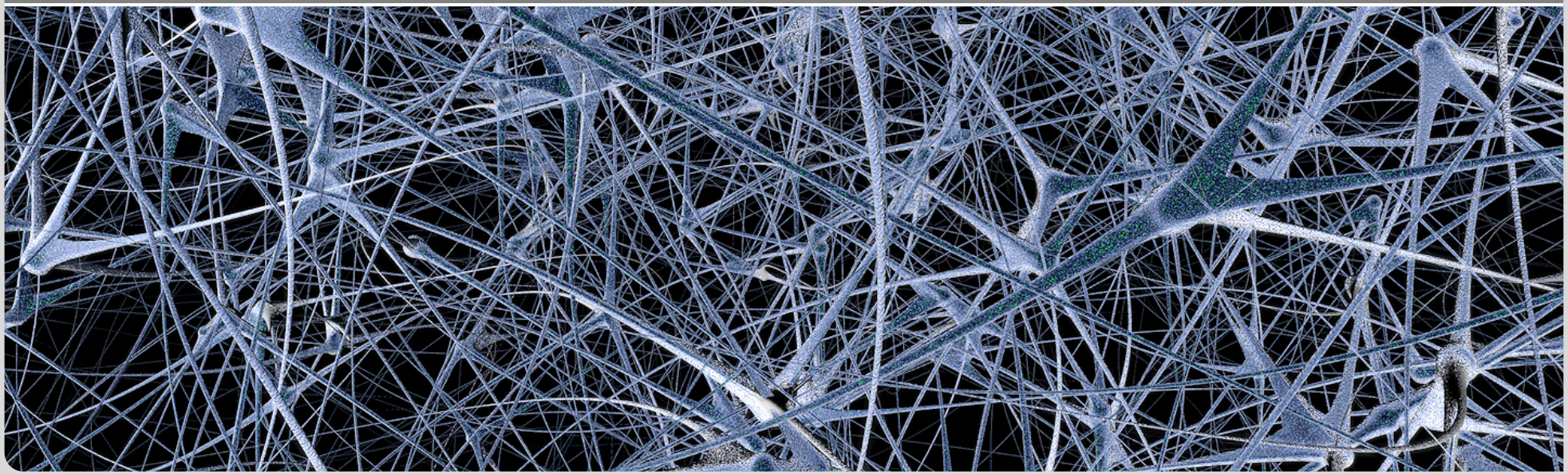
Karlsruher Institut für Technologie

The 2013 KIT IWSLT Speech-to-Text Systems

Kevin Kilgour, **Christian Mohr, Michael Heck, Quoc Bao Nguyen, Van Huy Nguyen, Evgeniy Shin, Igor Tseyzer, Jonas Gehring, Markus Müller, Matthias Sperber, Sebastian Stüker and Alex Waibel**

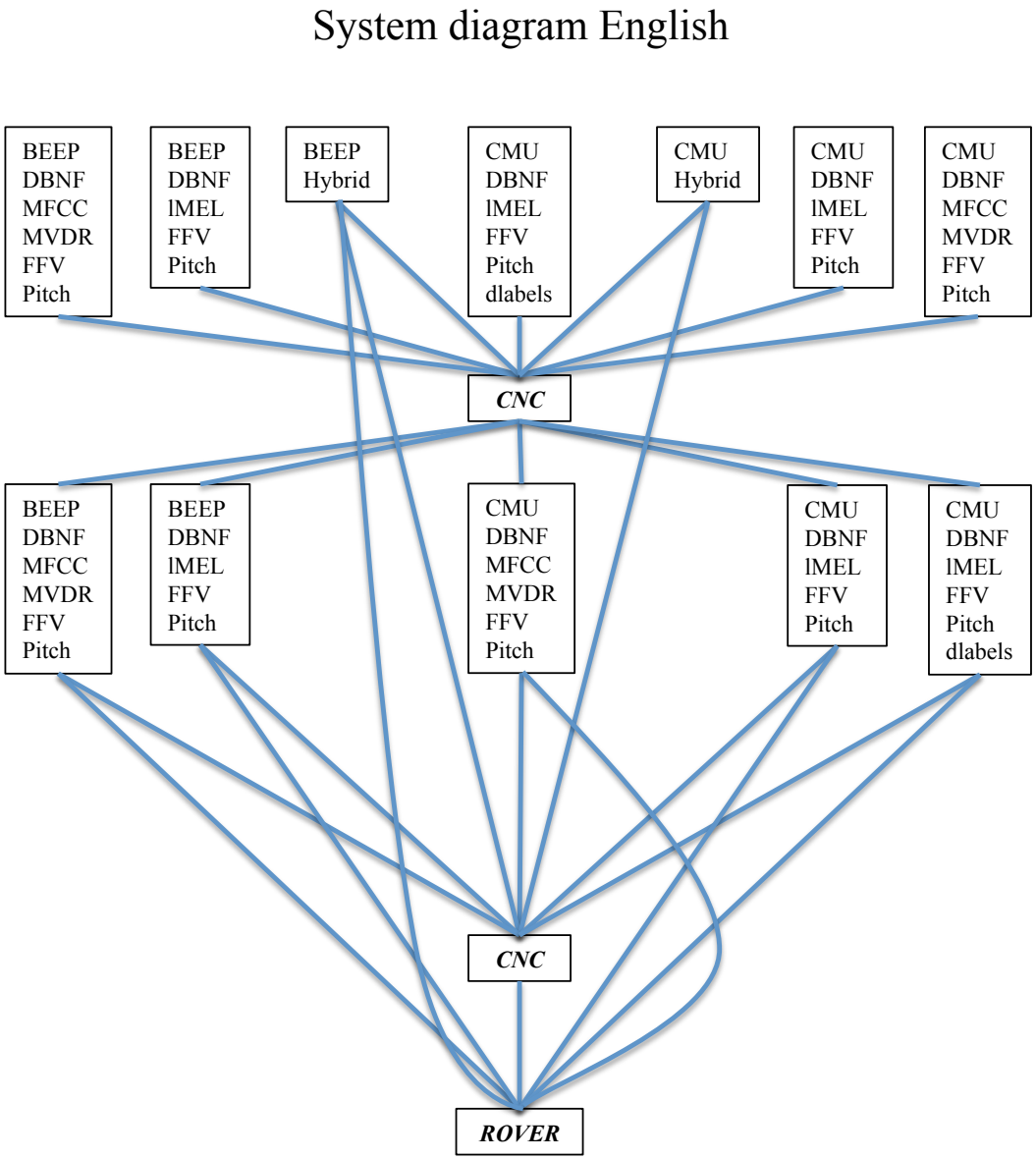
Institut für Anthropomatik, Interactive Systems Lab (Lst. Waibel)



Overview

- Big Picture
- Techniques
- Data
- System setup
- Results

Big Picture


System diagram English
BEEP DBNF MFCC MVDR FFV Pitch
BEEP DBNF lMEL FFV Pitch
BEEP Hybrid
CMU DBNF lMEL FFV Pitch dlabels
CMU Hybrid
CMU DBNF lMEL FFV Pitch
CMU DBNF MFCC MVDR FFV Pitch
CNC
BEEP DBNF MFCC MVDR FFV Pitch
BEEP DBNF lMEL FFV Pitch
CMU DBNF MFCC MVDR FFV Pitch
CMU DBNF lMEL FFV Pitch
CMU DBNF lMEL FFV Pitch dlabels
CNC
ROVER

Techniques

- Segmentation
- MVDR+MFCC BNFs
- DBNFs
- Tonal Features
- Hybrid Systems
- Grapheme Systems
- AM Training
- LM Training
- System Combination

Development Sets

- English:
  - IWSLT dev2012 - 10 talks - 1.7 hours
  - IWSLT eval2011 - 8 talks - 11 hours
  - IWSLT eval2012 - 11 talks - 1.7 hours
- German:
  - IWSLT dev2012 - 7 talks - 1.9 hours

Segmentation

- *Decoder based*
  - fast decoding pass
  - determine speech and non-speech regions
  - consecutively split segments at the longest non-speech region
- *GMM based*
  - speech, non-speech and silence GMM models
  - viterbi decoding
  - post-processing
- *SVM based*
  - frames labeled as speech and non-speech
  - frame stacking and LDA to incorporate temporal information
  - SVM classifier is trained with LIBSVM
  - 2-phase post-processing

SPECOM 2013 - Segmentation of Telephone Speech Based on Speech and Non-Speech Models

Segmentation

- SVM Segmentation is better in German, but slightly worse in English

English

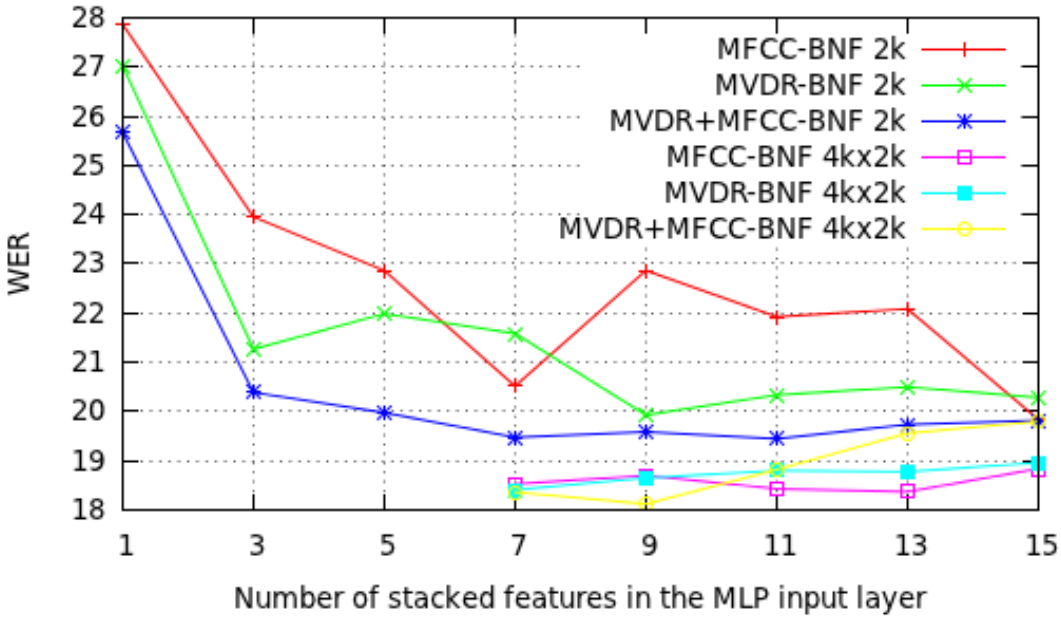
Segmentation	WER	#utt	dur	dur/utt
Manual	13.2%	1144	1.71h	5.4s
Decoder based	13.8%	594	1.83h	11.1s
SVM based	13.9%	431	1.78h	14.9s
GMM based	14.3%	695	1.77h	9.2s

German

Segmentation	WER
Decoder based	22.5%
SVM based	20.5%
GMM based	22.5%

MVDR+MFCC BNFs

- minimum variance distortionless response (MVDR)
- MVDR based BNFs are similar to MFCC based BNFs performance wise
- system combination useful
- idea: combine MFCC and MVDR features at the BNF input level



wMVDR+MFCC BNFs

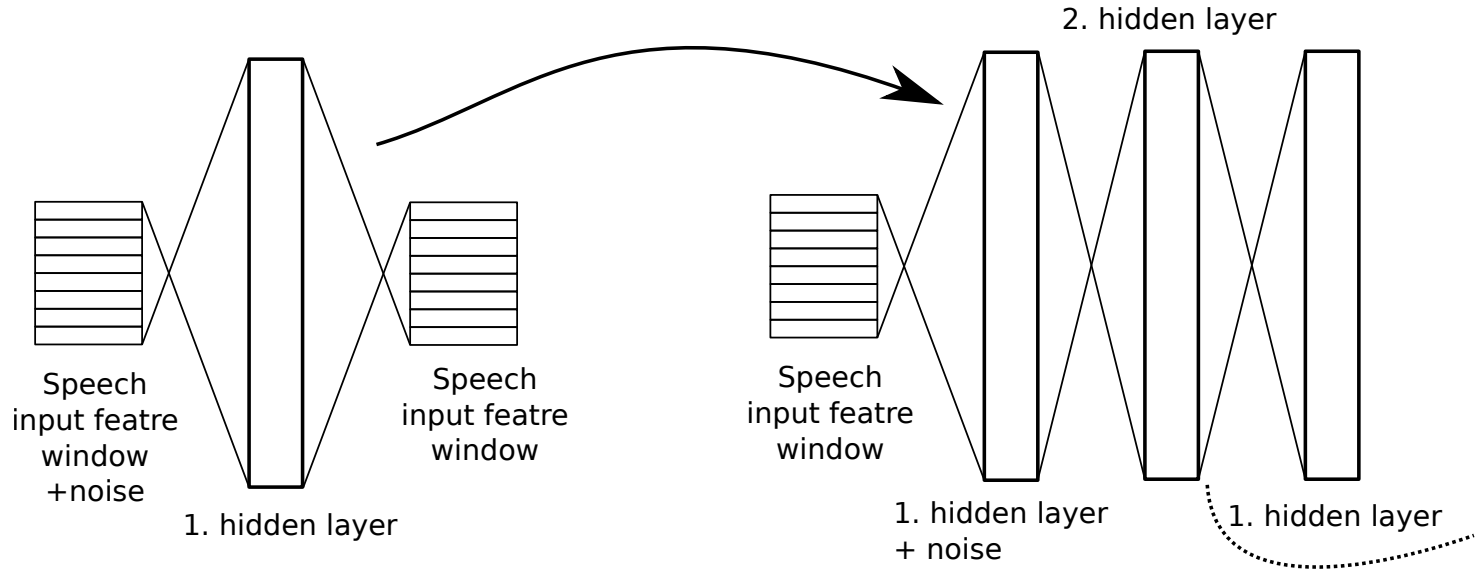
- combined MVDR+MFCC (**M2**) features reduce WER

feature	topology	9 frames	9 frames pretrained	15 frames	15 frames pretrained
MFCC-BNF	4kx2k	18.69	18.53	18.83	18.45
wMVDR-BNF	4kx2k	18.64	18.18	18.95	18.84
wMVDR+MFCC	4kx2k	18.11	17.81	19.79	18.38

ICASSP 2013 - Warped Minimum Distance Distortionless Response Based Bottle Neck Features for LVCSR

DBNFs

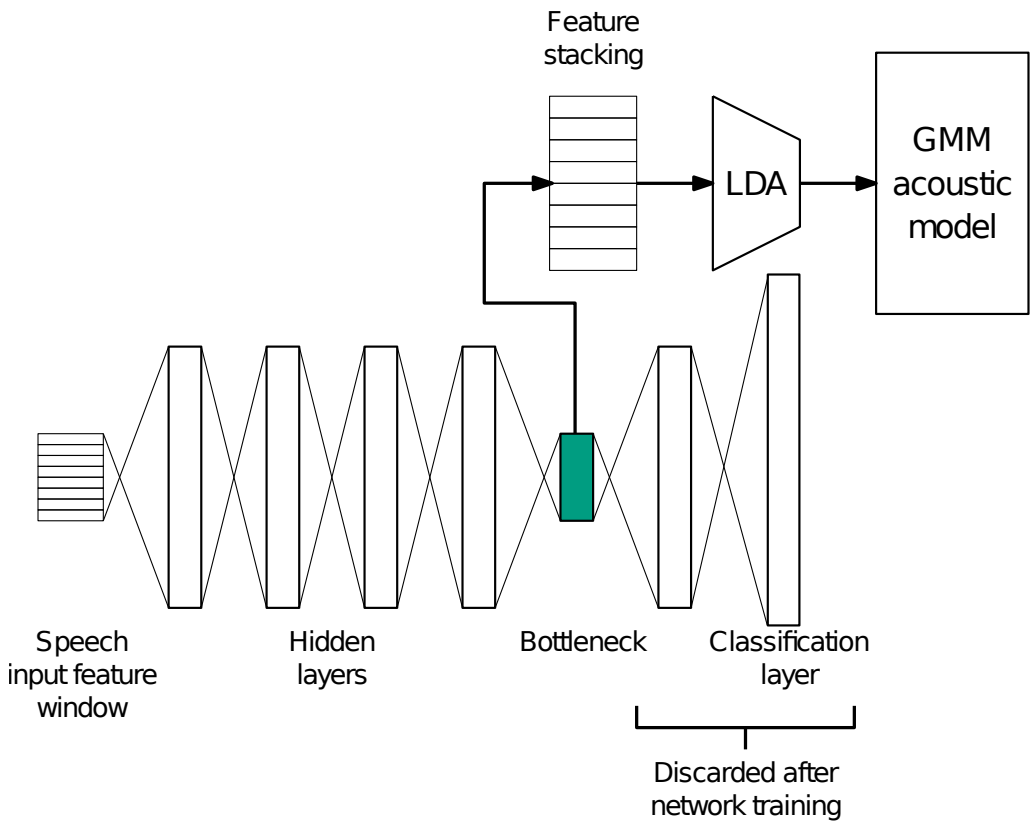
- deep neural networks have shown improvements almost everywhere
- train denoising auto-encoders layerwise



ICASSP 2013 - Extracting Deep Bottleneck Features Using Stacked Auto-Encoders

DBNFs

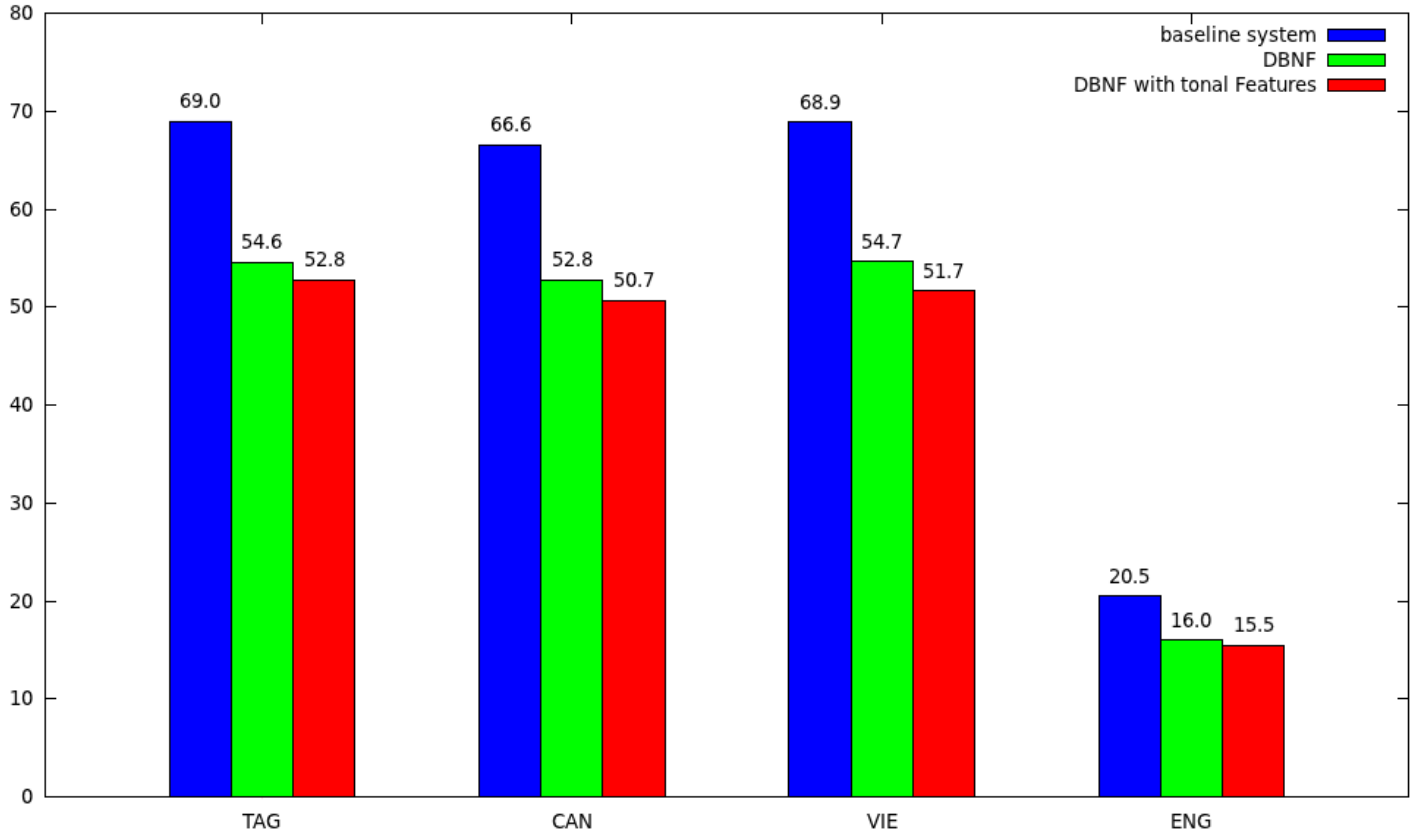
- input: 40 **IMEL** or
  20 MFCC + 20 MVDR (**M2**)
- frame stacking (15 frames)
- 4-6 hidden layers
- 1000 - 1600 nodes each
- output layer: 6000-8000 context dependent phone states



feature	WER
MFCC-BNF	18.53
MVDR-BNF	18.18
MVDR+MFCC BNF	17.81
MVDR+MFCC DBNF	16.41

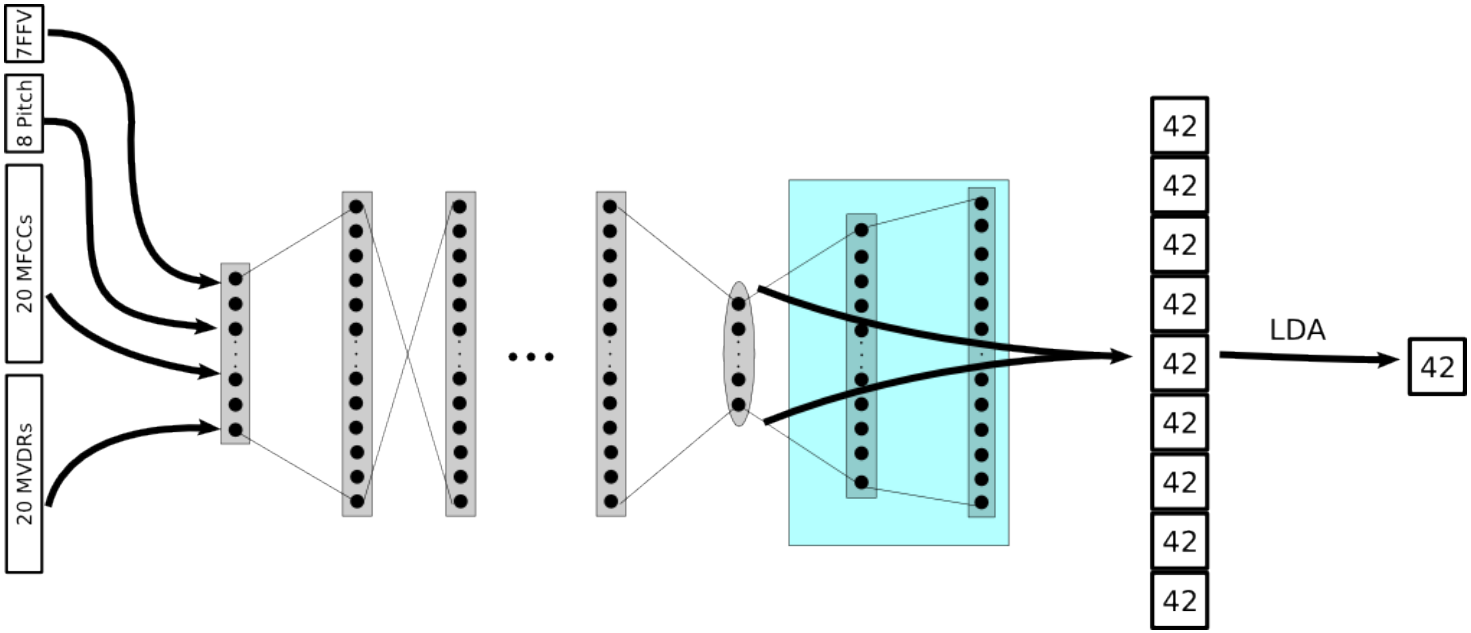
Tonal Features

- work well on tonal languages
- non-tonal languages also improve slightly



Tonal Features

- early DBNF integration works well
- tone features consist of pitch and FFV features

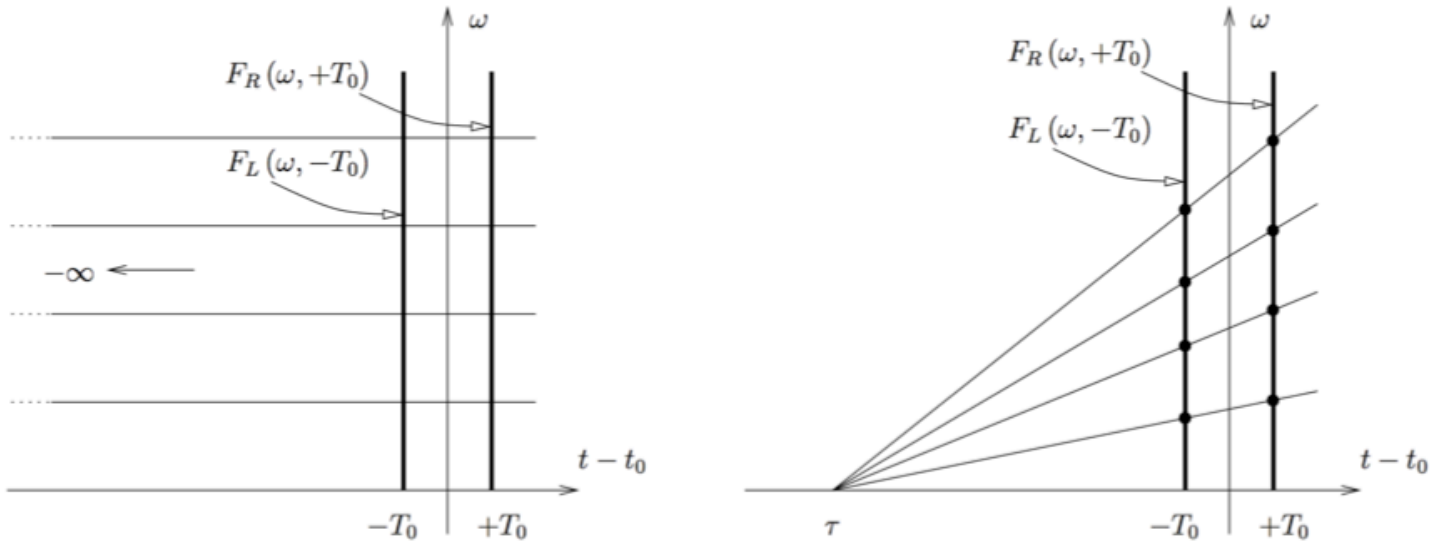


Tonal Features - Pitch

- select pitch candidates from a cepstrum on a 32ms window
- select further candidates from an autocorrelation
- use dynamic programming to pick the best candidate per frame
- add Δ and $\Delta\Delta$ features using the 3 right and 3 left neighbours

Tonal Features - FFV

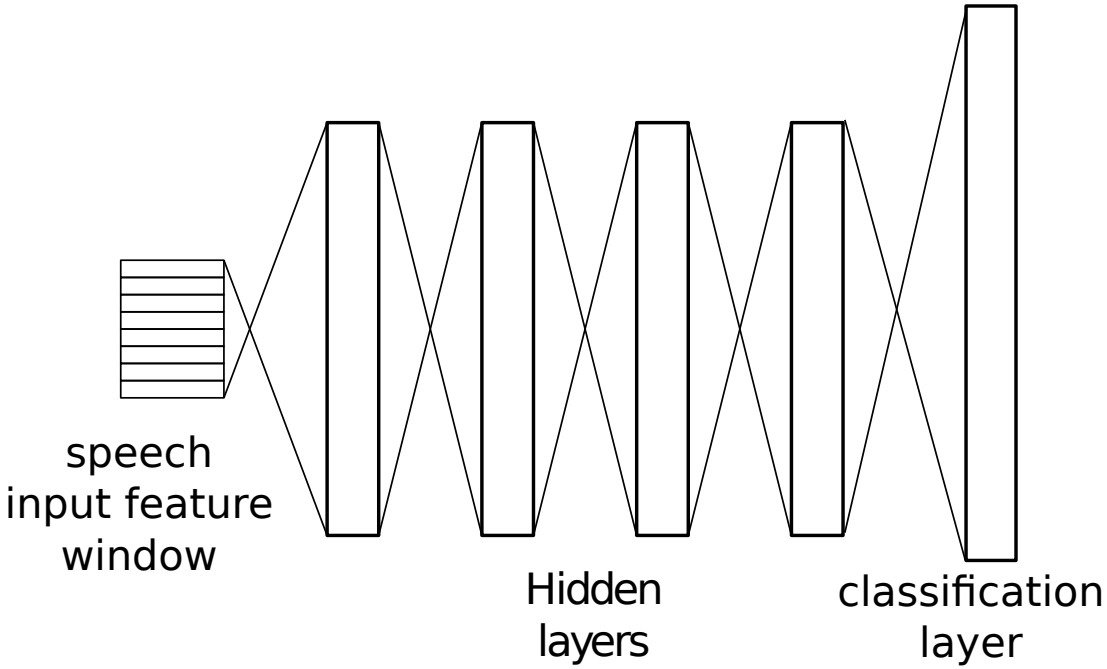
- FFV (Fundamental Frequency Variation)
- compute the vanishing-point products from two neighbouring spectra
- computing it for all τ leads to the FFV spectrum
- apply a filter bank containing 2 rectangular & 5 trapezoidal filters



Visualization of FFV features: The standard dot-product between two vectors is shown as an orthonormal projection onto a point at infinity (left panel), while the "vanishing-point product" for a point τ generalizes to the former when $\tau \to \infty$ (right panel).

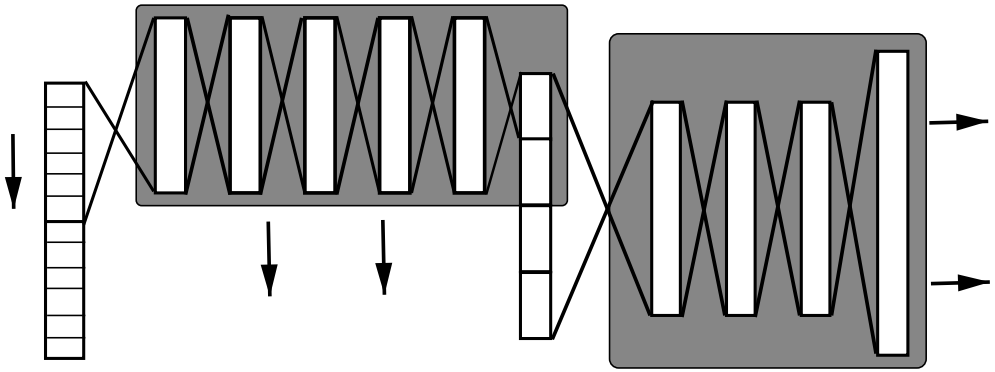
Hybrid systems

- use NN to compute AM scores instead of GMMs
- similar topology to DBNF networks
- input features: IMEL or MVDR+MFCC features



Shifted Hybrid Systems

- train on DBNF features instead of raw features
- German IMEL+T Hybrid 16.63%
- German IMEL+T shifted Hybrid 14.61% (**hyb**)
- optional: joint fine-tuning



ASRU 2013 - Modular Combination of Deep Neural Networks for Acoustic Modeling

Grapheme Systems

- alternative “phoneset”
- poor performance in English
- only slightly worse than a phoneme system in German (20.8% vs 21.7% WER)
- useful for system combination

AM Training

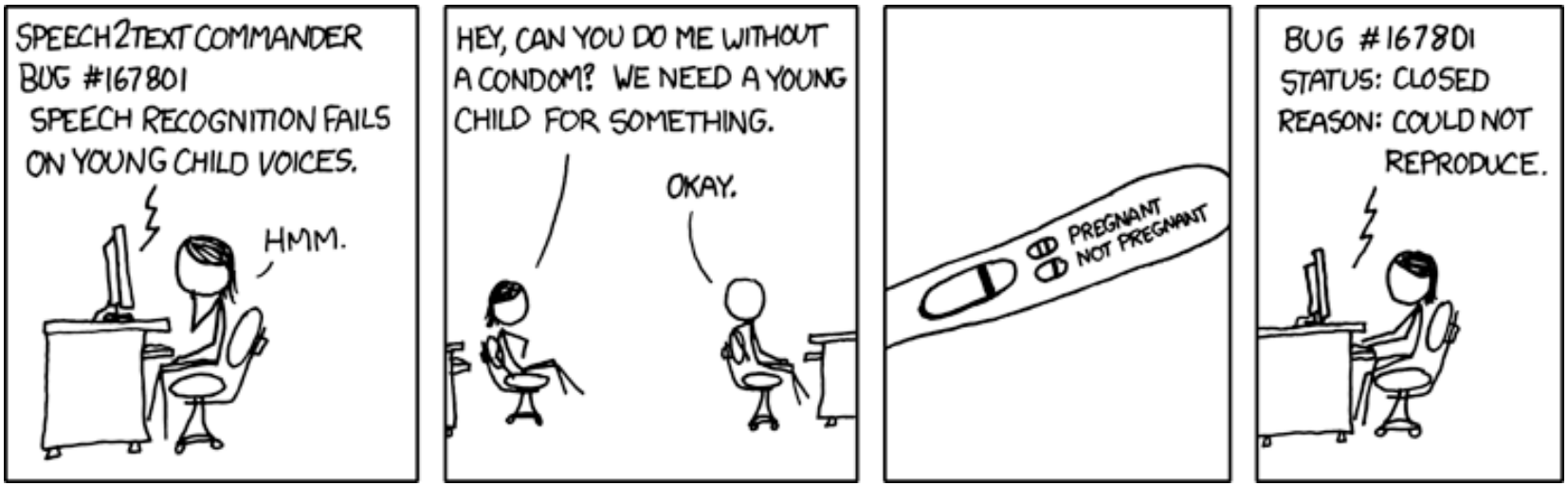
- for GMM Acoustic Models:
- all models use vocal tract length normalization (VTLN).
- incremental splitting of Gaussians training (MAS)
- optimal feature space training (OFS)
  - a *semi-tied covariance* (STC) variant
- one round of Viterbi
- boosted Maximum Mutual Information Estimation training (BMMIE)

LM Training

- tuning text: TED talk transcripts
- vocabulary selection
  - ML selection on the vocabularies of all data sources
  - use G2P tool for words not in the initial dictionary
    - German: Festival + Mary
    - English CMU: Festival
    - English Beep: Sequitur
- train 4gram smoothed LMs for each source
- estimate interpolation parameters
- interpolate into a large LM ($>$10 GBytes)
- use a memory mapped LM

System Combination

- CNC
- Cross Adaptation
- Rover



English Training Data

AM Training Data

- 200 hours of Quaero training data from 2010 to 2012.
- 18 hours of various noise data, such as applause and music.
- 158 hours of data downloaded from the TED talks website

LM Training Data

Text corpus	# Words
TED	3M
News + News commentary	2,114M
GIGA parallel	523M
Gigaword 4	1,800M
UN + Europarl	376M
Google Books Ngrams (subset)	(1000M ngrams)

Dictionary

- **CMU** Dictionary: American English
- **BEEP** Dictionary: British English

German Training Data

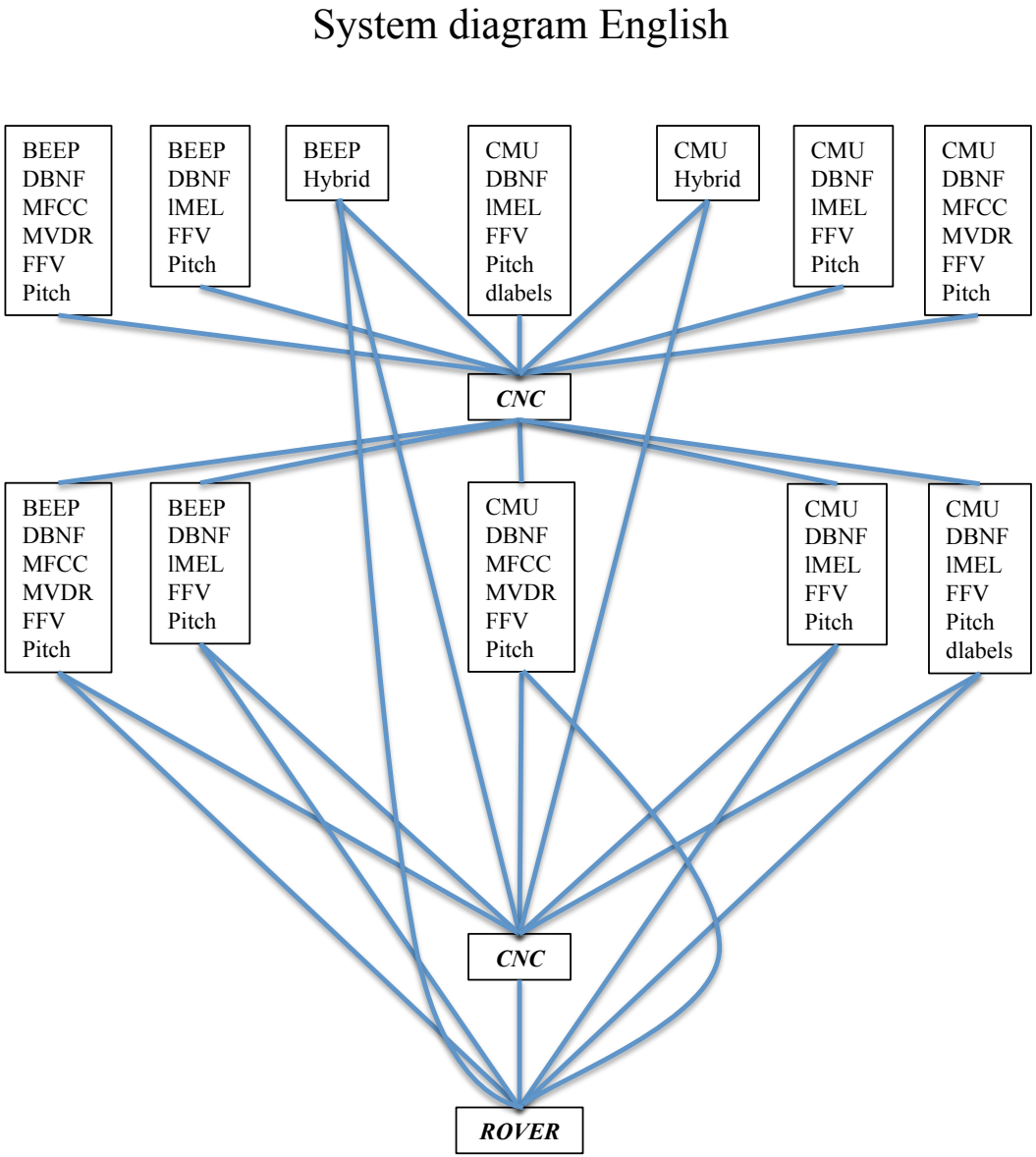
AM Training Data

- 179 hours of Quaero training data from 2010 to 2012.
- 24 hours of broadcast news data

LM Training Data

Text corpus	# Words
TED (translated)	2,259k
Callhome	150k
Europarl	47,306k
HUB5	19k
MultiUN	5,849k
News+News Commentary	284,415k
ECI	12,652k
Euro Language Newspaper	86,785k
German Political Speeches	5,514k
Common Crawl	47,046k
Google Web Ngrams	1.3T

Decoding Strategy



German Systems

System	Dev2013	Eval2013*
M2-P	21.00	29.40
M2+T-P	20.80	30.80
M2+T-G	21.70	29.80
M2-hyb-P	21.40	30.50
IMEL+T-P	21.10	29.70
IMEL-hyb-P	20.20	29.20
M2-G	22.90	30.70
CNC-01	18.60	26.70
M2-P	19.90	27.90
M2+T-P	19.60	27.80
M2+T-G	20.50	27.90
IMEL+T-P	20.10	27.80
M2-G	21.70	29.00
CNC-02	18.30	26.40
ROVER	18.30	26.30

*unoffical: using the adjudication stm and glm files

English Systems

System	Dev2012	Eval2011	Eval2012
M2+T-CMU	15.9	11.6	11.7
IMEL+T-CMU	16.1	11.4	11.4
M2+T-DLabel-CMU	15.8	11.2	11.5
M2+T-BEEP	16.2	12.0	12.6
IMEL+T-BEEP	16.1	12.2	12.6
M2+T-hyb-CMU	16.5	11.9	11.6
M2+T-hyb-BEEP	16.9	12.4	12.4
CNC-BEEP-01	13.7	9.8	9.5
M2+T-CMU	14.7	10.3	10.3
IMEL+T-CMU	15.0	10.2	10.1
M2+T-DLabel-CMU	14.5	10.3	10.1
M2+T-BEEP	14.7	10.8	10.5
IMEL+T-BEEP	14.4	10.6	10.6
CNC-BEEP-02	13.3	9.3	9.2
ROVER	13.3	9.2	9.0

Results

- English:
  - eval2011: 9.3% (KIT 2012: 12.0%)
  - eval2012: 9.6% (KIT 2012: 12.4%)
  - eval2013 primary: 14.4%
  - eval2013 contrastive1: 14.3%
- German:
  - eval2013 primary: 25.7%
  - eval2013 contrastive1: 24.7%
  - eval2013 contrastive2: 25.8%