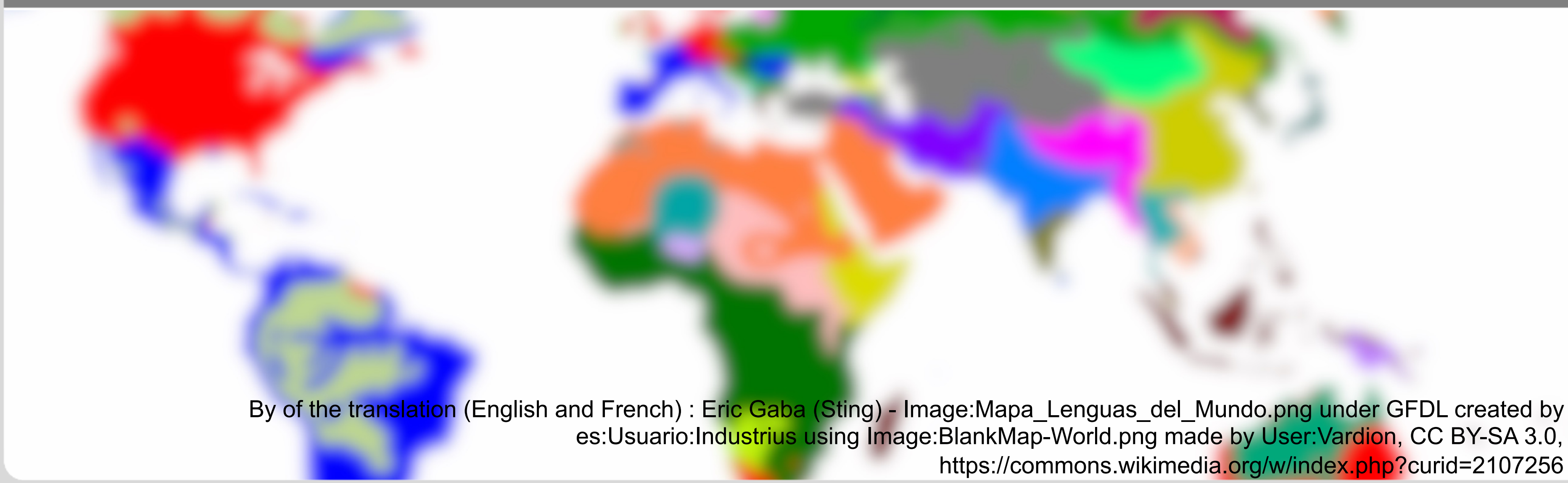
Karlsruhe Institute of Technology

# Towards Improving Low-Resource Speech Recognition Using Articulatory and Language Features

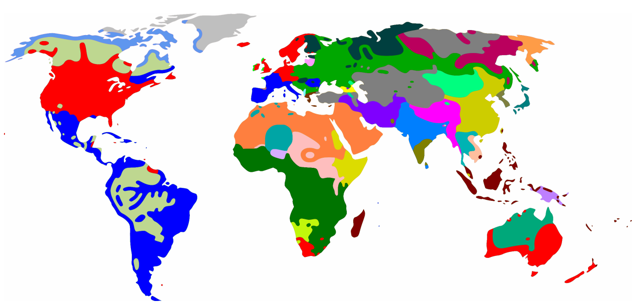
**Markus Müller,** *Sebastian Stüker and Alex Waibel*

Institute for Anthropomatics and Robotics, Interactive Systems Lab



# Low-Resource Speech Recognition

- Long tail of languages with only limited data available
- Train multilingual speech recognition systems
  - Merge training data from multiple languages
  - Built system with multilingual phone set
- Adapt neural networks to languages
  - Language Feature Vectors, similar to i-Vectors
  - Append language information to acoustic features
- Use articulatory features (AFs) as additional input features
  - Phoneme inventory is limited
  - Phonemes represent certain AFs configuration
  - Detecting AFs: No limitation to configurations

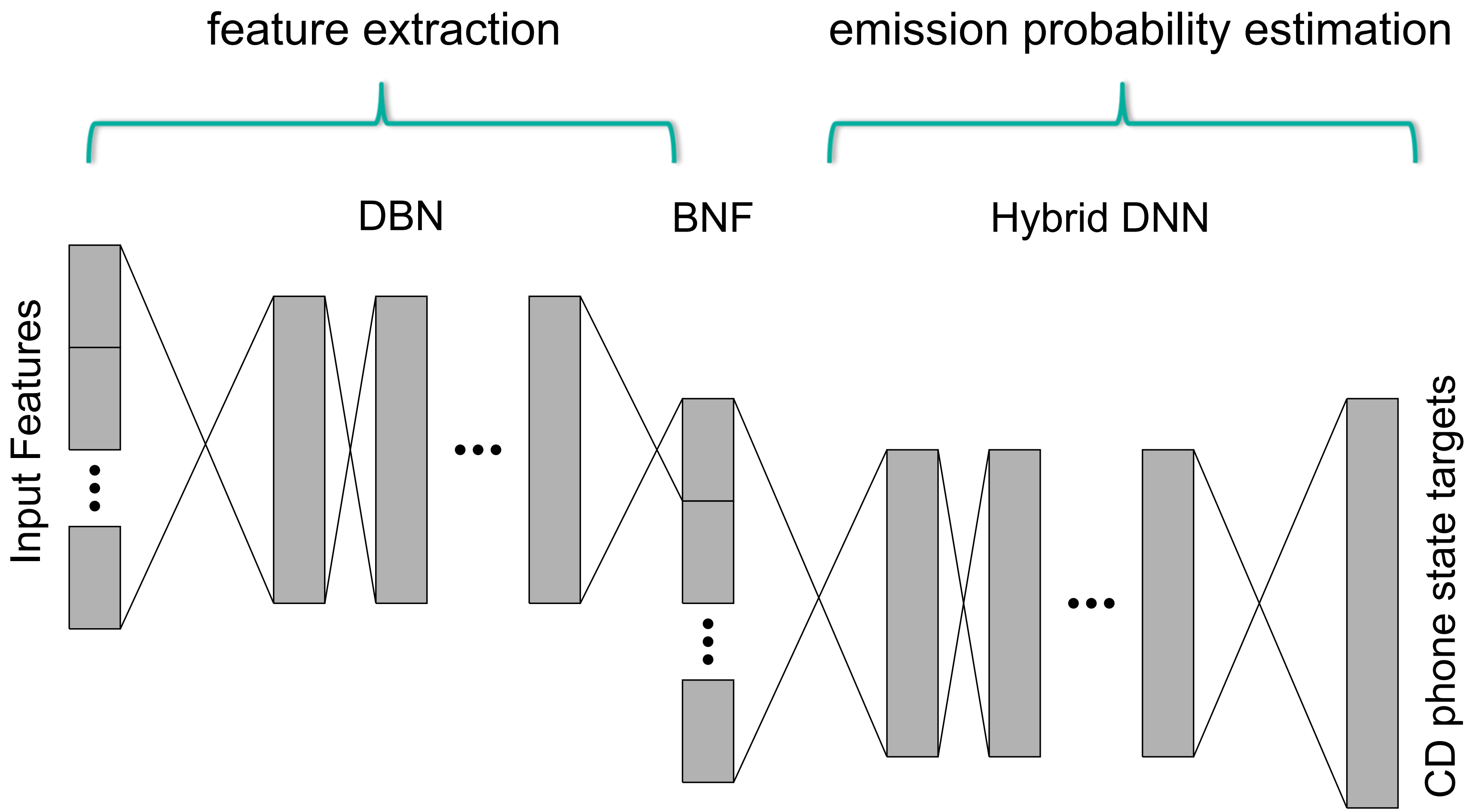


# Training Data

- TV broadcast news from Euronews
- Multilingual speech corpus
- 70h per language

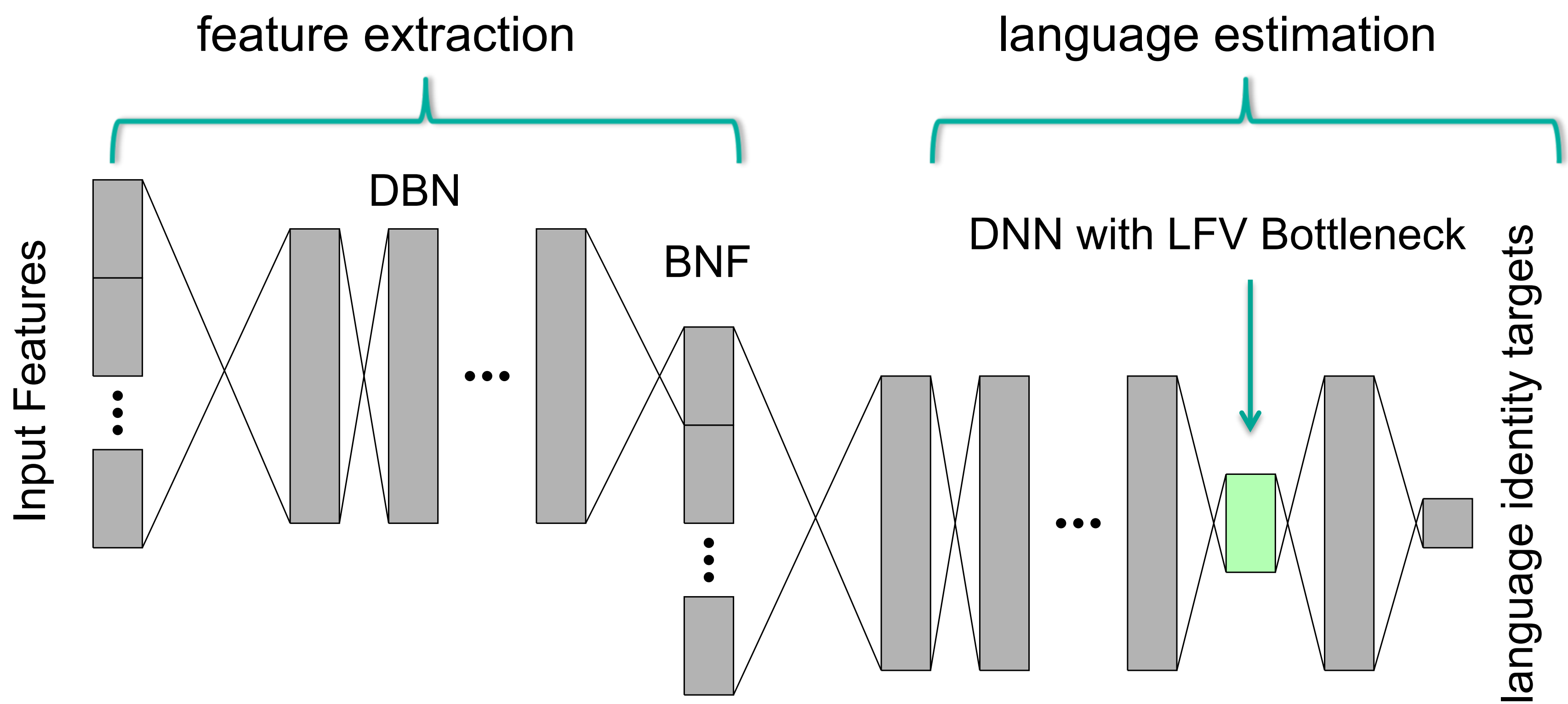
| Language | Audio Data | # Recordings |
|---|---|---|
| Arabic | 72.1h | 4,342 |
| English | 72.8h | 4,511 |
| French | 68.1h | 4,434 |
| German | 73.2h | 4,436 |
| Italian | 77.2h | 4,464 |
| Polish | 70.8h | 4,576 |
| Portuguese | 68.3h | 4,456 |
| Russian | 72.2h | 4,418 |
| Spanish | 70.5h | 4,231 |
| Turkish | 70.4h | 4,385 |
| Total | 715.6h | 44,253 |

# Our HMM/ANN Hybrid Architecture



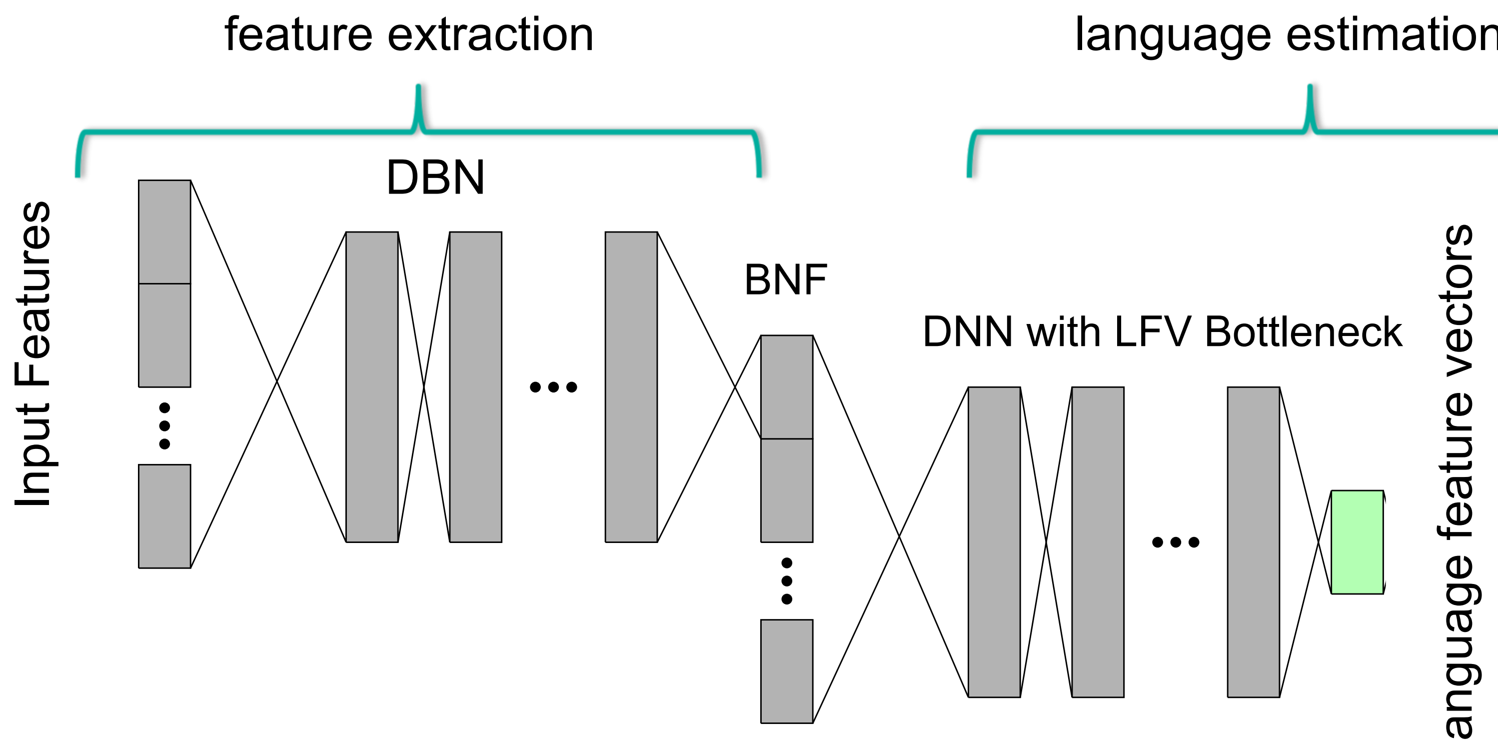
# Architecture for LFV Extraction

- Increased context width → language information long-term in nature
- LFV extraction: Discard layers after bottleneck
- Trained on 70h per language on 9 languages



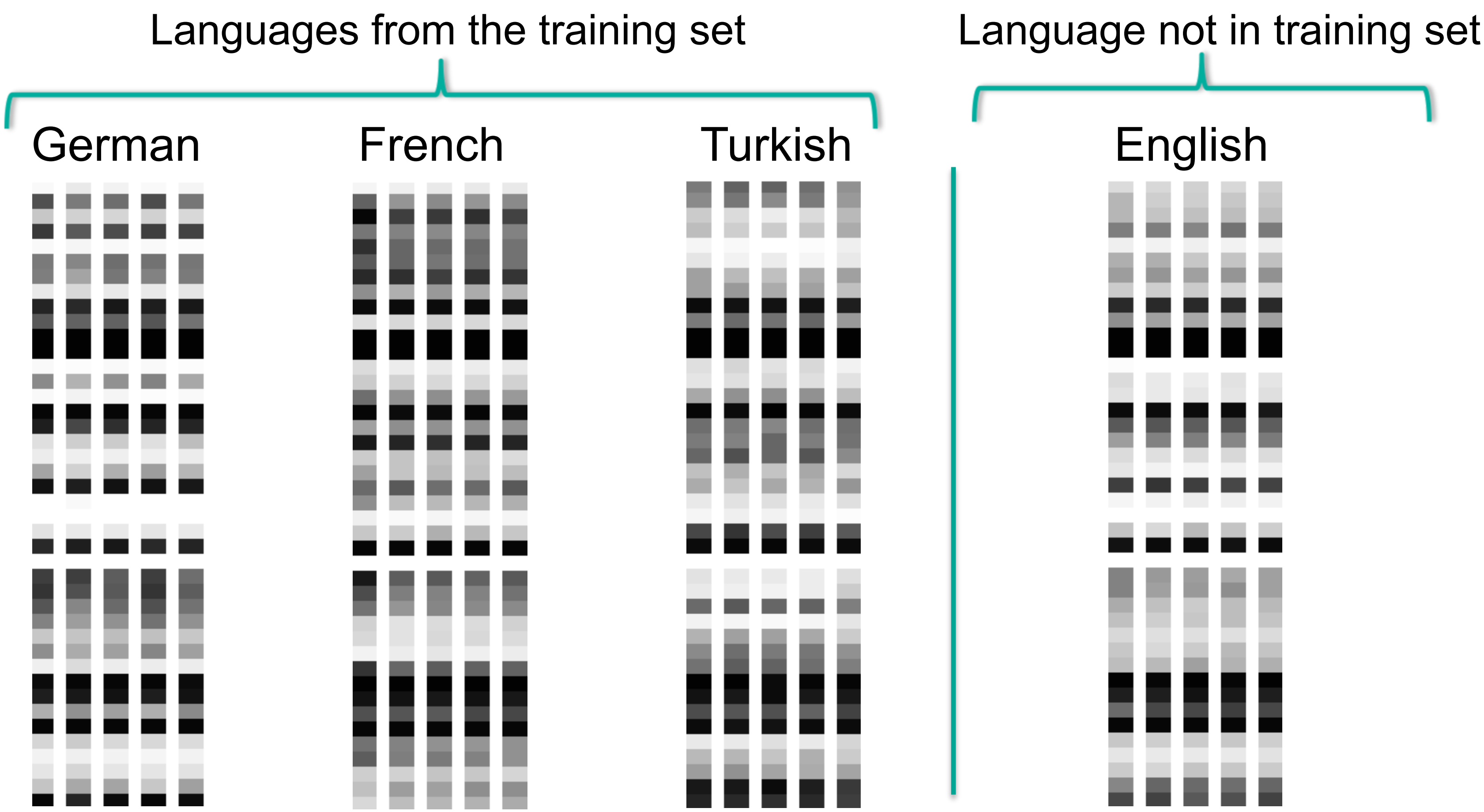
# Architecture for LFV Extraction

- Increased context width → language information long-term in nature
- LFV extraction: Discard layers after bottleneck
- Trained on 70h per language on 9 languages
  - TV broadcast news from Euronews



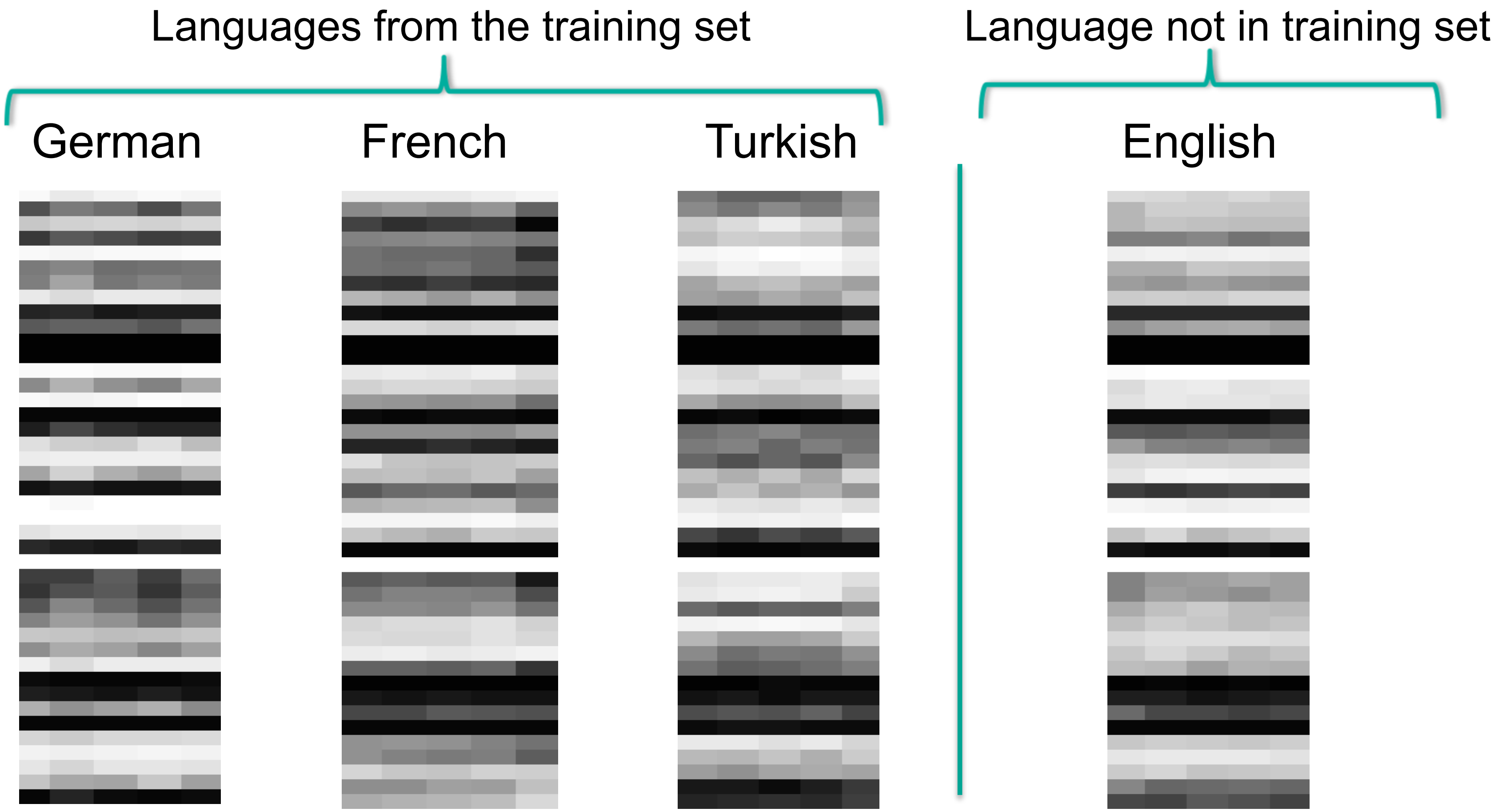
# Example Language Feature Vectors

- 5 examples per language



# Example Language Feature Vectors

- 5 per language

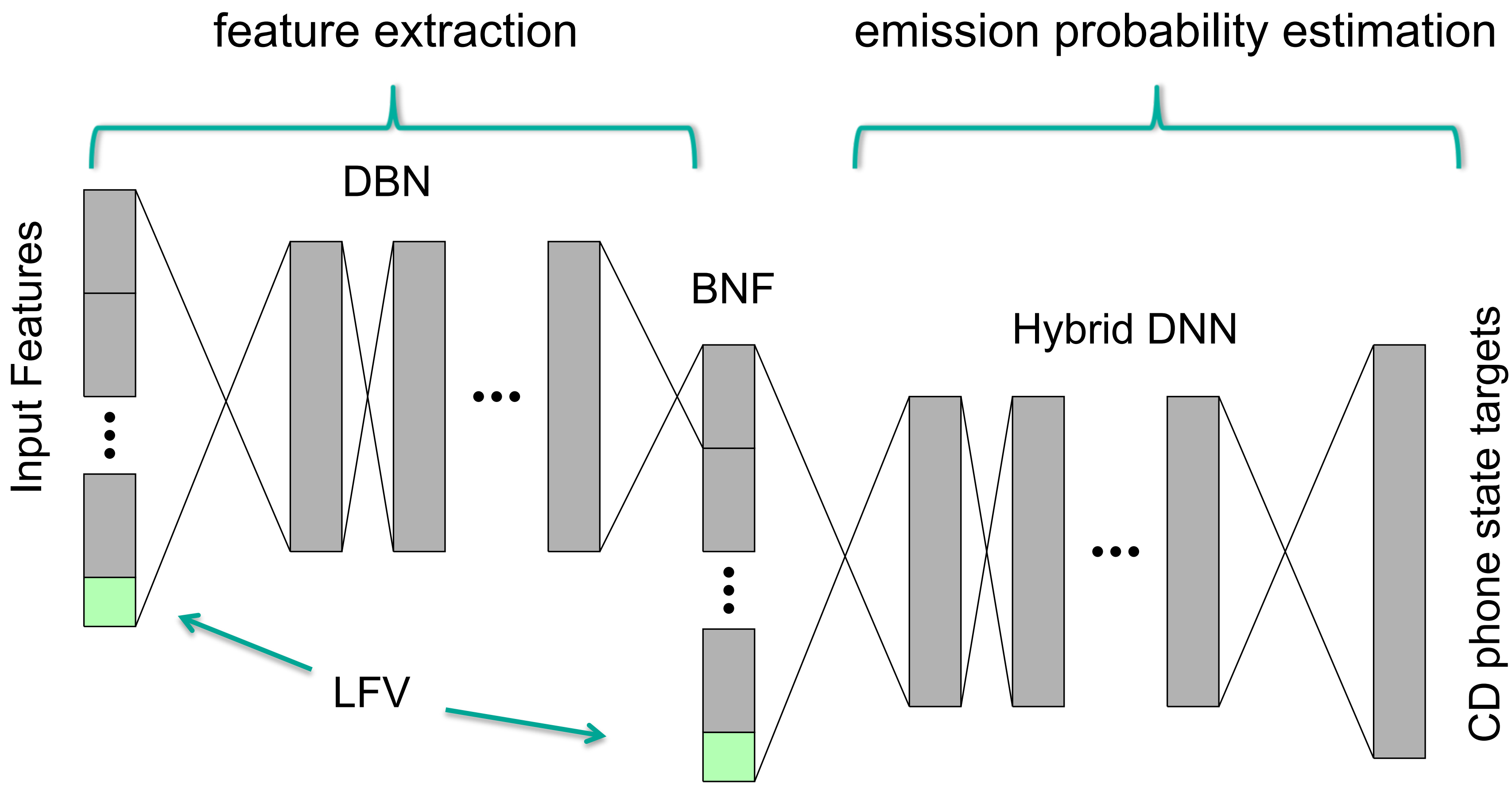





Institute for Anthropomatics and Robotics
Interactive Systems Lab

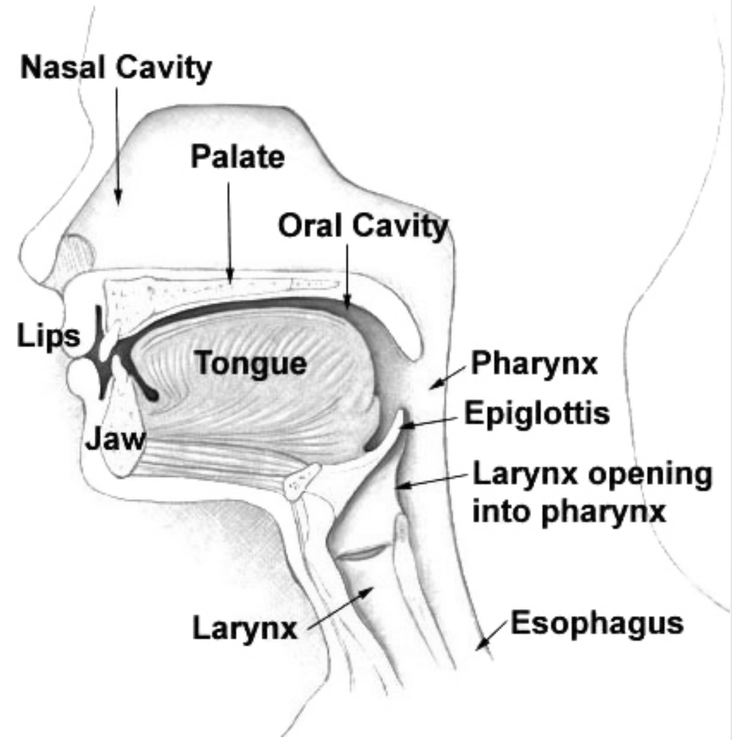
# Adding LFVs to ASR Systems

- LFVs added to acoustic input and bottleneck features
- Provide implicit language information to networks



# Articulatory Features (AFs)

- Represent state of articulators from the human vocal tract
  - e.g. place or articulation, tongue position
- Phonemes represent certain configuration of articulators
  - Configurations limited by phoneme inventory
  - Phoneme inventory limited by languages seen during training
- Detecting AFs directly allows for unlimited configurations
- Using AFs as additional input feature
  - Language universal




By Arcadian - http://training.seer.cancer.gov/head-neck/anatomy/overview.html, Public Domain, https://commons.wikimedia.org/w/index.php?curid=1678037

# Articulatory Features (AFs) 2

- 7 types of AFs
  - 3 for consonants (cplace, ctype, cvox)
  - 4 for vowels (vfront, vheight, vlng, vrnd)
  - Added additional target “does not apply”
- Additional: Detect type of phoneme
  - Consonant, vowel, noise, silence
- Discrete valued AFs

| Name | Description | # Classes |
|---|---|---|
| cplace | Place of articulation | 8 |
| ctype | Type of articulation | 6 |
| cvox | Voiced | 2 |
| ptype | Type of phoneme | 4 |
| vfront | Tongue back / front | 3 |
| vheight | Height of tongue | 3 |
| vrnd | Lips rounded | 2 |
| vlng | Type of vowel | 4 |

# AF Training Data

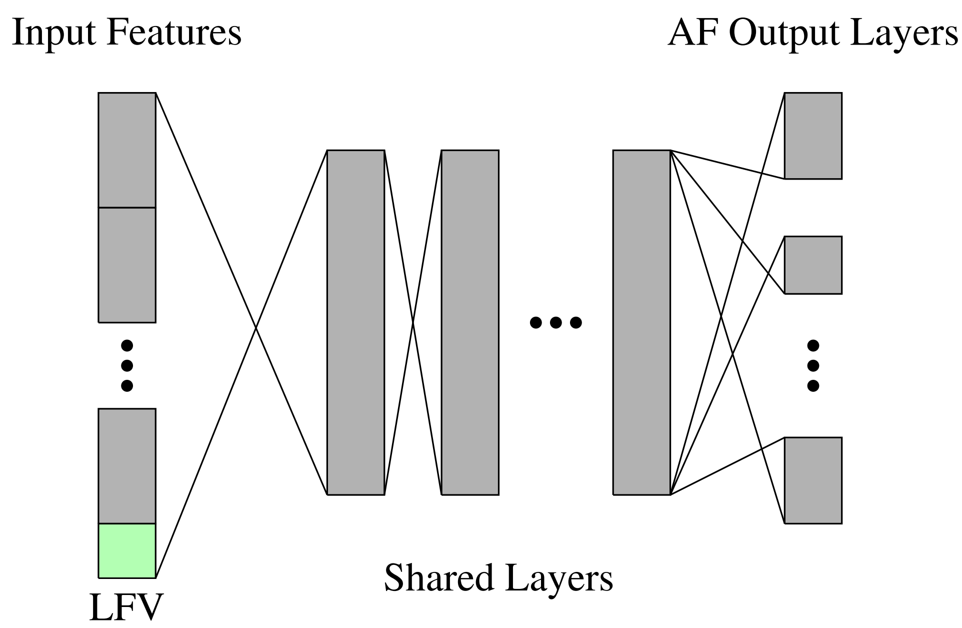
- Created phoneme / AF mapping using definitions from MaryTTS

```
<vowel     ph="Y" vlng="s" vheight="1" vfront="2" vrnd="+"/>
<consonant ph="p" ctype="s" cplace="l" cvox="-"/>
```

- Obtained AF training data based on labels from ASR systems
  - Phonemes modeled by 3 sub-phone states (begin, middle, end)
  - Mapped phonemes to AFs
  - Extracted data only from middle sub-phone states
  - Articulators in static position, do not move from one target to another
- Trained networks on data from 4 languages
  - English, French, German, Turkish

# AF Network Training

- Trained networks for AF extraction independent of each other
  - Prevent co-adaption based on combinations present in languages
- Multi-task Learning
  - Shared Hidden Layers
  - One output per AF



# Evaluation of AF Extraction

- Networks trained on 70h of French, German and Turkish
- Frame error rate (FER) on validation set
- Adding LFVs to input features lowers FER
- Mixed results for Multi-task Learning

| Setup | LFV | MTL | cplace | ctype | cvox | ptype | vfront | vheight | vlng | vrnd |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | - | - | 8.4 | 8.2 | 7.8 | 14.8 | 7.2 | 7.9 | 7.3 | 6.2 |
| 2 | • | - | **7.0** | **6.8** | 6.3 | 12.7 | 5.8 | **6.6** | 5.7 | 5.0 |
| 3 | • | • | 7.3 | 6.9 | **6.2** | **12.6** | **5.7** | **6.6** | **5.5** | **4.9** |

# Evaluation of AF Extraction (2)

- Networks trained on 4 languages, with LFVs
  - English, French, German, Turkish
- FER on English validation set
- Setup 1
  - Trained on 10h per language
- Setup 2
  - Trained nets first on 70h of French, German, Turkish
  - Additional fine-tuning on 10h of all 4 languages, reduced learning rate

| Setup | 3L pre-train | cplace | ctype | cvox | ptype | vfront | vheight | vlng | vrnd |
|---|---|---|---|---|---|---|---|---|---|
| 1 | - | 9.1 | 9.7 | 9.5 | 16.4 | 8.8 | 7.9 | 8.3 | 6.0 |
| 2 | • | **8.8** | **8.2** | **8.2** | **15.2** | **7.8** | **7.2** | **7.5** | **5.3** |

# AF Based ASR Systems

- Systems trained on 4 languages, 10h per language
  - English test set
- Multilingual system
- Using AFs as input features
  - Concatenating outputs of networks
  - 39 dimensional feature vector
- Replacing lMel + tone with AFs does not lead to improvements
- Adding LFVs increases performance

| Setup | Features | LFV | WER |
|---|---|---|---|
| 1 | lMel+T | - | 20.2% |
| 2 | AF (3L) | - | 22.6% |
| 3 | lMel+T | • | 18.7% |
| 4 | AF (3L) | • | 21.8% |
| 5 | AF (4L) | • | 20.2% |

# Combining Multiple Input Features

- Combine lMel + tone with AFs
- Stacked input features
- All systems using LFVs
- Adding AFs trained on 3 languages decreases performance
- Adding AFs trained on 3 languages and fine-tuned on 4 increases performance

| System | AF | WER |
|---|---|---|
| 1 | - | 18.7% |
| 2 | AF(3L) | 19.0% |
| 3 | AF(4L) | 18.5% |

# Combining Outputs of Different Systems (CNC)

- Trained systems using different types of input features
  - lMel + tone (lMel), MFCC + MVDR + tone (M2), AF
  - All systems use LFVs
- Confusion network combination
  - Same improvements by combining two systems
  - AFs contribute to CNC equally as M2
- Combining all 3 systems leads to best results

| Setup | lMel | M2 | AF | WER |
|---|---|---|---|---|
| 1 | ● | - | - | 18.7% |
| 2 | - | ● | - | 18.7% |
| 3 | - | - | ● | 20.2% |
| 4 | ● | ● | - | 18.1% |
| 5 | - | ● | ● | 18.1% |
| 6 | ● | - | ● | 18.1% |
| 7 | ● | ● | ● | 17.3% |

# Conclusion

- Neural networks for articulatory feature extraction benefit from LFVs
- Adding AFs to lMel + tone shows slight improvement
- Incorporating AF based ASR system in CNC shows improvements
  - AFs contribute as much as, e.g., MFCC + MVDR in system combination

# Thank you!