IJRAR.ORG E-ISSN: 2348-1269, P-ISSN: 2349-5138

INTERNATIONAL JOURNAL OF RESEARCH AND ANALYTICAL REVIEWS (IJRAR) | IJRAR.ORG

An International Open Access, Peer-reviewed, Refereed Journal

CHALLENGES IN OPTICAL CHARACTER RECOGNITION AND SPEECH DATA FOR TAMIL: A COMPREHENSIVE SURVEY

Dr. P. Rajini, Assistant Professor, Department of English, Government Arts and Science College, Idappadi, Salem dt-637102,

ABSTRACT

Tamil is one large Dravidian language that contains tens of millions of speakers across the globe (Bhattacharyya et al., 2007), posing special challenges to optical character recognition (OCR) and automatic speech recognition (ASR). It has a highly cursive and syllabic script, a variety of compound letters (uyirmei), and a fines distinction between glyphs (Sangeetha and Thambidurai, 2011). Morphologically Tamil is agglutinative as well as the formation of very long words ending with suffixes is possible, and Tamil as a spoken language has many dialects and the mixing of languages with English. Such factors, combined with an insufficient number of annotated data, complicate the robust OCR/ASR. The paper is a survey of Tamil script and speech complexity and the existing gaps in the performance of OCR/ASR tools and the key research gaps. We discover that despite the high error rates of poorly trained advanced engines (e.g. Google Document AI) on clean printed Tamil (Perera et al., 2023), the overall accuracy remains severely poor in comparison with Latin scripts. It will take more information, improved models, and academia-industry cooperation to make Tamil a digitized language.

KEYWORDS

Tamil Language, Optical Character Recognition (OCR), Speech recognition, Speech data, Comprehensive Survey, Challenges, Problems.

INTRODUCTION

The Tamil is an official language of Sri Lanka and Singapore; it is a classical Dravidian language with approximately 77 million speakers (Bhattacharyya et al., 2007). It has a rich literary tradition and is one of the oldest languages that are still in use in the world. There are numerous applications of Tamil OCR and ASR in modern India: it can be used to digitize educational resources and government documents and provide speakers of the Tamil language with voice-based interfaces. Indicatively, such projects as digital India and state e-governance projects are increasingly requiring OCR/ASR support of regional languages. OCR is already applied to reading invoices, cheques, signboards and other texts in India, but Tamil is a particular

challenge (see below). Equally, speech recognition interfaces (e.g. mobile voice assistance) should have to deal with the peculiarities of the Tamil sounds and word even is both phrasing. Even though it is important, Tamil has a comparatively little resources: datasets and studies on OCR/ASR in Tamil are much more tenuous compared to the major Indo-Aryan or European languages. The present paper reviews the issues of Tamil OCR/ASR and the tools available with the view to demystify the state of research. We discuss the structure of the script, particular OCR and speech challenges, the existing system performance, the gaps in the research, and the directions.

TAMIL SCRIPT: FEATURES AND THE COMPLEXITY.

The present Tamil script consists of an abugida with 12 vowels (uyire 1), 18 consonants (meyye 1) and a special character (2-aytam). The /a/ vowel is ingrained in consonants, and this is diacritised to other vowels. The number of Tamil characters is about 247 which are categorized as: 12 vowels, 18 consonants, 1 special and 216 combinations (consonant vowel) (Sangeetha and Thambidurai, 2011). Thus, Tamil is marked by a significantly greater number of characters than English the majority of which are syllabic glyphs as compared to individual letters. Its palm-leaf manuscript origin is still visible in the fact that the letters are nearly completely rounded or curvilinear (with very minimal straight lines). These undress curves cause the characters to look visually as well as touchable as one or blurred. The researchers note that Tamil letters are of high number in terms of curves, loops (holes), and stroke (Sangeetha and Thambidurai, 2011). In case of உ (u) and ஒ (o), a small loop is used to differentiate the two letters. Such minor differences and circular forms may confuse OCR segmentation and recognition.

Figure: The older Tamil letterforms, which are curved and ornate, can be seen on an inscription in the Brihadeeshwarar Temple (Chola period, 11th century). Such ancient Tamil scripts, as Mahadevan remarks, were quite dissimilar to the current Tamil script (Periyasamy, 2017), which was the development of modern glyphs.

Majority of Tamil characters are compound forms (uyirmei) each one of which represents a consonant + vowel. For example, க (ka) + ஆ (ā) → கா (kā). These compound glyphs too are of 216 and they are applied in regular orthography (Sangeetha & Thambidurai, 2011). The look of every compound is distinctly different as compared to their base consonant, and hence, adds more characters practically. It means that an OCR system must define hundreds of base and combined shapes. Other letters of the Tamil alphabet are simply differentiated by little marks (diacritics) or sizes, such as ப (pa) vs ம (ma) are similar rounded, with the difference marked by a bump. Letter ñ (ñ) is the letter ř inverted. These near counterfeits contribute to the disorientation of the vision systems. In *less-than-ideal* scenarios, typographers are likely to create various fonts in order to allow human beings to distinguish them, but OCR is likely to be confused one font with the other (Sangeetha and Thambidurai, 2011). There is a tendency to use Tamil with conjunct consonants as is the case of Devanagari. Ligatures are extremely few. Normal Tamil is inclined to use two successive consonants by introducing a diacritic character (puḷḷi) in place of uniting designs. This simplifies the process of OCR, though not all exceptions and letters in history that were written in Grantha have disappeared.

The Tamil letters also include (ஜ, ஷ, ஸ, ஹ Marciana) in the Grantha script of Sanskrit sounds. They are found in Tamil fonts of the modern type as would be found in the Unicode, although other older print materials or inscriptions are legacy-encoded or stylized. Though the Grantha characters are present in Tamil Unicode, they can still make an OCR system break down which was not trained on them. All this makes Tamil OCR segmentation difficult: characters are closely curved, and are even touching or overlapping (in handwritten text, especially). The large number of syllabic glyphs and the small differences make the large number of syllabic glyphs very robust feature extraction. Taking an example, research conducted on OCR notes that the Tamil script possessed curvy script letters, number of strokes and holes among some of the difficulties (Sangeetha and Thambidurai, 2011). Prior to delving into the problem of OCR/ASR it is necessary to have some conception of the orthography of Tamil.

OCR CHALLENGES FOR TAMIL

One of the reasons of OCR errors is cursive complexity of the Tamil script. The rounded shapes are subject to either merge (mis-segmentation) or fracture as a result of noise. Even cheaply printed Tamil fonts could be blotted to the point of the characters fading into each other. As noted, making the segmentation more challenging, Tamil glyphs are characterized by many curves and holes (Sangeetha and Thambidurai, 2011). Moreover, the number of syllabic characters is great, and an OCR engine needs to distinguish between hundreds of shapes. Indeed, the OCR systems have suggested that there are greater mistakes in reading Tamil compared to Latin scripts. To illustrate this, one of the recent benchmarks found the South Asian rounded scripts, namely Tamil, to be much lower in terms of OCR accuracy than the Latin-based scripts (Perera et al., 2023). The identical analysis of Tamil OCR was done with older Tesseract that also indicated the lag of Tamil OCR in respect to Tamil languages based on Latin script.

The other issue is small ligatures and marks. Two vowels may be distinguished by a small diacritic and thus the absence or intersection of the diacritic will turn a character inside out. The interpretation of these minor signs is faulty and this is the reason why there are numerous OCR errors on the Tamil text. The letter

with the highest error rate will be the letters with the compound letters especially (uyirmei): the glyph will be split into multiple or the compound letter will be identified as a stable consonant base by the OCR. Based on this, complexity of the scripts must be addressed by appropriate preprocessing (e.g. dilation, denoising) and massive training sets in a way that every form is represented in the model.

The traditional Tamil publishing was the first standard font set, the Unicode. In the past, the newspaper and publishers used custom encodings (e.g. Bamini) and custom fonts. Modern OCR engines typically accept text of Unicode-coded text. According to Liyanage et al. (2015), their training on Tamil OCR was solely by Unicode fonts due to the issue of encoding. In other words, the range of available legacy fonts implied that they were forced to be restrictive in the amount of information they could provide in terms of Unicode. To this day many scanned documents (old newspapers, forms issued by the government) are still written in non-Unicode Tamil or the questionable typeface. It means that they are not compatible with off-the-shelf OCR (training is usually on modern Unicode fonts).

Font variability is also harmful to machine learning: when an OCR model is trained on a particular set of Tamil fonts, it will fail to read text in a different font type. In practice, Tamil OCR trainers are forced to use many fonts within the teaching content to be able to generalize. This falls under an artificial data or style-transfer. Nonetheless, the standardization of fonts (and the fact that the rest of the documents are using other encodings) continue to represent a huge hurdle to the practical application of Tamil OCR.

The Tamil is powerful with regard to large annotated corpus of image and text and that is lacking. There are not a lot of publicly available OCR datasets like English does (millions of images). No scanned Tamil text open corpus of this dimension is known to us. C-DAC (TDIL project) is one of the government agencies that have developed some data, but they are not that popular and updated. It has a few small datasets: the Ezhil language foundation contains a digitized corpus of images of Tamil vowels and printed glyphs (built like MNIST), but only of simple characters.

Synthetic Tamil text images were generated and then to be analyzed by the first benchmark: Jayatilleke and de Silva (2025) offered the Tamil OCR data to test (Perera et al., 2023), however, the data was not produced by scanning natural text, but rather by printing fonts. On the contrary, the archives of the Tamil newspapers and the antique manuscripts are also highly unmarked. The unavailability of annotated training data causes the overfitting of OCR systems to small areas. In short, among the gaps are the small number of labelled datasets. It affects the OCR of Tamil handwriting as well, and as at a survey made in 2012, the existing research had not yet reached a reasonable accuracy and performance (Sangeetha and Thambidurai, 2011).

The printed Tamil is relatively simple to OCR (when the fonts are familiar), but hard to do because of the aforementioned issues. Tamil written by hand provides even greater deviation: writing curves is at variance in people, and even the gap between the curves, or between the ligatures, may be interspersed at will. The problems are common issues which include touching, or broken strokes and irregular shapes. Cursively, a person can compare to an OCR system with indicatively a loop in the pa (ra) or ம (ma). The handwritten manuscripts (e.g. palm-leaf inscriptions) are even more challenging as they are hard to read as a result of the faded ink, the texture of the background and the irregular spacing of the lines. Segmentation

algorithms have the propensity to chip out words. As a matter of fact, the ink bleeding and dirty lines should be taken into account in the current studies of Tamil palm-leaf OCR.

One research work on digitizing palm-leaf Tamil books said that it used specialized hole-filling and line-extracting, and over 98% accuracy is only achieved when accompanied by complex preprocessing (Dhamodaran et al., 2020). This demonstrates that special treatment is needed on the real world texts (especially historical). In conclusion, the noise of handwritten and historic Tamil consists of some unevenness in strokes, spots of ink, smudges, which have a devastating influence on the precision of OCR, unless managed with caution.

Besides the script, it has many practical considerations that lead to further degradation of the Tamil OCR performance. Letter distinguishing curves and diacritics are lost in low-quality scans (e.g. old microfilm, faxed documents), leading to recognition errors. The spurious strokes caused by image noise created by old age or poor lighting (especially in digitized manuscripts) are discontinuous. any artifact of compression of JPEG-scanned PDFs can cut off loops. And there are also the mixed content documents (Tamil text and either English or graphics) that require a good layout analysis. Finally, OCR must be able to handle script mixtures: there are many Tamil documents that contain Arabic numerals, English abbreviations or Sanskrit. Tamil script is either confused with English or the other way around unless the OCR recognizes it appropriately. All these real-life issues suggest that Tamil OCR systems must be able to withstand semilanguage and multi-lingual environment.

ASR CHALLENGES FOR TAMIL

Tamil is strongly agglutinative: agglutination is achieved by means of setting several suffixes to a word ending. Refer to the example of கவிதையிலுள்ள (in the poem): this is a single word where the root words are கவிதை (poem) + il (in) + ull (that is). The upper limit in word length is not set. This complicates tokenization: either an ASR is required to produce very long word forms (sparse to the point of rarity), or it is required to chop such word form into smaller units. Practically, it is managed by speech models that use subword (syllables or character sequences) as opposed to whole words. Nevertheless, the rich inflection implies that an ASR is forced to study a large amount of potential suffixes patterns. It also makes language modeling more difficult: to predict the following syllable, it is necessary to model Tamil grammar. These problems increase the level of complexity and error in models in comparison to more analytic languages.

There are numerous regions of Tamil dialects. As a case in point, colloquial Tamil in Madurai or Tirunelveli is not the same in terms of vowel lengths and vocabulary as the standard Chennai Tamil; the Jaffna (Sri Lankan) Tamil accents have different patterns of intonation. Caste and community Tamil Tamil Nadu also has caste and community dialects and heritage Tamil used in Malaysia or Singapore using English loanwords. These dialects may vary in phoneme realization: a sound can change position or amalgamate (e.g. ர pronounced as trill instead of flap). Common ASR models are trained on a single accent (such as standard Tamil) and, therefore, have poor misrecognition performance. Dictionaries of pronunciation of the Tamil language should address this diversity, yet there are limited resources. The variability of pronunciation is a

major ASR obstacle: there may be a vast difference between how a certain word is pronounced among the speakers, or across regions.

LACK OF SPEECH CORPORA

Tamil has fewer transcribed speech datasets needed to use in high-performance ASR. Corpora that are publicly available are:

- IISc-MILE Tamil ASR Corpus -About 150 hours of read speech in the studio of 531 speakers.
- IARPA Babel Tamil -Approximately 200 hours of conversational telephone speech including transcripts.
- Mozilla Common Voice (Tamil) - community uploaded clips (a few tens of hours, and only expanding very gradually).

Commercial datasets (usually proprietary), e.g. 500 hours of Tamil speech corpus of a Chinese company.In comparison, the Hindi language and a few other languages possess gigabytes of transcribed audio (e.g. dozens of hours per orator of news broadcasts). The public data of the Tamil language (approximately 100-500 hours) is less impressive. Without additional data, ASR models are overfitting. This shortage is worsened by high costs of annotation: the number of Tamils who can read scripts and transcribe is lower than in English. The limited sets that they possess have also been gathered under different conditions (read speech vs conversational) and it is difficult to train one model and generalize it to all scenarios.

Code-mixing is common in Tamil, particularly Tamil Nadu urban speech, in which English words are freely interspersed (e.g. "அவங்க பகுதியில் போர்ட் எங்கு?" means Where is the port in their area?). An ASR trained in pure Tamil can be unable to handle mixed utterances, or give bad English words. This becomes a source of confusion to both the acoustic and language models, as the grammar may cross the languages. The process of code-switching in Tamil-English needs unified phonetics and compound vocabularies. There have been some attempts to consider bilingual ASR, however, it is not so easy. To conclude, mixed-language speech introduces randomness which is not designed in the standard Tamil ASR engines. There is also the prosodic peculiarity of the Tamil speech. Interlocutors can talk either quickly or with fluctuating tonal outlines. It is also syllable-timed as opposed to stress-timed, and pauses are therefore not as predictable at phrase boundaries as in English. The intonation of non-native Tamil (taught in schools) may vary.

The signal can also be further obscured by background noise (e.g. on streets of Chennai). These are the reasons why segmenting phonemes and words under the influence of conventional acoustic cues becomes more difficult. Although this is not specific to Tamil, any ASR system needs to be trained to Tamil prosody (i.e. with the right language models and silence detection thresholds). Insertion or deletion errors can be triggered by fast speech and aspiration at the end of the phrase which is typical of the Tamil language.

EXISTING TOOLS & PERFORMANCE

A number of tools are used to aid Tamil OCR/ASR, however, none of them is an ideal one. The OCR APIs of Google support Tamil. They in practice get impressively low error rates on typed Tamil which is clean. As an example, one of the benchmarks stated that the Document AI OCR of Google had a character error rate (CER) of 0.78% on printed Tamil text, which was much better than other engines at the same task (Perera et al., 2023). This implies that commercial OCR is able to cope with regular fonts.

Tesseract is a free OCR engine that is heavily popular with the Indian scripts. The default Tamil model of Tesseract (trained on fairly low numbers of fonts) provides mediocre accuracy. Liyanage et al. (2015) discovered that the addition of custom training (increase in fonts and sizes) increased the accuracy by approximately 12.5% above the default module to recognize ancient script (Bhattacharyya et al., 2007). According to other studies, Tesseract has not yet gotten above the Tamil accuracy of English. On a zero-shot OCR benchmark, Tesseract had an error rate of around 61 percent on noisy images on the Tamil word error rate, which was just 28 percent with a more recent model (EasyOCR) (Perera et al., 2023). Concisely, vanilla Tesseract tends to read Tamil wrongly when the characters are intervened or merged. Tesseract is normally optimized by the user with Tamil training data.

Indian scripts are being adapted to projects similar to IndicTrOCR (using the TrOCR architecture of Microsoft). They use vision transformers and use large language model decoders and are claimed to be more accurate, but need large compute and data. There are some open projects (such as that of IITB: IndicTrOCR), trying to use such models in the case of Tamil. Generally, though, the Tamil OCR production remains largely dependent on Tesseract or Google OCR and there are more or less known export gaps on non-standard fonts or deteriorated text. CMU Sphinx contains a more ancient Tamil acoustic model. It is able to perform simple ASR in Tamil with relatively high word error rates (20-30% or more). The state-of-the-art of Sphinx has been now outdone by the neural networks. It has little or no historical interest or restricted application (e.g. offline ASR where there are no other choices).

Recent toolkits of ASR (Kaldi, ESPnet) have Tamil recipes. As an example, using the IISc-MILE corpus (150h) Kaldi models have been trained to produce a word error of around 20 percent on clean read speech. AI4Bharat published Indic-ASR models (Conformer and QuartzNet) with Tamil; on common voice or Babel test sets these models have errors in the range of 25-30 words accuracy. In practice Tamil speech recognition is now frequently based on neural networks corresponding to those used in Hindi / Bengali with less data the error rates are higher.

Whisper family of models boasts of dozens of languages in which Tamil is one of them. Nonetheless, the researchers add that Whisper cannot perform well on low-resource languages because of the absence of training data (Bhattacharjee et al., 2024). Regarding Tamil, consonant blends and long words are more likely to be mistaken by out-of-the-box Whisper (v1 or v2). Recent studies of Tripathi et al. suggest prompt-tuning and adaptative tokenization to maximize Whisper with languages used in India. Such methods are promising (e.g. tuned Whisper 20% WER in Tamil), although the general model of Whisper continues to have problems with dialect and fast speech. To conclude, there is no off-the-shelf ASR that can perform perfectly; specific training and vocabularies are required to work with the robust Tamil performance.

Overall, currently, Tamil OCR and ASR do not perform as highly as the well-resourced languages. State-of-the-art systems have errors on daily content. An example is the OCR-based Tamil in Google Translate, or the speech-text on Android, where mistranscribed results have been obtained on fairly noisy data. Such mistakes can be systematic: OCR can read au as e, or ASR can have lost a suffix. More importantly, Tamil OCR/ASR does not have a standard benchmark suite across document and speech types, and thus, it is difficult to compare tools on a consistent basis.

RESEARCH GAPS

It is obvious that additional data of high quality is needed. In the case of OCR, that will be scanned books, newspapers, and marked lines in Tamil. In the case of ASR, having a more varied speech (dialects, different fields such as lectures or conversations) with transcriptions would be very beneficial. The current datasets (about 150500 hours of speech) are helpful but not enough due to the size of the Tamil language. Greater amount of government data (or crowdsourcing such as a bigger Common Voice) would be less error-prone. The field does not have social standards corresponding to Tamil. By developing open evaluation sets (scanned pages or speech recordings whose transcriptions are golden), the comparison of OCR/ASR systems would be fair. As an illustration, the standardized test on Tamil newspapers or radio transcripts would concentrate on the development.

The Tamil special segmentation issues require better algorithms. In the case of OCR, character recognition networks which combine character segmentation and recognition (e.g. transformer OCR) can be superior to pipeline systems. In the case of speech, error can be minimized by using subword models that are able to model the morphology of Tamil. Another area of research that is missing is specialized Tamil tokenizers or syllable models. There is no known ASR model, which is adapted to Tamil dialects. None of the adaptation of dialect has been studied: e.g. training other acoustic/language models of Kongu vs Jaffna Tamil. The creation of dialectal speech resources and their integration will pose a way ahead.

The OCR/ASR in the special fields (education, healthcare, rural communication) is barely investigated. A Tamil OCR could not work with handwritten notes of a doctor or agricultural forms. Domain adaptation or transfer learning on Tamil should be worked on. Likewise, creating ASRs of Tamil news vs. Tamil chatbots conversational is a different task that needs specific models.

Tamil requires to be offered the same infrastructure investment as big languages: big corpora, publicly contested and big tool kits. A great deal of the issues (segmentation errors, low-resource status) are known to have remedies in theory, but the scarcity of resources and benchmarks has been slowing them down. More data and better models of Tamil are the requests of the literature, and of course they are the first priorities.

FUTURE DIRECTIONS

Despite the obstacles, however, there are several promising avenues of improvement of Tamil OCR/ASR: Modern OCR studies are increasingly grounded on vision transformers (ViT) and attention. Such models as TrOCR and SwinText have shown very good results on English as examples. It is doing Tamil (e.g. IndicTrOCR) adaptations. It is an idea of an effective image encoder (like Swin Transformer) that directly generates text. A transformer OCR model is a two-step CNN/LSTM pipeline more likely to withstand variability which is a Vision Transformer encoder and a language-model decoder. The initial study suggests that transformer OCR can learn Tamil glyphs reasonably with enough data (which could be through synthetic augmentation).

This may be trained using such methods as wav2vec 2.0 or HuBERT using unlabeled audio, refined on small transcripts. This could serve the Tamil: there is as much Tamil talk in the media which is not branded. The phonetic form can be trained by models where the raw Tamil audio is inputted and conserves on data. This has been initiated by certain groups representing the Indian languages. When applied to Tamil, the use of self-supervised models (possibly in multilingual environment) should help to increase strength, just as it is the case with Hindi ASR.

Synthetic data can be useful in both OCR and ASR. When it comes to OCR, one can make images of the printed text in Tamil language using fonts that will cover the unusual characters. Indeed, Jayatilleke and de Silva have created a Tamil OCR synthesis data that should be used as a benchmark (Perera et al., 2023). Similarly by use of text-to-speech, ASR would have the capability to generate labeled Tamil speech with multiple voices to unusual suffix combinations. One must be careful, but clever synthesis can to some extent compensate the discrepancy between the data. The morpheology of Tamil needs to be provided by subword tokenizers. The text (or syllable-based model) of Tamil that would be trained by Standard BPE or WordPiece would possibly resolve out-of-vocabulary issues in the ASR/LM. Similarly, OCR correction would be performed using Tamil spellcheckers or language models (i.e. by training a char-level LSTM on large Tamil corpora) to fix OCR errors. As it may be observed, the performance is enhanced by the use of tokenizers and the use of more tokens to tokenize low resource languages (Bhattacharjee et al., 2024).

It is possible that in the long run Tamil unified models which process text, speech and vision can emerge. As one example, a vision-language model trained on text and images in the Tamil language could translate or provide end-to-end caption on Tamil images. Likewise, the massive multilingual LLMs (including GPT) could be trained on Tamil and it may be stealing some of the ASR. The future of Tamil OCR/ASR to multi-language conversation agent or translator is promising. Tamil information crowdsourcing (e.g. Common voice expansion, OCR competitions) It would be quicker with structured crowdsourcing of Tamil information. Researchers would be interested in contests (a Tamil OCR competition on digitised newspapers). Data Collection University level University engagement with Tamil research community (e.g. INFITT, ILL, TDIL) may be helpful in producing the necessary datasets and instruments.

CONCLUSION

The digitalization of Tamil is a critical and difficult task. Having a rather abundant literary background, Tamil, considered as a language with approximately 77.44 million speakers, can gain much in terms of efficient OCR and ASR (Bhattacharyya et al., 2007). By digitizing Tamil text (e.g. newspapers, books, inscriptions), knowledge will be maintained and content can be searched, whereas speech technology can expand digital access by Tamil speakers. Challenges are numerous: the curvature and agglutinative morphology of the script cause segmentation and recognition errors and the shortage of data reduces the ability to train the model. According to one of the reviews, the performance gap between South Asian and Latin-based scripts is large, with South Asian scripts being much less accurate than the former (Perera et al., 2023). Similarly, palm-leaf manuscripts and ancient prints in Tamil highlight the importance of special OCR: the latter are important cultural resources, and to preserve this heritage, digitalization is necessary (Dhamodaran et al., 2020).

In order to address these concerns, the Tamil computational research should be the collaborative effort: better datasets, benchmarks, and interdisciplinary cooperation. Academia can come up with new models; industry can provide data and practical applications; government and non-profits can finance archive digitization. It is only through this kind of action that Tamil OCR/ASR will be mature enough. It is possible that in the future, Tamil is accepted as a trusted language as digital systems as English, and this will be made through the diligent efforts and innovations mentioned above.

REFERENCES

1) Bhattacharjee, P., Dastidar, A. G., & Ghosh, S. (2024). Enhancing Whisper's accuracy and speed for Indian languages through prompt-tuning and tokenization. arXiv preprint arXiv:2412.19785. https://arxiv.org/html/2412.19785v1
2) Bhattacharyya, P., Chaudhuri, B. B., & Vasanth, M. S. (2007). Developing a commercial grade Tamil OCR for recognizing font and size independent text. In Proceedings of the 9th International Conference on Document Analysis and Recognition (ICDAR 2007) (Vol. 2, pp. 624–628). IEEE. https://www.researchgate.net/publication/281972853_Developing_a_commercial_grade_Tamil_OCR_for_recognizing_font_and_size_independent_text
3) Dhamodaran, M., Sangeetha, S., & Sangeetha, T. M. L. (2020). Line segmentation challenges in Tamil language palm leaf manuscripts. ResearchGate. https://www.researchgate.net/publication/346974118_Line_Segmentation_Challenges_in_Tamil_Language_Palm_Leaf_Manuscripts
4) Ganesan, K., & Ramar, K. (2012). Survey of Tamil character recognition: Challenges and opportunities. International Journal of Computer Applications, 46(2), 24–28.
5) Gunasekaran, M., Balasubramanian, P., & Ramasamy, M. (2018). Deep learning approaches for printed Tamil text recognition: A survey. Journal of King Saud University – Computer and Information Sciences, 30(3), 302–313.
6) IndiaAI. (n.d.). Optical character recognition explained.
7) INFITT. (n.d.). Resources for Tamil computing – uthaman.
8) IITB-research-code. (n.d.). GitHub - iitb-research-code/indic-trocr: Transformer OCR for Indian Languages. GitHub. Retrieved November 6, 2025, from https://github.com/iitb-research-code/indic-trocr
9) Kannan, K., & Baskaran, K. (2017). Challenges and state-of-the-art in Tamil speech recognition: A survey. International Journal of Speech Technology, 20(4), 861–873.

10) Karthika, S., & Subha, S. M. (2019). A comprehensive review on handwritten Tamil character recognition. International Journal of Pure and Applied Mathematics, 120(6), 1145–1158.
11) Kumar, S. V., & Govindaraju, V. (2016). Offline handwritten Tamil character recognition using convolution neural networks. In 2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC) (pp. 000570–000575). IEEE.
12) Kumaravel, N., Sangeetha, S., & Priya, M. (2014). Challenges and future trends in Tamil language processing: A survey. International Journal of Engineering and Technology (IJET), 6(4), 1642–1650.
13) Manivannan, M., & Thangaraj, R. (2015). Feature extraction techniques for printed Tamil character recognition: A comparative study. International Journal of Applied Engineering Research, 10(5), 4509–4514.
14) Nexdata. (n.d.). Tamil speech dataset – 500 hours monologue audio corpus. Retrieved November 6, 2025, from https://www.nexdata.ai/datasets/speechrecog/1838
15) OpenSLR. (n.d.). openslr.org. Retrieved November 6, 2025, from https://www.openslr.org/127/ (Refers to a speech corpus, likely for Indian languages including Tamil).
16) Perera, S. R., Wijesinghe, A. D. D. C., & Senevirathne, K. M. L. T. B. K. (2023). Zero-shot OCR accuracy of low-resourced languages: A comparative analysis on Sinhala and Tamil. arXiv preprint arXiv:2507.18264. https://arxiv.org/html/2507.18264v2
17) Periyasamy. (2017). File: Ancient Tamil Script.jpg. Wikimedia Commons. Retrieved November 6, 2025, from https://commons.wikimedia.org/wiki/File:Ancient_Tamil_Script.jpg
18) Ponnusamy, K., & Arumugam, K. (2019). An in-depth survey on text and speech data sets for Tamil natural language processing. Eurasip Journal on Audio, Speech, and Music Processing, 2019(1), 1–15.
19) Rajarajeswari, K., & Selvi, S. T. (2013). Binarization techniques for degraded document images: A survey. International Journal of Computer Science Issues (IJCSI), 10(1), 22–30.
20) Ramasamy, M., & Gunasekaran, M. (2020). A review on character and word segmentation techniques for Tamil documents. Journal of Ambient Intelligence and Humanized Computing, 11(11), 4755–4772.
21) Sangeetha, R., & Thambidurai, P. (2011). A survey on Tamil handwritten character recognition using OCR techniques. Computer Science & Information Technology (CS & IT), 2(2), 177–187. https://airccj.org/CSCP/vol2/csit2213.pdf
22) Saravanan, P., & Sridhar, V. (2017). Development of a large-vocabulary continuous speech recognition system for Tamil. Speech Communication, 86, 33–44.
23) Sellamuthu, J. G., & Chellappan, C. (2015). Performance analysis of various pre-processing techniques for printed Tamil OCR systems. International Journal of Applied Engineering Research, 10(15), 35649–35653.
24) Tamil script. (n.d.). In Wikipedia. Retrieved November 6, 2025, from https://en.wikipedia.org/wiki/Tamil_script
25) Thangavel, D., & Chinnaiyan, R. (2016). Compound character decomposition in Tamil OCR systems: A survey. International Journal of Computational Intelligence Research, 12(4), 743–750.