MPEG-4 Audio Synchronization

Masayuki Nishiguchi, Shusuke Takahashi, Akira Inoue

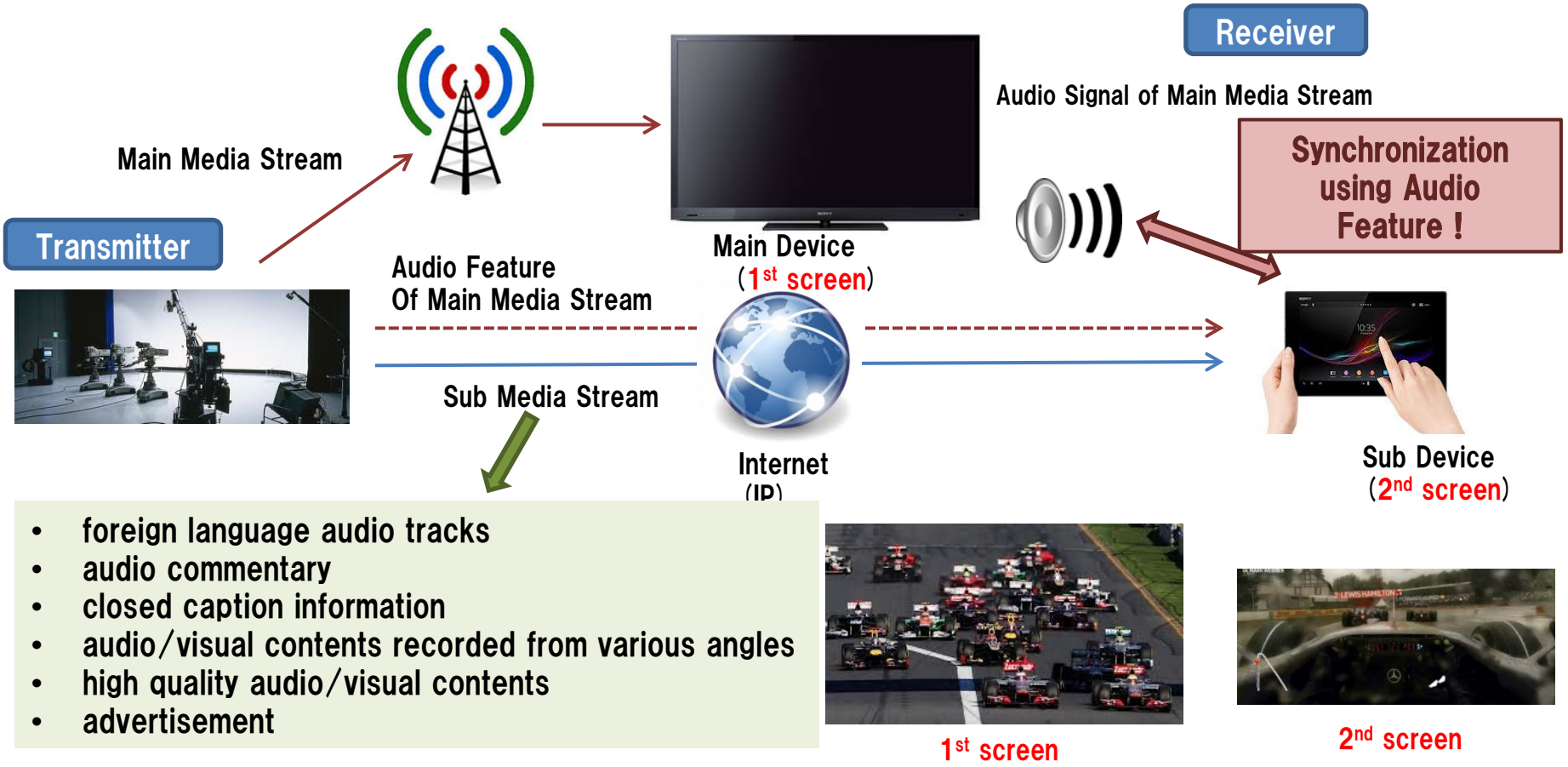
Oct 22, 2014

Sony Corporation

Agenda

- Use case
- Synchronization Scheme
- Audio Feature Extraction tool (Normative)
- Audio Feature Similarity Calculation Tool (Informative)
- Performance evaluation
- Conclusion

Audio Synchronization
Use case of “Second Screen” Application



Synchronization Scheme

Transmitter

Receiver (Sub device)

Main Device

Main Media Stream

(Synchronous in time)

Sub Media Stream

Audio Signal

Audio Feature Extraction

Audio Feature of Main Media Stream

MUX

DeMUX & Playback

Audio Playback device (such as speaker)

1st screen

Main Media Presentation

noise

Audio Recording Device (such as microphone)

Audio Signal of Main Media Stream

Sub Device

2nd screen

Audio Feature Extraction

Audio Feature of Main Media Stream (Transmitted)

Audio Feature of Main Media Stream (Extracted)

DeMUX

Audio Feature Similarity Calculation

Timing Adjustment & Playback

Sub Media Presentation

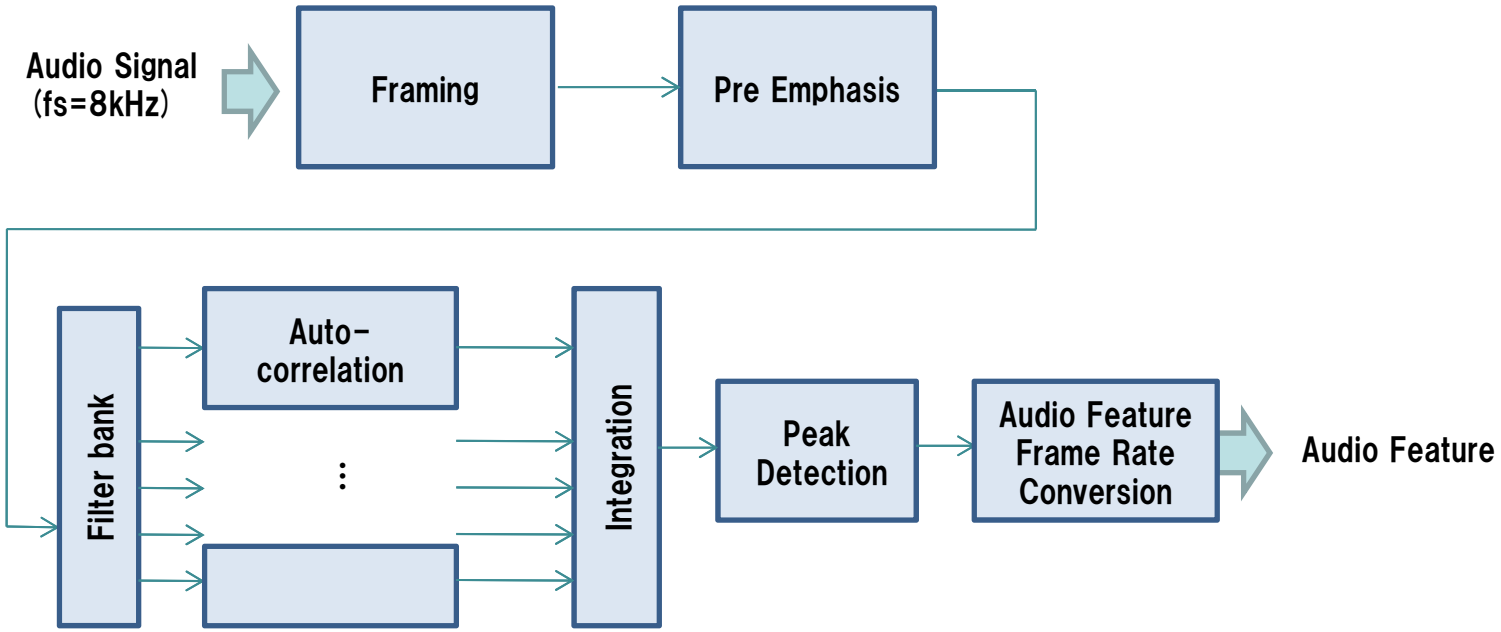
Sub Media Stream

Synchronization Information

Multiplexed Data Stream
- Sub Media Stream
- Audio Feature of Main Media Stream

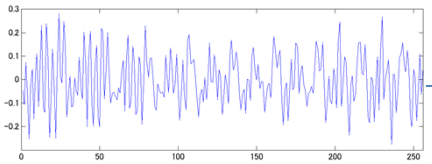
Audio Feature Extraction tool (Normative)

Block Diagram of Audio Feature Extraction tool



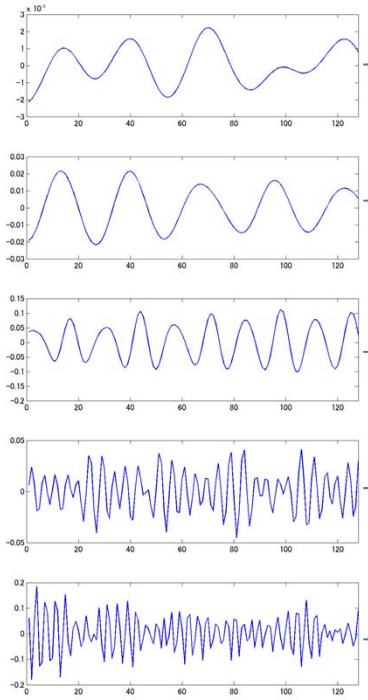
Overall Signal Flow

split the audio signals into 5 equally spaced frequency bands in log frequency domain

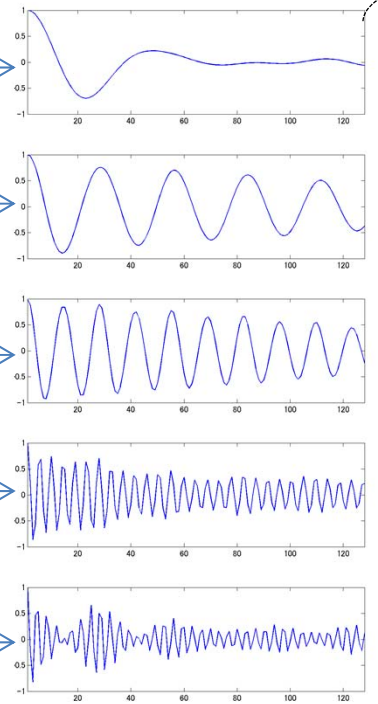


Input Signal
(After Pre emphasis filter)

$H(z) = 1 - 0.97 \cdot z^{-1}$



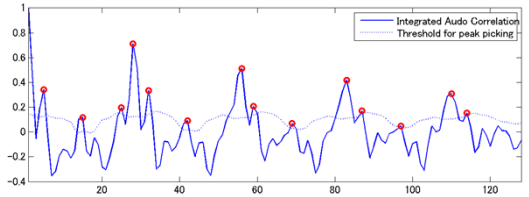
Band split signal



Auto Correlation

Confidence Measure for this band is less than threshold

Integrate together into single Auto Correlation



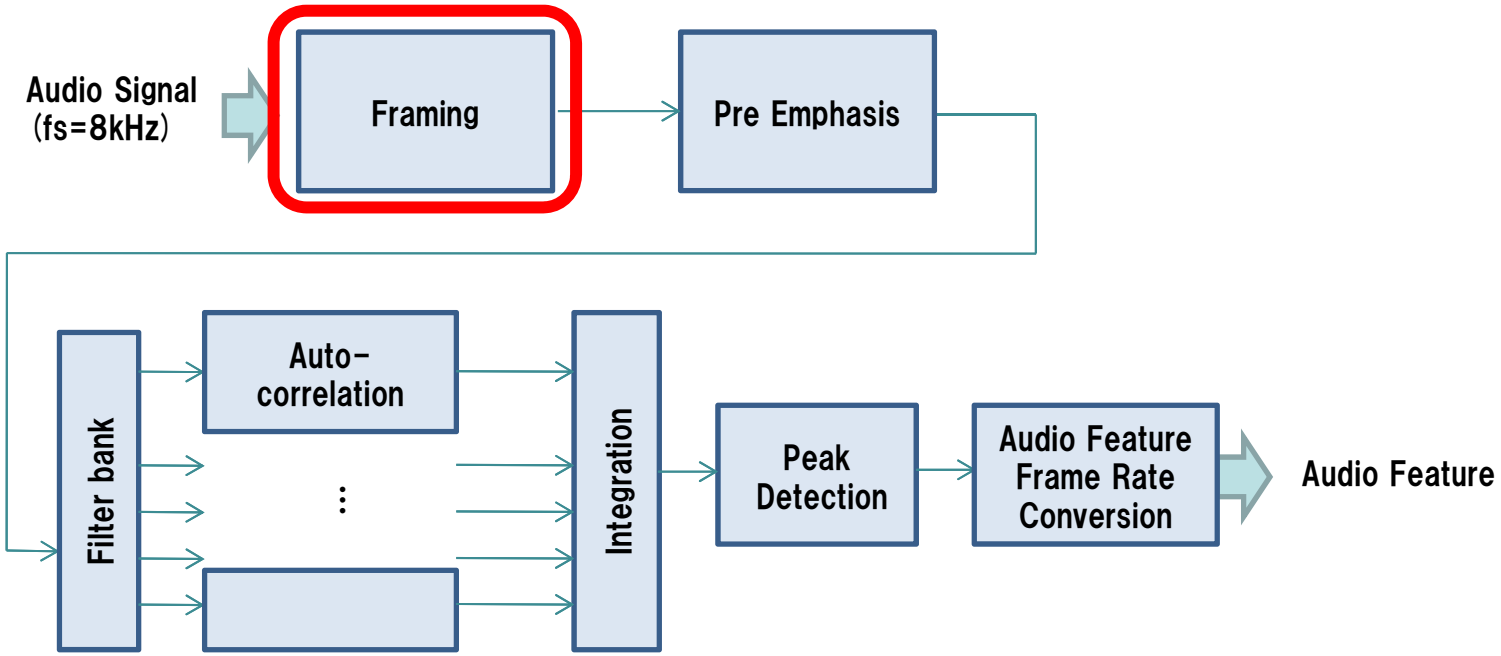
Integrated Auto Correlation

converted into a 128-bit length feature vector $f(k)$

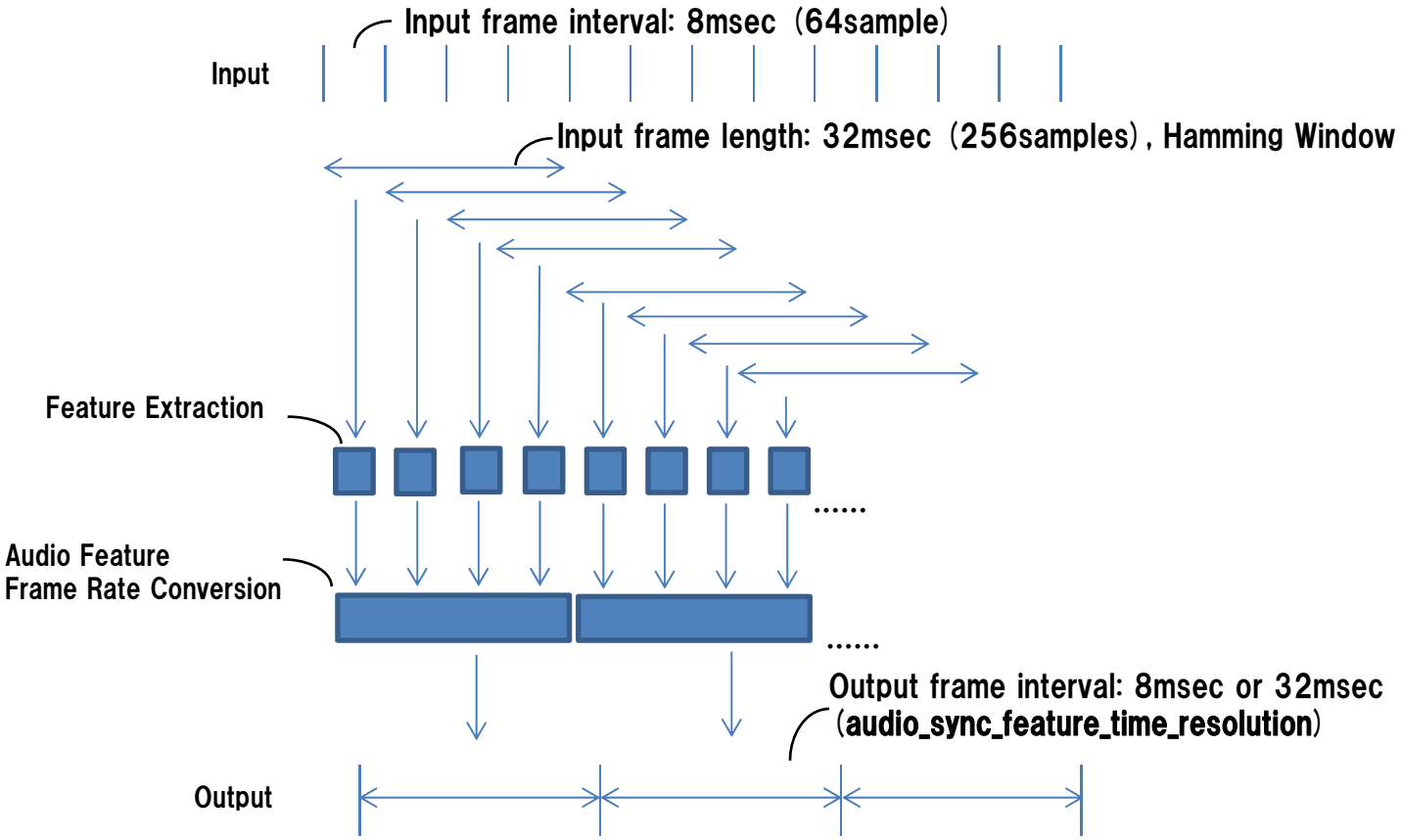
Prominent peak : 1

Otherwise : 0

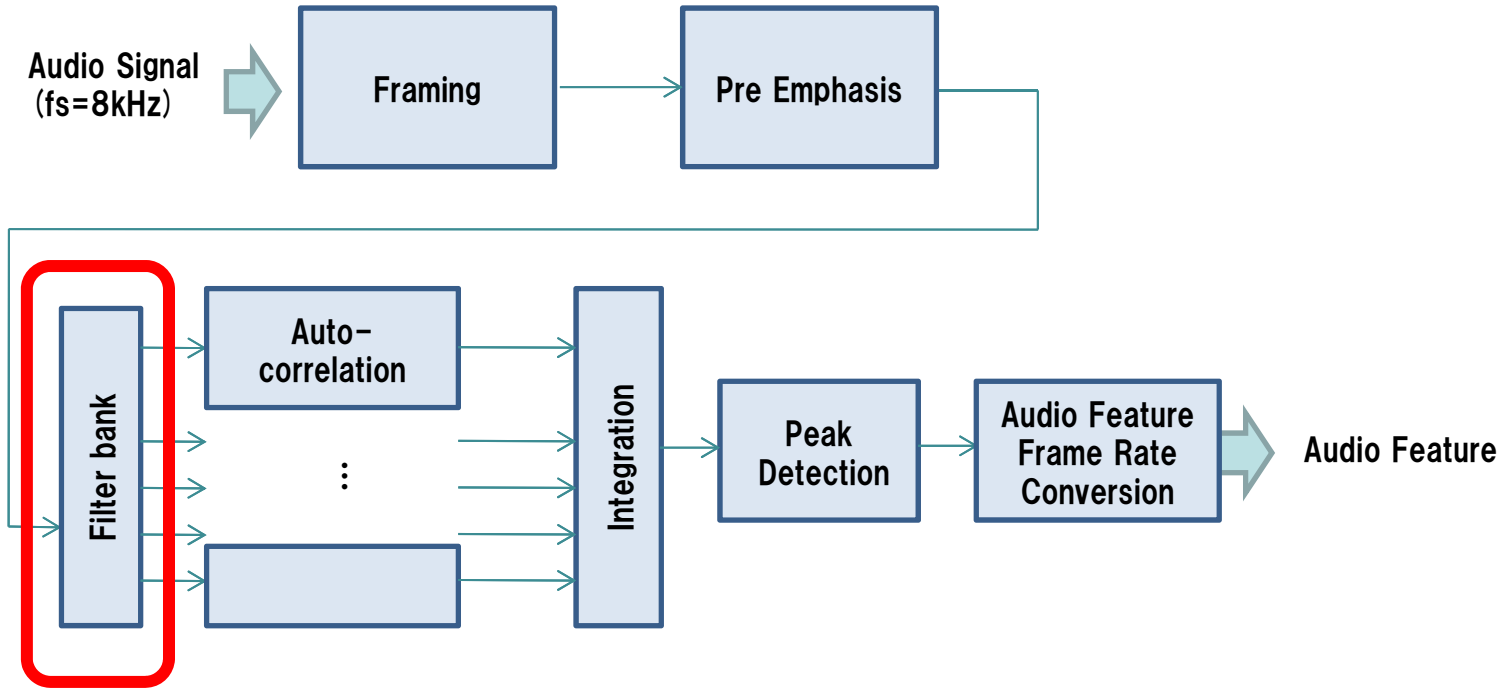
Block Diagram of Audio Feature Extraction tool



Framing

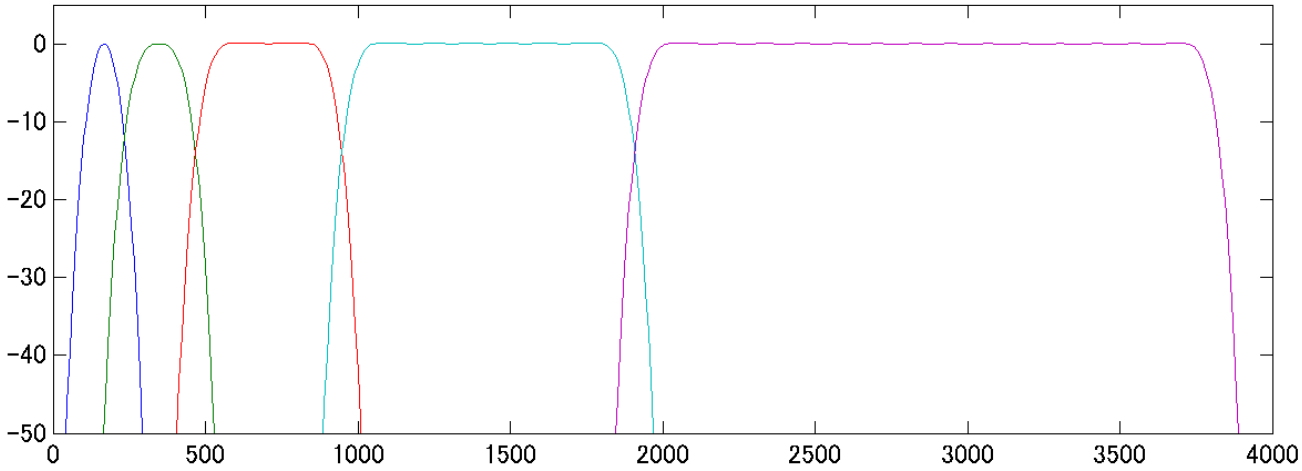

Input frame interval: 8msec (64sample)
Input
Input frame length: 32msec (256samples), Hamming Window
Feature Extraction
.....
Audio Feature
Frame Rate Conversion
.....
Output frame interval: 8msec or 32msec
(audio_sync_feature_time_resolution)
Output

Block Diagram of Audio Feature Extraction tool

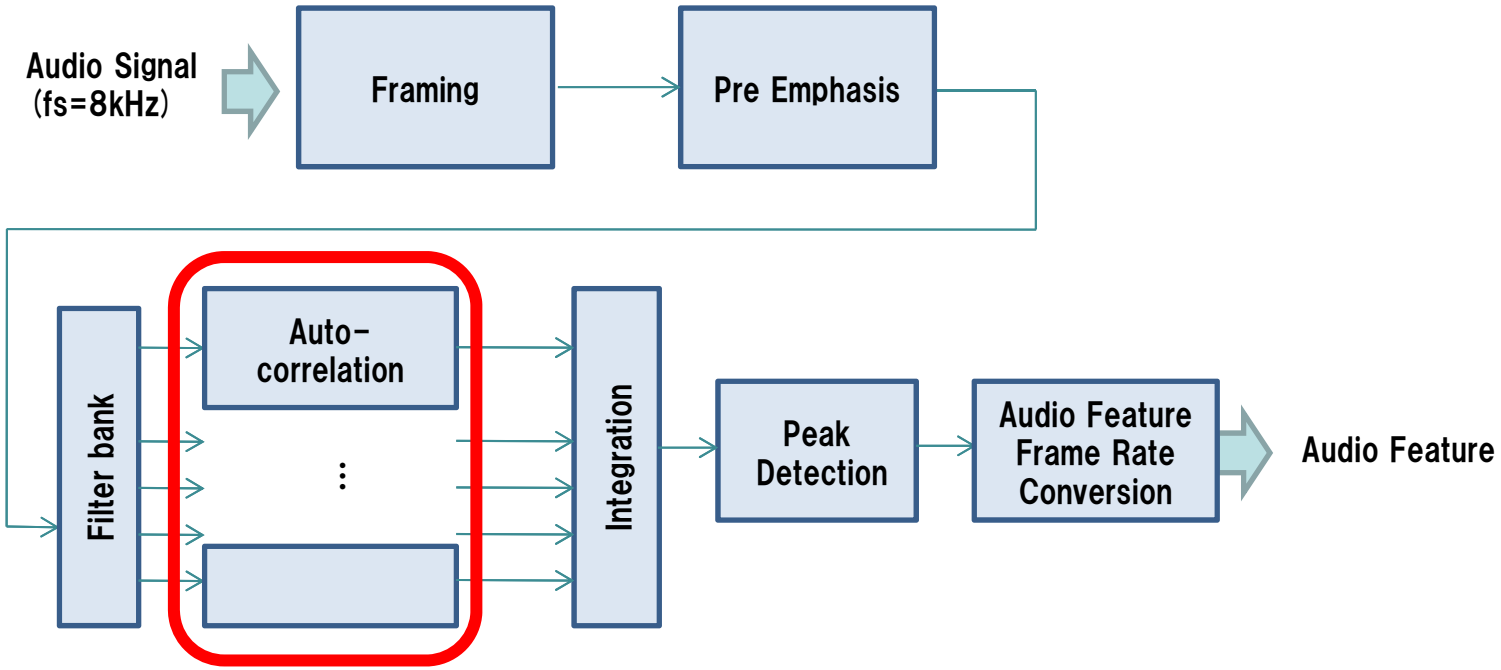


Filter Bank

For each audio frame, a pre-emphasis filter is applied to emphasize the high frequency, then band pass filtering is applied in order to split the audio signals into 5 equally spaced frequency bands in log frequency domain.



Block Diagram of Audio Feature Extraction tool



Auto-correlation

For each band, Auto-correlation is calculated using:

$$ACF_m(k) = \sum_{n=0}^{N-1} x_m(n) \cdot x_m(n+k), \qquad 0 \le k < K, \qquad 0 \le m < M$$

The Auto-correlation is normalized using:

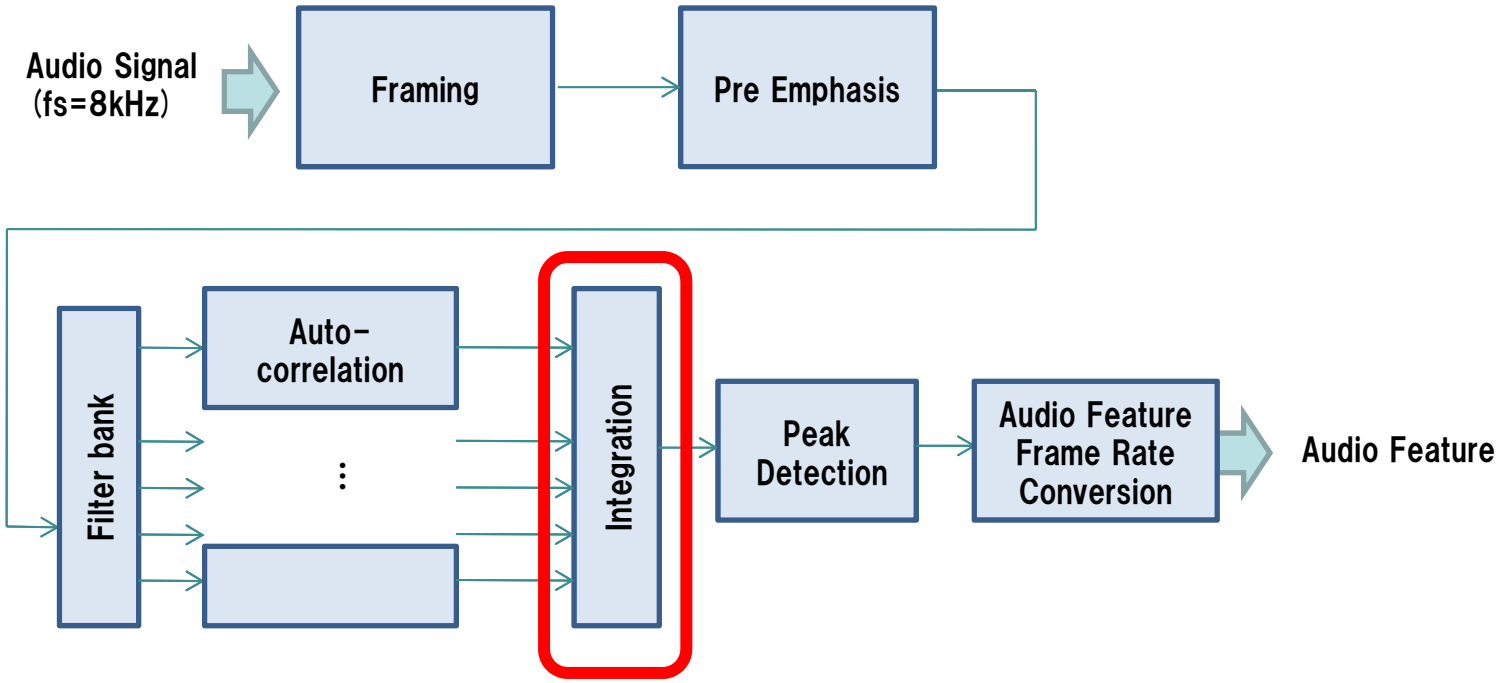
$$NACF_m(k) = \frac{ACF_m(k)}{ACF_m(0)} \quad 0 \le k < K, \qquad 0 \le m < M$$

N : input frame length,
m : index of frequency band
k : index of lag for autocorrelation
K : order of auto-correlation and is set to 128,
n : index of the input audio signal.
M : number of frequency bands and is set 5

For each frequency band m, confidence measure CM_m is calculated based on the auto-correlation value.

$$CM_m = \max_{10 \le k \le K-1} NACF_m(k), \qquad 0 \le m < M$$

Block Diagram of Audio Feature Extraction tool



Integration

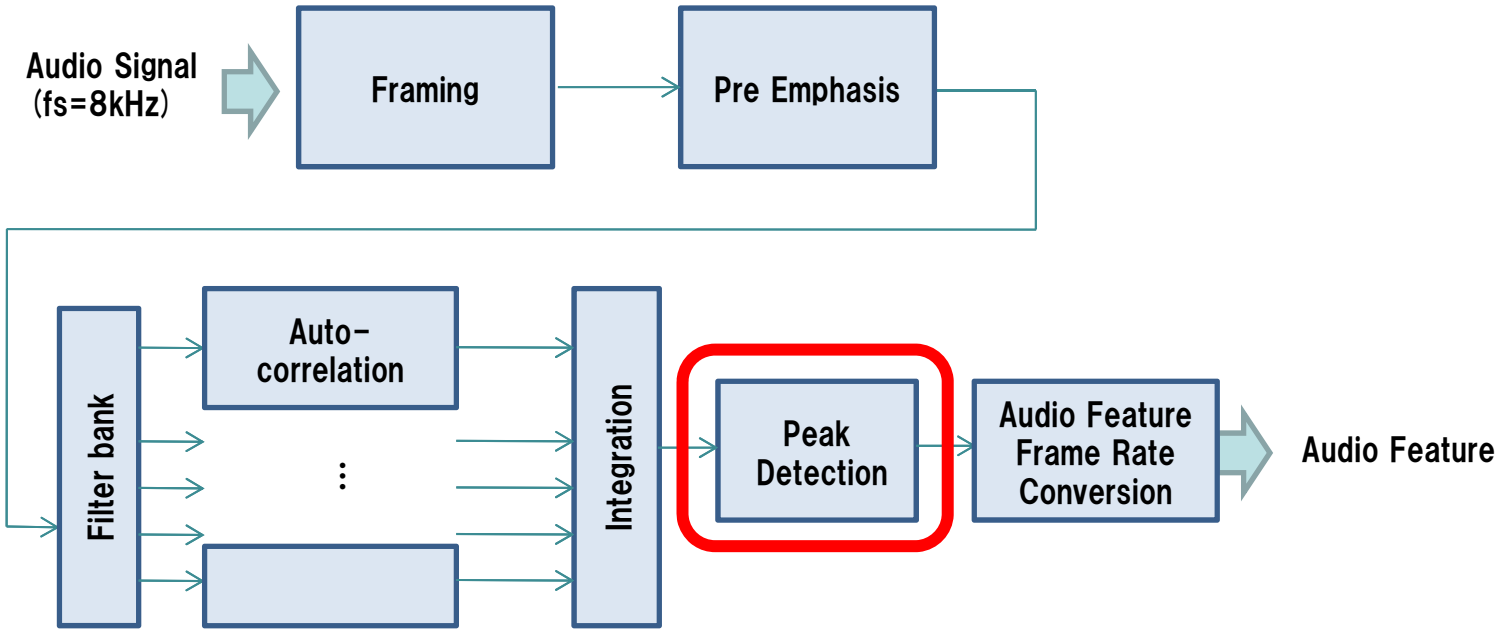
The normalized auto-correlation function $NACF_m(k)$ values derived from each sub-band are summed together into a single integrated auto-correlation function.

$$ACF_{integrated}(k) = \frac{\sum_{m=0}^{N_b-1} NACF_m(k) \cdot W_m}{\sum_{m=0}^{N_b-1} W_m}, \qquad 0 \le k < K$$

where W_m is defined as following

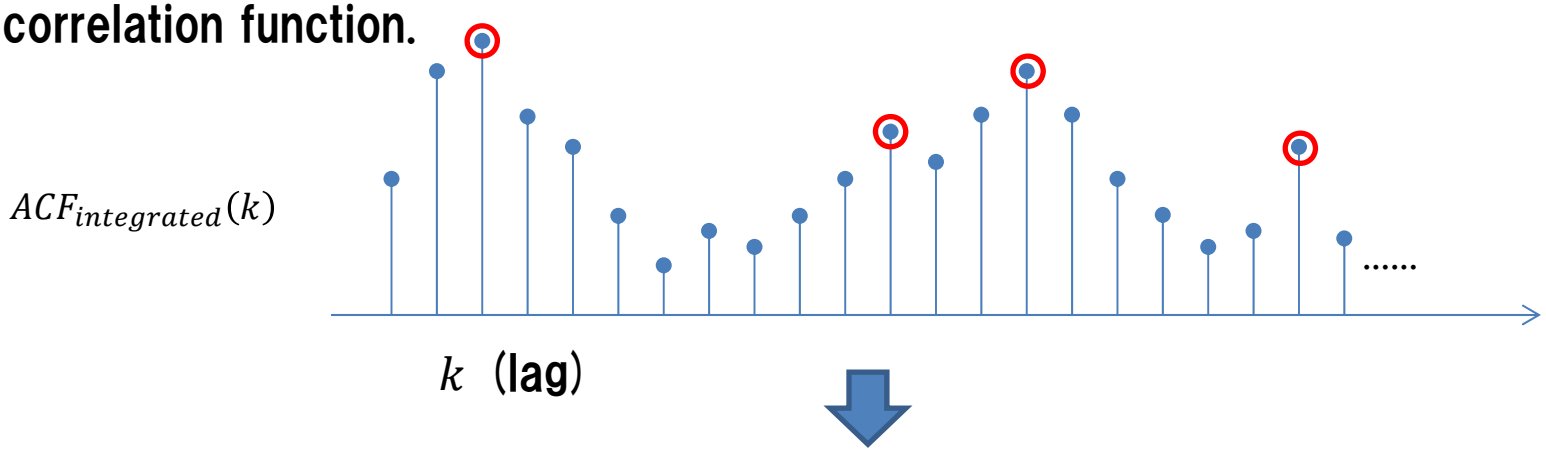
$$W_m = \begin{cases} 0, & CM_m < 0.3 \\ 1, & CM_m \ge 0.3 \end{cases}$$

Block Diagram of Audio Feature Extraction tool



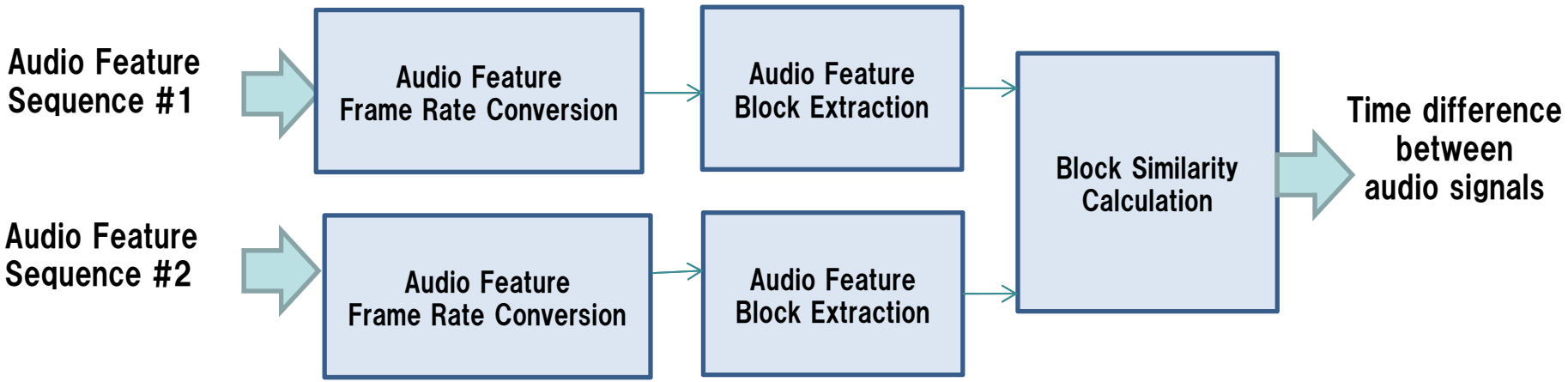
Peak Detection

The integrated auto-correlation function is converted into a 128-bit length feature vector $f(k)$ and each bit position corresponds to the lag of the auto-correlation function.

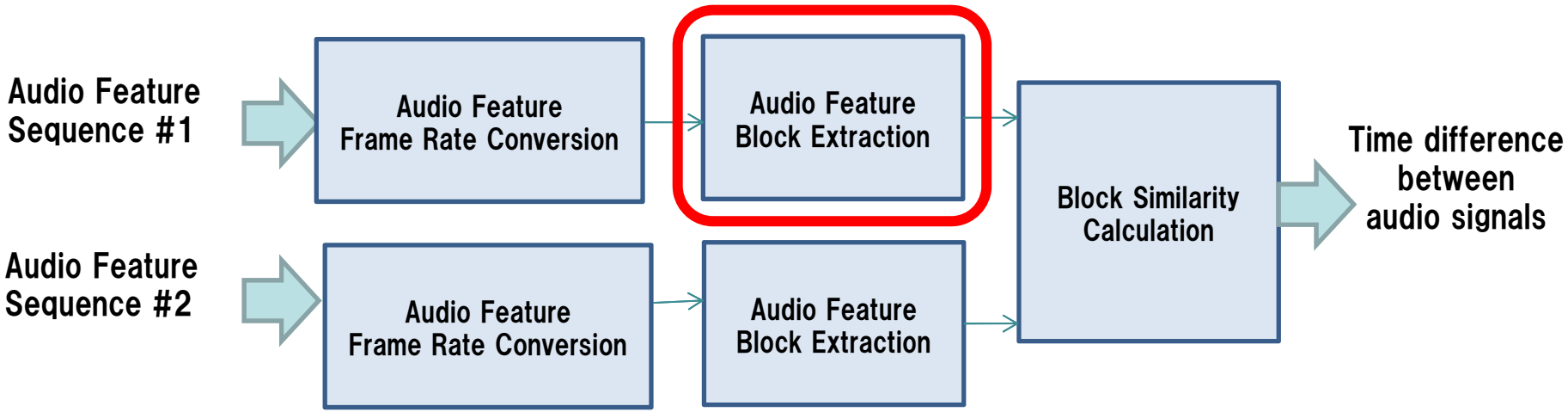


Audio Feature Similarity Calculation Tool (Informative)

Block Diagram Audio Feature Similarity Calculation Tool (Informative)

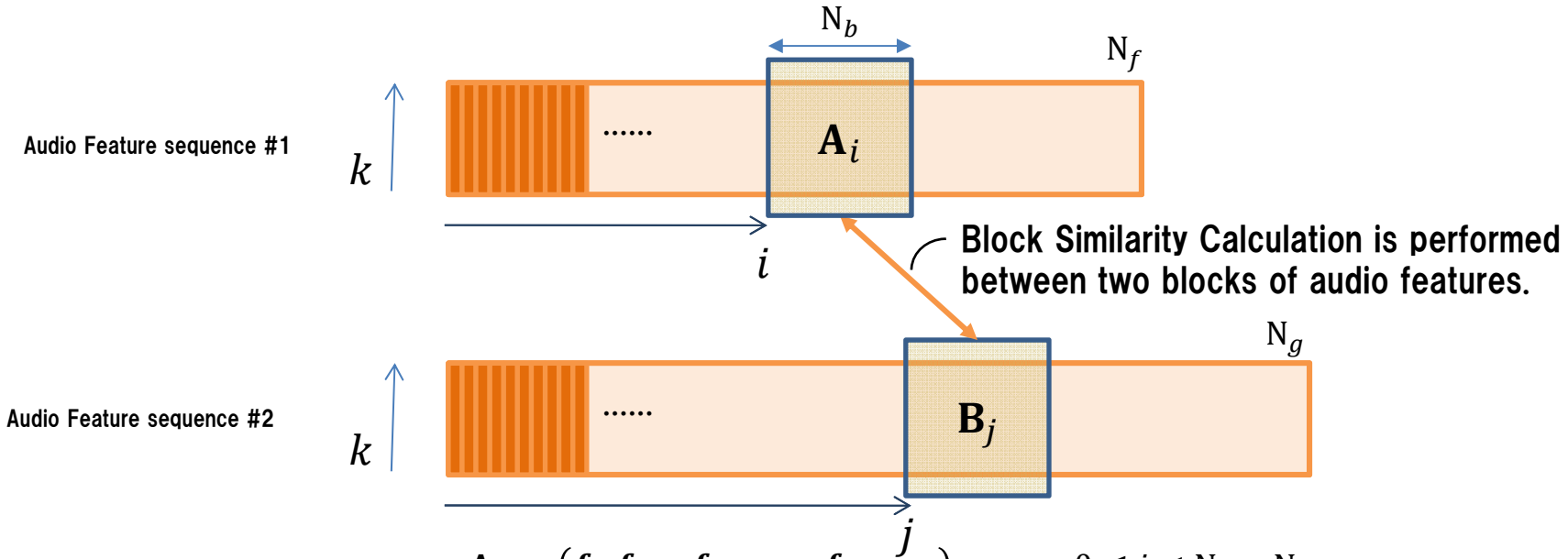


Block Diagram Audio Feature Similarity Calculation Tool (Informative)



Block Extraction

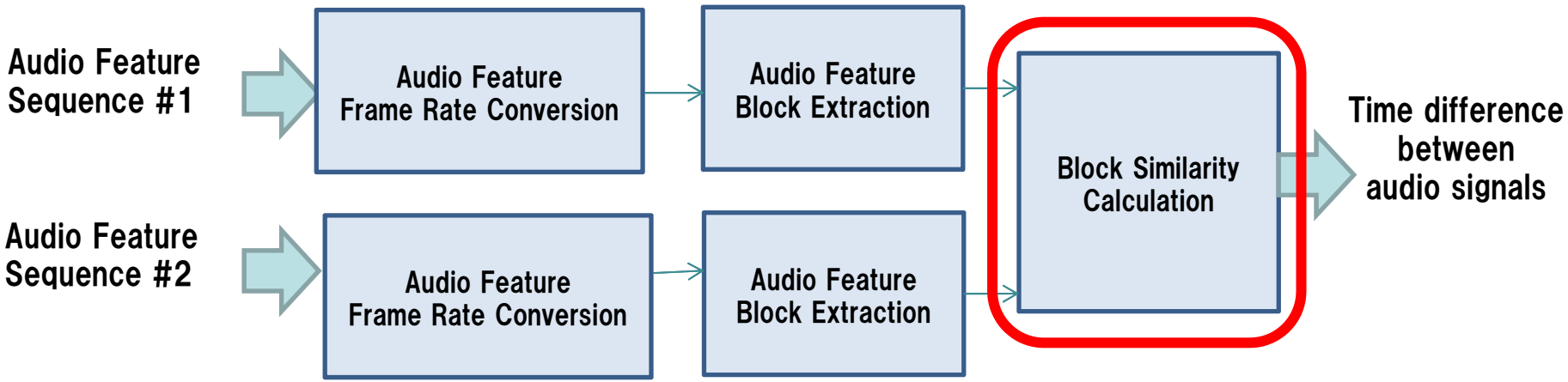
The blocks are generated by concatenating the consecutive audio features



$$\mathbf{A}_i = (\boldsymbol{f}_i, \boldsymbol{f}_{i+1}, \boldsymbol{f}_{i+2}, \ldots, \boldsymbol{f}_{i+N_b-1}), \qquad 0 \le i < N_f - N_b$$

$$\mathbf{B}_j = (\boldsymbol{g}_j, \boldsymbol{g}_{j+1}, \boldsymbol{g}_{j+2}, \ldots, \boldsymbol{g}_{j+N_b-1}), \qquad 0 \le j < N_g - N_b$$

Block Diagram Audio Feature Similarity Calculation Tool (Informative)

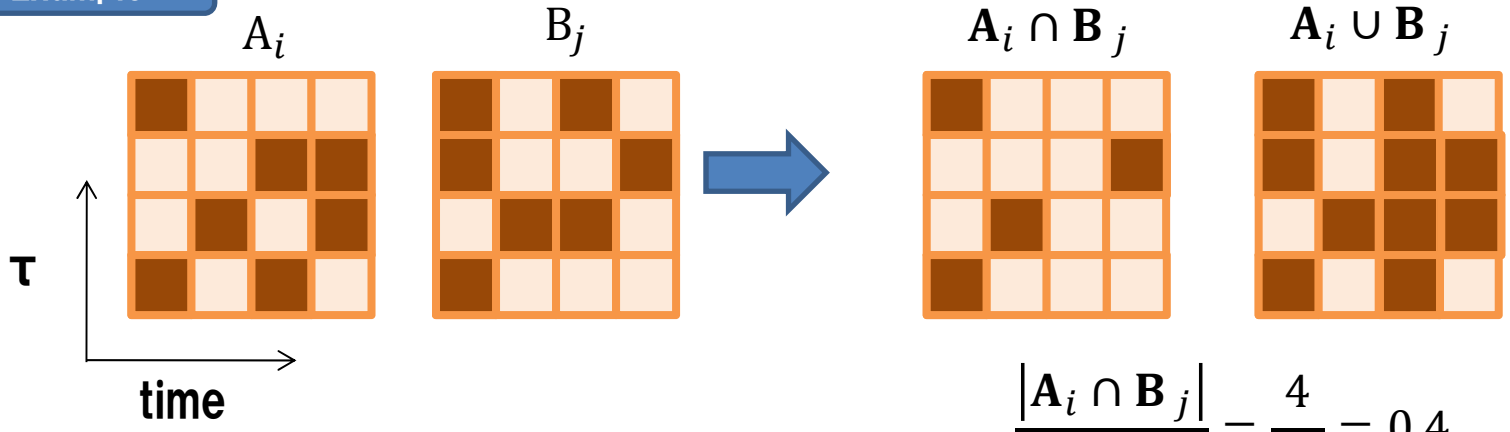


Block Similarity Calculation

Block Similarity between $\mathbf{A}_i$ and $\mathbf{B}_j$ is calculated as follows:

$$J(A_i, B_j) = \frac{|\mathbf{A}_i \cap \mathbf{B}_j|}{|\mathbf{A}_i \cup \mathbf{B}_j|}$$

Example



$$\frac{|\mathbf{A}_i \cap \mathbf{B}_j|}{|\mathbf{A}_i \cup \mathbf{B}_j|} = \frac{4}{10} = 0.4$$

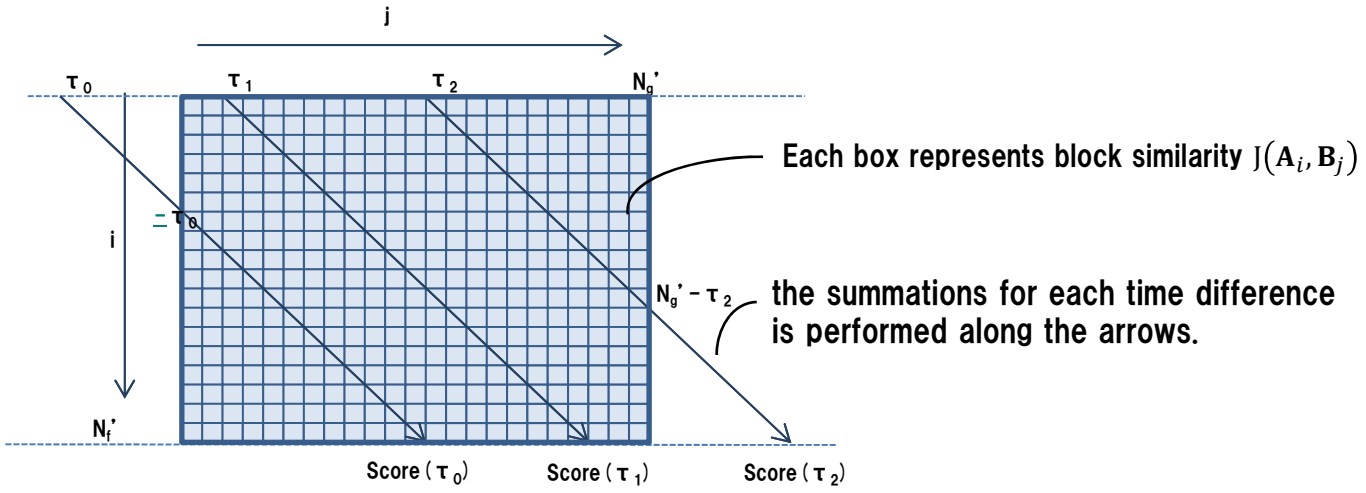
Time Difference Estimation

For each time difference τ, a score is calculated by using the block similarity as follows:

$$\mathrm{Score}(\tau) = \frac{1}{\min(N_g{}' - \tau, N_f{}') - \max(-\tau, 0)} \sum_{i=\max(-\tau,0)}^{\min(N_g{}'-\tau, N_f{}')} \mathrm{J}(\mathbf{A}_i, \mathbf{B}_{i+\tau})$$

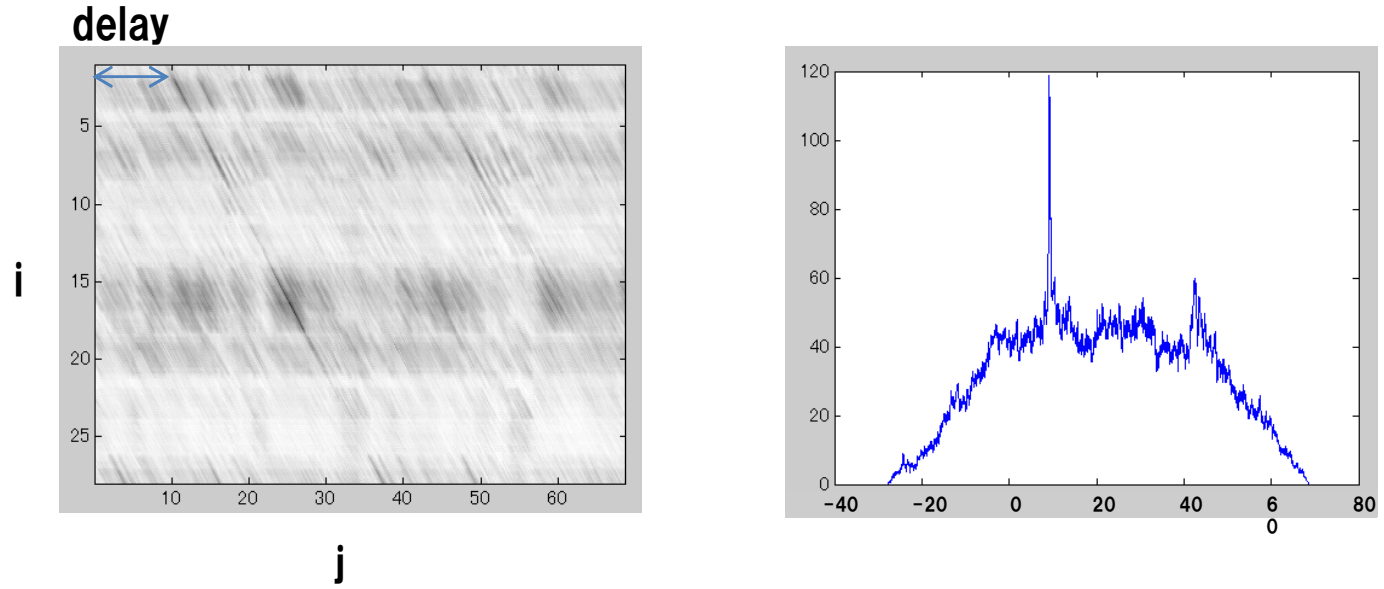
The time difference which has the largest score is regarded as the time difference between two audio feature sequences:

$$delay = \underset{-N_f{}' \le \tau < N_g{}'}{\mathrm{argmax}} \ \mathrm{Score}(\tau)$$

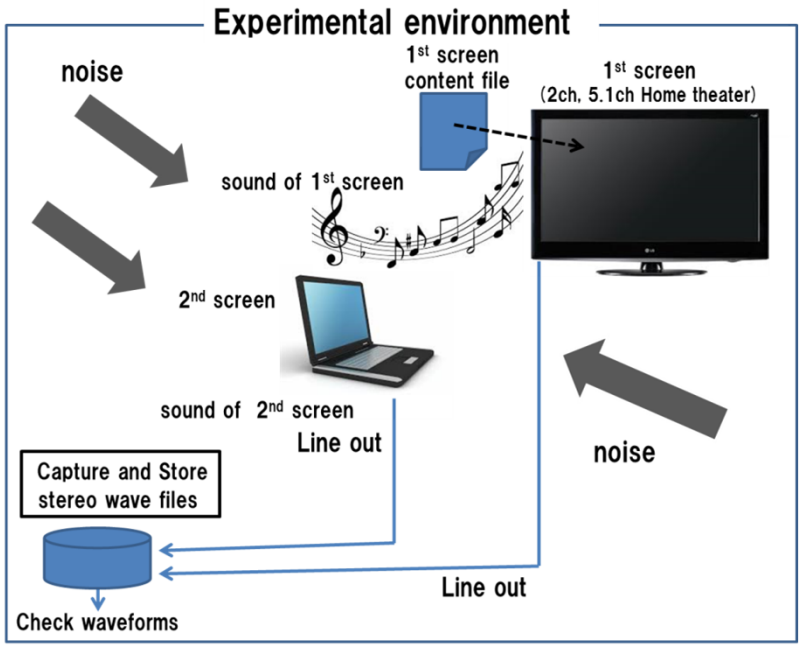


Time Difference Estimation

Example



Performance evaluation



- ✓ Capturing the 1st screen content and additive noise sound at the 2nd screen.
  → The noise contaminated 1st screen content files

- ✓ Line-out of the 1st screen and the 2nd screen are captured as a single stereo wave file.
  → Time difference between the L-ch and the R-ch in the file is measured

1st Screen content and additive noise files

The 1st Screen Content Files

	Filename of 1st screen content files	Description
1st_betty	5.1	down mix (according to ARIB STD-B32) version of CO_11_Betty3b_output
1st_speech	2	Speech (German Male, SQAM track 54)
1st_music	2	Music (Wind ensemble, SQAM track 67)

Additive Noise Sound Files

	Filename of additive noise sound files	Description
File4	noise_pinknoise	
File5	noise_speech	Speech (English Female, SQAM track 49)
File6	noise_music	Music (Eddie Rabbitt, SQAM track 70)

Result

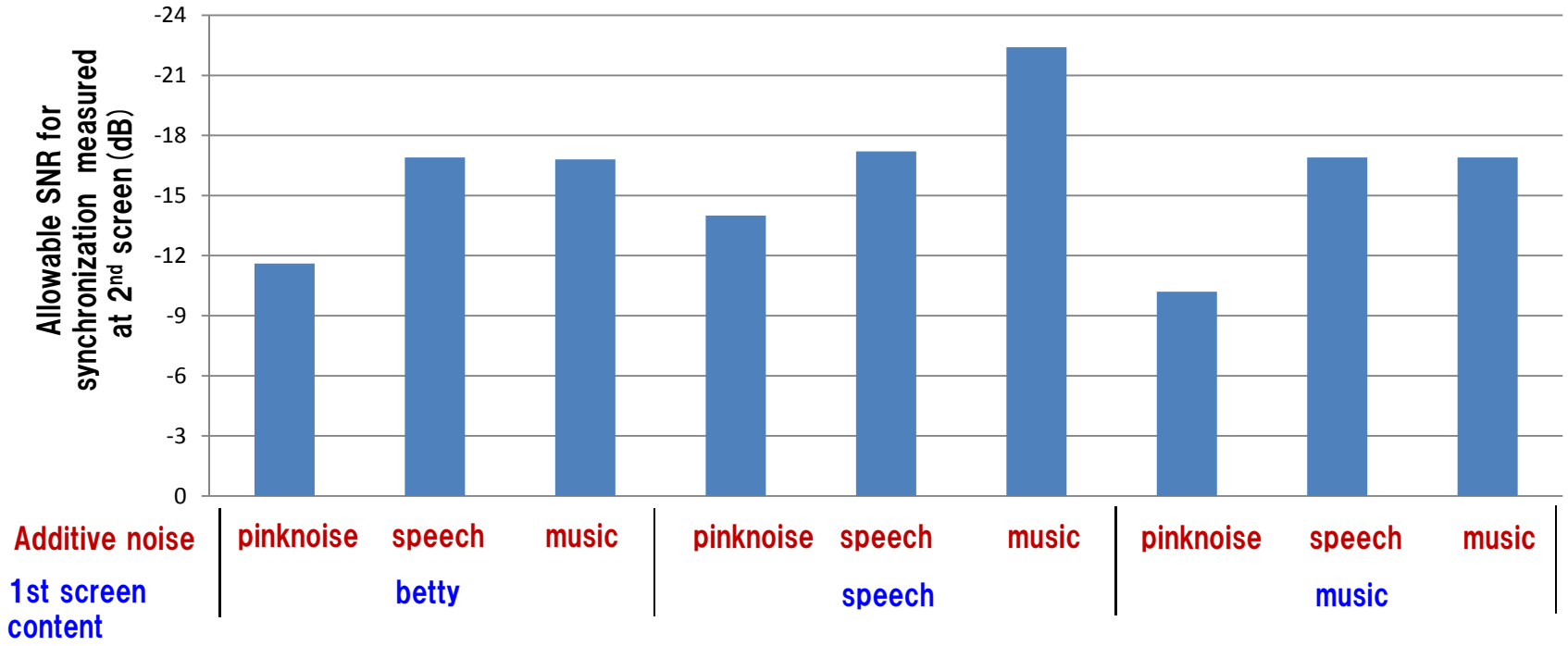
Time Difference between 1st Screen and 2nd Screen Line-Out Signals (sec)

Filename of 1st screen content files	Filename of additive noise sound files	Signal level of the 1st screen content (dB)						
		0	-6	-12	-18	-24	-30	-36
1st_betty	noise_pinknoise	0.003	0.003	0.007	N/A	N/A	N/A	N/A
	noise_speech	0.019	-0.009	-0.014	-0.014	N/A	N/A	N/A
	noise_music	-0.002	0.001	0.018	0.007	N/A	N/A	N/A
1st_speech	noise_pinknoise	-0.004	0.007	0.012	-0.001	N/A	N/A	N/A
	noise_speech	0.002	0.014	-0.013	0.003	0.008	N/A	N/A
	noise_music	-0.005	0.018	0.018	0.000	0.014	-0.011	-0.009
1st_music	noise_pinknoise	0.007	-0.007	0.015	N/A	N/A	N/A	N/A
	noise_speech	0.002	0.008	-0.004	0.007	-0.016	N/A	N/A
	noise_music	0.002	0.016	-0.007	0.015	0.018	N/A	N/A

The figures with orange background is approximately within 1frme length (32ms).
→ Synchronization is successful !

Result (cont.)

Synchronization robustness against interference noise



MPEG-4 Audio Object Type (ISO/IEC 14496-3:2009)

Object Type ID	Audio Object Type	gain control	[...]	Remark
0	Null			
[..]	[...]			
43	SAOC			
44	LD MPEG Surround			
45	SAOC-DE			
46	Audio Sync			
47 -95	(reserved)			

Demonstration

- 1st screen (blue walkman) : Instrument only

Same song

- 2nd screen (my note PC) : Vocal only

- Noise (white walkman) : Female speech

Conclusion

- MPEG-4 Audio Synchronization standard defines:

  - ✓ Audio Feature Extraction tool and syntax of the feature stream (Normative)
  - ✓ Feature Similarity Calculation Tool (Informative)
  - ✓ The Audio Object Type (AOT=46) "Audio Sync"

  to allow transmission of audio feature for synchronization as elementary stream

- The MPEG-4 synchronization mechanism works with highly noisy environment and proven that the scheme is useful under practical conditions.

End