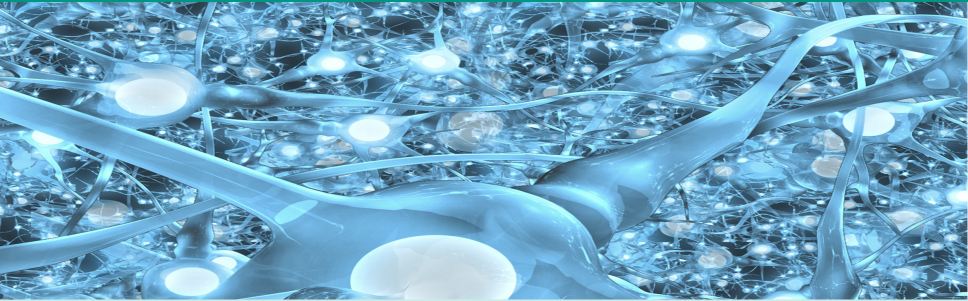
Multi-Feature Modular Deep Neural Network Acoustic Models

Kevin Kilgour & Alex Waibel
3. Dezember 2015

Institut für Anthropomatik und Robotik, Interactive Systems Lab (Lst. Waibel)



Overview

- Introduction
- Feature combination in neural networks
  - Used features
  - Combination approaches
- Modular deep neural network acoustic models
  - Motivation
  - Topology
  - Multiple modules
- Results

Introduction

- Goal: Combination of multiple input features
- Approach: Modular Deep Neural Network Acoustic Models
- Evaluated on the following German test sets
  - IWSLT dev2012:
    - TED and TEDx talks
    - 2 hours of audio from 7 speakers with a total of 18k words
    - Word error rate measured using 3 significant figures
  - Quaero eval2010:
    - Podcasts, talkshows, broadcast news
    - 3.5 hours of audio from 135 speakers with a total of 32k words
    - Word error rate measured using 4 significant figures
- Baseline system: KIT 2014 IWSLT system

Feature combination in neural networks

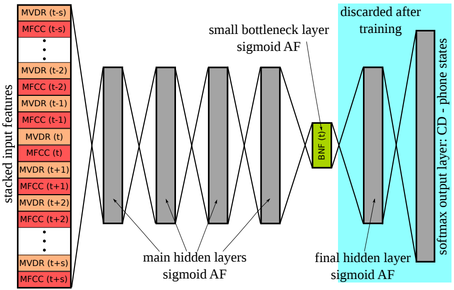
- Many approaches
  - Disregard irrelevant information: speaker, background noise, ...
  - Fundamentally similar and often equally useful
- Can be complementary
- ASR systems using different features can be combined for better results
- Neural networks can be used to combine multiple features in a single ASR system
  - *C. Plahl, R. Schlüter, and H. Ney, "Improved acoustic feature combination for lvcsr by neural networks.", INTERSPEECH, 2011*
  - *K. Kilgour, T. Seytzer, Q. Nguyen, and A. Waibel, "Warped minimum variance distortionless response based bottle-neck features for LVCSR," ICASSP, 2013*
  - *C. Plahl, M. Kozielski, R. Schlüter, and H. Ney, "Feature combination and stacking of recurrent and non-recurrent neural networks for lvcsr," ICASSP, 2013*
  - *F. Metze, Z. A. Sheikh, A. Waibel, J. Gehring, K. Kilgour, Q. B. Nguyen, and V. H. Nguyen, "Models of tone for tonal and non-tonal languages," ASRU, 2013*

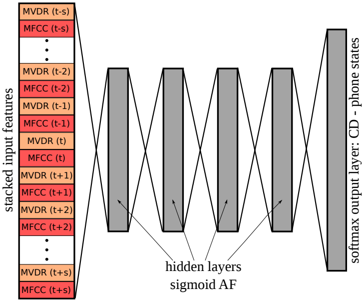
Features

- MFCC
  - 20 dimensional feature vector
  - Standard ASR feature for the past two decades
- MVDR
  - 20 dimensional feature vector
  - Improves on linear prediction features
  - *M. Wölfel, J. W. McDonough, and A. Waibel, "Minimum variance distortionless response on a warped frequency scale." INTERSPEECH, 2003*
- lMEL:
  - 40 dimensional feature vector
  - Precursor feature to MFCC features
  - Typically outperform MFCCs in large DNNs
- Tonal:
  - 14 dimensional feature vector
  - Combination of pitch (7) & FFV (7) feature vectors
  - Can not be used as stand alone features
  - *F. Metze, Z. A. Sheikh, A. Waibel, J. Gehring, K. Kilgour, Q. B. Nguyen, and V. H. Nguyen, "Models of tone for tonal and non-tonal languages," ASRU, 2013*

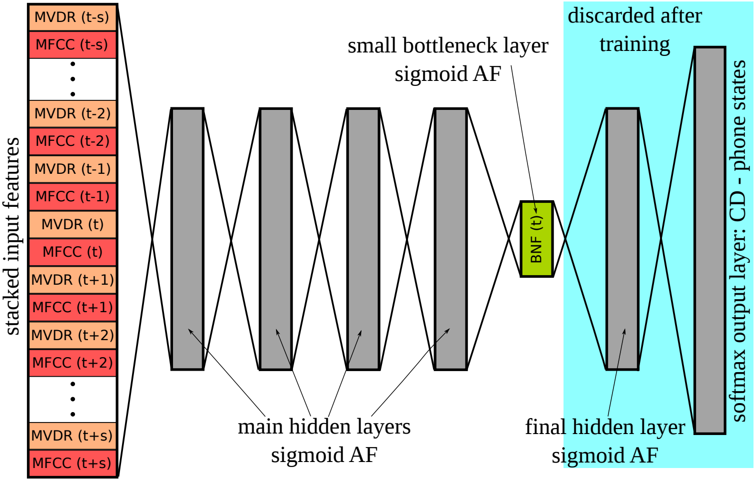
Neural Network Feature Combination Approaches

- Deep bottle neck features
- Deep neural network acoustic models

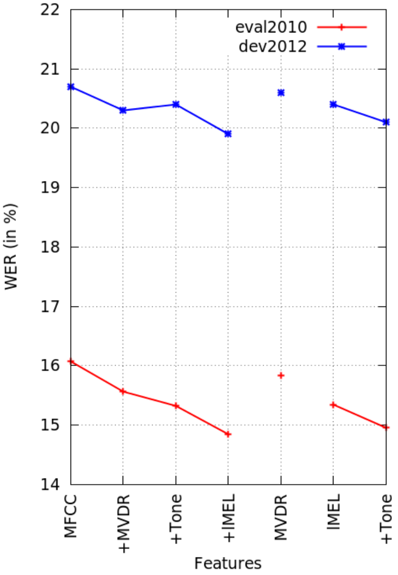




Multi-Feature DBNF

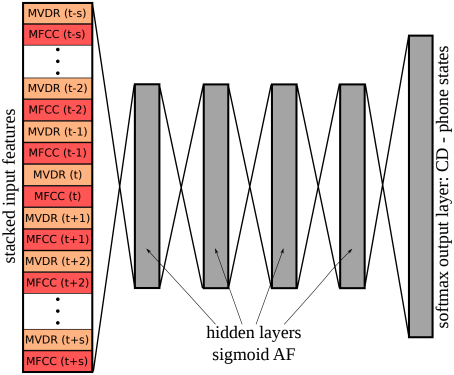


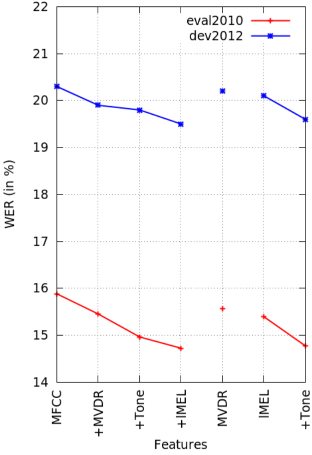
Multi-Feature DBNF Results



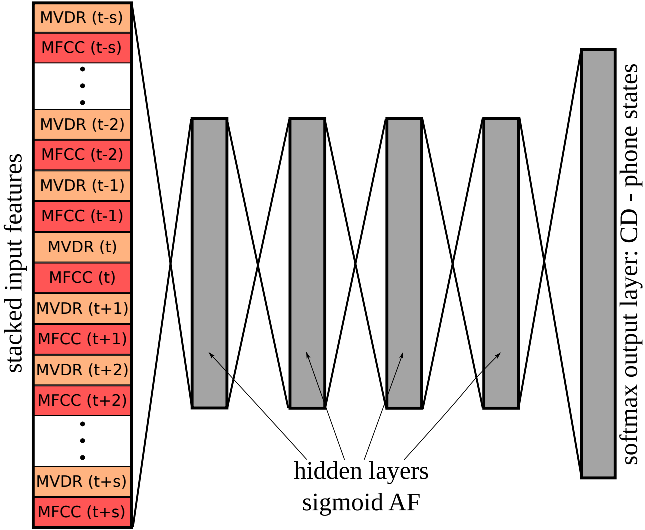
- Multiple input features are beneficial
- Significant improvements on both test sets:
  - dev2012: 0.8% (vs. baseline) & 0.5% (vs. best single feature)
  - eval2010: 1.23% (vs. baseline) & 0.5% (vs. best single feature)

Multi-Feature DNN AM

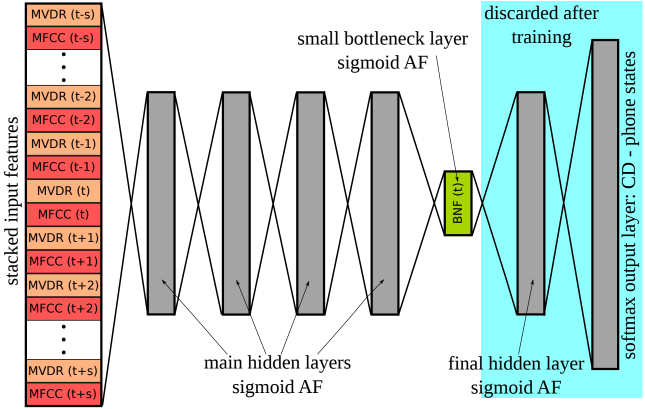




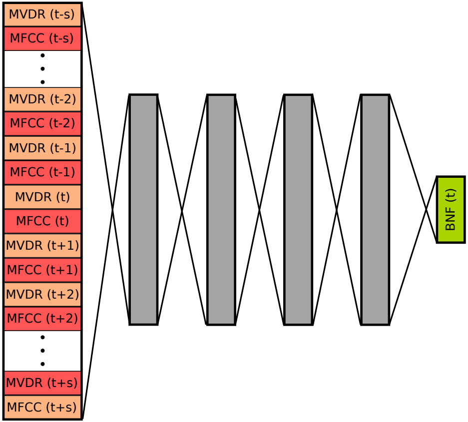
Modular Deep Neural Network Acoustic Models



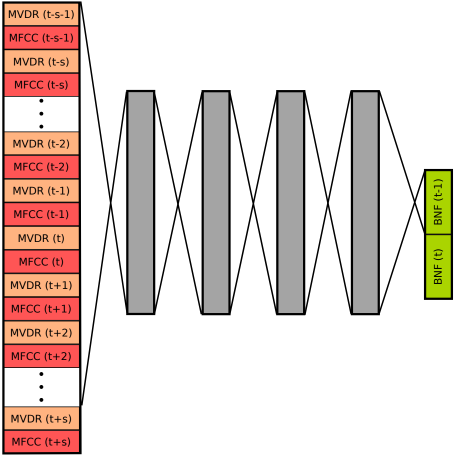
Modular Deep Neural Network Acoustic Models



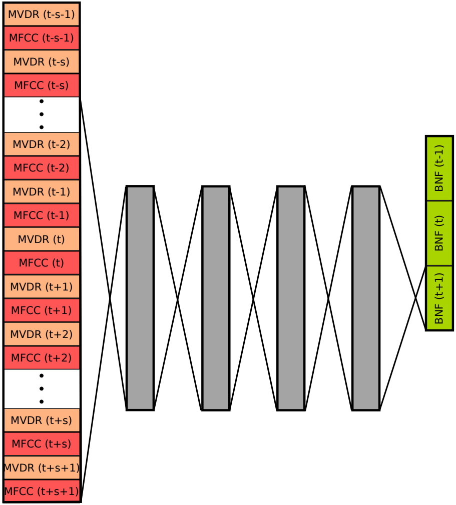
Modular Deep Neural Network Acoustic Models



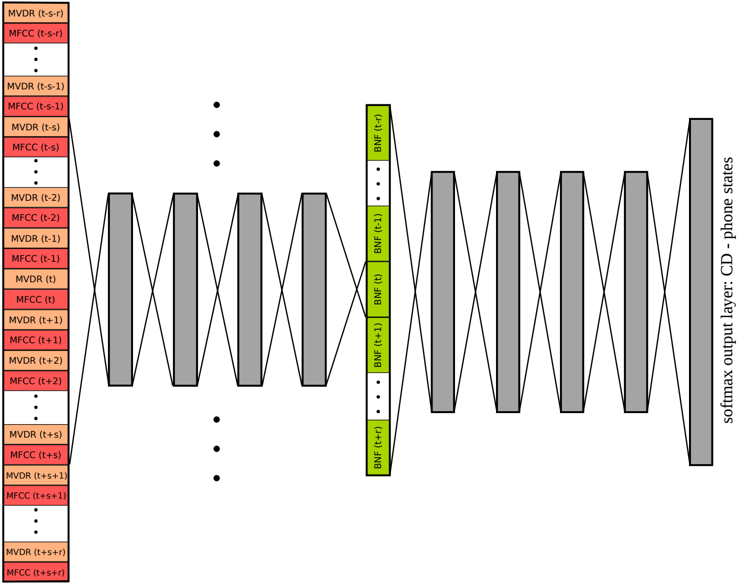
Modular Deep Neural Network Acoustic Models



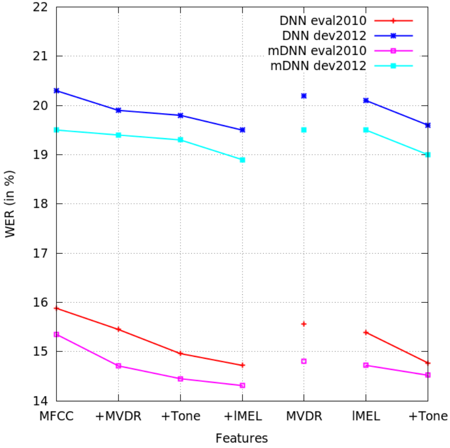
Modular Deep Neural Network Acoustic Models



Modular Deep Neural Network Acoustic Models

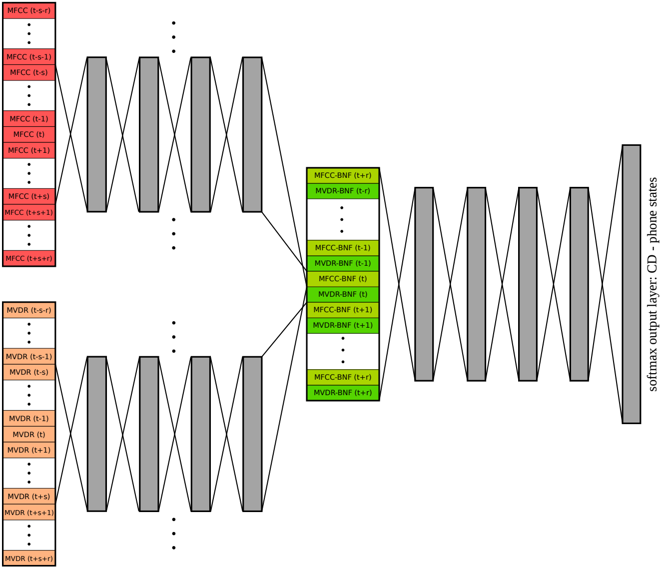


Modular DNN AM Results



	eval2010	dev2012
MFCC	15.35	19.5
+MVDR	14.71	19.4
+Tone	14.54	19.3
+lMEL	**14.31**	**18.9**
lMEL	14.72	19.5
+Tone	14.52	19.0
MVDR	14.81	19.5

mDNN AM with Multiple BNF Modules



mDNN AM with Multiple BNF Modules Results

	BNF modules	eval2010	dev2012
lMEL+Tone	1	*14.52*	*19.0*
MFCC+MVDR+Tone	1	*14.54*	*19.3*
MFCC+MVDR+Tone+lMEL	1	*14.31*	*18.9*
MFCC	1	15.35	19.5
$\oplus$ MVDR	2	14.54	19.2
$\oplus$ lMel	3	14.73	19.3
MFCC $\oplus$ MVDR $\oplus$ lMel+Tone	3	14.24	18.7
lMEL+Tone $\oplus$ MFCC+MVDR+Tone	2	14.19	18.8
$\oplus$ MFCC+MVDR+Tone+lMEL	3	**14.06**	**18.7**
$\oplus$ MFCC $\oplus$ MVDR	5	14.33	18.9
$\oplus$ lMEL $\oplus$ MFCC+MVDR	7	14.44	18.8
lMEL $\oplus$ MFCC+MVDR	2	14.34	19.1

Results Summary

	eval2010	dev2012
baseline MFCC DNN	*15.88*	*20.3*
best single-feature DNN	*15.31*	*20.1*
best DNN system combination (CNC)	*14.45*	*19.2*
best multi-feature DNN	*14.71*	*19.4*
best mDNN with a single module	*14.31*	*18.9*
best mDNN with multiple modules	*14.06*	*18.7*

Conclusion

- DNNs can benefit from multiple input features
- A modular DNN topology can improve its quality
- Multiple feature modules can outperform networks with only a single module
- simply concatenating all features in the input layer is no longer the best approach