2004 TRECVID Workshop

TRECVID Story Segmentation based on Content-Independent Audio-Video Features

Keiichiro Hoashi, Masaru Sugano, Masaki Naito, Kazunori Matsumoto, Fumiaki Sugaya, Yasuyuki Nakajima

KDDI R&D Laboratories, Inc.

Outline

- Introduction
- System description
  - Baseline story segmentation method
    - SVM-based segmentation w/ low-level features
  - System components:
    - Section-specific segmentation
    - Anchor shot segmentation
    - Post-filtering
- Experiment results
- Conclusion

Introduction

- Motivation
  - Development of a *generic* story segmentation algorithm applicable to non-news video contents
- Requirements
  - Utilize only low-level audio-video features which can be extracted from any video data
    - Restricted use of news-specific features (e.g., anchor shots)
    - Restricted use of text information (e.g., ASR results)

➡ Main focus: Story segmentation based on "Audio+Video" experiment condition

Introduction (cont'd)

- However, content-specific features *are* necessary to achieve accurate segmentation
  - ➡ Content-specific components developed to complement weak points of baseline method

- Highly accurate story segmentation achieved!

Overview: Experiment results

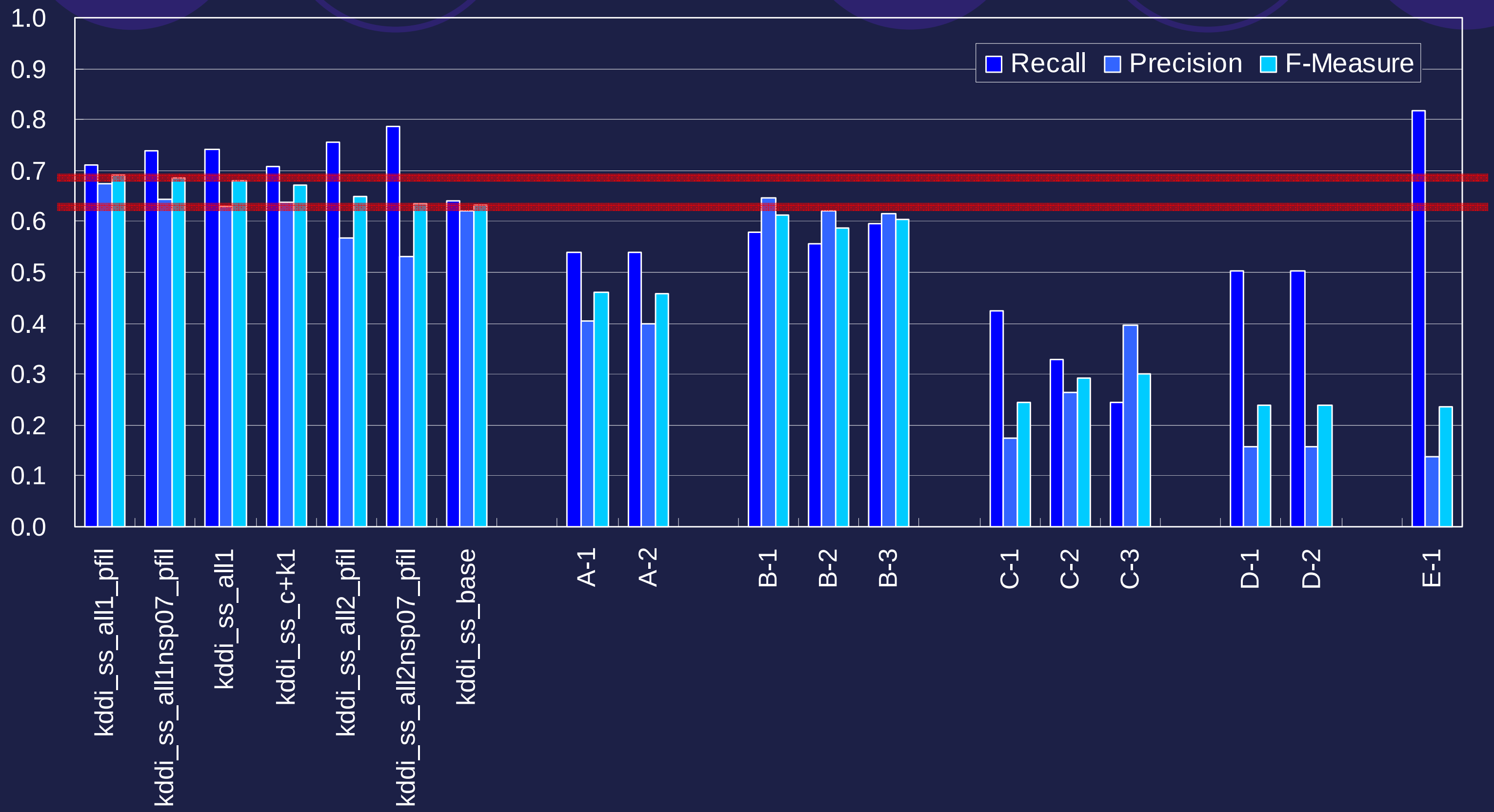
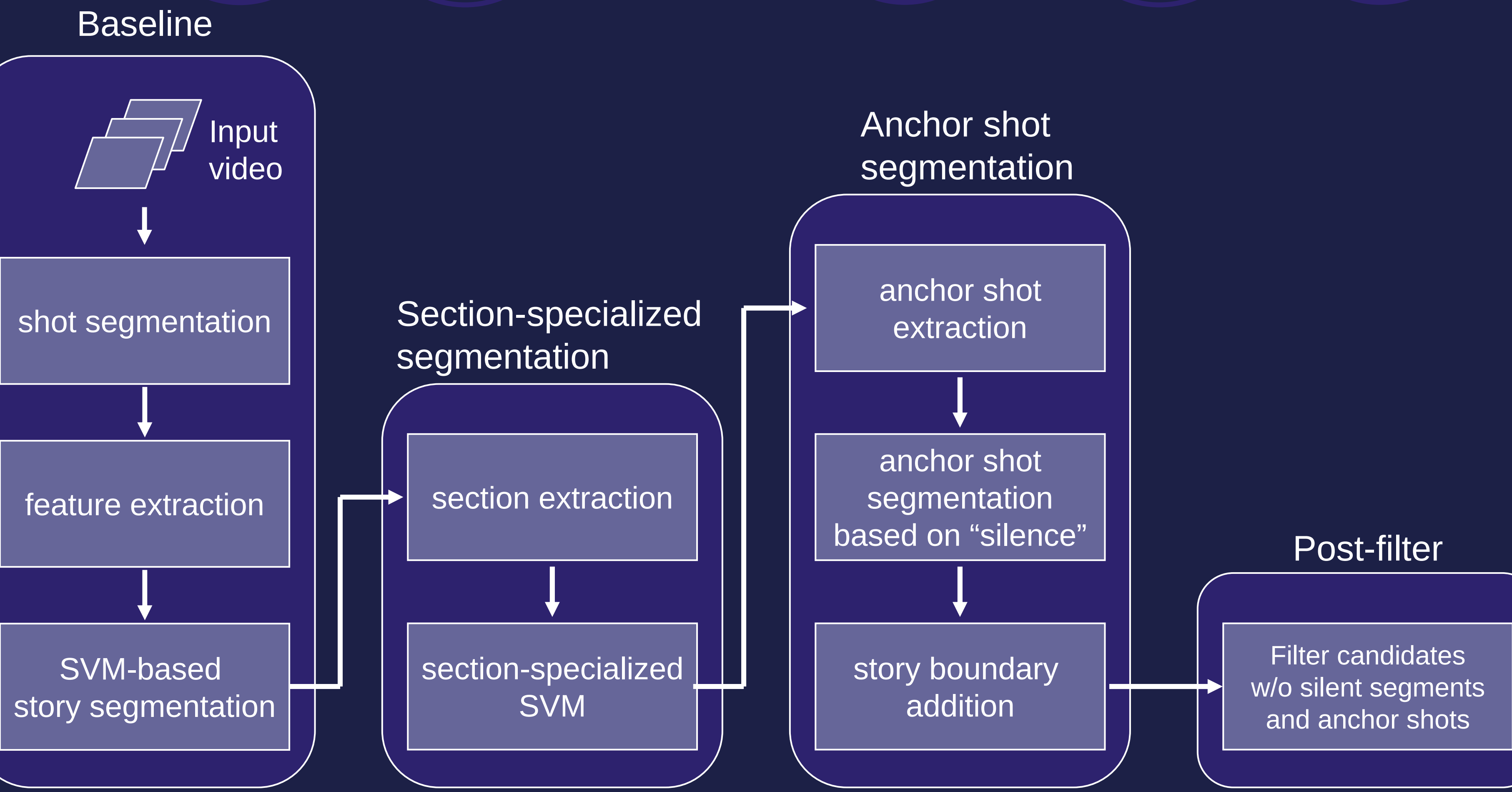


Figure 1. Recall, precision and F-measure of all “Audio+Video” TRECVID submissions

Outperformed all non-KDDI runs!

System Description

System outline



Baseline

Input video

shot segmentation

feature extraction

SVM-based story segmentation

Section-specialized segmentation

section extraction

section-specialized SVM

Anchor shot segmentation

anchor shot extraction

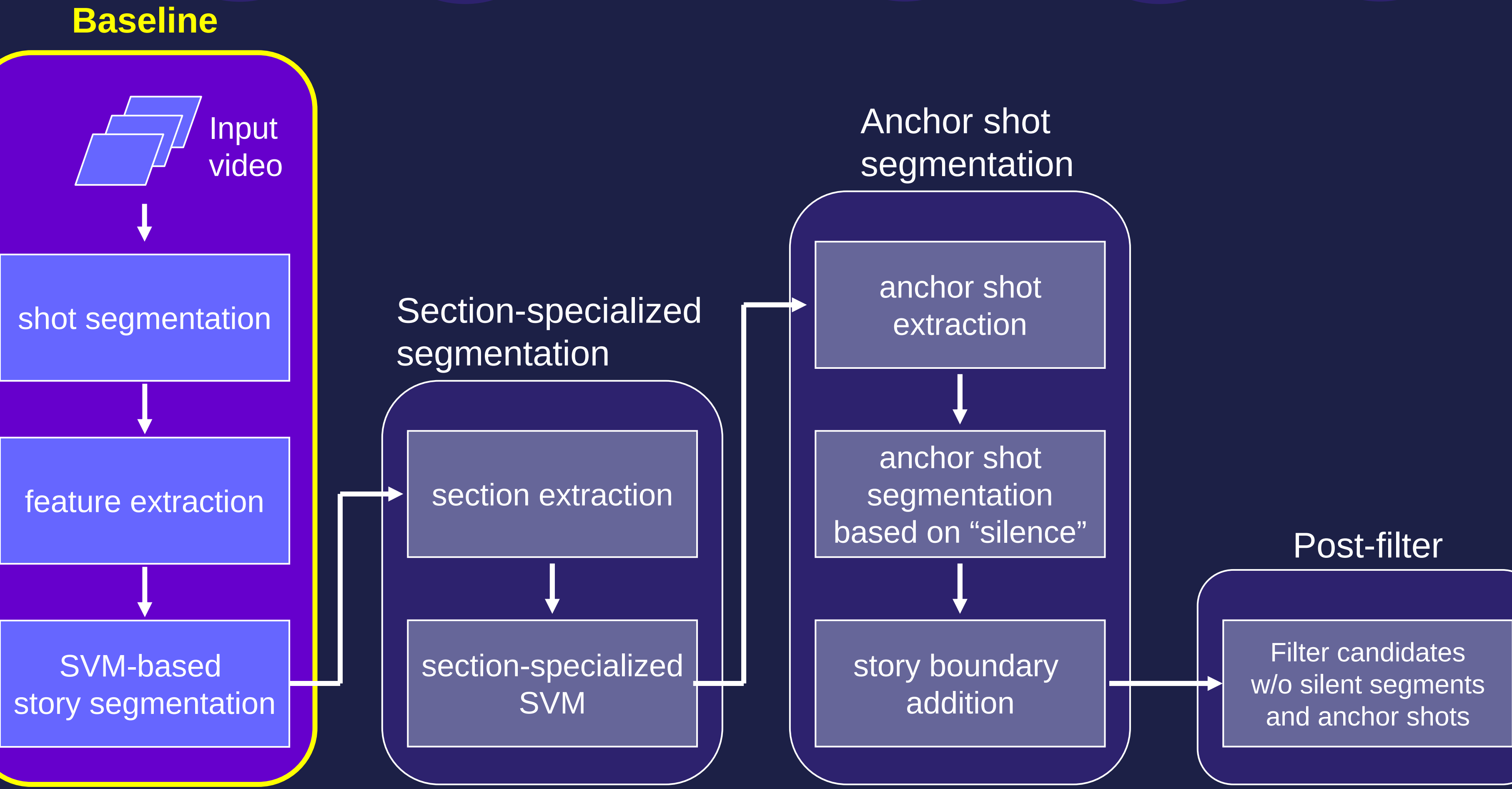
anchor shot segmentation based on “silence”

story boundary addition

Post-filter

Filter candidates w/o silent segments and anchor shots

“Baseline” component



Baseline

Input video

shot segmentation

feature extraction

SVM-based story segmentation

Section-specialized segmentation

section extraction

section-specialized SVM

Anchor shot segmentation

anchor shot extraction

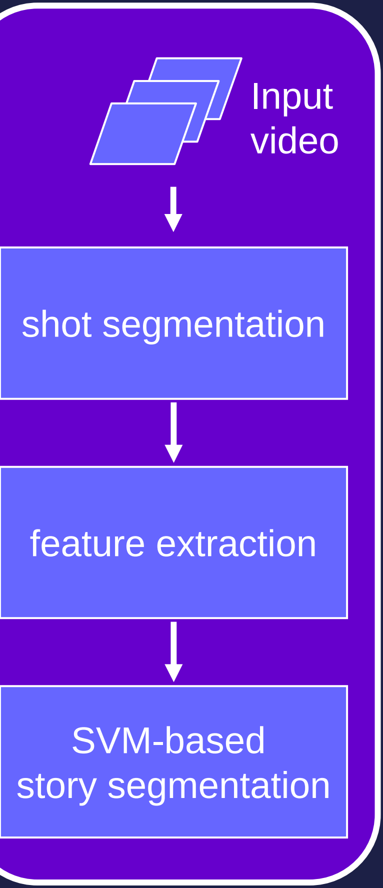
anchor shot segmentation based on “silence”

story boundary addition

Post-filter

Filter candidates w/o silent segments and anchor shots

Baseline story segmentation



- Procedures:
  - Shot segmentation
    - Merged TRECVID common shot boundaries with shot segmentation results of IBM *VideoAnnEx* tool
    - Applied “curtain-type” wipe detection method
  - Feature extraction
    - Extracts low-level audio-video features from each shot, and generates “shot vectors”
  - SVM-based story segmentation
    - Discriminates shots which contain story boundaries

Extracted audio-video features

- Audio
  - ○ Average RMS
  - ○ Avg RMS of first n frames
  - ○ Frequency of audio class (silence, speech, music, noise)
    - Details in Reference [4]
- Motion
  - ○ Horizontal motion
  - ○ Vertical motion
  - ○ Total motion
  - ○ Motion intensity
- Color
  - ○ Color layout of first, middle, and last frame (6*Y, 3*Cb, 3*Cr)
  - ○ Color layout distance between first, middle and last frames
- Temporal
  - ○ Shot duration
  - ○ Shot density

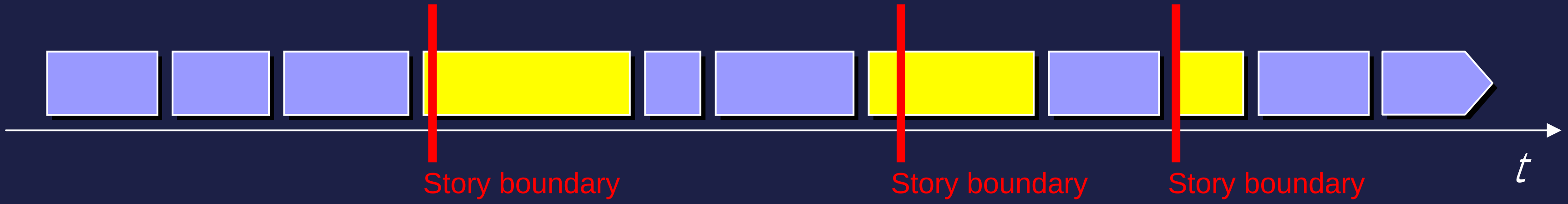
➡ Total number of elements: 51

⬇

51-dimensional "shot vector"

SVM-based story segmentation

- **Apply SVM to discriminate shots w/ story boundary**
- Training phase
  - Shots which contain story boundary ⇒ Positive
  - All other shots ⇒ Negative



- Evaluation phase
  - Extract *N* shots based on distance from SVM hyperplane
    - *N* = Average number of stories in ABC, CNN (Baseline)
    - *N* = Average number of stories x 1.5 (Extended baseline)
  - Set story boundary at beginning of each extracted shot

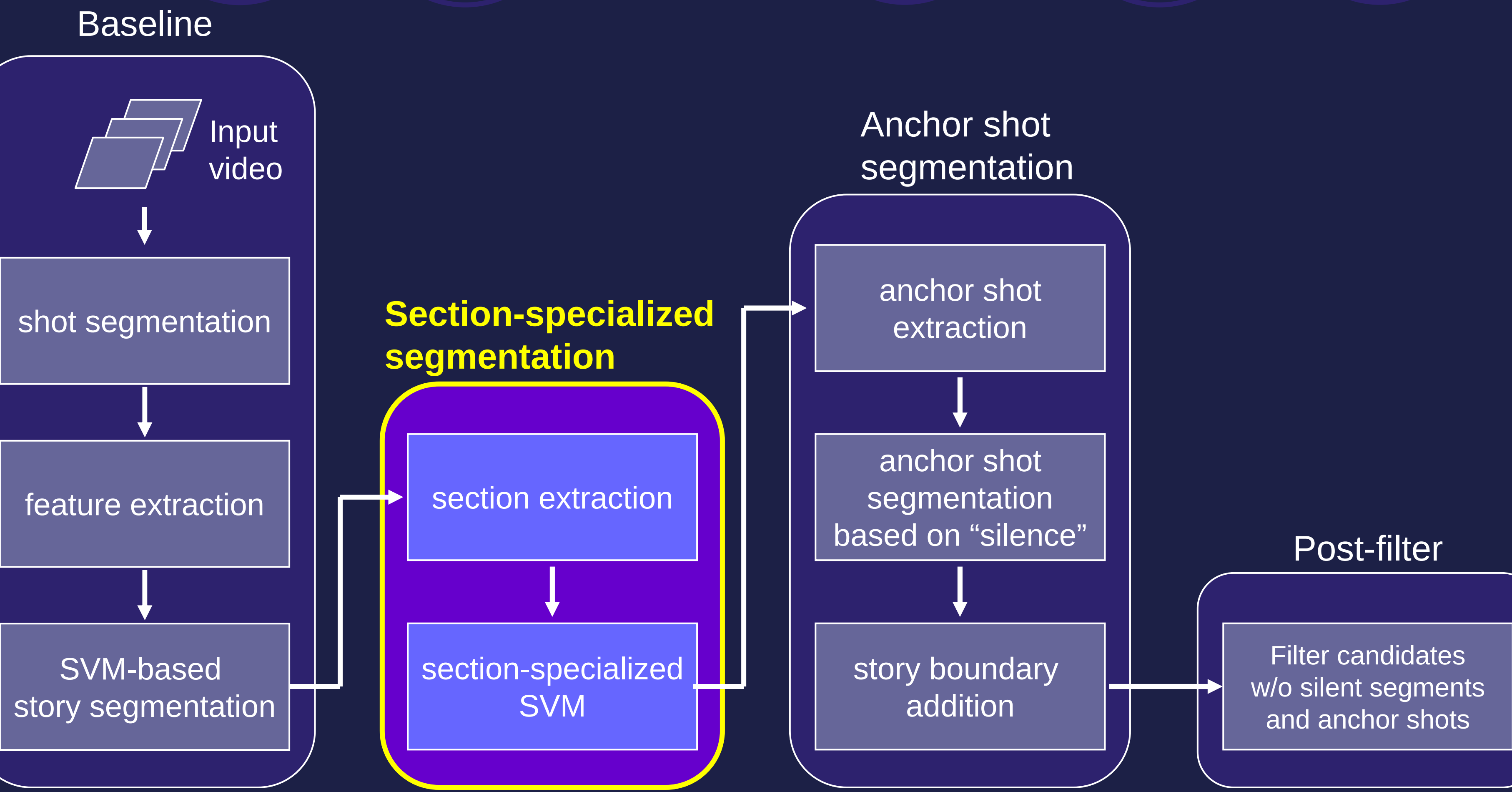
Problems of baseline method

- Although baseline results were satisfactory, several weak points were observed…
- Poor recall in various “sections”
  - e.g., *Top Stories*, *Headline Sports* of CNN
  - Cause: Different characteristics compared to general content
    - No anchor shots, background music, etc.
  - SVM unable to adapt to various features
- Impossible to detect multiple story boundaries that occur within a single shot
  - Baseline can only set one story boundary per shot

Additional system components

- Section-specialized segmentation
  - Objective:
    - Improvement of recall in specific sections which have different characteristics
- Anchor shot segmentation
  - Objective:
    - Detection of multiple story boundaries which occur within a single shot
- Post-filter
  - Objective:
    - Improvement of precision

Component 1: Section-specialized segmentation



Baseline

Input video

shot segmentation

feature extraction

SVM-based story segmentation

Section-specialized segmentation

section extraction

section-specialized SVM

Anchor shot segmentation

anchor shot extraction

anchor shot segmentation based on “silence”

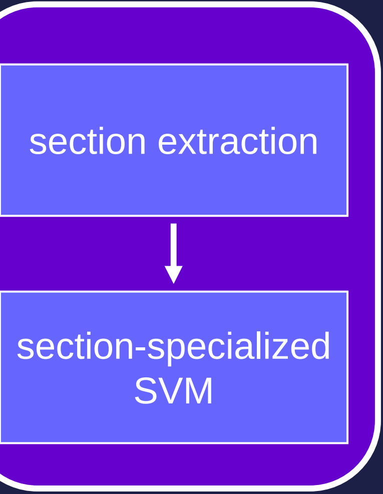
story boundary addition

Post-filter

Filter candidates w/o silent segments and anchor shots

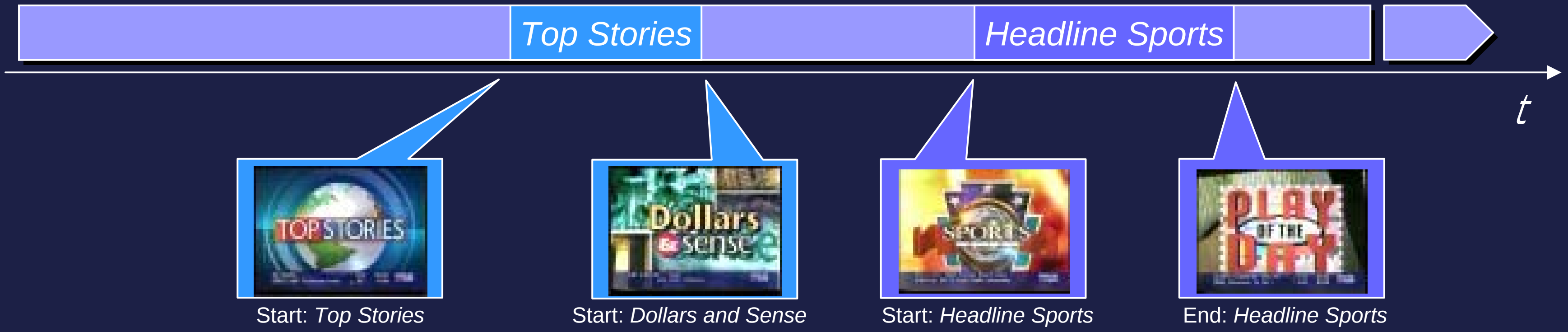
Section-specialized segmentation

- General approach:
  - Construct SVM specialized for story segmentation within specified sections
- Procedures:
  - Section extraction
    - Extraction based on “jingles”, i.e., audio-video sequences which initiate sections
  - Section-specialized SVM
    - Construct SVM specialized to conduct story segmentation on extracted sections



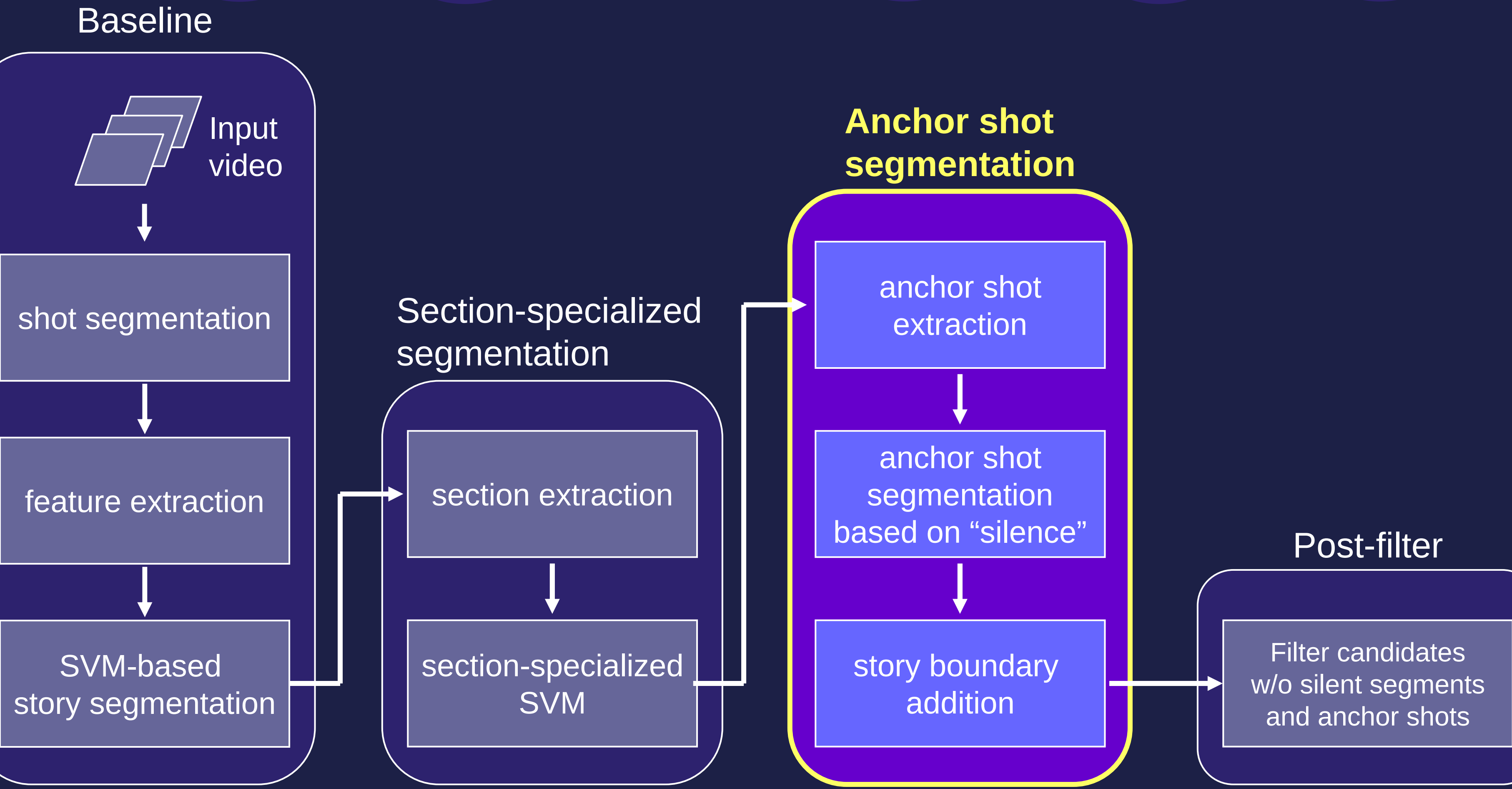
Section extraction

- Automatic detection of “jingles” based on reference audio signals
  - Based on “Time-series active search” algorithm [Kashino]
- Extract sections based on position of extracted jingles

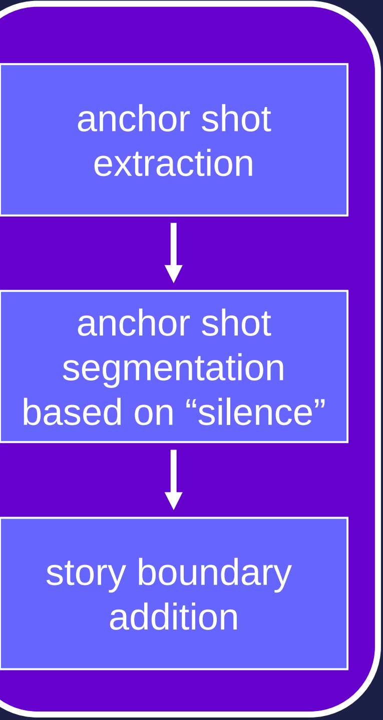


- Apply section-specialized SVM to set story boundaries within each extracted section

Component 2: Anchor shot segmentation

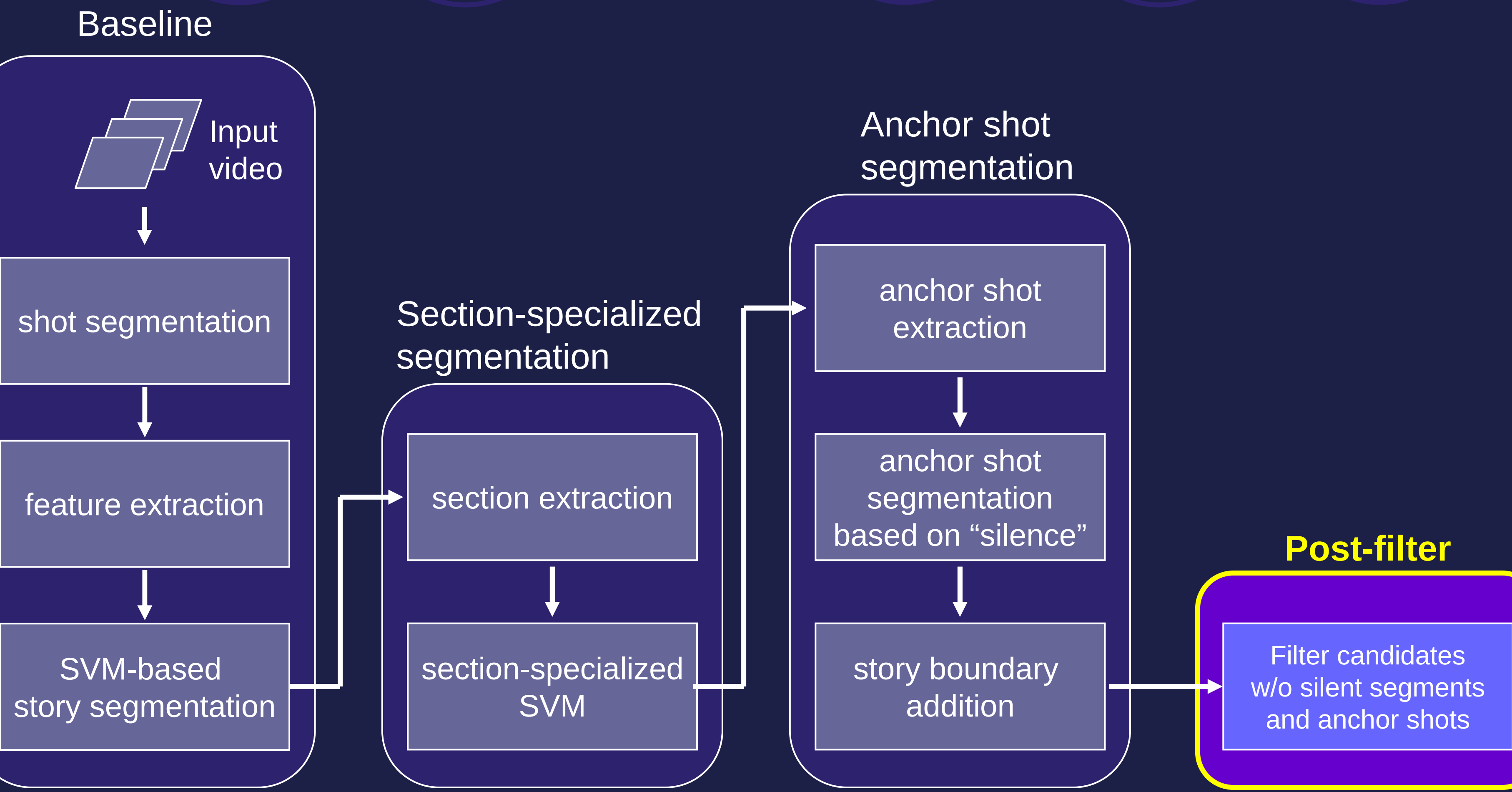


Anchor shot segmentation



- General approach:
  - Extract shots which are expected to contain multiple stories (anchor shots), and insert additional boundaries
- Procedures:
  - Anchor shot extraction
    - Construct SVM to discriminate anchor shots based on audio-video features
  - Extraction of “silent sections”
    - Two methods:
      - Audio classification results
      - HMM-based non-speech detector
  - Story boundary addition
    - Insert story boundaries at detected silence sections

Component 3: Post-filter



Post-filter

- Objective:
  - Improvement of story segmentation precision
    - *Objective of previous components is improvement of recall*
- Procedure:
  - Omission of questionable story boundary candidates based on:
    - Silence section extraction
      - Hypothesis: Story transitions are accompanied with significant pause = silence
    - Anchor shot detection
      - Hypothesis: Story boundaries accompanied with *non*-anchor shots are probably mistaken
  - Utilizes features used in in previous components

Filter candidates w/o silent segments and anchor shots

Experiment Results

Description of KDDI Audio+Video runs

Table 1. Summary of KDDI “Audio+Video” story segmentation runs

Run ID	Baseline	SS-S	Anchor SS	Post-filter
kddi_ss_base1	Base			
kddi_ss_c+k1	Base	✓		
kddi_ss_all1	Base	✓	Audio Class	
kddi_ss_all1_pfil	Base	✓	Audio Class	Audio Class
kddi_ss_all2_pfil	Ext	✓	Audio Class	Audio Class
kddi_ss_all1nsp07_pfil	Base	✓	HMM	HMM
kddi_ss_all2nsp07_pfil	Ext	✓	HMM	HMM

Evaluation results

Table 2. Results of KDDI “Audio+Video” story segmentation runs

Run ID	Recall	Precision	F-measure
kddi_ss_base1	0.640	0.622	0.631
kddi_ss_c+k1	0.707	0.637	0.670
kddi_ss_all1	0.741	0.630	0.681
kddi_ss_all1_pfil	0.710	0.675	0.692
kddi_ss_all2_pfil	0.756	0.567	0.648
kddi_ss_all1nsp07_pfil	0.738	0.642	0.687
kddi_ss_all2nsp07_pfil	0.786	0.531	0.634

Contribution of each system component

- Section-specialized segmentation (SS-S)
  - Baseline → Baseline + SS-S
    - Recall: +0.123 (0.605 → 0.728)
    - Precision: +0.026 (0.596 → 0.625)
  - *Comparison based only on CNN data*
    - Specific sections could not be defined for ABC…
- Anchor shot segmentation (ASS)
  - Baseline + SS-S → Baseline + SS-S + ASS:
    - Recall: +0.034 (0.707 → 0.741)
    - Precision: -0.007 (0.637 → 0.630)
- Post-filter (PF)
  - Baseline + SS-S + ASS → Base + SS-S + ASS +PF
    - Recall: -0.031 (0.741 → 0.710)
    - Precision: +0.045 (0.630 → 0.675)

Summary of system component contributions

- Section-specialized segmentation
  - Highly effective *(if sections are definable and extractable)*
- Anchor shot segmentation
  - Effective for recall improvement
  - Decrease of precision was not as significant as predicted
- Post-filter
  - Precision improved, recall decreased
  - Overall improvement (F-measure) was minimal

Conclusion

- Proposed SVM-based story segmentation method based on low-level audio-video features
  - ○ Applicable to video of any domain
  - ○ Significantly efficient compared to conventional methods which utilize sophisticated feature extraction
  - ○ Achieves highly accurate story segmentation!
- Various content-specific components also effective
  - ○ Generality of audio-video features enabled easy implementation of system components

Future work

- Segmentation on video w/ insufficient training
  - Recall was poor on video files recorded in environment that did not appear in development data

Normal studio setting
(Recall: approx. 80%)

19981216~18_ABCa.mpg
(Recall: 13~36%)

- Automatic extraction of reference signals for jingle detection
  - Enables application of section-specialized segmentation for various news programs

Thanks ☺