

# The Influence of Intra-Speaker Variability in Automatic Speaker Identification

Timo Becker

Austrian Academy of Sciences  
Acoustics Research Institute



IAFPA 2007

The College of St Mark & St John  
Plymouth, UK



# Outline

Intra-Speaker  
Variability

Timo Becker

Introduction

Method

Results

Discussion

Conclusion

**1** Introduction

**2** Method

**3** Results

**4** Discussion

**5** Conclusion

# Introduction

## Need for Investigation

### Intra-Speaker Variability

Timo Becker

### Introduction

Method

Results

Discussion

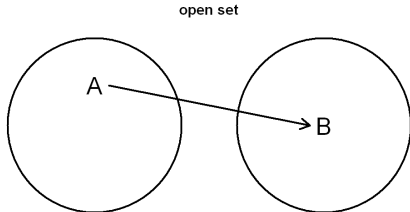
Conclusion

- Ongoing application of automatic speaker recognition in forensic domain
- Influencing factors? Coverage?
- Main influencing factors (channel mismatch, amount of data) already investigated in huge automatic tests (e. g. by NIST)
- Some factors from real forensic casework remain

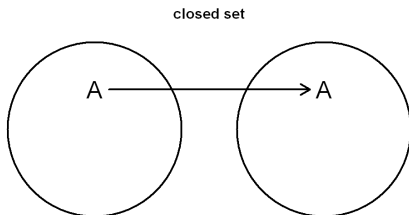
- *Inter-speaker variability*
  - Variation of speech observable as acoustic differences of speech from different speakers
- *Intra-speaker variability*
  - Changes of speech caused by different speaking styles  
→ voice quality, articulation rate, stress, pitch contour etc.
  - *Not:* external influences (e. g. transmission channel)

- Intra-speaker variability can be mistaken for inter-speaker variability → misinterpretations
- Applies to both human and machine
- Machine: unsupervised feature and model generation → not visible
- Feature extraction parameters for machines derived experimentally → robustness known?
- Mismatch of speaking styles in forensic casework:
  - Spontaneous vs. read speech
  - non-Lombard vs. Lombard speech

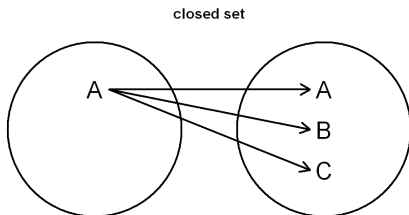
- Automatic speaker identification task:  
*associate a speaker with a speaker from a set of speakers*



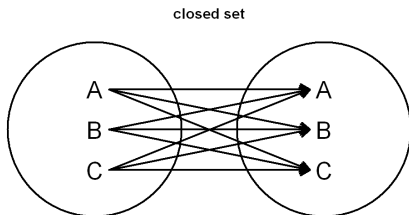
- Automatic speaker identification task:  
*associate a speaker with a speaker from a set of speakers*



- Automatic speaker identification task:  
*associate a speaker with a speaker from a set of speakers*



- Automatic speaker identification task:  
*associate a speaker with a speaker from a set of speakers*



- 'Pool 2010' data base (Department of Speaker Identification and Audio Analysis, Bundeskriminalamt, Germany)
- 100 male German speakers
- GSM-encoded telephone channel recordings
- Different experimental settings
  - Reading (R)
  - Spontaneous (S)
  - Free (F)
  - Lombard (L)
- Combinations
  - Reading Free (RF)
  - Reading Lombard (RL)
  - Spontaneous Free (SF)
  - Spontaneous Lombard (SL)

- Reading: 'The North Wind and the Sun'
- Spontaneous: Description of pictures
- Lombard: 80 dB<sub>SPL</sub> white noise over headphones

For details see:

Jessen, M., Köster, O. and Gfroerer, S. (2005). Influence of vocal effort on average and variability of fundamental frequency. *Speech, Language and the Law*, 12, 174-213.

# Method

## Data Base

### Mismatch Conditions

Intra-Speaker  
Variability

Timo Becker

Introduction

Method

Results

Discussion

Conclusion

	RF	SF	RL	SL
RF	-	SINGLE	SINGLE	DOUBLE
SF	SINGLE	-	DOUBLE	SINGLE
RL	SINGLE	DOUBLE	-	SINGLE
SL	DOUBLE	SINGLE	SINGLE	-

R/S mismatch

F/L mismatch

R/S & F/L mismatch

- 8 kHz, 16 bit signals
- 20 ms frames every 10 ms
- Hamming window
- Pre-emphasis 0.95
- Automatic speech detection
  - 36 s average duration after detection for reading
  - 119 s average duration after detection for spontaneous
  - 27 s overall minimum duration after detection
- Power spectrum (FFT)
- 23 triangular Mel filters (300–3370 Hz)
  - logarithmic filter coefficients
- Cepstral coefficients 1–14 (DCT), discarding  $c_0$ 
  - Mel frequency cepstral coefficients (MFCC)

- Speakers models: Gaussian mixture models (GMM)
- $d$ -variate Gaussian distribution function

$$f_i(x) = \frac{1}{(2\pi)^{d/2} \det(\Sigma_i)^{1/2}} e^{-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1} (x-\mu_i)}, \quad (1)$$

where  $\mu$  is mean and  $\Sigma$  is covariance matrix

- Gaussian mixture density function is weighted sum of  $M$  Gaussian distribution functions

$$f(x) = \sum_{i=1}^M p_i f_i(x), \quad \sum_{i=1}^M p_i = 1 \quad (2)$$

- GMM  $\lambda$  consisting of  $M$  Gaussians

$$\{(p_i, \mu_i, \Sigma_i) : i = 1, \dots, M\} \quad (3)$$

- Model estimation by expectation maximisation
- 32 Gaussians per model
- Diagonal covariance matrices
- Similarity of feature vectors  $X = \{x_k, \dots, x_n\}$  and speaker model  $\lambda$  by likelihood

$$l(X|\lambda) = \prod_{k=1}^n f(x_k) \quad (4)$$

- Performance measured as the identification rate

$$\text{IR} = \frac{\text{number correct assignments}}{\text{number total assignments}} \quad (5)$$

# Results

## Identification Rates

Intra-Speaker  
Variability

Timo Becker

Introduction

Method

Results

Discussion

Conclusion

	RF	SF	RL	SL
RF	-	0.88	0.69	0.53
SF	0.92	-	0.44	0.57
RL	0.79	0.66	-	0.84
SL	0.57	0.69	0.92	-

R/S mismatch

F/L mismatch

R/S & F/L mismatch

# Results

## Average Identification Rates

### Intra-Speaker Variability

Timo Becker

Introduction

Method

Results

Discussion

Conclusion

	IR	cases
overall	0.71	12
double mismatch	0.55	4
single mismatch	0.79	8
only R/S mismatch	0.89	4
only F/L mismatch	0.68	4

- Intra-speaker variabilites cause performance losses
- High variability → low identification rate
- Lowest identification rate when both conditions mismatch

- GMM: MFCC feature vectors statistically independent  
→ no temporal information
- No inclusion of  $c_0$  → no energy information
- Removal of higher cepstral features → smoothed spectrum (on cosine base) → little pitch information included
- Summary: Unordered vocal tract transfer function information (distribution) → low identification rates caused by spectral changes and/or changes in distribution of spectral information

- Speakers apply rules and processes when speaking style is changed
- Rules are individual → differ for speakers
- Different deviations between reading/spontaneous
  - Changes of speech segment durations
  - Changes of speech segment length relations
  - Changes of spectral properties
  - Extent of changes individual
- Different deviations for free/Lombard
  - Changes of intensity
  - Changes of vowel duration
  - Changes of spectral properties
  - Extent of changes individual
- → changes influence MFCC properties and GMM parameters (supported by results)
- More changes in F/L mismatch than in R/S mismatch
- F/L and R/S different changes

# Conclusion

## Intra-Speaker Variability

Timo Becker

Introduction

Method

Results

Discussion

Conclusion

- Mismatch in speaking styles leads to performance loss in automatic speaker identification
- Speaker specific strategies cause intra-individual variability (non-systematic and unpredictable)
- Since automatic speaker identification is unsupervised, only exclusion of mismatches can guarantee system stability
- In forensic settings, careful investigation of speaking styles should precede application of automatic systems

Thanks to the

*Department of Speaker Identification and  
Audio Analysis  
Bundeskriminalamt, Germany*



# The Influence of Intra-Speaker Variability in Automatic Speaker Identification

Intra-Speaker  
Variability

Timo Becker

Introduction

Method

Results

Discussion

Conclusion

Thank you  
for your attention.