The Munich 2011 CHiME Challenge Contribution: BLSTM-NMF Speech Enhancement and Recognition for Reverberated Multisource Environments

Felix Weninger, Jürgen Geiger, Martin Wöllmer, Björn Schuller, Gerhard Rigoll

Institute for Human-Machine Communication, Technische Universität München

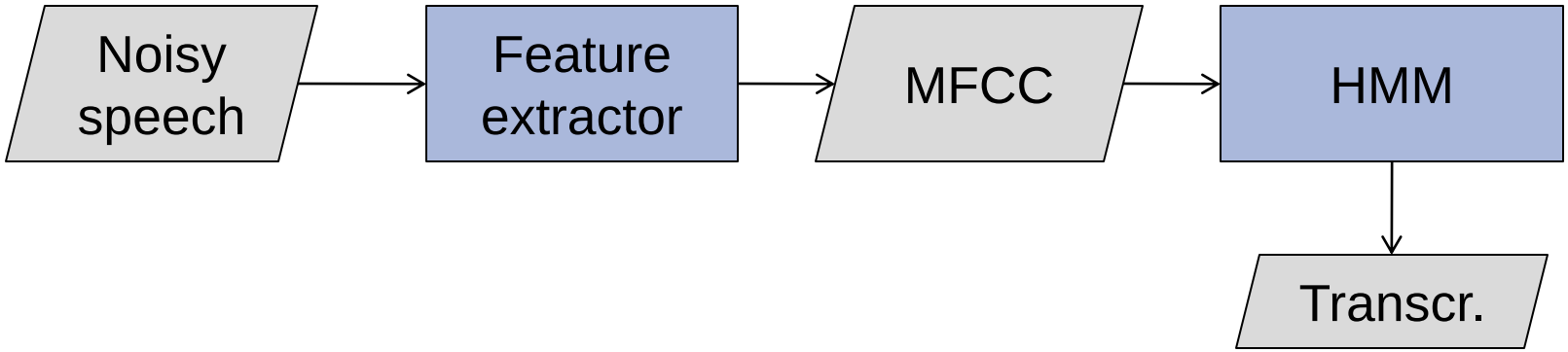
September 1st, 2011


+Gerhard Rigoll
Björn Schuller
Martin Wöllmer
Jürgen Geiger
Felix Weninger

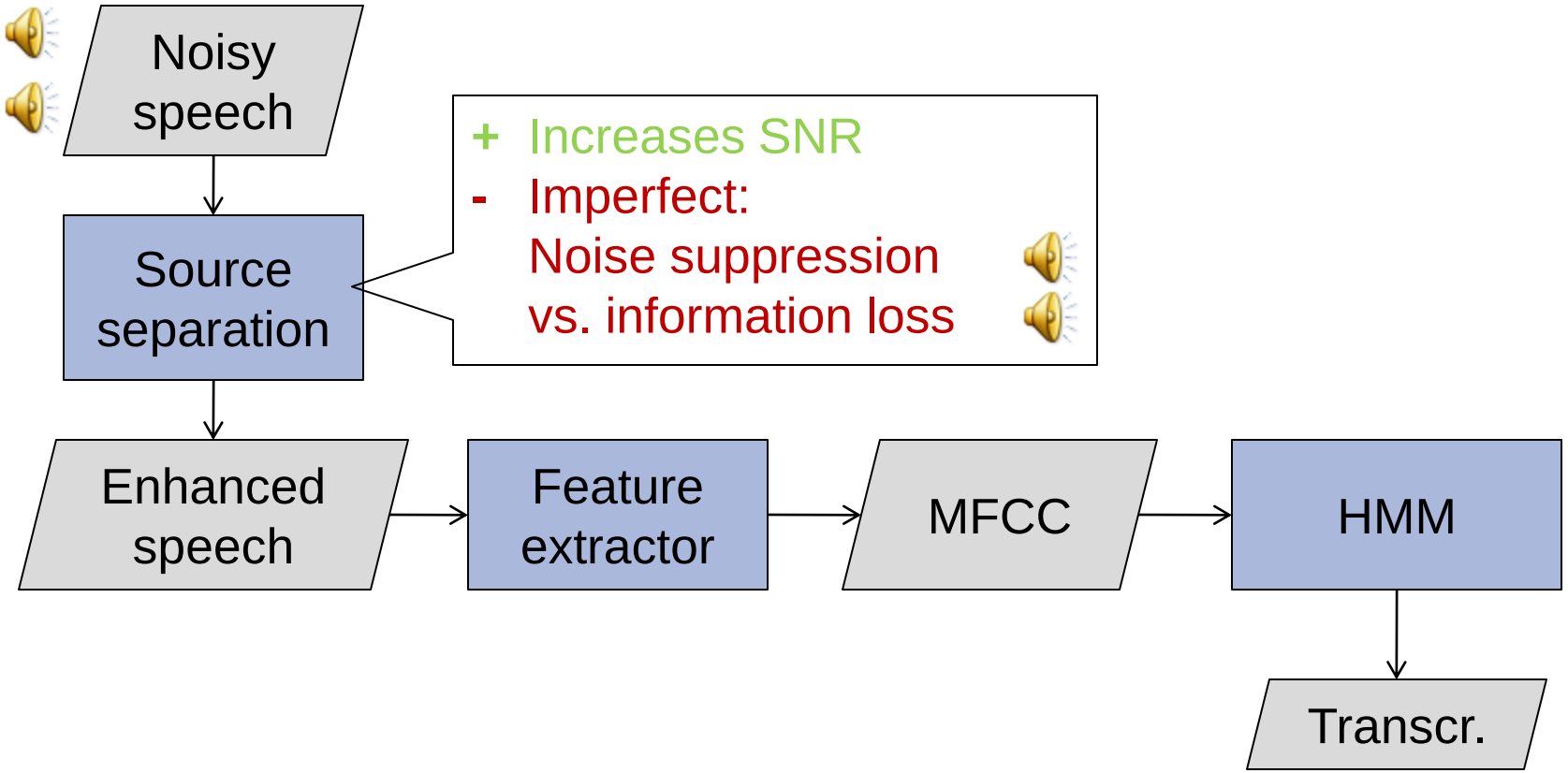
Outline

- Motivation
- Our ASR Architectures:
  - Speech Enhancement by Convolutive NMF
  - BLSTM Speech Recognition
  - Single- and Multi-Stream Recognisers
- Development Results
- Our Final Challenge Result
- Outlook

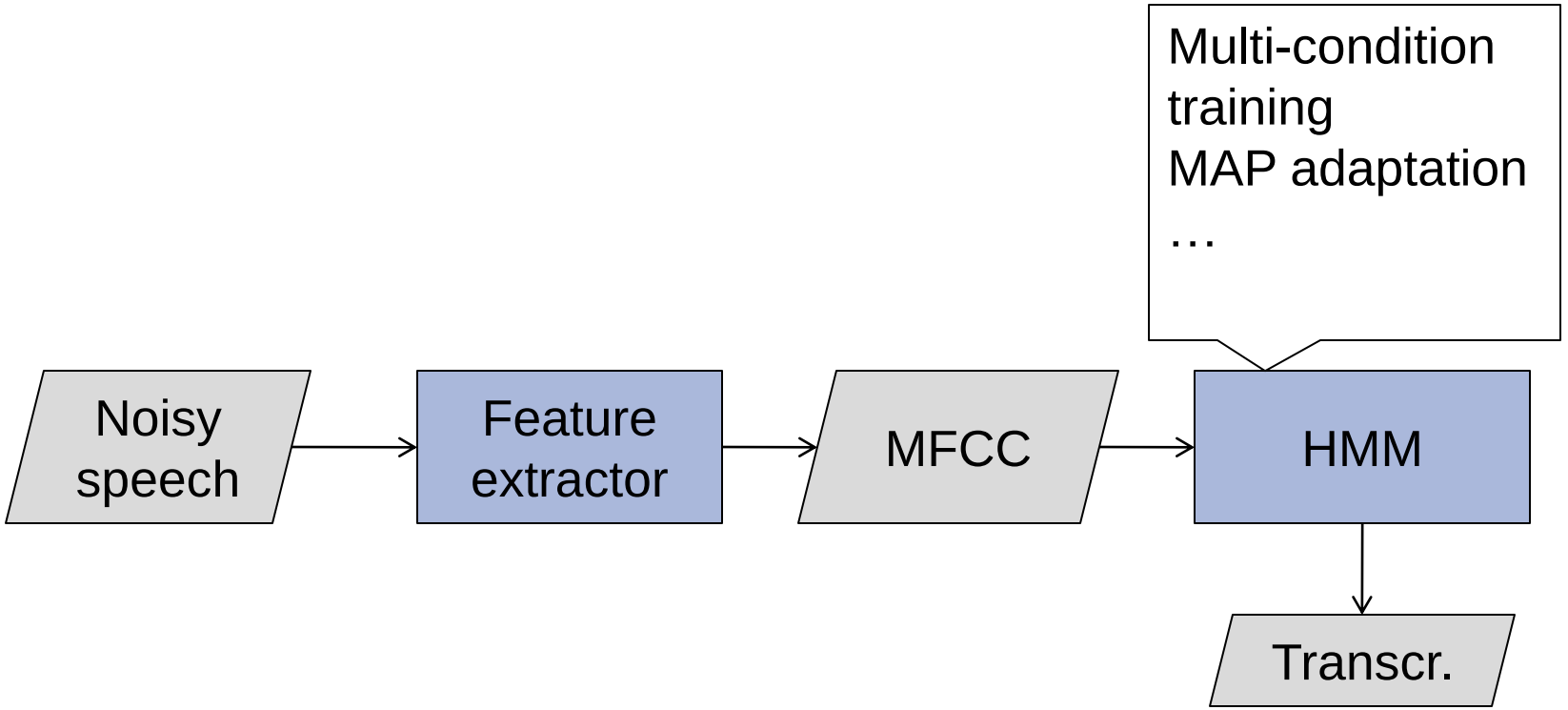
ASR in Noisy Conditions



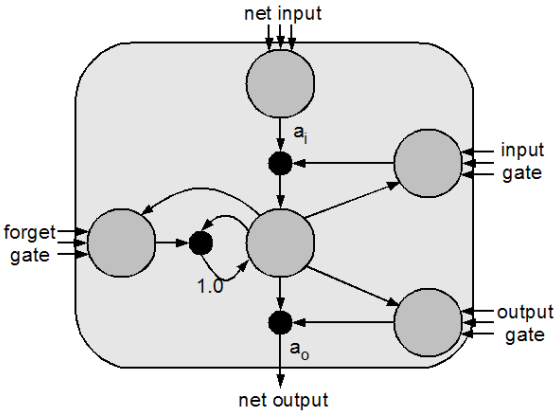
Solution 1: Front-End Enhancement



Solution 2: Robust Back-Ends



Solution 2: Robust Back-Ends



BLSTM-RNN

Word prediction

Noisy speech

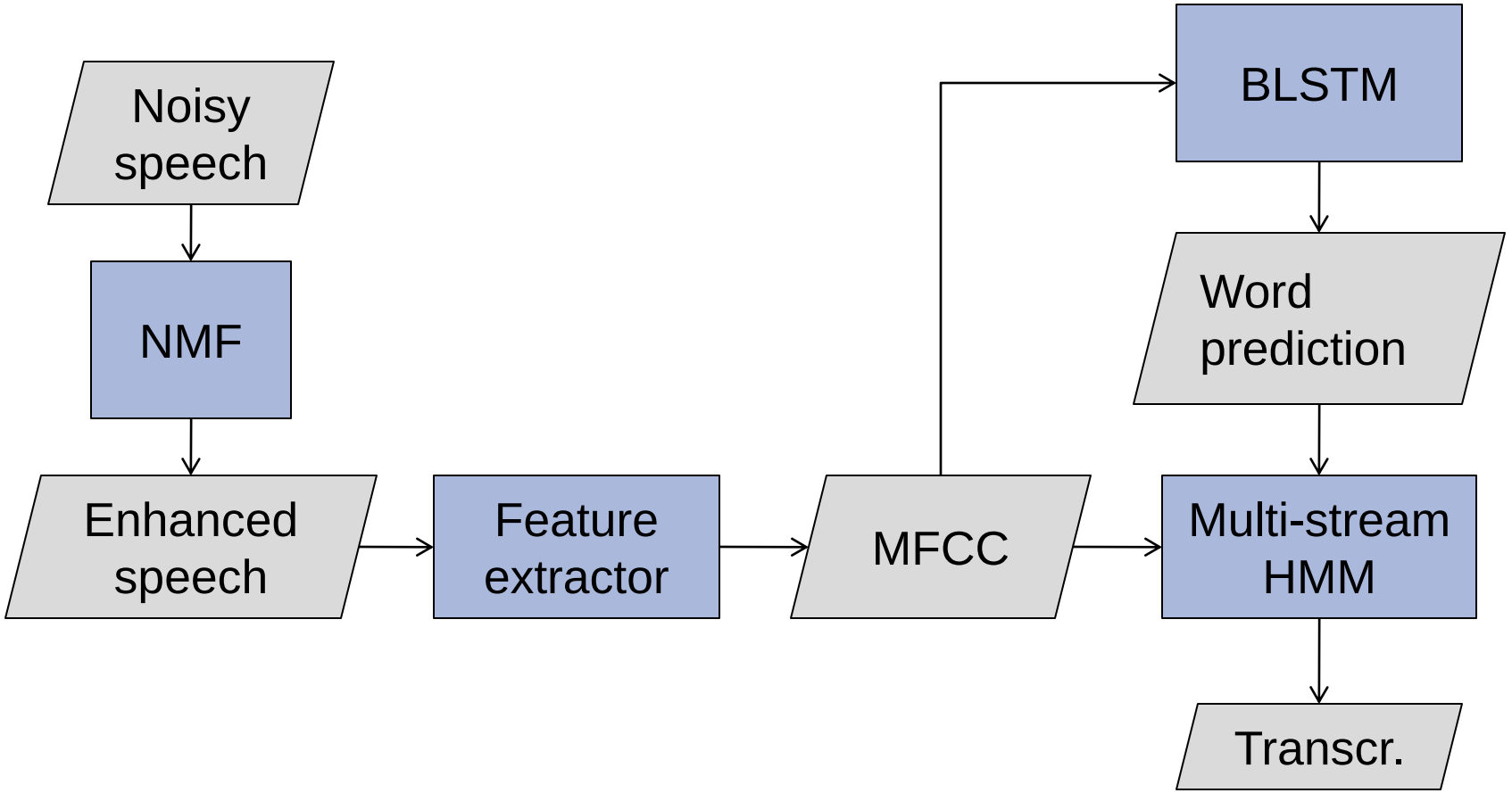
Feature extractor

MFCC

Multi-stream HMM

Transcr.

Proposed ASR Architecture



Speech Enhancement: Convolutive NMF

- Assumption of additive noise
- Observed magnitude spectrogram = Convolution of
  - Speech and noise spectrograms
    - P = 13 frames @ 64 ms frame size, 16 ms shift = 256 ms
  - Non-negative activations
- Dictionaries (`bases') of speech and noise computed from training data

Convolutive signal model

- Modelling of true speech spectrogram:

$$\mathbf{V}^{(s)}_{:,t} \approx \sum_{j=1}^{R^{(s)}} \sum_{p=1}^{\min\{P,t\}} \mathbf{H}^{(s)}_{j,t-p+1} \mathbf{X}^{(s)}_{:,p}(j)$$

- Modelling of true noise spectrogram:

$$\mathbf{V}^{(n)}_{:,t} \approx \sum_{j=1}^{R^{(n)}} \sum_{p=1}^{\min\{P,t\}} \mathbf{H}^{(n)}_{j,t-p+1} \mathbf{X}^{(n)}_{:,p}(j)$$

- $R^{(s)}$, $R^{(n)}$ = 51 (102 NMF “components”)

Speech Enhancement: Convolutive NMF

- Matrix formulation:

$$\begin{aligned}\mathbf{V} &\approx \mathbf{\Lambda}^{(s)} + \mathbf{\Lambda}^{(n)} \\ &= \sum_{p=0}^{P-1} \mathbf{W}^{(s)}(p) \overset{p\rightarrow}{\mathbf{H}}^{(s)} + \sum_{p=0}^{P-1} \mathbf{W}^{(n)}(p) \overset{p\rightarrow}{\mathbf{H}}^{(n)}\end{aligned}$$

- Determine $\mathbf{H}^{(s)}$, $\mathbf{H}^{(n)}$ by multiplicative updates
  - Minimize KL divergence $d(\mathbf{V}, \mathbf{\Lambda}^{(s)}+\mathbf{\Lambda}^{(n)})$

- Estimate $\widehat{\mathbf{V}}^{(s)} = \frac{\mathbf{\Lambda}^{(s)}}{\mathbf{\Lambda}^{(s)} + \mathbf{\Lambda}^{(n)}} \otimes \mathbf{V}$ (soft masking)

Convolutive Speech and Noise Bases

- Speaker-dependent **speech bases**:
  - Convolutive NMF on training set for speakers *k* and words *w*,

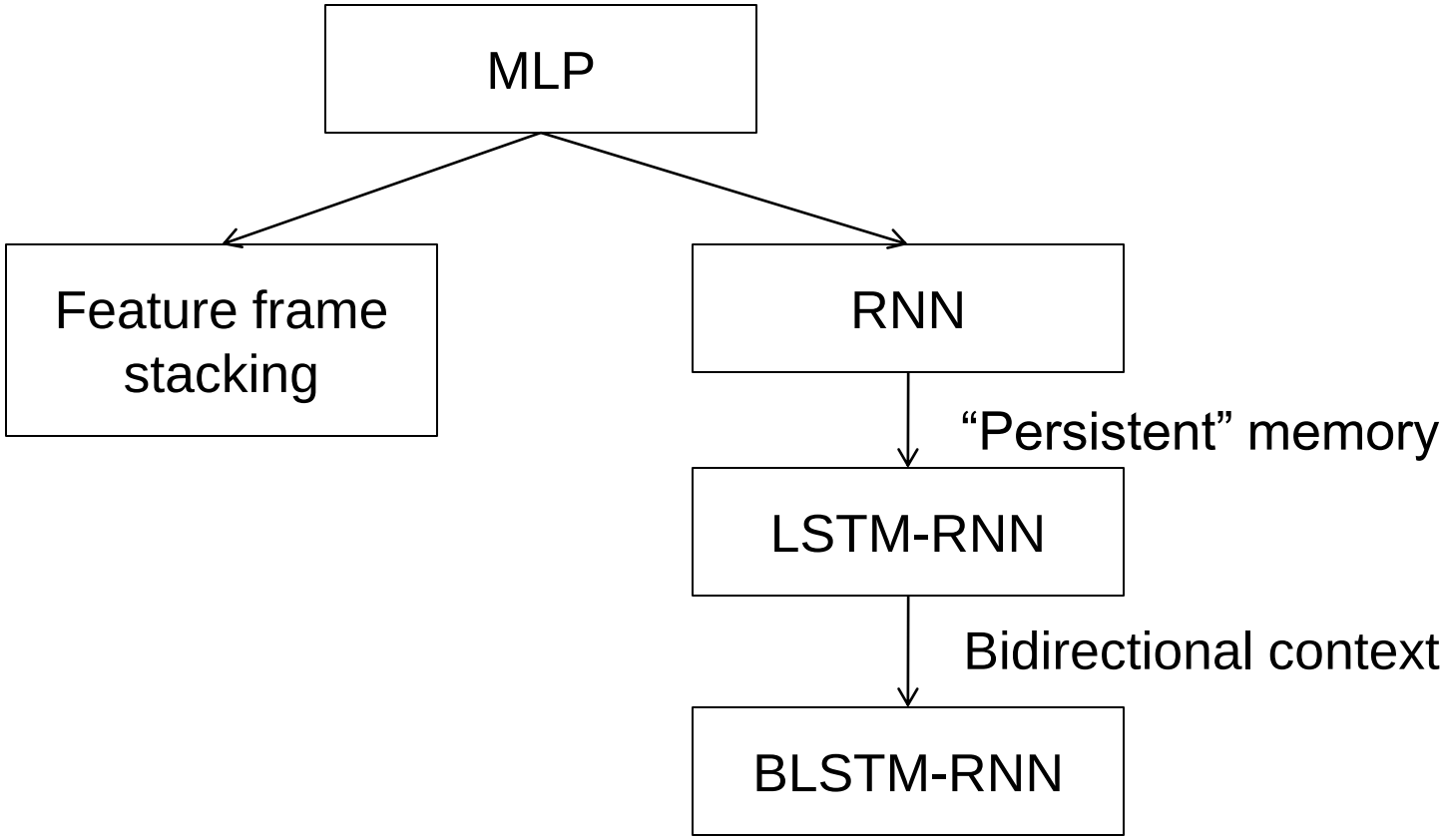
$$\mathbf{T}^{(s,k,w)} \approx \sum_{p=0}^{P-1} \mathbf{w}^{(s,k,w)}(p)\mathbf{h}^{(s,k,w)}_{p\rightarrow}$$

  - Build $\mathbf{W}^{(s,k)}(p) = [\mathbf{w}^{(s,k,1)}(p) \ \cdots \ \mathbf{w}^{(s,k,51)}(p)]$

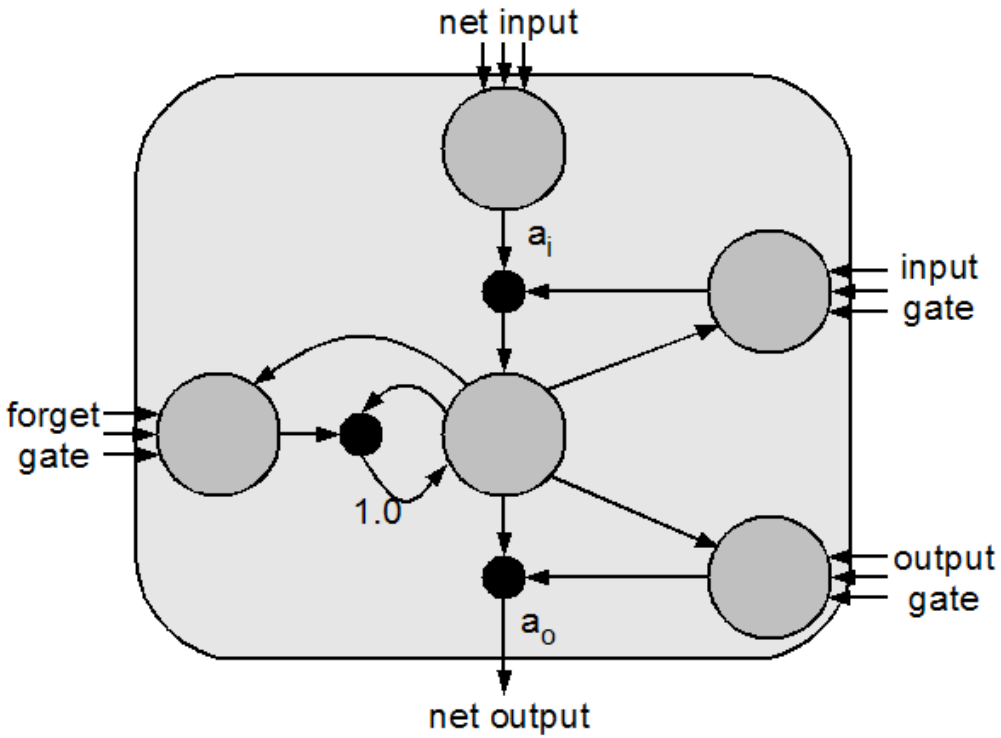
- General **noise base**:
  - Sub-sample training noise
  - Build $\mathbf{W}^{(n)}(p)$ by convolutive NMF

Back-End:
Multi-stream Tandem BLSTM-HMM

Context Modelling in Neural Networks



Word Predictions by BLSTM-RNNs



- Bi-directionally context-sensitive prediction
- Amount of context learned automatically during training
- Superior to (R)NN feature frame stacking [Woellmer, 2011]

BLSTM Training and Classification

- Dimension:
  - 39 input units (one per feature)
  - 3 hidden layers per direction (78 / 150 / 51 LSTM units)
  - 51 output units (one per word)
- Training:
  - Frame-wise word targets by forced alignment
  - Early stopping strategy (use best network on development set)
- Classification:
  - Input: (NMF-enhanced) speech
  - Output: Index of output unit with highest activation

Multi-Stream Hidden Markov Modelling

- GMM (M=7 mixtures) for MFCCs $\boldsymbol{x}_t$
- CPT for discrete BLSTM word prediction b_t
- Mitigate BLSTM misclassifications by Viterbi decoding
- HMM emission probability in state s_t:

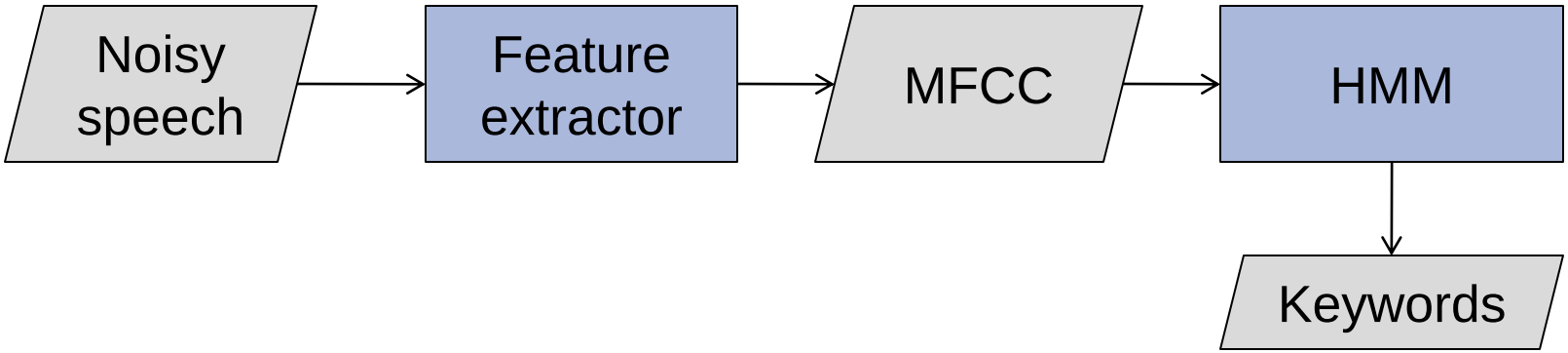
$$p(\mathbf{y}_t|s_t) = \left[\sum_{m=1}^{M} c_{s_t m}\mathcal{N}(\mathbf{x}_t;\boldsymbol{\mu}_{s_t m},\boldsymbol{\Sigma}_{s_t m})\right]^{a} \times p(b_t|s_t)^{2-a}$$

- MFCC stream weight a = *1.3* (tuned on devel. set)
- Superior to GMM feature fusion [Woellmer, 2011]

Results [Development Set]

CHiME baseline:

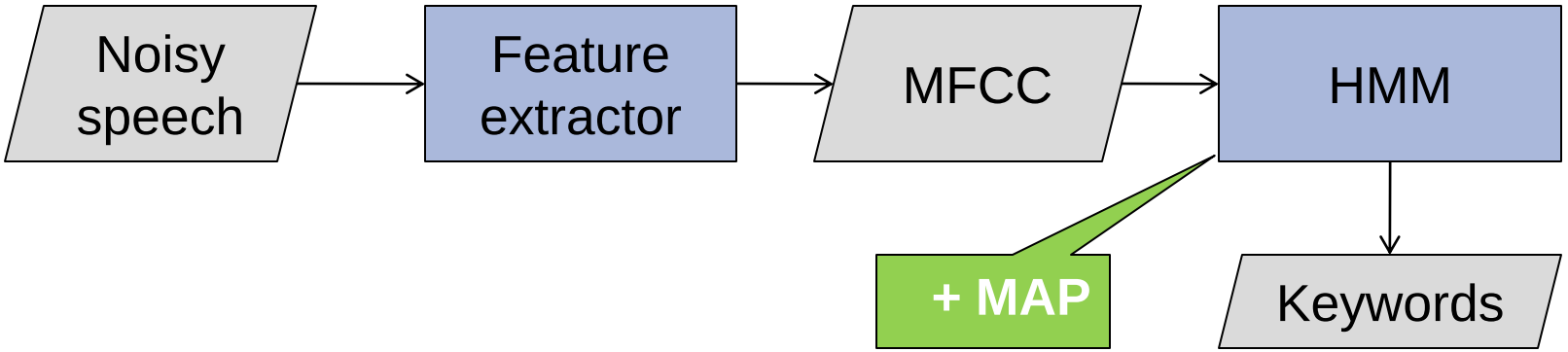
-6 dB	-3 dB	0 dB	3 dB	6 dB	9 dB	Mean
31.1	36.8	49.1	64.0	73.8	83.1	**56.3**



Results [Development Set]

With MAP speaker adaptation:

-6 dB	-3 dB	0 dB	3 dB	6 dB	9 dB	Mean
46.6	52.1	63.8	74.6	82.3	89.0	**68.1**

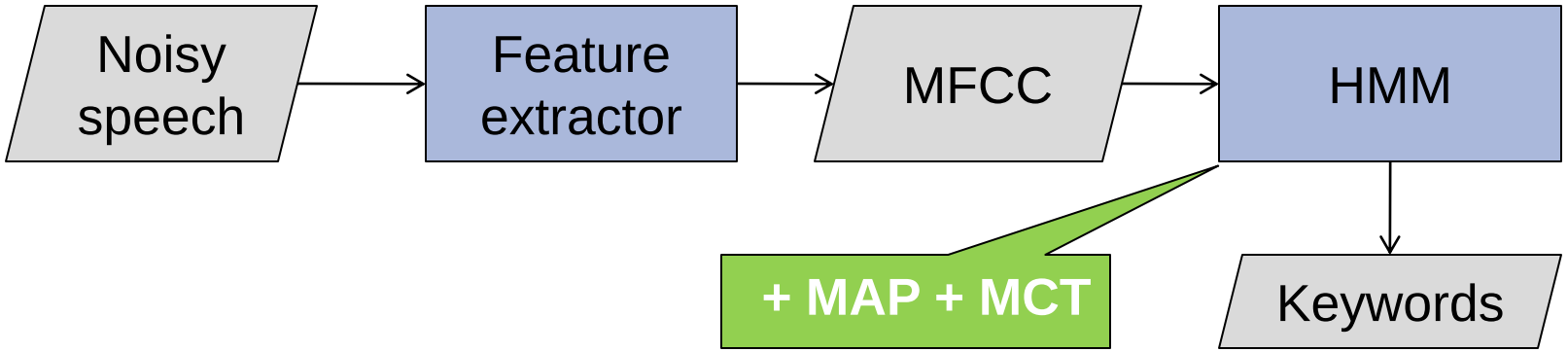


Results [Development Set]

With MAP and multi-condition training:

-6 dB	-3 dB	0 dB	3 dB	6 dB	9 dB	Mean
54.8	62.4	72.0	80.5	87.0	90.8	**74.6**

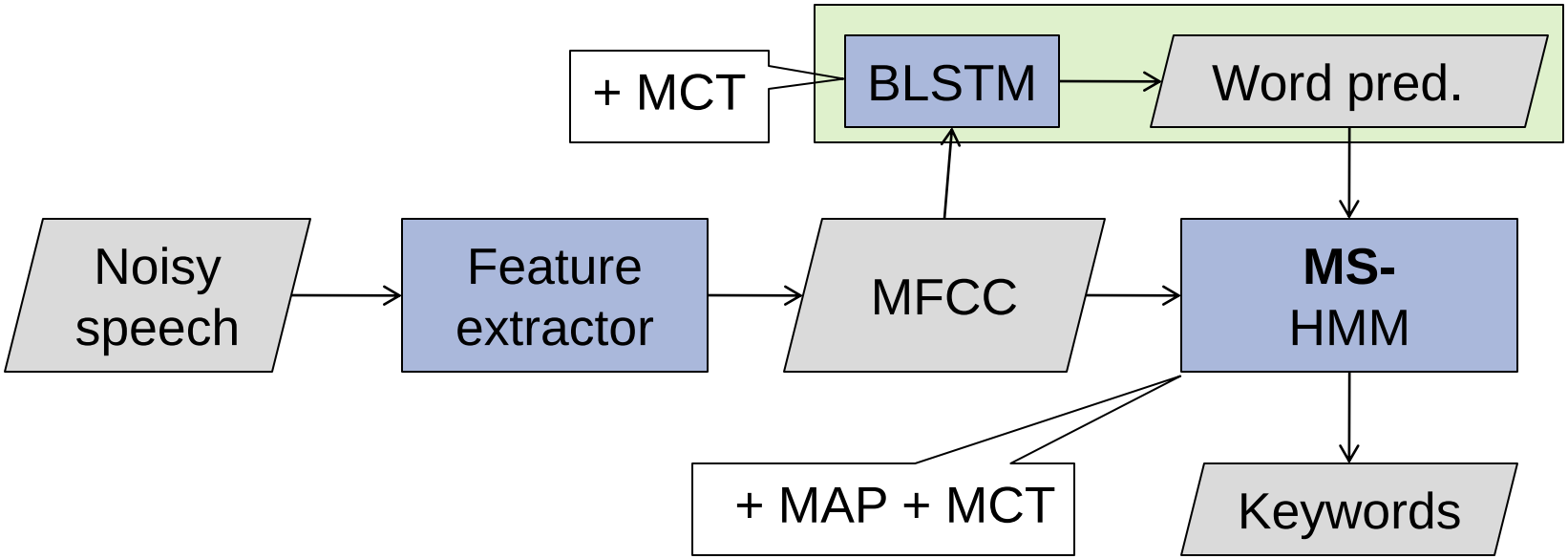
- Noise-free training set overlaid with CHiME training noise
- Select random segments to provide various SNRs
- Include noise in MAP



Results [Development Set]

Multi-stream HMM recogniser:

-6 dB	-3 dB	0 dB	3 dB	6 dB	9 dB	Mean
69.8	75.8	83.7	88.8	92.6	94.8	**84.2**

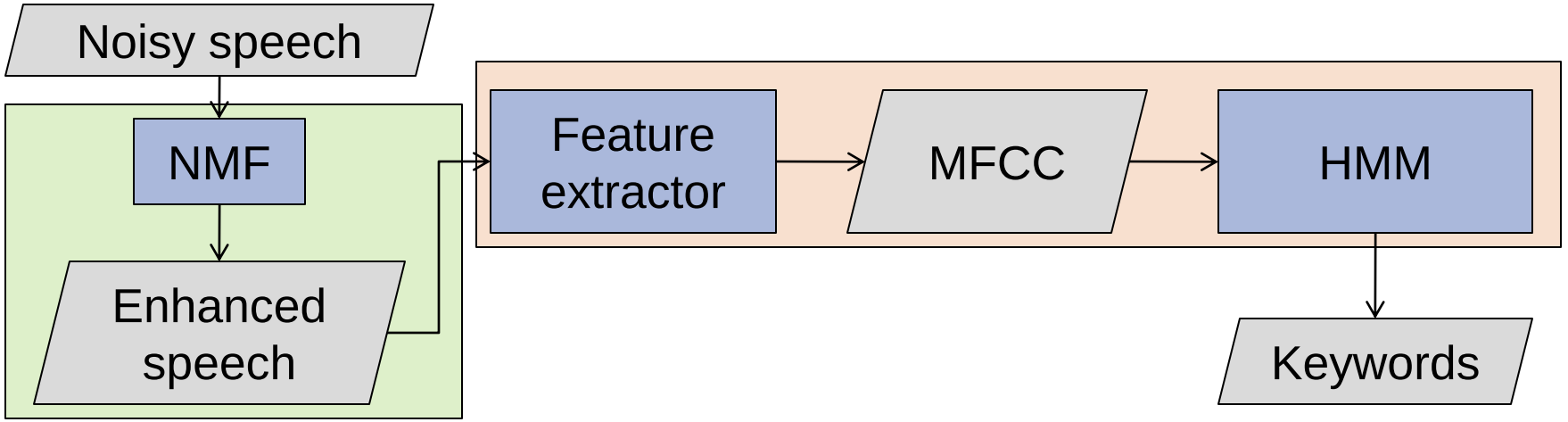


… What about Speech Enhancement?

Results [Development Set]

Baseline recogniser:

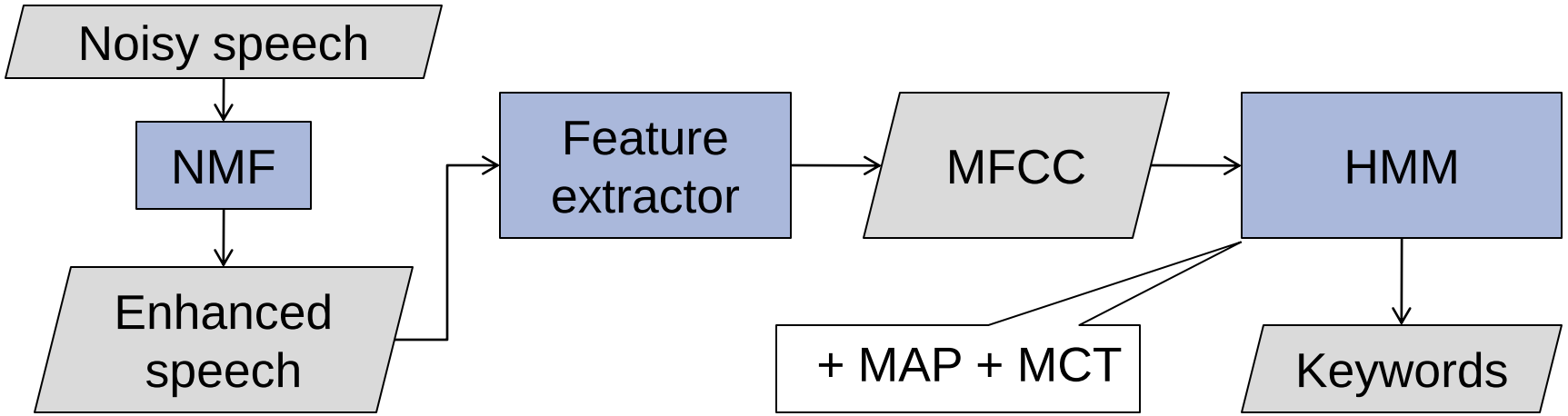
	-6 dB	-3 dB	0 dB	3 dB	6 dB	9 dB	Mean
w/o NMF	31.1	36.8	49.1	64.0	73.8	83.1	**56.3**
w/ NMF	62.2	67.7	73.2	78.5	83.8	86.1	**75.2**



Results [Development Set]

With MAP+MCT:

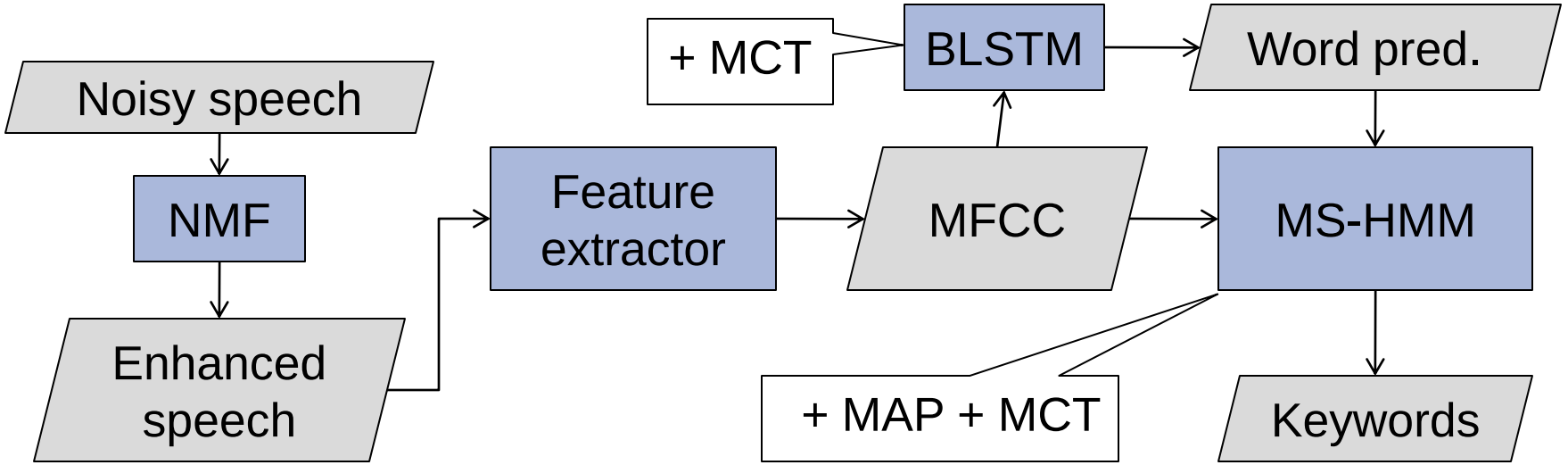
	-6 dB	-3 dB	0 dB	3 dB	6 dB	9 dB	Mean
w/o NMF	54.8	62.4	72.0	80.5	87.0	90.7	**74.5**
w/ NMF	73.6	77.3	82.2	84.2	88.6	90.0	**82.7**



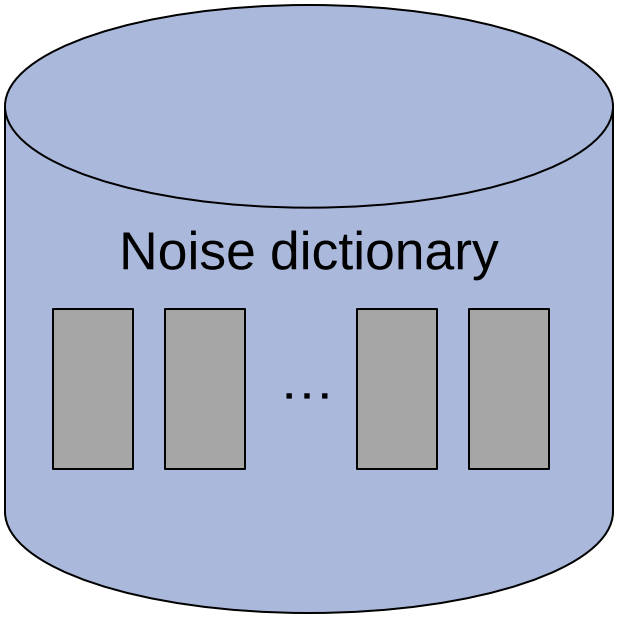
Results [Development Set]

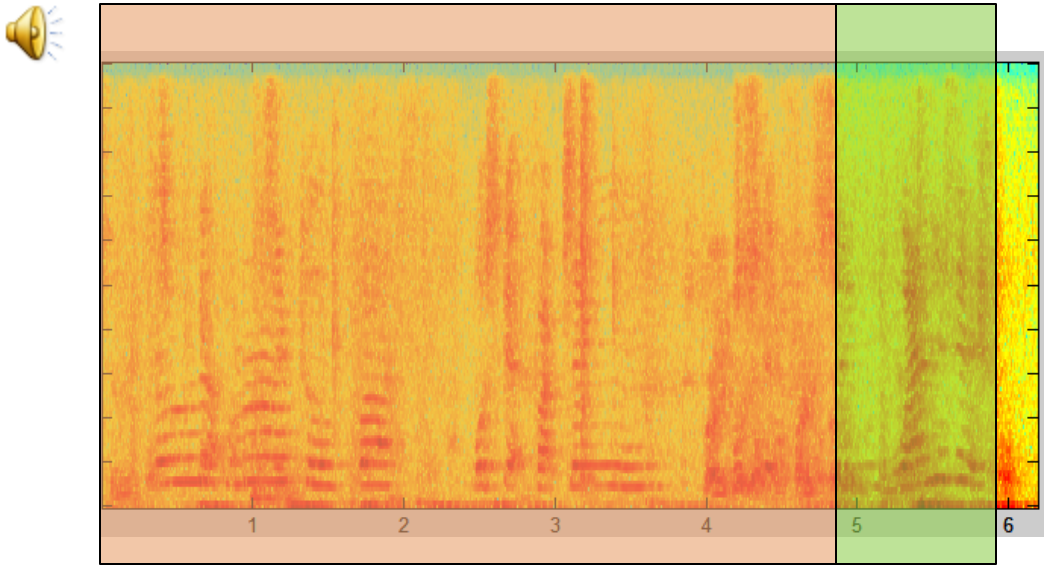
Multi-Stream Recogniser:

	-6 dB	-3 dB	0 dB	3 dB	6 dB	9 dB	Mean
w/o NMF	69.8	75.8	83.7	88.8	92.6	94.8	**84.2**
w/ NMF	81.5	83.0	86.8	90.6	92.2	93.7	**88.0**



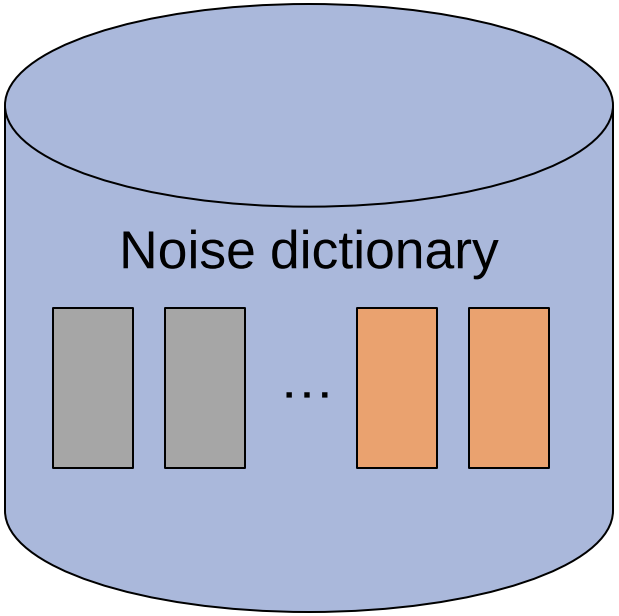
Noise-Adaptive Speech Enhancement

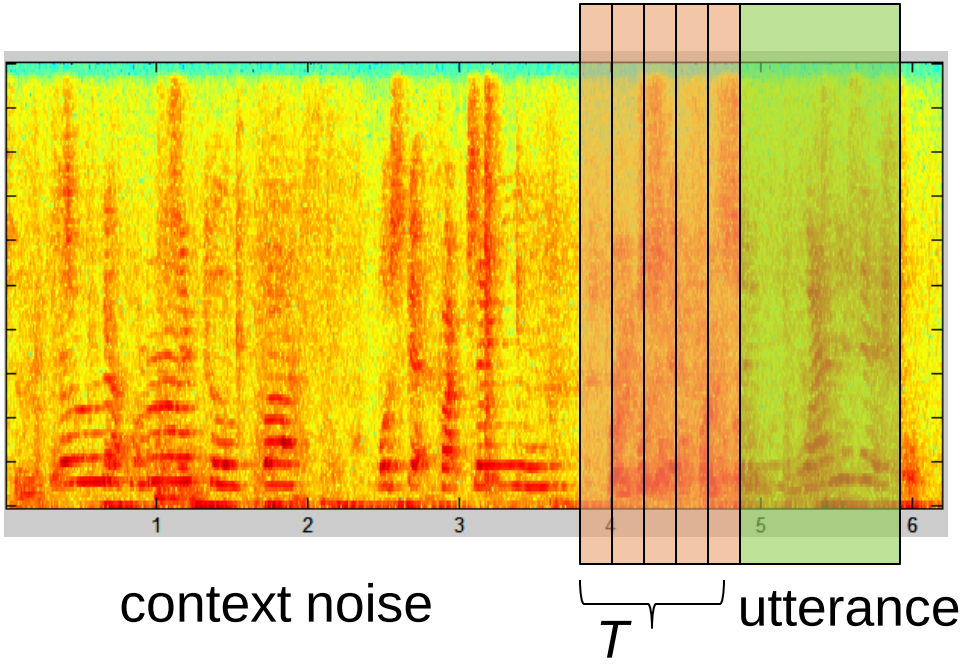




context noise utterance

Noise-Adaptive Speech Enhancement





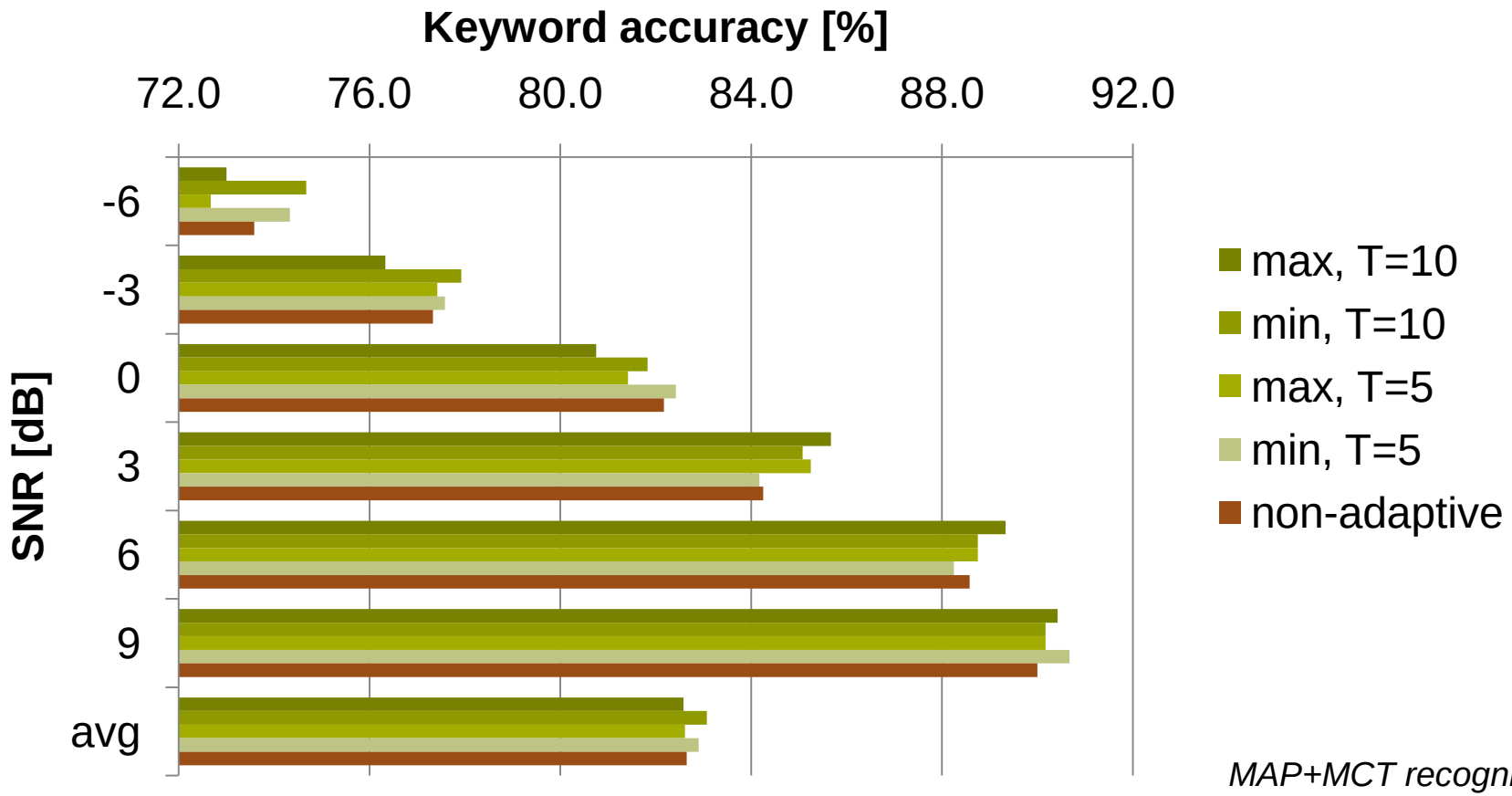
Replace *T* dictionary entries with

a) Minimum KL divergence

b) Maximum KL divergence

d(context noise | dictionary)

Noise-Adaptive Speech Enhancement: Results [Development Set]



MAP+MCT recogniser

TUM Challenge Results [Test Set]

Multi-stream HMM recogniser, MCT + MAP

	-6 dB	**-3 dB**	**0 dB**	**3 dB**	**6 dB**	**9 dB**	**Mean**
w/o NMF	68.5	75.6	82.2	88.3	90.6	93.9	83.2
w/ NMF	80.3	83.5	86.7	90.0	90.3	92.9	87.3
w/ ANMF	79.8	84.0	87.9	90.7	91.8	92.9	87.9

- 87.3% accuracy in full realism
- 87.9% using oracle for VAD

Conclusions

- Reduction of KW error rate:
  44.1% (baseline)
  → 15.6% (single-stream)
  → 12.7% (multi-stream)

- Front-end enhancement and refined back-ends:
  Complementary approaches to ASR robustness

Outlook

- Speaker-dependent BLSTM
  - First results on test (non-adaptive NMF):

-6 dB	-3 dB	0 dB	3 dB	6 dB	9 dB	Mean
82.9	87.2	90.3	93.7	93.9	94.8	**90.5**

- Pure BLSTM modelling
- Multi-stream modelling of (sparse) NMF activations
- NMF dictionary optimization

Do it Yourself!

- cNMF enhancement by openBliSSART [Weninger, 2011]
  - http://openblissart.github.com/openBliSSART
- Feature extraction: openSMILE [Eyben, 2010]
  - http://opensmile.sourceforge.net
- Multi-stream HMM:
  - HTK
  - BLSTM implemented using RNNLIB by Alex Graves http://rnnl.sourceforge.net/

Thank you.