



Bundeskriminalamt



# FORENSIC VOICE COMPARISONS IN GERMAN WITH PHONETIC AND AUTOMATIC FEATURES USING VOCALISE SOFTWARE



**AES 54th International  
Conference, London, UK, 2014  
June 12–14**

**Michael Jessen,  
Anil Alexander &  
Oscar Forth**

[michael.jessen@bka.bund.de](mailto:michael.jessen@bka.bund.de)  
[{anil|oscar}@oxfordwaveresearch.com](mailto:{anil|oscar}@oxfordwaveresearch.com)





# Structure

1. Introduction and theory
  - a. Forensic Voice Comparisons and different traditions of performance testing: proficiency testing and system evaluations
  - b. Overview of VOCALISE and its main design features
2. Demonstration of software operation and results
  - a. System evaluations with VOCALISE and Bio-Metrics on **lab-speech data** based on MFCC and long-term formants
  - b. System evaluations with VOCALISE and Bio-Metrics on **real-case data** (MFCC)



# Forensic Voice Comparison: Methods

1. auditory-phonetic and linguistic analysis  
(regional/social varieties and „idiolect“; „paralinguistic“ features, such as voice quality, fluency interruptions, breathing patterns, speech pathology)

2. acoustic-phonetic analysis (e.g. f0, formants, articulation rate)

3. Automatic speaker recognition

(cf. Gold & French 2011)

auditory-acoustic approach

THE INTERNATIONAL PHONETIC ALPHABET (revised to 2005)											
© 2005 IPA											
	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b		t d		t̪ d̪	c j	k g	q G		?	
Nasal	m	n̪	n		n̪	j̪	ŋ	ŋ̪	N		
Till		B		r					R		
Tap or Flap		v̪		t̪		t̪̪					
Fricative	f β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ɿ	x ɣ	χ ʁ	h ɦ	
Lateral fricative			l̪	ɬ							
Approximant	w		r		ɻ	ɻ̪	ɻ̪̪				
Lateral approximant			l̪		ɬ	ɬ̪	ɬ̪̪				

Where symbols appear in pairs, the one to the right represents a voiced consonant. Shaded areas denote articulations judged impossible.

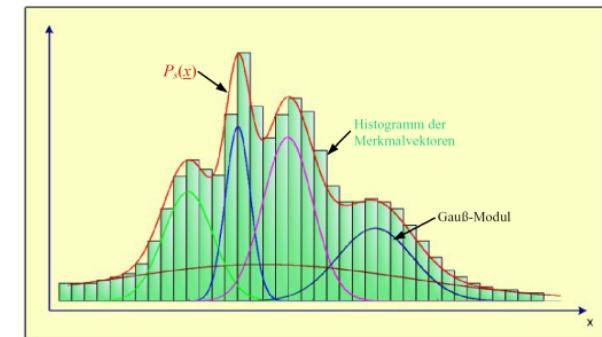
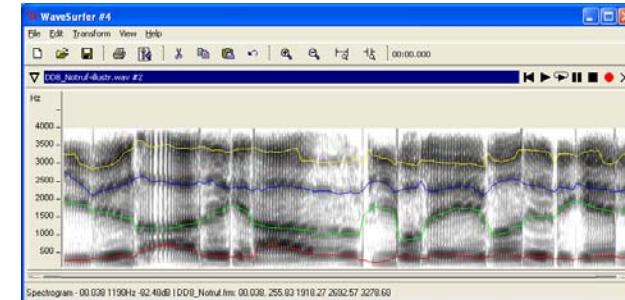


Abbildung 3.5: schematische Darstellung eines GMM-Modells



# Forensic Voice Comparison: Traditions of performance testing

## I. Proficiency tests and collaborative exercises (cf. Cambier-Langeveld 2007; various ENFSI documents)

- Concept: Inter-laboratory tests, limited to a few comparisons, using the full range of methods used in casework.
- Advantage: high representativeness for casework.
- Disadvantage: very limited statistical robustness (very few comparisons per test; test about once per year, but often less frequently than that).

## II. System evaluations (cf. many papers in automatic speaker recognition; papers by Rose, Morrison et al. on LR-based acoustic-phonetic analysis)

- Concept: Many comparisons, based on a restricted number of features that can be processed in an semiautomatic or automatic fashion.
- Advantage: high statistical robustness (many tests; many comparisons per test); many meaningful, performance indicators (e.g. EER, Cllr, Tippett plots).
- Disadvantage: Only some of the features applied in casework are tested.



## Forensic Voice Comparison: Traditions of performance testing

- Both proficiency tests/collaborative exercises and system tests are important due to their mutual advantages and disadvantages.
- The goal should be to increase the number of features that can undergo system evaluations.
- System evaluations should not be limited to automatic speaker recognition (where they are most well-known), but should also include acoustic-phonetic or even auditory-phonetic / linguistic features.
- VOCALISE (along with Bio-Metrics) is a tool that enables system evaluations based on automatic speaker recognition and phonetics



# Design features of VOCALISE I

(Voice Comparison and Analysis of the Likelihood of Speech Evidence)

I. Common platform for automatic speaker recognition and phonetics-based methods of forensic voice comparison

- **Spectral**: extraction of the kind of features that are most commonly used in automatic speaker and speech recognition (currently MFCCs).
- **User (-defined)**: users upload their own stream(s) of independently measured phonetic values, such as formant frequencies, fundamental frequency, or durations of sounds.
- **Autophonetic**: automatic (unsupervised) extraction of phonetic features (currently formants F1 to F4 selected in any combination for analysis).

These different features types undergo modelling (GMM) and likelihood score calculation within the same methodological framework.



## Design features of VOCALISE II

II. Control over different relevant analysis parameters, including, but not limited to:

- Number of Gaussians
- Number of MFCCs (in the Spectral mode)
- In- or exclusion of Delta features
- In- or exclusion of various forms of Channel Normalisation
- Specification of a file minimum duration threshold



## Design features of VOCALISE II

II. Control over different relevant analysis parameters, including, but not limited to:

- Providers of automatic speaker recognition software usually have their parameter settings “hardwired” into their system. This is based on solid research, using speaker corpora.
- However, the type of audio material found in casework might differ from the development data of the software providers.
- This is an argument to give the user the opportunity to find their own best parameter settings based on the audio data that they encounter in their casework.
- Furthermore, still very little is known about the best parameter settings in the processing of *phonetic* data (e.g. how many Gaussians should be used?) This is another argument for user-access to the parameters.



## Design features of VOCALISE III

### III. User-friendliness and audio interface

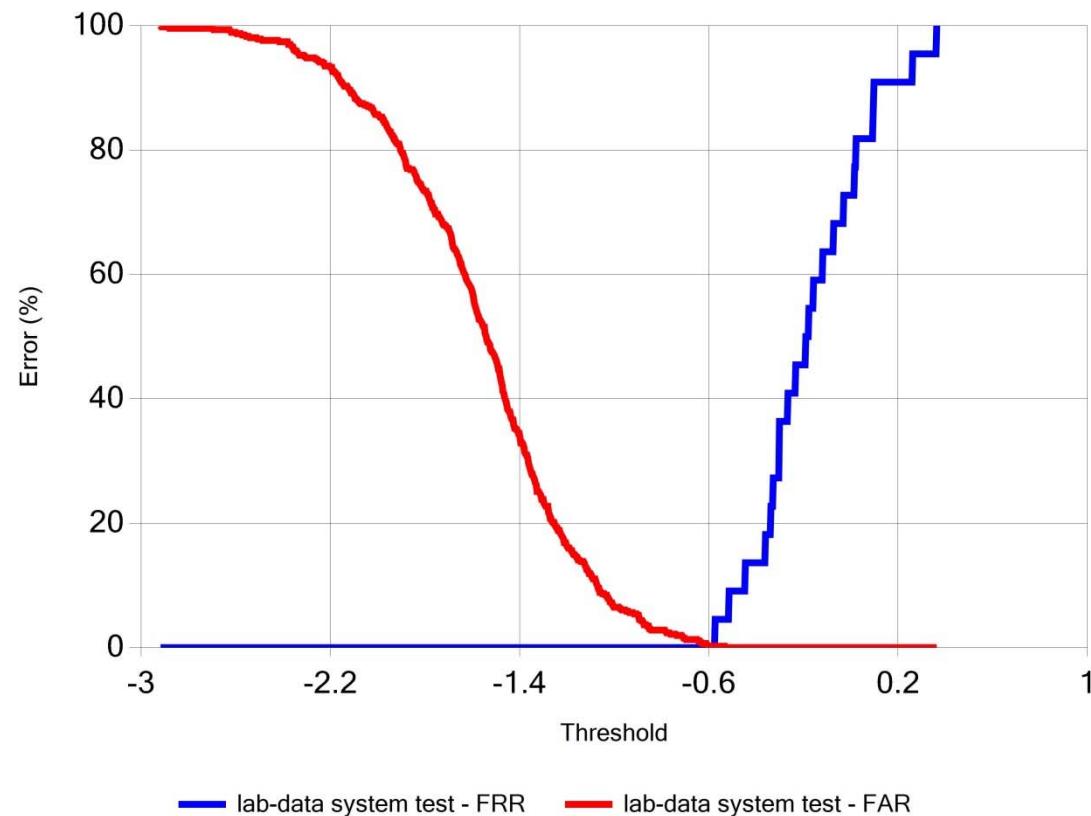
- Some freeware for system evaluations based on phonetic features such as e.g. formant measurements is available as but requires in-depth knowledge of R, Matlab or other R&D environments.
- Most forensic practitioners lack the knowledge, time or enthusiasm to make use oft these resources.
- If the software isn't user-friendly the methods (such as Likelihood Ratio-based evaluations of formant measurements or f0) will simply not be used at all, although they might be important.
- Access to the audio files during all stages of the analysis can help in the interpretation of the results.

# Lab-speech data: Speech corpus Pool 2010

- 21 male adult speakers of the West-Central regional variety of German
- From each speaker, one questioned recording and one suspect recording, resulting in 22 same-speaker comparisons and 462 different-speaker comparisons. Studio recordings which were subsequently transmitted via authentic mobile phone connections.
  - Questioned recordings from a (nearly) spontaneous task in Pool 2010 (commenting on the experiment)
  - Suspect recordings from a semi-spontaneous task in Pool 2010 (describing pictures while avoiding certain keywords)
- UBM based on 22 other speakers of the same variety speaking in semi-spontaneous style
- The net duration of the files was between about 20 and 40 seconds.
- Vowel set F1, F2, F3 was used; the original studio recordings were mobile-phone transmitted
- For GMM, the number of Gaussians was varied.



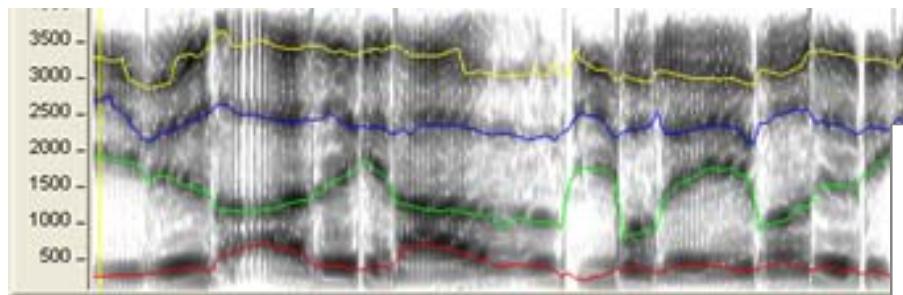
## Results Spectral (MFCC-based): Tippett plot



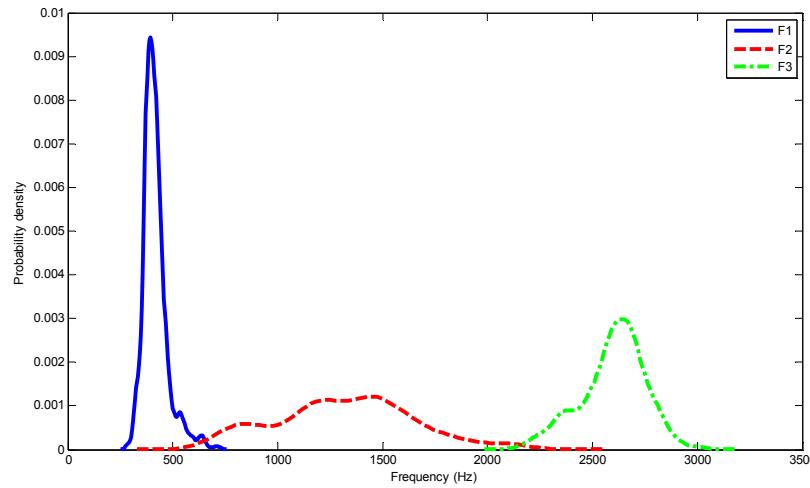
Very good speaker separation, EER close to zero

# Results User (Long-term formants): Methods and EER with different parameter settings

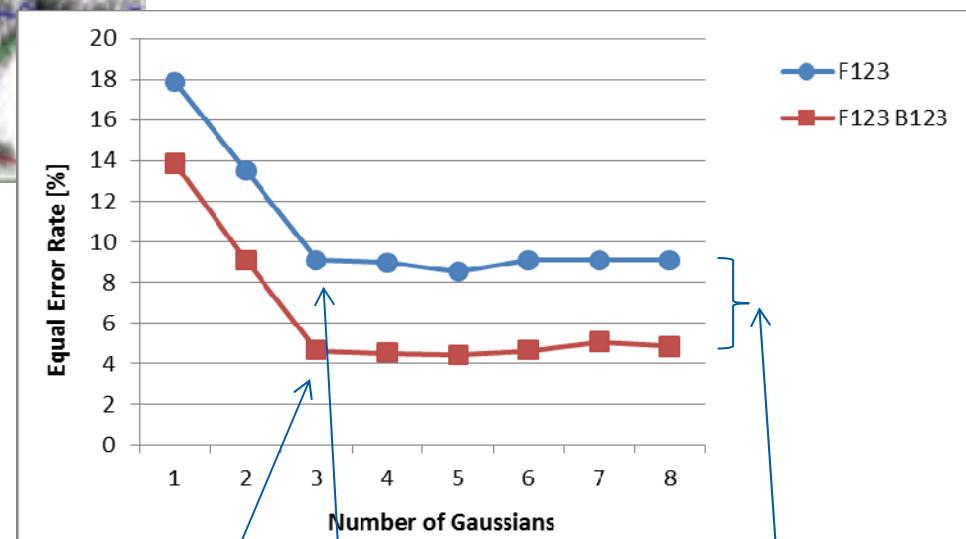
## Method



Typical long-term-formant distribution  
of a speaker



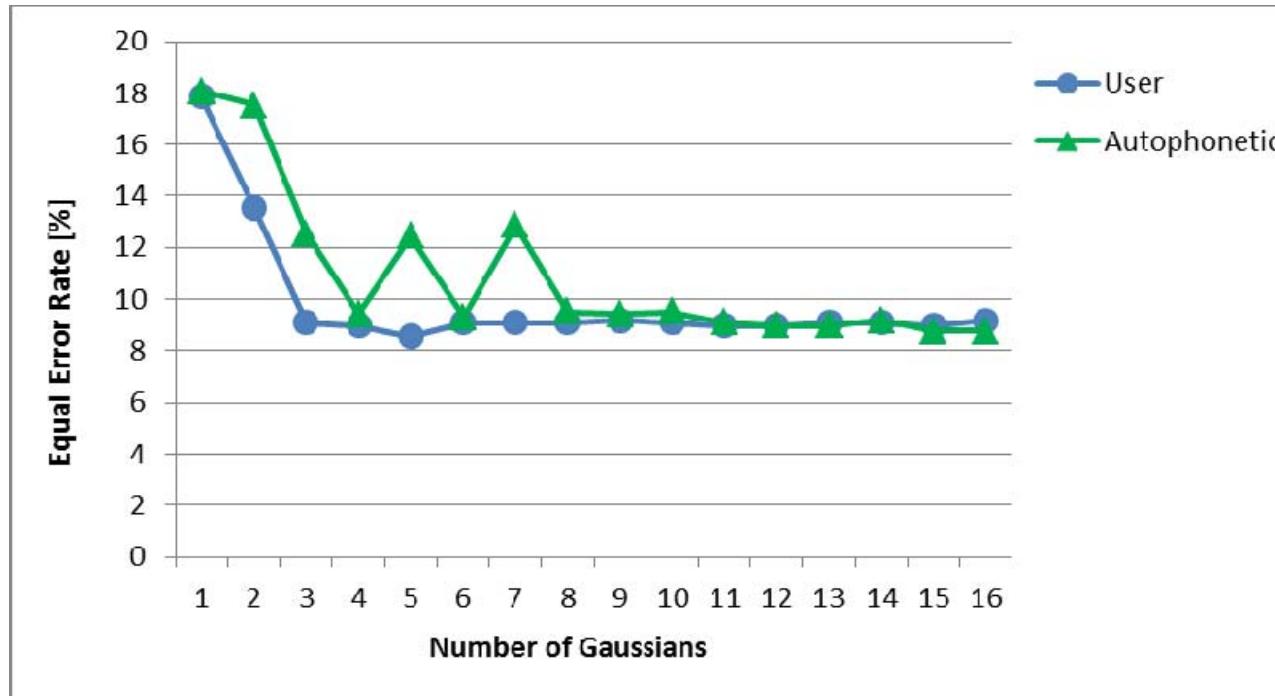
## Results



At least 3 Gaussians necessary

Better results with bandwidths included (this does not carry over to real-case data)

## Results User compared to Autophonetic (Long-term formants)



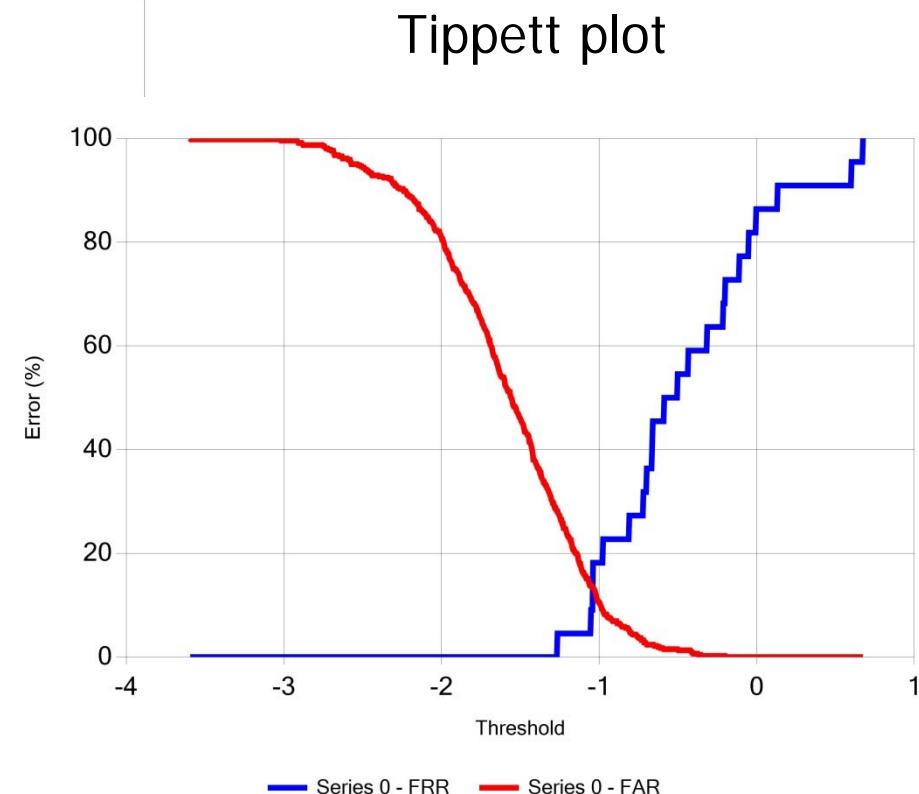
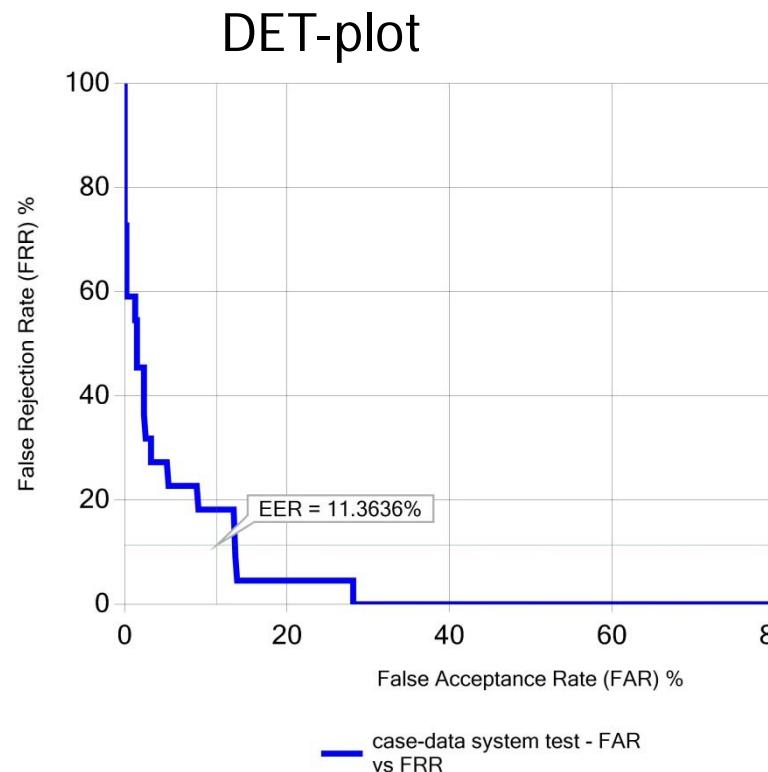
With good-quality data like in Pool 2010 (though still GSM-transmitted) automatic and manual formant analysis yield equivalent results with # Gaussians > 7.



## Real-case data: Telephone interception

- Adult males and speaking German, some of whom had regional or ethnic accent.
- From each speaker, one questioned recording and one suspect recording, resulting in 22 same-speaker comparisons and 462 different-speaker comparisons.
- UBM based on 22 other speakers from a telephone recordings of male adult speakers with regional accents; quality is roughly equivalent to the case recordings.
- The net duration of the files was between about 20 and 60 seconds.
- Spectral (MFCC-based) module was used.

# Results Spectral (MFCC-based): DET-Plot and Tippett plot



EER 11.3: result in line with other studies on real-case data (e.g. NFI-TNO-Test)