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
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

Noisy speech → Feature extractor → MFCC → HMM → Transcr.
Solution 1: Front-End Enhancement

- Increases SNR
- Imperfect: Noise suppression vs. information loss
Solution 2: Robust Back-Ends

Noisy speech → Feature extractor → MFCC → HMM → Transcr.

Multi-condition training
MAP adaptation ...

Solution 2: Robust Back-Ends

Noisy speech → Feature extractor → MFCC → BLSTM-RNN

BLSTM-RNN → Word prediction

Word prediction → Multi-stream HMM

Multi-stream HMM → Transcr.
Proposed ASR Architecture

Noisy speech → NMF → Enhanced speech

Feature extractor → MFCC

BLSTM

Word prediction

Multi-stream HMM

Transcr.
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:

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

• Modelling of true noise spectrogram:

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

• \( R^{(s)}, R^{(n)} = 51 \) (102 NMF “components”)
Speech Enhancement: Convolutive NMF

- Matrix formulation:

\[
V \approx \Lambda^{(s)} + \Lambda^{(n)}
\]

\[
= \sum_{p=0}^{P-1} W^{(s)}(p) H^{(s)} + \sum_{p=0}^{P-1} W^{(n)}(p) H^{(n)}
\]

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

- Estimate \(\hat{V}^{(s)} = \frac{\Lambda^{(s)}}{\Lambda^{(s)} + \Lambda^{(n)}} \otimes V\) (soft masking)
Convolutional Speech and Noise Bases

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

\[
T(s,k,w) \approx \sum_{p=0}^{P-1} w(s,k,w)(p) h(s,k,w)
\]

  – Build \( W(s,k)(p) = [w(s,k,1)(p) \cdots w(s,k,51)(p)] \)

• General *noise base*:
  – Sub-sample training noise
  – Build \( W^{(n)}(p) \) by convolutional NMF
Back-End:
Multi-stream Tandem BLSTM-HMM
Context Modelling in Neural Networks

- MLP
- Feature frame stacking
- RNN
  - “Persistent” memory
  - LSTM-RNN
  - Bidirectional context
  - BLSTM-RNN
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 $x_t$
- CPT for discrete BLSTM word prediction $b_t$
- Mitigate BLSTM misclassifications by Viterbi decoding
- HMM emission probability in state $s_t$:

$$ p(y_t|s_t) = \left[ \sum_{m=1}^{M} c_{s_tm} \mathcal{N}(x_t; \mu_{s_tm}, \Sigma_{s_tm}) \right]^a \times p(b_t|s_t)^{2-a} $$

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

CHiME baseline:

<table>
<thead>
<tr>
<th>Noise Level</th>
<th>31.1</th>
<th>36.8</th>
<th>49.1</th>
<th>64.0</th>
<th>73.8</th>
<th>83.1</th>
<th>56.3</th>
</tr>
</thead>
<tbody>
<tr>
<td>-6 dB</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-3 dB</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0 dB</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 dB</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6 dB</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9 dB</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Results [Development Set]

With MAP speaker adaptation:

<table>
<thead>
<tr>
<th>Noisy speech</th>
<th>Feature extractor</th>
<th>MFCC</th>
<th>HMM</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCC</td>
<td>HMM              + MAP</td>
<td>Keywords</td>
<td></td>
</tr>
</tbody>
</table>

-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:

<table>
<thead>
<tr>
<th>-6 dB</th>
<th>-3 dB</th>
<th>0 dB</th>
<th>3 dB</th>
<th>6 dB</th>
<th>9 dB</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>54.8</td>
<td>62.4</td>
<td>72.0</td>
<td>80.5</td>
<td>87.0</td>
<td>90.8</td>
<td>74.6</td>
</tr>
</tbody>
</table>

- 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:

<table>
<thead>
<tr>
<th>Noisy speech</th>
<th>Feature extractor</th>
<th>MFCC</th>
<th>BLSTM</th>
<th>Word pred.</th>
</tr>
</thead>
<tbody>
<tr>
<td>+ MAP + MCT</td>
<td>+ MCT</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>-6 dB</th>
<th>-3 dB</th>
<th>0 dB</th>
<th>3 dB</th>
<th>6 dB</th>
<th>9 dB</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>69.8</td>
<td>75.8</td>
<td>83.7</td>
<td>88.8</td>
<td>92.6</td>
<td>94.8</td>
<td>84.2</td>
</tr>
</tbody>
</table>
… What about Speech Enhancement?
Results [Development Set]

Baseline recogniser:

<table>
<thead>
<tr>
<th></th>
<th>-6 dB</th>
<th>-3 dB</th>
<th>0 dB</th>
<th>3 dB</th>
<th>6 dB</th>
<th>9 dB</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>w/o NMF</td>
<td>31.1</td>
<td>36.8</td>
<td>49.1</td>
<td>64.0</td>
<td>73.8</td>
<td>83.1</td>
<td>56.3</td>
</tr>
<tr>
<td>w/ NMF</td>
<td>62.2</td>
<td>67.7</td>
<td>73.2</td>
<td>78.5</td>
<td>83.8</td>
<td>86.1</td>
<td>75.2</td>
</tr>
</tbody>
</table>
Results [Development Set]

With MAP+MCT:

<table>
<thead>
<tr>
<th></th>
<th>-6 dB</th>
<th>-3 dB</th>
<th>0 dB</th>
<th>3 dB</th>
<th>6 dB</th>
<th>9 dB</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>w/o NMF</td>
<td>54.8</td>
<td>62.4</td>
<td>72.0</td>
<td>80.5</td>
<td>87.0</td>
<td>90.7</td>
<td>74.5</td>
</tr>
<tr>
<td>w/ NMF</td>
<td>73.6</td>
<td>77.3</td>
<td>82.2</td>
<td>84.2</td>
<td>88.6</td>
<td>90.0</td>
<td>82.7</td>
</tr>
</tbody>
</table>

Noisy speech → NMF → Enhanced speech → Feature extractor → MFCC → HMM → Keywords

+ MAP + MCT
Results [Development Set]

Multi-Stream Recogniser:

<table>
<thead>
<tr>
<th></th>
<th>-6 dB</th>
<th>-3 dB</th>
<th>0 dB</th>
<th>3 dB</th>
<th>6 dB</th>
<th>9 dB</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>w/o NMF</td>
<td>69.8</td>
<td>75.8</td>
<td>83.7</td>
<td>88.8</td>
<td>92.6</td>
<td>94.8</td>
<td>84.2</td>
</tr>
<tr>
<td>w/ NMF</td>
<td>81.5</td>
<td>83.0</td>
<td>86.8</td>
<td>90.6</td>
<td>92.2</td>
<td>93.7</td>
<td>88.0</td>
</tr>
</tbody>
</table>

Noisy speech → NMF → Enhanced speech

NMF → Feature extractor

Feature extractor → MFCC

MFCC → BLSTM

BLSTM + MCT → Word pred.

BLSTM + MAP + MCT

Word pred. → Keywords

MFCC + MAP + MCT → Keywords
Noise-Adaptive Speech Enhancement

Noise dictionary

context noise

utterance
Noise-Adaptive Speech Enhancement

Noise dictionary

Replace $T$ dictionary entries with:

a) Minimum KL divergence
b) Maximum KL divergence
d($\text{context noise | dictionary}$)
Noise-Adaptive Speech Enhancement: Results [Development Set]

Keyword accuracy [%]

SNR [dB]

-6
-3
0
3
6
9
avg

max, T=10
min, T=10
max, T=5
min, T=5
non-adaptive

MAP+MCT recogniser
TUM Challenge Results [Test Set]

Multi-stream HMM recogniser, MCT + MAP

<table>
<thead>
<tr>
<th></th>
<th>-6 dB</th>
<th>-3 dB</th>
<th>0 dB</th>
<th>3 dB</th>
<th>6 dB</th>
<th>9 dB</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>w/o NMF</td>
<td>68.5</td>
<td>75.6</td>
<td>82.2</td>
<td>88.3</td>
<td>90.6</td>
<td>93.9</td>
<td>83.2</td>
</tr>
<tr>
<td>w/ NMF</td>
<td>80.3</td>
<td>83.5</td>
<td>86.7</td>
<td>90.0</td>
<td>90.3</td>
<td>92.9</td>
<td>87.3</td>
</tr>
<tr>
<td>w/ ANMF</td>
<td>79.8</td>
<td>84.0</td>
<td>87.9</td>
<td>90.7</td>
<td>91.8</td>
<td>92.9</td>
<td>87.9</td>
</tr>
</tbody>
</table>

- 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):

<table>
<thead>
<tr>
<th>-6 dB</th>
<th>-3 dB</th>
<th>0 dB</th>
<th>3 dB</th>
<th>6 dB</th>
<th>9 dB</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>82.9</td>
<td>87.2</td>
<td>90.3</td>
<td>93.7</td>
<td>93.9</td>
<td>94.8</td>
<td>90.5</td>
</tr>
</tbody>
</table>

• 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]

• Multi-stream HMM:
  – HTK
  – BLSTM implemented using RNNLIB by Alex Graves
    http://rnnl.sourceforge.net/
Thank you.