



INDIAN LANGUAGE RESOURCES— SPEECH SUBCOMMITTEE REPORT

Authored by

IEEE SA Speech Subcommittee

TRADEMARKS AND DISCLAIMERS

IEEE believes the information in this publication is accurate as of its publication date; such information is subject to change without notice. IEEE is not responsible for any inadvertent errors.

The ideas and proposals in this specification are the respective author's views and do not represent the views of the affiliated organization.

The Institute of Electrical and Electronics Engineers, Inc. 3 Park Avenue, New York, NY 10016-5997, USA

Copyright © 2023 by The Institute of Electrical and Electronics Engineers, Inc.

All rights reserved. 18 September 2023. Printed in the United States of America.

PDF: STDVA26427 979-8-8557-0066-4

IEEE is a registered trademark in the U. S. Patent & Trademark Office, owned by The Institute of Electrical and Electronics Engineers, Incorporated. All other trademarks are the property of the respective trademark owners.

IEEE prohibits discrimination, harassment, and bullying. For more information, visit <http://www.ieee.org/web/aboutus/whatis/policies/p9-26.html>.

No part of this publication may be reproduced in any form, in an electronic retrieval system, or otherwise, without the prior written permission of the publisher.

Find IEEE standards and standards-related product listings at: <http://standards.ieee.org>.

NOTICE AND DISCLAIMER OF LIABILITY CONCERNING THE USE OF IEEE SA INDUSTRY CONNECTIONS DOCUMENTS

This IEEE Standards Association (“IEEE SA”) Industry Connections publication (“Work”) is not a consensus standard document. Specifically, this document is NOT AN IEEE STANDARD. Information contained in this Work has been created by, or obtained from, sources believed to be reliable, and reviewed by members of the IEEE SA Industry Connections activity that produced this Work. IEEE and the IEEE SA Industry Connections activity members expressly disclaim all warranties (express, implied, and statutory) related to this Work, including, but not limited to, the warranties of: merchantability; fitness for a particular purpose; non-infringement; quality, accuracy, effectiveness, currency, or completeness of the Work or content within the Work. In addition, IEEE and the IEEE SA Industry Connections activity members disclaim any and all conditions relating to: results; and workmanlike effort. This IEEE SA Industry Connections document is supplied “AS IS” and “WITH ALL FAULTS.”

Although the IEEE SA Industry Connections activity members who have created this Work believe that the information and guidance given in this Work serve as an enhancement to users, all persons must rely upon their own skill and judgment when making use of it. IN NO EVENT SHALL IEEE OR IEEE SA INDUSTRY CONNECTIONS ACTIVITY MEMBERS BE LIABLE FOR ANY ERRORS OR OMISSIONS OR DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO: PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS WORK, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE AND REGARDLESS OF WHETHER SUCH DAMAGE WAS FORESEEABLE.

Further, information contained in this Work may be protected by intellectual property rights held by third parties or organizations, and the use of this information may require the user to negotiate with any such rights holders in order to legally acquire the rights to do so, and such rights holders may refuse to grant such rights. Attention is also called to the possibility that implementation of any or all of this Work may require use of subject matter covered by patent rights. By publication of this Work, no position is taken by the IEEE with respect to the existence or validity of any patent rights in connection therewith. The IEEE is not responsible for identifying patent rights for which a license may be required, or for conducting inquiries into the legal validity or scope of patents claims. Users are expressly advised that determination of the validity of any patent rights, and the risk of infringement of such rights, is entirely their own responsibility. No commitment to grant licenses under patent rights on a reasonable or non-discriminatory basis has been sought or received from any rights holder. The policies and procedures under which this document was created can be viewed at <http://standards.ieee.org/about/sasb/iccom/>.

This Work is published with the understanding that IEEE and the ICom members are supplying information through this Work, not attempting to render engineering or other professional services. If such services are required, the assistance of an appropriate professional should be sought. IEEE is not responsible for the statements and opinions advanced in this Work.

TABLE OF CONTENTS

ABSTRACT.....	6
1. INTRODUCTION	7
2. SCOPE	7
3. TECHNOLOGY COMPONENTS OF A SPEECH SOLUTION.....	9
3.1. VOICE ACTIVITY DETECTION	9
3.1.1. APPROACHES	9
3.1.2. EVALUATION METRIC	10
3.2. LANGUAGE IDENTIFICATION	10
3.2.1. APPROACHES	10
3.2.2. EVALUATION METRIC	11
3.3. SPEECH ENHANCEMENT	11
3.3.1. APPROACHES	12
3.3.2. EVALUATION METRICS.....	12
3.4. SPEAKER DIARIZATION	12
3.4.1. APPROACHES	12
3.4.2. EVALUATION METRIC	13
3.5. SPEECH RECOGNITION.....	13
3.5.1. APPROACHES	13
3.5.1.1. CASCADED MODELS.....	13
3.5.1.2. END-TO-END MODELS.....	13
3.5.2. DATA REQUIREMENT	14
3.5.3. EVALUATION METRIC	14
3.6. SPEECH SYNTHESIS	14
3.6.1. APPROACHES	15
3.6.1.1. PARAMETRIC SPEECH SYNTHESIS	15
3.6.1.2. END-TO-END MODELS.....	15
3.6.2. EVALUATION METRIC	16
3.7. SPEAKER IDENTIFICATION AND VERIFICATION.....	16
3.7.1. APPROACHES	16
3.7.2. EVALUATION METRICS.....	16
3.8. SPEAKER PROFILING	17
3.8.1. APPROACHES	17
3.8.2. EVALUATION METRIC	17

TABLE OF CONTENTS

4. SPEECH USE CASES.....	18
4.1. USE CASE 1: FINANCIAL SERVICES	18
4.2. USE CASE 2: CUSTOMER SUPPORT AT CONTACT CENTERS	21
5. SURVEY OF CURRENT SPEECH STANDARDS.....	25
5.1. EXISTING W3C STANDARDS FOR SPEECH TECHNOLOGIES	25
5.2. STANDARDS FOR EUROPEAN LANGUAGES.....	26
5.3. STANDARDS AVAILABLE FOR INDIAN LANGUAGES	27
5.3.1. TEXT-TO-SPEECH TESTING STRATEGY VERSION 2.1.....	27
5.3.2. SPEECH TRANSCRIPTION GUIDELINES FOR INDIAN LANGUAGES	27
5.3.3. TESTING, EVALUATION, AND BENCHMARKING FOR SPEECH-BASED ACCESS FOR AGRICULTURAL COMMODITY PRICES IN SIX INDIAN LANGUAGES	28
5.4. LIST OF SPEECH ASSETS AVAILABLE FOR INDIAN LANGUAGES.....	28
6. GAPS IN STANDARDS FOR INDIAN LANGUAGES.....	30
6.1. GAPS IN W3C STANDARDS SPECIFIC TO INDIAN LANGUAGES	30
6.2. GAPS IN EVALUATION STANDARDS SPECIFIC TO INDIAN LANGUAGES	30
6.3. PRIVACY REQUIREMENTS	32
6.3.1. PRIVACY POLICY IN SPEECH	32
7. REFERENCES.....	33

SPEECH SUBCOMMITTEE REPORT

ABSTRACT

This report summarizes the findings and recommendations of the speech subcommittee of the IEEE prestandardization effort on standards for Indian language resources. Speech processing technology is all about recognizing spoken words. Technologies like language detection, speech transcription, speech synthesis, etc., are part of speech processing technology. In the context of the official Indian languages, there is an identified gap in the status of such technologies for Indian languages. The report focuses exclusively on official Indian languages. It summarizes the available standards and practices, key use cases driving the deployment of such technology, priority gaps that need to be addressed to promote the adoption of speech technologies, resources available, and resources required toward the development of speech technologies, current metrics, and datasets used for evaluation of various technologies and identifies gaps that need to be addressed.

1. INTRODUCTION

Language-enabled products are gaining wide traction in every single sector of the economy. It is acknowledged that the next billion internet users in India will come because of the localization of content in Indian languages. Automatic translation, speech, and text recognition, context-aware automation and extraction, etc., are the key use cases driving the deployment. The biggest challenge, and hence an area of active innovation for Indian languages, is the reality that most of these languages are under-resourced. The challenge opens a big opportunity for standardization and fits very nicely with why IEEE drives standardization activities. Standardization here will spur innovation, fuel new growth of technology consumers, expand customer choice, reduce cost structures, and support interoperability between multiple languages among other benefits.

Identification of relevant use cases that drive the deployment of technology, gaps in standardization that need to be addressed for wider adoption of technology, and providing proposals justifying the need for standards to govern language resources for Indian languages is the motivation of this effort. Language resources include speech, text, language data, and language descriptions which are made available in a machine-readable form and used for developing solutions, metrics used for evaluations of solutions, and algorithms in natural language and speech processing. Such standards can also be used for language studies, localization of software, electronic publishing and any purpose for researchers, subject area specialists, etc. Examples of language resources are spoken and written corpora, computational lexica, terminology databases, etc. The work will also include any tools that are used for the above-stated purposes.

The overall investigation to identify the gaps in standards is organized into multiple subgroups with each subgroup looking into specific areas of language resources. This document represents the summary report from the language processing (speech) workgroup following investigations in speech resources. The document is not intended to be a standalone document but will be an input into the final report coming from the overall investigation. This document will also serve as input for other subgroups to understand the key findings from the speech workgroup and, in turn, use it for both their own consumption as well as provide valuable feedback to the work of the speech workgroup.

2. SCOPE

The document represents the output from the speech workgroup covering details of use cases, the lay of the land for standards, and the gaps that exist in speech-language resources for Indian languages. The scope of

Indian languages covered in this document is limited to the 22 officially recognized scheduled languages of the country (explicitly listed below) along with the acknowledged different dialects of the same language. The output of the document will cover the following areas:

1. Dominant use cases drive the deployment of technology in speech-language for the scheduled Indian languages, and the majority of which belong to the Indo-Aryan or Dravidian language families. The use cases can both be driven from public utility and industry perspectives.
2. List of speech-language resources required for the identified use cases for the officially recognized scheduled languages of India: Assamese, Bengali, Bodo, Dogri, Gujarati, Hindi, Kannada, Kashmiri, Konkani, Maithili, Malayalam, Manipuri, Marathi, Nepali, Odia, Punjabi, Sanskrit, Santali, Sindhi, Tamil, Telugu, and Urdu.
3. Identification of required standards based on the use cases driving the deployment. This should lead to a detailed description of the gaps that exist. The study can incorporate results from similar undertakings from around the world. A good example is the study undertaken by the European Union, which the authors will use as a reference for their work.
4. Detailed listing of the available resources for speech resources for the listed official languages. The listing should provide a view of the conformance of such resources to the required standards. A good representation of the same is the language matrix as used by W3C [1]. The resources should also include any computer code for the identified use cases.
5. The document can also make a proposal for required workshops, conferences, etc., for detailed discussions, deliberations, trigger research initiatives, contests, etc., to promote the necessary investigations to expose the gaps.

A critical disclaimer is that this study is only about the standard gaps and will not make any effort to define the required standards. The authors do not make any commitments that the work will equally apply to other nonofficial languages of India.

Before delving into details of the above-mentioned points, the authors will first outline the main technology components typically implemented in speech-based solutions. This will highlight the main requirements and evaluation metrics used for various speech-based components which will further help contextualize the need for standardization.

3. TECHNOLOGY COMPONENTS OF A SPEECH SOLUTION

A speech solution is essentially a combination of various speech technology components integrated to form a complete solution. The predominantly used speech-based technology components are listed in 3.1–3.8.

3.1. VOICE ACTIVITY DETECTION

Voice activity detection (VAD) is an integral part of different speech communication systems that require real-time audio-processing capabilities. It refers to the task of determining whether an audio signal contains speech or not. Depending on the nature of the data, an audio signal can generally consist of speech and often have pauses or nonspeech sounds between sentences or even within sentences. Using VAD as a preprocessing step helps the subsequent modules process speech and nonspeech content separately, thus allowing for the possibility of better performance on downstream tasks.

3.1.1. APPROACHES

The simplest and most straightforward approach to determine the voicing activity of the signal is to use the signal energy. The time-varying energy of the speech signal can be used as an indicator of speech. Due to the nonstationary nature of speech, a windowed region of speech is chosen, and time-varying energy is estimated for the entire audio. Thereafter, a threshold is chosen such that the high-energy regions above the threshold of the signal are likely to be speech and the remaining can be pauses or silences.

While this approach is very simple, it is not easy to identify a threshold that works well broadly. In cases of whispered speech, noise, etc., the threshold setting becomes difficult and is prone to error and can affect the accuracy of the following speech modules. Other approaches that include thresholding are as follows:

- Estimation of zero crossing rate.
- Fundamental frequency.
- Autocorrelation.

More complex approaches consist of training a classifier that can effectively discriminate between the speech and nonspeech parts of the audio waveform.

3.1.2. EVALUATION METRIC

Evaluation of the performance of VAD systems requires ground truth labels explicitly listing both speech and nonspeech segments. Thus, errors are when speech is identified as nonspeech and vice versa. Therefore, the VAD performance can be estimated in terms of the percentage of true positives and true negatives and/or false negatives and false positives. The threshold can be set where the error rate is equal, i.e., at the equal error rate (EER) point, or can be set depending on the ability of the subsequent downstream application to either handle true positives or false negatives.

3.2. LANGUAGE IDENTIFICATION

Automatic spoken language identification (LID) refers to identifying the language being spoken. For a known and familiar language, humans can identify a language within seconds after hearing an utterance. For an unknown or unfamiliar language, humans can make subjective judgments based on the similarities to a language that they know, e.g., “sounds like Marathi.” An ideal LID system uses different aspects of speech information that can discriminate different languages from a huge amount of target languages, and flexible enough to accommodate the variations of different speakers. The LID task has applications in multilingual services, e.g., routing an incoming telephone call to a human operator knowing the corresponding language or the audio can be redirected to other language-specific speech technologies such as speech recognition.

At a low level, speech features such as acoustic, phonetic, phonotactic, and prosodic information are widely used in LID tasks. At a higher level, the difference between languages can be exploited, based on morphology and sentence syntax. An ideal LID system should be robust to speaker and channel variations and other noise.

3.2.1. APPROACHES

Language recognition is like speaker recognition in terms of problem formulation, system approaches, and methodologies, including modeling, classification and fusion strategies, and system evaluation measures.

Traditionally, phoneme recognition followed by language modeling (PRLM) was one of the popular methods that use a multilingual phoneme recognizer to generate phoneme sequences which are converted to language model (n-gram) features for the LID classifier. The acoustic information is generally considered the first level of analysis of speech production. The most widely used parameterization techniques are linear prediction, mel frequency cepstral coefficient (MFCC), perceptual linear prediction (PLP), and linear prediction cepstral

coefficient (LPCC) features. Once the basic acoustic features have been obtained, additional features are appended to each feature vector with the intention of incorporating the temporal aspects of the speech signal. Some commonly utilized additional features are the delta and acceleration cepstrum and the shifted delta cepstrum (SDC). These features have been found to be beneficial for LID.

3.2.2. EVALUATION METRIC

LID evaluations are carried out using varying test data lengths. LID systems are known to have good performance for test data >30 s, and as the duration of the test data reduces, the LID performance degrades. The National Institute of Standards and Technology (NIST) has conducted a series of evaluations of LID technology in 1996, 2003, 2005, 2007, and 2009. Language recognition evaluations (LREs) focus on language and dialect detection in the context of conversational telephony speech. NIST LRE is formulated as a language detection or verification task, where each trial is a hypothesis test as to whether the test sample belongs to a claimed language identity. Therefore, there are two types of errors, i.e., detection miss in which a true claim is denied, and false alarm in which a false claim is accepted. The primary evaluation metric is taken as the average cost performance (Cavg), which indicates the pair-wise language recognition performance, represented as a function of the detection miss and false alarm probabilities for all target/nontarget language pairs. In addition to Cavg, the authors also compute the average EER for each of the target languages and take their average as the performance measure.

3.3. SPEECH ENHANCEMENT

In many real-case scenarios, the speech signals are degraded by the ambient interfering noises present in the environment while recording. The noise can be either stationary or nonstationary in nature. This can cause a loss of quality and intelligibility, and it is known that the performance of speech-processing techniques degrades in noisy scenarios. Therefore, enhancement of speech is normally required to reduce or decrease the distortion in the speech signal and to improve the perceptual aspects of the speech, i.e., mainly to improve the quality and/or intelligibility of the speech signal. It is observed that denoising or enhancement improves the intelligibility and the quality of the recorded signal. However, in certain cases, enhancement introduces distortions in the denoised speech and this residual noise is unseen by the following speech modules unless these modules are specifically retrained with the enhanced speech [i.e., signal-distortion-independent training (SDIT)].