Speech Bandwidth Extension Based on Speech Phonetic Content and Speaker Vocal Tract Shape Estimation

Itai Katsir, Israel Cohen and David Malah
Outline

• Introduction
  ➢ Bandwidth Extension (BWE) Background
  ➢ General BWE Block Diagram

• Proposed BWE Algorithm
  ➢ Vocal Tract Modeling
  ➢ Features Extraction
  ➢ WB Spectral Envelope Estimation

• Performance Evaluation

• Conclusion
Introduction - Background

- **Fact** – Growing consumer demand for HD media → high quality speech communication.
- **Problem** – Today’s analog telephone and PSTN limit the speech to narrowband frequency range of about 300-3400Hz → lower speech quality compared to wideband speech of range 50-7000Hz.

![Graph showing Intelligibility and Quality](image)

(a) Intelligibility of meaningless syllables

(b) Subjective speech quality
Introduction - Background

- **Solution** – Artificially extend speech bandwidth to achieve speech quality enhancement.
  - 3.4-7kHz – Higher intelligibility and quality
  - 0-0.3kHz – Higher naturalness and quality

- **+0.3 MOS** - state of the art published BWE algorithms
**Application** – In the transition time to full WB communication networks, BWE can be used in mix NB-WB communication networks.

**Option 1** BWE in the Gateway and its output is WB produced by BWE

**Option 2** BWE locally in the IPP/DECT/Mobile which receives NB and by BWE produces WB.
Introduction - General BWE Scheme

- Using the Source-filter model: separate, independent algorithms for excitation extension and spectral envelope extension.

**Spectrum of \( s_{NB}(n) \)**

**Spectrum of estimated \( s_{WB}(n) \)**

**Filter path**
- NB Spectral envelope extraction
- LB/HB spectral envelope estimation
- Gain adjustment and post-processing of LB/HB signals

**Source path**
- NB Excitation extraction
- LB/HB excitation generation

**Spectrum of NB excitation**

**Spectrum of estimated WB envelope**

**Spectrum of generated WB excitation**
Proposed BWE Algorithm

Vocal Tract Modeling

- The proposed algorithm tries to estimate the speaker physical vocal tract shape
- Atal and Wakita showed the equivalence of acoustic tube model and linear prediction (LP) under certain conditions [Atal, 1970; Wakita, 1973]
- The $M^{th}$ order filter transfer function derived through LP is equivalent to the transfer function of an acoustical tube made up of $M$ equal length sections of variable areas.
- This is referred to as Vocal Tract Area Function (VTAF)

\[
M = f_s \frac{2L}{c}
\]

\[
A_{nA} = \frac{1 + r_{nA}}{1 - r_{nA}} A_{nA+1}
\]
Proposed BWE Algorithm

Algorithm stages:

I. NB signal preprocessing and features extraction
II. HB spectral envelope estimation and postprocessing
III. HB excitation generation
IV. Wideband signal synthesis
Proposed BWE Algorithm

Algorithm stages:

1. **NB signal preprocessing and features extraction**
   - $x_1$ - Features for speech state estimation (MFCC, spectral centroid, spectral flatness, spectral slope and normalized energy)

2. **HB spectral envelope estimation and postprocessing**
   - $x_2$ - NB VTAF for WB VTAF estimation

3. **HB excitation generation**
   - $x_3$ - NB excitation for WB excitation generation

4. **Wideband signal synthesis**
Proposed BWE Algorithm

Algorithm stages:

I. NB signal preprocessing and features extraction

II. HB spectral envelope estimation and postprocessing

III. HB excitation generation

IV. Wideband signal synthesis
WB Spectral Envelope Estimation

**Phoneme estimation:** using HMM to estimate each speech frame linguistic state

- **Off-line process:** using TIMIT transcription to build HMM statistical model using phoneme based states. Calculating the following PDFs:
  - $p(S_i)$ - Initial probability of each state
  - $p(S_i(m)|S_j(m-1))$ - Transition probability of the Markov chain from state $j$ to state $i$
  - $p(x_1|S_i)$ - Observation probability for each state. Approximated by GMM parameters using the EM algorithm
WB Spectral Envelope Estimation

Phoneme estimation: using HMM to estimate each speech frame linguistic state

- On-line process: making a decision on current frame state (phoneme) by maximizing the a-posteriori PDF:

\[
p(S_i(m)|X_1(m)) = p(x_1(m)|S_i(m)) \cdot \sum_{j=1}^{N_s} p(S_i(m)|S_j(m-1)) p(S_j(m-1)|X_1(m-1))
\]
WB Spectral Envelope Estimation

Estimate WB VTAF: using codebook mapping of calculated NB VTAF to WB VTAF

- **Off-line process:** for each speech state, clustering of $N_{CB}$ WB VTAF using vector quantization of speech frames training set
- **On-line process:** finding closest WB VTAF to extracted NB VTAF using Euclidean distance

$$\tilde{A}_{WB}^S_i = A_{WB}^S(j^{opt})$$

$$j^{opt} = \arg \min_{j=1}^{N_{CB}} \left\| \log(A_{NB}) - \log(A_{WB}^S(j)) \right\|_2^2$$

$$s_{WB}(n)$$

Preprocessing and Feature Extraction

Stage I

\[ x_2 \]

Stage III

Speech State Estimation

\[ x_2 \]

Stage IV

WB VTAF Estimation

Statistical model #1

Postprocessing

WB Speech Synthesis

Statistical model #2
**WB Spectral Envelope Estimation**

**Postprocessing:** reduce artifact due to erroneous estimation

- Reduce artifacts due to erroneous state estimation by using $N_{\text{best}}$ highest probability states for VTAF estimation

$$\tilde{A}_{\text{WB}} = C \cdot \left( p_1 \cdot \tilde{A}_{\text{WB}}^{S_i} + \ldots + p_{N_{\text{best}}} \cdot \tilde{A}_{\text{WB}}^{S_{i_N}} \right)$$

- Estimated WB envelope fit to NB envelope by formant frequencies tuning of estimated WB VTAF to allow better gain adjustment to NB envelope. Iterative tuning by VTAF perturbation

- Iterative VTAF perturbation based on the sensitivity function:

$$\frac{\Delta f_{n_f}}{f_{n_f}} = \sum_{n_A} S_{n_f,n_A} \frac{\Delta A_{n_A}}{A_{n_A}}$$
WB Spectral Envelope Estimation

Vocal Tract Sensitivity Function

Relate changes in area to changes in formant frequencies

\[
\frac{\Delta f_{n_f}}{f_{n_f}} \leftrightarrow \frac{\Delta A_{n_A}}{A_{n_A}}
\]

Sensitivity function

\[
\frac{\Delta f_{n_f}}{f_{n_f}} = \sum_{n_A}^{N_A} S_{n_f,n_A} \frac{\Delta A_{n_A}}{A_{n_A}}
\]
**WB Spectral Envelope Estimation**

**Postprocessing:** reduce artifact due to erroneous estimation

- Stopping condition for iterative process is formant frequencies difference
- Smoothing in time of tuned estimated VTAF
  \[
  \tilde{A}_{WB}'(m) = \beta \cdot \tilde{A}_{WB}'(m-1) + (1 - \beta) \cdot \tilde{A}_{WB}(m)
  \]
- Converting WB VTAF to WB spectral envelope
- Gain adjustment of estimated WB spectral envelope to match the input NB spectral envelope

![Diagram of WB Spectral Envelope Estimation Process]
BWE Quality Evaluation

Log Spectral Distance for phone category

- Reference algorithm: *Evaluation of an Artificial Speech Bandwidth Extension Method in Three Languages* [Pulakka et al., 2008]

\[
LSD = \sqrt{\frac{1}{k_{\text{high}} - k_{\text{low}} + 1} \sum_{k=k_{\text{low}}}^{k_{\text{high}}} 10\log_{10} \frac{P(k)}{\bar{P}(k)}^2}
\]

- voiced fricative
- unvoiced fricative
- affricates
- diphthongs
- vowels
- unvoiced stops
- voiced stops
- semivowels
- nasals

<table>
<thead>
<tr>
<th></th>
<th>Proposed algorithm</th>
<th>Reference algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td>voiced fricative</td>
<td>approximately 10</td>
<td>approximately 15</td>
</tr>
<tr>
<td>unvoiced fricative</td>
<td>approximately 10</td>
<td>approximately 15</td>
</tr>
<tr>
<td>affricates</td>
<td>approximately 5</td>
<td>approximately 10</td>
</tr>
<tr>
<td>diphthongs</td>
<td>approximately 15</td>
<td>approximately 20</td>
</tr>
<tr>
<td>vowels</td>
<td>approximately 10</td>
<td>approximately 15</td>
</tr>
<tr>
<td>unvoiced stops</td>
<td>approximately 5</td>
<td>approximately 10</td>
</tr>
<tr>
<td>voiced stops</td>
<td>approximately 15</td>
<td>approximately 20</td>
</tr>
<tr>
<td>semivowels</td>
<td>approximately 5</td>
<td>approximately 10</td>
</tr>
<tr>
<td>nasals</td>
<td>approximately 10</td>
<td>approximately 15</td>
</tr>
</tbody>
</table>
BWE Quality Evaluation

Histogram of estimated formant frequencies error

- Reference algorithm: *Low-Complexity Feature-Mapped Speech Bandwidth Extension* [Gustafsson et al., 2006]

\[
\text{Error} = \left| \tilde{f}_{HB} - f_{HB} \right|
\]
BWE Quality Evaluation

Subjective quality evaluation

- **MUSHRA (MUltiple Stimuli with Hidden Reference and Anchor)** – ranks several speech samples for score between 0-100.
- Recommendation ITU-R BS.1116-1
- Test setup: 11 listeners, 6 different experiments (English, 3 male, 3 female) – each with WB reference signal, NB anchor signal, proposed BWE signal and a reference BWE signal from Geiser [Geiser et al., 2010]

<table>
<thead>
<tr>
<th></th>
<th>Proposed</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean MUSHRA score</td>
<td>60.5</td>
<td>52</td>
</tr>
</tbody>
</table>

![Chart showing MUSHRA scores for WB, NB, Proposed, and Reference signals]
BWE Quality Evaluation - Male

Spectograms of original WB signal, NB signal and proposed BWE signal.
BWE Quality Evaluation - Female

Spectograms of original WB signal, NB signal and proposed BWE signal.
Conclusion

- Proposed BWE algorithm innovations
  - Phonetic content estimation of each speech frame
  - WB VTAF estimation for specific speaker using codebook mapping
  - Perturbation of estimated WB VTAF for better gain adjustment

- Algorithm advantages
  - Reduce estimation error for unvoiced frames
  - Reduce estimation error artifacts

- Algorithm shortcomings
  - Using VTAF is lacking in modeling nasal and unvoiced sounds
  - Sensitivity function calculation and the iterative VTAF tuning cause high algorithm complexity

- Future work
  - Reduce algorithm complexity by using sensitivity functions tables
  - Use the postprocessing iterative procedure for better refinement and control of estimated spectral envelope
Speech Bandwidth Extension Based on Speech Phonetic Content and Speaker Vocal Tract Shape Estimation

Thank You