# Multimodal Person Recognition using Unconstrained Audio and Video

28. January 2003

Presented by: Stefan Rondinelli

# Outline

1. Motivation
2. Face Recognition
3. Speaker Identification
4. Bayes Net
5. Results

# Motivation

- Automatic banking
- Password-free computer login
- Person dependent behavior

# Face Recognition

## What are the requirements?

- Face must be found in any kind of background
- Should recognize a person despite wide variations in pose and facial expression
- Don`t let fool the system by a photograph

# Face Recognition

## Face Detection and Tracking

**1. Detect the face using skin color information**

- The skin color is modeled with a mixture of Gaussians
- The model is trained with faces with varying skin tone and under different lighting conditions

# Face Recognition

## Face Detection and Tracking

**2. Detect the features (eyes, mouth, etc.)**

**=>> The positions of the features give an estimate of the pose**

**3. Warp the detected face to a frontal view**

**=>> Use the pose estimate and a 3D head model**

# Face Recognition

## Eigenspace Modeling

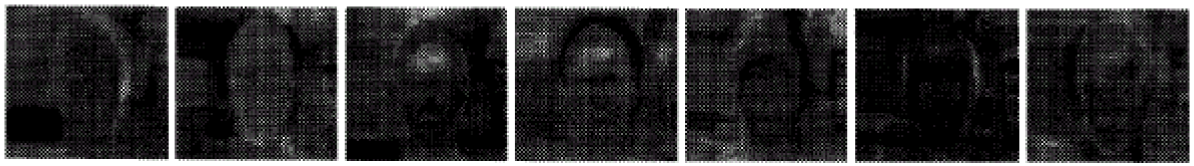
**Preparation**

- Search the face for exact positions of the features
- Normalize the face such that eyes and mouth are at fixed locations

# Face Recognition

## Eigenspace Modeling

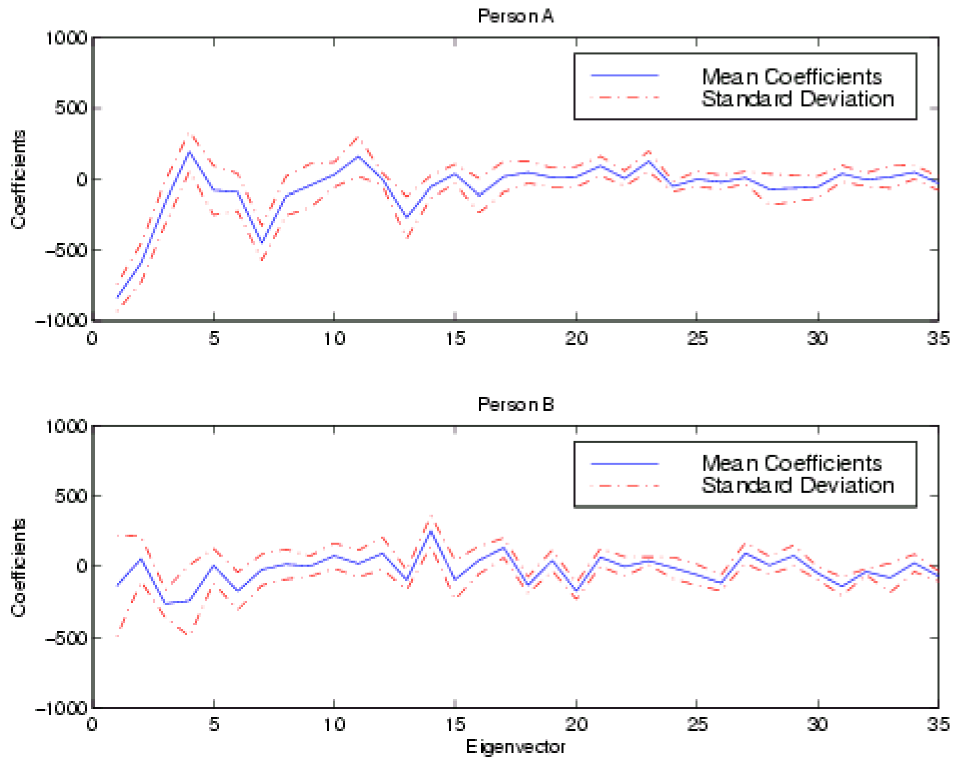
Convert from „pixel space“ into „face space“



Basis vectors of the „face space“

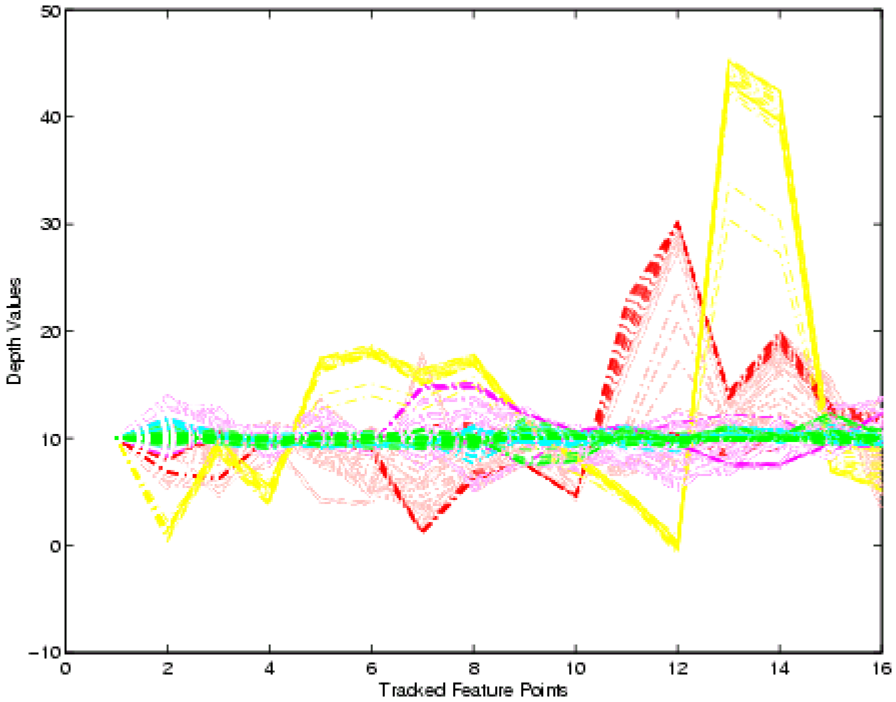
# Face Recognition

The first 35 coefficients of two persons



# Face Recognition

## Depth Estimate



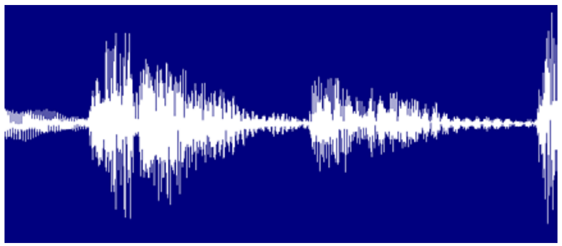
# Speaker Identification

What are the requirements?

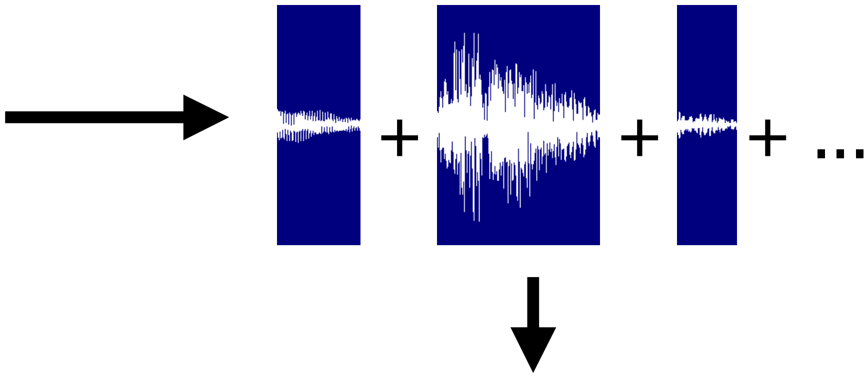
- Should work also in a noisy environment

# Speaker Identification

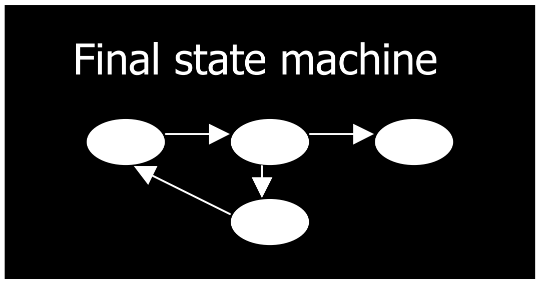
## Event Detection



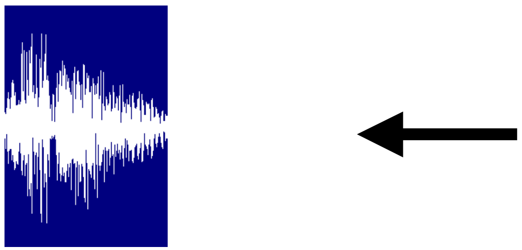
Audio stream









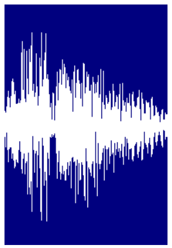


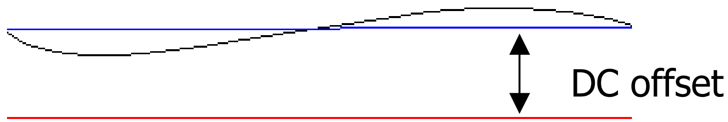
Segment containing speech

- Threshold total energy
- Constraints on event length and surrounding pauses

# Speaker Identification

## Feature Extraction






Remove the DC offset

Calculate Mel-scaled frequency coefficients for frames that are spaced 16ms apart and 32ms long

# Speaker Identification

## Modeling

One HMM (Hidden Markov Model) for each person

→ **Initialization**

by using segmental k-means

**Maximization of the model likelihood**

by using the EM algorithm

(Expectation-Maximization)

# Speaker Identification

## Background Adaption

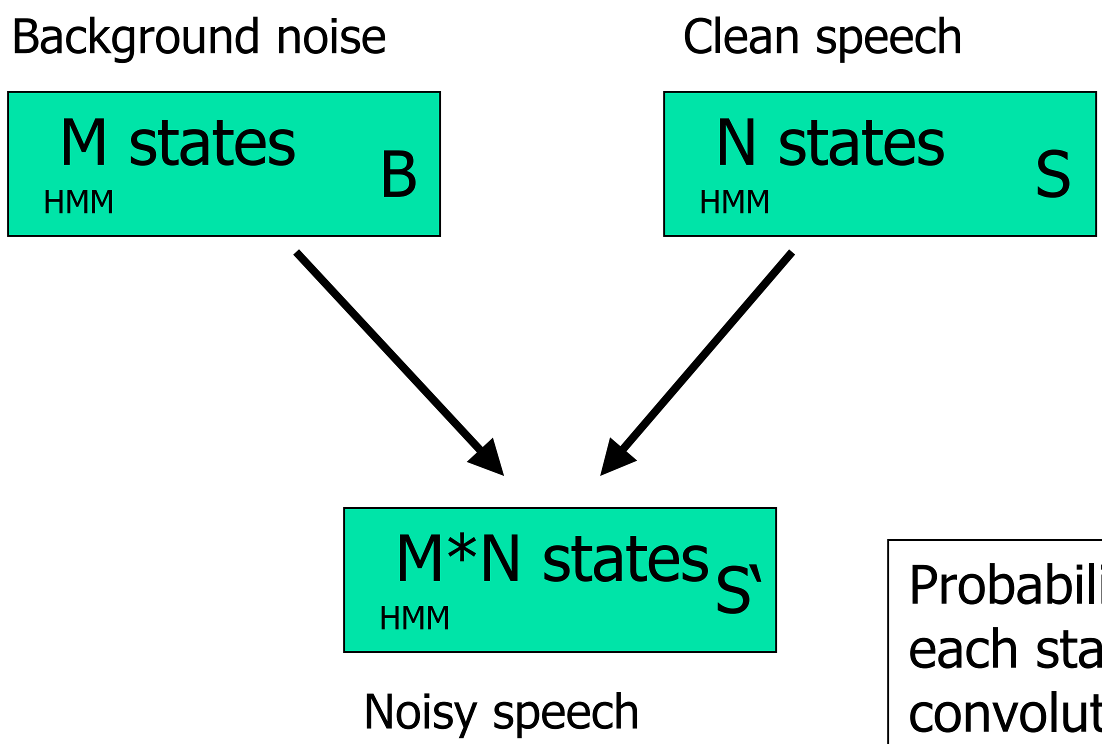
**2 types of noise**

- Convolutional noise → Equipment Assumed to be constant
- Additive noise → Repetitive noise / Randomly occurring noise

| Repetitive noise | Randomly occurring noise |
|---|---|
| eg. motor noise | eg. thunder in a rain storm |

# Speaker Identification

## Background Adaption




Probability distribution for each state in S` is the convolution of the distributions in S and B

# Speaker Identification

## Background Adaption

| HMM Models | Speech Only | Speech + Noise |
|---|---|---|
| Speech Only ($S$) | 71.5% | 23.1% |
| Adapted ($S'$) | N/A | 65.4% |
| Corrupted ($C$) | N/A | 69.2% |

HMM S: Clean Speech

HMM S‘: Clean Speech * Noise

HMM C: Clean Speech and Noise (for evaluation)

# Bayes Net

## Confidence Scores

Distance from Face Space (DFFS)

$$DFFS(x) = \| x - \bar{x} \|_{Eigenspace}$$

Aggregate Model Likelihood (AML)

$$AML(x) = \log\left( \sum_j P\left(x \mid Model_j\right) \right)$$

Maximum-Probability to Average-Probability Distance (MPAP)

$$MPAP(x) = \max_j \{\mathrm{P}(\mathrm{X} = \mathrm{j})\} - \frac{1}{\mathrm{N}} \sum_j P(X = j)$$

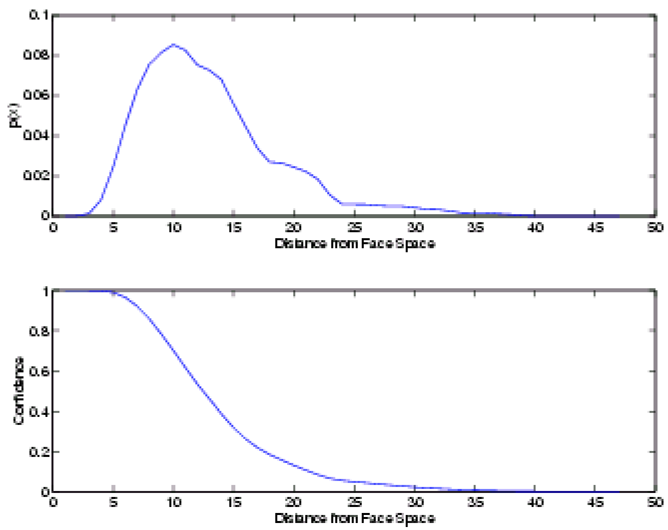
# Bayes Net

Convert measures into probabilities:

Let $p(M(x)) = pdf$

$\rightarrow$ $$confidence(\omega_0) = P(\omega < \omega_0) = \int_{\infty}^{\omega_0} p(\omega) d\omega$$

$$\omega = M(x)$$



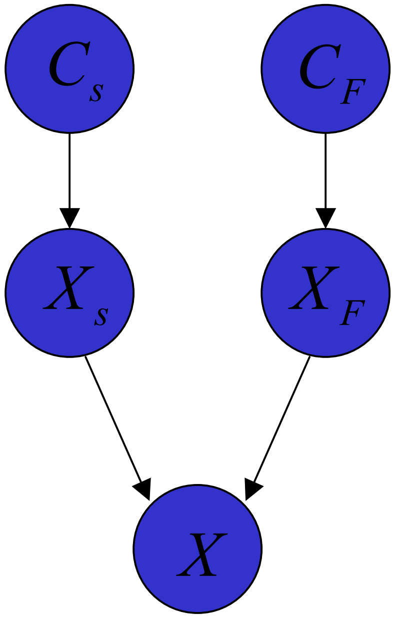
# Bayes Net

| Confidence Score | Speech | Face |
|---|---|---|
| DFFS | N/A | 55.3%,90.0% |
| AML | 50.2%,47.6% | N/A |
| MPAP | 71.4%,50.3% | 99.1%,53.4% |

Comparison of Confidence Scores: Prediction rates of Correct Recognition (left) and Wrong Recognition (right)

- The percentages are based on the correlation between the confidence scores and the correctly or incorrectly recognized test cases.
- 50% (chance) means that the confidence score is uncorrelated with recognition

# Bayes Net




$C_s$ Speech Confidence

$C_F$ Face Confidence

$X_s$ Speaker Identity

$X_F$ Face Identity

$X$ Person Identity

**Other probabilities:**

$P(C_i)$ Recognition rate for each classifier

**Knowledge sources:**

$P(X \mid X_i)$ Classifier's probability for each person

$P(X_i \mid C_i)$ Confidence in the classifier

Where $C_i$ = {reliable, not reliable}, $X_i$ = {j|j∈ Client database}

$$P(X) = P(X \mid X_S)P(X_S \mid C_S)P(C_S) + P(X \mid X_F)P(X_F \mid C_F)P(C_F)$$

# Results

Using only the most reliable image/clip pair

Recognition Rates (Zero Rejection Threshold)

| Modality | Per Image/Clip | Per Session |
|---|---|---|
| Audio | 71.2 % | 80.8 % |
| Video | 83.5 % | 88.4 % |
| Audio + Video | 93.5 % | 100 % |

Recognition Rates (Optimal Rejection Threshold): the rejection rates are in parentheses

| Modality | Per Image/Clip |
|---|---|
| Audio | 92.1% (28.8%) |
| Video | 97.1% (17.7%) |
| Audio + Video | 99.2% (55.3%) |

Verification Rates (Optimal Rejection Threshold): false acceptance rates are in parentheses

| Modality | Per Image/Clip | Per Session |
|---|---|---|
| Audio | 97.8 % (0.2%) | 98.5 % (0%) |
| Video | 99.1 % (0.2%) | 99.6 % (0%) |
| Audio + Video | 99.5 % (0.3%) | 100 % (0%) |