ETSI TS 123 042 V14.0.0 (2017-04)

**Digital cellular telecommunications system (Phase 2+) (GSM);
Universal Mobile Telecommunications System (UMTS);
LTE;
Compression algorithm for text messaging services
(3GPP TS 23.042 version 14.0.0 Release 14)**

Reference
RTS/TSGC-0123042ve00

Keywords
GSM,LTE,UMTS

ETSI

650 Route des Lucioles
F-06921 Sophia Antipolis Cedex - FRANCE

Tel.: +33 4 92 94 42 00 Fax: +33 4 93 65 47 16

Siret N° 348 623 562 00017 - NAF 742 C
Association à but non lucratif enregistrée à la
Sous-Préfecture de Grasse (06) N° 7803/88

Intellectual Property Rights

IPRs essential or potentially essential to the present document may have been declared to ETSI. The information pertaining to these essential IPRs, if any, is publicly available for **ETSI members and non-members**, and can be found in ETSI SR 000 314: *"Intellectual Property Rights (IPRs); Essential, or potentially Essential, IPRs notified to ETSI in respect of ETSI standards"*, which is available from the ETSI Secretariat. Latest updates are available on the ETSI Web server (https://ipr.etsi.org/).

Pursuant to the ETSI IPR Policy, no investigation, including IPR searches, has been carried out by ETSI. No guarantee can be given as to the existence of other IPRs not referenced in ETSI SR 000 314 (or the updates on the ETSI Web server) which are, or may be, or may become, essential to the present document.

Foreword

This Technical Specification (TS) has been produced by the ETSI 3rd Generation Partnership Project (3GPP).

The present document may refer to technical specifications or reports using their 3GPP identities, UMTS identities or GSM identities. These should be interpreted as being references to the corresponding ETSI deliverables.

The cross reference between GSM, UMTS, 3GPP and ETSI identities can be found under http://webapp.etsi.org/key/queryform.asp.

Modal verbs terminology

In the present document "**shall**", "**shall not**", "**should**", "**should not**", "**may**", "**need not**", "**will**", "**will not**", "**can**" and "**cannot**" are to be interpreted as described in clause 3.2 of the ETSI Drafting Rules (Verbal forms for the expression of provisions).

"**must**" and "**must not**" are **NOT** allowed in ETSI deliverables except when used in direct citation.

Contents

Foreword

This Technical Specification (TS) has been produced by ETSI 3rd Generation Partnership Project (3GPP).

The contents of the present document are subject to continuing work within the TSG and may change following formal TSG approval. Should the TSG modify the contents of this TS, it will be re-released by the TSG with an identifying change of release date and an increase in version number as follows:

Version x.y.z

where:

- x the first digit:
  - 1 presented to TSG for information;
  - 2 presented to TSG for approval;
  - 3 Indicates TSG approved document under change control.
- y the second digit is incremented for all changes of substance, i.e. technical enhancements, corrections, updates, etc.
- z the third digit is incremented when editorial only changes have been incorporated in the specification;

Introduction

This clause introduces the concepts and mechanisms involved in the compression and decompression of a stream of data.

Overview

Central to the compression of a stream of data and the subsequent recovery of the original data is the that both sender and receiver have information that not only describes the content of the data stream, but how the stream is encoded.

For example, a simple rule such as "it's 8 bit data" is enough to transport any character value in the range 0 to 255 with 8 bits being required for each and every character. In contrast if both sender and receive know that some characters are more frequent than others, then the more frequent might be encoded in fewer bits while the less frequent in more - resulting in a net reduction of the total number of bits used to express the data stream.

This knowledge of the nature of the data stream can be established in two ways. Either both sender and receiver can agree some key aspects of the data stream *prior* to it being processed or key aspects of the data can be garnered *dynamically* during its processing.

The disadvantage of an approach based on "prior information" is that it must be known. It can either be carried as a header to the data stream, in which case it adds to the net size of the compressed stream. Or it can be fixed and known to the (de)compression algorithm itself in which case compression performance degrades as a given stream diverges in nature from these fixed and known states. In contrast, the disadvantage of "dynamic information" is that it must be discovered; typically this means a greater processing requirement for the (de)compressor. It also implies that compression performance is initially poor as the algorithm has to "learn" about the data stream before it can apply this knowledge. It will also require greater working memory to store its knowledge about the data stream.

The choice of compression algorithms is always a balancing of compression rate (in terms of fewer output bits), working memory requirements of the (de)compressor and CPU bandwidth. For the compression of SMS messages, there is the additional requirement that it should work well (in terms of compression rate) even on short data streams.

Compression / Decompression is an optional feature but when implemented, the only mandatory requirement is ‘Raw Untrained Dynamic Huffman’ . The default initialisation for the Huffman Encoder / Decoder operating in the Raw Untrained Dynamic Huffman mode are defined in annex R. (See also subclause 4.1.)

i.e. There is no need for any pre-defined attributes such as language dependency to be included. This is of particular significance for entities such as an MS which may have memory storage constraints.

1 Scope

The present document introduces the concepts and mechanisms involved in the compression and decompression of a stream of data.

2 References

The following documents contain provisions which, through reference in this text, constitute provisions of the present document.

- References are either specific (identified by date of publication, edition number, version number, etc.) or non-specific.
- For a specific reference, subsequent revisions do not apply.
- For a non-specific reference, the latest version applies. In the case of a reference to a 3GPP document (including a GSM document), a non-specific reference implicitly refers to the latest version of that document *in the same Release as the present document*.

2.1 Normative references

[1] 3GPP TS 23.038: "Alphabets and language-specific information".

2.2 Informative references

[2] "The Data Compression Handbook 2nd Edition" by Mark Nelson and Jean-Loup Gailly, published by M&T Books, ISBN 1-22851-434-1.

3 Abbreviations

For the purposes of the present document, the following abbreviations apply.

CD	Compressed Data
CDS	Compressed Data Stream
CDSL	Compressed Data Stream Length
CF	Compression Footer
CG-ID	Character Group ID
CH	Compression Header
CLC	Compression Language Context
HI-ID	Huffman initialization ID
KD-ID	Keyword Dictionary ID
PU-ID	PUnctuator ID

4 Algorithms

The compression algorithm comprises a number of components that may be combined in a variety of configurations. The discrete algorithms are discussed in the following subclauses.

4.1 Huffman Coding

The base compression algorithm is a Huffman coder, whereby characters in the input stream are represented in the output stream by bit sequences of variable length. This length is inversely proportional to the frequency with which the character occurs in the input stream.