Your Host

Joel E. Welch
Director of Education
SMPTE

Housekeeping

- Please use the text chat/Questions box to submit questions to the speaker
- The moderator will pose questions to the presenters on your behalf
- PDF of slides will be provided in exchange for your feedback about the webcast
- Webcast will be recorded and posted for on-demand playback
Today's Speaker

Paul Briscoe

We are not talking about these kinds of lipsync...
What exactly are we talking about?

Maintaining Audio-Video synchronization throughout the media ecosystem from acquisition to consumption.

Acquisition is the easy part
The rest, not so much

What We Will Cover

- Nature of the problem
- Sources of AV sync errors
- Legacy solutions
- Candidate technology
- Fingerprinting the audio and video
- Binding to the essence
- Fingerprinting at the system level
- Standards activity
- Q&A
Nature of the Problem

- Sound and picture are not temporally aligned
- Audience impacts vary with content
  - Live music, sports, SFX, close-up dialogue
  - Depends on what you can see – need motion and sound
- Direction of error has different human sensitivities
  - Late is natural (1 foot = ~1 ms) – brain is trained from birth
  - Early is unnatural - no natural phenomenon (yet)
- Human susceptibility threshold hysteresis effect
  - Once it’s noticed, it’s hard to miss – “can’t be unseen”

Problem? What Problem?

- Diminished Viewer Experience
  - Esthetic enjoyment, QoE
  - Irritation – annoying to watch and listen to
    - Change channel
    - Pay less attention or completely ignore
      - Advertisers don’t like this!
  - Loss of viewer suspension of disbelief
    - Probably why you were watching it
  - Loss of believability of content
    - Advertisers and politicians really don’t like this!
More of a Problem Today Than Ever

- Rich content / less rigid shooting styles
- HD / larger screens (more to see)
- Complex / large-span system designs
  - Uncorrected signal processing, physical delay
- Use of compression – storage and transport
- Complex and multiple distribution architectures
- Modern TV ‘set’ technology – complex = delay
- The internet
- Viewer awareness

Where Can it Creep in?

- Special event venue
  - Mobile studio Encode + Tx

- Network Facility
  - Server IRD Facility Infrastructure Branding + Proc Encode + Tx
  - Lip Sync

- Local Station
  - Local Process Encode + Tx
  - Rx + Decode

- Service Provider (Cable, DTH, IPTV)
  - Encode + Tx
  - Rx + Decode

© 2013 SMPTE
Opportunity for Errors

- Video and Audio may travel separate roads in life
  - Discrete AV signals vs. embedded
  - Different signal processing paths and processes
  - Many forms of transport
  - Compression encoding and decoding
  - In and out of storage
- Human error – routing, patching, QC snoozing
- Can happen in multiple places enroute to consumption
  - Right down to the viewer’s set.

Example of a Subtle Contributor

- Framesyncs ‘fix’ video timing – lock to new reference
- One framesync in the path may not be a problem
- Concatenated framesyncs can produce peak delays which can occasionally trigger the viewer
- Framesyncs require tracking audio delay to maintain lipsync
And it’s Not Just Framesynccs

- Vision mixers (production switchers) – DVE, reentry
- Routing switchers – FS + proc integration
- Cameras – sensor processing latency, rig setup
- Up / Down / Cross format / frame rate conversion
- Compression technologies – takes time on both ends
  - Contribution, storage, editing, distribution
- Viewing environment variability
- System design
- **IP Infrastructures?**

Downstream Used to be so Simple

- After broadcast, there was little opportunity to break it
  - Immediate viewing and listening, essentially zero delay
- Today’s distribution ecosystem is far more complex
  - Codecs - many codecs
  - Satellite, cable, IP, public internet distribution
- New ‘value adds’
  - Commercial substitution
  - Logo / GFX insertion
- Viewing environment variables
  - More vendor choices
  - Set-top boxes, home theatres, media players
### How Bad is it?

- **MPEG1 LII**
  - 41% at +/- 20 ms (past 49%)
  - 52% at +/- 40 ms (past 24%)
  - 4% at +/- 80 ms (past 24%)
  - 3% > +/- 80 ms (past 3%)
  - Max values at -160 ms and +80 ms
- **Dolby Digital/Dolby Digital Plus**
  - 53% < +/- 20 ms (past 44%)
  - 39% at +/- 40 ms (past 34%)
  - 7% at +/- 80 ms (past 23%)
  - 1% > +/- 80 ms
  - Max values at -160 ms and +80 ms
- **HE-AAC**
  - 46% < +/- 20 ms
  - 34% at +/- 40 ms
  - 13% at +/- 80 ms
  - 7% > +/- 80 ms
  - Max values at -160 and +160 ms

### Why is it This Bad?

- **Lack of Standards for broadcasters**
  - Measurement usually involves fallible human interaction
    - Spot measurement, if at all
  - No single interoperable measurement method available
    - Particularly necessary for in-service usage
- **Lack of Standards for distributors and consumers**
  - MPEG encoding is normative, decoding is not
    - CEA CEB-20 addresses improvement in decoder behaviour
  - Internet-type codec / player behaviour is highly variable
Measurement Today 1

- Out-of-service measurement
  - Simple techniques
  - Some can be automated
    - Many vendor solutions
  - Relatively foolproof at time of use
    - If things change later, you don't know
  - Human involvement
  - Can't be used within content or on-air
    - Generally for acquisition, editing and system testing

Measurement Today 2

- In-service measurement
  - Can be used on air within content
  - Various manufacturers
  - Various techniques, varying capabilities
  - Intended for use within broadcast plant
  - Varying degrees of complexity
  - May not traverse all processing
  - Box or module level solutions
  - NON-INTEROPERABLE among manufacturers
    - Non-Standardized

- Manufacturers
  - Dolby
  - Evertz
  - Miranda
  - Sigma
  - K-Will
  - Astro Design
  - Asaca
    - And more
Desired Measurement Capability

- **Standardized in-service measurement**
  - Can be used on air within live content at anytime
  - Medium-agnostic
  - Low degree of complexity – can be on a port of any device
  - INTEROPERABLE among manufacturers
  - Works throughout chain, not just in-plant
  - Traverses all processing, even when concatenated
    - *Potentially right to the home viewer*

Candidate Technologies

- **Watermarking**
  - Insertion of invisible / inaudible information into video and audio essence for later detection and extraction.
    - Complex processing required to generate and extract
    - Modifies content, may not survive all signal processing
    - May not coexist well with other watermarks
    - Can be removed

- **Fingerprinting**
  - Measurement of specific properties of video and audio essence and coding into metadata for downstream use.
    - Simple processing to generate fingerprints
    - Does not modify content
    - Can survive processing (incl. compression, ARC, etc..)
    - Cannot be removed (because it’s not there!)
Winning Technology

- Watermarking
  - Insertion of invisible / inaudible information into video and audio essence for later detection and extraction.
    - Complex processing required to generate and extract
    - Modifies content, may not survive all signal processing
    - May not coexist well with other watermarks
    - Can be removed

- Fingerprinting
  - Measurement of specific properties of video and audio essence and coding into metadata for downstream use.
    - Simple processing to generate fingerprints
    - Does not modify content
    - Can survive processing (incl. compression, ARC, etc..)
    - Cannot be removed (because it's not there!)

Fingerprint Robustness

- Up/Down conversion
- Encoding/Decoding for video and audio compression
- Audio Level Change, Up/Downmixing, Loudness
- Insertion of logos, lower thirds, and other graphics
- Resizing and spatial format conversion
- Frame rate/temporal conversion
- Cropping
- Signal distortions – filtering
Fingerprinting Steps

- Generate audio and video fingerprints (in sync)
- Generate fingerprint metadata payload
- Deliver audio and video to endpoint (or midpoint)
- Deliver fingerprint metadata to endpoint
- Generate audio and video fingerprints (unknown sync)
- Correlate source and local fingerprints in one essence
- Measure differential delay in the other essence
- Use measured delay to drive display / correction

Generalized System View
What’s in the Standard (and what’s not)

Stream Transport

- Extract Audio Signature
- Extract Video Signature
- Generate A/V Sync Signature

V/A Sync Signature Generator

Content Distribution Network

- Variable Processing and Delays
- Stream Transport File Binding

Variable Delay

A/V Sync Detector

- Extract Audio Signature
- Extract Video Signature
- Signature Comparisons

Meter Display

Unknown A/V sync

Audio and Video are again “in sync”

Variable Delay

Audio and Video are known to be “in sync”

Audio

Video

Functional View

A/V Sync Signature maintains time relationship of individual audio and video signatures

Audio delay

Video delay

Sent in A/V Sync Signature

A/V Sync Signature

Ref Audio

Ref Video

Audio and Video in Sync

Audio and Video Unknown Sync

Compare Signatures

Reliability Estimation

Compare Delays

A/V Sync Signatures

Audio delay

Video delay

Ref Video Signature

Difference between audio delay and video delay is the A/V sync error

© 2013 SMPTE
Fingerprinting Algorithms

- Measure simple characteristics of change over time
  - Frame to frame (field to field) for video
  - Sample to sample for (digital) audio
- Cost effective to implement
  - No big DSPs, FPGA logic commitment, software
- Tolerant of essence processing
- Generate low data rates (metadata)

Video Fingerprinting

- Use Fields (-i) or frames (-p) for processing
- Simple H downscaling of HD to SD
- Establish central window
- Measure specific samples within window (960)
- Compare with samples from prior field / frame
- Count number samples with change >32
- Divide by 4 (to get a single byte)
- This is the video fingerprint value for that field / frame
SMPTE Standards Update: The Lip-Sync Challenge

Video Prefiltering

- Simple downscale from HD to SD (no change to SD)
- Horizontal only
- Fields if interlaced
- Frames if progressive

<table>
<thead>
<tr>
<th>Video Format</th>
<th>Prefilter Used</th>
</tr>
</thead>
<tbody>
<tr>
<td>1080i, 1080p</td>
<td>[1 1 1] / 3</td>
</tr>
<tr>
<td>720p</td>
<td>[1 1 0] / 2</td>
</tr>
<tr>
<td>SD 525, SD 625</td>
<td>[0 1 0] / 1</td>
</tr>
</tbody>
</table>

Video Windowing

- Fixed coordinates for each standard
- Scaled images overlay each other geometrically
- Start and end samples stay away from picture edges

<table>
<thead>
<tr>
<th>Video Format</th>
<th>Window</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>VStart (f1)</td>
</tr>
<tr>
<td>720 X 480i</td>
<td>123</td>
</tr>
<tr>
<td>720 X 576i</td>
<td>123</td>
</tr>
<tr>
<td>1280 X 720p</td>
<td>256</td>
</tr>
<tr>
<td>1920 X 1080i</td>
<td>399</td>
</tr>
<tr>
<td>1920 X 1080p</td>
<td>399</td>
</tr>
</tbody>
</table>

© 2013 • Powered by SMPTE Professional Development Academy • www.smpte.org
Video Motion Detection

- Compares current frame / field to second previous
- *Keeps –i / -p conversions happy*

![Diagram of Video Motion Detection](image)

Video Motion Detection

- Y channel (luma) data is truncated to 8-bits
- Count number of selected samples which have changed by >32, divide by 4
- Result (byte) is video fingerprint

\[
VS(f) = \sum_{k=1}^{N} \frac{\text{abs}(P_k - C_k)}{4} \begin{cases} 1 & (\text{if } \text{abs}(P_k - C_k) \geq 32) \\ 0 & \text{otherwise} \end{cases}
\]
Audio Fingerprinting

- Use 48 KHz samples, SRC other rates
- Truncate to 16 bits
- Generate mean value (long time constant)
- Generate envelope value (short time constant)
- For each sample,
  - If envelope > mean => 1
  - If envelope < mean => 0
- Accumulate 1’s and 0’s for entire video frame
- Decimate to 1 ms resolution

Audio Mean Detection

\[ a_{\text{wav}} = \text{Absolute value of original unfiltered audio} \]
\[ K_m = \text{Local mean IIR filter coefficient} \]
\[ M_s = \text{Mean signal} \]
\[ M_s(0) = 0; \quad \text{// local mean detector IIR filter coefficient} \]
\[ K_m = 8192; \quad \text{// init first value to a known state} \]

\[
\text{for( } i = 1; i < \text{max_sample}; i++ \) \\
\{ \\
\quad M_s(i) = a_{\text{wav}}(i) + M_s(i-1) - \text{floor}(M_s(i-1) / K_m); \\
\}
\]

\text{Km is set to 8192, which is large enough to emulate a local mean function.}
Audio Envelope Detection

\[ a_{wav} = \text{Absolute value of original unfiltered audio} \]
\[ K_e = \text{Envelope detector IIR filter coefficient} \]
\[ E_s = \text{Envelope signal} \]

// Envelope detector IIR filter
Ke = 1024; // envelope detector IIR filter coefficient
Es[0] = 0; // init first value to a known state
for (i = 1; i < max_sample; i++)
{
    Es[i] = a_wav(i) + Es[i-1] - floor(Es[i-1] / Ke);
}

Ke is set to 1024 which is small enough to reproduce the audio envelope.

Envelope / Mean Comparison

- Envelope is compared to mean
- Output is stream of bits at sample rate (48K)

// extract fingerprint by comparing envelope with local mean
for (i=0; i < max_sample; i++)
{
    if (M_s(i) < (E_s(i) * Km / Ke))
        comp_bit(i)=1;
    else
        comp_bit(i)=0;
}
Audio Sample Data Reduction

- Decimator loop reduces amount of data
  
  ```
  For (i = 0; i < max_sample; i += decimator_factor)
  {
    result(i / decimator_factor) = comp_bit(i);
  }
  
  Algorithm keeps 1 bit per decimator loop.
  
  Result is some number of bytes per video frame
  ```

<table>
<thead>
<tr>
<th>Video Frame Rate</th>
<th>Decimator Factor</th>
<th>Bits per x frames</th>
<th>Bitrate per Second</th>
</tr>
</thead>
<tbody>
<tr>
<td>23.98</td>
<td>52</td>
<td>616 per 16 frames</td>
<td>~923 b/s</td>
</tr>
<tr>
<td>29.97</td>
<td>52</td>
<td>616 per 20 frames</td>
<td>~923 b/s</td>
</tr>
<tr>
<td>59.94</td>
<td>52</td>
<td>616 per 40 frames</td>
<td>~923 b/s</td>
</tr>
<tr>
<td>24</td>
<td>50</td>
<td>640 per 16 frames</td>
<td>960 b/s</td>
</tr>
<tr>
<td>25</td>
<td>50</td>
<td>768 per 20 frames</td>
<td>960 b/s</td>
</tr>
<tr>
<td>30</td>
<td>50</td>
<td>640 per 20 frames</td>
<td>960 b/s</td>
</tr>
<tr>
<td>50</td>
<td>50</td>
<td>640 per 40 frames</td>
<td>960 b/s</td>
</tr>
<tr>
<td>60</td>
<td>50</td>
<td>640 per 40 frames</td>
<td>960 b/s</td>
</tr>
</tbody>
</table>

What Audio to Fingerprint?

- Multichannel audio is downmixed to mono
  - Within a soundfield, it’s all the same audio
  - Individual soundfield channels can also be fingerprinted

- Multi-track / Multi-language audio supported
  - Up to 32 audio fingerprints may be associated with a video

- Not part of the standard, to be determined by individual operating practice.
Fingerprint Encapsulation

- Containerizing fingerprints and helper data
- One container per video field / frame
  - Protocol Version
  - Video fingerprint
  - Audio fingerprint
  - Sequence count
  - Status bits
  - ID Descriptor
  - Checksum
- Same container used for all binding applications
- Enables easy inter-media interchange

Transport and Binding

- Same payload delivered by different means
- SDI VANC Metadata
- MPEG-2 Transport Stream
- UDP/IP Packets
- File Binding
- Same metadata in all domains
- Easy to move between
SDI Transport

- Carried in ST291 VANC Packets
- Unique / registered DID / SDID
- Inherently bound to the essence
- Essentially lossless and error-free
  As much as SDI is error-free

UDP / IP Transport

- Raw UDP Packets
- Indirectly bound to essence
  - IP address
  - ID Descriptor
- Requires re-association at receiver
- May be errored / lossy, out-of-order
MPEG-TS Transport

- Private user data in TS
- Unique PID
- Inherently bound to essence (via maps)

- Essentially lossless and error-free
  As much as an MPEG TS is error-free

File Binding

- MXF Files
  - MXF-specific method
- Arbitrary media files
  - File-agnostic method

Under development!
Prototyping and Testing

- Test Road Kit built and deployed for testing
- Testplan / results spreadsheet
- Sent to several large users with the instruction “see if you can break it, and if so, how”.
- Results very successful – no reasonable failure modes

An Example Test Configuration
SMPTE Standards Activity

- Lipsync Ad-Hoc Group 24TB-01 AHG Lipsync
- Weekly meetings, strong core group of SMEs
- ST2064 Document Suite
- Developing multiple documents:
  - Fingerprint generation and encapsulation (Part 1)
  - Fingerprint transport binding (Part 2)
  - Fingerprint file binding (Parts 3+)

Document Status:
- Parts 1 and 2 – Committee Drafts ready for FCD Ballot
- Part 3 – work underway

Beyond SMPTE

- After emission, then what?
  - "looks good here, must be your set"
- Method will work right down to the point of consumption
  - Cost-effective enough to endure downstream price points
- Additional binding and transport standards may be required
- Standards in other bodies maybe required
  - CEA, DVB, other Liaisons established
- Can be used in walled garden ecosystems
  - Netflix, YouTube, etc..
- Can be added to content retrospectively, used anywhere
Where Does It All End?

• Maybe here.

• This standard offers a toolset to solve lipsync issues independently of underlying media systems and technologies

• Metadata from the upstream essence is made available to downstream devices for measurement and correction - a “wrapper” around the system

• Can offer ‘self-healing’ system behaviour

In Conclusion

• Standardized interoperable lipsync measurement. Coming soon to a signal near you.

• Will enable vendors to offer broad range lipsync measurement and correction capabilities across many devices at low cost

• If deployed end-to-end in a system, can enable assured AV synchronization throughout.
Thank-you!

Questions, comments, further dialogue? Please feel free to drop me a note.
Paul Briscoe
hdtvpeb@gmail.com