Speech Compression

Application Scenarios

<table>
<thead>
<tr>
<th>Multimedia application</th>
<th>Live conversation?</th>
<th>Real-time network?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Video telephony/conference</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Business conference with data sharing</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Distance learning</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Multimedia messaging</td>
<td>No</td>
<td>Possibly</td>
</tr>
<tr>
<td>Voice annotated documents</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>

- **Key Attributes of a Speech Codec**
  - Delay
  - Complexity
  - Quality
  - Bit-rate
Key Attributes (cont’d)

- Delay
  - One-way end-to-end delay for real-time should be below 150 ms (at 300 ms becomes annoying)
  - If more than two parties, the conference bridge (in which all voice channels are decoded, summed, and then re-encoded for transmission to their destination) can double the processing delay
  - Internet real-time connections with less than 150 ms of delay are unlikely, due to packet assembly and buffering, protocol processing, routing, queuing, network congestion…
  - Speech coders often divide speech into blocks (frames) e.g. G.723 uses frames of 240 samples each (30 ms) + look-ahead time

Key Attributes (cont’d)

- Complexity
  - Can be very complex
  - PC video telephony: the bulk of the computation is for video coding/decoding, which leaves less CPU time for speech coding/decoding

- Quality
  - Intelligibility + naturalness of original speech
  - Speech coders for very low bit rates are based on a speech production model (not good for music - not robust to extraneous noise)
ITU Speech Coding Standards

<table>
<thead>
<tr>
<th>Standard</th>
<th>Bit rate</th>
<th>Frame size/ Look-ahead</th>
<th>Complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>G.711 PCM</td>
<td>64 Kb/s</td>
<td>0 / 0 ms</td>
<td>0 MIPS</td>
</tr>
<tr>
<td>G.726, G.727</td>
<td>16,24,32,40 Kb/s</td>
<td>0.125 / 0 ms</td>
<td>2 MIPS</td>
</tr>
<tr>
<td>G.722</td>
<td>48,56,64 Kb/s</td>
<td>0.125 / 1.5 ms</td>
<td>5 MIPS</td>
</tr>
<tr>
<td>G.728</td>
<td>16 Kb/s</td>
<td>0.625 / 0 ms</td>
<td>30 MIPS</td>
</tr>
<tr>
<td>G.729</td>
<td>8 Kb/s</td>
<td>10 / 5 ms</td>
<td>20 MIPS</td>
</tr>
<tr>
<td>G.723</td>
<td>5.3 &amp; 6.4 Kb/s</td>
<td>30 / 7.5 ms</td>
<td>16 MIPS</td>
</tr>
</tbody>
</table>

ITU G.711

- Designed for telephone bandwidth speech signal (3KHz)
- Does direct sample-by-sample non-uniform quantization (PCM)
- Provides the lowest delay possible (1 sample) and the lowest complexity
- Not specific for speech
- High-rate and no recovery mechanism
- Default coder for ISDN video telephony
**ITU G.722**

<table>
<thead>
<tr>
<th>Bit-rate</th>
<th>Frame size</th>
<th>Look-ahead</th>
<th>Complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>48,56,64 Kb/s</td>
<td>0.125 ms</td>
<td>1.5 ms</td>
<td>5 MIPS</td>
</tr>
</tbody>
</table>

- Designed for transmitting 7-KHz bandwidth voice or music (sampled at 16 KHz)
- Divides signal in two bands (high-pass and low-pass) which are then encoded with different modalities
- Quality not perfectly transparent, especially for music. Nevertheless, for teleconference-type applications, G.722 is greatly preferred to G.711 PCM because of increased bandwidth

**ITU G.726, G.727**

<table>
<thead>
<tr>
<th>Bit-rate</th>
<th>Frame size</th>
<th>Look-ahead</th>
<th>Complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>16,24,32,40 Kb/s</td>
<td>0.125 ms</td>
<td>0 ms</td>
<td>2 MIPS</td>
</tr>
</tbody>
</table>

- **ADPCM** (Adaptive Differential PCM) codecs for telephone bandwidth speech
- Can operate using 2, 3, 4 or 5 bit per sample
**ITU G.729, G.723**

<table>
<thead>
<tr>
<th></th>
<th>Bit-rate</th>
<th>Frame size</th>
<th>Look-ahead</th>
<th>Complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>G.723</td>
<td>5.3, 6.4 Kb/s</td>
<td>30 ms</td>
<td>7.5 ms</td>
<td>16 MIPS</td>
</tr>
<tr>
<td>G.729</td>
<td>8 Kb/s</td>
<td>10 ms</td>
<td>5 ms</td>
<td>20 MIPS</td>
</tr>
</tbody>
</table>

- **Model-based coders**: use special models of production (synthesis) of speech
  - Linear synthesis: feed a “noise” signal into a linear LPC filter (whose parameters are estimated from the original speech segment).
  - Analysis by synthesis: the optimal “input noise” is computed and coded into a multipulse excitation
  - LPC parameters coding
  - Pitch prediction
- Have provision for dealing with frame erasure and packet-loss concealment (good on the Internet)
- G.723 is part of the standard H.324 standard for communication over POTS with a modem

**ITU G.723 Scheme**

![ITU G.723 Scheme Diagram](image-url)
**ITU G.728**

<table>
<thead>
<tr>
<th>Bit-rate</th>
<th>Frame size</th>
<th>Look-ahead</th>
<th>Complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>16 Kb/s</td>
<td>0.625 ms</td>
<td>0 ms</td>
<td>30 MIPS</td>
</tr>
</tbody>
</table>

- Hybrid between the lower bit-rate model-based coders (G.723 and G.729) and ADPCM coders
- Low-delay but fairly high complexity
- Considered equivalent in performance to 32 Kb/s G.726 and G.727
- Suggested speech coder for low-bit rate (64-128 Kb/s) ISDN video telephony
- Remarkably robust to random bit errors

---

**Application Examples**

- **Video telephony/teleconference**
  - Higher rate, more reliable networks (ISDN, ATM) logical choice is G.722 (best quality - 7KHz band)
  - 56-128 Kb/s: G.728 is a good choice because of its robust performance for many possible speech and audio inputs
  - Telephone bandwidth modem, or less reliable network (e.g., Internet): G.723 is the coder of choice
Application Examples (cont’d)

- **Multimedia messaging**
  - Speech, perhaps combined with text, graphics, images, data or video (asynchronous communication). Here delay is not an issue.
  - Message may be shared with a wide community. The speech coder ought to be a commonly available standard.
  - For the most part, fidelity will not be an issue
    - G.729 or G.723 seem like good candidates

**Structured Audio**
What Is Structured Audio?

- Description format that is made up of **semantic information** about the sounds it represents, and that makes use of **high-level** (algorithmic) models

- **Event-list representation**: sequence of control parameters that, taken alone, do not define the quality of a sound but instead specify the ordering and characteristics of parts of a sound with regards to some external model

Event-list representation

- Event-list representations are appropriate to soundtracks, piano, percussive instruments. Not good for violin, speech and singing
- **Sequencers**: allow the specification and modification of event sequences
**MIDI**

- **MIDI (Musical Instrument Digital Interface)** is a system specification consisting of both hardware and software components that define interconnectivity and a communication protocol for electronic synthesizers, sequencers, rhythm machines, personal computers and other musical instruments.
- **Interconnectivity** defines standard cabling scheme, connectors and input/output circuitry.
- **Communication protocol** defines standard multibyte messages to control the instrument’s voice, send responses and status.

**MIDI Communication Protocol**

- The MIDI communication protocol uses multibyte messages of two kinds: **channel messages** and **system messages**.
- **Channel messages**: address one of the 16 possible **channels**.
  - **Voice Messages**: used to control the voice of the instrument.
    - Switch notes on/off
    - Send key pressure messages indicating the key is depressed
    - Send control messages to control effects like vibrato, sustain and tremolo
    - Pitch-wheel messages are used to change the pitch of all notes
- **Channel key pressure** provides a measure of force for the keys related to a specific channel (instrument).
**MIDI Files**

- **MIDI messages** are received and processed by a MIDI sequencer asynchronously (in real time)
  - When the synthesizer receives a “note on” message it plays the note
  - When it receives the corresponding “note off” it turns it off

- If MIDI data is stored as a **data file**, and/or edited using a sequencer, some form of “time stamping” for the MIDI message is required and is specified by the Standard MIDI file specifications.

---

**Sound Representation and Synthesis**

- **Sampling**
  - Individual instrument sounds (notes) are digitally recorded and stored in memory in the instrument. When the instrument is played, the note recording are reproduced and mixed to produce the output sound
  - Takes a lot of memory! To reduce storage:
    - Transpose the pitch of a sample during playback
    - Quasi-periodic sounds can be “looped” after the **attack transient** has died
  - Used for creating sound effects for film (Foley)
• **Original sound**

Sound Representation and Synthesis (cont’d)

• **Additive and subtractive synthesis**
  – Synthesize sound from the superposition of sinusoidal components (additive) or from the filtering of a harmonically rich source sound (subtractive)
  – Very compact but with “analog synthesizer” feel

• **Frequency modulation synthesis**
  – Can synthesize a variety of sounds such as brass-like and woodwind-like, percussive sounds, bowed strings and piano tones
  – No straightforward method available to determine a FM synthesis algorithm from an analysis of a desired sound
Application of Structured Audio

- Low-bandwidth transmission
- Interactive music applications
- Content-based retrieval

References

Immersive Audio

Applications of Immersive Audio

- Teleconferencing
- Telepresence
- Augmented and virtual reality for manufacturing and entertainment
- Air-traffic control, pilot warning and guidance systems
- Display for the visually or aurally impaired
- Home entertaining
Sound Localization

• The human ear-brain interface is uniquely capable of **localizing** and **identifying** sounds in a 3-D environment with remarkable accuracy
  
  • E.g., human listeners can detect time-of-arrival differences of about 7 μs

• Human hearing process is based on the analysis of input signals to the two ears for differences in:
  
  – **Intensity**
  – **Time of arrival**
  – **Direction filtering** by outer ear

Sound Localization (cont’d)

• For **short wavelength** (4-20 KHz) the listener’s head casts an acoustical shadow giving rise to a lower sound level at the ear farthest from the sound sources
  
  – Therefore for 4-20 KHz sound localization is based on **Interaural Level Difference (ILD)**

• At **long wavelength** (20 Hz - 1 KHz) the head is very small compared to wavelengths
  
  – In this case localization is based on perceived **Interaural Time Differences (ITD)**
Sound Localization (cont’d)

- Time or intensity differences provide source direction information only in the horizontal (azimuthal) plane
- In the medial plane time differences are constant and localization is based on spectral filtering
- Reflection and diffraction of sound waves from the head, torso, shoulder and pinnae, combined with resonances caused by the ear canal, form the physical bases for the Head-Related Transfer Function (HRTF)

Historical Overview

- **Stereo**: comes from the Greek στερεός meaning “solid or three-dimensional”
  - The two-channel association came about in the 1950’s because records could only encode two channels
- In actuality, film stereo sound started out with (and continues to use) a minimum of **four channels**
- Researchers at Bell Labs showed in the 1930s that a 3-channel system (left, center and right in the azimuthal plane) can represent lateralization and depth of desired sound with acceptable accuracy
**Historical Overview (cont’d)**

- Cinema stereo in the 1950s was using no less than 4 channels and as many as 7
- The ill-fated *quadraphonic* (4-channel) system (promoted in early 1970) attempted to capture and transmit information about the *direct sound* and the *reverberant sound* field.
  - Had standardization problems with encoding-decoding
  - Few consumers perceived any real advantage
  - Producers and recording engineers couldn’t agree on how to best use the extra channels

**Multichannel Surround Sound**

- The combination of wide-screen formats (such as CinemaScope) with *multichannel sound* was the film industry’s response in the early 1950s to the growing threat of television
- A monophonic channel reproduced over two loudspeakers behind the audience was known as the *effect channel*
- To create a more diffuse sound, a second channel reproduced over an array of loudspeakers along the side of the theater
**Dolby Stereo Digital 5.1**

- In 1992, Dolby create a new format called **Dolby Stereo Digital (SR•D)** that provided 5 discrete channels (left, center, right, and independent left and right surround) in a configuration known as **stereo surround**
- A sixth, low-frequency enhancement (**LFE**) channel was introduced
  - The frequency range of the LFE (0-120 Hz) is outside the localization range for human listeners in a reverberant room
  - LFE prevents the main speakers from overloading at low frequencies
  - Since it only needs about 1/10 the bandwidth of the others, the LFE channel is referred to as .1 channel

**Spatial Audio Rendering**

- **Goal of immersive audio**: reproduce 3-D sound fields that preserve the desired spatial location, frequency response, and dynamic range
- **General methods**:
  - **Head-related** (based on headphone reproduction)
  - **Nonhead-related** (based on loudspeaker reproduction)
Head-related Binaural Recording

- Also called **dummy-head stereophony**
- Attempts to accurately reproduce at each eardrum of the listeners the sound pressure generated by a set of sources and their interaction with the environment
- Such recordings can be made with specially designed probe microphones that are inserted in the listener’s ear canal or by using a dummy head microphone system that is based on average human characteristics

Head-related Binaural Recording (cont’d)

- Head-related binaural recordings are not widely used primarily due to limitations that are associated with headphone listening:
  - Individual HRTF information does not exist for each listener
  - There are large errors in sound position perception associated with headphones (especially for the most important visual direction, out in front)
  - Headphones are uncomfortable for extended periods of time
Nonheadroom-related Methods

- Use multiple loudspeakers to reproduce multiple channels
- Can convey precisely localized sound images that are primarily confined to the horizontal plane and diffuse (ambient) sound to the sides of and behind the listener
- They use left, right and center loudspeaker to help create a solidly anchored center-stage sound image plus two loudspeakers for the ambient surround sound field
- Problem: need to eliminate cross-talk to deliver the appropriate binaural sound field to each ear

Audio-visual Spatial Localization

- Vision also plays an important role in localization and can overwhelm the aural impression
- A mismatch between the aurally perceived and visually observed position of a particular sound causes a cognitive dissonance that can seriously limit the visualization enhancement produced by a immersive sound
- For professional sound engineers, a mere 4° offset in the horizontal plane between the visual and aural image is perceptible, whereas it takes a 15° before the average layperson will notice
References