QUALITY ENHANCEMENT OF CELP CODED SPEECH
OUTLINE

- INTRODUCTION (CELP CODING OF SPEECH)
- PROBLEM IN CELP CODED SPEECH
- PROPOSED ENHANCEMENT SYSTEM
- EXPERIMENT RESULTS
- CONCLUSION
INTRODUCTION

- **CELP (Codebook Excited Linear Prediction)** Coder is one of the most widely accepted coder in the telecommunication industry that can compress speech signals to 4.8kbps.

- CELP coder is basically a hybrid coder which exploits features from waveform coders and from parameter coders. It can operate at low bit rate while preserving the coded speech quality within acceptable limits.

- CELP coders operate on narrowband (0-4khz) or telephone speech signal
CELP coding is based on a source-excitation model of human vocal system. The excitation signal is selected from a stochastic codebook containing sequences of noise which mimic the residual signal from linear prediction process.

Standards of CELP coders that are being used in telecommunication industry, e.g., FS 1016 CELP coder, QCELP coder of North American Cellular (CDMA) IS96, CS-ACELP coder, LD-CELP coder of ITU-T G.728 etc.
PROBLEM WITH CELP CODED SPEECH

- CELP coding would introduce degradation in the original speech. The two distinct artifacts are known as Hoarse and Muffing characteristics.
Hoarse characteristics are mainly due to coding noise inherent in the CELP coded speech, in which the noise is generated as a result of stochastic excitation. The pitch periodicity in voiced speech is considerably weakened by this coding noise.
Muffing characteristics are due to the lack of high frequency components in the signal. The removal of the high frequencies from 4khz to 8khz prior to coding severely affects the voice quality of the high frequency consonants such as /s/, /f/, /sh/ in the decoded speech.
Proosed Enhancement System

- **Hoarse characteristics** from high level of coding noise can be reduced by removing the noise from the coded speech or, on the other hand, by improving the pitch periodicity of coded speech in the voiced region of speech spectrum. This can be achieved by using the harmonic-plus-noise model which is very well known for its capability of generating “clean” synthetic speech.
Muffling characteristics can be reduced by re-inserting appropriately measured high frequency components (4-8 khz) in the narrowband CELP coded speech.
The narrowband CELP decoded speech is first analyzed by a Harmonic plus Noise analyzer and lowband information are extracted.

By exploiting the lowband spectrum envelope and V/UV information, the highband (4-8kHz) spectrum envelope is recovered by using a Codebook or Gaussian mixture model.

Both the lowband and highband information are then combined and a wideband (0-8kHz) harmonic plus noise synthesizer is then used to generate speech.
During lowband harmonic analysis, the fundamental frequency, harmonic magnitudes, phases, gain, and 10 LSP representing the spectrum envelope are obtained.
The highband information is estimated from the highband (4kHz-8kHz) spectrum envelope.
Different techniques are used to recover the highband spectrum envelope in which a straightforward codebook mapping approach and statistical based prediction models are used.

The codebooks and prediction models are trained on wideband speech parameters (0-8khz).
CODEBOOK APPROACH

- A Codebook is trained on the speech database which contains lowband (0-4kHz) feature vectors (MFCC) and their corresponding wideband (0-8kHz) spectrum envelope.

- Codebooks of different sizes (128, 256, 512) are trained and average spectral distortion in estimating the highband spectrum envelope is calculated by the following equation

\[
SD_{AVE} = \left[ \frac{1}{N} \sum_{n=1}^{N} \frac{1}{2\pi} \int_{-\pi/2}^{\pi/2} \left(10\log|H_n(\omega)| - 10\log|\hat{H}_n(\omega)| \right)^2 d\omega \right]^{1/2}
\]  

(1.a)
Wideband Speech Signal (0-8 kHz) → LP Filtering + Decimation by factor of 2 → Lowband Speech Signal (4-8 kHz)

Wideband 
LPC Analysis

CELP CODER

Channel

CELP DECODER

128-Point Original WSE

Wideband Codebook

21-Point MFCC Vectors

128-Point Estimated WSE

Compute:

- Average Highband (4kHz-8kHz) Spectral Distortion

Performance evaluations of trained codebooks

<table>
<thead>
<tr>
<th>Codebook Size</th>
<th>Average Highband (4kHz-8kHz) Spectral Distortion $S_{DAVE}$ dB</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Inside the Training Data</td>
</tr>
<tr>
<td>128</td>
<td>3.71</td>
</tr>
<tr>
<td>256</td>
<td>3.63</td>
</tr>
<tr>
<td>512</td>
<td>3.55</td>
</tr>
</tbody>
</table>
CODEBOOK LIMITATIONS

- Hard classification procedure
- Codebooks are lack of statistical information
- No solid criteria to generate the new templates
A finite mixture of gaussian densities can be expressed mathematically as

\[ p(x) = \sum_{j=1}^{q} \pi_j p(x; \theta_j) \quad (1) \]

\[ \sum_{j=1}^{q} \pi_j = 1 \quad (2) \]

Where \( q \) is the number of normal component densities and \( \pi_j \) are the mixing proportions, component weights or prior probabilities.
The individual component density is completely defined by the parameter vector $\theta_j$ as given below

$$\theta_j = (\mu_j, \Sigma_j), \quad j = 1, 2, 3 \ldots \ldots q \quad (3)$$

Where $\mu_j$, and $\Sigma_j$ are the mixing proportions, mean vectors and covariance matrix of $j_{th}$ component density in GMM.

The probability of an $n$-dimensional input vector $x$ for $j_{th}$ component density can be determined by the equation given below

$$p(x; \theta_j) = \frac{\pi_j}{(\sqrt{2\pi})^n |\Sigma_j|^{\frac{1}{2}}} \exp \left[ -\frac{1}{2} (x - \mu_j)^T \Sigma_j^{-1} (x - \mu_j) \right] \quad (4)$$
The feature vectors are obtained first from the speech database and a GMM model is trained over feature vector distribution.

The mean vectors $\mu$, covariance matrices $\Sigma$, and prior probabilities $\pi$ for each component density are estimated during the training of a Gaussian Mixture Model.

Maximum likelihood method is applied which uses the well-known EM (Expectation and Maximization) algorithm iteratively to estimate the parameters.

The LBG algorithm is used with split initialization to classify the speech database into different number of classes and to obtain the initial parameter values for EM algorithm.
The EM algorithm has two steps; the E-step or expectation step uses the equation (5) to classify the input vectors.

\[ y_i \in C_k \iff k = \arg \max_j \left[ \log \pi_j - \log |\Sigma_j| - (y_i - \mu_j)^T \Sigma_j^{-1} (y_i - \mu_j) \right] \quad (5) \]

M-step updates the parameters by using the following equations (6) and (7)

\[ \pi_k = \frac{N_k}{N}, \quad \mu_k = \frac{1}{N_k \sum_{y_i \in c_k} y_i} \]

\[ \Sigma_k = \frac{1}{N_k \sum_{y_i \in C_k} (y_i - \mu_k)(y_i - \mu_k)^T} \]
RECOVERING HIGHBAND SPECTRUM ENVELOPE

Let \( x \) be the lowband and \( y \) be the highband spectral vector then the joint density of wideband vector

\[
z = \begin{bmatrix} x \\ y \end{bmatrix}
\]

(8.a)

is modeled as a mixture of \( q \) component of \((2 \times n)\)-variate Gaussian function and mathematically given below

\[
p(z \mid \theta_j) = \sum_{j=1}^{q} \frac{\pi_j}{(2\pi)^{n/2} |\Sigma_j|^{1/2}} \exp \left[ -\frac{1}{2} (z - \mu_j)^T \Sigma_j^{-1} (z - \mu_j) \right]
\]

(8.b)

\[
\sum_{j=1}^{q} \pi_j = 1, \quad \pi_j \geq 0
\]

(9)
The highband spectrum envelope is obtained by interpolating among the component densities having the highest MAP (Maximum a Posterior Probabilities) for lowband vector $x$ by using equation given below

$$p(x) = p(\theta_j | x) = \frac{\frac{\pi_j}{(2\pi)^{\frac{q}{2}} |\Sigma_j|^{\frac{1}{2}}} \exp \left[ -\frac{1}{2} (x - \mu_j)^T \Sigma_j^{-1} (x - \mu_j) \right]}{\sum_{j=1}^{M} \frac{\pi_j}{(2\pi)^{\frac{q}{2}} |\Sigma_j|^{\frac{1}{2}}} \exp \left[ -\frac{1}{2} (x - \mu_j)^T \Sigma_j^{-1} (x - \mu_j) \right]}$$

$$\alpha_i = \frac{\left[1 - p_l(x)\right]}{\sum_{i=1}^{M} \left[1 - p_l(x)\right]}, \quad E(y) = \sum_{l=1}^{M} \alpha_i \times \mu_l$$

Where $M$ is the total number of component densities used for the interpolation and has range $1 < M < 5$. 
The efficiency of trained voiced and unvoiced Gaussian mixture model is also evaluated. The average spectral distortion in lowband and highband between the original wideband and predicted wideband spectrum envelope is shown in table below.

<table>
<thead>
<tr>
<th>GMM MODELS</th>
<th>Average Highband (4kHz-8kHz) Spectral Distortion $SD_{AVE}$ dB</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Inside the Training Data</td>
</tr>
<tr>
<td>Voiced</td>
<td>2.01</td>
</tr>
<tr>
<td>Unvoiced</td>
<td>3.11</td>
</tr>
</tbody>
</table>

Performance evaluations of trained voicing GMM models
The voicing GMM model is further modified to multilevel GMM model and trained with MFCC coefficients and their corresponding wideband spectrum envelope.

<table>
<thead>
<tr>
<th>GMM MODELS</th>
<th>Average Highband (4kHz-8kHz) Spectral Distortion SD(_{AVE}) dB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Types</td>
<td>Classes</td>
</tr>
<tr>
<td>Voiced</td>
<td>128</td>
</tr>
<tr>
<td>Mixed</td>
<td>128</td>
</tr>
<tr>
<td>Unvoiced</td>
<td>128</td>
</tr>
</tbody>
</table>

Performance evaluations of trained multi-voicing GMM models
VOICED SPECTRUM ENVELOPE

Original wideband spectrum envelope
Predicted wideband spectrum envelope

Magnitude dB

Frequency Hz

0 1000 2000 3000 4000 5000 6000 7000 8000

-20 -10 0 10 20 30
Estimated wideband speech

CELP coded speech

**SHORT TIME MAGNITUDE SPECTRUM**

**Frequency Hz**

**Magnitudes dB**
SHORT TIME MAGNITUDE SPECTRUM

Estimated speech signal
CELP coded speech signal

Frequency Hz

Magnitudes dB

0 1000 2000 3000 4000 5000 6000 7000 8000

-60 -50 -40 -30 -20 -10 0 10 20 30

29
Original wideband speech signal
Estimated wideband speech signal
Short Time Magnitude Spectrum

Original wideband speech signal
CELP coded narrowband speech signal
**Short Time Magnitude Spectrum**

- **Original Wideband Speech Signal**
- **Estimated Wideband Speech Signal**
SUBJECTIVE MEASURE OF QUALITY

A : Synthesis Wideband Speech
B : CELP Coded Narrowband Speech

OBSERVATION
A much better than B
A better than B
A slightly better than B
A same as B
A slightly worse than B
A worse than B
A much worse than B

GRADING
3
2
1
0
-1
-2
-3

The ITU-R 7 point comparative scale for grading
CONCLUSION

- An enhancement system is proposed to improve the quality of narrowband CELP coded speech via lowband Harmonic plus Noise analysis and wideband extension by using Gaussian Mixture Models.

- The quality of CELP coded speech after the harmonic analysis and wideband extension has been improved significantly.

- The wideband speech is pleasant to listen, but it shows some buzzing artifacts.