ATSC Standard: Captions and Subtitles (A/343)

Doc. A/343:2017
18 September 2017

Advanced Television Systems Committee
1776 K Street, N.W.
Washington, D.C. 20006
202-872-9160

The Advanced Television Systems Committee, Inc., is an international, non-profit organization developing voluntary standards for digital television. The ATSC member organizations represent the broadcast, broadcast equipment, motion picture, consumer electronics, computer, cable, satellite, and semiconductor industries.

Specifically, ATSC is working to coordinate television standards among different communications media focusing on digital television, interactive systems, and broadband multimedia communications. ATSC is also developing digital television implementation strategies and presenting educational seminars on the ATSC standards.

ATSC was formed in 1982 by the member organizations of the Joint Committee on InterSociety Coordination (JCIC): the Electronic Industries Association (EIA), the Institute of Electrical and Electronic Engineers (IEEE), the National Association of Broadcasters (NAB), the National Cable Telecommunications Association (NCTA), and the Society of Motion Picture and Television Engineers (SMPTE). Currently, there are approximately 150 members representing the broadcast, broadcast equipment, motion picture, consumer electronics, computer, cable, satellite, and semiconductor industries.

ATSC Digital TV Standards include digital high definition television (HDTV), standard definition television (SDTV), data broadcasting, multichannel surround-sound audio, and satellite direct-to-home broadcasting.

Note: The user's attention is called to the possibility that compliance with this standard may require use of an invention covered by patent rights. By publication of this standard, no position is taken with respect to the validity of this claim or of any patent rights in connection therewith. One or more patent holders have, however, filed a statement regarding the terms on which such patent holder(s) may be willing to grant a license under these rights to individuals or entities desiring to obtain such a license. Details may be obtained from the ATSC Secretary and the patent holder.

Revision History

Version	Date
Candidate Standard approved	23 December-2015
Standard approved	21 December 2016
A/343:2017 Revision approved	18 September 2017

Table of Contents

Index of Figures

ATSC Standard:
Captions and Subtitles

1. SCOPE

This standard defines the required technology for closed caption and subtitle tracks over ROUTE-DASH and MMT transports. This includes the content essence and the packaging and timing.

1.1 Organization

This document is organized as follows:

- Section 1 – Outlines the scope of this document and provides a general introduction.
- Section 2 – Lists references and applicable documents.
- Section 3 – Provides a definition of terms, acronyms, and abbreviations for this document.
- Section 4 – System overview
- Section 5 – Content Essence description
- Section 6 – Packaging and timing in the ISO BMFF
- Section 7 – Signaling
- Section 8 – Decoder Recommendations
- Annex A – Live and Broadcast Boundary Considerations

2. REFERENCES

All referenced documents are subject to revision. Users of this Standard are cautioned that newer editions might or might not be compatible.

2.1 Normative References

The following documents, in whole or in part, as referenced in this document, contain specific provisions that are to be followed strictly in order to implement a provision of this Standard.

[1] IEEE: "Use of the International Systems of Units (SI): The Modern Metric System," Doc. SI 10, Institute of Electrical and Electronics Engineers, New York, N.Y.

[2] W3C: "TTML Profiles for Internet Media Subtitles and Captions 1.0 (IMSC1)", Recommendation, W3C, www.w3.org.

[3] W3C: "Timed Text Markup Language 2 (TTML2)", 06-01-2017 Public Working Draft, W3C, https://www.w3.org/TR/2017/WD-ttml2-20170106/.

[4] ATSC: ATSC Proposed Standard: "Signaling, Delivery, Synchronization, and Error Protection (A/331)." Doc. A331(S33-174r7), Advanced Television Systems Committee, Washington, D.C., 4 May 2017. (work in process)

[5] SMPTE: "ST 2052-11:2013, Conversion from CEA-708 Caption Data to SMPTE-TT," Society of Motion Picture and Television Engineers, White Plains, NY, https://www.smpte.org/standards.

2.2 Informative References

The following documents contain information that may be helpful in applying this Standard.

[6] SMPTE: "ST 2052-1:2013, Timed Text Format (SMPTE-TT)," Society of Motion Picture and Television Engineers, White Plains, NY, https://www.smpte.org/standards.

[7] SMPTE: Webcasts: https://www.smpte.org/standards-webcasts-on-demand (second webcast from the bottom).
[8] W3C: "Timed Text Markup Language 1 (TTML1) (Second Edition)", Recommendation, W3C, www.w3.org.
[9] W3C: "TTML Simple Delivery Profile for Closed Captions (US)", Recommendation, W3C, www.w3.org.
[10] DECE: Common File Format and Media Formats Specification, DECE, www.uvcentral.com.
[11] CTA: 608-E, "Line 21 Data Services," Consumer Electronics Association, Arlington, VA, www.ce.org.
[12] CTA: 708.1, "Digital Television (DTV) Closed Captioning: 3D Extensions," Consumer Technology Association, Arlington, VA, www.ce.org.
[13] EBU: TECH 3381, "EBU-TT-D SUBTITLING DISTRIBUTION FORMAT VERSION: 1.0 https://tech.ebu.ch/publications/tech3381

3. DEFINITION OF TERMS

With respect to definition of terms, abbreviations, and units, the practice of the Institute of Electrical and Electronics Engineers (IEEE) as outlined in the Institute's published standards [1] shall be used. Where an abbreviation is not covered by IEEE practice or industry practice differs from IEEE practice, the abbreviation in question will be described in Section 3.3 of this document.

3.1 Compliance Notation

This section defines compliance terms for use by this document:

shall – This word indicates specific provisions that are to be followed strictly (no deviation is permitted).

shall not – This phrase indicates specific provisions that are absolutely prohibited.

should – This word indicates that a certain course of action is preferred but not necessarily required.

should not – This phrase means a certain possibility or course of action is undesirable but not prohibited.

3.2 Treatment of Syntactic Elements

This document contains symbolic references to syntactic elements used in the audio, video, and transport coding subsystems. These references are typographically distinguished by the use of a different font (e.g., restricted), may contain the underscore character (e.g., sequence_end_code) and may consist of character strings that are not English words (e.g., dynrng).

3.2.1 Reserved Elements

One or more reserved bits, symbols, fields, or ranges of values (i.e., elements) may be present in this document. These are used primarily to enable adding new values to a syntactical structure without altering its syntax or causing a problem with backwards compatibility, but they also can be used for other reasons.

The ATSC default value for reserved bits is '1.' There is no default value for other reserved elements. Use of reserved elements except as defined in ATSC Standards or by an industry standards setting body is not permitted. See individual element semantics for mandatory settings and any additional use constraints. As currently-reserved elements may be assigned values and

meanings in future versions of this Standard, receiving devices built to this version are expected to ignore all values appearing in currently-reserved elements to avoid possible future failure to function as intended.

3.3 Acronyms and Abbreviations

The following acronyms and abbreviations are used within this document:

ABNF – Augmented Backus–Naur Form
ATSC – Advanced Television Systems Committee
BMFF – Base Media File Format
CFF – Common File Format
CTA – Consumer Technology Association
DASH – Dynamic Adaptive Streaming over HTTP
DASH-IF – DASH Industry Forum
DECE – Digital Entertainment Content Ecosystem
EBU – European Broadcast Union
FCC – Federal Communications Commission
HTTP – Hyper-Text Transport Protocol
IETF – Internet Engineering Task Force
IMSC1 – Internet Media Subtitles and Captions Version 1
ISO – International Standards Organization
MMT – MPEG Media Transport
MMTP – MPEG Media Transport Protocol
MPD – Media Presentation Description
MPU – Media Processing Unit
SMPTE – Society of Motion Picture and Television Engineers
TT – Timed Text
TTML – Timed Text Markup Language
URI – Uniform Resource Identifier
USBD – User Service Bundle Description
W3C – World Wide Web Consortium
XML – Extensible Markup Language

3.4 Terms

The following terms are used within this document:

reserved – Set aside for future use by a Standard.

3.5 Extensibility

This ATSC 3.0 specification is based on W3C IMSC1, an XML-based representation of captions. XML is inherently extensible and can be enhanced over time by ATSC retaining compatibility with earlier versions. For example, user systems can extend it using their own namespaces and retain compatibility with the core feature set defined here.

3.6 XML Schema and Namespace

The schema is available at W3C and the namespace is defined there. There are currently no ATSC-defined namespaces or schemas.

4. SYSTEM OVERVIEW

4.1 Features

The technology is SMPTE Timed Text (SMPTE-TT) as defined in SMPTE 2052-1 [6]. SMPTE-TT was chosen as it:

- Supports world-wide language and symbol tables (specifically including non-Latin)
- Supports world-wide image glyph delivery
- Is in use today by various "media delivery silos", including broadcaster internet-delivered services
- Is US FCC closed caption safe harbor for IP-delivered content
- Supports FCC requirements for both 708[1] and IP captions (See US 47CFR§79)
- Compatible with DECE (UltraViolet) Common File Format Timed Text (CFF-TT) at [10]

All of SMPTE-TT is complex and not required to meet closed captions and subtitle requirements. A simpler subset is desirable for practical implementation. Therefore, W3C's new "TTML Text and Image Profiles for Internet Media Subtitles and Captions (IMSC1)" [2] is selected having been designed specifically for needs like broadcast as well as broadband delivery. In summary:

- Superset of DECE/Ultraviolet CFF-TT (TTML + SMPTE-TT extensions)
- Superset of EBU-TT-D being deployed in Europe (see EBU Tech 3381 [13])
- Two profiles are included
  - Text Profile requiring a font rendering engine in the decoder
  - Image Profile with PNG files

The rough feature relationships of the TTML profiles mentioned above are shown in Figure 4.1. The ATSC 3.0 elements are the "IMSC1" (bright green) ovals.

[1] Drop-shadow is not exact.

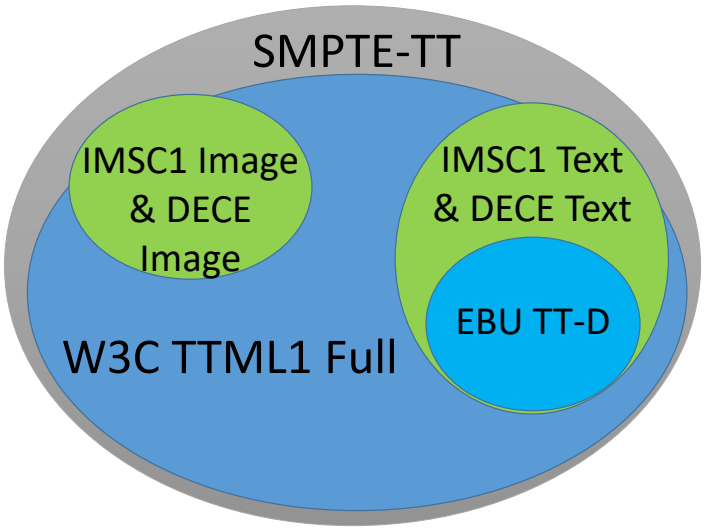


Figure 4.1 Venn diagram of TTML profiles.

4.2 System Architecture

When present, the content essence for captions and subtitles is formed using one or more ISO BMFF track files each containing one or more XML documents. The XML documents conform to W3C TTML IMSC1 profiles as constrained and extended in this specification. Each track contains only one set of "timed text" corresponding to a set of metadata "signaling".

4.3 Central Concepts

A tutorial on TTML in general and SMPTE-TT specifically can be found in SMPTE Webcasts at [7].

Examples of using TTML for US closed caption scenarios can be found in the underlying TTML1 specification at [8].

Additional background on using TTML1 for the conversion from CTA 608 [11] can be found in a W3C profiles, called SDP-US at [9].

A graphical description of timing for the live content scenario (see Section 6.3) is shown in Figure 4.2.

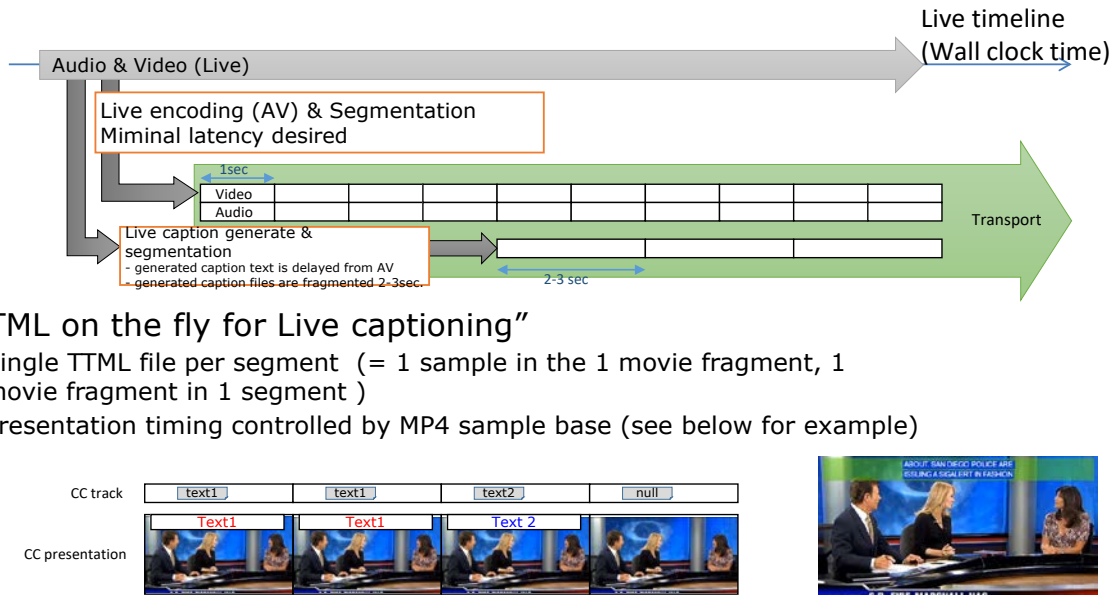


Figure 4.2 Live caption timing model.

5. CONTENT ESSENCE SPECIFICATION

The content essence for closed captions and subtitles shall be IMSC1 as defined at [2] including either text profile or image profile or both.

When caption or subtitle tracks are present, the decoder cannot be expected to provide viewer override of IMSC1 attributes for images, e.g., color, opacity, edge effects. The Service shall provide at least one IMSC1 text profile track when such viewer override capability is required.

5.1 Extensions

This section contains extensions to the IMSC1 XML language.

5.1.1 3D

Extensions for 3D allow caption authors to correctly place caption regions over 3D video. When 3D disparity is used, it shall be as described in TTML2 at [3] Section 10.2.10.

> Note: Section 10.2.10 of TTML2 provides static disparity found in CTA 708.1 [12].
>
> Note: Although reference is made to TTML2 for tts:disparity syntax and semantics, it does not imply the adoption of any new TTML2 features.
>
> Note: Although this feature is not in IMSC1 or TTML1, the TTML1 (which is also the TTML2) namespace is still used, http://www.w3.org/ns/ttml#styling.

When the disparity value is specified as a percentage of the video width format, it can scale properly for any resolution image. The range should be from +/– 0.0% to +/– 10.0% of picture width.

5.1.2 High Dynamic Range (HDR) and Wide Color Gamut (/WCG)

Note: IMSC1 image subtitles use the PNG image file format. PNG is "SDR" using the sRGB color model. Compositing these image files with the underlying video is up to the decoding device.

5.2 Conversion and Carriage of Legacy Caption Data

Conversion of CTA 708 into IMSC1 requires well defined procedures to ensure interoperability and consistency. Additionally, to provide more options for downstream processing, e.g. at MVPD interfaces, interoperability there would be improved by having the original 708 information available within IMSC1.

When the source of IMSC1 captioning information is a translation from CTA 708 (or CTA 608 carried in 708 compatibility bytes), then:

1) the conversion into IMSC1 shall follow the recommendations of SMPTE RP2052-11 [5]; and
2) the original 708 caption channel packet (ccdata()) should be included in the IMSC1 document according to SMPTE RP2052-11 [5], with the additional provisions below.

When carrying the 708 caption channel packet data (cc_data()), it shall be temporally co-located and interspersed with the IMSC1 information in order to facilitate synchronization, fragmentation, random access and live broadcast requirements.

6. PACKAGING AND TIMING IN THE ISO BASE MEDIA FILE FORMAT (ISO BMFF)

Caption and Subtitle Elementary Streams shall be packaged and signaled as defined in ATSC A/331 [1].

6.1 Pre-recorded Broadband Content

For broadband delivery, the DASH segment size shall be less than 500K bytes. This is needed to bound the amount of decoder memory needed to decode a document and also provide a reasonable startup acquisition time at the beginning of a program.

> Note: The caption ISO BMFF sample length can be the length of the program; i.e., a single file.

6.2 Pre-recorded Broadcast Content

For pre-recorded broadcast, caption ISO BMFF segments (i.e., IMSC1 documents) should be relatively short in duration. This is needed to allow decoders to join an in-progress broadcast and acquire and present caption content concurrent with AV program content.

The time for acquisition and presentation of captions (if present at that moment) should be on the order of the time for acquisition and presentation of video and audio. The IMSC1 document duration therefore typically varies from ½ to 3 seconds. Longer IMSC1 documents, while being more efficient, could result in objectionable delays to the first presentation of caption content.

The IMSC1 timebase shall be “media”.

Note: When fragmenting a caption file it is sufficient to just include all IMSC1 content elements that are active during the sample time period. This will, in the general case, result in begin and end times that are outside the sample duration. It is not necessary when fragmenting the file to clip the begin and end times. This overrides the recommendation in ISO BMFF Part 30 [1] Section 6.3.

6.3 Live Content (Broadband and Broadcast)

For live content, i.e. content that is authored in real time without prediction of the future layout (see Figure 4.2), packaging shall conform to the provisions in this section.

For broadcast, when MMTP is used, an MPU containing the content essence of Section 5 shall have only one sample, a single IMSC1 document per MPU.

Each document shall initially contain a recreation of the previous document's last Intermediate Synchronic Document (see W3C TTML1 [8]). See Annex A. When creating this, encoders should not include content that is entirely outside the sample duration.

When an IMSC1 content element's end time is coincident with the ISO BMFF sample boundary, any such content elements shall be repeated in the following sample's first Intermediate Synchronic Document. This is needed for the decoder to observe the "scroll event" to properly manage smooth scrolling. Without this, the decoder would "jump scroll". See Annex A.

IMSC1 content elements should specify a maximum duration (i.e., not indefinite) up to 16 seconds. This will ensure that the text is automatically erased according to current industry practice (see CTA 608 [11], Section C.9) should there not be a follow on document, in order to avoid "stuck captions".

7. SIGNALING

7.1 Metadata

The following closed caption metadata is signaled:

- Language: the dominant language of the closed caption text.
- Role: the purpose of the closed caption text; e.g., main, alternate, commentary.
- Display aspect ratio: the display aspect ratio assumed by the caption authoring in formatting the caption windows and contents.
- Easy reader: this metadata, when present, indicates that the closed caption text tailored to the needs of beginning readers.
- Profile: this metadata indicates whether text or image profile is used.
- 3D support: this metadata, when present, indicates that the closed caption text is tailored for both 2D and 3D video.

8. DECODER RECOMMENDATIONS

Decoders need to:

- Be able to decode and present both IMSC1 Profiles (text and image) content
- Support smooth scrolling as described in TTML1

Annex A – Live and Broadcast Document Boundary Considerations (Informative)

A.1 INTRODUCTION

This Annex provides information about how to encode several types of caption "modes" found in CTA 608 (and 708). "Pop-on" captioning is straight forward and is not covered here. This annex addresses "paint-on" and "roll-up."

A.2 DISCUSSION

A typical caption timeline is shown in Figure A.2.1. This diagram has a time axis (in seconds) at the top and 2-second duration samples described further below.

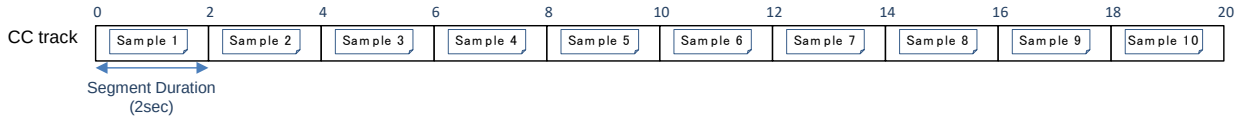


Figure A.2.1 Caption Track Time line.

The contents of sample 1 and the resulting display at the end of the sample period (2 sec.) is shown in Figure A.2.2.

Sample 1 (*)

```
<tt>
  <body>
    <div>
      <p>
        <span begin="0s">Lorem</span>
        <span begin="1s">ipsum</span>
      </p>
    </div>
  </body>
</tt>
```

(*) Following <head> part is common and omitted

```
<head>
  <layout>
    <region xml:id="r1" tts:color="white" tts:origin="10c 4c" tts:extent="40c 2c"/>
  </layout>
</head>
```

Display Image (0-2sec)

- "Lorem" and "ipsum" shall displayed paint-on style

Figure A.2.2 Sample 1 file and resulting display (0–2 sec.).

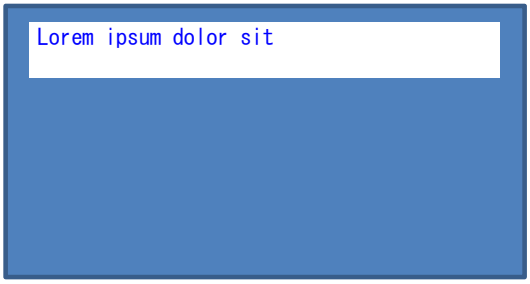
The contents of sample 2 and the resulting display at the end of the sample period (4 sec.) are shown in Figure A.2.3. Note that the first two span lines replicate the display at the end of the previous sample, before adding the new text for current period 2–4 seconds. The new text is "paint on," appended to the prior text.

Sample 2

```
<tt>
  <body>
    <div>
      <p>                                redundant text (*)
        <span begin="0s">Lorem</span>
        <span begin="1s">ipsum</span>
        <span begin="2s">dolor</span>
        <span begin="3s">sit</span>
      </p>
    </div>
  </body>
</tt>
```

(*) When TTML decoder recreates Intermediate Synchronic Document (ISD), This redundant text can tell it is the same as previous document's last ISD and not "flash" the screen.

Display Image (2-4sec)



- TV keeps presenting "Lorem ipsum"
- "dolor" and "sit" shall display after previous text with paint-on style (It shall not display in new line.)

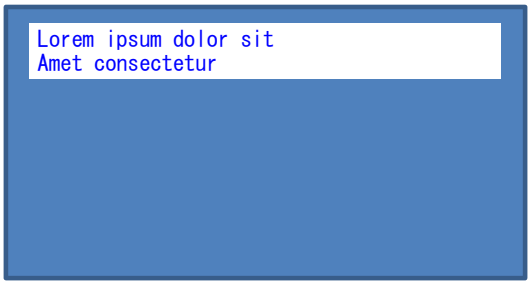
Figure A.2.3 Sample 2 file and resulting display (2–4 sec.).

The contents of sample 3 and the resulting display at the end of the sample period (6 sec.) is shown in Figure A.2.4. Note that the first paragraph (<p>) line replicates the display at the end of the previous sample, before adding the new text for current period 4-6 seconds. (<p> injects a newline.)

Sample 3

```
<tt>
  <body>
    <div>                                  redundant text
      <p begin="4s">
          Lorem ipsum dolor sit
      </p>
      <p> <span begin="4s">Amet</span>
          <span begin="5s">consectetur</span> </p>
    </div>
  </body>
</tt>
```

Display Image (4-6sec)



- TV keeps presenting "Lorem ipsum dolor sit"
- "Amet" and "consectetur" shall display as a new line since begin="indefinite" is not set.

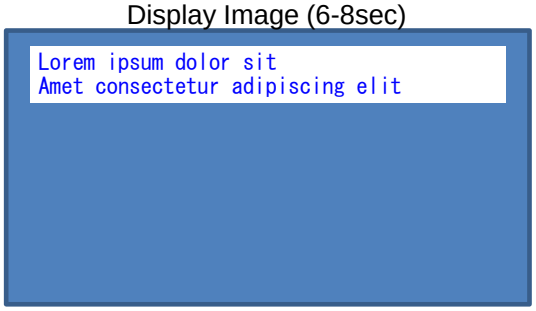
Figure A.2.4 Sample 3 file and resulting display (4–6 sec.).

The contents of sample 4 and the resulting display at the end of the sample period (8 sec.) is shown in Figure A.2.5. This shows additional "paint on" text.

Sample 4

```
<tt>
  <body>
    <div>                                  redundant text
      <p begin="6s">
          Lorem ipsum dolor sit
      </p>
      <p> <span begin="4s">Amet</span>
          <span begin="5s">consectetur</span>
          <span begin="6s">adipiscing</span>
          <span begin="7s">elit</span> </p>
      </p>
    </div>
  </body>
</tt>
```

Display Image (6-8sec)



- TV keeps presenting "Lorem ipsum dolor sit" and "Amet consectetur"
- "adipiscing" and "elit" shall display after previous text with paint-on style

Figure A.2.5 Sample 4 file and resulting display (6–8 sec.)

The contents of sample 5 and the resulting display at the end of the sample period (10 sec.) is shown in Figure A.2.6. In this sample, the prior first line "expires" and scroll off. The old second line and the new text scrolls up. Note that **how** it scrolls (jump versus smooth) is entirely up to the decoder.

Sample 5

```
<tt>                Delete this text for Roll-up
 <body>             (inform to TTML decoder)
   <div>                                  redundant text
     <p end="8s">
        Lorem ipsum dolor sit
     </p>
     <p begin="8s">
        Amet consectetur adipiscing elit
     </p>
     <p> <span begin="8s">Sed</span>
         <span begin="9s">do</span>          </p>
   </div>
 </body>
</tt>
```

Display Image (8-10sec)

- Line of text "Lorem ipsum dolor sit" is disappear because end time is set to 8s.
- TTML decoder recognizes this is a trigger of Roll-up
- "Amet consectetur adipiscing elit" is Roll-up to 1st line.
- "Sed" and "do" shall display in 2nd line.

Figure A.2.6 Sample 5 file and resulting display (8–10 sec.).

A.3 EFFICIENCY

The above examples show how to generate a stream that has minimum latency to displaying the captions at any random access point. This comes at the cost of frequently duplicating information in the transmission and using more bandwidth.

A more bandwidth efficient alternative would be to send larger groups of words at a time. This comes at the cost of increased latency. However, stepping from 2 words at a time (representing about one second of speech) to four words at a time (representing 2 seconds) can cut the needed bandwidth in half. Selecting groups of words larger than a sentence would probably introduce unacceptable delay.

End of Document