IEC TR 63558

Edition 1.0 2025-01

TECHNICAL REPORT

Automatic speech recognition: Classification according to acoustic and linguistic indicators in real-life applications

IEC TR 63558:2025-01(en)

IEC Secretariat
3, rue de Varembé
CH-1211 Geneva 20
Switzerland

Tel.: +41 22 919 02 11
info@iec.ch
www.iec.ch

About the IEC
The International Electrotechnical Commission (IEC) is the leading global organization that prepares and publishes International Standards for all electrical, electronic and related technologies.

About IEC publications
The technical content of IEC publications is kept under constant review by the IEC. Please make sure that you have the latest edition, a corrigendum or an amendment might have been published.

IEC publications search - webstore.iec.ch/advsearchform
The advanced search enables to find IEC publications by a variety of criteria (reference number, text, technical committee, …). It also gives information on projects, replaced and withdrawn publications.

IEC Just Published - webstore.iec.ch/justpublished
Stay up to date on all new IEC publications. Just Published details all new publications released. Available online and once a month by email.

IEC Customer Service Centre - webstore.iec.ch/csc
If you wish to give us your feedback on this publication or need further assistance, please contact the Customer Service Centre: sales@iec.ch.

IEC Products & Services Portal - products.iec.ch
Discover our powerful search engine and read freely all the publications previews, graphical symbols and the glossary. With a subscription you will always have access to up to date content tailored to your needs.

Electropedia - www.electropedia.org
The world's leading online dictionary on electrotechnology, containing more than 22 500 terminological entries in English and French, with equivalent terms in 25 additional languages. Also known as the International Electrotechnical Vocabulary (IEV) online.

IEC TR 63558

Edition 1.0 2025-01

TECHNICAL REPORT

Automatic speech recognition: Classification according to acoustic and linguistic indicators in real-life applications

INTERNATIONAL
ELECTROTECHNICAL
COMMISSION

ICS 33.160.01

ISBN 978-2-8327-0129-4

CONTENTS

INTERNATIONAL ELECTROTECHNICAL COMMISSION

_____

AUTOMATIC SPEECH RECOGNITION: CLASSIFICATION ACCORDING TO ACOUSTIC AND LINGUISTIC INDICATORS IN REAL-LIFE APPLICATIONS

FOREWORD

1) The International Electrotechnical Commission (IEC) is a worldwide organization for standardization comprising all national electrotechnical committees (IEC National Committees). The object of IEC is to promote international co-operation on all questions concerning standardization in the electrical and electronic fields. To this end and in addition to other activities, IEC publishes International Standards, Technical Specifications, Technical Reports, Publicly Available Specifications (PAS) and Guides (hereafter referred to as "IEC Publication(s)"). Their preparation is entrusted to technical committees; any IEC National Committee interested in the subject dealt with may participate in this preparatory work. International, governmental and non-governmental organizations liaising with the IEC also participate in this preparation. IEC collaborates closely with the International Organization for Standardization (ISO) in accordance with conditions determined by agreement between the two organizations.
2) The formal decisions or agreements of IEC on technical matters express, as nearly as possible, an international consensus of opinion on the relevant subjects since each technical committee has representation from all interested IEC National Committees.
3) IEC Publications have the form of recommendations for international use and are accepted by IEC National Committees in that sense. While all reasonable efforts are made to ensure that the technical content of IEC Publications is accurate, IEC cannot be held responsible for the way in which they are used or for any misinterpretation by any end user.
4) In order to promote international uniformity, IEC National Committees undertake to apply IEC Publications transparently to the maximum extent possible in their national and regional publications. Any divergence between any IEC Publication and the corresponding national or regional publication shall be clearly indicated in the latter.
5) IEC itself does not provide any attestation of conformity. Independent certification bodies provide conformity assessment services and, in some areas, access to IEC marks of conformity. IEC is not responsible for any services carried out by independent certification bodies.
6) All users should ensure that they have the latest edition of this publication.
7) No liability shall attach to IEC or its directors, employees, servants or agents including individual experts and members of its technical committees and IEC National Committees for any personal injury, property damage or other damage of any nature whatsoever, whether direct or indirect, or for costs (including legal fees) and expenses arising out of the publication, use of, or reliance upon, this IEC Publication or any other IEC Publications.
8) Attention is drawn to the Normative references cited in this publication. Use of the referenced publications is indispensable for the correct application of this publication.
9) IEC draws attention to the possibility that the implementation of this document may involve the use of (a) patent(s). IEC takes no position concerning the evidence, validity or applicability of any claimed patent rights in respect thereof. As of the date of publication of this document, IEC had not received notice of (a) patent(s), which may be required to implement this document. However, implementers are cautioned that this may not represent the latest information, which may be obtained from the patent database available at https://patents.iec.ch. IEC shall not be held responsible for identifying any or all such patent rights.

IEC TR 63558 by IEC technical committee 100: Audio, video and multimedia systems and equipment. It is a Technical Report.

The text of this Technical Report is based on the following documents:

Draft	Report on voting
100/4214/DTR	100/4263/RVDTR

Full information on the voting for its approval can be found in the report on voting indicated in the above table.

The language used for the development of this Technical Report is English.

This document was drafted in accordance with ISO/IEC Directives, Part 2, and developed in accordance with ISO/IEC Directives, Part 1 and ISO/IEC Directives, IEC Supplement, available at www.iec.ch/members_experts/refdocs. The main document types developed by IEC are described in greater detail at www.iec.ch/publications.

The committee has decided that the contents of this document will remain unchanged until the stability date indicated on the IEC website under webstore.iec.ch in the data related to the specific document. At this date, the document will be

- reconfirmed,
- withdrawn, or
- revised.

INTRODUCTION

With the development of network and information technology, people are relying more and more on smart equipment, such as smart speakers, smart service robots and so on. Speech recognition technology is the main means to realize man-machine communication. Speech recognition is the process of converting a voice into digital data. Popular use in recent years has pushed improvements in its algorithm and increased accuracy. But the performance of different speech recognition solutions differs greatly and is sometimes a source of confusion for users. The factors used to evaluate the performance of speech recognition technology need more discussion.

This document mainly aims to set up a set of parameters which can be used to reflect the complexity of real-life applications, by means of a classification using scenarios.

AUTOMATIC SPEECH RECOGNITION: CLASSIFICATION ACCORDING TO ACOUSTIC AND LINGUISTIC INDICATORS IN REAL-LIFE APPLICATIONS

1 Scope

This document describes the factors related to classification of the real-life environment according to acoustic indicators and linguistic indicators. The set of factors can be used to describe complexities of use scenarios, from level 1 to 4, and can be helpful when setting up the testing environment.

This document applies for evaluating automatic speech recognition technology which is widely used for smart equipment, such as smart speakers.

2 Normative references

There are no normative references in this document.

3 Terms and definitions

For the purposes of this document, the following terms and definitions apply.

ISO and IEC maintain terminology databases for use in standardization at the following addresses:

- IEC Electropedia: available at https://www.electropedia.org/
- ISO Online browsing platform: available at https://www.iso.org/obp

3.1
automatic speech recognition
ASR
process of converting human voice signals into digital data

NOTE The output of the speech recognition process is a unique series of data, which can match a predefined word set then make the machine "understand" the meaning of speaker.

3.2
word set
all the words which can be processed by the speech recognition system

NOTE A word set may contain several languages.

3.3
discrete speech recognition
speaker pronounces each word separately, inserting pauses between each one

EXAMPLE When the sentence "Good to know" is pronounced as [gud], [tu] and [neu], just like three separate words, this way of speaking is regarded as separated word speech. For a beginner learning a foreign language, this way of speaking is normal.

3.4
continuous speech recognition
speaker's speech rate is close to normal, with no obvious intentional pauses between each word

EXAMPLE When the sentence "Good to know" is pronounced as [gu-tu neu], there is no clear pause between "Good" and "to"; this way of speaking is like natural speech.

3.5
linguistic feature
phonetic feature of non-segment in speech and various non-speech signals

EXAMPLE Almost everyone has their own feature words or phrases in daily speaking, such as "well", "OK", "you know", etc.

4 Use of automatic speech recognition (ASR) system

ASR is a high-tech system by which a machine converts the speech signal into corresponding text or commands after recognizing and understanding the speech signal. It includes the extraction and determination of the acoustic feature, the acoustic model, and the language model.

A wide number of industries are utilizing different applications of speech technology today, helping businesses and consumers save time and even lives. Some examples include:

– Automotive: ASR system improves driver safety by enabling voice-activated navigation systems and search capabilities in car radios, instead of using hands. It can make driving both more convenient and safer.

– Digital assistant: Virtual digital assistant is increasingly used in our daily lives, particularly on our mobile devices. Voice commands are used to access applications in smartphones, such as voice searching, or playing music. Survey statistics [1][1] show that voice search is the most common use of ASR, in 2022, in the United States alone, 135,6 million users used a digital assistant at least once a month.

– Healthcare: Doctors and nurses leverage dictation applications to capture and log patient diagnoses and treatment notes. It saves doctors a significant amount of time and allows them to spend more time with patients and reduce the mistakes made when hand-written diagnoses are misunderstood.

– Sales: ASR has a couple of applications in sales. It can help a call center transcribe thousands of phone calls between customers and agents to identify common call patterns and issues. AI chatbots can also talk to people via a webpage, answering common queries and solving basic requests without needing to wait for a contact center agent to be available. It both instances speech recognition systems help reduce time to resolution for consumer issues.

NOTE In this case, communication between customer and call center sometimes using text instead of talking.

– Security: Voice-based authentication adds a viable level of security. Combine this with other identification methods, such as fingerprint and face recognition, to increase the security level.

5 Needs for standards when using ASR

Currently, when ASR is used in commercial situations, accuracy is still the key factor. And there are many discussions on interfaces between users and ASR models.

1 Numbers in square brackets refer to the Bibliography.

When considering standards at the application level for ASR systems, problems in real use cases are discussed first.

Rapidly developed ASR technology makes it possible for almost all smart devices to be equipped with ASR functions, such as smart phones, smart speaks, smart televisions, smart home facilities, etc. But on the other hand, users are sometimes unsatisfied with the performance of such smart devices. Accuracy is the key problem, and complaints about misunderstanding are often heard.

Concerns around the challenge of accuracy, which is the top priority, is addressed for all ASR systems. According to a survey published by AIMultiple Research [2], 73 % of users raised the issue of accuracy. Even if in a laboratory setting, the ASR system can respond well to the testing content, in a real-life scenario, the performance will be far worse. The difference in performance is related to the differences between real-world conditions and the laboratory environment.

When trying to improve the accuracy of an ASR model, the feature of the background noise can be a significant barrier. When the system is exposed to real-world situations, there is a lot of background noise, such as crosstalk, white noise, and other distortions that can reduce its accuracy.

To overcome the challenge of ASR accuracy, it is important to know the user's environment before developing the ASR model. Specifically, this means having a reference testing environment with typical background noise features.

6 Definition and classification of factors affecting speech recognition

6.1 General

Considering the complexity for AI based automatic speech recognition, the real-life applications can be divided into four levels as follows:

L1: Limited sentence pattern (such as discrete speech recognition) and high acoustic quality.

L2: Unrestrained syntactical structure and high acoustic quality.

L3: Unrestrained syntactical structure with a small number of difficult entries (such as multilingual hybrid words, continuous speech recognition, etc.), and high-quality acoustic scene.

L4: Unrestrained syntactical structure with many difficult entries, and low-quality acoustic scene.

6.2 Acoustic indicators

6.2.1 Signal-to-noise ratio (SNR)

A higher SNR will help the speech recognition system to identify the relevant content. For example, the speech recognition system will have a higher level of performance in an ideal laboratory than in a real-life use scenario, such as in a market, street or some other public place.

6.2.2 Reflections

In physics, reflection is defined as the change in the direction of a wavefront at the interface between two different media, bouncing the wavefront back into the original medium. A common example of reflection is reflected light from a mirror or a still pool of water, but reflection affects other types of waves beside light. Water waves, sound waves, particle waves, and seismic waves can also be reflected.

In this document, reflection refers to sound waves derived from the sound source. When combined with noise, it will be more difficult for the ASR system to deal with reflection.

6.2.3 Reverberation

A reverberation, or reverb, is produced when a sound or signal is reflected producing many reflections to form and then decay as the sound is absorbed by the objects in the relevant space, which could contain people and material objects, including air. This is clearly experienced when the sound origin stops but the reflections continue, reducing the amplitude gradually to zero.

Reverberation is frequency dependent on reverberation time or the length of the decline.

6.2.4 Data compression

Data compression is the key factor related to the performance of speech recognition systems, especially for system delay.

When the data compression rate is low, speech recognition systems can work well, but the performance will decrease with a high data compression rate.

6.3 Linguistic indicators

6.3.1 Unrestrained syntactical structure

Speech recognition uses the basic framework of pattern identification, divided into data preparation, feature extraction, model training, test application, and a well-trained language model reflects whether the identification result meets the linguistic features and reflects semantic information.

Therefore, it is difficult to adjust a trained ASR model to adopt linguistic changes when using a changed scenario. A better approach is to select an appropriate training set and to test the set based on a deep understanding of the use scenario.

6.3.2 Vocabulary list size

Vocabulary list size will affect the identification ability of the speech recognition system in a real-life use case. Usually, the larger the vocabulary list size, the better the ability to adapt to the complexities of a real-life scenario.

6.3.3 Homonyms

Homonyms appear frequently in every language system. The same pronunciation can signify different words with different meanings. The performance in dealing with homonyms depends on the size of the vocabulary list and the ability to understanding the sentence's meaning.

6.3.4 Multilingual words

With the trend of cultural integration all over the word, a user can use multiple language systems to make their own language system. In this case, multilingual words will appear frequently, which can also require the speech recognition system to have the ability to identify and process such words. Multiple vocabulary lists can be a reasonable solution.

6.3.5 Speaking speed

Speaking speed will be a challenge for speech recognition systems, the higher the speaking speed is, the more difficult it is for the system to achieve performance requirements, especially for latency.

6.3.6 Accent

The accent will differ greatly due to distinctive modes of pronunciation in a language, especially when associated with a particular country, area, or social class.

It depends on the way of developing the speech recognition system, as to what kind of data is selected to train the recognition model. If only standard or normal accent data is selected, the system sometimes cannot work well in a scenario where multiple accents are present.

6.3.7 Speaking behavior

The user's speaking behavior will affect the performance of the speech recognition system. Speech behaviors such as mumbling, stammering, unconventional pauses, etc., can affect the system's recognition ability.

See Annex A for a sample of an indicator set for classification.

7 Existing International Standards

7.1 Inside IEC

IEC Technical Committee 29 prepares documents related to instruments and methods of measurement in the field of electroacoustics. It defines the requirements for devices used in acoustic testing, such as in the IEC 61094 series, IEC 61230, and IEC 61252.

In IEC Technical Committee 100, TA 20 focuses on developing standards with analogue and digital audio, such as the IEC 60268 series. An International Standard for micro-speakers is also developed in TC 100 as the IEC 61034 series.

7.2 In ISO/IEC Joint Technical Committee 1

JTC 1/SC 35 develops International Standards in the field of user-system interfaces in information and communication technology (ICT) environments and support for these interfaces to serve all users, including people having accessibility or other specific needs, with a priority of meeting the JTC 1 requirements for cultural and linguistic adaptability.

ISO/IEC 24661 covers information technology-user interfaces-full duplex speech interaction and was published in 2022. ISO/IEC 24661 specifies user interfaces (UIs) designed for full duplex (FDX) speech interaction. It also specifies the FDX speech interaction model, features, functional components and requirements, thus providing a framework to support natural conversational interfaces between humans and machines. It also provides privacy considerations for applying FDX speech interaction.

ISO/IEC 24661 is applicable to UIs for speech interaction and communication protocols for setting up a session-oriented FDX interaction between humans and machines.

8 Potential items in TC 100

As discussed in this document, setting up a reference testing set is necessary to evaluate the accuracy of the ASR. The reference testing set includes the following:

– Feature of background noise.
– Allocation of device under test and noise source.
– Reference testing contents.

Annex A
(informative)

A sample of an indicator set for classification

A sample classification for an acoustic scenario where automatic speech recognition system really works; see Table A.1 and Table A.2.

Table A.1 – Acoustic indicators and classification

Influence factors			Classification			
			L1	L2	L3	L4
Acoustics	SNR	> 15 dB	√			
		> 10 dB		√		
		> 5 dB			√	
		≤ 5 dB				√
	Data compression	Uncompressed or opus format compressed to bitrate ≥ 16 kbps opus	√	√	√	
		Heavily compressed acoustic data: opus format compressed to bitrate < 16 kbps or other format				√

Table A.2 – Linguistic indicators and classification

Influence factors			Classification			
			L1	L2	L3	L4
Linguistics	Unrestrained syntactical structure	No	√			
		Yes		√	√	√
	Vocabulary size	≤ 1 000	√			
		> 1 000		√	√	√
	Sentence length	≥ 3	√	√		
		≤ 2			√	√
	Similar-sounding words	Yes				√
		No	√	√	√	
	Multilingual hybrid words	Yes (Multilingual finite entries)			√	√
		No	√	√		
	Speaking behavior	Good speaking behavior	√	√	√	
		Behaviors which do not conform to semantic norms, e.g. repetition of single characters or words, long pauses, stuttering				√
	Speed	Moderate speed, 3 to 6 words per second	√	√	√	
		Speed less than 3 words per second or greater than 6 words per second				√
	Accent	Standard accent	√	√	√	
		Non-standard accent				√