoral project will leverage strong acoustic features and knowledge learned from this work to attempt bias mitigation in SSL at the pre-training stage.

6. References

[1] Y. K. Singla, J. Shah, C. Chen, and R. R. Shah, "What do audio transformers hear? probing their representations for language delivery & structure," in *2022 IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE, 2022, pp. 910–925.

[2] R. Tatman, "Gender and dialect bias in youtube's automatic captions," in *Proceedings of the first ACL workshop on ethics in natural language processing*, 2017, pp. 53–59.

[3] J. L. Martin and K. Tang, "Understanding racial disparities in automatic speech recognition: The case of habitual" be"." in *Interspeech*, 2020, pp. 626–630.

[4] Z. Mengesha, C. Heldreth, M. Lahav, J. Sublewski, and E. Tuennerman, ""i don't think these devices are very culturally sensitive."—impact of automated speech recognition errors on african americans," *Frontiers in Artificial Intelligence*, vol. 4, p. 169, 2021.

[5] C. Liu, M. Picheny, L. Sarı, P. Chitkara, A. Xiao, X. Zhang, M. Chou, A. Alvarado, C. Hazirbas, and Y. Saraf, "Towards measuring fairness in speech recognition: Casual conversations dataset transcriptions," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6162–6166.

[6] W. T. Hutiri and A. Y. Ding, "Bias in automated speaker recognition," in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022, pp. 230–247.

[7] M. Garnerin, S. Rossato, and L. Besacier, "Investigating the impact of gender representation in ASR training data: a case study on librispeech," in *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*, M. Costa-jussa, H. Gonen, C. Hardmeier, and K. Webster, Eds. Online: Association for Computational Linguistics, Aug. 2021, pp. 86–92. [Online]. Available: https://aclanthology.org/2021.gebnlp-1.10

[8] L. Maison and Y. Estève, "Some voices are too common: Building fair speech recognition systems using the common voice dataset," *arXiv preprint arXiv:2306.03773*, 2023.

[9] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," in *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, 2020, pp. 4211–4215.

[10] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 12 449–12 460. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/file/92d1e1eb1cd6f9fba3227870bb6d7f07-Paper.pdf

[11] A. Baevski, S. Schneider, and M. Auli, "vq-wav2vec: Self-supervised learning of discrete speech representations," *arXiv preprint arXiv:1910.05453*, 2019.

[12] 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.

[13] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.