Revisiting Automatic Speech Recognition for Tamil and Hindi Connected Number Recognition

Rahul Mishra*, **Senthil Raja Gunaseela Boopathy, Manikandan Ravikiran**
Shreyas Kulkarni, Mayurakshi Mukherjee, Ananth Ganesh, Kingshuk Banerjee
R&D Centre, Hitachi India Pvt Ltd, Bangalore, India
rahul.mishra@hitachi.co.in, senthil.raja@hitachi.co.in

Abstract

Automatic Speech Recognition and its applications are rising in popularity across applications with reasonable inference results. Recent state-of-the-art approaches, often employ significantly large-scale models to show high accuracy for ASR as a whole but often do not consider detailed analysis of performance across low-resource languages applications. In this preliminary work, we propose to revisit ASR in the context of Connected Number Recognition (CNR). More specifically, we (i) present a new dataset HCNR collected to understand various errors of ASR models for CNR, (ii) establish preliminary benchmark and baseline model for CNR, (iii) explore error mitigation strategies and their after-effects on CNR. In the due process, we also compare with end-to-end large scale ASR models for reference, to show its effectiveness.

1 Introduction

Automatic Speech Recognition is a wide variety with a majority of them claiming that these speech-to-text systems are able to deliver high accuracy on some of the well-established benchmarks (Prabhavalkar et al., 2023). Connected Number Recognition (CNR) is a subproblem of ASR that focuses on recognizing spoken numbers that are connected in a continuous sequence. For example, Toḷḷāyirattu aintu is a CNR speech sample representing number 905. Extracting such numbers from the speech is helpful in multiple applications (Vajpai and Bora, 2016) ranging from assisting senior citizens to make online purchases to simplification of complex banking functions. Recently there are many works that further augment ASR systems into low-resource languages such as Tamil (Diwan et al., 2021). The motivation for this research work lies in the use of ASR techniques in the sectors (like finance etc.) where the frequency of speaking connected numbers and the importance of each utterance is very high.

However many of these methods largely focus on reporting exclusively the overall Word Error Rate (WER) of the whole without discerning application-specific results and corner cases. These results are not transferable across the subset of ASR applications with different human perceptions (Kim et al., 2022). As such to advance applications of the ASR further there is a need to understand the impact of input in realistic application settings on the final ASR output across specific applications. Moreover, such analysis of errors, will in turn help introduce better post-editing mechanisms that are dynamic and selectable for specific inputs, leading to improved effectiveness of such systems. Specifically, benchmarking of CNR will economic progress by enabling technology for people across the spectrum. Thus, in this work, we analyze ASR systems for the problem of CNR in Tamil and Hindi languages.

Specifically, this work focuses on benchmarking ASR systems for CNR in Tamil, and Hindi and the impact of input data-related errors on the final performance of CNR. For the former case, we study the performance of 4 different models that currently exist for ASR. Accordingly, we find that all the existing models show significant performance degradation for CNR in Tamil and Hindi. In the latter case, we find very few works to focus on input data-related errors of ASR systems, with a majority of them concentrating on the English Language and establishing WER and few of them on other languages (Choudhary et al., 2023; Singh et al., 2020), but not from the point of CNR. Overall the contributions of this paper are as follows.

*Corresponding Author

Third Workshop on Speech and Language Technologies for Dravidian Languages,
pages 116–123, Varna, Bulgaria, Sep 7, 2023.
https://doi.org/10.26615/978-954-452-085-4_015

- We create a new CNR dataset (HCNR) for Tamil and Hindi in line with guidelines of Bakhturina et al. (2021) and present comprehensive error analysis.
- We establish preliminary baselines on HCNR with existing state-of-the-art models.
- We identify various errors and associate them with data characterstics.
- Finally, we explore some error mitigation strategies of spectral gating (Sainburg et al., 2020), spectral subtraction (Martin, 1994), speaker diarization (Bredin et al., 2020) and PESQ (Perceptual Evaluation of Speech Quality) (Rix et al., 2001) to reduce impact of few of the common errors to help researchers understand strengths and weaknesses of the developed baseline.

The rest of the paper is organized as follows. In section 2, we present the existing literature, followed by 3 showing HCNR dataset used in this work. Meanwhile in section 4, we discuss various methods, followed by results and key findings in section 5. We conclude with implications on future work in section 6.

2 Related Work

Automatic Speech Recognition systems often aim to learn end to end with output directly conditioned on raw input sample (Schneider et al., 2019). To achieve this, many works add variety of dense architectures (Povey et al., 2011), weak supervision (Radford et al., 2022) and unique components (Kaur et al., 2023). More recently there are a plethora of ASR systems for Indian languages despite low resource constraints (Gupta et al., 2023; Kumar and Mittal, 2021; Sharma et al., 2023; Madhavaraj and Ramakrishnan, 2017; Choudhary et al., 2022).

Number Recognition using speech samples, often limited to recognizing single digits with shallow analysis on few samples. Notable of these include Muhammad et al. (2009) which identifies digits spoken in Bangladesh, Alotaibi (2005) investigated the recognition of Arabic digits from the speech signals using artificial neural network and attempts of Mishra et al. (2011), Krishnamurthy and Prasanna (2017) and Patel and Patel (2017) for languages of Hindi, Malayalam, Gujarati respectively. In this work, we focus on establishing a comprehensive benchmark for connected number recognition using Povey et al. (2011) and Radford et al. (2022).

Datasets often used to train these models are trained on large, clean, and very generic. Few of the notable datasets for Indian languages include Bansal et al. (2023) for Hindi, Rakib et al. (2023) for Bengali, Manjutha et al. (2019) for Tamil, Banga et al. (2019) for emotion-based speech recognition and accented speech data by Rajaa et al. (2022). However to date, there aren't any large datasets specifically developed for connected numbers, accordingly in this work, we create new dataset catering to CNR in Tamil and Hindi.

3 Dataset

Characteristics	Values
Languages Selected	Tamil, Hindi
Number of Train Samples	56000, 35000
Number of Test Samples	8000, 5000
Sampling Rate	16 KHz
Preprocessing at collection	None
Maximum SNR	60%
Dual Talk	Yes
Background Noise	Yes
Inaudible Sound	Yes
Clipping	Yes
Repeated Numbers	Yes
Pitch Variations	Yes
Long Pauses	Yes
SNR	$\geqslant$20

Table 1: Dataset Characteristics of HCNR

The overall HCNR dataset characteristics are as shown in Table 1. Specifically, we collected datasets for two languages namely Tamil and Hindi, with the former used for the main evaluation and the latter to test the scalability of results. For Tamil, we explored various districts of Tamil Nadu, while for Hindi we collected data across Northern states of India. Moreover, each person was randomly shown a number and was asked to repeat the same as per day-to-day usage and these were recorded at the specific sampling frequency. Each of the collected samples was re-sampled at 16KHz inline with Radford et al. (2022).

The dataset was separated into the train, validation, and test splits as shown in Table 1, without any overlap between the speakers themselves. Besides, the speakers used across the sets included both male and female genders. Also, the dataset collected was made sure to include (a) Dual talk (more than one person speaking) (b) Background noise below 300 Hz (c) Inaudible sound, where

Sample Characteristics	Number of Samples	
	Tamil	Hindi
Clean	4721	4432
Dual Talk	11	254
Background Noise	116	78
Inaudible Sound	40	28
Missing Segments	52	18
Repetitions	2	5
Others	58	185

Table 2: Characteristics of HCNR from randomly drawn 5000 samples from training set.

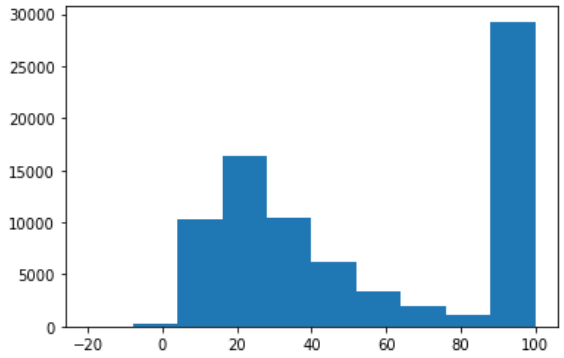

Figure 1: Histogram of SNR of Hindi data.

the quality of spoken number is poor (d) Missing digit Segments where speech doesn't include any digits (e) Clipping in Spoken Number where complete instance of spoken number is not present (f) Repetition of spoken number (g) Long pauses between spoken numbers (h) Pitch variations leading to changes of speech within a sample (See Table 2). Figures 2 and 1, show the signal-to-noise ratio (Kim and Stern, 2008) of the collected dataset. From the histograms, we can see that for the Hindi dataset, around 45% of the total speech samples are having SNR less than 40 db while for the Tamil dataset, around 57% of the total speech samples are having SNR less than 40 dB.

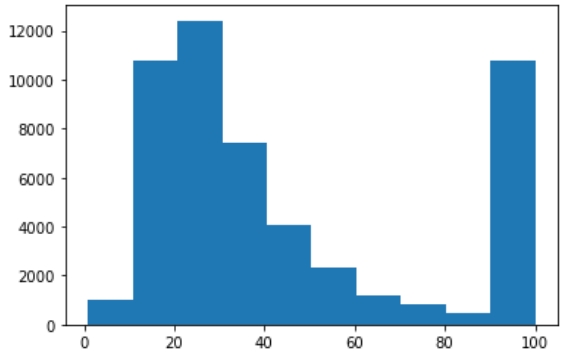

Figure 2: Histogram of SNR of Tamil data.

4 Experimental Setup

Our experimental setup is split into two parts which aim to establish a strong baseline for the problem of connected number recognition. More specifically, we assess WER for the baseline methods that are not end-to-end ASRs, rather widely used hybrid deep learning LSTM-TDNN model. Following this, we understand various errors and relate them to the characteristic of the dataset in turn highlighting the strength and weaknesses of the said baseline methods. Following this, we explore certain mitigation strategies to further ground the method so selected as a strong baseline candidate for connected number recognition. Throughout this work, we employ Word Error Rate (WER) and Sentence Error Rate (SER) inline with Klakow and Peters (2002).

4.1 Methods

Following are the various models used in this work.

- **Baseline Hybrid ASR:** In this work, we use LSTM-TDNN **(LT-Kaldi)** architecture that is part of Kaldi Speech Toolkit (Povey et al., 2011). This baseline model is composed of size convolutional layers and 15 factorized time-delay neural networks with a total of 31M parameters. We follow the standard Kaldi training recipe. The input to this model is high-resolution MFCCs with cepstral mean normalization. The **LT-Kaldi-F** model is trained for a total of 5 epochs on the training samples from Table 1. Additionally, **LT-Kaldi-P** model is trained on 50% of total training samples. This is a standard setup taken from Kaldi Speech Toolkit (Povey et al., 2011).

- **End-to-End ASR:** Though our end goal, is to establish a baseline benchmark for CNR and analyze it thoroughly, we however debate on the merits of end-to-end ASR models specifically, Wav2Vec 2.0 and Whisper. From now on we refer to these models with the following tags (i) **w2v2**: Fine-tuned `wav2vec2-large-xlsr-53` in Tamil and Hindi using the Common Voice (ii) **WH**: This model is a fine-tuned version of openai/whisper-small on the Tamil and Hindi data available from multiple publicly

available ASR corpora (For fine tuning Tamil ASR models, Tamil characters are used).

- **Error Mitigation Methods in ASR:** Additionally, we also explore (i) **P1**: Spectral Gating, (ii) **P2:** Spectral Subtraction as two measures to see effect of background noise reduction, that is part of the input, (iii) **P3:** Speaker Diarization to remove samples that consists of more than one speaker and (iv) **P4:** PESQ based score assignment to remove poor quality samples.

5 Results and Discussion

5.1 Evaluation measure

For evaluating the performance of the system, we used **World error rate (WER)**. The motivation for using WER as a performance measure comes from the type of output we are getting. As we are getting the translated text from the process of recognizing speech there is a possibility that some words may be left out or mistranslated. WER can be calculated by taking into account all of these possibilities. Mathematically, WER can be calculated as

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

Here, S is the number of substitutions, D is the number of deletions, I is the number of insertions, C is the number of correct words, N is the number of words in the reference ($N = S + D + C$). Sentence error rate (SER) is the number of incorrect sentences divided by the total number of sentences.

We structure the discussion of results by focusing on establishing the suitability of the simple hybrid method of LT-Kaldi as the baseline for the task of CNR. Although ASR-based methods are used for a subset of ASR problems, the overall results for CNR are not well-established with a significantly large dataset as approached by this work. Accordingly, in Table 3, we compare the results of LT-Kaldi across different settings mentioned earlier. From the results we can explicitly see that for both Tamil and Hindi with full data, the individual word error are 15% and 7% respectively, indicating the simple methods indeed show strong performance on the overall dataset with a variety of characteristics. SER depends on the correctness of each word in a complete sentence. If there is a prediction error only in one word of a full sentence, it will make the prediction of the entire sentence wrong. That's why the SER is relatively higher than the WER.

Table 4 shows error statistics of LT-Kaldi in the test set with a breakdown across sample characteristics. Meanwhile, Table 5, shows example predictions and errors in Tamil (Transliterated).

	WER (%)		SER (%)	
Method	**Tamil**	**Hindi**	**Tamil**	**Hindi**
LT-Kaldi	15.11	7.63	25.64	15.40

Table 3: Baseline results on `HCNR` across different methods.

Sample Characteristics	**Erroneous Samples**	
	Tamil	**Hindi**
Dual Talk	52	329
Background Noise	514	328
Inaudible Sound	273	138
Missing Segments	160	46
Repetitions	33	8
Others	216	241

Table 4: Error statistics across languages with LT-Kaldi on test set.

From the table 4, we can see that for Tamil, the overall WER is majorly dominated by samples with background noise, repetition of spoken numbers, and Inaudible sound respectively. Meanwhile, in the case of Hindi, the resulting errors are heavily concentrated in background noise, dual talk, and repetition of spoken numbers. For this analysis, we considered all samples of the test set for both Hindi as well as Tamil language. Further contrasting the two languages, one would argue that background noise and dual talk are vital to be handled in the problem of CNR, followed by repetition of spoken number and Inaudible sound respectively. Thus, the languages despite being different the model shows common behavior across its errors, indicating its potential generalization.

Meanwhile, to further verify the effectiveness of LT-Kaldi, we subject the samples of the test set to noise removal using P1 and P2 respectively. For P1, we compare the spectrogram of the input speech and estimate a noise threshold (SNR Threshold) to gate out the unnecessary signals. In this work, we test with three different SNR threshold values namely `None`, `<15`, `<30` respectively. Meanwhile, for P2, we subtract the current speech

spectrum with noise to estimate a clean signal. The result with these methods is as shown in Table 6 and 7 respectively. From the tables, we can see that with both spectral gating and spectral subtraction on the overall data, there is indeed a negative impact on the overall results across both languages. This is because of models like LT-Kaldi tempo-spectral properties of any type of speech and noise and adding noise removal method indeed distorts speech samples and in turn effectively removes useful parts. While noise removal using spectral gating had a negative impact, we argue that the method is crude in removing noise and rather verify the same using spectral subtraction to obtain results as shown in Table 7. From the results, we can see that indeed removing noise improves the results of Tamil CNR by 1% with a still negative impact on the Hindi language. The potential reason behind this may be the degradation of signal power with respect to the noise power. In few samples (% of total samples) the SNR is low, which essentially signifies that the signal power is nearly equal to the noise power. In these cases the application of noise cancellation techniques may also result in the degradation of necessary signal information. However, more investigation is required in this regard.

Meanwhile, we also argue that the removal of dual talk would improve the overall results due to inherent distortion created by the voices of multiple people. Accordingly, we employ works of Bredin et al. (2020) where we remove samples that have more than one identified speaker and accordingly obtain results as shown in Table 8. Besides combining all the pre-processing methods shows additional improvement as shown in Table 9.

Apart from the speaker diarization to remove samples, we also employ PESQ based technique to assign score to each speech sample. Its value lies in between -0.5 to 4.5. A higher score indicates a better signal quality. If the score is greater than a threshold, it means that the quality of the speech sample is good and we can consider that sample for further proceedings. This score has been calculated between the raw speech sample and its processed version (speech signal after passing through the spectral subtraction based noise cancellation pipeline). For more details refer to this work. For getting the threshold, we create the histogram of scores of all speech samples of Hindi and Tamil languages (Fig. 3 and Fig. 4). After manually visualizing the histograms, we decided to keep 2.3 and 1.9 as the threshold for the Hindi and Tamil speech samples respectively. However, more experiments can be done to get a more robust threshold value.

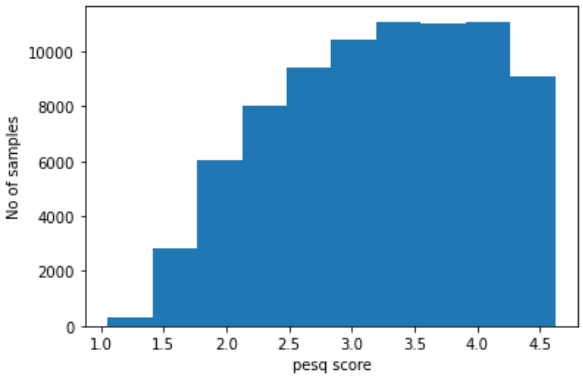

Figure 3: Histogram of PESQ score Hindi data.

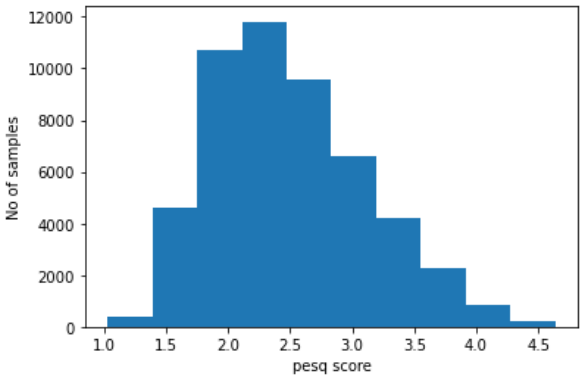

Figure 4: Histogram of PESQ score of Tamil data.

From the result again we can see net effectiveness restricted to only around 1% indicating the effectiveness of preprocessing methods on the problem of CNR is not high, warranting more study in the training process and sample processing. Overall from results across Table 3-9, we can conclude that LT-Kaldi is a descent baseline for CNR with various preprocessing methods having a negligible effect on the results. To further, establish the effectiveness of the results of LT-Kaldi, we compare the results of LT-Kaldi with Wav2vec2 and whisper respectively. To this end, we compare the results of LT-Kaldi trained with `HCNR` against pretrained models Wav2vec2 and whisper in Table 10. The poor performance is because the models are trained on general speech data. It also signifies that if we want to use the model for utilizing connected numbers for any specific task, we need to finetune the publicly available SOTA models on

Input Tamil Sample	Prediction	
	LT-Kaldi	WH
Muppattu mūu āyirattu eṇṇūu aupattiraṇṭu	mūu āyirattu muppattu mūu āyirattu eṇṇūu aupattiraṇṭu	nukarppu eṇṇūi aupatti iraṇṭu
Toḷḷāyirattu aintu	eupattaintu	vārttai
Ainūu irupattou	eṭṭu	ea
Muppattu nāku āyirattu eṇpattu mūu	nāku āyirattu eṇpattu mūu	tōṭṭi nāku āyirattu topatti mūu
Ōr āyirattu nūi nāpattu mūu	ōr āyirattu nāpattu mūu āyirattu nāpattu mūu	āyiratti nāppatti mūṇu āyiratti nāppatti mūṇu
Toḷḷāyirattu eṇpattu nāku	toūu eṇpattaintu	coatu

Table 5: Example Errors from Tamil with LT-Kaldi and WH models

Method	SNR Threshold	WER (%)		SER (%)	
		Tamil	Hindi	Tamil	Hindi
LT-Kaldi + P1	None	15.34	8.23	29.23	16.84
LT-Kaldi + P1	<15	15.52	9.40	30.17	16.70
LT-Kaldi + P1	<30	15.81	9.09	30.54	17.09

Table 6: Results on HCNR for LT-Kaldi with spectral gating.

Method	SNR Thresholding	WER (%)		SER (%)	
		Tamil	Hindi	Tamil	Hindi
LT-Kaldi + P2	None	14.37	9.41	26.35	18.09

Table 7: Results on HCNR for LT-Kaldi with spectral subtraction.

	WER (%)		SER (%)	
Method	**Tamil**	**Hindi**	**Tamil**	**Hindi**
LT-Kaldi + Diarization	14.03	6.35	26.31	12.40

Table 8: Results on HCNR across with LT-Kaldi and Diarization

	WER (%)		SER (%)	
Method	**Tamil**	**Hindi**	**Tamil**	**Hindi**
LT-Kaldi + P2 + P3	13.57	7.85	26.35	15.62
LT-Kaldi + P2 + P4	13.72	8.49	27.21	16.52

Table 9: Results on HCNR across with LT-Kaldi with P2, P3 and P4

the specific dataset.

From the results it is evident both the models are directly not suitable for CNR with high WER, indeed indicating that simply trained LT-Kaldi is a more suitable baseline method. Table 5, shows various errors obtained using **WH** on Tamil language. From the results it is evident that E2E models indeed being unable to understand the spoken language, indicating the need for domain adaptation. This is in contrast with other applications of ASR where E2E show high results.

	WER (%)		SER (%)	
Method	**Tamil**	**Hindi**	**Tamil**	**Hindi**
w2v2	98.63	71.08	99.12	87.07
WH	93.37	85.70	97.80	93.70

Table 10: Results on HCNR across Wav2vec2 and Whisper models.

6 Conclusion and Future Work

Overall in this work, we study the problem of CNR by creating a new HCNR dataset and report baseline results with the LT-Kaldi model across Tamil and Hindi languages. In the process, we find that the baseline LT-Kaldi shows WER of around 15% and 7% respectively across the languages. In the due process, we conjectured the sample characteristics might be the key reason leading to higher WER through analysis, for which we studied spectral gating, spectral subtraction, and diarization methods for further improvement. However, we could see that the overall results improved only by 2% for Tamil and 1% for Hindi CNR. Most importantly, we could also see that compared to LT-Kaldi, the pretrained models performed significantly worse, unlike prior works. However, we think this may be attributed to the case of out-of-domain samples, needing further studies. In this regard, a possible question to explore includes evaluating the effect of training the E2E model with HCNR and the effect of in-domain data on CNR performance. Additionally, we plan to explore other methods to remove negative sample characteristics and study their impact on overall CNR results.

Acknowledgements

We thank our anonymous reviewers for their valuable feedback.

References

Yousef Ajami Alotaibi. 2005. Investigating spoken arabic digits in speech recognition setting. *Information Sciences*, 173(1-3):115–139.

Evelina Bakhturina, Vitaly Lavrukhin, and Boris Ginsburg. 2021. A toolbox for construction and analysis of speech datasets. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.

Subham Banga, Ujjwal Upadhyay, Piyush Agarwal, Aniket Sharma, and Prerana Mukherjee. 2019. Indian emospeech command dataset: A dataset for emotion based speech recognition in the wild. *CoRR*, abs/1910.13801.

Vansh Bansal, T. Thishyan Raj, Nagarathna Ravi, Shubham Korde, Jaskaran Kalra, Sudha Murugesan, B. Ramkrishnan, Aboli Gore, and Vipul Arora. 2023. Parturition hindi speech dataset for automatic speech recognition. In *28th National Conference on Communications, NCC 2023, Guwahati, India, February 23-26, 2023*, pages 1–6. IEEE.

Hervé Bredin, Ruiqing Yin, Juan Manuel Coria, Gregory Gelly, Pavel Korshunov, Marvin Lavechin, Diego Fustes, Hadrien Titeux, Wassim Bouaziz, and Marie-Philippe Gill. 2020. Pyannote. audio: neural building blocks for speaker diarization. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7124–7128. IEEE.

Tripti Choudhary, Atul Bansal, and Vishal Goyal. 2022. Investigation of cnn-based acoustic modeling for continuous hindi speech recognition. In *IoT and Analytics for Sensor Networks: Proceedings of ICWSNUCA 2021*, pages 425–431. Springer.

Tripti Choudhary, Vishal Goyal, and Atul Bansal. 2023. WTASR: wavelet transformer for automatic speech recognition of indian languages. *Big Data Min. Anal.*, 6(1):85–91.

Anuj Diwan, Rakesh Vaideeswaran, Sanket Shah, Ankita Singh, Srinivasa Raghavan, Shreya Khare, Vinit Unni, Saurabh Vyas, Akash Rajpuria, Chiranjeevi Yarra, Ashish R. Mittal, Prasanta Kumar Ghosh, Preethi Jyothi, Kalika Bali, Vivek Seshadri, Sunayana Sitaram, Samarth Bharadwaj, Jai Nanavati, Raoul Nanavati, Karthik Sankaranarayanan, Tejaswi Seeram, and Basil Abraham. 2021. Multilingual and code-switching asr challenges for low resource indian languages. *ArXiv*, abs/2104.00235.

Astha Gupta, Rakesh Kumar, and Yogesh Kumar. 2023. An automatic speech recognition system in indian and foreign languages: A state-of-the-art review analysis. *Intell. Decis. Technol.*, 17(2):505–526.

Amrit Preet Kaur, Amitoj Singh, Rohit Sachdeva, and Vinay Kukreja. 2023. Automatic speech recognition systems: A survey of discriminative techniques. *Multim. Tools Appl.*, 82(9):13307–13339.

Chanwoo Kim and Richard M Stern. 2008. Robust signal-to-noise ratio estimation based on waveform amplitude distribution analysis. In *Ninth Annual Conference of the International Speech Communication Association*. Citeseer.

Suyoun Kim, Duc Le, Weiyi Zheng, Tarun Singh, Abhinav Arora, Xiaoyu Zhai, Christian Fuegen, Ozlem Kalinli, and Michael L. Seltzer. 2022. Evaluating user perception of speech recognition system quality with semantic distance metric. In *Interspeech 2022, 23rd Annual Conference of the International Speech Communication Association, Incheon, Korea, 18-22 September 2022*, pages 3978–3982. ISCA.

Dietrich Klakow and Jochen Peters. 2002. Testing the correlation of word error rate and perplexity. *Speech Commun.*, 38(1-2):19–28.

S. Krishnamurthy and S. R. Mahadeva Prasanna. 2017. A hybrid feature extraction technique for continuous number speech recognition in malayalam. In *Proceedings of the 24th International Conference on Machine Learning*, pages 33–40.

Ashok Kumar and Vikas Mittal. 2021. Hindi speech recognition in noisy environment using hybrid technique. *International Journal of Information Technology*, 13:483–492.

A Madhavaraj and AG Ramakrishnan. 2017. Design and development of a large vocabulary, continuous speech recognition system for tamil. In *2017 14th IEEE India Council International Conference (INDICON)*, pages 1–5. IEEE.

Manavalan Manjutha, Parthasarathy Subashini, Marimuthu Krishnaveni, and V. Narmadha. 2019. An optimized cepstral feature selection method for dysfluencies classification using tamil speech dataset. In *2019 IEEE International Smart Cities Conference, ISC2 2019, Casablanca, Morocco, October 14-17, 2019*, pages 671–677. IEEE.

Rainer Martin. 1994. Spectral subtraction based on minimum statistics. *power*, 6(8):1182–1185.

AN Mishra, Mahesh Chandra, Astik Biswas, and SN Sharan. 2011. Robust features for connected hindi digits recognition. *International Journal of Signal Processing, Image Processing and Pattern Recognition*, 4(2):79–90.

Ghulam Muhammad, Yousef A Alotaibi, and Mohammad Nurul Huda. 2009. Automatic speech recognition for bangla digits. In *2009 12th international conference on computers and information technology*, pages 379–383. IEEE.

P. Patel and P. Patel. 2017. A comparative study of continuous digit recognition using mfcc and lpc features for gujarati language. In *Proceedings of the 24th International Conference on Machine Learning*, pages 33–40.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Kumar Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely. 2011. The kaldi speech recognition toolkit. IEEE Signal Processing Society.

Rohit Prabhavalkar, Takaaki Hori, Tara N. Sainath, Ralf Schluter, and Shinji Watanabe. 2023. End-to-end speech recognition: A survey. *ArXiv*, abs/2303.03329.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *CoRR*, abs/2212.04356.

Shangeth Rajaa, Swaraj Dalmia, and Kumarmanas Nethil. 2022. Skit-s2i: An indian accented speech to intent dataset. *CoRR*, abs/2212.13015.

Fazle Rabbi Rakib, Souhardya Saha Dip, Samiul Alam, Nazia Tasnim, Md. Istiak Hossain Shihab, Md. Nazmuddoha Ansary, Syed Mobassir Hossen, Marsia Haque Meghla, Mamunur Mamun, Farig Sadeque, Sayma Sultana Chowdhury, Tahsin Reasat, Asif Shahriyar Sushmit, and Ahmed Imtiaz Humayun. 2023. Ood-speech: A large bengali speech recognition dataset for out-of-distribution benchmarking. *CoRR*, abs/2305.09688.

Antony W Rix, John G Beerends, Michael P Hollier, and Andries P Hekstra. 2001. Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. In *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221)*, volume 2, pages 749–752. IEEE.

Tim Sainburg, Marvin Thielk, and Timothy Q Gentner. 2020. Finding, visualizing, and quantifying latent structure across diverse animal vocal repertoires. *PLoS computational biology*, 16(10):e1008228.

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. 2019. wav2vec: Unsupervised pre-training for speech recognition. *CoRR*, abs/1904.05862.

Usha Sharma, Hari Om, and AN Mishra. 2023. Hindispeech-net: a deep learning based robust automatic speech recognition system for hindi language. *Multimedia Tools and Applications*, 82(11):16173–16193.

Amitoj Singh, Virender Kadyan, Munish Kumar, and Nancy Bassan. 2020. Asroil: a comprehensive survey for automatic speech recognition of indian languages. *Artif. Intell. Rev.*, 53(5):3673–3704.

Jayashri Vajpai and Avnish Bora. 2016. Industrial applications of automatic speech recognition systems.