Speech-Based Content Indexing System for Broadcast News

Yoshihiko Hayashi

NTT Cyberspace Laboratories

mailto:hayashi.yoshihiko@lab.ntt.co.jp

Thanks to: Katsutoshi Ohtsuki, Katsuji Bessho, Osamu Mizuno, Yoshihiro Matsuo, Shoichi Matsunaga, Takaaki Hasegawa, Naruhiro Ikeda, and Minoru Hayashi

Content

- Motivation
  - Metadata Bottleneck
  - Why News Programs?
- System Overview
  - Demonstration
- Techniques for News Story Segmentation
  - Audio-based Segmentation
  - Automatic Speech Recognition
  - **Topic Segmentation from ASR transcription**
  - **Integrated Segmentation: *Media Fusion***
- Evaluation
- Conclusions and Future Works

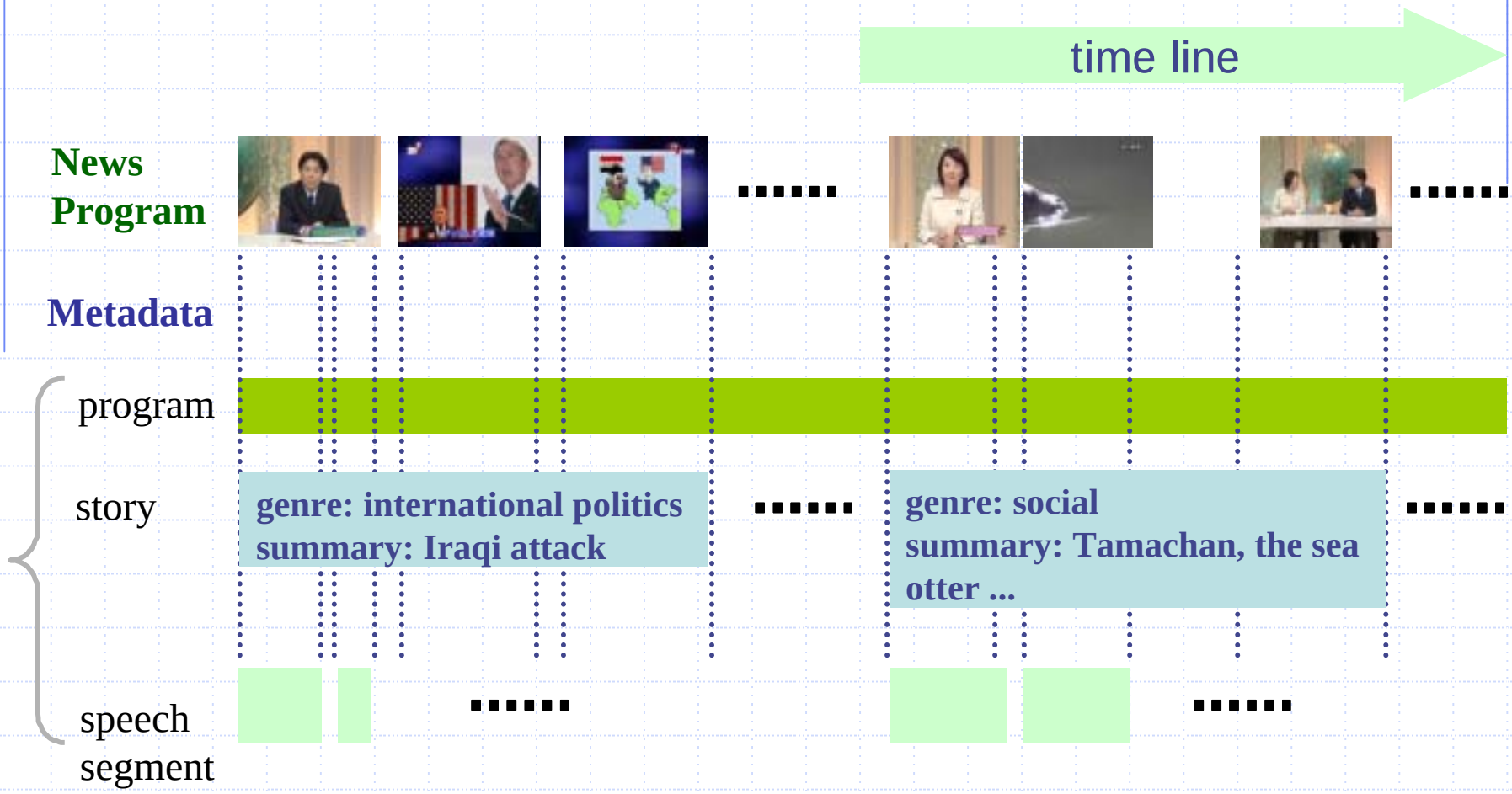
Motivation *- Metadata Bottleneck -*

- Metadata is
  - **useful** in discovering and utilizing existing information resources (contents)
    - **essential** for non-textual content
  - however, **costly** if we are to create it
    - **inevitably so** for non-textual content
- Metadata Bottleneck
  - there **is a bottleneck in metadata creation**
  - media processing (speech recognition, character recognition, image recognition, etc.) and natural language processing are key to break the bottleneck
- to begin with the project
  - primal target: broadcast news program

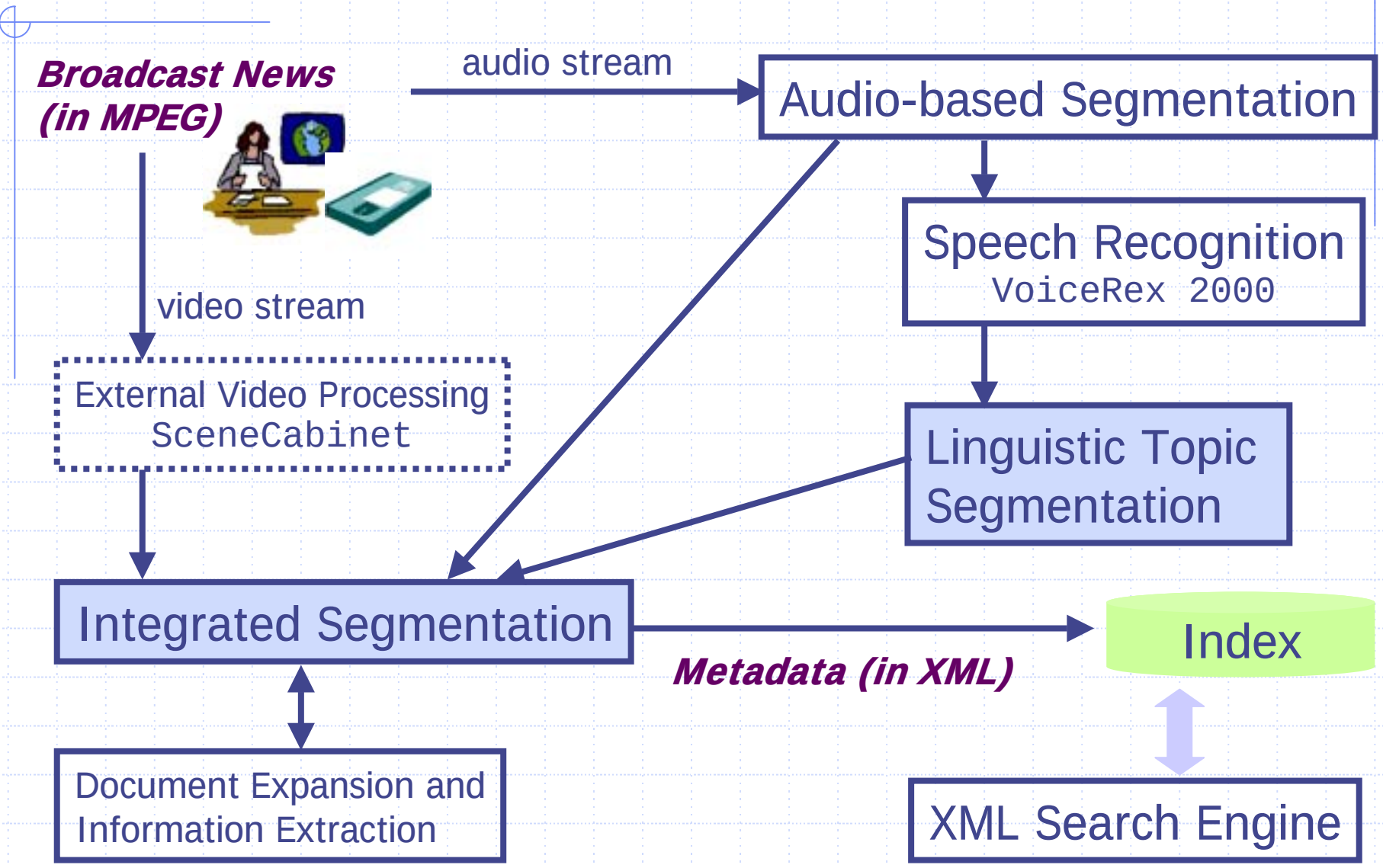
Why Broadcast News Program?

- The semantic intent is mainly conveyed by the speech uttered by the anchors
- The speech is usually fluent and clear (acoustic/linguistic)
  - adequate for applying current-level automatic speech recognition (ASR)
  - NLP can be applicable after the speech-to-text process
- News program has relatively clear structure
  - Example: Opening, Leading index, Story-1, Story-2, ..., Story-n, Weather forecast, Market information, Closing
  - Chance to apply NLP for structuring the entire program (automatic story segmentation)

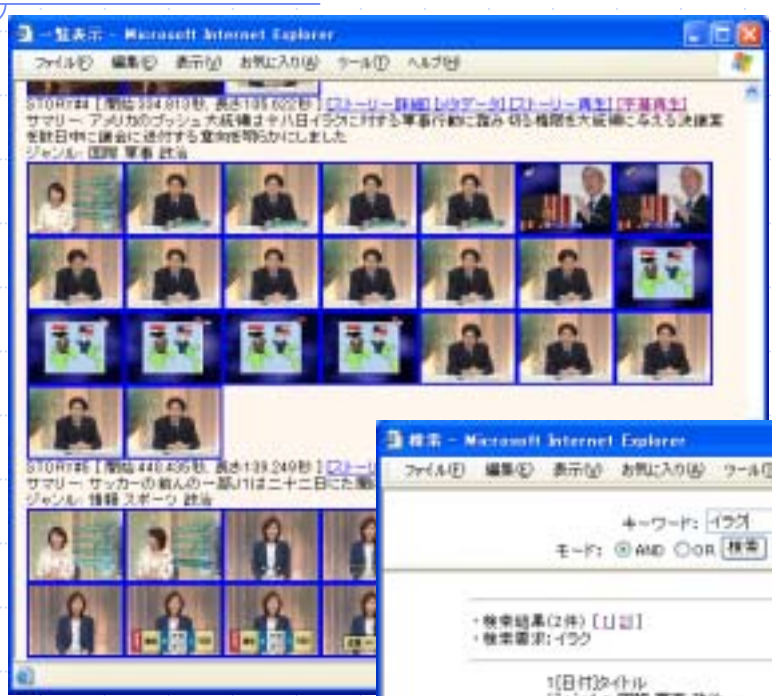
Structure of the News Program Metadata



Prototype System Overview



Demonstration



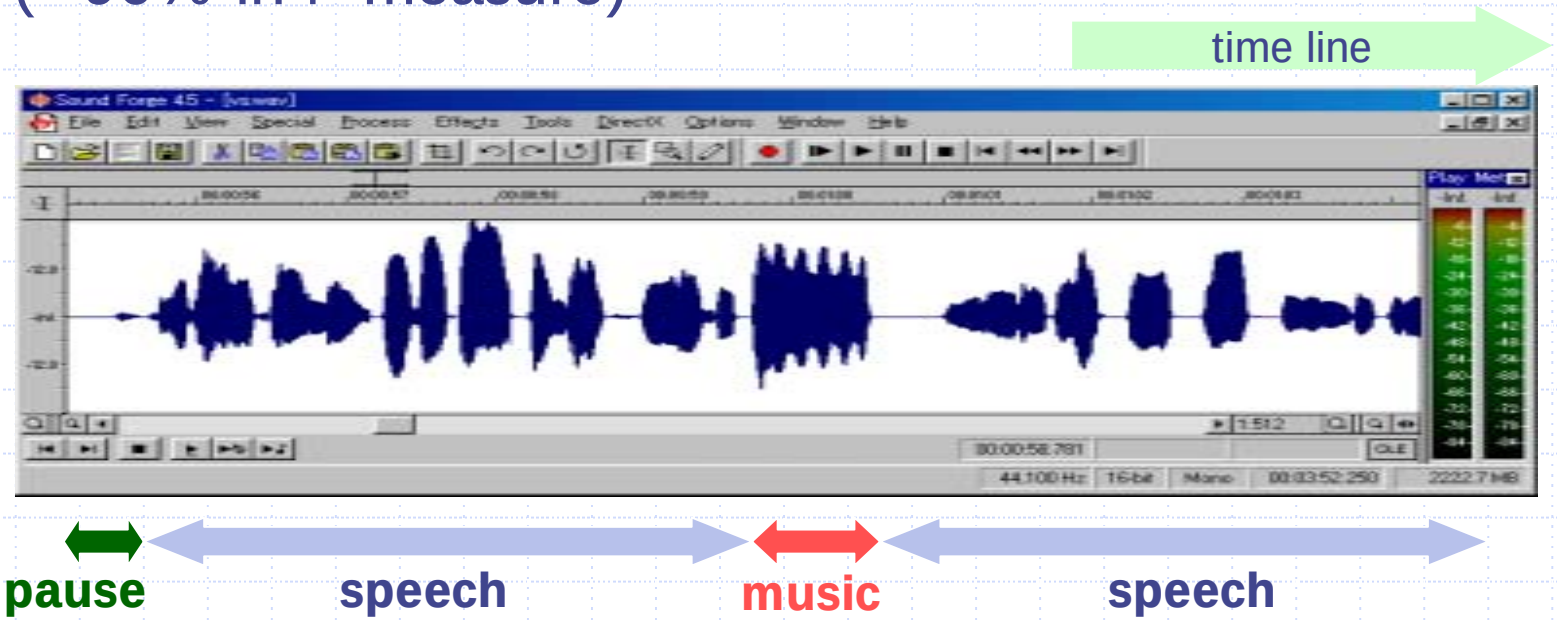
Segmentation View

KW Search

Playback a news story with open-caption (speech-recognized)

Audio-based Segmentation

- segments audio stream into a set of intervals, each of them is assigned one of the following class labels: speech, music, noise, silence
- based on supervised learning method
- using some distinctive acoustic features
- highly accurate, particularly in extracting speech intervals (~96% in F-measure)



Automatic Speech Recognition - **VoiceRex** -

- Input: a speech segment
- Output: time-stamped morphological information
  - pronunciation, part-of-speech, information from the dictionary
  - confidence measure (acoustic/linguistic)
- Acoustic model
  - can utilize multiple models (male/female/gender-free)
  - chooses the best one by GMM (Real Time Factor: 1.4)
- Language model
  - tri-gram learned from 600k sentences
  - vocabulary: 30k words
- Recognition accuracy (in WER)
  - speaker: anchor (15.7%) ~ reporter (28.4%)
  - noise: clean (19.2%) ~ very-noisy (35.4%)
  - style: read (18.1%) ~ spontaneous (30.4%) ~ free (59.7%)

Topic Segmentation - Algorithm Overview -

- Input: a sequence of ASR transcriptions, each of them comes from a speech segment
- Output: topic boundaries
- Procedure
  - preparation: constructs "concept-base"
  - pre-processing:
    - remove "stop-words" from the input
    - assign "concept-vector" to each remaining content word
  - for each sliding windows
    - compute centroid concept vector for the window
    - compute similarity values with the adjacent window, via cosine measure
  - smooth the similarity values
  - choose the topic boundaries by looking at the "depth scores"

Topic Segmentation - An Example -

depth score

$$d_i = (S_l - S_i) + (S_r - S_i)$$

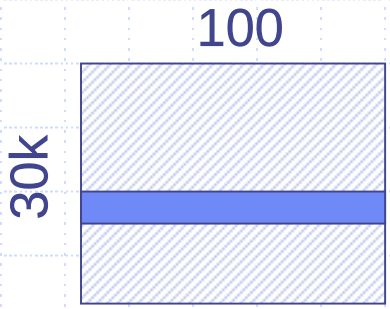
Topic Segmentation - Concept-Base -

content bearing words (1.5k)

high-frequency words (30k)

	...	computer	...	disease	...
...	...	...	...	...	...
Internet		403	...	3	...
...	...	...	...	...	...
virus	...	61	...	312	...
...	...	...	...	...	...

Singular Value Decomposition



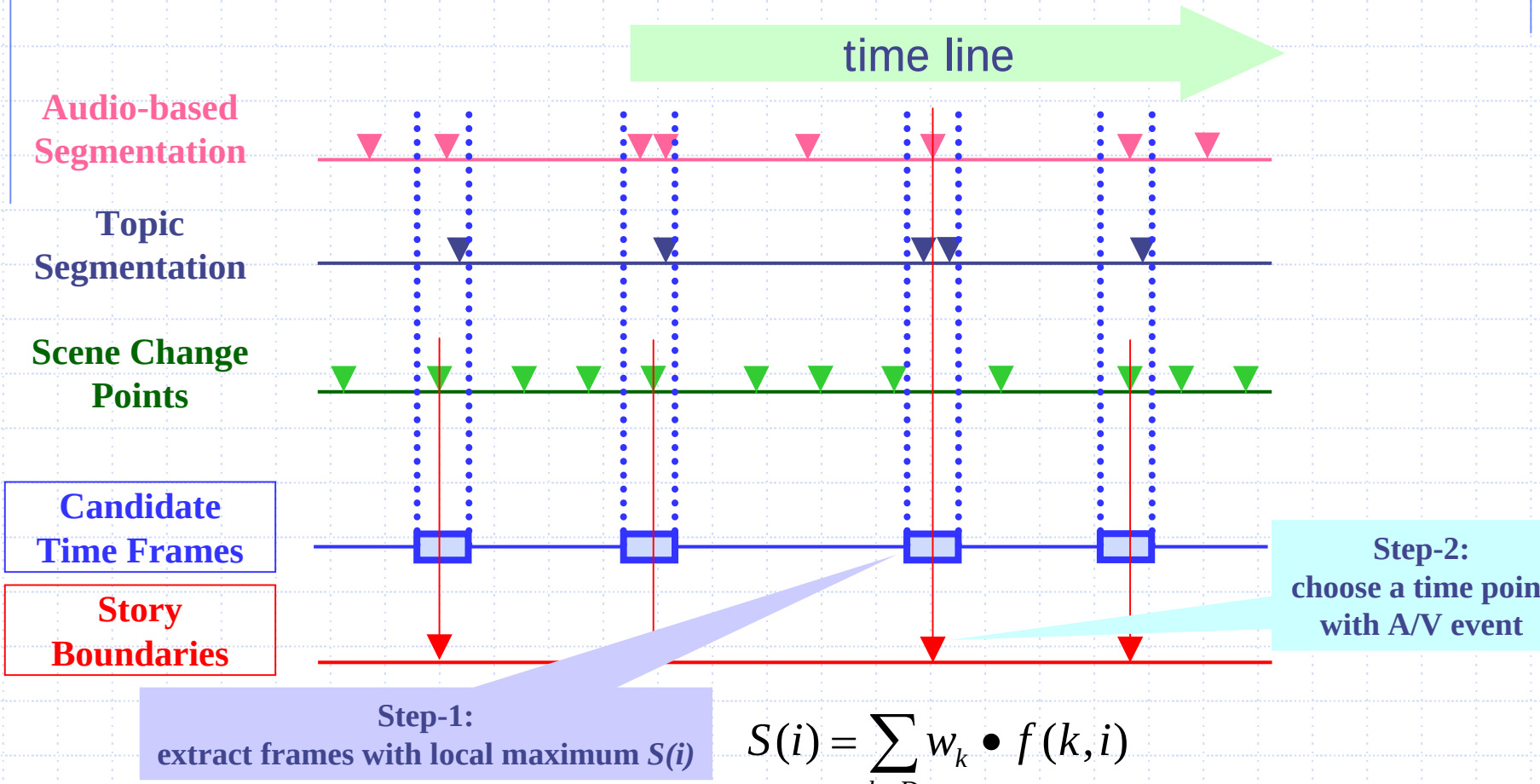
Matrix: concept-case
Row Vector: concept vecto

reduces a high-dimensional sparse matrix to lower-dimensional dense matrix

Integrated Segmentation - Algorithm Overview -

Two Step Algorithm:

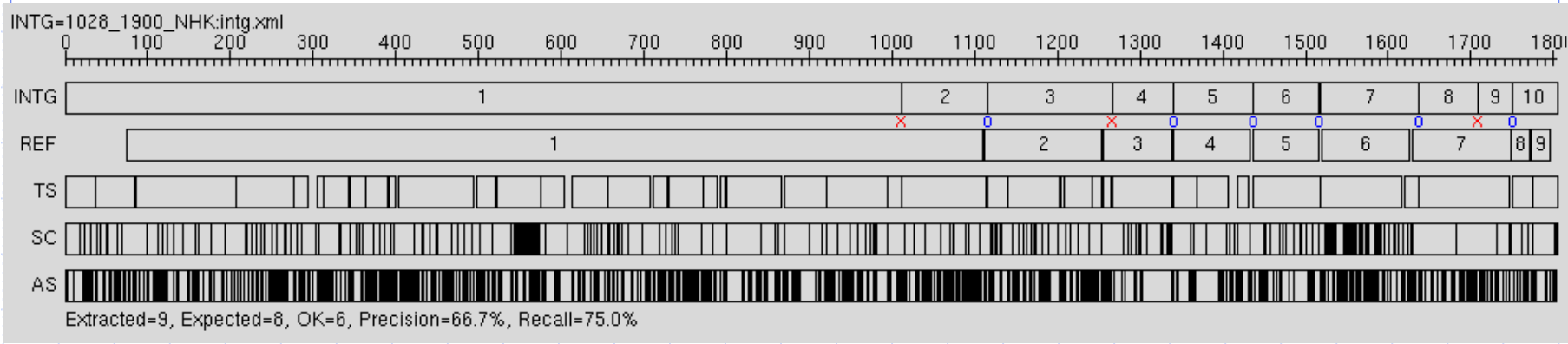
- 1-st Step: extract candidate time frames by calculating score values
- 2-nd Step: choose a boundary time point from each candidate



$$S(i) = \sum_{k \in D} w_k \bullet f(k,i)$$

Integrated Segmentation - An Example -

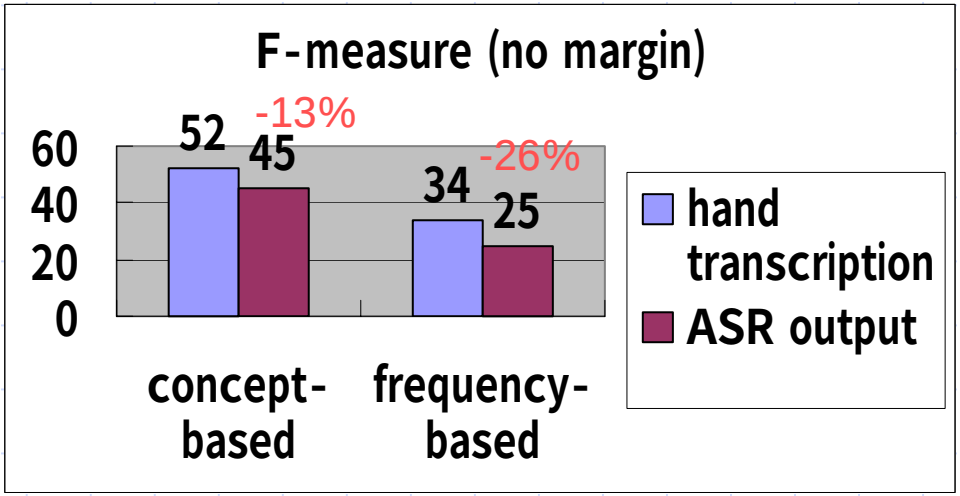
time line

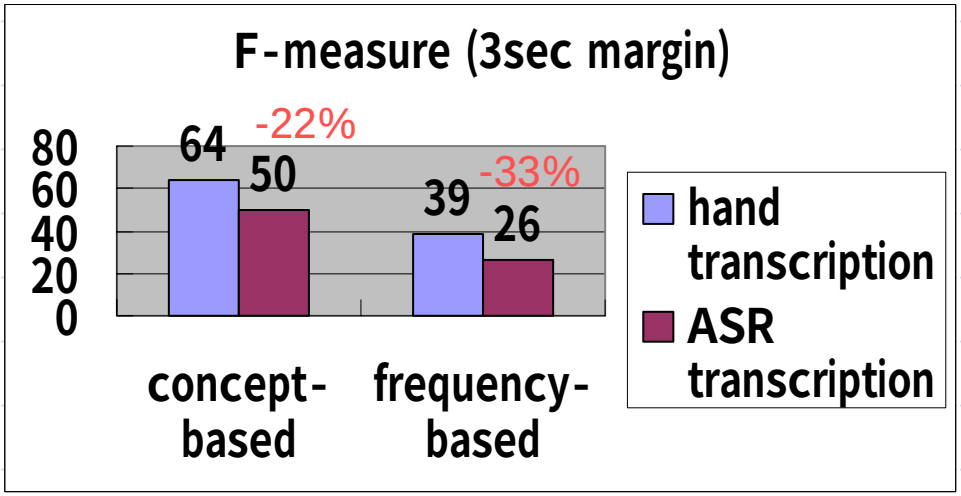


A 30-min. TV News without CM, 2002 10/28

INTG: result from the integrated segmentation
REF: human-annotated reference (correct)
TS: Topic Segmentation result
SC: Scene Change detection result
AS: Audio-based Segmentation results

Evaluation - Topic Segmentation -

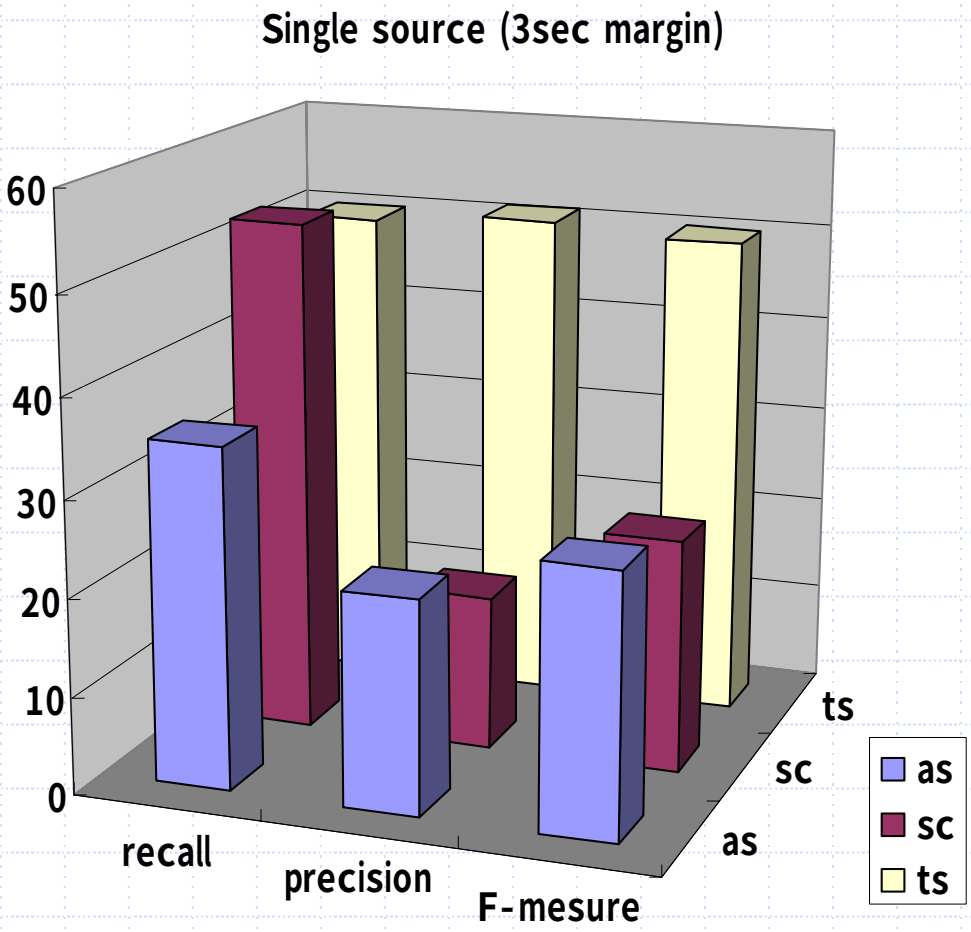




- Data: twelve TV news programs (total: 195 minutes, almost 1,000 sentences, 26k tokens)

- not accurate as artificial data (cf. ~80% in [Bessho 2003])
- degradation in ASR output cannot be ignored (13%~33%)
- concept-based method is more robust to the ASR errors than the frequency-based (Hearst) method (13%-26%, 22%-33%)

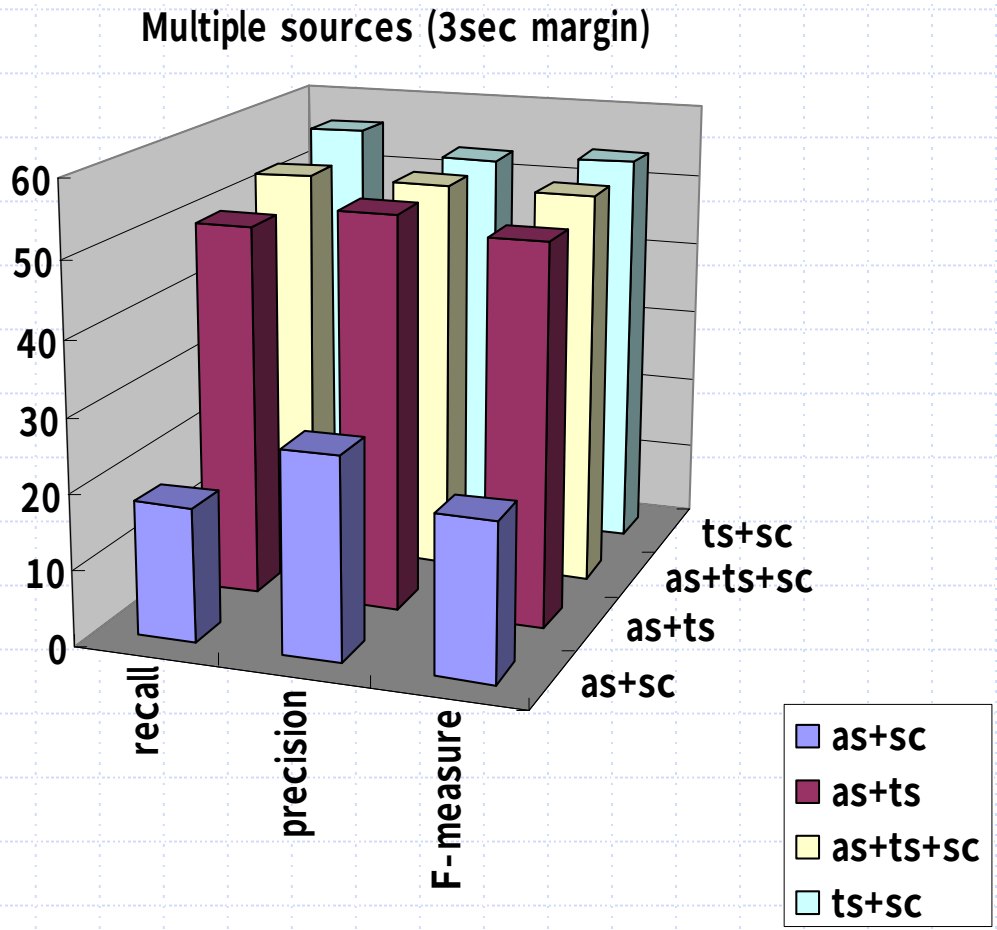
Evaluation - Integrated Segmentation (1/2) -



- non-linguistic processing, such as (audio-based segmentation) and sc (scene-change based segmentation) are not powerful as the linguistic processing ts (topic segmentation)
- non-linguistic processing (as and ts) are recall-oriented methods

- These are reasonable, because news programs are speech/language-centered contents

Evaluation - Integrated Segmentation (2/2) -



- non-linguistic processing (as+sc) is clearly insufficient
- speech-based and audio/visual-supported processing (as+ts+sc) (as+ts), (sc+ts) are slightly better than linguistic processing (ts) alone! (but not very significant...)

- probably, there are plenty of rooms to improve the accuracy by integrating the audio/visual results in better ways

Conclusions

- A prototype indexing system with search/access interface is developed and demonstrated
  - ASR and succeeding NLP play an essential role for speech-centered news programs
- Evaluation results from the small-sized preliminary experiments are shown
  - not perfect yet, but promising
  - audio/visual information may further improve the story segmentation accuracy
- More efforts are necessary for realizing automatic content metadata creation and the associated advanced search/access functions
  - *even if, "ad hoc" SDR (Spoken Document Retrieval) in news domain is a solved problem, as claimed by TREC people*

Future Works

- More evaluations...
- Improve segmentation accuracy
  - improve topic segmentation accuracy by using reliably recognized tokens (use of ASR confidence measure)
  - seek better "blend" in the integrated segmentation
  - last resort?: use of pre-knowledge about structure of the target news programs
- More NLP
  - Event extraction and tracking (TDT)
  - Summary generation (from collapsed ASR transcriptions)
- Search/Access system design
- Other types of content: documentaries, lectures, meeting, etc.
  - improve the robustness of the ASR for spontaneous/free speech
  - beyond the topic segmentation...

References

- in English
  - [Hayashi 2003a] Y.Hayashi, et.al: Speech-based and Video-supported Indexing of Multimedia Broadcast News, SIGIR 2003 poster, to appear, 2003
  - [Hayashi 2003b] Y.Hayashi, et.al: Speech and Language Processing for Content Description Metadata for Broadcast News, NTT Technical Review, Vol.1, No.3, pp.62-65, 2003
- in Japanese
  - [大附 2003] 大附, 別所, 水野, 松尾, 林: 音声認識を用いたマルチメディアコンテンツのインデクシング, 情報処理学会 第47回音声言語情報処理研究会, to appear, 2003
  - [水野 2003] 水野, 大附, 松永, 林: ニュースコンテンツにおける音響信号自動判別の検討, 電子情報通信学会 総合大会, D-14-19, 2003
  - [別所 2003] 別所, 大附, 松永, 林: 概念ベクトルの結束性によるトピックセグメンテーション精度の評価, 言語処理学会 第9回年次大会, B7-7, 2003