July 16th - August 10th, 2007 - Istanbul, Turkey

eNTERFACE '07

The SIMILAR NoE Summer Workshop on Multimodal Interfaces

Project 3
A Multimodal Framework for the Communication of Disabled

Part 2
Retrieval from Hearing Impaired News Videos

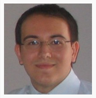
Lale Akarun, Murat Saraçlar

Oya Aran, Ismail Ari, Pavel Campr, Erinç Dikici,
Marek Hruz, Deniz Kahramaner, Siddika Parlak

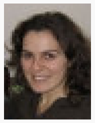
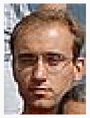
hosted by Boğaziçi University, Multimedia Group

Retrieval from Hearing Impaired News Videos

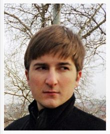
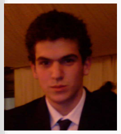
- Modalities



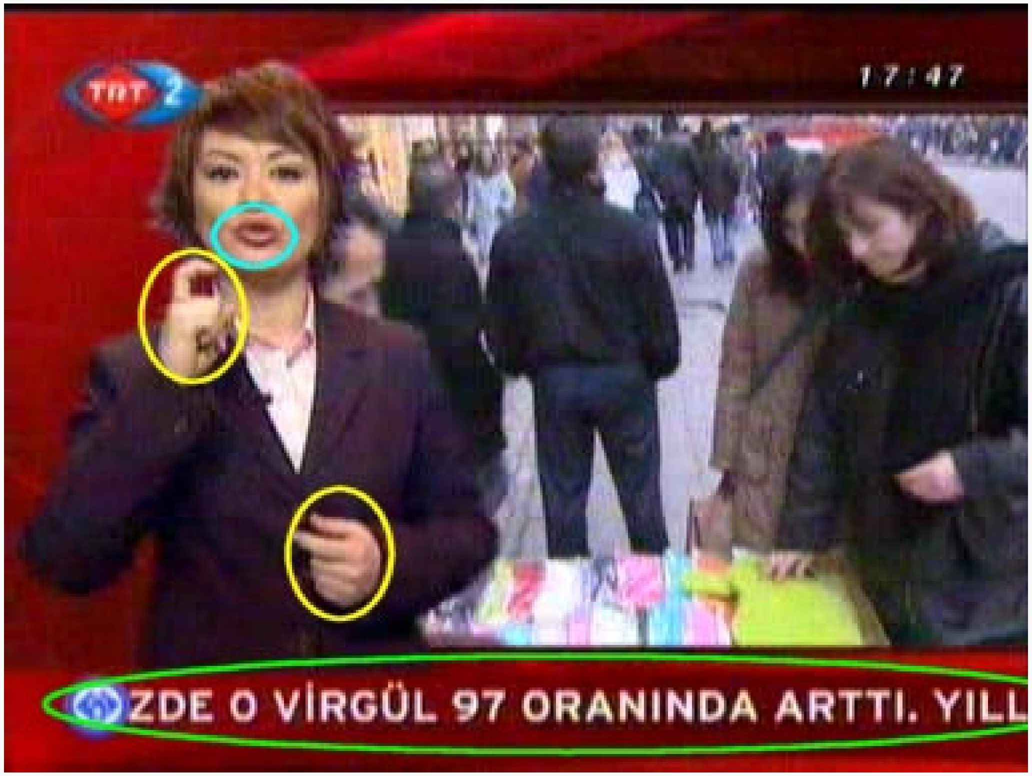
  - Speech
  - Lips
  - Text
  - Sign

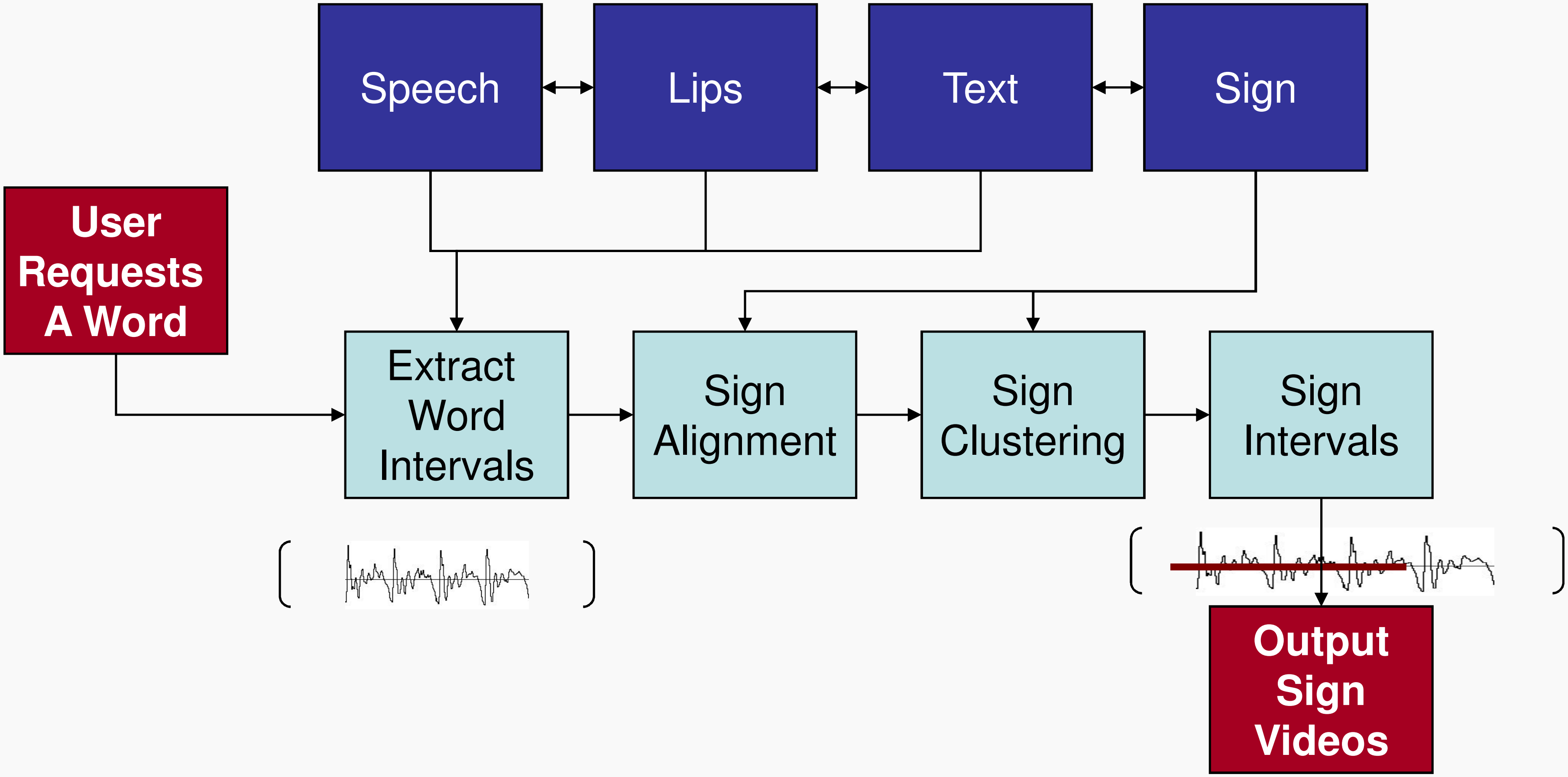
- GUI

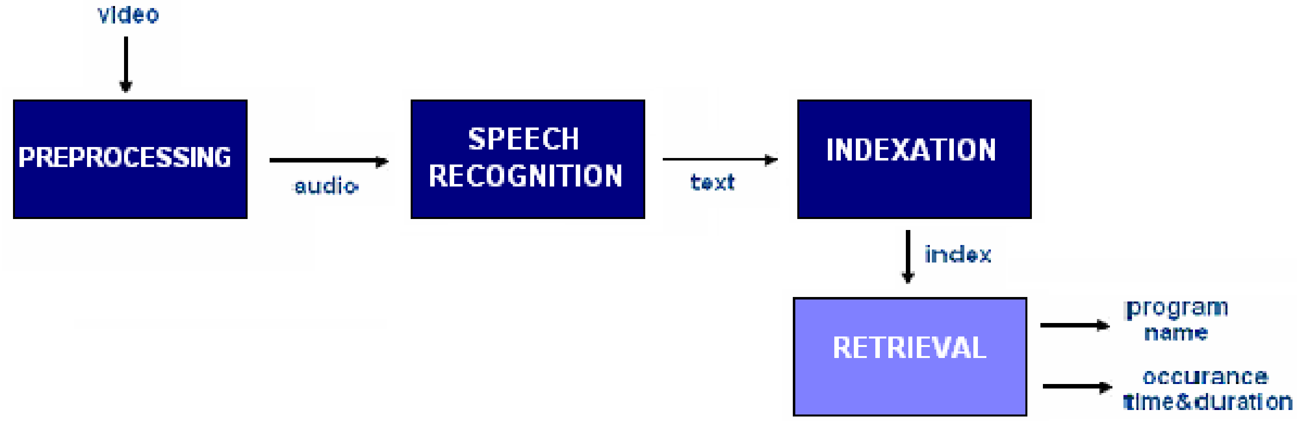
  - Stand alone application
  - Web site



Retrieval Application



SpokenTerm Detection(STD)



video

PREPROCESSING

audio

SPEECH RECOGNITION

text

INDEXATION

index

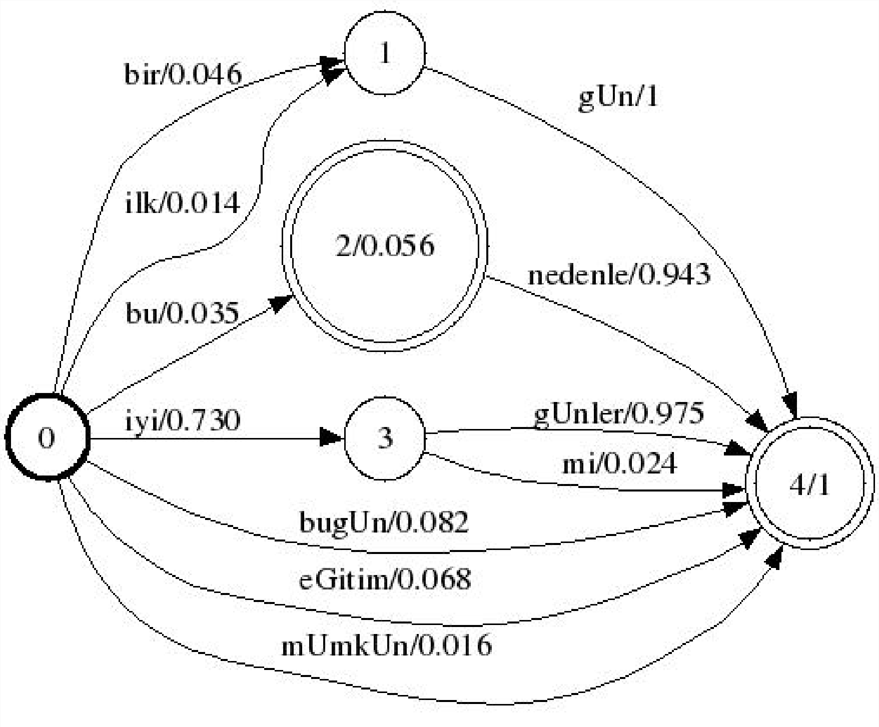
RETRIEVAL

program name

occurance time&duration

- **Speech Recognition**
  - Automatic segmentation based on energy
  - HMM-based LVCSR system
  - ASR output is a weighted finite state automata in form of lattice.
- **Indexation**
  - a weighted finite-state transducer mapping each factor of each utterance to the indices of the automata where it appears and the expected count of the factor.
  - WFST index is beneficial since ASR output is uncertain. It is also optimizes search time.
- **Retrieval**
  - Query is composed with the index FSM, resulting in utterance indices.
  - Forced alignment to find the beginning and duration of query.
- **Server Application**
  - Socket connection is established between client and server. After the query is received from client, server searches for it in the database and returns the results back to client.

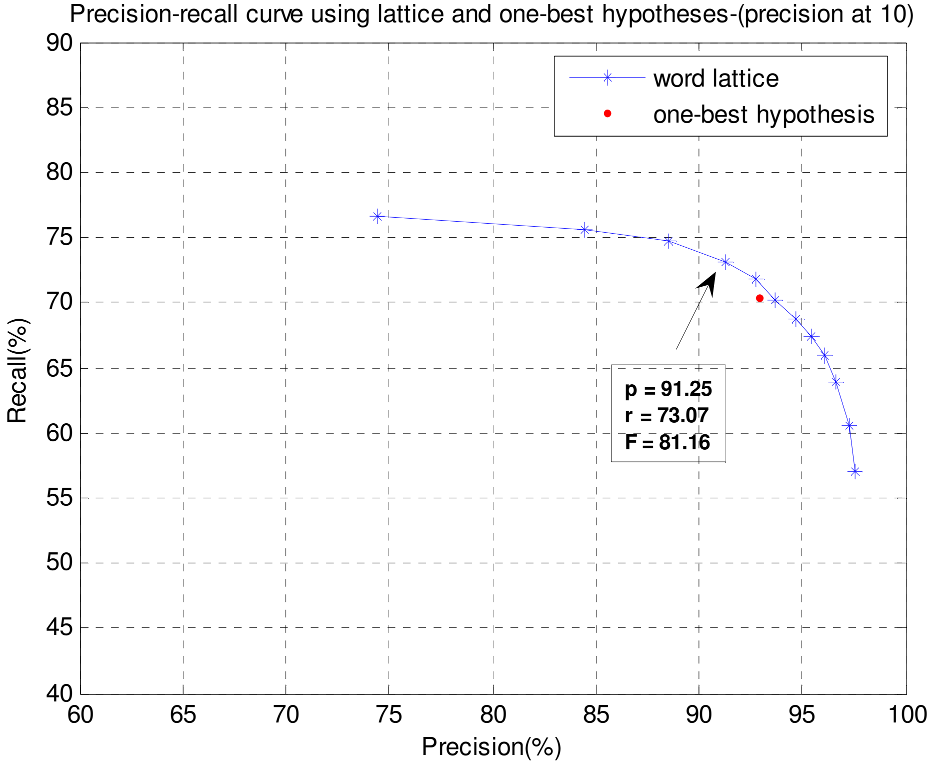
Lattices: Usage&Benefits

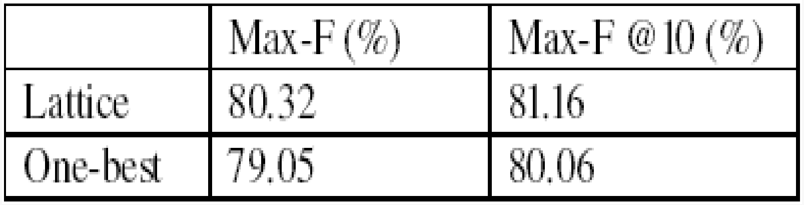


- Arc labels are word hypotheses.

  Arc weights are path probabilities.
- Indexation process estimates expected counts from path probabilities.
- By setting a threshold on expected counts, different precision-recall points can be obtained.
- This corresponds to a curve, while one-best output results in only one point.
- FLEXIBILITY:
  - If precision is more important→ increase the threshold (recall falls)
  - If recall is more important → decrease the threshold (precision falls)

STD RESULTS



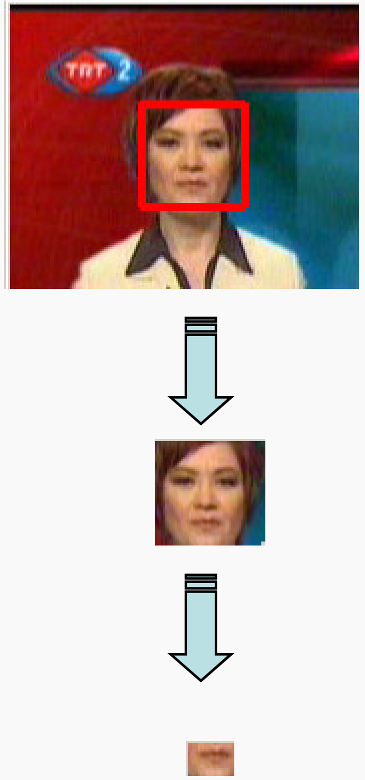


	Max-F (%)	Max-F @10 (%)
Lattice	80.32	81.16
One-best	79.05	80.06

- Metrics:
  - Precision-Recall
  - F-measure
- Evaluation corpora includes 15 of the videos.
- Maximum F-measure (81.16%) is achieved at P=91.25%, R=73.07% point.
- One-best point, indicated with red, is below the lattice curve.
- Use of lattice introduces 1-1.5% improvement on max-F.
- For the sign language tutor application, it may be desirable to operate on high-precision regions.

Lip Reading
Detection & Feature extraction

- Viola and Jones' method for detecting face
- Correlation of each frame with a lip template to extract the lip region with size 24x16
- Future work
    - Use DCT coefficients extracted from the lip region as the visual features to be combined with the audio features



Sliding Text Retrieval

Baseline Method

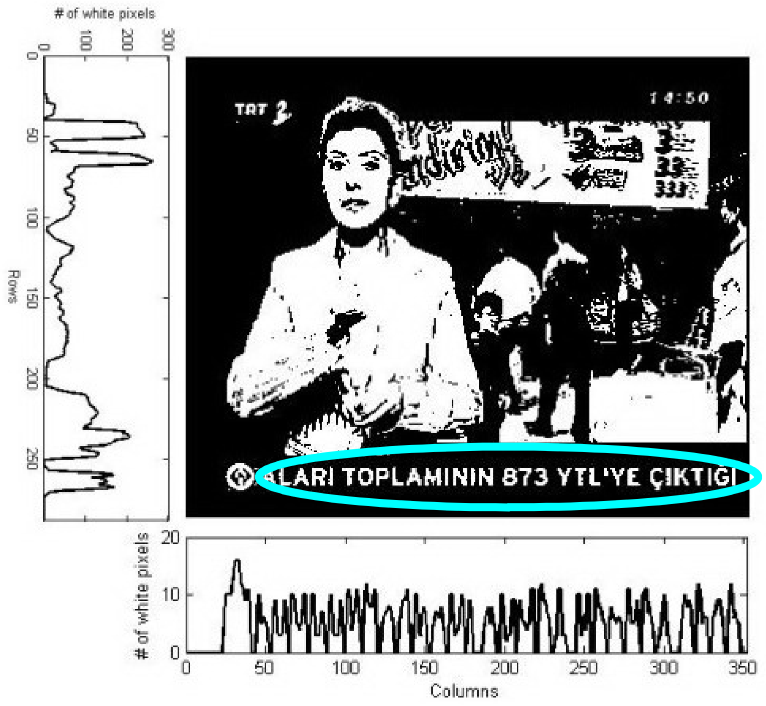
- Text band extraction
- Determination of word and space positions
- Template matching
- Text alignment in every 10 frames
- Noise removal
  - Averaging / Smoothing
  - Morphological operations

Improvements

- Jaccard's binary template match score

$$d_J = \frac{n_{11}}{n_{11} + n_{10} + n_{01}}$$

- Integrating heuristics
- Incorporating language model



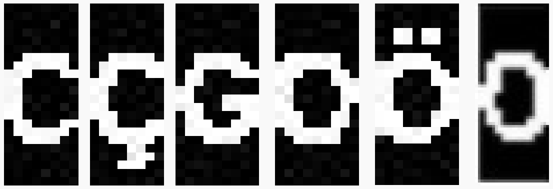
TOPLAMININ 873 YTL'YE ÇIKTIĞI

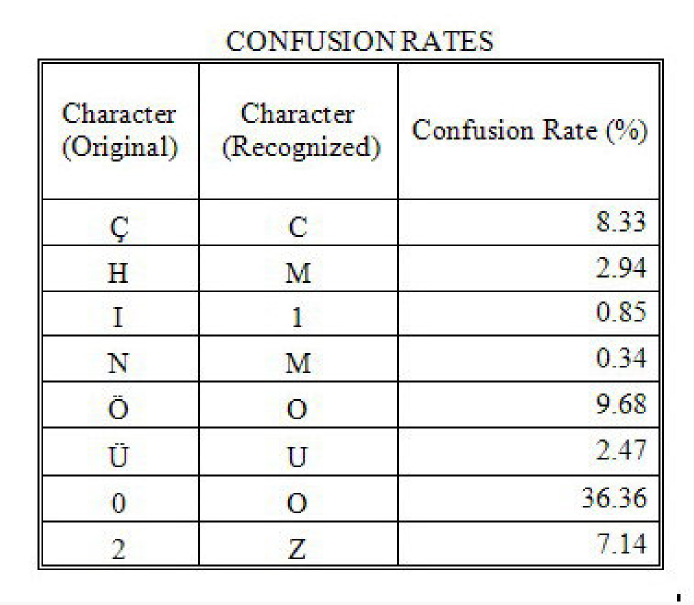
248 253 254 261 263 271 274 282 284 290 294 302 305 314 ...

Sliding Text Retrieval (II)

Problems

- **Low resolution and noise**
  - Distorted images
  - Losing distinctive parts of some Turkish characters
- **Temporal alignment between successive frames**
  - Merged / divided characters
  - Pixel shifts





CONFUSION RATES

Character (Original)	Character (Recognized)	Confusion Rate (%)
Ç	C	8.33
H	M	2.94
I	1	0.85
N	M	0.34
Ö	O	9.68
Ü	U	2.47
0	O	36.36
2	Z	7.14

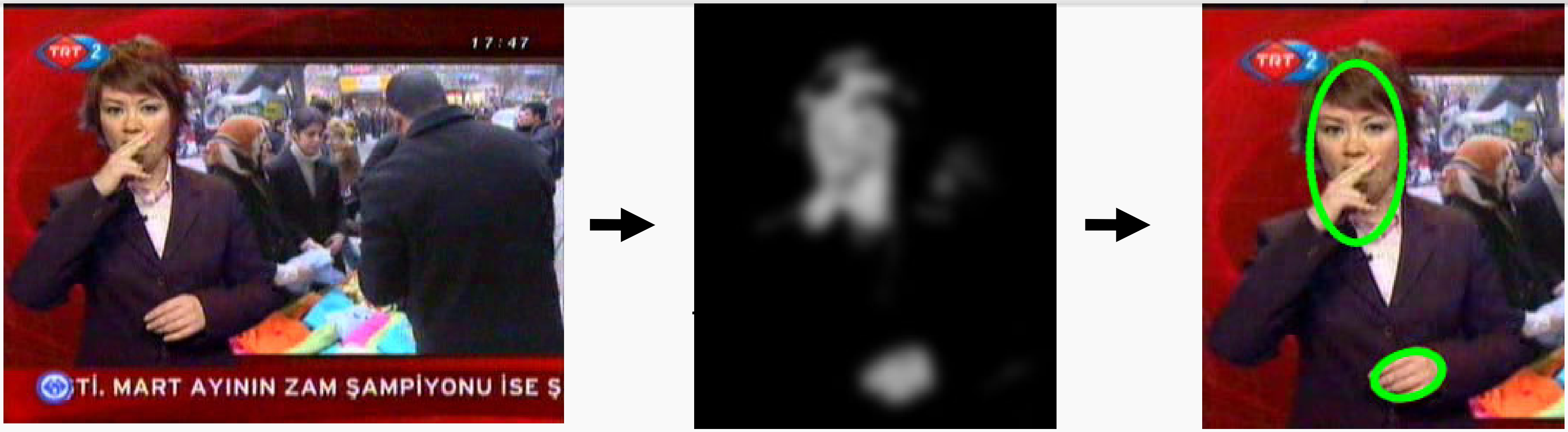
Performance

- Character Recognition Accuracy

  94.0% ➔ 98.5%
- Word Recognition Accuracy

  70% ➔ 90%

Image segmentation by using skin color model



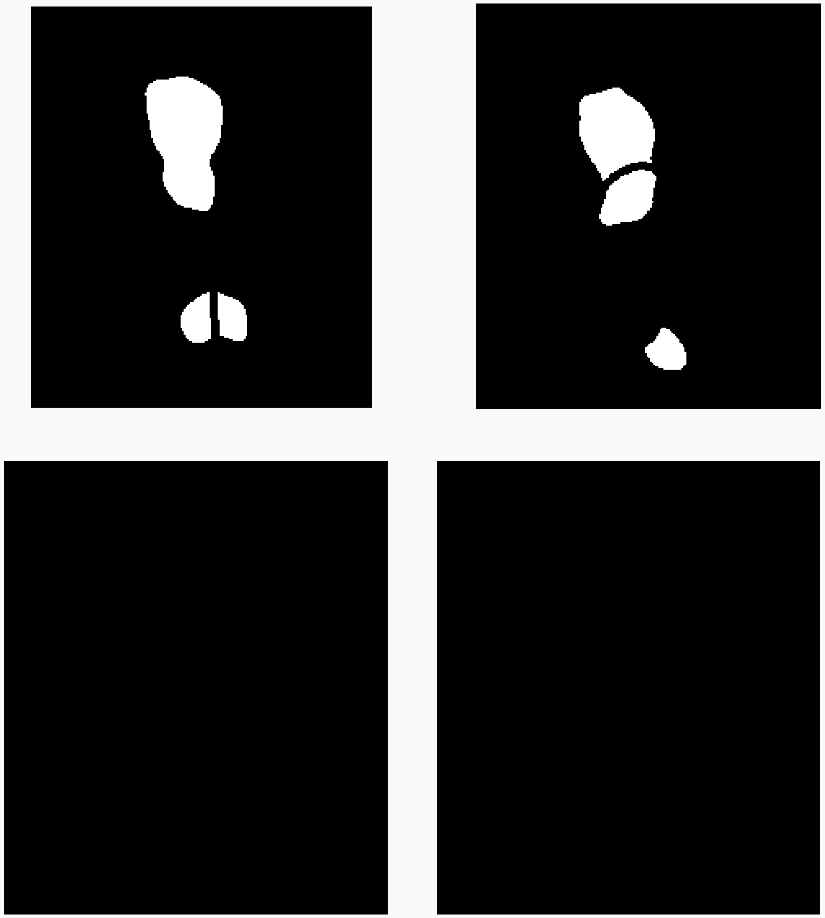
Input: Single frame from TV news

Output: Segmented blobs (head and hands of speaker and people in background)

- GMM of skin color distribution in RGB space
- Adaptation for each speaker
- Connected Components Labeling

Tracking Face and Hands

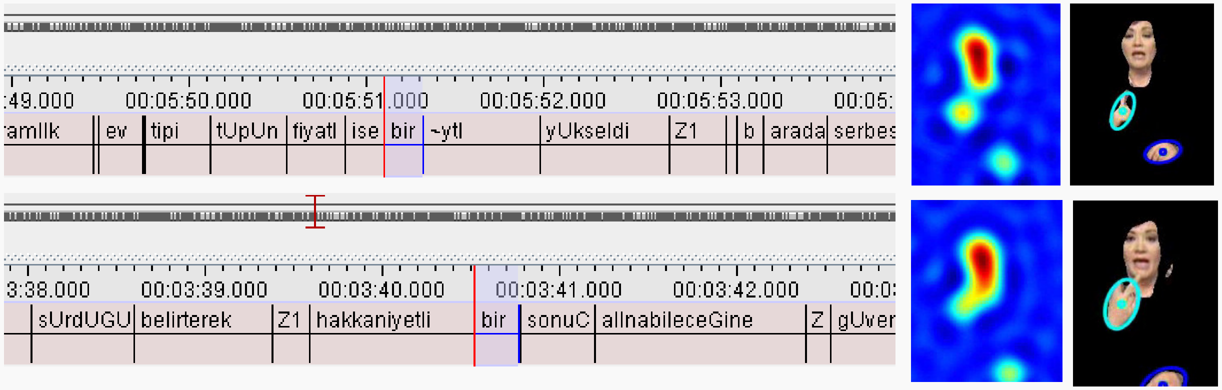
- Handling blobs
  - Blob filters
  - Template acquisition
  - Occlusion prediction & detection
  - Separation of blobs
- Blob tracking
  - Rule based blob classification



Feature extraction

- Tracking algorithm
  - Five features for every blob
  - Position, size and angle of bounding ellipse
- DCT
  - DCT of hand template
  - DCT of whole image
- Together 258 features
  - 15 tracking features
  - 243 DCT features (108 for hands and 135 for image)

Clustering of signs



```
:49.000    00:05:50.000    00:05:51.000    00:05:52.000    00:05:53.000    00:05:
amlIk | ev | tipi | tUpUn | fiyatl | ise | bir | ~ytl | yUkseldi | Z1 | b | arada | serbes

3:38.000    00:03:39.000    00:03:40.000    00:03:41.000    00:03:42.000    00:0
sUrdUGU | belirterek | Z1 | hakkaniyetli | bir | sonuC | allnabileceGine | Z | gUver
```

? → same signs in sequences

? → different signs

Input:

- Two (or more) short video sequences, in which same word was pronounced and where same sign is expected to occur (cca 0.4 seconds each sequence)
- Features extracted from image data for each sequence

Our task:

Cluster the sequences, i.e. determine whether they contain same sign.

(homonyms – same pronouncing in speech but different sign and meaning)

Clustering of signs: Sequences distance estimation

1) Subtraction + Sum

subtract calculated features of two sequences, sum this difference

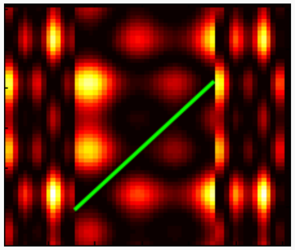
- borders of signs aren‘t accurate, sequences can contain part of previous / following sign -> the distance increases
- time warped data -> the distance increases

2) DTW (multidimensional), in progress…

- possibility to detect borders of same signs, align them and cut previous / following sign
- problems: short sequences, hard to estimate the borders of signs

3) HMM – another possibility, future work…

- could detect borders of signs too



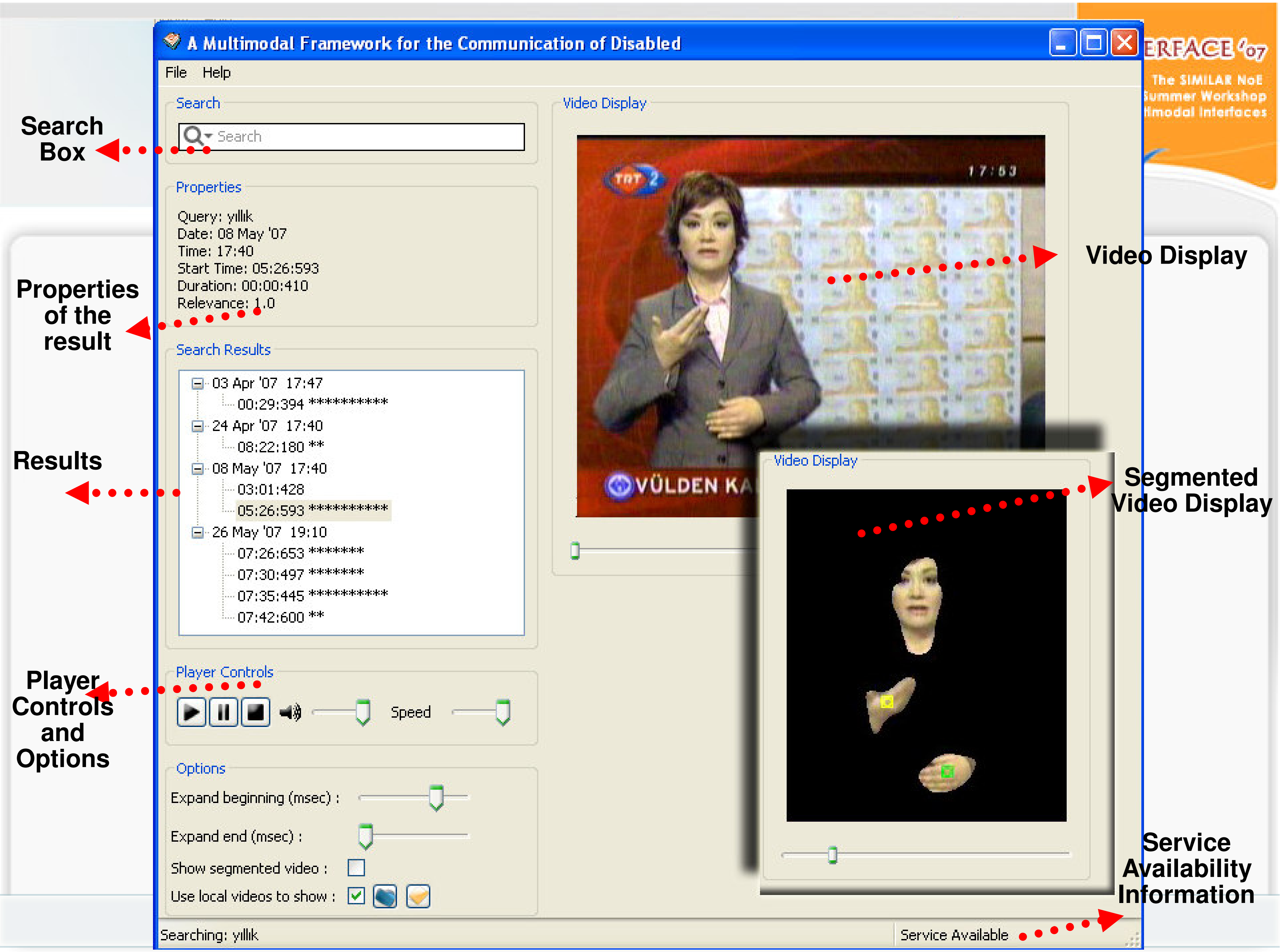
DTW cost matrix

green line: mapping one sign to another

Ideal case

Retrieval Application & GUI

- Seperate GUI and core program:
  - Search engine is on the server
  - Connection via TCP/IP socket
- Used tools:
  - wxPython (GUI widgets)
  - wxFolmBuilder (Separate UI files)
  - py2exe (creating exe from python code)
  - nsis (standalone program setup)



Search Box

Properties of the result

Results

Player Controls and Options

Video Display

Segmented Video Display

Service Availability Information

Properties

Query: yıllık
Date: 08 May '07
Time: 17:40
Start Time: 05:26:593
Duration: 00:00:410
Relevance: 1.0

Search Results

- 03 Apr '07 17:47
  - 00:29:394 **********
- 24 Apr '07 17:40
  - 08:22:180 **
- 08 May '07 17:40
  - 03:01:428
  - 05:26:593 **********
- 26 May '07 19:10
  - 07:26:653 *******
  - 07:30:497 *******
  - 07:35:445 **********
  - 07:42:600 **

Expand beginning (msec) :

Expand end (msec) :

Show segmented video : □

Use local videos to show : ☑

Searching: yıllık

Service Available

DEMO