# Localization of Non-Linguistic Events in Spontaneous Speech by Non-Negative Matrix Factorization and Long Short-Term Memory

Felix Weninger, Björn Schuller,
Martin Wöllmer, Gerhard Rigoll

Institute for Human-Machine Communication

Technische Universität München, Germany

# Outline

- Background and Motivation
- The Features:
  Non-Negative Matrix Factorization
- The Classifier:
  Long Short-Term Memory
- Evaluation: Buckeye Corpus
- Conclusions

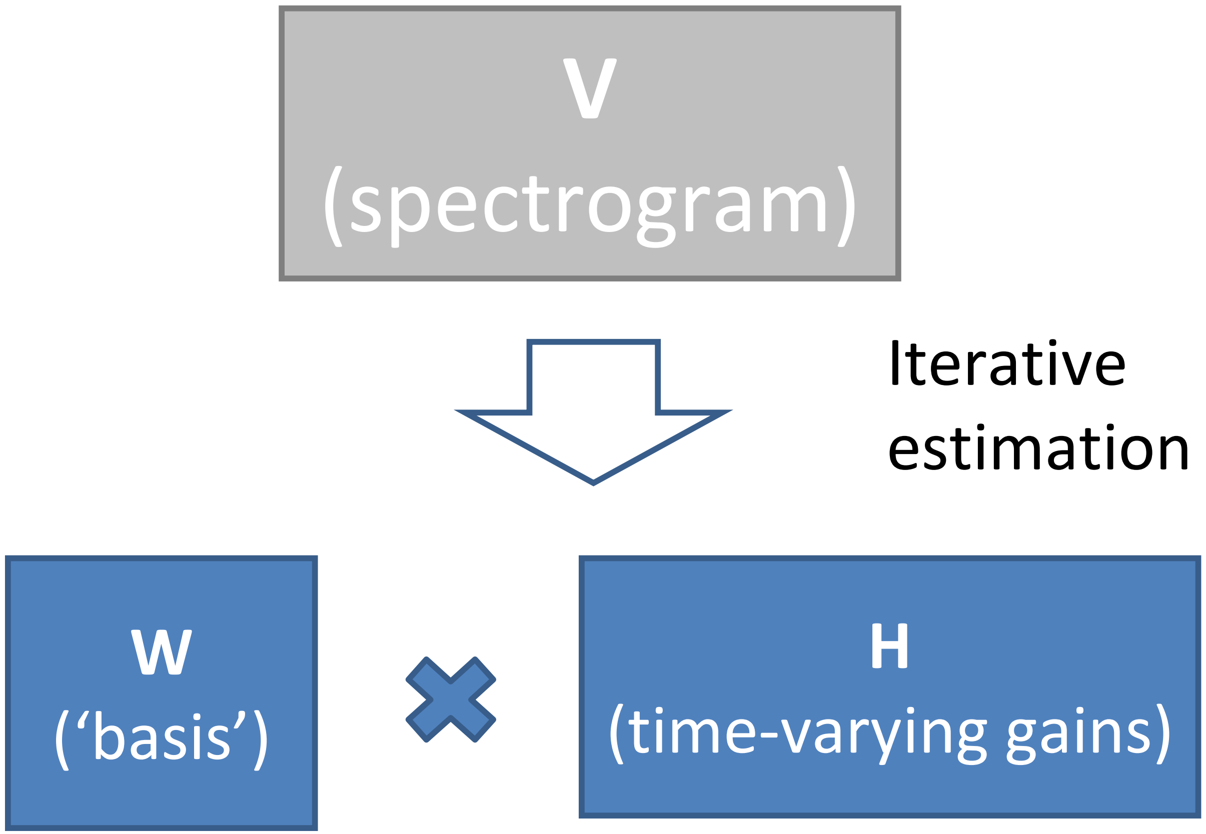
# Background and Motivation

- Localization of non-linguistic segments: laughter, vocal noise, environmental noise, …
- Gain paralinguistic information
- Increase word accuracy
- Inside / outside ASR framework?
- Here:
  - Data-based approach
  - Frame-wise context-sensitive classification

# Background and Motivation (2)

- NMF+SVM classification (Schuller and Weninger 2010)
  - Speech / non-linguistic segments
  - Functionals of NMF activations
  - Outperformed MFCC features
- Now: Include segmentation
  - Bidirectional Long Short-Term Memory RNN
  - Successfully used for phoneme recognition

# NMF for Audio Processing




Open-source toolkit: openBliSSART
(http://openblissart.github.com/openBliSSART)

# NMF Algorithm

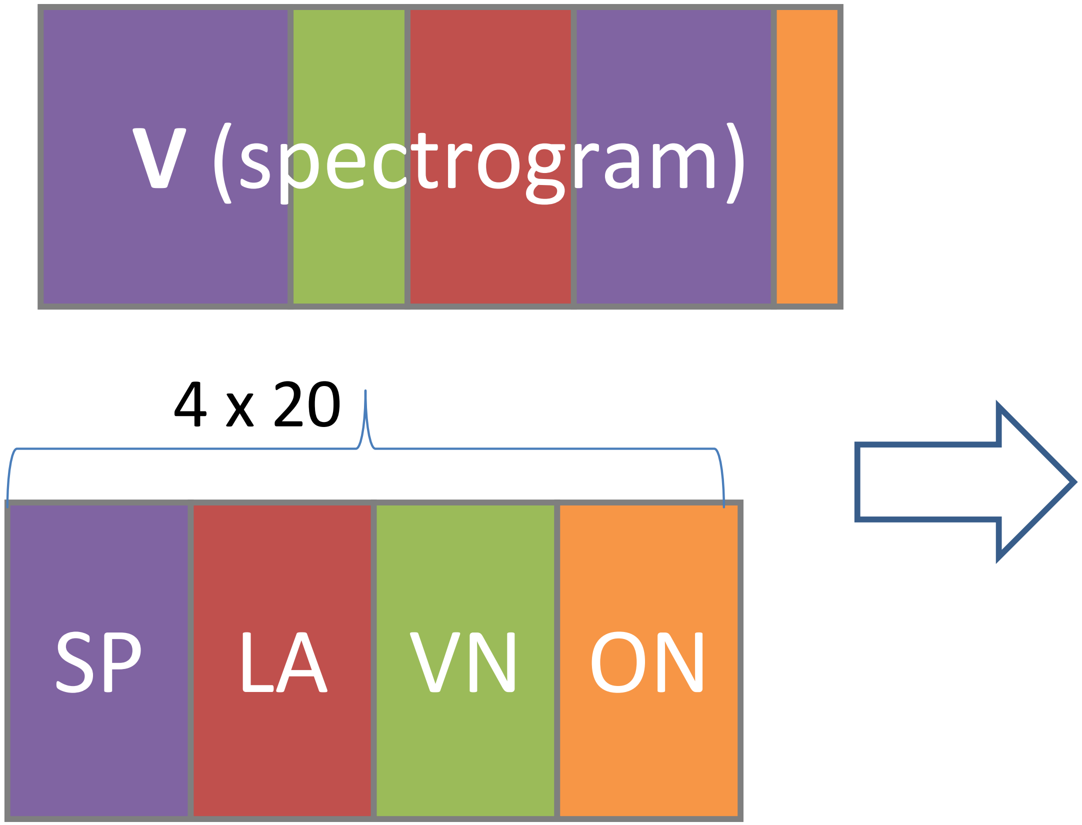
- Multiplicative updates for **W** and **H**
- Minimization of cost d(**V**, **WH**)
  – [Euclidean distance] (Schuller and Weninger 2010)
  – Kullback-Leibler divergence (NMF-KL)

$$d_1(\mathbf{V}|\mathbf{WH}) \quad = \quad \sum_{i,j} [\mathbf{V}]_{ij} \log \frac{[\mathbf{V}]_{ij}}{[\mathbf{WH}]_{ij}} - [\mathbf{V} - \mathbf{WH}]_{ij}$$

  – Itakura-Saito divergence (NMF-IS)

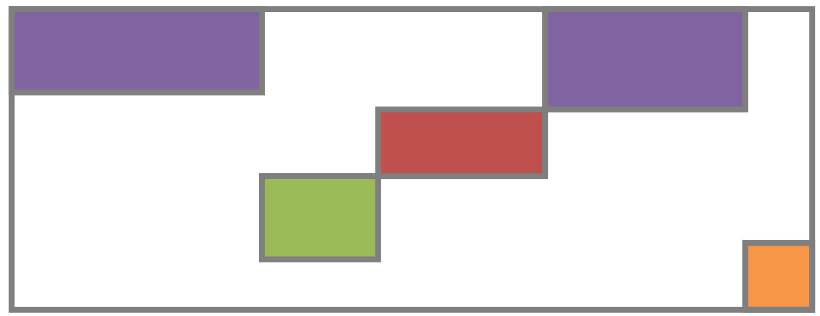
$$d_0(\mathbf{V}|\mathbf{WH}) \quad = \quad -MN + \sum_{i,j} \frac{[\mathbf{V}]_{ij}}{[\mathbf{WH}]_{ij}} - \log \frac{[\mathbf{V}]_{ij}}{[\mathbf{WH}]_{ij}}$$

# NMF Likelihood Features

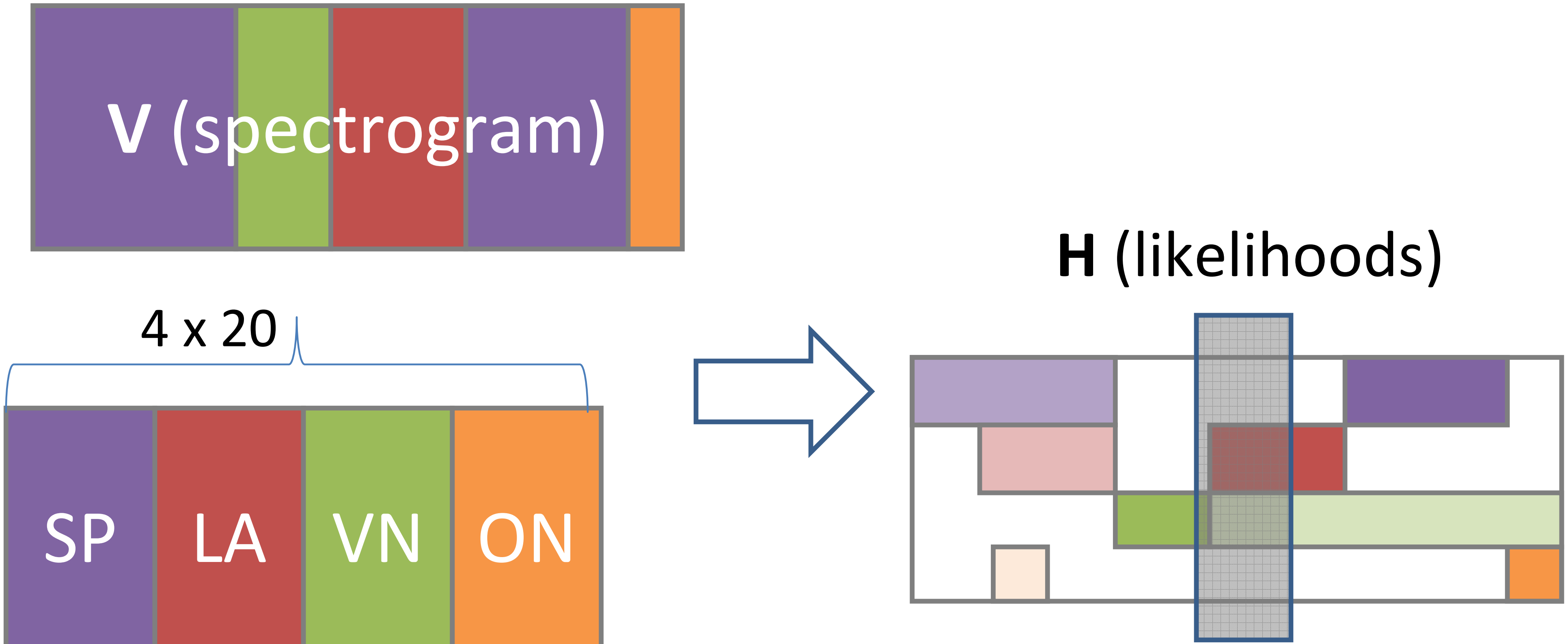



**W** (predefined, NMF on training data)

**H** (estimated activations)



# NMF Likelihood Features




**W** (predefined, NMF on training data)

$\Sigma$=1

+energy

# Long Short-Term Memory

- Conventional RNN: Context range limited
  - Weights  decay exponentially over time
  - “Vanishing gradient problem”
- Solution: LSTM memory cells
  - Internal state maintained by 1.0-connection
  - Input, output, memory controlled by multiplicative gate units
  - Automatically learn required amount of context

# RNN Configurations

| | Unidir. | Bidir. |
|---|---|---|
| Logistic units | RNN | BRNN |
| LSTM cells | LSTM | BLSTM |

- 3 Layers:
  - Input: 39 (PLP) / 83 (NMF)
  - Hidden: 80 / 120
  - Output: 4 (posterior prob.)
- Bidirectional: 2 input / hidden layers

# Evaluation: Buckeye Corpus

- > 25 hours of spontaneous speech
- 40 speakers (20 male, 20 female)
- Speaker-independent evaluation
  - Training, validation, test set stratified by age / gender
  - Subdivision in ascending order of speaker ID
- Automatic alignment
  - **SP**eech, **LA**ughter, **V**ocal **N**oise, **O**ther **N**oise

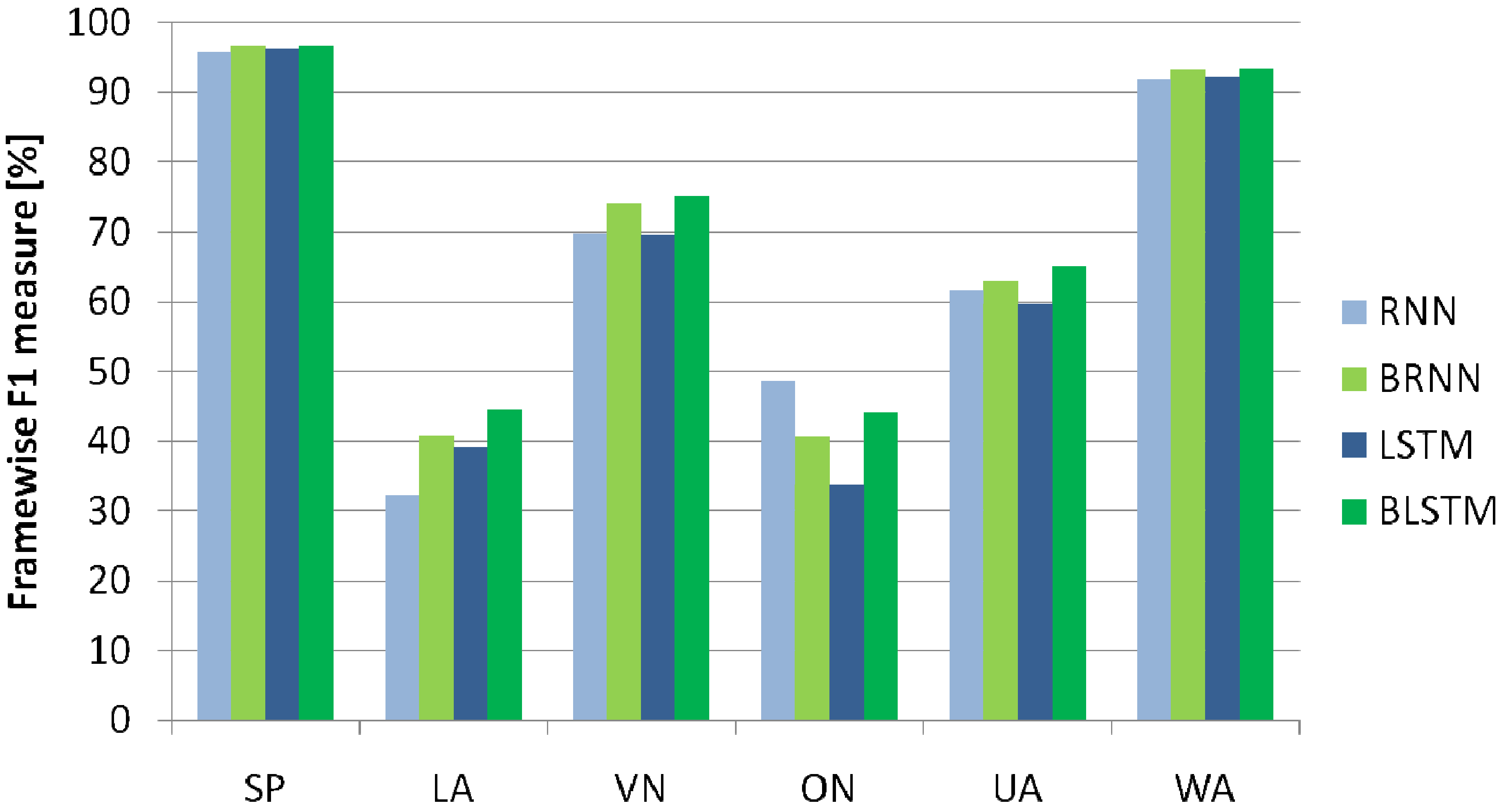
# Evaluation: Buckeye Corpus

| [sec] | **train** | **valid** | **test** | Σ |
|---|---|---|---|---|
| SP | 62 974 | 7 050 | 7 960 | 77 983 |
| LA | 1 562 | 252 | 104 | 1 918 |
| VN | 9 444 | 1 336 | 1 087 | 11 867 |
| ON | 398 | 94 | 30 | 522 |
| Σ | 74 378 | 8 732 | 9 181 | 92 290 |

# Evaluation: Baseline HMM-ASR

- PLP coefficients 1-12 + RMS Energy + δ+ δ δ
- 9.1 K Back-off bi-gram language model (Buckeye training set)
- Monophones:
  - 39 phoneme models (3 states); silence + sp
  - 3 non-linguistic models (LA, VN, ON) with 6 states
- State-clustered triphones, 16/32 mixtures
- Word accuracy: 50.0%

# Results (1): Types of RNNs



# Results (2): BLSTM Size and Features

| **F1 [%]** | **PLP** | | **NMF-IS** | | **NMF-KL** | |
|---|---|---|---|---|---|---|
| # units | 80 | 120 | 80 | 120 | 80 | 120 |
| SP | 96.69 | 96.67 | 96.77 | 96.81 | 96.80 | **96.96** |
| LA | 44.54 | 44.53 | 37.59 | 35.83 | 40.01 | **45.95** |
| VN | 75.08 | 75.07 | 73.54 | 72.41 | 72.64 | **75.79** |
| ON | 38.29 | 44.12 | 39.31 | 32.54 | 39.09 | **50.76** |
| **UA** | 63.65 | 65.10 | 61.80 | 59.40 | 62.14 | **67.37** |
| **WA** | 93.39 | 93.38 | 93.26 | 93.16 | 93.28 | **93.82** |

# Results (3): BLSTM-NMF vs. HMM-PLP

| [%] | HMM-ASR (PLP) | | | BLSTM (NMF-KL) | | |
|---|---|---|---|---|---|---|
| | REC | PR | F1 | REC | PR | F1 |
| SP | 93.85 | 97.68 | 95.72 | 97.62 | 96.31 | **96.96** |
| LA | 50.63 | 45.47 | **47.91** | 61.70 | 36.61 | 45.95 |
| VN | 78.41 | 63.84 | 70.38 | 69.58 | 83.22 | **75.79** |
| ON | 39.92 | 14.78 | 21.57 | 49.98 | 51.56 | **50.76** |
| **UA** | 65.70 | 55.44 | 58.90 | 69.72 | 66.92 | **67.37** |
| **WA** | 91.35 | 92.79 | 92.06 | 93.71 | 93.92 | **93.82** |

WA REC 91.35 → 93.71% : $p < .001$

# Conclusions

- BLSTM-NMF vs. HMM-ASR: 37.5% relative reduction of frame-wise error rate
- Best results with KL divergence
- Future work:
  - Use BLSTM prediction / NMF likelihoods in multi-stream HMM-ASR
  - Context-sensitive NMF features (deconvolution algorithm, etc.)

Thank you.