IEC 62503

Edition 1.0 2008-09

INTERNATIONAL STANDARD

Multimedia quality – Method of assessment of synchronization of audio and video

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 corrigenda or an amendment might have been published.

- Catalogue of IEC publications: www.iec.ch/searchpub

The IEC on-line Catalogue enables you to search by a variety of criteria (reference number, text, technical committee,…). It also gives information on projects, withdrawn and replaced publications.

- IEC Just Published: www.iec.ch/online_news/justpub

Stay up to date on all new IEC publications. Just Published details twice a month all new publications released. Available on-line and also by email.

- Electropedia: www.electropedia.org

The world's leading online dictionary of electronic and electrical terms containing more than 20 000 terms and definitions in English and French, with equivalent terms in additional languages. Also known as the International Electrotechnical Vocabulary online.

- Customer Service Centre: www.iec.ch/webstore/custserv

If you wish to give us your feedback on this publication or need further assistance, please visit the Customer Service Centre FAQ or contact us:

Email: csc@iec.ch
Tel.: +41 22 919 02 11
Fax: +41 22 919 03 00

IEC 62503

Edition 1.0 2008-09

INTERNATIONAL STANDARD

Multimedia quality – Method of assessment of synchronization of audio and video

INTERNATIONAL ELECTROTECHNICAL COMMISSION

PRICE CODE M

ICS 33.160.01

ISBN 2-8318-9986-9

CONTENTS

INTERNATIONAL ELECTROTECHNICAL COMMISSION

_____

MULTIMEDIA QUALITY – METHOD OF ASSESSMENT OF SYNCHRONIZATION OF AUDIO AND VIDEO

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 provides no marking procedure to indicate its approval and cannot be rendered responsible for any equipment declared to be in conformity with an IEC Publication.
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) Attention is drawn to the possibility that some of the elements of this IEC Publication may be the subject of patent rights. IEC shall not be held responsible for identifying any or all such patent rights.

International Standard IEC 62503 has been prepared by technical area 11: Quality for audio, video and multimedia systems, of IEC technical committee 100: Audio, video and multimedia systems and equipment.

The text of this standard is based on the following documents:

CDV	Report on voting
100/1277/CDV	100/1358/RVC

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

This publication has been drafted in accordance with the ISO/IEC Directives, Part 2.

The committee has decided that the contents of this publication will remain unchanged until the maintenance result date indicated on the IEC web site under "http://webstore.iec.ch" in the data related to the specific publication. At this date, the publication will be

- reconfirmed,
- withdrawn,
- replaced by a revised edition, or
- amended.

A bilingual version of this publication may be issued at a later date.

INTRODUCTION

Contemporary multimedia systems are realized by digital technology. Depending on what digital processing is being applied, time delays differ among medium channels for reproduction as perceptible stimulus for a human audience. An example is video delay against audio, which is identified by such terms as lip sync or AV-sync. Video delay against audio will be inevitable for large sized displays, because necessary time for rendering and visualization will be proportional to the number of picture elements.

There should also be additional factors to be considered. They include synchronization problem during medium gathering, production, post-production, processing in various aspects to combine these multiple media and to be sent out or recorded as “multimedia”.

There is a need for international standards to provide the following three related methodologies:

a) an objective method of measurement for difference of delays between reproduced audio and video (lip sync) by multimedia systems and equipment,

b) a subjective (or perceptible) and statistical method of assessment of overall difference of delays between a real world and a reproduced scene and sound,

c) a method of estimation of implied difference of delays inherent in multimedia received, recorded or under reproduction.

This International Standard addresses the item b) using typical multimedia content such as bust shots of news casters because of easiness in defining synchronization of audio and video. Since a range of perceptibly allowable miss-synchronization and sensitivity of human audience for lead and delayed audio against accompanied video depends on human sensation and the conditions for assessment, a clearly defined method of assessment of such characteristics should be standardised.

This International Standard is intended to supplement Recommendation ITU-R BT.1359-1, [1] [1], as well as partly answer the request of ITU-R to IEC that has been stated in Recommendation ITU-R BT.1377 [2].

Technical contents are based on a study in Faculty of Engineering, Chiba University in Japan, conducted in April 2006.

---

1 Numbers in square brackets refer to the Bibliography.

MULTIMEDIA QUALITY – METHOD OF ASSESSMENT OF SYNCHRONIZATION OF AUDIO AND VIDEO

1 Scope

This International Standard provides a subjective (or perceptible) and statistical method of assessment of overall, or end-to-end, difference of delays between real world and reproduced scenes in terms of video and accompanying audio recoded in a medium.

This International Standard does not specify limiting values for those results obtained by the application of the provisions in this standard. It excludes applications to professional broadcast systems.

2 Normative reference

The following referenced document is indispensable for the application of this document. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

ITU-R BT.500-11:2002, *Methodology for the subjective assessment of the quality of television pictures*

3 Terms and definitions

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

3.1
lip sync
video delay against accompanying audio

3.2
outlier
subjective opinion score outside $m \pm s$, where m is a sample mean of the original scores of a set of subjects for the same video delay and s is a standard deviation of the scores

3.3
subject
ordinary untrained human audience of audio and video reproduction; random sample of individual members of general public

3.4
test video clip
short duration of video frames with accompanying audio to be used as original

3.5
test video sequence
random series of test video clips where the audio channels are shifted in time compared to the original

4 Overview of methods of assessment

Figure 1 depicts overview of possible objective methods of measurement and subjective method of assessment to acquire necessary parameters corresponding to lip sync.

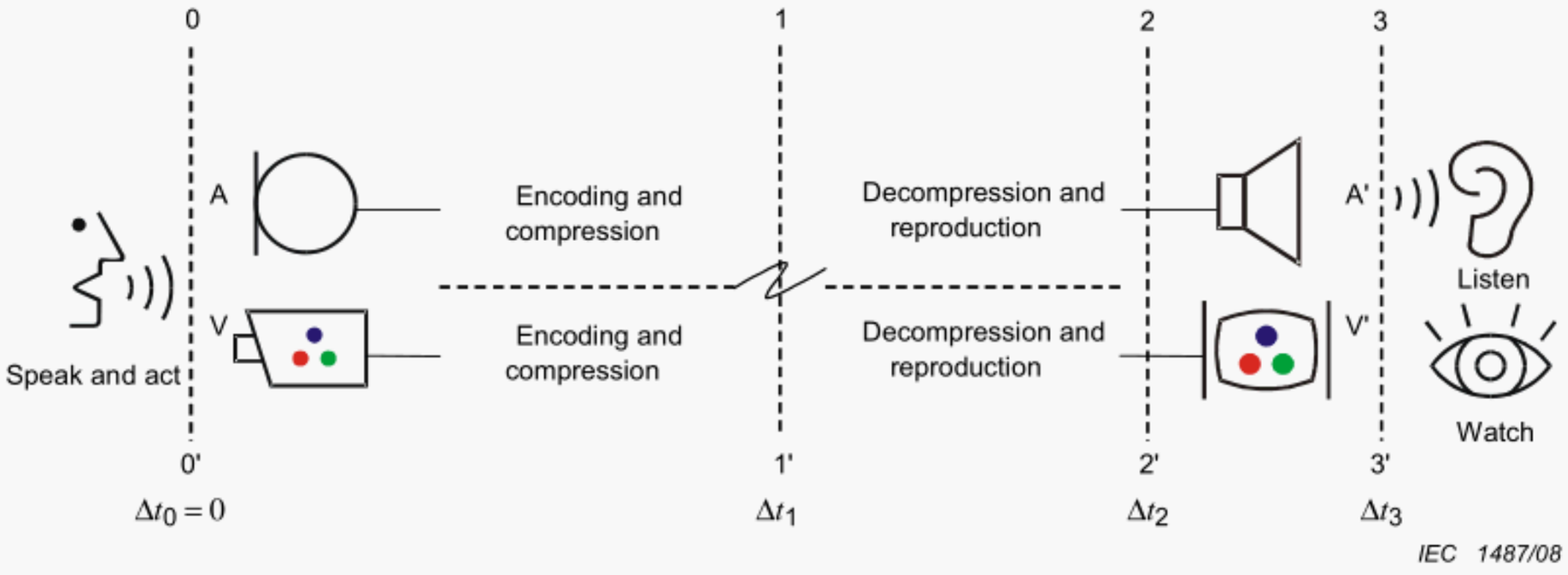

Figure 1 – Overview

The leftmost in Figure 1 is a real world and the rightmost in Figure 1 is a reproduced world.

Lip sync at the section 0-0', i.e. scene delay in reference to accompanying sound, is normally zero; in other words null video delay against accompanying audio is expected; $\Delta t_0 = 0$. Where $\Delta t_0 \neq 0$ is foreseen, it shall also be taken into account.

Lip sync at the section 1-1' is supposed to be introduced by separate acquisition of physical phenomena by microphones and video cameras followed by yet further separate digital processing for audio and video data. It will cause lip sync of $\Delta t_1 \neq 0$.

NOTE In case of MPEG-2 encoding, there is the scheme of synchronization using Decoding Time Stamp (DTS) as well as Presentation Time Stamp (PTS) embedded in the header of Packetized Elementary Stream (PES). See ISO/IEC 13818-1 [11].

Lip sync at the section 2-2' is supposed to be introduced by reproduction process for audio and video channels separately such as decompression, rendering and reproduction. It will cause lip sync of $\Delta t_2 \neq 0$, which can be measured using a reference test multimedia material with $\Delta t_1 = 0$.

Lip sync at the section 3-3' is in the reproduced multimedia world and assessed by human subjects. Subjective opinion scores on lip sync are statistically analyzed to find estimated value for $\Delta t_3 \neq 0$; corresponding to the amount of compensation for just-synchronized reproduction.

5 Subjective assessment of lip sync

5.1 Items to be assessed

Subjective grading level of miss-synchronization of video and audio.

5.2 Preparation of test video clips and test video sequence

5.2.1 Selection of content of a test video clip

Since lip sync is a kind of human perception, it may depend on the contents of the video and accompanying audio. Especially when it is related to movement of lips of a human speaker, a match between a spoken language and a mother tongue may affect the result.

NOTE In this International Standard, in order to provide worked examples, speech in Japanese language uttered by a well trained processional news reader is watched and listened to by the subjects with the same mother tongue.

A bust shot of a news reader shall be extracted, duration of which should be around 10 s to 20 s. Data of audio channel of the video clip shall be taken as the timing reference.

Possible amount of time caused by miss-synchronization in this original video clip, Δt_1 at the section 1-1', is unknown. However, this international standard provides the method to estimate overall lip sync Δt_3 including Δt_0 and Δt_1. Namely, $\Delta t_3 = \Delta t_0 + \Delta t_1 + \Delta t_2$.

5.2.2 Creation of a test video sequence

The test video sequence shall be a randomised series of the video clip selected in 5.2.1, in which each of the audio channels shall be replaced by time-shifted audio data with necessary duration of padding as a leader or a trailer depending on the direction of the time shift. Preparation of such video clips is show in Figure 2 as in the image frames with delayed audio and with led audio. The amount of time shifts T_l and T_d is subject to be adjusted.

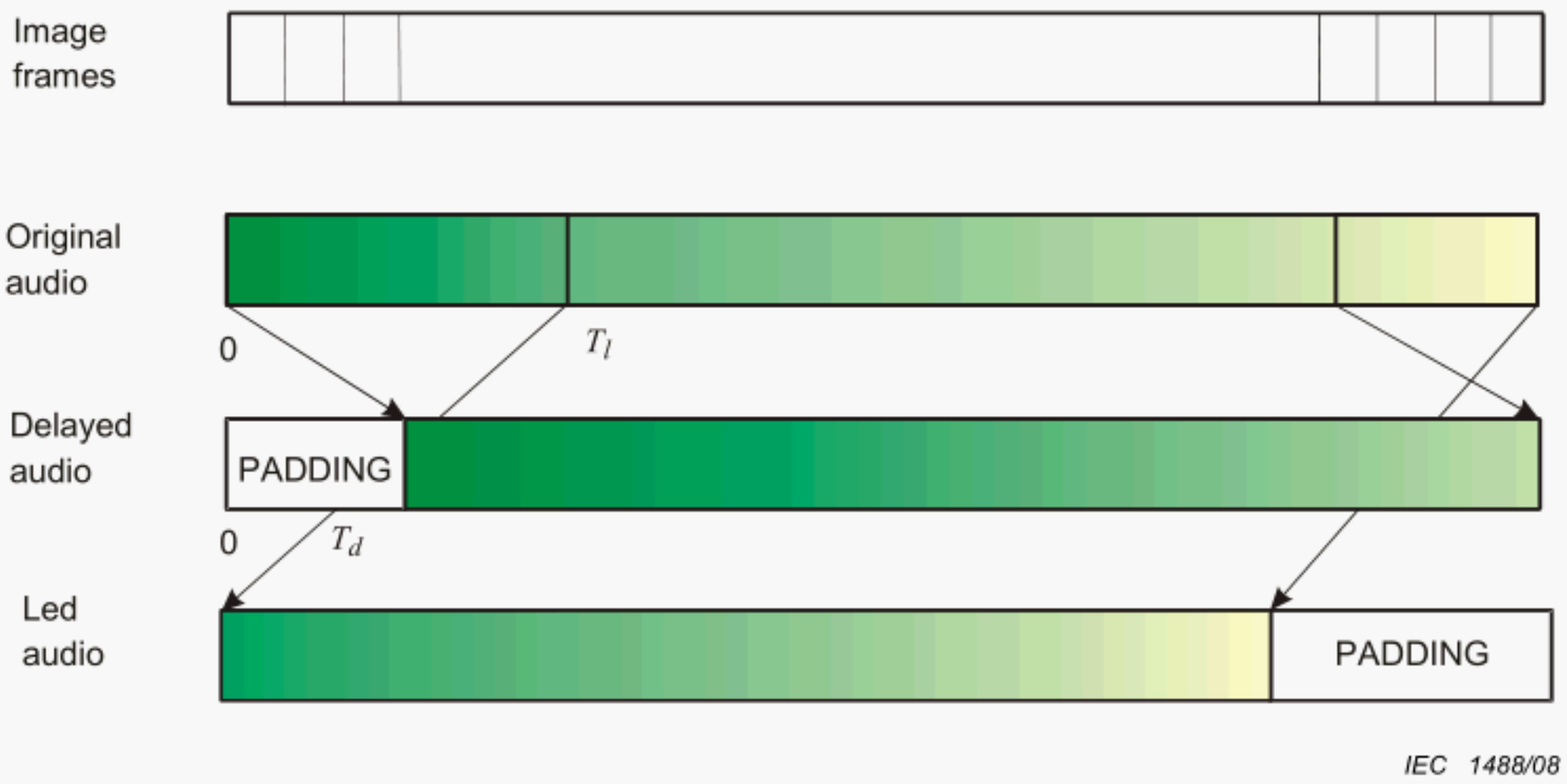

Figure 2 – Preparation of test video clips with time shifted audio

To allow for each of the video clips with the time-shifted audio composing the test video sequence to be visually identified by a subject, each video clip prepared in accordance with Figure 2 should be preceded by a necessary number of title frames which include a sequence number.

The amount of time shifts for audio data, T_l and T_d, shall be determined taking into account the sum of the lip sync in reproduction system, Δt_2, and possible lip sync in the original video clip, Δt_1. The amount of increment and decrement of the time shift ΔT for T_l or T_d shall be decided in accordance with precision of assessment.

In this standard, ΔT = 10 ms is recommended.

The test video sequence should be stored in a medium such as CD-ROM for use in 5.3 without losing audio-video synchronization.

5.3 Procedures and condition for assessment of lip sync at the section 3-3'

The procedures described below shall be followed.

a) The test video sequence being composed of randomized order of the same short video clips of different time-shifted audio (plus and minus) in reference to video shall be reproduced to subjects.

b) The number of the subjects shall be at least 15. Each of them shall be asked to report its subjective opinion score for each of the video clips in the test video sequence under a fixed viewing and audible condition.

c) An advance instruction to the subjects shall be provided on the five-grade impairment scale recommended by ITU-R BT.500-11 for subjective judgement, as shown in Table 1.

Table 1 – Five-grade impairment scale and explanation of subjective opinion score

Grade	Explanation
5	**Imperceptible** (Complete agreement of audio and video)
4	**Perceptible but not annoying** (miss-synchronization is within the acceptable degree)
3	**Slightly annoying** (slight miss-synchronization is recognized)
2	**Annoying** (apparent miss-synchronization is recognized and beyond acceptable degree)
1	**Very annoying** (not synchronized at all)

d) Viewing distance of the reproduced test video sequence L shall be four times of reproduction height H of video frames: $L = 4H$.

e) Audio shall be reproduced by a headphone or a loudspeaker. Use of the headphone is recommended to prevent any multi-path echo in a test room.

NOTE Taking into account the travel speed of sound waves, the time for the sound pressure of a speaker to reach to ears of the observer should be regarded as an additional audio delay.

f) Illumination of a test room where the audio-video reproduction system is installed should be around 300 lx with correlated colour temperature of about 6 500 K.

5.4 Reporting of the result of assessment

Taking into account outliers from the original five-grade opinion scores, averaged subjective opinion scores shall be plotted against predetermined audio shift, T_l or $-T_d$, as exemplified in Figure 3, with error bars of $m \pm c$, where m is a sample mean of the opinion scores of a set of subjects for the same video clip and c is a 95 % confidence interval of each of the respective mean opinion scores.

The horizontal axis shall be the amount of time shifts of the audio channel or the value of T_l in Figure 2.

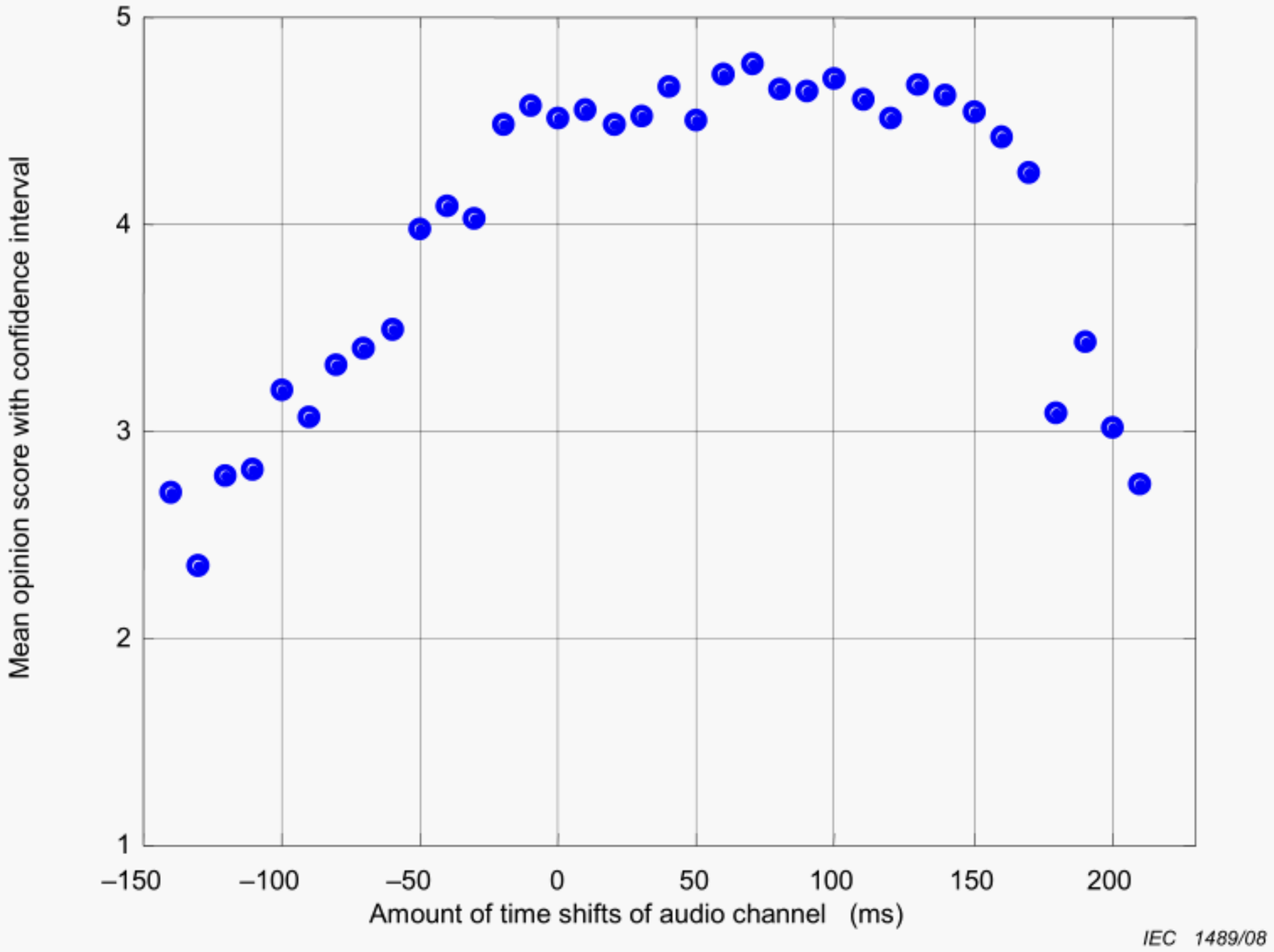

Figure 3 – An example of subjective assessment of lip sync

The report shall also include the following information:

a) description of a kind of audio and video source used in the assessment, such as spoken language in case of a speech and duration of the test video clip;

b) the number of the subjects;

c) illumination level and correlated colour temperature of lighting used; and

d) the position and distance from the display and speakers to the observer, in case of headphone not being used.

6 Data processing

6.1 Items to be assessed

Subjective time difference in milliseconds between reproduced video and associated audio.

6.2 Method of assessment

The mean opinion scores in the five-grade impairment scale reported in figure 3 shall be regressed by the stepwise linear functions as shown by the set of equations (1) to (3).

$$g = a_1 (t - t_1) + g_0, \quad \text{for} \quad t < t_1 \tag{1}$$

$$g = g_0, \qquad \text{for} \quad t_1 \le t \le t_2 \tag{2}$$

$$g = a_2 (t - t_2) + g_0, \quad \text{for} \quad t_2 < t \tag{3}$$

A horizontal coordinate value, or an estimate for Δt_3 in terms of Figure 1, should be calculated as an intersection of the two lines defined by equation (1) and equation (3). Namely, Δt_3 is expressed by equation (4).

$$\Delta t_3 = \frac{a_1 t_1 - a_2 t_2}{a_1 - a_2} \tag{4}$$

NOTE An amount of the lip sync, i.e. video delay against audio needed to be compensated falls between t_1 and t_2 in equation (2). The value of Δt_3 calculated by equation (4) is one of possible estimates.

Figure 4 explains the scheme of the method of estimation by an example, in which the vertical line at the intersection coordinate of the two lines corresponds to an estimated lip sync compensation Δt_3.

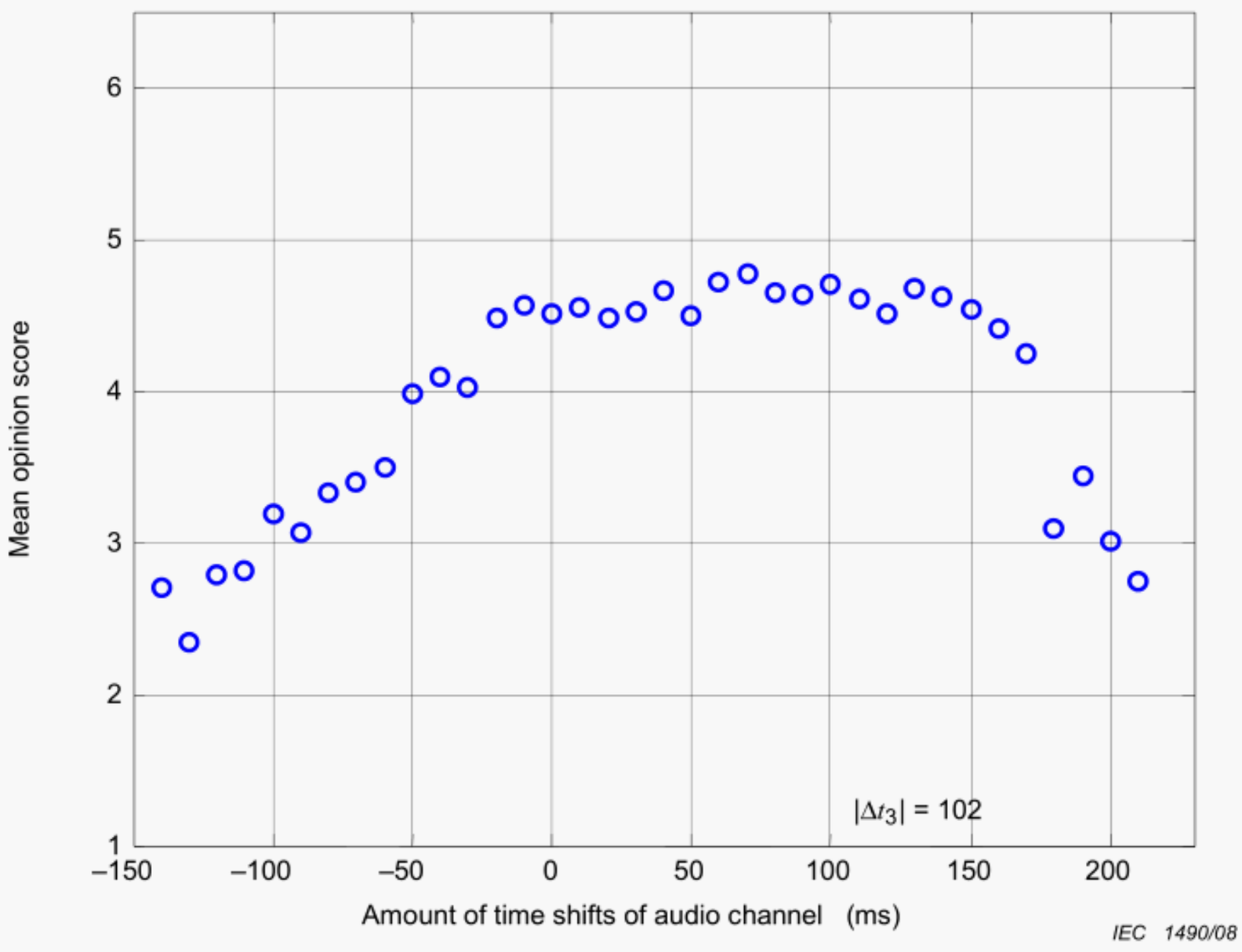

Figure 4 – Normalized response for grading impairment caused by lip sync mismatch

6.3 Reporting of the result of estimation

The amount of overall audio time shift to compensate lip sync obtained by the regression and the extrapolation shall be reported in milliseconds together with the time parameters t_1 and t_2 as shown below.

$$\begin{aligned} t_1 &= -3 \text{ ms} \\ t_2 &= 152 \text{ ms} \\ \Delta t_3 &= 102 \text{ ms} \end{aligned}$$

Bibliography

[1] Recommendation ITU-R BT.1359-1: Relative timing of sound and vision for broadcasting (1998)

[2] Recommendation ITU-R BT.1377: Labelling of video and audio apparatus throughput (processing) delay (1998)

[3] ATSC Implementation Subcommittee Finding: Relative timing of sound and vision for broadcast operations, Advanced Television Systems Committee Doc. IS-191 (26 June 2003).

[4] EBU Technical Recommendation R37-2007: The relative timing of the sound and vision components of a television signal, the European Broadcasting Union.

[5] IEC 60417, Graphical symbols for use on equipment, http://www.graphical-symbols.info/equipment

[6] Hiroaki IKEDA, Reiko IWAI and Junichi YOSHIO: Kansei on the time difference between audio and video, Hoso Gigyutsu (Broadcasting Technologies), Vol.59, No.10, pp.151-155 (2006-10).

[7] Hiroaki IKEDA and Junichi YOSHIO: Synchronization in multimedia – Technical review, Journal of IEEJ, Vol. 126, No.5, pp.288-291 (May 2006).

[8] Question ITU-R 68/6: Synchronization necessary for the satisfactory reception of sound and picture signals (2003-10) – still under study as of 2006.

[9] Question ITU-T 11/9: Requirements and methods for sound and television transmission over IP networks “webcasting” (Study period: 2005-2008).

[10] ISO/IEC 13818-1 Ed.2:2000, Amendments 1, 2, 3, 4, 5, Information technology – Generic coding of moving pictures and associated audio information – Part 1: systems.

[11] IEC/TS 62312-2, Guideline for synchronization of audio and video – Part 2: Methods for synchronization of audio and video systems

INTERNATIONAL
ELECTROTECHNICAL
COMMISSION

3, rue de Varembé
PO Box 131
CH-1211 Geneva 20
Switzerland

Tel: + 41 22 919 02 11
Fax: + 41 22 919 03 00
info@iec.ch
www.iec.ch