Karlsruhe Institute of Technology

Neural modulation for multilingual speech recognition

Markus Müller, Sebastian Stüker and Alex Waibel

Institute for Anthropomatics and Robotics, Interactive Systems Lab

Introduction

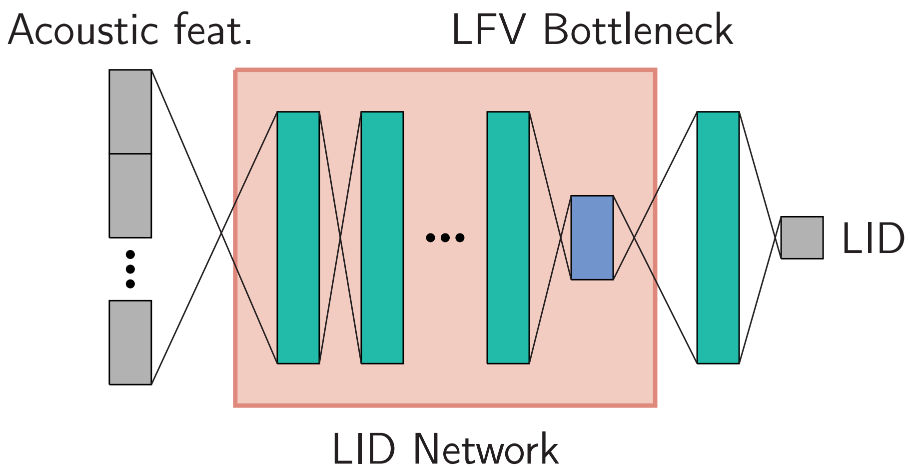
- Automatic speech recognition (ASR): Costly AI problem
  - 7,000+ living languages, each requires own acoustic model
- How to train a system for a language?
  - EN on EN (monolingual): best performance
  - L_x on EN (cross-lingual): worst performance
  - $L_1, L_2, \ldots L_n$ on EN (multilingual): mediocre performance
- Monolingual setup wins
- Multilingual training
  - Train model on multiple languages
  - Fine-tune on target language
- Want: Quick adaptation to languages
  - Monolingual performance multilingually

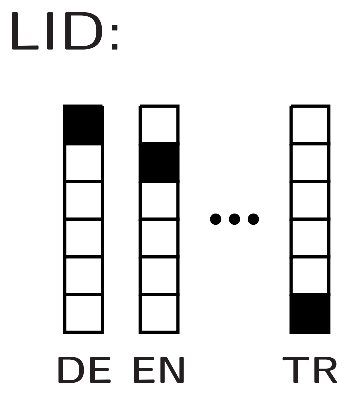
Multilingual Neural Network Adaptation

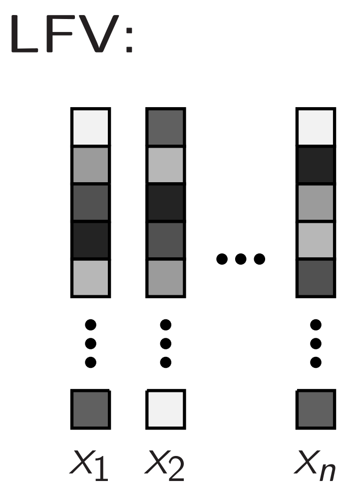
- Multilingual acoustic model: Multilingual set of acoustic units
  - IPA: Same symbols across languages, language specific contexts
  - Multilinguality adds more ambiguity, performance loss
- Adaptation method: Networks modulated by language codes
  - Extracted via ancillary network
- Stimulate networks to learn features depending on language properties
- Optimized neural network architecture and application of language codes
- Achieved and exceeded parity with monolingual setups
- Instantly adapts to languages

Neural Network Language Adaptation

- Supply additional language code
- Language identity (LID)
  - One-hot encoding of identity
- Language Feature Vectors (LFV)
  - Encoding of language properties
  - Extracted via bottleneck layer

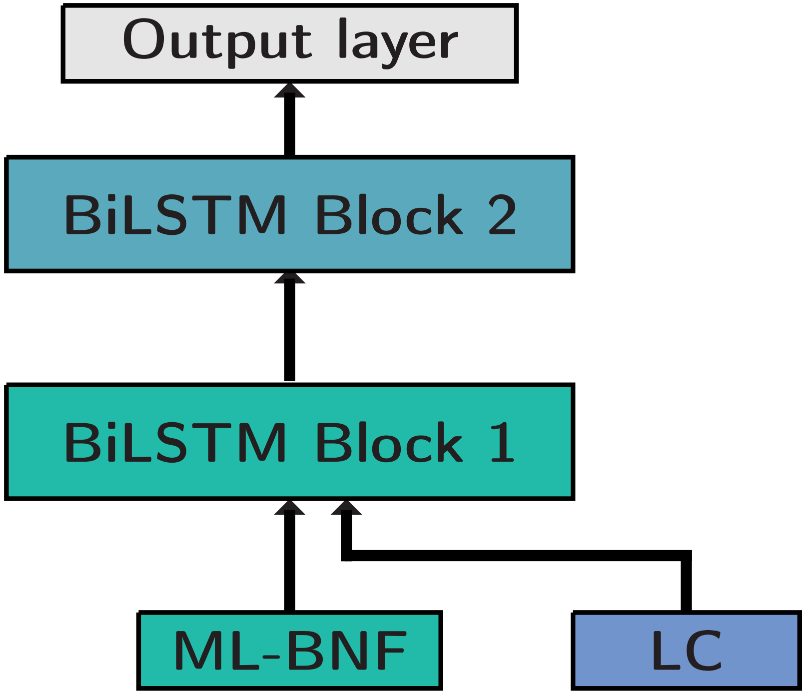




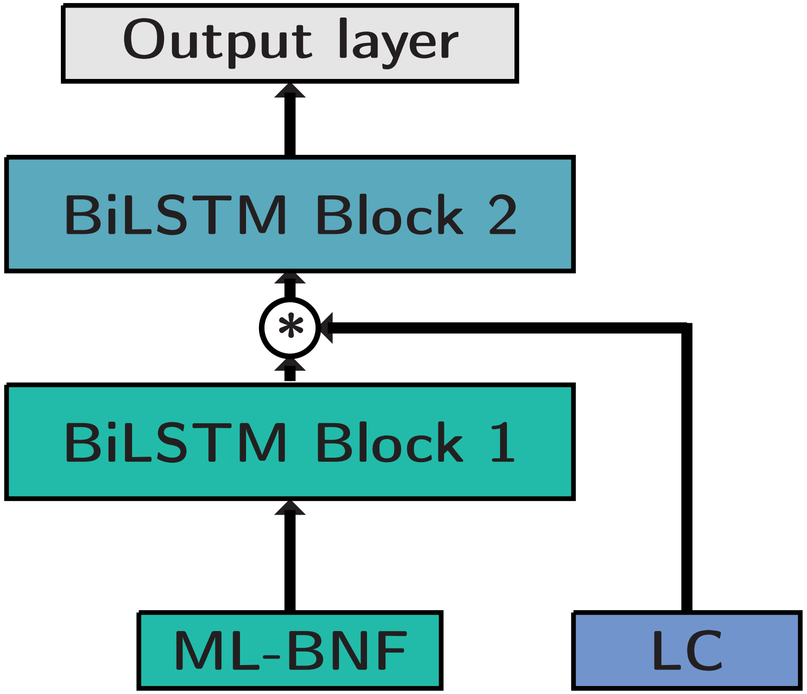


Comparison of Network Architectures

- Additive language codes

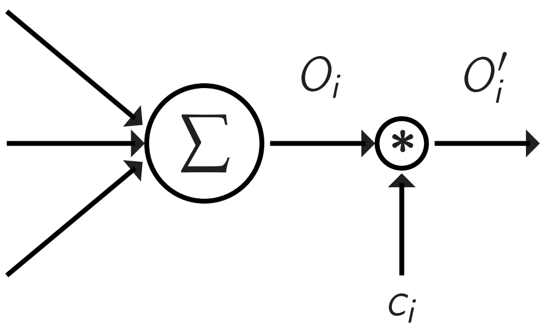


- Multiplicative language codes



Multiplicative Language Codes

- Language properties not as signal related as speaker properties
- Integrate language adaptation deeper into the network
- Neural network modulation related to modulation in Meta-PI
- Outputs weighted by language codes
  - Emphasized / attenuated based on language properties
  - Forces neural units to learn features depending on LCs
  - Network instantly adapts to languages



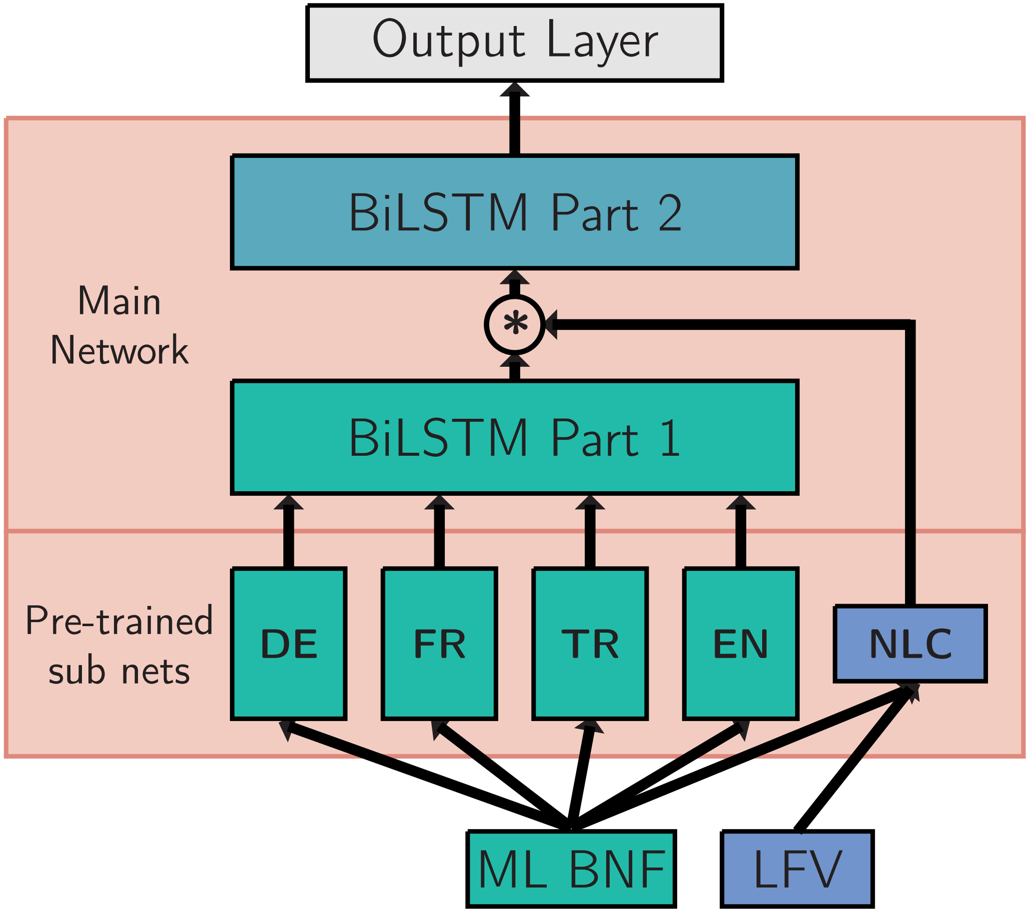
"**The meta-pi network: Building distributed knowledge representations for robust multisource pattern recognition.**" Hampshire, John B., and Alex Waibel. IEEE Transactions on Pattern Analysis and Machine Intelligence 14, no. 7 (1992): 751-769.

Network Superstructure for Multilingual ASR

- Modulation (already covered)
  - Apply weights to outputs of neural units
- Train smaller subnets on individual tasks
  - Language dependent subnets
- Learn mixture weights of subnets based on final task
  - Train adaptive neural language codes (NLCs) based on LFVs
- Joint training of entire network superstructure
  - Parameters of individual networks updated
  - Monolingual subnets adapted to multilingual speech recognition

Network Architecture

- Stack outputs of subnets
  - Language dependent
  - Remove output layers
  - Stack outputs of last hidden layers
- Main network
  - 2 BiLSTM blocks
- Joint training of *all* networks
  - Update pre-trained language dependent networks
  - Update NLCs



Experimental Setup BiLSTM/CTC Systems

- Trained on 4 languages (English, French, German, Turkish)
  - TV broadcast news (Euronews TV station)
  - 45h per language
- No pronunciation dictionaries used
  - Trained on characters only
    $\rightarrow$ Network has to infer pronunciations automatically
- Character based RNN language model
  - Trained on 0.5 million words of training transcripts
- Evaluation metrics
  - WER: Word error rate

Results

- Network superstructure and NLCs improve performance
  - Evaluation on English

Setup	WER LM1	WER LM2
Monolingual baseline	25.3%	24.2%
No adaptation	27.4%	–
LFV Modulation	26.3%	–
Phonetic pre-training	25.4%	–
Network Superstructure	**24.2%**	**23.5%**

- LM1: Baseline
- LM2: Optimized number of BiLSTM cells

Conclusion

- Language adaptation of neural networks
  - Language codes extracted by ancillary network
- Modulation stimulates neural networks to learn features depending on language properties
- Network superstructure with pre-trained sub nets
  - Joint optimization for best recognition performance
  - Multilingual acoustic model achieves and exceeds parity with monolingual counterpart
- Modulation enables mode dependent networks
  - Intelligent "dropout"
  - Apply method to other domains

Thank you.

More details can be found in
“Neural Language Codes for Multilingual Acoustic Models”
Accepted at Interspeech 2018
Pre-print available at: https://arxiv.org/pdf/1807.01956.pdf