Recognition Compatible Voice Coder (RECOVC)

Multimedia & Signal Processing Group

IBM Research Lab in Haifa
The Application

- A low-complexity (possibly mobile) pervasive device captures voice
- Voice needs to be converted to text by an Automatic Speech Recognition (ASR) Server
- Voice needs to be compressed in order to reduce storage space or transmission bandwidth
- Playback of the voice on the pervasive device may be required (e.g. digital voice recorders)
- Playback of the voice on the server may be required
  - Interactive Voice Response (IVR) services of "sensitive type", e.g. banking and brokerage transactions (playback is a legal requirement)
  - Human verification of collected speech databases, used to retrain the engine and improve recognition rates
- Other applications: real-time voice communication with ASR-based computer monitoring (e.g. key word spotting)
Low bit rate speech compression methods available today are tuned for the human listener.

ASR performance substantially degrades when processing compressed speech.

Especially true for:
- Large vocabulary tasks,
- Continuous speech recognition,
- Noisy recording environments,
- Channels prone to bit-errors and packet loss.
Possible Solutions

- Do not use low bit rate compression when ASR is to be performed (e.g. use G.711 at 64 kbps for telephony bandwidth speech)
  - Not applicable for most applications

- Develop new ASR engines that are tuned for specific low bit rate speech coders (e.g. AMR-GSM)
  - Requires a new engine when the network coder changes
  - Questionable performance for large vocabulary tasks, noisy recordings and channels prone to bit-errors and packet loss
  - Not extendable to wideband speech
Distributed Speech Recognition (DSR)

- ASR features are calculated on the pervasive device (ASR "front-end")
- ASR features are compressed at low bit rate such that the performance of the speech recognizer does not degrade
- Compressed ASR features are packed with error detection and recovery bits, and transmitted over a data channel to the Server
- ASR "back-end" performed on the server
- DSR Standardization Activity:
  - ETSI/Aurora
  - ITU-T SG16 (new activity)
- **The problem: Speech playback capability is lost!**
  - Transmission of an additional compressed voice stream for playback purposes only increases complexity & bandwidth
What is RECOVC?

- **IBM proprietary technology for DSR**
  - Suitable for large-vocabulary, continuous speech recognition (as well as small vocabulary, command & control type applications)
  - Interoperable with the IBM ViaVoice recognition engine (8 - 16 kHz speech)
  - Enables playback of speech on the pervasive device and on the Server, with minimal overhead

- **Includes:**
  - Extraction of Mel-Frequency Cepstral Coefficient (MFCC) on the pervasive device (common ASR features)
  - Low complexity fundamental-frequency ("pitch") extraction on the pervasive device
  - Compression of ASR features such that recognition rates do not degrade, and compression of fundamental-frequency
  - Transmission format suitable for packet-based data channels
  - Speech reconstruction from the decoded ASR and fundamental-frequency, for playback purposes
Speech Recognition Front-end

## RECOVC Encoder (Client Side)

- **Input Speech (8-16 kHz)**
- **STFT**
- **Pitch Detection**
- **Pitch**
- **Pitch Compression**
- **Feature Vector Compression**
- **MUX**
- **Bitstream for storage/transmission (4.5-7 kbps)**

- **MFCC**
- **500 bps**
- **4-6.4 kbps**
**RECOVC Decoder (Server Side)**

- **DE MUX**: Bit-stream -> **Pitch Decompression**
- **Feature Vector Decompression**
- **Speech Recognition Back-end**
- **Speech Reconstruction (sinusoidal model)**
- Output speech for playback

- Pitch
- MFCC
- Text
Mel-Frequency Cepstral Coefficients (MFCC)

★ Phase lost while taking the ABS of STFT
★ Spectral resolution lost while integrating
★ Only part of the MFCCs are maintained (13 out of 24)

Q: Can intelligible speech be regenerated from MFCC?
A: Yes, if the pitch and voicing information is present
# Implementation Results

## Bit rates:

<table>
<thead>
<tr>
<th>Mode</th>
<th>13 MFCCs:</th>
<th>24 MFCCs:</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCC only:</td>
<td>4 kbps</td>
<td>6.4 kbps</td>
</tr>
<tr>
<td>MFCC+pitch*:</td>
<td>4.5 kbps</td>
<td>6.9 kbps</td>
</tr>
</tbody>
</table>

* Reconstruction enabled

## Floating point run-time*:

<table>
<thead>
<tr>
<th>MFCC Compression/ Decompression</th>
<th>Full RECOVC Encoder</th>
<th>Full RECOVC Decoder</th>
</tr>
</thead>
<tbody>
<tr>
<td>~3.5 %</td>
<td>~9 %</td>
<td>~4 % (8 kHz)</td>
</tr>
<tr>
<td>-MFCC calculation (24 dim)</td>
<td>- Pitch Detection</td>
<td>~6.5 % (22 kHz)</td>
</tr>
<tr>
<td>-MFCC compression &amp; decompression</td>
<td>- MFCC calculation</td>
<td>- MFCC &amp; Pitch decompression</td>
</tr>
<tr>
<td></td>
<td>- MFCC &amp; Pitch compression</td>
<td>- Speech reconstruction</td>
</tr>
</tbody>
</table>

* Real-time percentage on Pentium II 266 Mhz running on Windows NT 4.0

## 32 bit fixed point implementation:

<table>
<thead>
<tr>
<th>Mode</th>
<th>CPU (MIPS*)</th>
<th>ROM (KB)</th>
<th>RAM (KB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full RECOVC</td>
<td>15</td>
<td>90</td>
<td>40</td>
</tr>
<tr>
<td>MFCC Comp./decomp. only</td>
<td>4</td>
<td>70</td>
<td>10</td>
</tr>
</tbody>
</table>

* Estimated, encoder plus decoder
- Signals are sampled at 8 kHz

<table>
<thead>
<tr>
<th>Original</th>
<th>Reconstructed (6.9 kbps)</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Original Image" /></td>
<td><img src="image2.png" alt="Reconstructed Image" /></td>
</tr>
<tr>
<td><img src="image3.png" alt="Original Image" /></td>
<td><img src="image4.png" alt="Reconstructed Image" /></td>
</tr>
</tbody>
</table>
Enabling DSR services for mobile devices

2.5 & 3G handsets, VoIP (RTP), ...

Members:

- IBM, Nokia, Motorola, Intel, Alcatel, Ericsson, Qualcomm, TI, FT, HP, Siemens, SpeechWorks, ...

April 2000 - First front-end standard for MFCC feature extraction & compression (ETSI ES 201 108)

1Q 2002 - Advanced front-end standard (noise robust)

New activity - "extension of front-end for tonal-language recognition and speech reconstruction"

- IBM is a major contributor to this activity (standardization of RECOVC technologies)
RECOVC Web Site: http://www.haifa.il.ibm.com/recovc/

- Overview
- Presentation
- Demonstration
- Conference papers

IBM Conference Papers:
- D. Chazan, G. Cohen, R. Hoory and M. Zibulski, "Speech reconstruction from mel-frequency cepstral coefficients and pitch frequency", ICASSP 2000

Contact person: Gilad Cohen (giladc@il.ibm.com)