

Language Technology Research at MILE Laboratory

Ramakrishnan A G and Madhavaraj A

MILE Laboratory, Department of Electrical Engineering

Indian Institute of Science, Bangalore, India

(agr_madhavaraja@iisc.ac.in)

Abstract

Medical Intelligence and Language Engineering Laboratory at the Indian Institute of Science developed Kannada and Tamil OCR and TTS systems and deploying them, created Braille and audio books for blind students and won Manthan Awards. Algorithms proposed for scene and born-digital word image recognition consistently retained the top positions in ICDAR Robust Reading Competitions since 2011. The TTS systems were adjudged second best in Blizzard TTS Challenge in 2013 and 2014. Current research is on solving the problems of unlimited vocabulary in building ASR systems for Dravidian languages of Kannada and Tamil, by using sub-word units for language modeling.

Keywords: OCR, TTS, ASR, transcription, Tamil, Kannada, Braille books, audio books, Manthan award, script recognition.

Résumé

இந்திய அறிவியல் கழகத்தின் மருத்துவ நுண்ணறிவு மற்றும் மொழிப் பொறியியல் (மறுமொழி) ஆய்வகம், கன்னடம் மற்றும் தமிழ் ஓ.சி.ஆர் மற்றும் டி.டி.எஸ் மென்பொருள்களை உருவாக்கி அவற்றின் மூலம் பார்வையற்ற மாணவர்களுக்கு பிரெய்ல் மற்றும் பேசும் புத்தகங்களை உருவாக்கி, மந்தன் விருதுகளை வென்றது. செல்பேசியில் பிடிக்கப்பட்ட மற்றும் பிறப்பு-டிஜிட்டல் சொல் படங்களை அடையாளம் கண்டு கொள்வதற்காக முன்மொழியப்பட்ட மறுமொழி வழிமுறைகள் 2011 முதல் ஐசிடிஏஆர் வலுவான வாசிப்பு போட்டிகளில் தொடர்ந்து முதலிடங்களைத் தக்க வைத்துக் கொண்டன. 2013 மற்றும் 2014 ஆம் ஆண்டுகளில் ப்ளிஸ்ஸார்டு டிடிஎஸ் சவாலில் மறுமொழி டிடிஎஸ் மென்பொருள்கள் இரண்டாவது சிறந்த இடத்தைப் பிடித்தன. தற்போதைய ஆராய்ச்சி, மொழி மாடலிங் செய்வதற்கு துணை சொல் அலகுகளைப் பயன்படுத்துவதன் மூலம் கன்னட மற்றும் தமிழ் ஆகிய திராவிட மொழிகளுக்கு ஏ.எஸ்.ஆர் அமைப்புகளை உருவாக்குவதில், வரம்பற்ற சொற்களஞ்சியத்தின் சிக்கல்களைத் தீர்ப்பதாகும்.

1. Introduction

The MILE laboratory at the Department of Electrical Engineering has made significant contributions to analysis of speech signals, text-to-speech conversion systems for Indian languages, document and scene image analysis and recognition and online handwriting recognition in Indian languages and is currently working on automated speech recognition for Tamil, Kannada and Hindi languages. The research focuses on real life applications of signal, image processing and pattern recognition in solving crucial problems in language technology, of particular relevance to Indian scenario, which is followed up with actual technology development. The work on document image analysis, script recognition, OCR of printed text, scene text recognition and text-to-speech conversion have been motivated from the commitment to develop deployable technology to enable visually challenged people to be able to have access to (by machine reading of) any printed book in Indian languages. In addition, free tools have been developed to read web text in any major Indian language in one's own script. Also, Indic kBD, a tool for typing in any Indian script on Linux & Windows using QWERTY keyboard using anyone of many keyboard mappings has been created and made available for download.

2. Speech Analysis, Synthesis & Recognition

2.1 Time-Domain Features for Speech Analysis

Novel techniques have been proposed based on knowledge-based acoustic-phonetic approach to detect stop closure-burst transitions and epochs in speech. The new nonlinear feature, defined in the time-domain, called plosion index, 75

is robust in detecting stop closure-burst transitions and performs much better than complex feature vectors of a large dimension. Extension of this, called the dynamic plosion index, has been shown to be robust in detecting instants of significant excitations in voiced speech and also the QRS complexes of noisy ECG.

2.2 DCT based Pitch-Synchronous Pitch Modification for Prosody Modification

A novel algorithm has been proposed for pitch modification. The linear prediction (LP) residual is obtained from pitch synchronous frames. The dimension of the DCT coefficients of the residual is modified by truncating or zero padding, and then the inverse DCT is obtained. This period-modified residual is then forward filtered to obtain the pitch modified speech. The radii of the poles of the filter are modified to smoothen the LP spectrum. This minimizes the mismatch between the pitch modified signal and the LP spectrum due to the change in the positions of the pitch harmonics. The technique has been applied to create interrogative sentences from affirmative ones in our Tamil TTS.

2.3 TTS Systems for Kannada and Tamil

The concatenative TTS system uses a basic unit, that is distinctly different from all the TTS systems in the world. This is very close to what is known as akshara in Indian languages and the concatenation is performed only across similar vowels, which makes it smooth and glitch-free. The *Thirukkural* Tamil TTS system developed is being used by over 1000 blind student members of Anna Centenary Library, Chennai. The *Madhura Vaachaka* Kannada TTS system has been used to convert high school and PUC books to audio books by www.kannadapustaka.org.

3. Document Image Reconstruction, Analysis, and Recognition

3.1 Knowledge-Driven Deep Models for Superresolution of Low-Resolution, Scanned Binary Document Images

While the widely accepted view in computer vision today is to use end-to-end approaches using deep neural networks (DNN), this work has convincingly shown that the performance of existing state-of-the-art DNN models for super resolution can be significantly improving by suitably modifying the objective function, driven clearly by the knowledge of the specific vision problem in question. Computationally efficient superresolution models are obtained by nonlinear fusion of the outputs of well-known image interpolation techniques. This method has been used to significantly improve the resolution of binary document images so that human readability as well as OCR recognition accuracy improve appreciably. This has been filed as an Indian patent and also a PCT application.

3.2 Script Recognition at the Word Level in Multilingual Documents

An algorithm has been proposed to identify the script of each word in a multiscript document image. Gabor and DCT features were independently evaluated for their effectiveness using different classifiers. Gabor features with support vector machine classifier has given promising results; i.e., over 98% for bi-script and tri-script cases.

3.3 Kannada OCR with Performance Better Than Google's Tesseract OCR

Inspired by the rich feedback in the ascending visual pathway in higher mammals, attention -feedback has been proposed for improving the performance of printed text and handwriting recognition systems. Different types of recognition errors are identified at the different stages of the machine learning system, and this feedback is used effectively to revise the binarization, line segmentation, character segmentation and recognition in printed text. This innovative idea has been filed as a patent. This has resulted in a Kannada OCR (Lipi Gnani, which has been shown to perform better than the latest version of Tesseract OCR on 250 benchmarking images (Shiva Kumar and Ramakrishnan, 2020). Tesseract OCR is being developed for 3 decades, originally by HP Labs for a decade, and then taken over by Google.

3.4 Analysis and Recognition of Camera-Captured Document Images

Techniques have been proposed to binarize coloured documents captured by cameras. New approaches were proposed for recognition of script at the level of the word in multi-script documents and for text extraction from complex, colour document images. An edge-based connected component approach has been proposed for binarization of color documents. It handles documents with multi-colored texts with different background shades; deals with text of widely varying sizes, not handled by local binarization methods; automatically computes the

binarization threshold without requiring any input parameter.

4. Development of ASR Systems for Three Indian Languages

In this section, we elaborate the development of ASR systems for three Indian languages namely Hindi, Kannada and Tamil. In order to develop a good automated speech recognition (ASR) system, we require (i) high quality transcribed speech data in the order of several hundred hours and (ii) multi-domain text corpus containing several million words. Conventional ASR systems use graphical models like finite state transducers (FST) and stochastic and neural models like hidden Markov model (HMM) and deep neural networks (DNN) (Hinton et al., 2012). Recent end-to-end connectionist temporal classification (CTC) based techniques have been successfully applied to build large-scale ASR systems (Amodei et al., 2016). However, the size of the speech corpus required for the CTC models is much larger than that needed by the graphical models.

Due to limited data resources, we have used FST based models, which have been trained using 137, 280 and 180 hours of speech for Hindi, Kannada and Tamil, respectively. Our ASR systems have been built using the Kaldi open source toolkit (Povey et al., 2011). The building blocks of our ASR system are explained below.

4.1 Speech Data Collection and Correction

Since we require a large amount of transcribed speech corpus for training the acoustic model of the ASR, we have developed a speech recording tool that loads text prompts from our database and the volunteer reads the prompts one by one. Speech data has been collected using Sennheiser PC-8, Plantronics C320-M headphones and mobile handsets. The tool has provisions to rectify any errors in the recorded text/speech at the time of recording. Natural language processing tools for converting numbers, symbols and abbreviations have been integrated into the tool so that minimal manual effort is required to make the data ASR-ready. We have also developed an online transcript correction tool so that any errors uncorrected by the speaker while recording can be corrected at a later stage. After this, the transcribed speech data would be used by our ASR for training. Measures have been taken to ensure that the phone distribution in the spoken utterances matches that of the text corpus. Using this tool, we have collected speech from around 2500 native speakers of Hindi, Kannada and Tamil.

We have also collected a large amount of text corpus in these languages from Wikipedia articles, newspapers, magazines and books. The collected text has been pruned by removing Unicode errors and converting numerals, abbreviations and symbols. This is used to build the language model.

4.2 Design of the ASR Systems

The lexicon/pronunciation model has been created by getting all unique words from the text corpus and performing a grapheme to phoneme conversion. Schwa deletion in Hindi (Deepa et al., 2004) and voiced/unvoiced phonation rules for stop consonants in Tamil have been incorporated for better phone modeling. The lexicon model can be thought of as a map from word to phone sequences.

76 Alternate pronunciations have been included for relevant

words in the lexicon with appropriate pronunciation probabilities (Chen et al., 2015).

Using the transcribed speech, we train a series of models: monophone, triphone and DNN to get the speech transcribed at the phone-level (Madhavaraj and Ramakrishnan, 2017). Finally, the DNN model is used with the word level trigram language model for decoding. Viterbi decoding with beam search is used during testing to get the best possible sequence of words from the given speech. The size of the vocabulary of train and test data and word error rate performance of our ASR systems are given in Table 1.

Language	Training data	Test data	vocabulary size	WER
Hindi	137	45	65421	9.51
Kannada	280	67	200690	11.45
Tamil	180	54	189644	13.56

Table 1: Training and test data sizes (in hours), vocabulary size (in words) and word error rate (in %) of MILE ASR systems for Tamil, Kannada and Hindi.

5. Performance Enhancement of ASR Systems

Traditional ASR systems use mel frequency cepstral coefficients (MFCC) as speech features for acoustic modeling. These features are inspired by the human auditory mechanism and contain information about speaker identity, phone identity, stress, emotion, age and gender. However, for speech recognition, features are desired containing information only about phone identity. We have proposed two techniques: the first is a hybrid feature/model engineering technique based on scattering transform, and the second adapts the DNN model to suppress speaker variability. Both techniques succeed in extracting phonetic information as evident from the reduction in WER (see Table 2). These techniques are illustrated in the following two subsections.

Features/model	WER
Baseline MFCC	13.56
LFBE	13.48
Scattering transform (order 1)	13.16
Scattering transform (order 2)	12.36

Table 2: Comparison of word error rates (in %) of MILE Tamil ASR for different architectures and features.

5.1 Scattering Transform based Features for Better Acoustic Modeling

In this experiment, we have proposed a new DNN architecture employing a cascade of 1-D and 2-D filterbank layers which are essentially 1-D and 2-D convolution layers initialized with Gabor filter coefficients with various center frequencies and orientations. This architecture is motivated by filterbank learning techniques from raw speech waveform (Sainath, 2015) and uses scattering spectrum as front-end features (SainathScattering). The features obtained from 1-D and 2-D filterbank layers are combined and fed to a 7-layer feed-forward DNN for

predicting the phoneme labels. This architecture models the acoustic features better, since it learns the features directly from the raw waveform. Using these filterbanks for Tamil ASR, we get a relative WER reduction of 2.94% and 8.85%, respectively, compared to the baseline features as shown in Table 2. More details about this experiment can be found in (Madhavaraj and Ramakrishnan, 2019).

5.2 Speaker Adaptation using DNN Co-activation Modeling

This involves suppressing speaker-specific information contained in the speech signal and extracting features relevant only for phone identification. Here, we propose a supervised speaker adaptation technique for DNN, which estimates prior statistics of node activations in every DNN layer from the training data and adapts the weights based on the activations obtained from the adaptation data. The DNN weight update optimizes a loss function which combines cross-entropy loss and KL-divergence measure between the prior activation statistic and adaptation data's statistic. Just by modifying the loss function, we obtain an absolute WER reduction of 2.44% over the baseline model. The results of our experiments with other variants of this training strategy are listed in Table 3.

Model adaptation type	WER
Baseline architecture	13.56
Mean normalization at every DNN layer	13.48
Mean & variance normalization at every DNN layer	13.44
Mean normalization at the first affine layer of DNN	11.62
Mean & var. normalization at the first affine layer of DNN	11.12

Table 3: Comparison of performance of Tamil ASR for different speaker adaptation schemes for DNN-based acoustic models.

6. Extending the Vocabulary of ASR Systems using Subword Modeling

Handling the infinite vocabulary problem is a major task in improving the recognition accuracy of ASR systems for Tamil and Kannada. This problem arises due to morphology, inflexion and agglutination properties of the languages. Graphical model based ASRs require a finite set of words, and it is impossible to contain these languages within a finite vocabulary and build the system. To tackle this issue, we propose subword modeling, where the vocabulary contains only the subword prefixes, infixes and suffixes with proper identification markers for the language model to learn the order of subwords. The subword modeling experiments conducted for Tamil ASR are explained below.

6.1 Word Morphology based Language Modeling

In this experiment, we use the Morfessor toolkit (Smit et al., 2014) for subword modeling. Morfessor is a statistical

machine learning tool used for morphological analysis to segment a given word. These subwords are used as basic units in our lexicon for recognition. We learn the language model by converting the training corpus containing words into a sequence of subword tokens. The lexicon preparation for the subword dictionary is a straightforward task for Indian languages, since they have an almost one-to-one correspondence between graphemes and phonemes. The rest of the ASR systems are built as explained in (Madhavaraj and Ramakrishnan, 2017). During post-processing, the subwords are joined into words using the identification markers in the recognized text. We obtain a WER comparable to that of the word-level ASR as reported in Table IV.

6.2 Maximum Likelihood based Language Modeling

This uses byte pair encoding (Sennrich et al., 2016) to perform subword modeling, where the list of subwords, their occurrence and co-occurrence probabilities we are derived and based on a specifically designed subword finite state transducer (FST), the most likely segmentation for a given word is estimated. Since segmenting a given word into subwords is a combinatorial explosion problem, maximum likelihood (ML) estimation is employed through expectation-maximization procedure and this problem is posed as a weighted-FST (WFST) graph search problem. Two different methods are proposed for ML estimation, namely Viterbi and forward-backward techniques, whose performances are listed in Table 4.

6.3 Manual Modeling

The performances of both morphology-based and ML-based techniques depend highly on the quality and diverseness of the corpus. Yet, there is a high probability that some of the subwords derived by these techniques may not be valid prefixes, infixes or suffixes. So, we manually construct the lexicon graph for Tamil words for different nouns, pronouns, adjectives, adverbs, numbers, verbs and infinitives. The ASR now uses a lexicon of only hand-labeled subword units for recognition. We have also created a provision to add new words into this lexicon graph which cannot be analyzed morphologically. One advantage with this modeling is that the lexicon graph can be readily used for many NLP tasks such as part of speech tagging, lemmatization and text translation.

7. Conclusion

The commitments and contributions of MILE laboratory over the past two decades in performing fundamental research in technologies for Indian languages have been described briefly. All the data we use have been collected by us: India has a huge population and so, there is no dearth for creation of standard databases.

8. Acknowledgements

The first author gratefully acknowledges Tata Trust Travel Grant for funding him to travel and participate in this conference. Immense thanks are also due to the Technology Development for Indian Languages (TDIL), Ministry of Information Technology, Government of India, for funding

many of his projects in language technology, which has made it possible for him to be invited to this conference. Thanks are also due to many students and research staff, who enriched his knowledge and experience.

9. Bibliographical References

- Geoffrey Hinton, li Deng, Dong Yu, George Dahl, Abdelrahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Phuongtrang Nguyen, Tara Sainath, and Brian Kingsbury. (2012). Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. *IEEE Signal Processing Magazine*, vol. 29, pp. 82–97.
- Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen. (2016). Deep speech 2: End-to-end speech recognition in English and Mandarin. Proc. 33rd International Conf. on Machine Learning (ICML) Vol. 48, p. 173–182.
- Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely. (2011). The Kaldi speech recognition toolkit. Proc. IEEE Workshop Automatic Speech Recog. Understanding.
- S.R. Deepa, Kalika Bali, A.G. Ramakrishnan, and Partha Pratim Talukdar. (2004). Automatic generation of compound word lexicon for Hindi speech synthesis. Proc. Fourth International Conf. on Language Resources and Evaluation (LREC'04), May, ELRA.
- Guoguo Chen, Hainan Xu, Minhua Wu, Daniel Povey, and Sanjeev Khudanpur. (2015). Pronunciation and silence probability modeling for ASR,” Proc. 16th Interspeech.
- A. Madhavaraj and A. G. Ramakrishnan (2017). Design and development of a large vocabulary, continuous speech recognition system for Tamil. Proc. 14th IEEE INDICON, Dec, pp. 1–5.
- Tara N. Sainath, Ron J. Weiss, Andrew W. Senior, Kevin W. Wilson, and Oriol Vinyals. (2015). Learning the speech front-end with raw waveform CLDNNs. Proc. 16th Interspeech.
- V. Peddinti, T. Sainath, S. Maymon, B. Ramabhadran, D. Nahamoo, and V. Goel. (2014). Deep scattering spectrum with deep neural networks. Proc IEEE ICASSP, pp. 210–214.
- A. Madhavaraj and A. G. Ramakrishnan. (2019). Scattering transform inspired filterbank learning from raw speech for better acoustic modeling. Proc. IEEE Region 10 Conf. (TENCON), Oct, pp. 1154–1158.
- Peter Smit, Sami Virpioja, Stig-Arne Gronroos, and Mikko Kurimo. (2014). Morfessor 2.0: Toolkit for statistical morphological segmentation. Proc. Demonstrations 14th Conf. of European Chapter of the Association for Computational Linguistics. Apr., pp. 21–24, ACL.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. (2016). Neural machine translation of rare words with subword units. Proc. 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Germany, Aug., pp. 1715–1725, ACL.
- Shiva Kumar H R and A G Ramakrishnan. (2020). Lipi Gnani - A versatile OCR for documents in any language printed in Kannada script. ACM Transactions on Asian and Low-Resource Language Info Processing (TALLIP).