# Recent Progress in Prosodic Speaker Verification

*Marcel Kockmann*[1], *Luciana Ferrer*[2], *Lukáš Burget*[1], *Elizabeth Shriberg*[2], *and Jan "Honza" Černocký*[1]

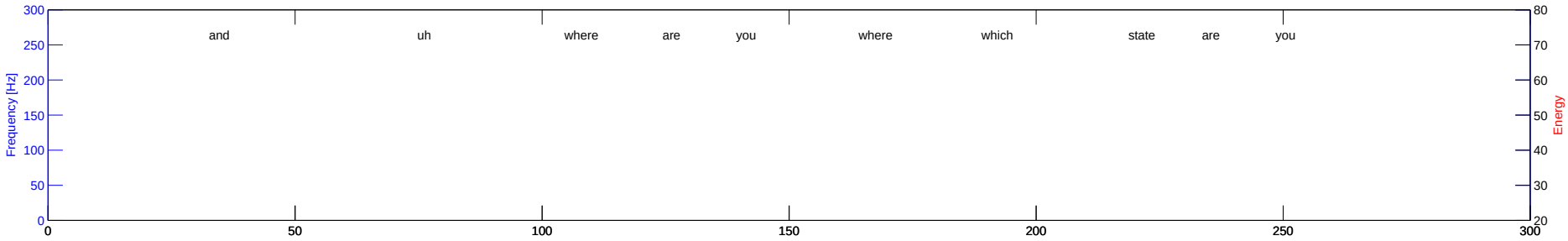
[1]Brno University of Technology, Speech@FIT, Czech Republic
[2]Speech Technology and Research Laboratory, SRI International, Menlo Park, CA, USA

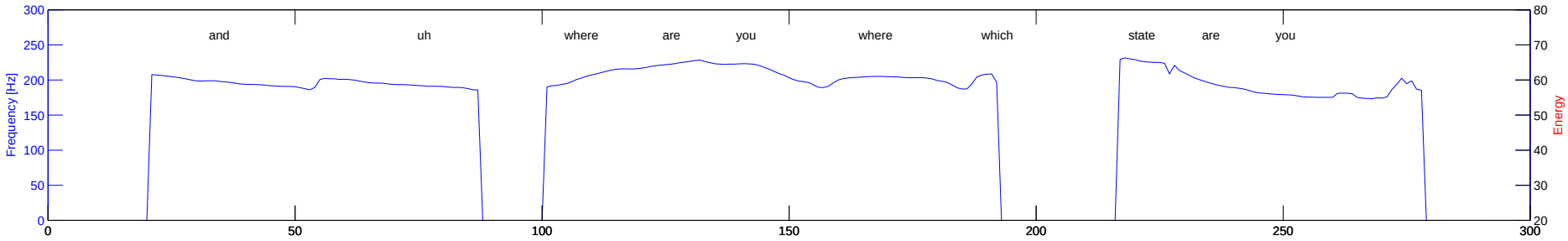
May 26, 2011

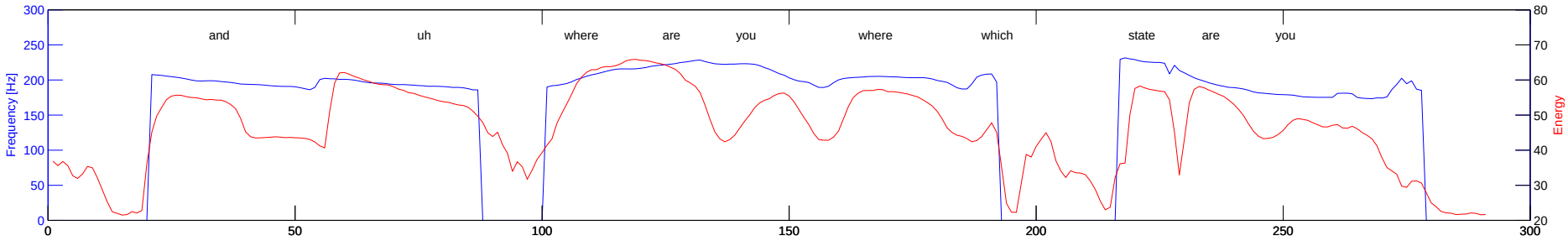
Follow-up work on Subspace Multinomial Models (SMM) to model SNERFs

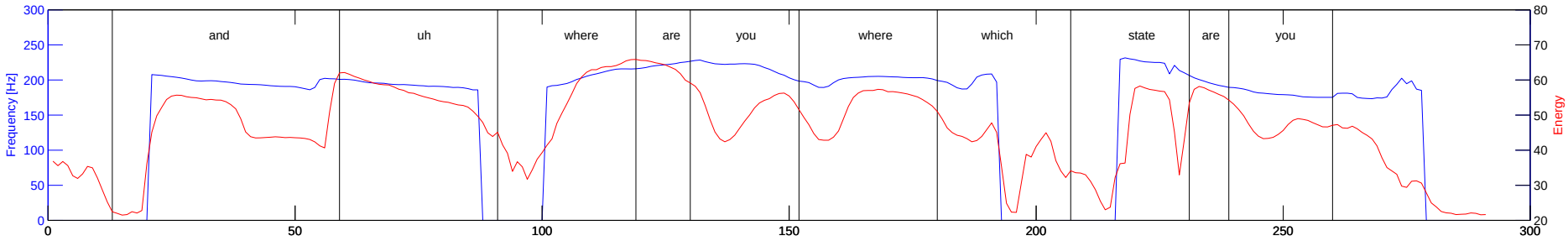
Claims:

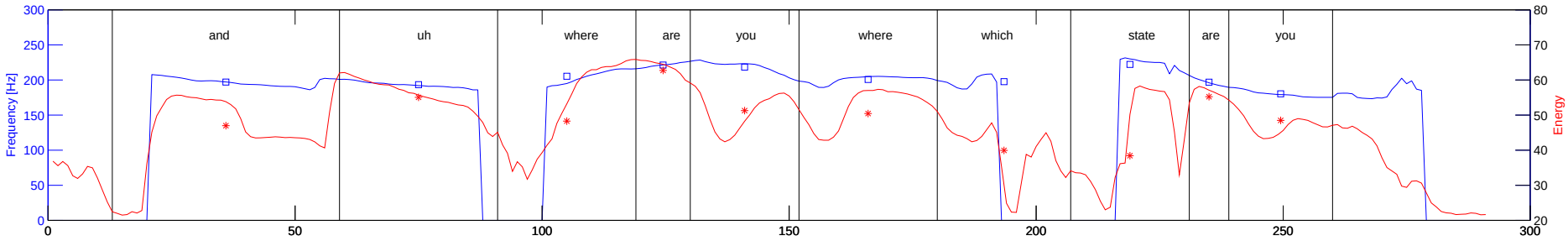
- Visualize complex system building process by a toy-example
- Introduce modeling of prosodic iVectors by Probabilistic Linear Discriminat Analysis (PLDA)
- Compare to State-of-the-Art prosodic speaker verification systems
- Combination with cepstral baseline system

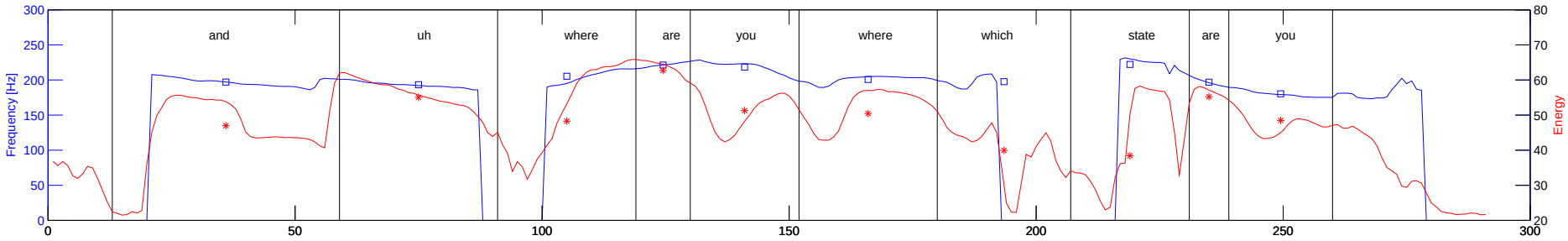

and
uh
where
are
you
where
which
state
are
you
Frequency [Hz]
Energy
0
50
100
150
200
250
300
20
30
40
50
60
70
80

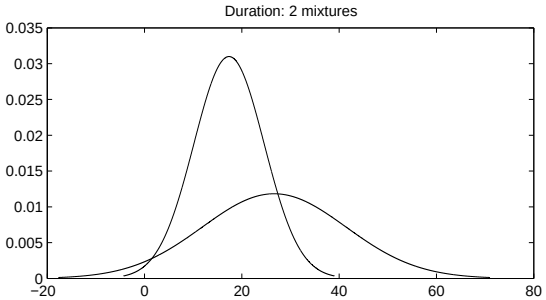

and
uh
where
are
you
where
which
state
are
you
Frequency [Hz]
Energy
0
50
100
150
200
250
300
20
30
40
50
60
70
80

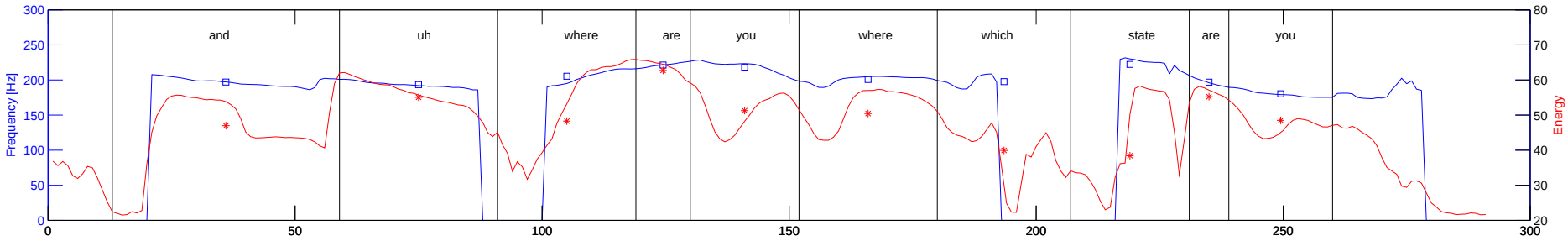

and
uh
where
are
you
where
which
state
are
you
Frequency [Hz]
Energy
0
50
100
150
200
250
300
20
30
40
50
60
70
80

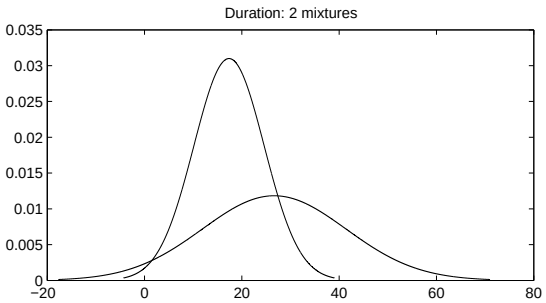

and
uh
where
are
you
where
which
state
are
you
Frequency [Hz]
Energy
0
50
100
150
200
250
300
20
30
40
50
60
70
80

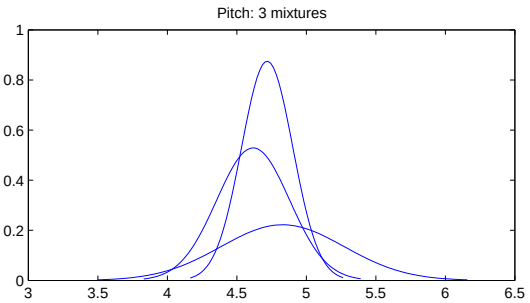

and
uh
where
are
you
where
which
state
are
you
Frequency [Hz]
Energy
0
50
100
150
200
250
300
20
30
40
50
60
70
80

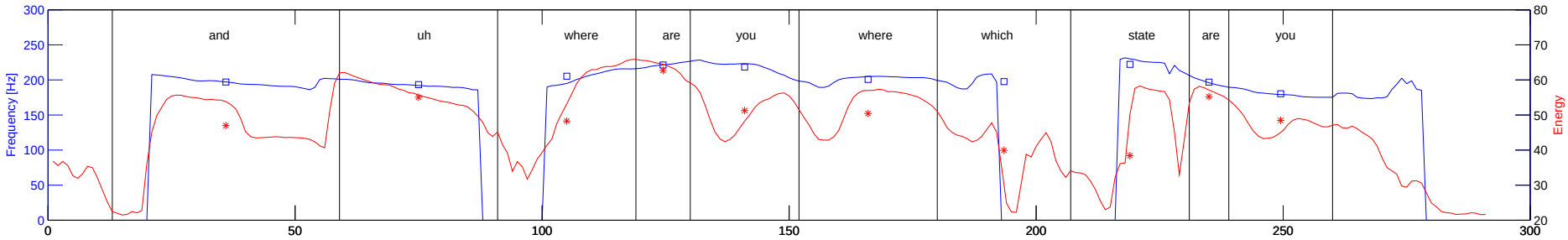

and
uh
where
are
you
where
which
state
are
you
Frequency [Hz]
Energy

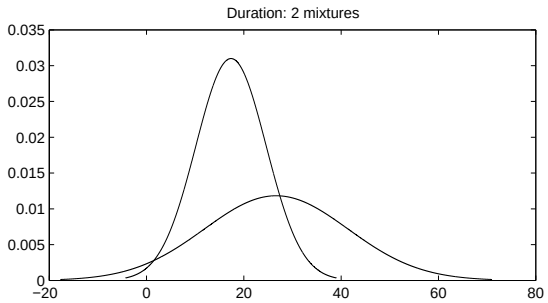

Duration: 2 mixtures

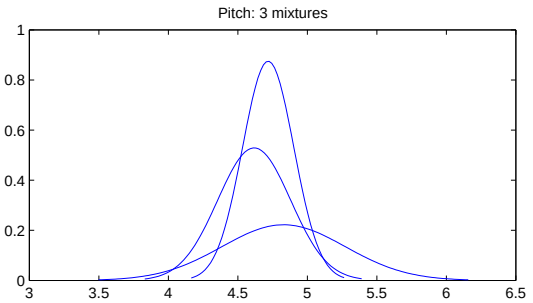

and
uh
where
are
you
where
which
state
are
you
Frequency [Hz]
Energy

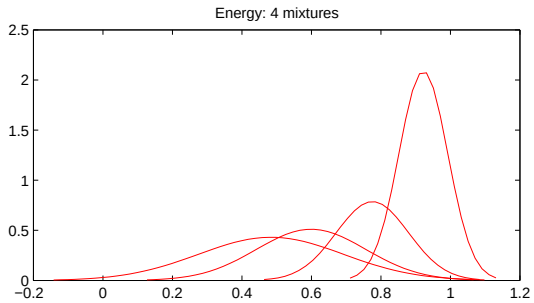

Duration: 2 mixtures

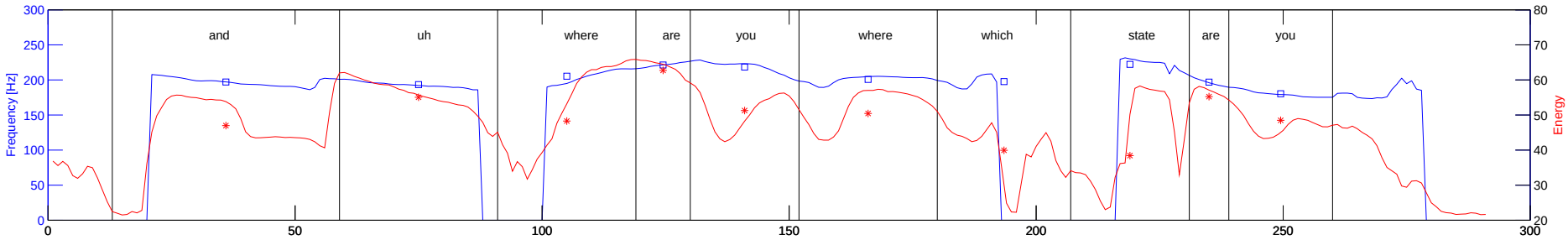

Pitch: 3 mixtures

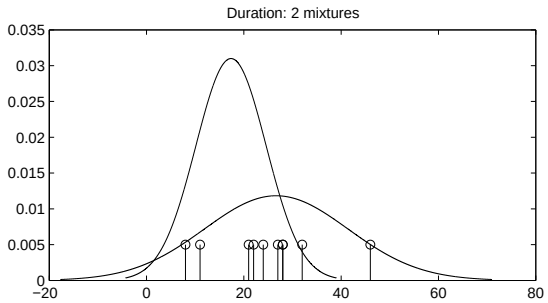

and
uh
where
are
you
where
which
state
are
you
Frequency [Hz]
Energy

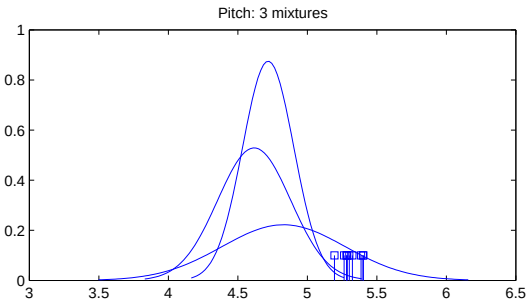

Duration: 2 mixtures

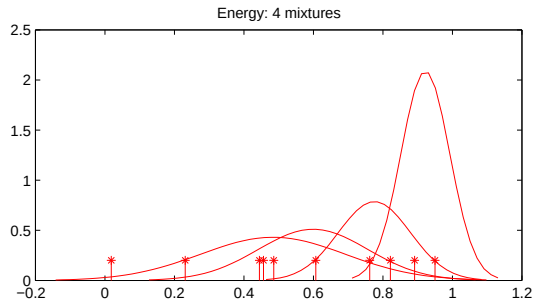

Pitch: 3 mixtures

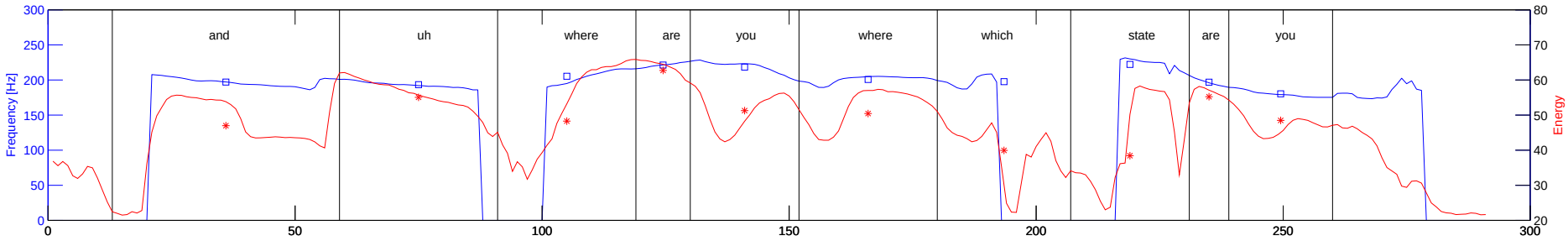

Energy: 4 mixtures

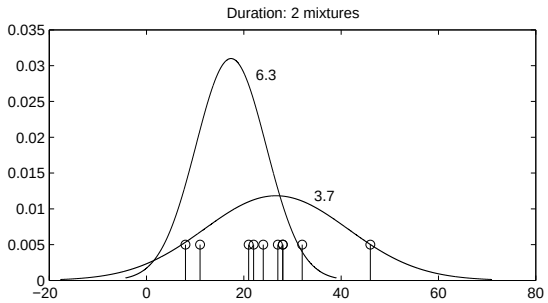

and
uh
where
are
you
where
which
state
are
you
Frequency [Hz]
Energy

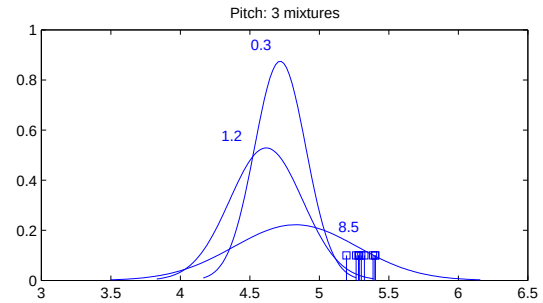

Duration: 2 mixtures

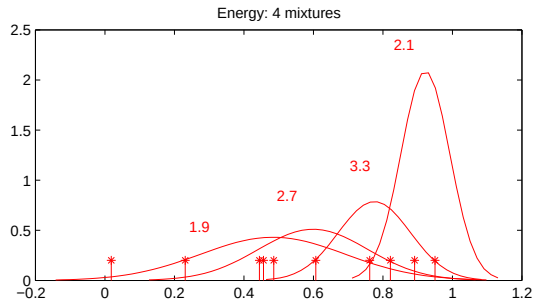

Pitch: 3 mixtures

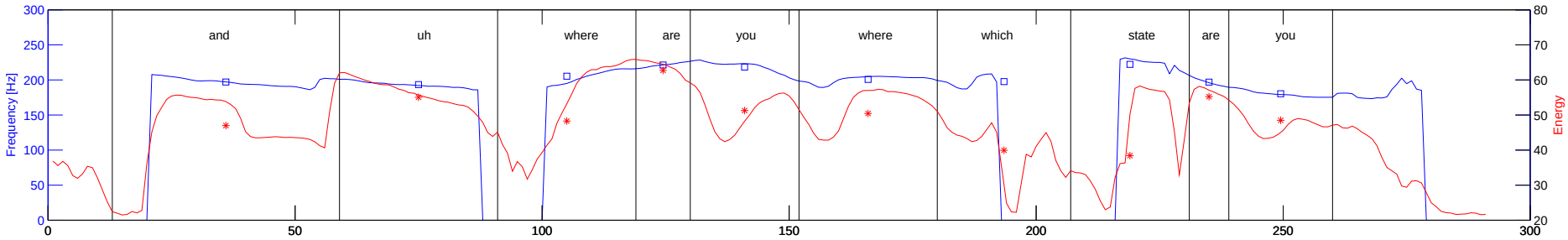

Energy: 4 mixtures

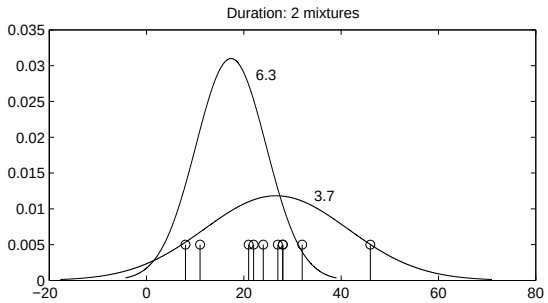

and
uh
where
are
you
where
which
state
are
you
Frequency [Hz]
Energy

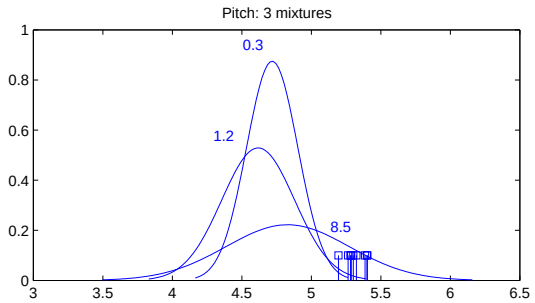

Duration: 2 mixtures
6.3
3.7

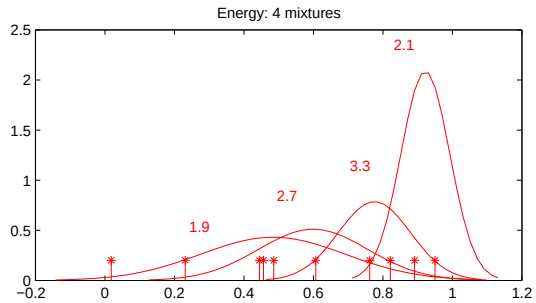

Pitch: 3 mixtures
0.3
1.2
8.5

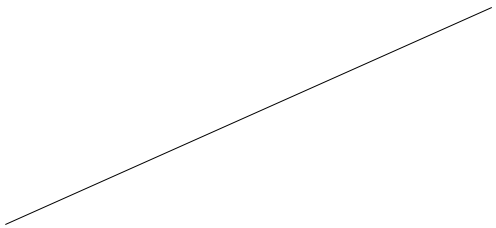

Energy: 4 mixtures
2.1
3.3
2.7
1.9

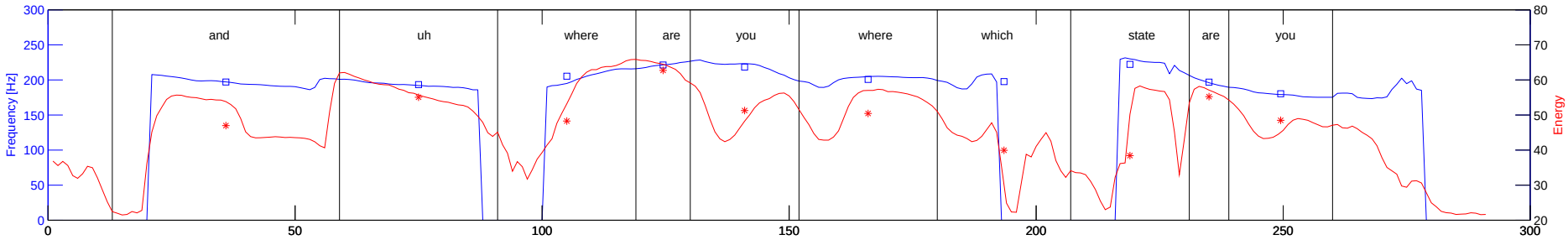

and
uh
where
are
you
where
which
state
are
you
Frequency [Hz]
Energy

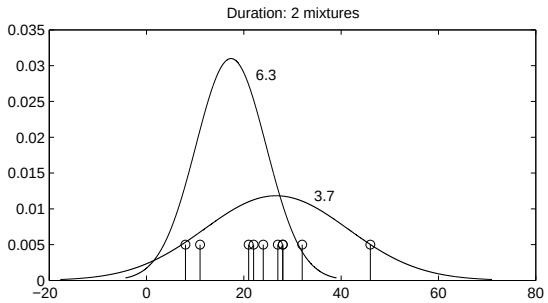

Duration: 2 mixtures
6.3
3.7

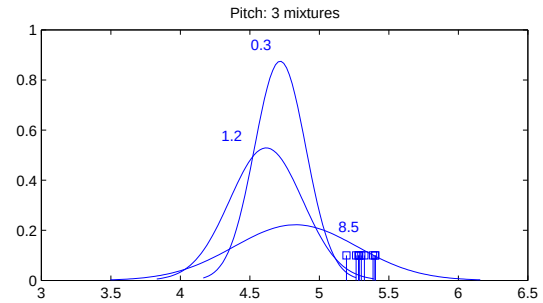

Pitch: 3 mixtures
0.3
1.2
8.5

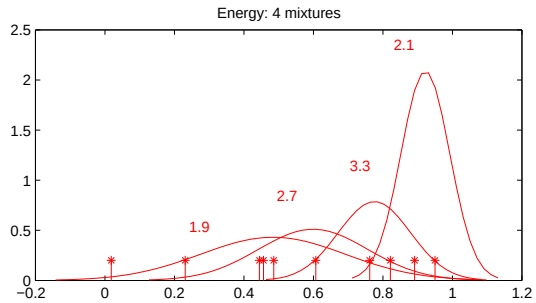

Energy: 4 mixtures
2.1
3.3
2.7
1.9

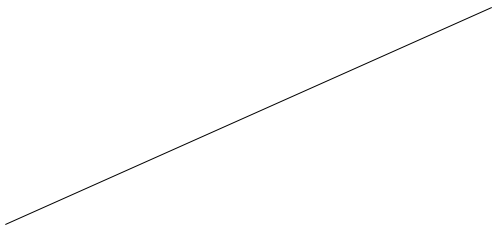


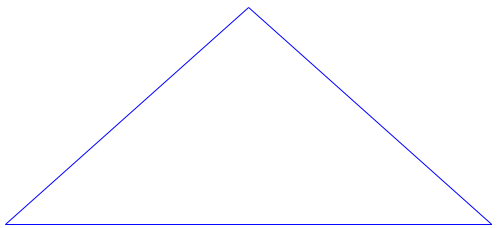

Frequency [Hz]
Energy
and
uh
where
are
you
where
which
state
are
you

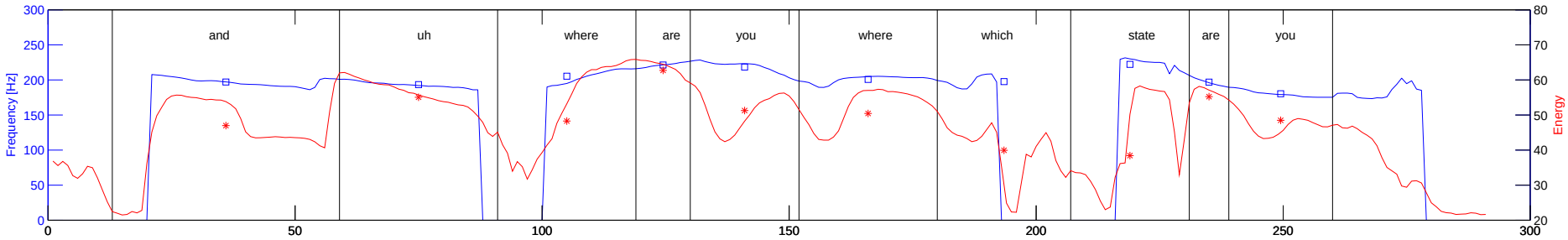

Duration: 2 mixtures
6.3
3.7

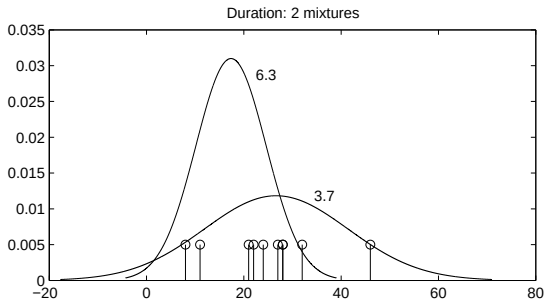

Pitch: 3 mixtures
0.3
1.2
8.5

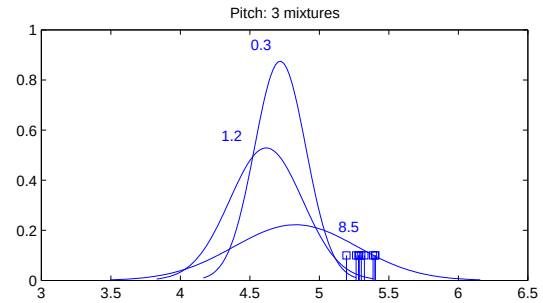

Energy: 4 mixtures
2.1
3.3
2.7
1.9

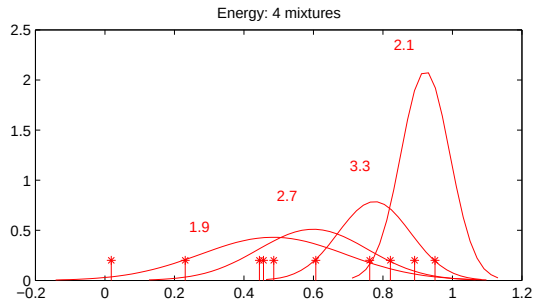


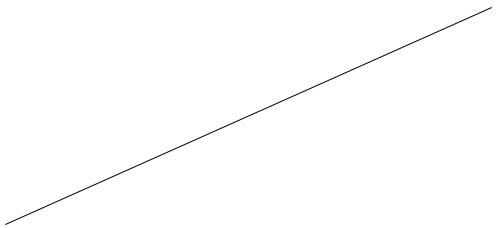


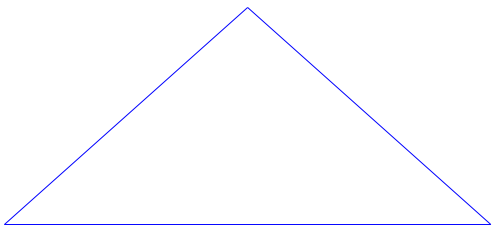

Frequency [Hz]
Energy
and
uh
where
are
you
where
which
state
are
you

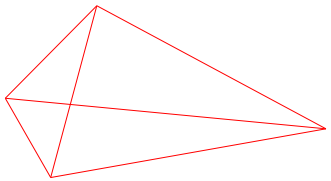

Duration: 2 mixtures
6.3
3.7

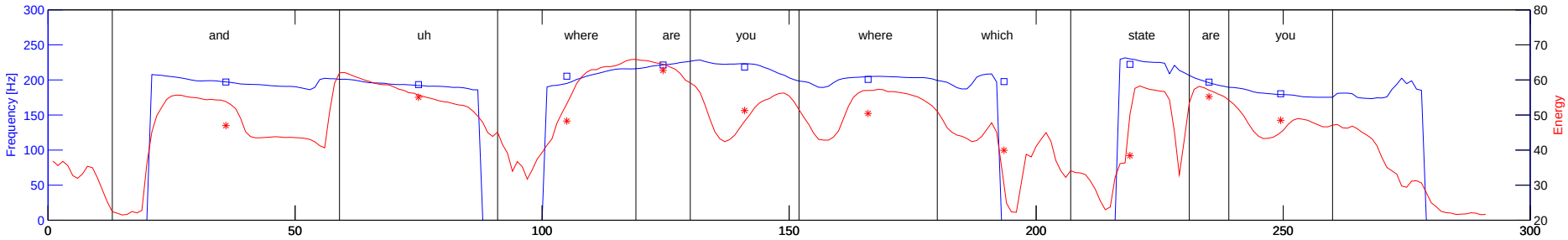

Pitch: 3 mixtures
0.3
1.2
8.5

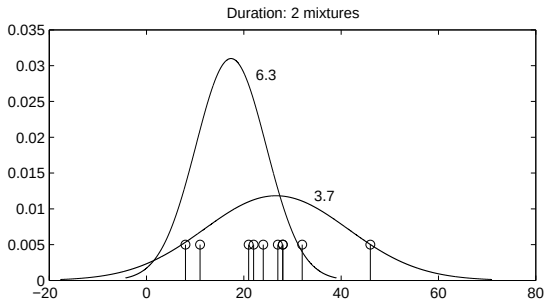

Energy: 4 mixtures
2.1
3.3
2.7
1.9

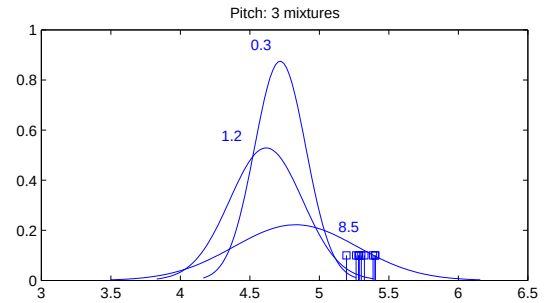


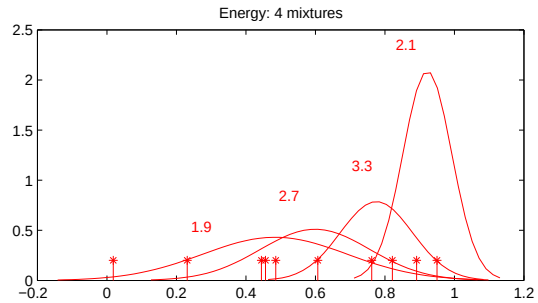


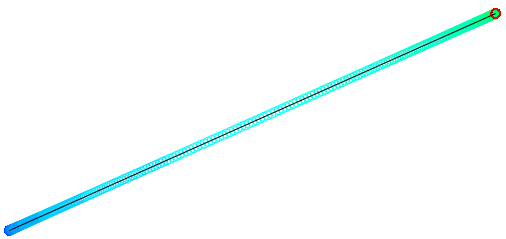


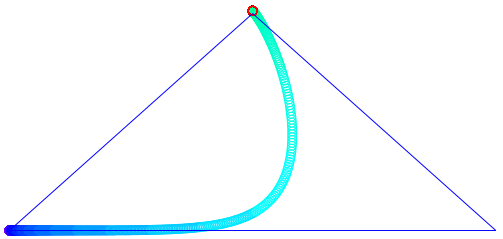

and
uh
where
are
you
where
which
state
are
you
Frequency [Hz]
Energy

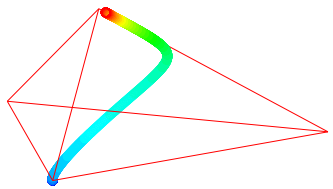

Duration: 2 mixtures
6.3
3.7

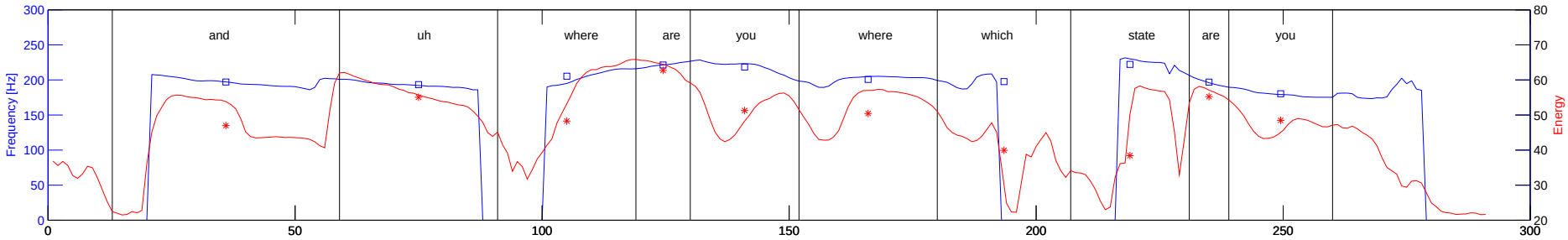

Pitch: 3 mixtures
0.3
1.2
8.5

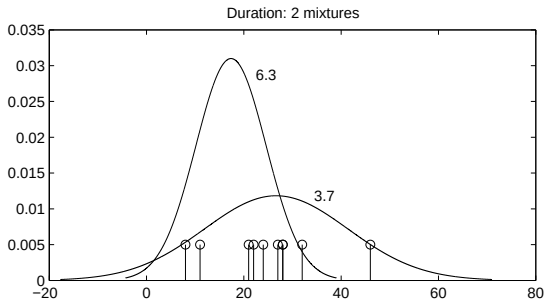

Energy: 4 mixtures
2.1
3.3
2.7
1.9

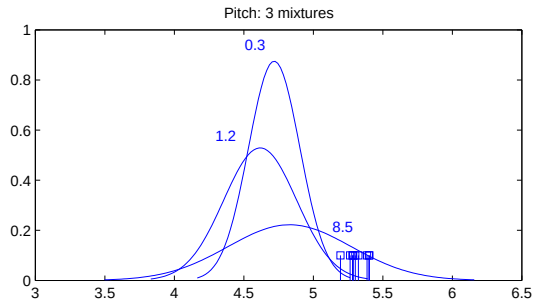


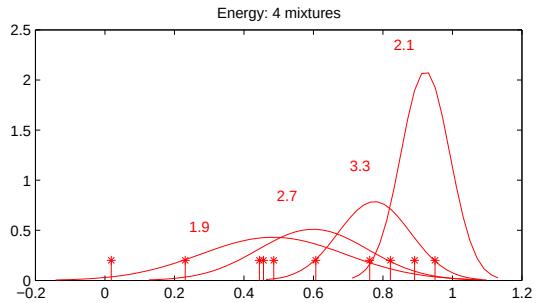


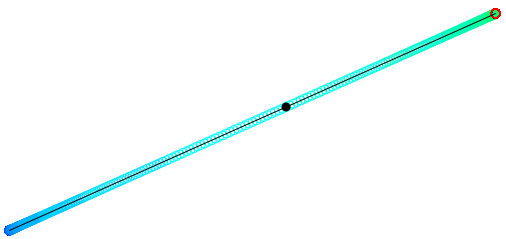


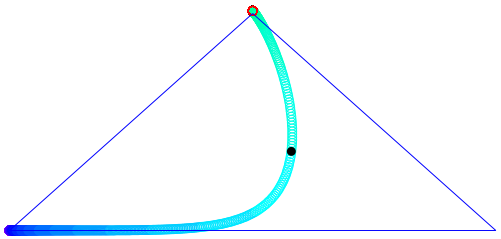

and
uh
where
are
you
where
which
state
are
you
Frequency [Hz]
Energy

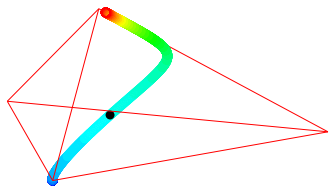

Duration: 2 mixtures
6.3
3.7

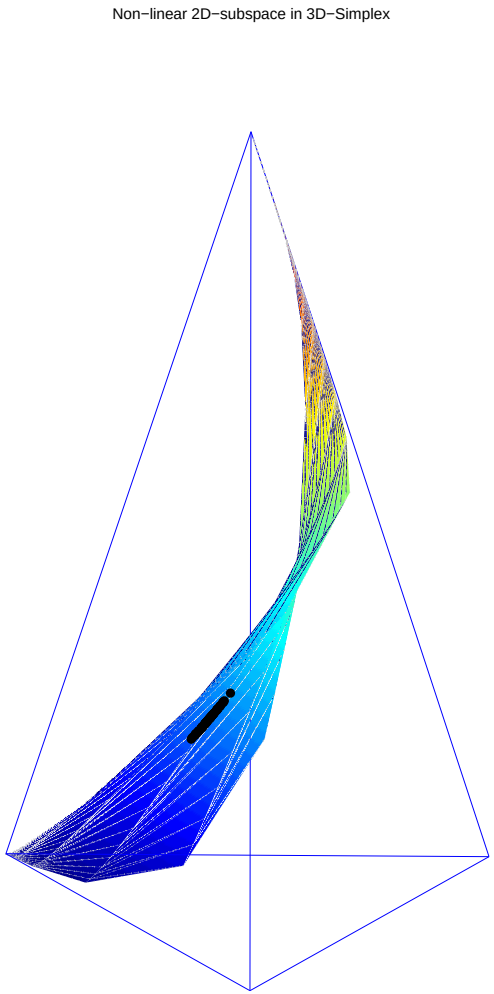

Pitch: 3 mixtures
0.3
1.2
8.5

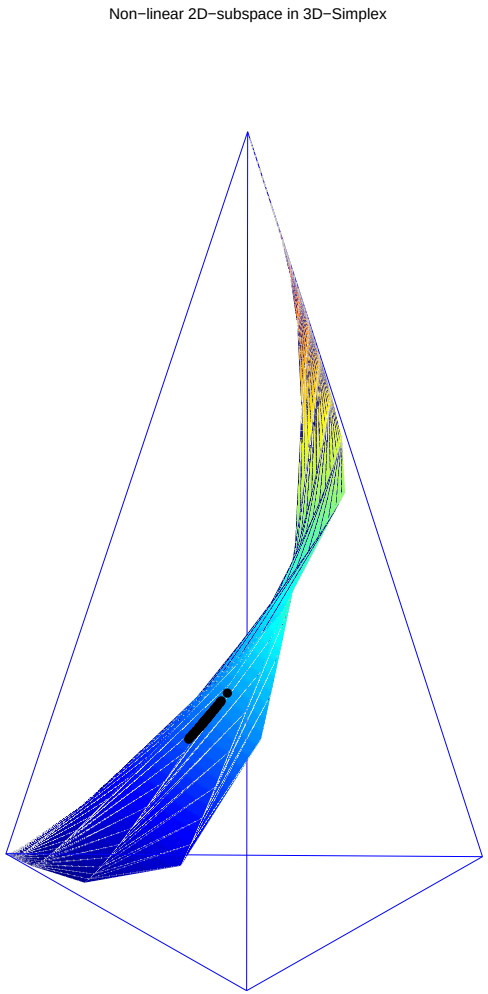

Energy: 4 mixtures
2.1
3.3
2.7
1.9

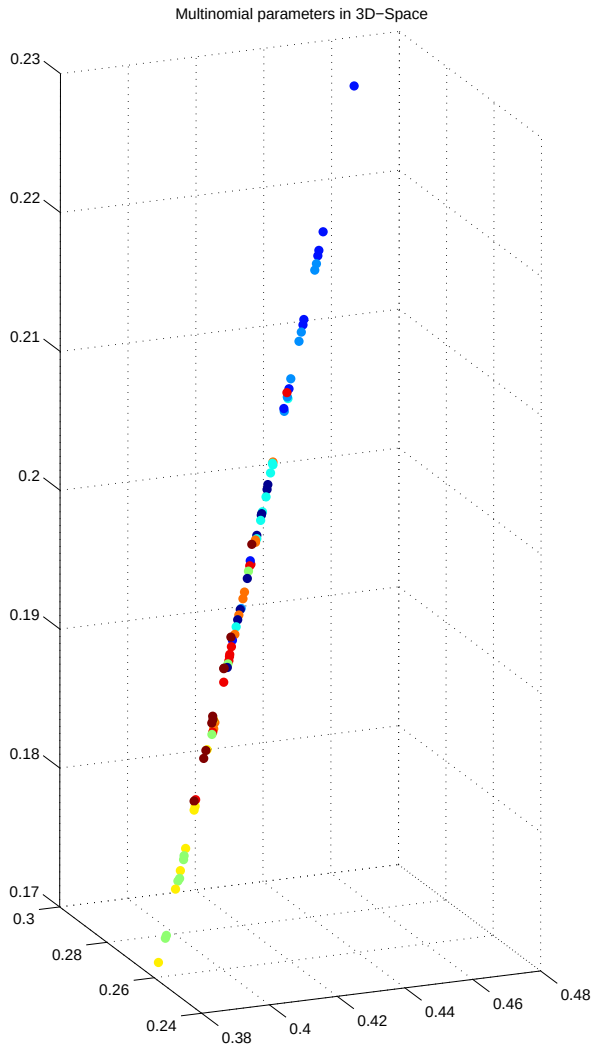


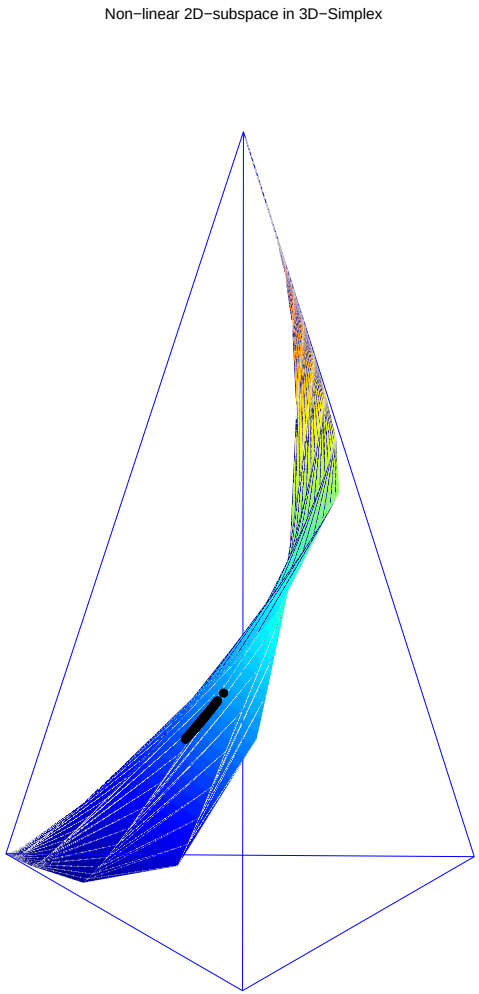


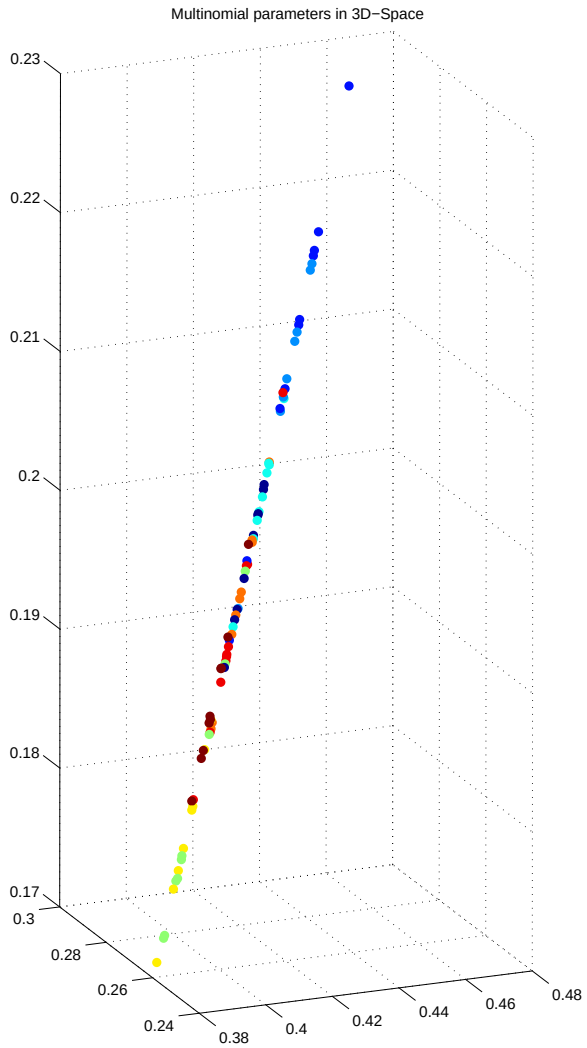


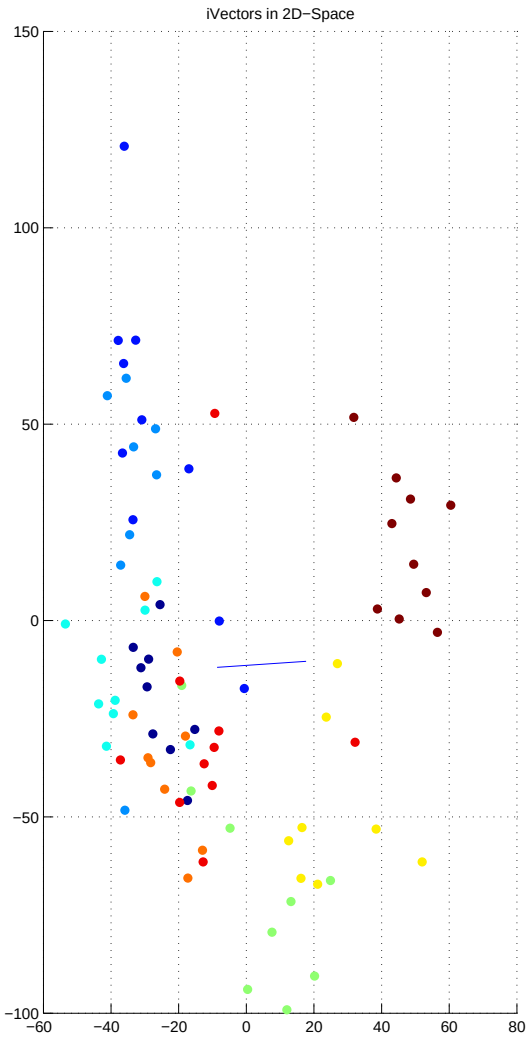

Non−linear 2D−subspace in 3D−Simplex

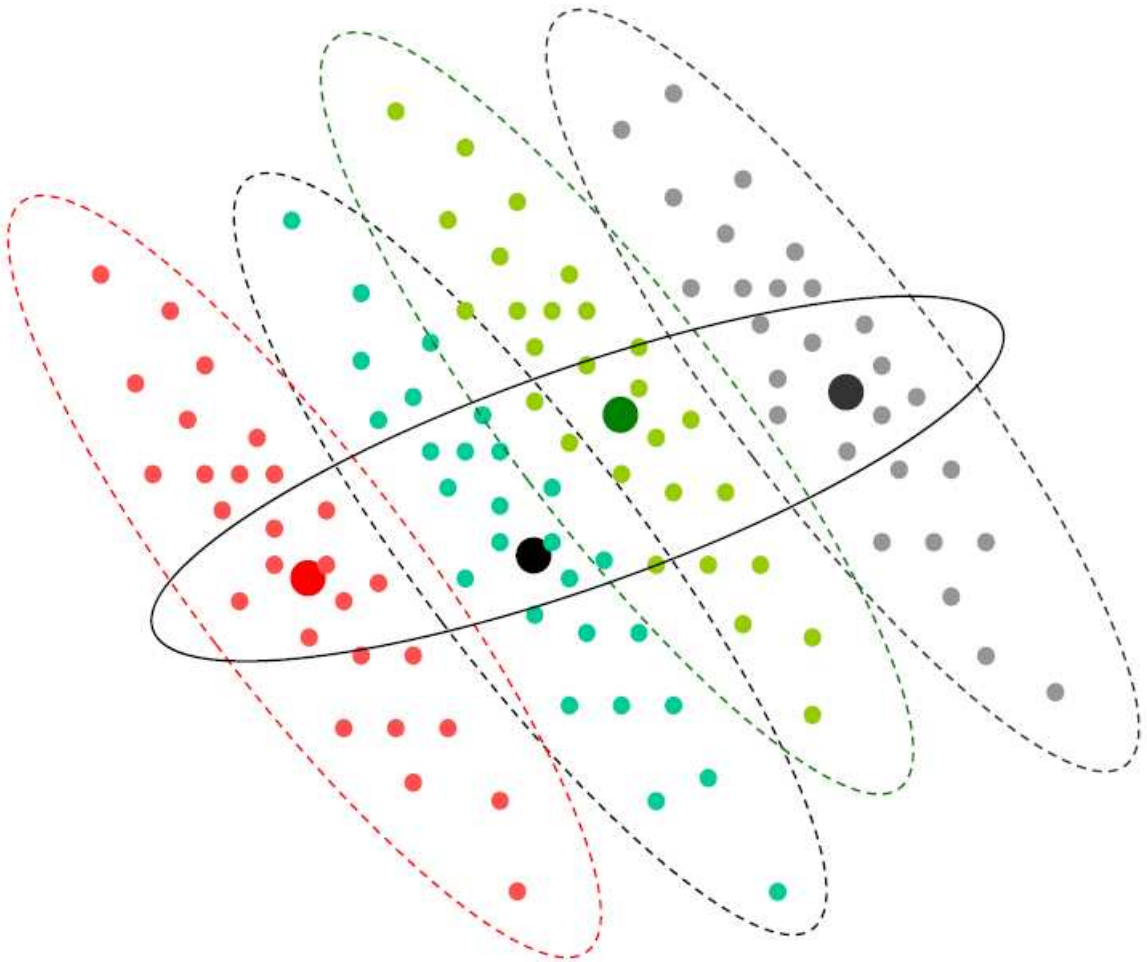

Non−linear 2D−subspace in 3D−Simplex


Multinomial parameters in 3D−Space
0.23
0.22
0.21
0.2
0.19
0.18
0.17
0.3
0.28
0.26
0.24
0.38
0.4
0.42
0.44
0.46
0.48


Non−linear 2D−subspace in 3D−Simplex


Multinomial parameters in 3D−Space
0.23
0.22
0.21
0.2
0.19
0.18
0.17
0.3
0.28
0.26
0.24
0.38
0.4
0.42
0.44
0.46
0.48


iVectors in 2D−Space
150
100
50
0
−50
−100
−60
−40
−20
0
20
40
60
80



Using the PLDA model, one can directly evaluate the log-likelihood ratio for the hypothesis test corresponding to "the two iVectors were generated by the same speaker or not":

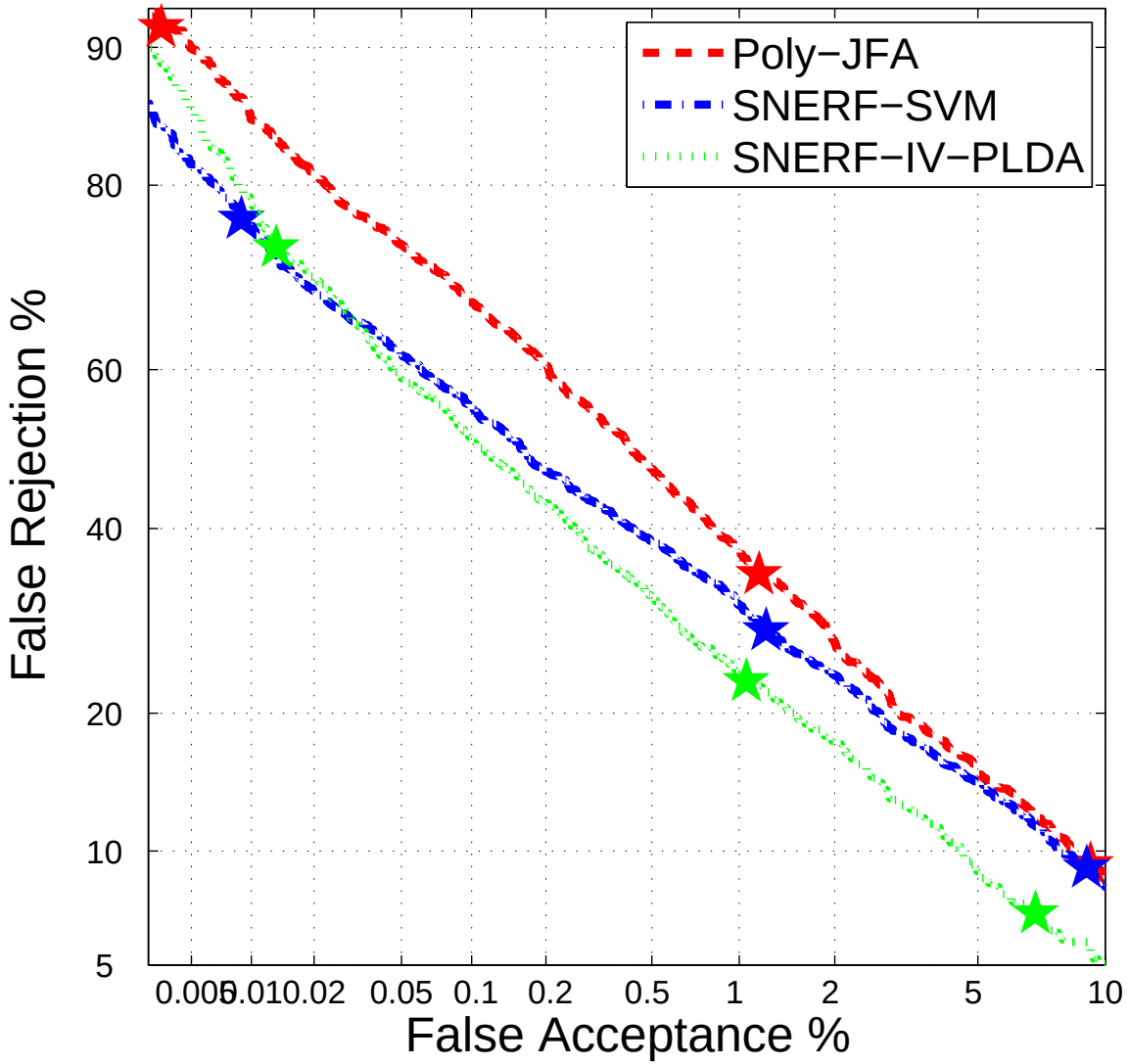
$$s = \log \frac{\int p(\mathbf{w}_1|\mathbf{y})p(\mathbf{w}_2|\mathbf{y})p(\mathbf{y})d\mathbf{y}}{p(\mathbf{w}_1)p(\mathbf{w}_2)} \tag{1}$$

The numerator gives the marginal likelihood of producing both iVectors from the same speaker, while the denominator is the product of the marginal likelihoods that both iVectors are produced from different speakers. The integrals can be evaluated analytically and scoring can be performed very efficiently.

NIST SRE 2008 task:

- Extended for SRE 2010
- Telephone condition
- 1,108 target samples
- 1,453,237 impostor samples
- UBM, JFA, SMM and PLDA on SRE 04 & 05 and SW

- **Poly-JFA**: Joint Factor Analysis (JFA) modeling of Gaussian mean parameters for simple prosodic polynomial contour features [Dehak07]
- **SNERF-SVM**: Support Vector Machine (SVM) modeling of SNERF soft counts [Ferrer07]
- **SNERF-IV-PLDA**: iVector modeling of SNERF soft counts with successive PLDA modeling
- **Baseline**: SRI cepstral system for NIST SRE 2010


Poly−JFA
SNERF−SVM
SNERF−IV−PLDA
False Rejection %
90
80
60
40
20
10
5
0.005
0.01
0.02
0.05
0.1
0.2
0.5
1
2
5
10
False Acceptance %

Baseline: EER 1.65%, old DCF 0.073, new DCF 0.42.

Table: *Score level fusion by Logistic Regression (LR).*

| | System | new DCF | old DCF | EER |
|---|---|---|---|---|
| Fusion | Baseline+Poly-JFA | 0.390 | 0.074 | 1.74 |
| | Baseline+SNERF-SVM | 0.386 | 0.070 | **1.47** |
| | Baseline+SNERF-IV-PLDA | **0.376** | **0.069** | 1.56 |

- PLDA highly outperforms simple LDA+WCCN (20% relative on EER)
- IV-PLDA system gives best overall performance (6.9% EER)
- Decrease of gain towards low false acceptance regions?
- Fusion with baseline gives 10% relative on new DCF measure
- Investigations in diverse channel and speech style conditions
- iVector modeling with PLDA for simple polynomial features
- Combination of Gaussian and Multinomial iVector modeling

# Thank you!

Probabilities of multinomial distribution $\phi_{nc}$ are represented by a log-linear model, where the vector of log-probabilities (natural parameters) is constrained to live in subspace (similar to standard iVectors).

$$\phi_{nc} = \frac{\exp(m_c + \mathbf{t}_c \mathbf{w}_n)}{\sum_i^C \exp(m_i + \mathbf{t}_i \mathbf{w}_n)}, \tag{2}$$

$\mathbf{t}_c$: $c$-th row of subspace matrix $\mathbf{T}$

$\mathbf{w}_n$: $r$-dimensional column vector (i-vector) representing speaker and channel of utterance $n$.

All Multinomial models (one for **each** SNERF token) are stacked into one super-vector, which is modeled by **single** subspace.