

Automatic Language Identification for India's Low-Resource and Regional Languages: A Review with a Case Study on Hindi's Native Language Family

Jogender Singh

Research Scholar, Department of Computer Science and Engineering, University Institute of Engineering and Technology, Maharshi Dayanand University, Rohtak, Haryana, India

Dr. Yudhvir Singh

Professor, Department of Computer Science and Engineering, University Institute of Engineering and Technology, Maharshi Dayanand University, Rohtak, Haryana, India

Abstract- Automatic Language Identification (LID) is the task of determining which language is being spoken in an audio signal, without reference to what is actually said. The field has advanced considerably over the past three decades. It has moved from classical statistical and phonotactic approaches, through factor-analysis techniques, to the transformer-based and knowledge-distillation architectures that define current practice. Most of this progress, however, has concentrated on a small set of international and major Indian languages. India's regional and low-resource languages, Angami, Mizo, Assamese, Bengali, and the large family of native languages spoken under the broader umbrella of Hindi among them, remain comparatively neglected, despite combined speaker populations running into the hundreds of millions. This paper reviews the LID literature from tracing the field's evolution from Zissman's foundational 1996 comparison through i-vector and deep neural approaches to the attention-based and knowledge-distilled architectures used today and then surveys the smaller but growing body of work addressing Indian regional and low-resource languages specifically, drawing out the recurring methodological lessons these efforts share, particularly around corpus scarcity and acoustic similarity between closely related languages. The review gives particular attention on the native and dialectal languages spoken under Hindi, several with populations exceeding fifty million, which remain almost entirely absent from the LID literature despite their scale. The paper closes by identifying open challenges and directions for future research across India's low-resource language landscape more broadly.

Keywords: *Language Identification, Low-Resource Languages, Indian Regional Languages, Native Languages of Hindi, Speech Corpus, Deep Learning, Knowledge Distillation, Survey*

1. Introduction

People tell languages apart before they can speak any of them. In the first few months of life, an infant already recognizes the rhythm and tune of its native language against a new one, without knowing a single word of either [1]. Adults do something similar every day. A person who steps off a train in an unfamiliar place can often tell within seconds whether the language around them is one they know, one close to their own, or something entirely foreign. This judgement takes almost no conscious effort. It happens long before anyone tries to work out what is actually being said.

Machines do not have built in ability to make LID. A computer given an audio recording has no inherent sense of language. It sees only a sequence of numbers representing sound pressure over time. Machine recognize which language is being spoken, purely from the acoustic signal and without relying on subtitles or transcripts, is the task known as Automatic Language Identification (LID). In practice, LID has proven to be one of the more persistently difficult problems in speech processing.

Speech-based technology has moved from research laboratories to everyday devices: phones, call centres, smart speakers, security systems. Alongside that move has come a practical need to know which language a speaker is using, automatically and before any further processing takes place. A voice assistant cannot select the correct speech recognizer without first knowing the language being spoken. A multilingual call centre cannot route a call to the right operator without first identifying the caller's language. A surveillance or forensic system reviewing a large volume of recorded speech cannot triage that material efficiently without an automatic means of sorting it by language. In each of these cases, language identification acts as a gatekeeper. It is the first decision, and mistakes made here carry forward through everything that follows.

LID has advanced significantly for global languages such as English, Mandarin and the major European languages. Progress for India's languages has been considerably more uneven. Several of the country's officially scheduled languages have received reasonable research attention. Its many regional and low-resource languages have not, and this unevenness runs deep. Angami and Mizo, spoken by populations in the low hundreds of thousands, have only recently gained dedicated speech corpora [2, 3]. Assamese and Bengali have somewhat better coverage but still lag far behind English or Mandarin [4]. Most strikingly, the large family of native and dialectal languages spoken under the broader umbrella of Hindi has received almost no dedicated attention at all, despite carrying a combined speaker population running into the hundreds of millions. What the Census of India records as "Hindi" is, in reality, a large family of native languages. Bhojpuri, Rajasthani, Chhattisgarhi, Magahi, Haryanvi, Marwari, Malvi, Awadhi and Pahari are among the largest, and dozens more exist beyond them. Each has its own phonology, vocabulary and speaker community, and several carry populations exceeding fifty million [5]. This gap exists largely because of their administrative subsumption under a single Hindi category, rather than any lack of linguistic distinctiveness or acoustic difficulty.

This paper provides a structured review of the LID literature with two goals. First, it traces the methodological evolution of the field over roughly three decades, identifying which techniques have proven durable and why. Second, it surveys the body of work addressing India's regional and low-resource languages specifically, drawing out the shared methodological lessons, around corpus scarcity in particular, that recur across these otherwise quite different languages. The remainder of this paper is organized as follows. Section 2 reviews the general LID pipeline and the metrics used to evaluate LID systems. Section 3 traces the methodological history of the field. Section 4 reviews literature specific to India's regional and low-resource languages. Section 5 examines the specific case of Hindi's native languages in more depth. Section 7 discusses open challenges and future directions across India's low-resource language landscape more broadly. Section 8 concludes.

2. The General LID Pipeline and Evaluation Metrics

2.1 Pipeline Overview

Most LID systems, whether explicit (relying on recognized phonetic units) or implicit (working directly with acoustic features), are organized around a common pipeline. This pipeline comprises preprocessing, feature extraction, feature normalization, language modeling, classification, and score-based identification [6]. This structure holds despite the wide range of specific techniques found across the literature.

Preprocessing converts raw audio into a consistent digital format that machine can understand. Then it resamples to a common rate, reduces background noise, and removes silence or non-speech segments. Feature extraction then divides the cleaned signal into short frames, typically around 25 milliseconds, and computes acoustic features for each frame. Mel Frequency Cepstral Coefficients (MFCC) remain the most widely used representation. They are computed via a mel-scale filter bank followed by a discrete cosine transform [7]. Alternatives include Linear Predictive Coding (LPC) coefficients, which model the vocal tract as a linear filter, and their cepstral variant (LPCC). Log Power Spectrum (LPS) features form a further alternative, retaining more raw spectral detail at the cost of higher dimensionality.

Feature normalization reduces channel and speaker-specific variation unrelated to language. Common techniques include Cepstral Mean Subtraction, RASTA filtering, and Gaussian Mixture Model (GMM) based normalization. Language modeling then builds a statistical representation of each candidate language, historically through GMMs and later through i-vector representations or deep neural embeddings. Classification compares a test utterance's representation against each candidate language model. A final scoring and calibration step produces the identification decision.

2.2 Evaluation Metrics

When we evaluating an LID system requires attention to the type of errors made, not only overall accuracy. Classification accuracy is the proportion of correctly identified utterances. It is intuitive but says little about how errors are distributed across languages or language pairs. The Equal Error Rate (EER) addresses part of this limitation. It captures the trade-off between false acceptance and false rejection as the decision threshold varies, and it is defined as the point where these two error rates become equal. The average detection cost (Cavg) extends this idea further. It weights different error types by a cost function and averages the result across all languages and confusable pairs, and it sits at the centre of NIST's Language Recognition Evaluations [8]. Confusion matrices offer a more granular view still. By tabulating how often each true language is misidentified as each other candidate language, they help diagnose which specific language pairs prove hardest to separate.

3. Methodological Evolution of the LID Field

3.1 Early Statistical and Phonotactic Approaches

Much of the modern LID literature traces its methodological roots to Zissman's 1996 comparison of four approaches on the OGI Multi-Language Telephone Speech corpus. These were GMM classification, single phone recognition followed by language modeling (PRLM), parallel PRLM, and language-dependent parallel phone recognition [1]. GMM classification trained fastest. Parallel PRLM delivered the strongest performance, however, and it did not require a labelled phonetic corpus for every target language, since it

reused phone recognizers trained on whichever languages had suitable resources available. On an eleven-language closed set, the system achieved eleven percent error. This fell to two percent for a two-language task. The pattern, degrading performance with increasing language count and acoustic similarity, recurs throughout the literature reviewed below.

The GMM tradition underlying much of this early work originates in speaker recognition rather than LID specifically. Reynolds and Rose's 1995 paper on Gaussian mixture speaker models established GMM-based acoustic modeling as a standard baseline. It was valued for requiring no explicit segmentation of speech into words or phones [9]. This shared ancestry explains why techniques such as i-vectors, Joint Factor Analysis, and Probabilistic Linear Discriminant Analysis appear in both the speaker recognition and LID literatures, as discussed below.

3.2 Feature Engineering and Indian Language Applications

Koolagudi, Rastogi and Rao's 2012 study represents one of the earliest serious applications of LID methodology to Indian languages specifically. It used MFCC features with a GMM classifier across fifteen languages, including Rajasthani, treated as distinct from Hindi rather than subsumed under it [10]. This was an early acknowledgment that some of what census data groups under Hindi carries sufficient acoustic distinctiveness for separate identification. The system achieved eighty-eight percent accuracy, with twenty-nine spectral features proving most effective. This remains, to the authors' knowledge, the only prior study to treat any native language of Hindi as a distinct LID target. It did not extend this treatment to Bhojpuri, Chhattisgarhi, Magahi, Haryanvi or the many other native languages this review is centrally concerned with.

3.3 Factor Analysis and i-vector Methods

The i-vector framework addresses a central limitation of per-language GMM training. With limited training data per language, estimating a full set of GMM parameters independently for each language becomes unreliable. i-vector approaches instead train a single Universal Background Model across all languages jointly, representing each utterance as a low-dimensional vector describing how its statistics deviate from this shared model. Lopez-Moreno et al.'s 2014 study marked a turning point. It demonstrated that deep neural networks could outperform this i-vector framework, with an eight-hidden-layer DNN achieving roughly seventy percent improvement in Cavg over the strongest i-vector baseline on the Google 5M LID corpus and NIST LRE 2009 [11]. Grisard et al.'s 2019 study pushed the feature representation itself further. Bottleneck features extracted from a neural network trained on a related task substantially outperformed conventional MFCC within an i-vector pipeline, reducing Cavg from 2.40 to 0.90 on a four-language European task evaluated across mixed fixed and mobile telephony channels [12].

3.4 Deep Learning Architectures

The deep learning era has produced several architectural families relevant to closely related, low-resource language identification. Geng et al.'s 2016 attention-based recurrent network combined an LSTM encoder with soft and hard attention mechanisms. It achieved a 14.72 percent EER on NIST LRE 2007, a marked improvement over a 20.39 percent i-vector baseline, by allowing the model to weight informative frames more heavily than uninformative ones [13]. This capacity to extract more discriminative value from limited acoustic evidence matters particularly for short-duration identification. Most of the literature does not treat this condition as a primary evaluation criterion, despite its practical importance.

Bartz et al.'s 2017 hybrid Convolutional Recurrent Neural Network treated spectrograms as images. It achieved 92 percent accuracy across six languages, rising to 96 percent when narrowed to the acoustically distant Mandarin-English pair [14]. This same pattern, stronger performance on acoustically distant pairs than on closely related ones, is confirmed repeatedly across the literature. Heracleous et al.'s comparison of SVM, CNN and DNN classifiers found EER rising from roughly two percent on a ten-language task to over three percent on a fifty-language task [15]. Mukherjee et al.'s CNN-on-spectrogram system achieved a near-perfect 99.96 percent on a controlled seven-language, vocabulary-restricted task, falling to 95.52 percent on a broader eleven-language condition [16]. Vuddagiri et al.'s multi-head self-attention network took a different approach to the variable-length input problem, converting sequences of arbitrary length into a fixed-dimension representation and achieving 6.44 percent EER on the ten-language AP17-OLR challenge [17].

NIST's own evaluation design reflects this same pattern. Its 2017 Language Recognition Evaluation deliberately targeted fourteen closely related languages across five clusters, a harder version of the problem than the more acoustically diverse language sets used in earlier evaluations. This choice recognized explicitly that within-family discrimination among closely related languages, rather than sheer language count, represents the field's hardest unsolved challenge [8].

4. Indian Regional and Low-Resource Language LID Literature

4.1 Language Identification Studies

Beyond Koolagudi et al., discussed above, a small but growing body of work addresses language identification for Indian languages specifically. Chakraborty, Nath, Nirmala and Samudravijaya's 2018 study built a Kaldi-toolkit-based system combining Hidden Markov Models with Gaussian Mixture Models and deep neural networks. It identified Assamese, Bengali and English speech read by multiple speakers over mobile phones, achieving 99.3 percent accuracy on test data [4]. English is acoustically far more distant from Assamese and Bengali than these two closely related eastern Indo-Aryan languages are from each other. This strong aggregate figure is therefore best read as reflecting the comparatively easy Assamese/Bengali-versus-English distinction, rather than necessarily reflecting equally strong separation between Assamese and Bengali themselves. This is consistent with the acoustic-distance pattern established throughout this review.

Two studies address corpus development for India's smaller, under-resourced languages directly, with language identification or recognition as an explicit downstream goal. Terhija, Sarmah and Vijaya built a speech corpus and automatic speech recognition system for Angami, a Tibeto-Burman language of Nagaland with a speaker population in the low hundreds of thousands [2]. Dey et al. built a phone recognition system for Mizo, a related language of comparable scale. They collected passages from 81 native speakers and identified 36 phonetic units for the language [3]. Both studies demonstrate that dedicated, ground-up corpus development is achievable for a previously unaddressed Indian language, a methodological precedent directly relevant to the corpus development summarized in Section 6 below, even though both languages are far smaller in speaker population than any native language of Hindi.

4.2 Early and Mid-Scale Indian Speech Corpora

Beyond studies built explicitly around language identification, a considerably larger body of work has built speech corpora for Indian languages in support of speech recognition, speaker recognition, and text-to-

speech synthesis. Shrishrimal, Deshmukh and Waghmare's 2012 survey remains the most comprehensive account of this earlier period, cataloguing thirteen Indian speech databases built during the 2000s and early 2010s [18]. Their survey is worth summarizing by name, since these databases collectively illustrate the single-language, single-application pattern that characterized Indian speech technology research before large multilingual efforts such as VAANI emerged.

Several databases targeted a specific application in a specific language. TIFR Mumbai and IIT Bombay jointly built a Marathi agricultural information corpus from roughly 1,500 speakers recorded over telephone lines. C-DAC Noida built a Hindi travel-domain corpus from 30 female speakers, recording around 26 hours of speech across 8,567 sentences. IIIT Hyderabad built a Telugu commodity-price-enquiry corpus from 96 speakers recorded on mobile phones in noisy conditions, and separately built a Telugu-Hindi-English travel and emergency-services corpus from 15 speakers under studio conditions. Government P.G. College, Rishikesh built a Garhwali corpus from 100 speakers across five districts of Uttarakhand, targeting automatic speech recognition for a language with no prior dedicated resource.

Other databases targeted text-to-speech synthesis specifically, generally using syllables or phonemes as the basic recording unit rather than full sentences. Punjabi University, Patiala built a Punjabi database of 3,312 syllables, selected to cover over 99 percent of syllable frequency in the language. Utkal University, Bhubaneswar built a four-language database spanning Hindi, Odia, Bengali and Telugu using the same syllable-concatenation approach. Rajarambapu Institute of Technology built a Konkani database of roughly 3,000 recordings covering vowels, characters and half-characters. A separate effort at SJCE Mysore built a Kannada phoneme database of 1,605 recordings.

A smaller number of databases pursued general-purpose coverage across multiple languages simultaneously. TIFR Mumbai and C-DAC Noida jointly built a general-purpose Hindi corpus from 100 native speakers, designed to capture inter- and intra-speaker phonetic variability rather than any single application. IIT Kharagpur built a four-language broadcast-news corpus spanning Hindi, Telugu, Tamil and Kannada, totalling 17.5 hours across 92 speakers, intended to support prosody modelling for recognition, synthesis, and language identification jointly. KIIT Bhubaneswar, sponsored by Nokia Research Centre, built a mobile-recorded Hindi and Indian-English corpus from 200 speakers using three simultaneous recording channels. IIIT Hyderabad, working with HP Labs Bangalore, built the largest of the thirteen databases surveyed, a 559-speaker corpus spanning Marathi, Tamil and Telugu, recorded over mobile phones and landlines.

Shrishrimal et al.'s comparison across these thirteen databases found that only five were recorded under genuinely noisy, real-world conditions, with the remaining eight recorded in noise-free or studio environments, and that the great majority were built for text-to-speech synthesis using phonemes or syllables rather than continuous speech suitable for recognition tasks. Only the Linguistic Data Consortium for Indian Languages (LDC-IL), which by the time of that survey had assembled speech corpora exceeding 100 hours each for eleven languages, Assamese, Bengali, Bodo, Gujarati, Hindi, Kannada, Konkani, Malayalam, Manipuri, Marathi and Nepali, approached the scale of comparable international speech resources, though even there coverage across India's full language inventory remained uneven [18].

A separate landscape study commissioned by GIZ under the FAIR Forward initiative reached a similar conclusion roughly a decade later, finding that open, freely licensed voice data for Indian languages remained scarce and fragmented even among the languages that did have some existing coverage [19]. Read alongside the corpus efforts catalogued above, this persistence of fragmentation across more than a decade

of Indian speech technology research reinforces the same underlying pattern: the barrier facing India's low-resource languages has been the absence of coordinated, adequately scaled data collection, not any inherent technical obstacle to building one once the data exists.

4.3 Recent Large-Scale Multilingual Indic Speech Datasets

A newer generation of substantially larger, multilingual Indic speech datasets has emerged more recently, most collected specifically to support deep learning models that require considerably more training data than the corpora surveyed in Section 4.2 could provide. Table 1 summarizes the major datasets in this generation.

Table 1: Recent large-scale Indic speech datasets

Dataset	Collection method	Duration (hours)	Languages	Speakers
IndicVoice	Mixed	23,700	22	51,000
RESPIN	Read	10,416	9	18,000+
Shrutilipi	Read	6,457	12	Not reported
Kathbath	Read	1,684	12	1,217
Spring-INX	Mixed	2,005	10	7,609
SYSPIN	Read	900	9	18
NPTEL	Read	857	8	Not reported
MSR Challenge	Mixed	150	3	1,286
Gram Vaani	Spontaneous	1,108	1	Not reported
IISc-MILE	Read	497	2	1,446
MUCS	Mixed	600	7	310
IIITH-ILSC	Mixed	103.5	23	1,150
CommonVoice (Indic subset)	Read	373	10	Not reported
IndicTTS	Read	225	13	26
FLEURS (Indic subset)	Read	163	13	39
NISP	Mixed	56.9	6	345
E&NE Languages	Mixed	19.75	4	Not reported
CMS	Read	35	6	243
IndicSpeech	Read	24	3	3
IIITH-ISD	Read	11	7	35

Among the twenty prior datasets listed, none covers more than 23 languages, and IndicVoice, the largest by duration among them, covers 22, the same number as the languages officially recognized under the Eighth Schedule of the Indian Constitution, though this review has not verified whether the two lists coincide exactly. What is clear is that the great majority of these datasets, like the LDC-IL holdings discussed in Section 4.2, were built primarily around India's officially recognized languages, leaving the

remainder of the country's roughly 1,369 recorded mother tongues, including every native language of Hindi discussed in Section 5, without comparable representation in any of the resources listed in Table 1.

4.4 Native Language Identification from Text

A further, related but methodologically distinct strand of work addresses Native Language Identification (NLI) from written text rather than speech. This task infers a writer's native language from patterns in their use of a second language. Markov and Sidorov's contribution to the 2018 Indian NLI shared task exemplifies this strand [20]. Despite the similarity in name, text-based NLI is a fundamentally different task from the acoustic, speech-based LID this review is concerned with. It works from orthographic and lexical patterns in written text rather than acoustic signal, and the underlying techniques, drawn from text classification, do not transfer directly to the speech-based problem this paper addresses.

4.5 A Common Pattern Across Three Generations of Indian Speech Resources

Read together, the studies discussed across Sections 4.1 through 4.3 span three generations of Indian speech resource development, and they share a common shape despite addressing quite different languages and quite different scales. The earliest generation, catalogued by Shrishrimal et al., built small, single-application, single-language corpora, each rarely exceeding a few hundred speakers. A second generation, exemplified by Angami and Mizo, applied the same ground-up, dedicated-corpus philosophy to specific under-resourced languages outside the officially scheduled set, at a similarly modest scale. A third, current generation, exemplified by the datasets in Table 1 and by VAANI specifically, pursues multilingual coverage at a scale of tens of thousands of hours, but even this generation, prior to VAANI, remained anchored to India's twenty-two scheduled languages almost without exception. Across all three generations, the same lesson holds: identification and recognition accuracy holds up reasonably well once a working corpus exists, for languages large and small alike, which suggests that the absence of dedicated speech resources for a given Indian language reflects a lack of prior investment far more than any inherent acoustic or technical barrier. This pattern matters directly for the case examined in the next section.

5. A Closer Look: The Gap Facing Hindi's Native Languages

Among the languages surveyed in Section 4, one case stands out for the scale of the mismatch between its speaker population and its research attention. Bringing together the review presented in Sections 3 and 4, four gaps in the existing literature can be stated precisely. First, no study reviewed here addresses more than one native language of Hindi as a distinct identification target. Koolagudi et al.'s treatment of Rajasthani, discussed above, is a single exception within a fifteen-language study otherwise focused on major scheduled languages of India. Second, no study addresses Bhojpuri, Chhattisgarhi, Magahi, Haryanvi, Marwari, Malvi, Awadhi, Pahari, or any of the further native languages of Hindi discussed below, whether individually or as a group. Third, the studies that do address closely related low-resource Indian languages work with speaker populations several orders of magnitude smaller than any native language of Hindi; Terhija et al.'s work on Angami and Dey et al.'s work on Mizo both illustrate this. Fourth, almost none of the literature reviewed treats utterance durations below three seconds as a primary evaluation condition, despite its practical importance for real deployment.

This gap is difficult to justify on demographic grounds. Table 2 lists several native languages of Hindi alongside their Census 2011 speaker populations [5]. It compares them against several of India's officially

scheduled languages, which have historically received considerably more institutional and research attention.

Table 2: Native languages of Hindi compared against selected scheduled languages of India (Census 2011)

Native language of Hindi	Speakers	Scheduled language	Speakers
Bhojpuri	50,579,447	Assamese	15,311,351
Rajasthani	25,806,344	Maithili	13,583,464
Chhattisgarhi	16,245,190	Santali	7,368,192
Magahi	12,706,825	Kashmiri	6,797,587
Haryanvi	9,806,519	Konkani	2,256,502

Several of Hindi's native languages carry larger speaker populations than half of India's officially scheduled languages. They lack comparable institutional recognition under the Eighth Schedule of the Constitution of India, however. This administrative asymmetry, rather than any linguistic property of the languages themselves, is the most direct explanation for the research gap this review has documented. Scheduled status has historically correlated with dedicated corpus development, through bodies such as the Linguistic Data Consortium for Indian Languages, and native languages of Hindi have not benefited from that same correlation.

The acoustic difficulty compounding this gap deserves explicit statement too. These native languages share a common Indo-Aryan ancestry with standard Hindi and, in several cases, with one another. They follow recognizable linguistic clusters. A Rajasthani cluster includes Rajasthani, Marwari, Mewari and Malvi. A Bihari cluster includes Bhojpuri and Magahi. An Eastern Hindi cluster includes Awadhi and Chhattisgarhi. A Pahari cluster includes Garhwali and Kumauni. A Western Hindi cluster includes Haryanvi, geographically and linguistically the closest to standard Hindi itself [6]. This is precisely the kind of acoustic overlap the literature reviewed in Section 3.4 identifies as the field's hardest unsolved challenge. Here it appears in a specific combination: substantial demographic scale, acute acoustic difficulty, and near-total research neglect. This review has not found that combination replicated for any other language group in the literature surveyed.

6. Open Challenges and Future Directions

Several open challenges emerge from this review which are following.

Corpus scale and coverage. Table 2 lists only five of the native languages addressed by the case study corpus. Marwari, Mewari, Malvi (beyond its inclusion in the VAANI-derived experiments), Khortha, Sadan, and numerous smaller native languages remain entirely uncovered by any dedicated corpus.

Short-duration robustness. As Section 3 established, the great majority of the literature reviewed evaluates on utterances of ten seconds or longer, despite the practical importance of short-duration performance for applications such as call routing. Systematic evaluation of how identification accuracy degrades as utterance duration shrinks toward one second remains largely unaddressed, for the specific case of closely related native languages.

Code-switching. Real-world speech in these languages frequently mixes native-language and standard-Hindi vocabulary within a single utterance. The literature reviewed in Sections 3 and 4 does not directly address this phenomenon for this language family. Word-level or segment-level identification within code-mixed utterances remains an open problem specific to this setting.

Institutional and policy dimensions. As Section 5 established, the research gap facing Hindi's native languages is substantially administrative rather than technical in origin. Closing it fully will likely require institutional recognition and dedicated funding comparable to what India's scheduled languages have historically received. This dimension sits outside the scope of technical research alone, but this review's demographic evidence directly supports it.

7. Conclusion

This paper has reviewed the automatic language identification literature from Zissman's foundational 1996 comparison through the attention-based and knowledge-distillation architectures that define current practice. A clear historical pattern emerges from this review. Acoustic similarity between target languages, rather than sheer language count, is the field's most persistent source of difficulty. Short-duration robustness remains comparatively under-addressed, too, despite its practical importance. Against this backdrop, the paper has surveyed the smaller but growing body of work addressing India's regional and low-resource languages, from Angami and Mizo through Assamese and Bengali, and has found the same corpus-scarcity and acoustic-similarity challenges recurring across otherwise quite different languages. Within this broader landscape, Hindi's native languages stand out for the scale of the mismatch between their demographic weight and the research attention they have received. Several carry speaker populations exceeding fifty million, using Census 2011 figures, yet the language family as a whole combines substantial scale with acute acoustic difficulty and near-total research neglect. Hindi's native languages specifically and for India's low-resource languages more broadly. Corpus scale and coverage, short-duration evaluation, cross-corpus generalization, code-switching, and institutional support all need further attention. Even so, the evidence assembled in this review suggests the technical barriers to serving these substantial, under-resourced language communities are lower than the current absence of dedicated research might otherwise suggest.

References

- [1] Zissman, M. A. (1996). Comparison of four approaches to automatic language identification of telephone speech. *IEEE Transactions on Speech and Audio Processing*, 4(1), 31-44.
- [2] Terhijja, V., Sarmah, P., & Vijaya, S. (2018). Development of speech corpus and automatic speech recognition of Angami. In *6th Workshop on Spoken Language Technologies for Under-Resourced Languages*.
- [3] Dey, A., Lalhminghlui, W., Sarmah, P., Samudravijaya, K., Prasanna, S. R. M., Sinha, R., & Nirmala, S. R. (2017). Mizo phone recognition system. In *2017 14th IEEE India Council International Conference (INDICON)* (pp. 1-5). IEEE.
- [4] Chakraborty, J., Nath, S., Nirmala, S. R., & Samudravijaya, K. (2018). Language identification of Assamese, Bengali and English speech. In *6th Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU 2018)* (pp. 177-181).

- [5] Census of India. (2011). *Census of India 2011: Language*. Office of the Registrar General and Census Commissioner, India.
- [6] Rao, K. S., & Nandi, D. (2015). Language identification: A brief review. In *Language Identification Using Excitation Source Features* (pp. 1-15). Springer.
- [7] Gupta, R. (2016). Mel-frequency cepstral coefficients (MFCCs) for text-independent speaker recognition. *International Journal of Advanced Research in Computer Science*.
- [8] Sadjadi, S. O., Kheyrkhah, T., Tong, A., Greenberg, C., Singer, E., Reynolds, D., Mason, L., & Hernandez-Cordero, J. (2018). Performance analysis of the 2017 NIST language recognition evaluation. In *Proceedings of Interspeech 2018* (pp. 1798-1802).
- [9] Reynolds, D. A., & Rose, R. C. (1995). Robust text-independent speaker identification using Gaussian mixture speaker models. *IEEE Transactions on Speech and Audio Processing*, 3(1), 72-83.
- [10] Koolagudi, S. G., Rastogi, D., & Rao, K. S. (2012). Identification of language using Mel-frequency cepstral coefficients (MFCC) and Gaussian mixture models (GMM). *Procedia Engineering*, 38, 3391-3398.
- [11] Lopez-Moreno, I., Gonzalez-Dominguez, J., Plhot, O., Martinez, D., Gonzalez-Rodriguez, J., & Moreno, P. (2014). Automatic language identification using deep neural networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 5337-5341).
- [12] Grisard, C., Ferras, M., Kanevsky, D., & Sahidullah, M. (2019). Bottleneck feature based i-vector approach for language identification. In *Proceedings of INTERSPEECH*.
- [13] Geng, W., Wang, W., Zhao, Y., Cai, X., & Xu, B. (2016). End-to-end language identification using attention-based recurrent neural networks. In *Interspeech 2016* (pp. 2944-2948).
- [14] Bartz, C., Herold, T., Yang, H., & Meinel, C. (2017). Language identification using deep convolutional recurrent neural networks. In *International Conference on Neural Information Processing* (pp. 880-889).
- [15] Heracleous, P., Takai, K., Yasuda, K., Mohammad, Y., & Yoneyama, A. (2018). Comparative study on spoken language identification based on deep learning. In *2018 26th European Signal Processing Conference (EUSIPCO)* (pp. 2265-2269). IEEE.
- [16] Mukherjee, H., Ghosh, S., Sen, S., Obaidullah, S. M., Santosh, K. C., Phadikar, S., & Roy, K. (2019). Deep learning for spoken language identification: Can we visualize speech signal patterns? *Neural Computing and Applications*, 31(12), 8483-8501.
- [17] Vuddagiri, R. K., Gurugubelli, K., Jain, P., Vydana, H. K., & Vuppala, A. K. (2018). IIITH-ILSC speech database for Indian language identification. In *6th Workshop on Spoken Language Technologies for Under-Resourced Languages* (pp. 56-60).
- [18] Shrishrimal, P. P., Deshmukh, R. R., & Waghmare, V. B. (2012). Indian language speech database: A review. *International Journal of Computer Applications*, 47(5), 17-21.

[19] Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) GmbH. (2020). *A study on open voice data in Indian languages*. FAIR Forward - Artificial Intelligence for All, on behalf of the German Federal Ministry for Economic Cooperation and Development (BMZ).

[20] Markov, I., & Sidorov, G. (2018). CIC-IPN@INLI2018: Indian native language identification. In *Proceedings of the Tenth Annual Meeting of the Forum for Information Retrieval Evaluation*.