Artificial Bandwidth Extension
Using Deep Neural Networks for Spectral Envelope Estimation

Johannes Abel and Tim Fingscheidt

Institute for Communications Technology,
Technische Universität Braunschweig
We Need More Acoustical Bandwidth!

**Problem**: Speech quality and intelligibility suffers from limited acoustical bandwidth

Conventional **narrowband (NB)** telephony call (acoustic bandwidth: 0.3<f<3.4 kHz)
- Speech quality: 3.2/5.0 Mean opinion score (MOS) points
- Intelligibility: 90% (Consonant-vowel-consonant test)

**Wideband (WB)** telephony call with acoustic bandwidth of 0.05<f<7 kHz
- Speech quality: 4.5/5.0 MOS points
- Intelligibility: 98%

**Problem solved?**

We Need More Acoustical Bandwidth!

Requirements for a WB call:
1. WB-capable mobile handsets (far-end and near-end)
2. All participants of a call need to be located within a WB-capable cell
3. The provider’s backbone network must be WB-capable
4. Further requirements for international WB calls and also for inter-operator connections

If the many requirements are not met at the beginning of a call, only NB mode is possible.

If requirements during a call are not met anymore, the call drops to NB mode. Typically, switching back to WB mode if requirements are met again is then disabled.

Solution: Artificial Bandwidth Extension (ABE)
Estimation of frequency components from 4 to 7 kHz, a.k.a. the upper band (UB), at the receiver-side for a more consistent and WB-like experience.
Outline

1. Motivation
2. ABE Framework
   - Overview
   - Statistical Models
     - Baseline: HMM/GMM
     - DNN and HMM/DNN
3. Simulations
4. Summary
2. ABE Framework

\[ f_s' = 8\text{kHz} \]
\[ f_s = 16\text{kHz} \]
\[ s^{\text{NB}}(n') \]
\[ s^{\text{NB}}(n) \]
\[ \tilde{\Phi}_\ell^{\text{UB}} \]
\[ \tilde{\Phi}_\ell^{\text{NB}} \]
\[ \tilde{\Phi}_\ell^{\text{WB}} \]
\[ \tilde{\alpha}_\ell^{\text{WB}} \]
\[ s^{\text{UB}}(n) \]
\[ s^{\text{WB}}(n) \]

Abbreviations:
- \( n' \): NB sample idx
- \( n \): WB sample idx
- \( \ell \): Frame index
- \( f_s, f_s' \): Sampling frequencies
- \( \tilde{\Phi} \): Power spectral density
- \( \Phi \): LP filter coef.
2. ABE Framework

UB Spectral Envelope Classification

Feature vec. Codebook entry idx
\( \mathbf{x}_\ell \)
A posteriori prob.
\( P(s_i | \mathbf{x}) \)
Est. UB cepstral vec.
\( \mathbf{\hat{y}}_i \)

UB Spectral Envelope Estimation

\( s_{NB}^{NB}(\eta') \rightarrow \) Feature Extraction \( \xrightarrow{\mathbf{x}_\ell} \) Statistical Model
\( \xrightarrow{P(s_i | \mathbf{x})} \sum_i P(s_i | \mathbf{x}) \cdot \mathbf{\hat{y}}_i \rightarrow \mathbf{\hat{y}}_\ell \xrightarrow{\text{Spectral Conversion}} \) Estimated UB Spectral Envelope

UB Envelope Codebook

\( \Phi_{UB}^{NB} \)

UB Envelope

\( \Phi_{UB}^{UB} \)

"UB energy"

\( \tilde{\Phi}_{UB}^{WB} = \mathbf{\hat{y}}_\ell(0) \)
2. ABE Framework

Statistical Model: HMM/GMM (Baseline)

- **LDA Matrix**
  - $H$
  - $x_t$ → LDA Transform

- **GMM Param.**
  - $w_{i,m}$, $\mu_{i,m}$, $\Sigma_{i,m}$
  - $p(x_t | s_i)$ → GMM

- **Forward Algorithm**
  - $P(s_i)$, $P(s_j | s_i)$
  - $P(s_i | x_t^f)$ → HMM/GMM

Linear discriminant analysis (LDA) for dimension reduction of features

Forward algorithm for HMM evaluation

GMM as **acoustic model**
2. ABE Framework
Statistical Model: HMM/DNN (new)

Deep neural network (DNN) as acoustic model

Forward algorithm for HMM evaluation

Posterior outputs from DNN are recalculated to likelihoods
2. ABE Framework
Statistical Model: DNN (new)

DNN as statistical model
Outline

1. Motivation
2. ABE Framework
   - Overview
   - Statistical Models
     - Baseline: HMM/GMM
     - DNN and HMM/DNN
3. Simulations
4. Summary
3. Simulations
Experimental Setup

DNN Experiments
- Initial weights for DNN training from restricted Boltzmann machine (RBM) pretraining
- DNN topologies under test:
  - Number of hidden layers: 1, 2, 3, 4, 5, 6
  - Number of units per layer: 512

Datasets

<table>
<thead>
<tr>
<th>Step</th>
<th>Speech Database</th>
</tr>
</thead>
<tbody>
<tr>
<td>Codebook, RBM pretraining, HMM/DNN/GMM training</td>
<td>TIMIT Train Set</td>
</tr>
<tr>
<td>DNN validation checks</td>
<td>TIMIT Test Set</td>
</tr>
<tr>
<td>Result reporting</td>
<td>NTT-AT Database (EN+DE)</td>
</tr>
</tbody>
</table>

Cepstral Distances for...
...estimated UB envelope: \( D_{\text{env}} = 10\sqrt{2} \cdot \log_{10}(e) \sqrt{\sum_{\nu=1}^{N_{\text{env}}} (y(\nu) - \tilde{y}(\nu))^2} \)

...estimated UB energy ratio: \( D_0 = 10\sqrt{2} \cdot \log_{10}(e) |y(0) - \tilde{y}(0)| \)
### 3. Simulations

Results – Cepstral Distances

<table>
<thead>
<tr>
<th>#Hidden Layer(s)</th>
<th>#Units</th>
<th>( D_{\text{env}} ) [dB]</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>DNN</td>
</tr>
<tr>
<td>1</td>
<td>512</td>
<td>5.34</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>5.41</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>5.38</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>5.44</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>5.40</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>5.39</td>
</tr>
<tr>
<td><strong>HMM/GMM</strong></td>
<td></td>
<td>5.31</td>
</tr>
<tr>
<td><strong>Oracle</strong></td>
<td></td>
<td>4.44</td>
</tr>
</tbody>
</table>

DNN topology has only small influence on evaluation metrics.

UB energy cepstral distance decreased by more than 2 dB (improvement!).

Still big potential for further improvement.

UB envelope reconstruction very similar in all cases, small potential for further improvement.
3. Simulations
Results – Speech Quality (WB-PESQ)

<table>
<thead>
<tr>
<th>Statistical Model</th>
<th>MOS_{LQO}</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMM/GMM (Baseline)</td>
<td>2.73</td>
</tr>
<tr>
<td>DNN</td>
<td>[3.05,3.08]</td>
</tr>
<tr>
<td>HMM/DNN</td>
<td>[2.99,3.02]</td>
</tr>
<tr>
<td>Oracle</td>
<td>3.26</td>
</tr>
</tbody>
</table>
3. Simulations
Latest ABE Approach and CCR-Test

UB SpectralEnvelopeEstimation → FeatureExtraction → DNN → DNN++ → SpectralConversion

<table>
<thead>
<tr>
<th>CCR Condition</th>
<th>CMOS</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMR vs. AMR-WB</td>
<td>2.15</td>
</tr>
<tr>
<td><strong>HMM/GMM</strong> vs. AMR-WB</td>
<td>1.48</td>
</tr>
<tr>
<td><strong>DNN++</strong> vs. AMR-WB</td>
<td>1.31</td>
</tr>
<tr>
<td><strong>HMM/GMM</strong> vs. <strong>DNN++</strong></td>
<td>0.13</td>
</tr>
<tr>
<td>AMR vs. <strong>HMM/GMM</strong></td>
<td>0.81</td>
</tr>
<tr>
<td>AMR vs. <strong>DNN++</strong></td>
<td>1.37</td>
</tr>
</tbody>
</table>
1. Motivation
2. ABE Framework
   - Overview
   - Statistical Models
     ▪ Baseline: HMM/GMM
     ▪ DNN and HMM/DNN
3. Simulations
4. Summary
4. Summary

- DNNs outperform GMMs as acoustic model for artificial bandwidth extension.

- Using DNNs led to an improvement of up to 0.35 MOS\textsubscript{LQO} points when ABE-processed speech is evaluated using WB-PESQ.

- A superior UB energy estimation $\tilde{y}_\ell(0)$ is responsible for the speech quality gain, rather than the UB envelope.

- The UB spectral envelope estimation performance of DNNs is similar compared to GMMs.

- Huge potential for further improvement of UB energy estimate.

- Superiority of using DNNs in ABE was proven by a clear 1.37 CMOS points advantage over AMR-coded narrowband speech.
Thank you for your attention

Johannes Abel

abel@ifn.ing.tu-bs.de
2. ABE Framework
UB Envelope Codebook

\[ \varphi_\ell = \begin{cases} 
1, & \text{if frame } \ell \text{ contains an } /s/ \text{ or } /z/ \text{ sound} \\
0, & \text{else}
\end{cases} \]

Prediction gain UB

Prediction gain NB

Relative energy ratio

\[ y_\ell(0) = \ln \left( \frac{g_{\ell}^{UB}}{g_{\ell}^{NB}} \right) \cdot \frac{1}{\sqrt{2}} \]

3. Simulations
Results – Phoneme Accuracy

Relative classification accuracy of **HMM/DNN** vs. **HMM/GMM** for phonemes (measured on validation set)

<table>
<thead>
<tr>
<th>Phoneme</th>
<th>/f/</th>
<th>/th/</th>
<th>/dh/</th>
<th>/t/</th>
<th>/zh/</th>
<th>…</th>
<th>/s/</th>
</tr>
</thead>
<tbody>
<tr>
<td>$ACC_{rel} [%]$</td>
<td>+83</td>
<td>+59</td>
<td>+56</td>
<td>+54</td>
<td>+52</td>
<td>…</td>
<td>+8</td>
</tr>
</tbody>
</table>

4 of 5 phonemes that profit most are fricative sounds

**All** phonemes take profit from DNN as acoustic model

$$ ACC_{rel} = \frac{ACC_{DNN} - ACC_{GMM}}{ACC_{GMM}} $$