Speaker Recognition with Phonetic and Automatic Features using VOCALISE software

July 22nd 2013

International Association for Forensic Phonetics and Acoustics (IAFPA) Conference 2013
Florida, USA

Anil Alexander[1], Oscar Forth[1], Marianne Jessen[2] and Michael Jessen[3]

[1]Oxford Wave Research Ltd, Oxford, United Kingdom
[2]Stimmenvergleich, Wiesbaden, Germany
[3]Department of Speaker Identification and Audio Analysis, Bundeskriminalamt, Germany.

OxfordWaveResearch

VOCALISE

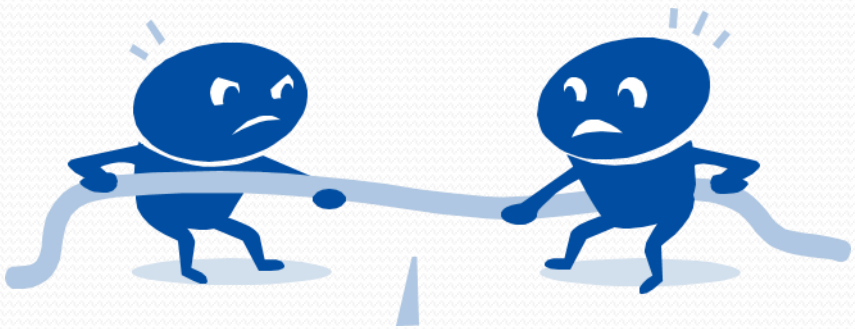
- A new system for forensic speaker recognition called
  - **VOCALISE** - Voice Comparison and Analysis of the Likelihood of Speech Evidence
- Provides the capability to perform comparisons using
  - ‘Automatic’ spectral features
  - ‘Traditional’ forensic phonetic parameters
  - ‘User’- provided features
  - Semi- or fully automatic comparisons





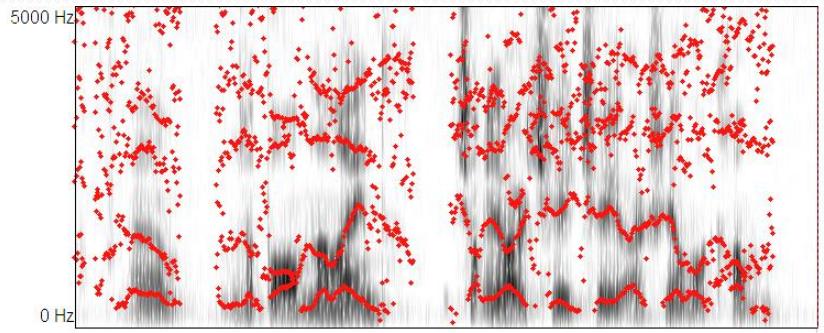
VOCALISE seeks to:

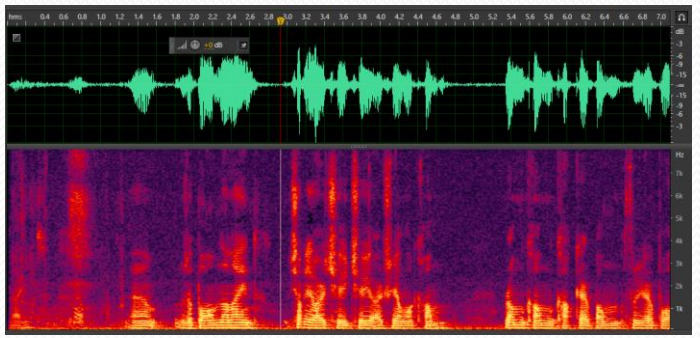
- Form a bridge between traditional forensic phonetics-based speaker recognition and forensic automatic speaker recognition
- Provide a common methodological platform for both classical automatic and phonetic speaker recognition
- Provide a coherent means of expressing the combined results.

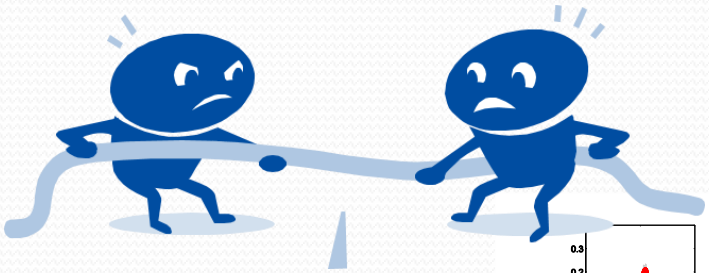


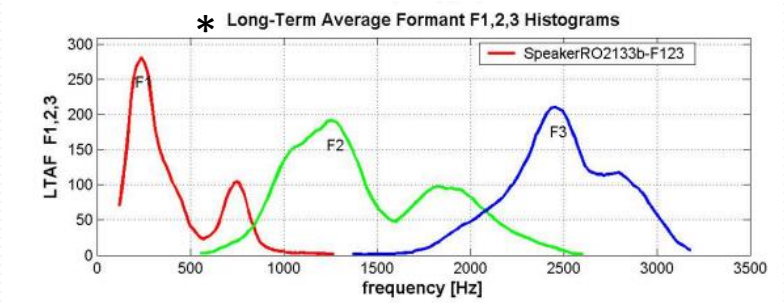


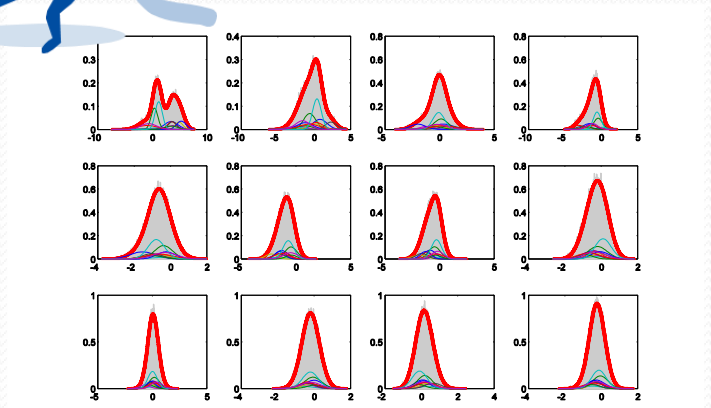
Towards a common methodological platform (1/2)







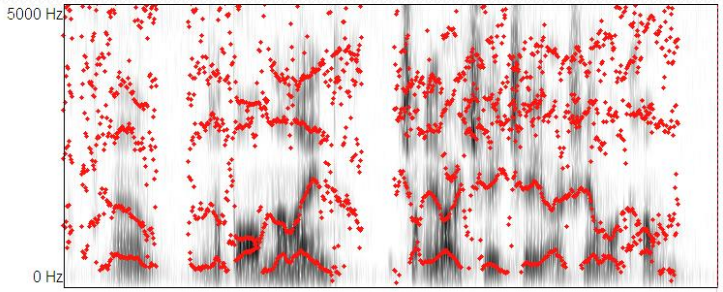




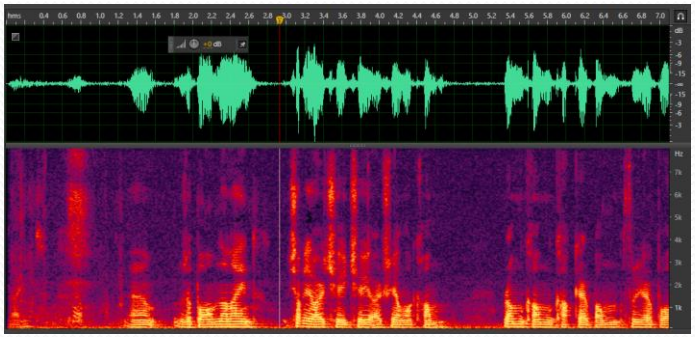
OxfordWaveResearch

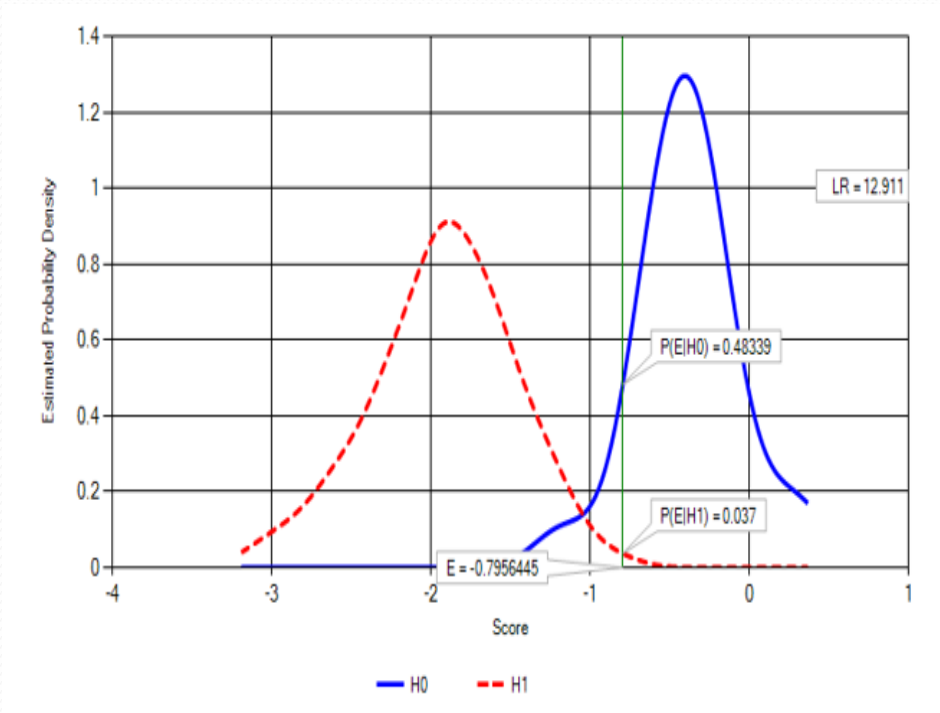
*LTF illustration from Catalina Manual

Towards a common methodological platform (2/2)





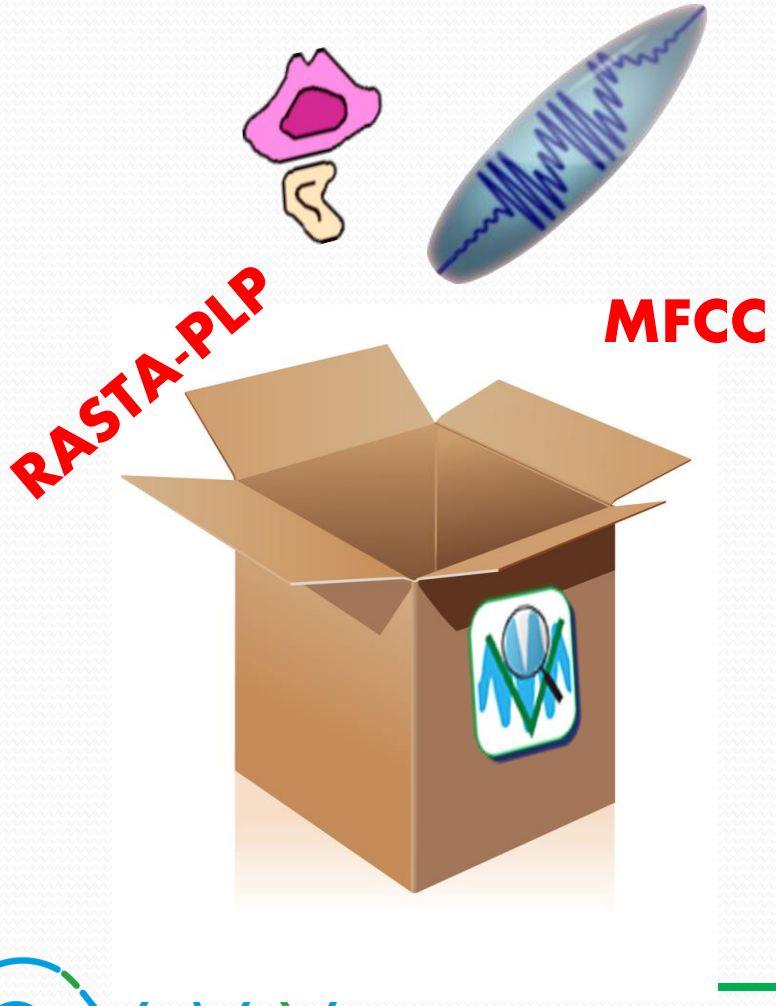




Bayesian Likelihood Ratios and
Log-likelihood Calculations made easy

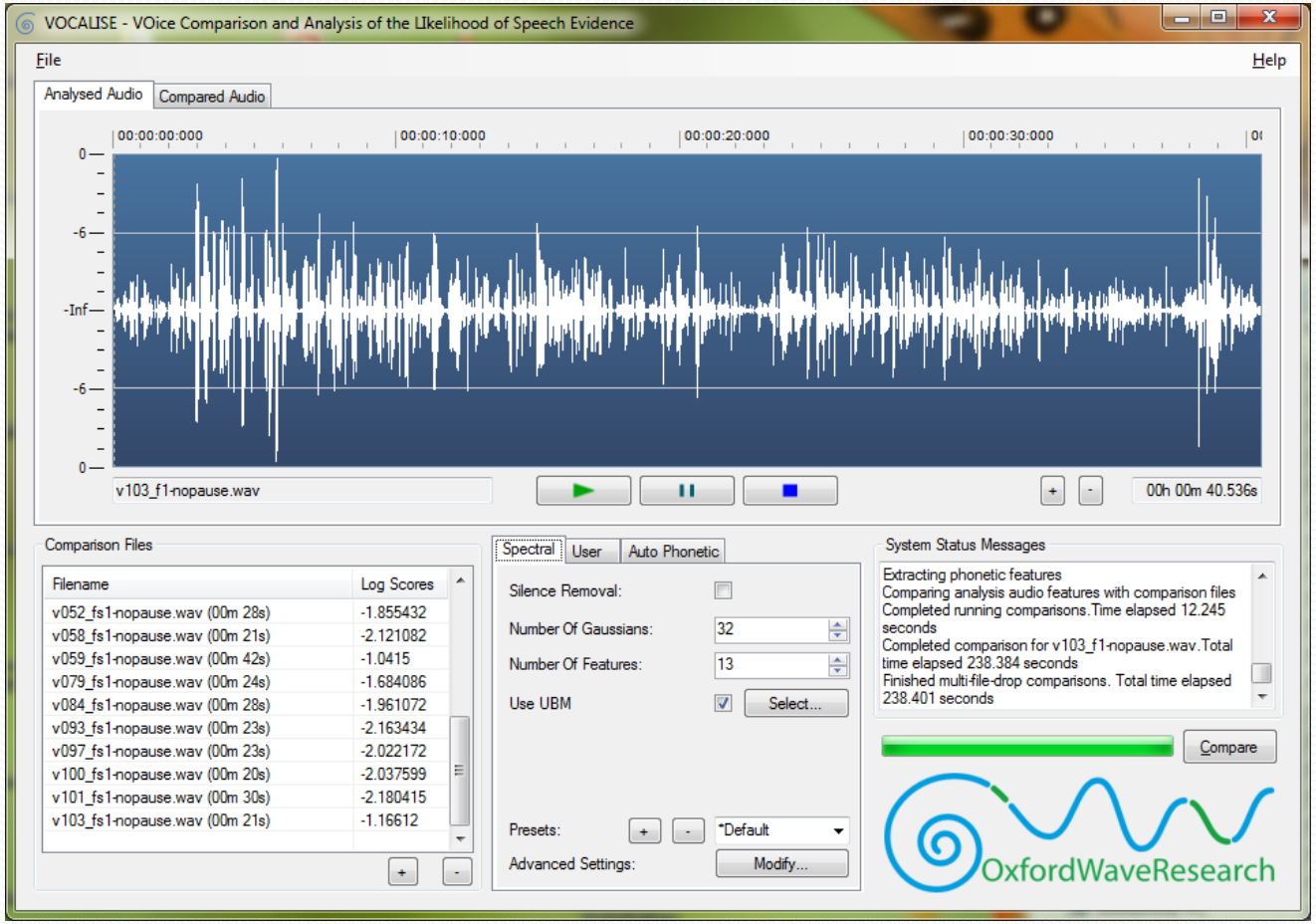
OxfordWaveResearch

Design Philosophy

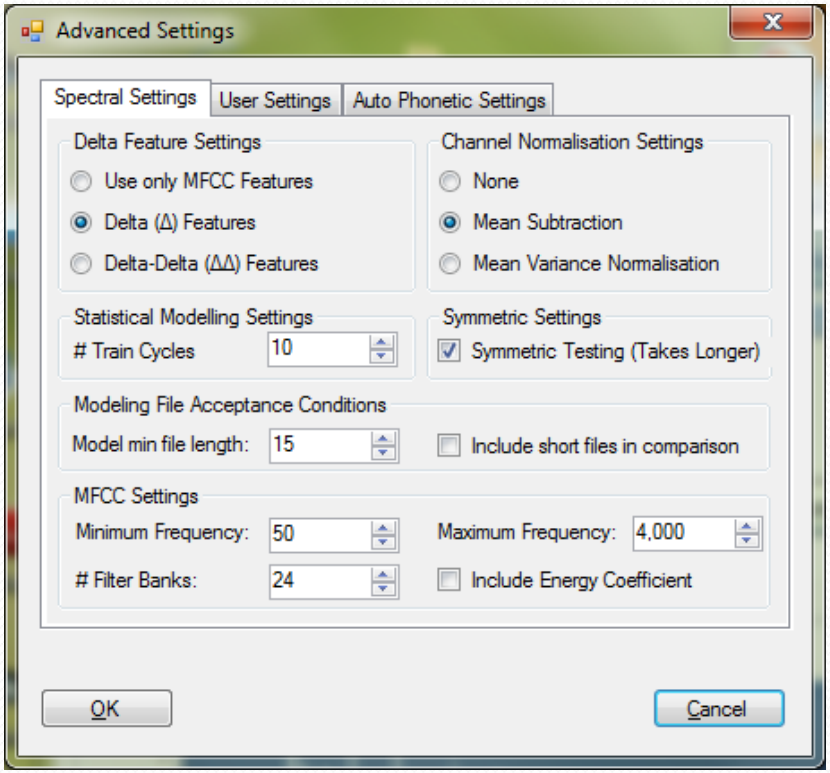


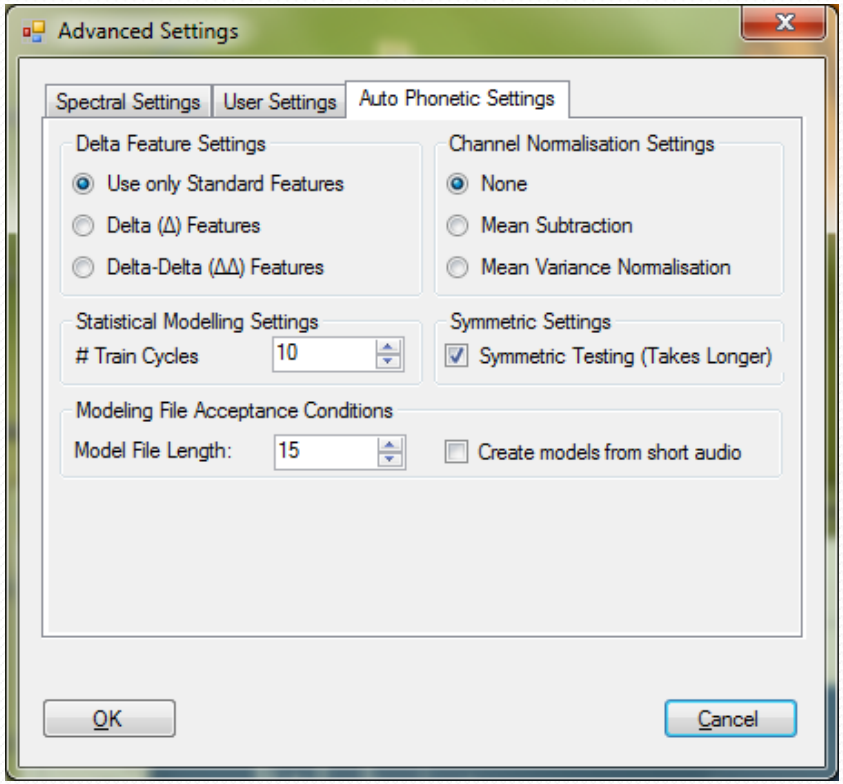
- Simple, ‘open’ box architecture
- Allows the user maximum control over their data and comparisons
- Interfaces with ‘trusted’ programs like Praat, Wavesurfer, etc
- Easy-to-use as a research tool
- With ‘great’ flexibility, greater onus on the user to use it intelligently

Voice Comparison and Analysis of the Likelihood of Speech Evidence (1/2)



Voice Comparison and Analysis of the Likelihood of Speech Evidence (2/2)





Spectral and Autophonetic modes (advanced settings screen)

OxfordWaveResearch

Why would you need VOCALISE?

The VOCALISE framework allows for:

- Benefitting from the expertise of forensic phoneticians
- Providing validation and safeguards for automatic and semi-automatic speaker recognition approaches
- Transparent and easy calculation

Benefitting from the expertise of forensic phoneticians

- Most of the forensic speaker recognition case-work is performed by forensic phoneticians who
  - Have a lot of experience and knowledge in voice comparison and an understanding of the legal requirements in their area
  - Are currently out of the loop in a fully automatic analysis
  - Want to include automatic methods, but do not have any straight-forward means of doing so
- Some of the experts would like to meet the challenge of making their speaker recognition analysis more objective using likelihood ratios and evaluating system performance
- They would like to provide an alternative opinion not entirely on 'autopilot' which includes results based on phonetic expertise.

Validation and Safeguards

- In certain cases the automatic system alone could be misleading, e.g. false identification results
  - based-on channel characteristics
  - between unusual and similar speech samples
- Automatic systems may depend on the assumptions made by the developers which may not always hold
  - It is useful to have a second opinion to confirm or deny
  - The corpora used to develop and test the systems may be different and often not representative of forensic conditions
- Adding forensic phonetic features will add a certain level of validation and safeguard
- Forensically relevant as adding phonetic information can prevent or at least reduce these errors

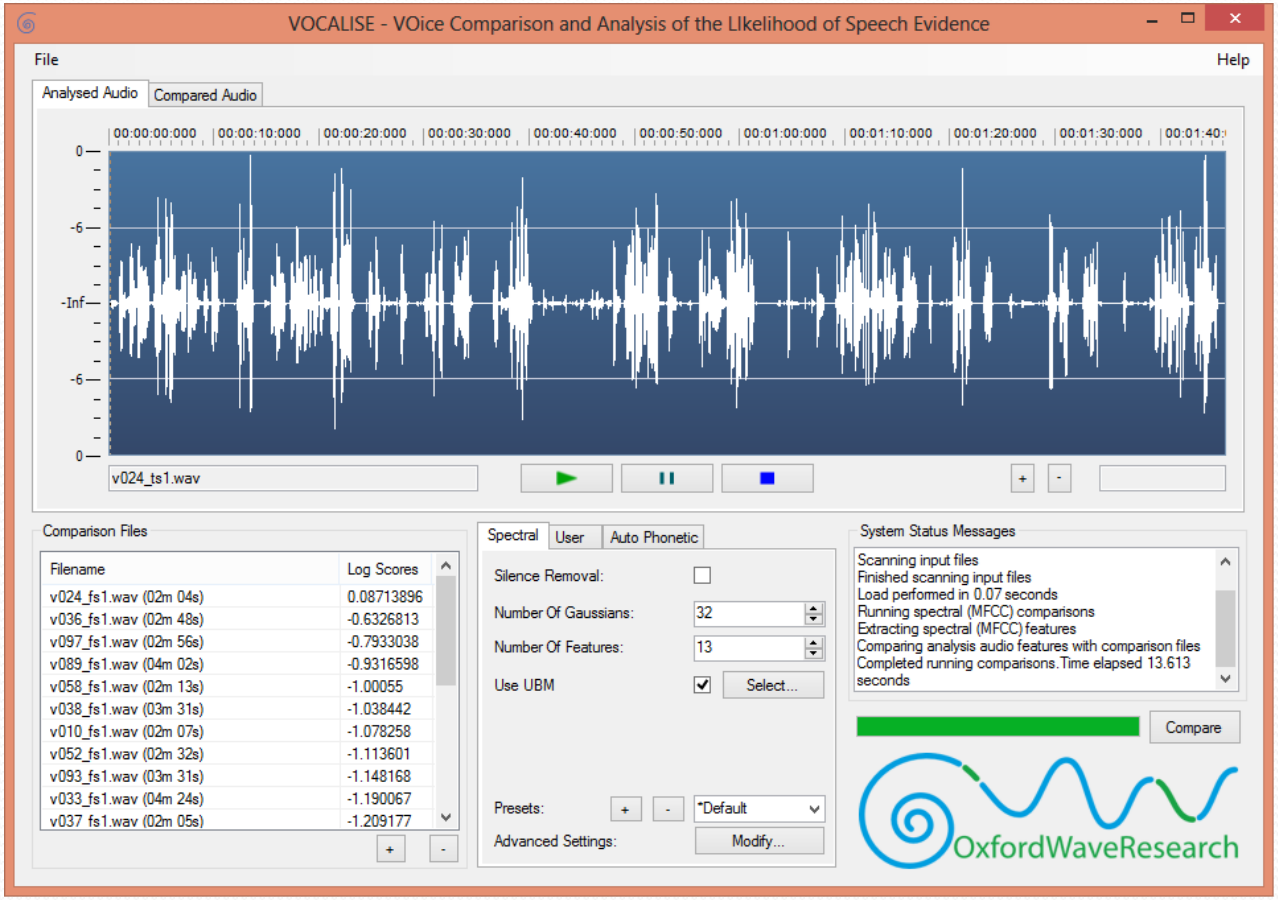
Need for transparent and easy calculations (1/2)

- Forensic labs could run automatic systems along with their 'traditional' analysis
  - if the necessary conditions of the material are met (duration, quality, etc.)
  - Use an automatic system to obtain result in the form of a LR (or score).
  - Phonetic and linguistic methods currently mostly used to obtain an opinion would benefit from being expressed in LR framework
- No reason why **everything** phonetic/linguistic stays outside the LR-framework.

Need for transparent and easy calculations (2/2)

- Currently to obtain LRs based on formants F1, F2, F3 of the vowels /i/, /a/, /u/ in a recording (York, Australia)
  - Use the Multivariate Kernel Density (MVKD) formula by Aitken & Lucy
  - Or use GMM-modeling for the distribution of F1-F2-F3 values of the suspect and a UBM
  - To obtain one LR from all three vowels you could average or use other measures
- Major impediment for including more phonetics into the LR-framework is the lack of appropriate and user-friendly tools
  - Requires dedicated knowledge of Matlab and R
  - Other phonetic measurements such as or durations of sounds, syllables or sub-syllabic constituents, fundamental frequency, etc. cannot be incorporated into the LR easily

VOCALISE - Operation Modes



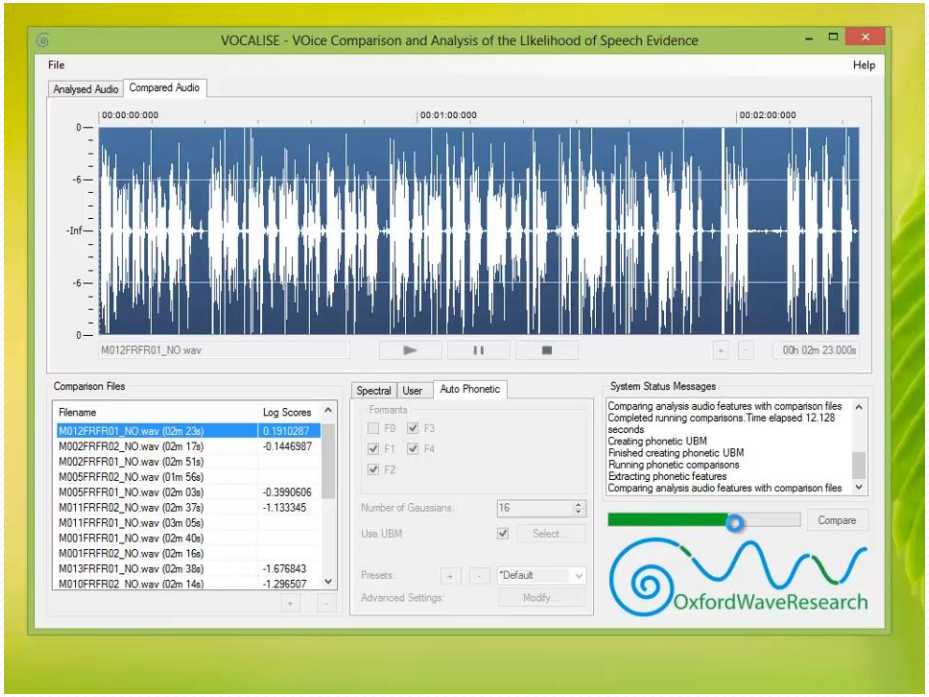
Three operation modes called ‘spectral’, ‘user’, and ‘auto phonetic’ are currently included in VOCALISE.

VOCALISE - Operation modes

- **Spectral** - automatic extraction of features most commonly used in automatic speaker and speech recognition
  - currently MFCCs
- **User** (-defined) refers to the option that lets the users use **any features** as their own stream(s) of values which can be manually measured, labelled, or corrected
  - e.g. formant frequencies, fundamental frequency, or durations of sounds, syllables or sub-syllabic constituents (units relevant to tempo and rhythm), or even auditory features.
- **Auto-phonetic** - refers to the automatic (unsupervised) extraction of phonetic features interfacing with Praat
  - currently formants F1 to F4 selected in any combination .

Software Features

- Windows-based software (Win7/Win8)
- Easy to listen and compare waveform viewer and playback controls
- Fast modelling and comparison using multi-threading
- Verbose system status messages
- All scores are calculated directly from audio files (no caching of files, except the UBM data)
- Currently 32 bit system (64 bit release version coming shortly)



OxfordWaveResearch

Algorithms provided

- VOCALISE allows for phonetic, spectral or user-defined features (interchangeably)
  - Normalization and extraction of dynamic information
  - Gaussian Mixture Modeling (GMM)
  - Creation of statistical models for populations using universal background models (UBMs).
- Incorporates ongoing areas of phonetic research:
  - Long-term formant information (Nolan & Grigoras 2005)
  - Formant dynamics (McDougall 2005)

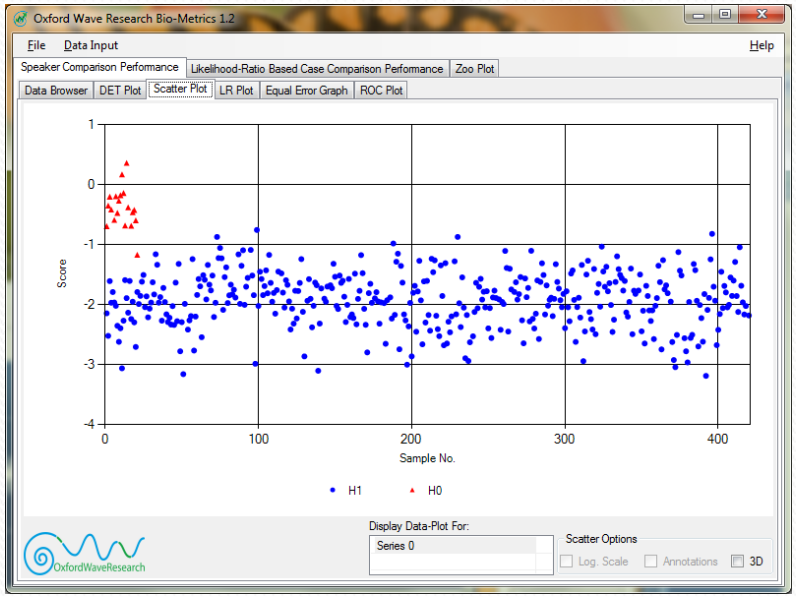
Demonstration

VOCALISE + BioMetrics as a Research tool

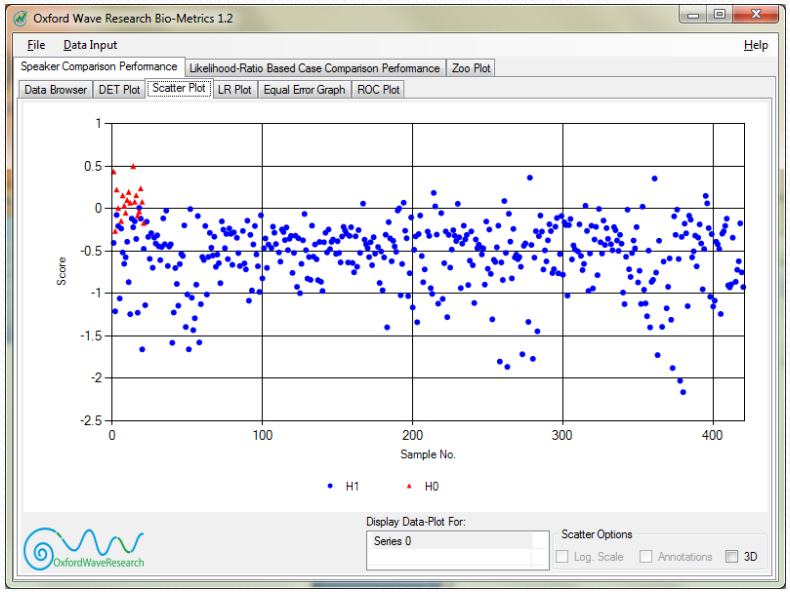
- VOCALISE allows for easy database comparisons using drag and drop
- Results provided in Excel readable table from which data can either be exported into your own statistical packages or into Oxford Wave Research's Bio-Metrics software.

	A	B	C	D	E	F	G	H
1		v007_fs1-r	v010_fs1-r	v014_fs1-r	v016_fs1-r	v018_fs1-r	v024_fs1-r	v033_fs1-r
2	v007_f1-nopause.wav	-0.68963	-2.14451	-2.52062	-1.60631	-1.96737	-1.78921	-1.96383
3	v010_f1-nopause.wav	-1.78876	-0.34617	-1.98914	-1.85355	-1.61697	-1.50373	-2.04166
4	v014_f1-nopause.wav	-2.291	-2.21757	-0.19732	-2.33149	-2.02399	-2.33804	-1.63257
5	v016_f1-nopause.wav	-1.57172	-1.69392	-2.54283	-0.41096	-1.50781	-1.58474	-1.91204
6	v018_f1-nopause.wav	-1.84382	-1.66616	-1.82773	-1.74988	-0.41765	-1.88168	-1.90524
7	v024_f1-nopause.wav	-1.45017	-1.5714	-1.84064	-1.68997	-1.42739	-0.58545	-1.80966
8	v033_f1-nopause.wav	-2.41442	-2.0675	-2.31681	-1.76713	-2.23074	-1.77208	-0.19001
9	v036_f1-nopause.wav	-1.72879	-1.75771	-1.89966	-1.95145	-1.69753	-1.68192	-1.67593
10	v037_f1-nopause.wav	-2.16273	-2.22291	-1.48212	-2.32782	-1.90332	-1.76887	-1.17248
11	v038_f1-nopause.wav	-1.98192	-1.97435	-2.6535	-1.92587	-1.8964	-1.87775	-2.23284
12	v047_f1-nopause.wav	-1.80103	-1.68507	-2.45179	-1.77869	-1.26774	-1.92723	-2.66163
13	v052_f1-mopause.wav	-2.67726	-1.30685	-2.64809	-2.28421	-2.46124	-2.19533	-2.17071
14	v058_f1-nopause.wav	-2.12128	-1.63481	-2.05904	-1.70212	-1.56814	-1.91536	-1.78395
15	v059_f1-nopause.wav	-1.10287	-1.38704	-2.44847	-1.39997	-1.74655	-1.73201	-1.78832
16	v079_f1-nopause.wav	-2.15057	-1.59572	-2.57283	-1.62138	-1.30857	-1.50514	-2.16316
17	v084_f1-nopause.wav	-1.77576	-2.04392	-2.27668	-1.48072	-1.42761	-1.91999	-2.62147
18	v093_f1-nopause.wav	-1.80245	-1.65457	-2.033	-1.03101	-1.44695	-1.70133	-1.36735
19	v097_f1-nopause.wav	-2.1996	-1.95469	-1.39777	-2.49283	-1.76707	-1.8199	-1.82903
20	v100_f1-nopause.wav	-1.35004	-1.62408	-2.74601	-1.25683	-1.74413	-1.67751	-1.80825
21	v101_f1-nopause.wav	-1.95505	-2.55599	-2.50331	-1.32165	-1.42351	-1.93445	-2.69674
22	v103_f1-nopause.wav	-2.15606	-1.45022	-2.05742	-1.68718	-1.79494	-2.04457	-2.00282

Scatter-plot (Spectral vs. Auto-phonetic)



Spectral (MFCC)



Auto-Phonetic (F1,F2,F3)

Pool 2010: 21 male adult speakers of the West-Central regional variety of German. Ref. Jessen et al IAFPA 2013)

OxfordWaveResearch

Some early results (1/2)

- Subsets of good quality databases like Pool 2010 (Jessen et al. 2005) (22 speakers) and DyVIS (Nolan et al. 2009) (17 speakers) have been very promising, achieving close to complete speaker separation
  - 0.1% and 0.374% equal error rate (EER) respectively, using spectral methods.
  - The auto-phonetic mode yielded approximately 8.9% EER for Pool 2010 and 5.88% for DyVIS.
  - A more challenging variation of this test is presented by Jessen et al, IAFPA 2013 (including Lombard speech)

Some early results (2/2)

- Real case data - 22 male speakers of German from authentic anonymized cases.
  - between 20 and 60s of speech, initial testing obtains an EER of 12.6% in the spectral mode.
  - Long-term formant analysis (6 Gaussians, MVN, Symmetric Testing and delta features EER was at 18.1%.
- Considering that this is about fully natural, quality reduced speech, unsupervised formant tracking, and no other information than F1 to F3, this is a promising result.

Conclusions

- Whereas features pertaining to the spectral envelope such as MFCCs are powerful, they are also very sensitive to channel effects and recording quality,
  - are mostly data-driven and less directly connected to the theory of speech production (Rose 2002).
- VOCALISE makes it possible to
  - apply classical automatic speaker recognition transparently
  - analyse the speaker-discriminative information of acoustic phonetic data such as formant frequencies, fundamental frequency or sound durations.
- Processing phonetic data will be in many ways complementary and will offer insights into the voice comparison analysis that the classical automatic methods cannot.

Future work

- Extending the user-defined feature set to auditory and linguistic features – obtaining a numerical likelihood ratio
- Handling mismatched conditions in a combined framework for phonetic and automatic speaker recognition
- Incorporating the ‘scoring method’ of the Bayesian framework for calculating likelihood ratios (Drygajlo et. al)

Special Thanks





- Department of Linguistics, University of Cambridge, for kindly providing data from the DyVIS database
- Bundeskriminalamt for kindly providing the Pool 2010 corpus used in the experiments

References

- A. Alexander, '*Forensic Automatic Speaker Recognition using Bayesian Interpretation and Statistical Compensation for Mismatched Conditions*'. PhD thesis, Swiss Federal Institute of Technology, Lausanne, 2005
- Jessen, M., O. Köster & S. Gfroerer (2005). Influence of vocal effort on average and variability of fundamental frequency. Intern. J. of Speech, Language and the Law, 12, 174–213.
- McDougall, K. (2005). The Role of formant dynamics in determining speaker identity. Ph.D. Dissertation, University of Cambridge.
- Nolan, F. & C. Grigoras (2005). A case for formant analysis in forensic speaker identification. Intern. J. of Speech, Language and the Law, 12, 143–173.
- Nolan, F., K. McDougall, G. de Jong & T. Hudson (2009). The DyViS database: style-controlled recordings of 100 homogeneous speakers for forensic phonetic research. Intern. J. of Speech, Language and the Law 16: 31–57.
- Rose, P. (2002). Forensic speaker identification. London: Taylor & Francis.