IT694 M.Tech. Seminar
“Very low bit rate Speech Coding”

Chetan Vaity (02329901)

KReSIT,
Indian Institute of Technology,
Mumbai.
http://www.it.iitb.ac.in/~chetanv

Overview

- Speech Coding - Introduction and Applications
- Traditional low bit rate coders
- Vector quantization and other optimizations
- Variable length segments - random quantizer and segment network
- Phonetic vocoder
- Concatenative synthesis of waveforms
- Concatenating segment waveforms
- Frame concatenation - Cost functions, Viterbi search
- Comments

Introduction

- Process of coding speech signals for efficient transmission/storage
- Non-compressed: 64 kbits/s (scalar quantization)
- Medium rate: 8-16 kbits/s
- Low rate: 2.4-8 kbits/s
- Very low rate: Below 2.4 kbits/s

Applications

- Wherever large amount of pre-recorded speech is involved (talking books etc)
- Voice messages over the net
- Multimedia applications
- Digital telephony (real time)

Speech production model based coding

- Physiology and movements of the vocal tract are modelled
- A source exciting a vocal tract filter
- Linear Predictive Coding (LPC)
- LPC yields a set of “features” or “coefficients” for each frame
- Typical LPC vector: 8-12 spectral coefficients with 5-6 bits/coefficient, gain and pitch with 2-4 bits each, voiced/unvoiced with 1 bit
- 25 ms frames and about 60 bits/frame gives a 2400 bits/s vocoder

Vector Quantization

- Clustering in vector space
- Codebook size - quality vs search and storage

Speech Coding based upon Vector quantization - Buzo et al (1980)

- Gain separation
- Binary tree search of the codebook
- Benefit: 10 bits/frame vs 37 bits/frame for scalar quantization

Recent developments in vector quantization for speech processing - Wong et al (1981)

- Pitch and voicing separation in addition to gain separation
- Pitch and gain - once per 3 frames
- 32 branches/node tree search

Segment quantization for very low rate speech coding - Roucos et al (1982)

- Obtain segments with average duration comparable to phonemes
- Automatic segmentation - Changes in spectral time derivatives
- Threshold chosen such that approximately 11 segments/frame are obtained
- Distance measure for segments (equi-spaced spatial sampling)
- 140 dimensional vectors (14 spectral values x 10 spatial samples)
- Instead of clustering, random quantization
- 8000 segments (13 bits). Gain, pitch, timing (8 bits). Resulted in 231 bits/s for 11 segments/s

A segment vocoder at 150 bits/s - Roucos et al (1983)

- Segment network - restrict number of segment templates that can follow a given segment template
- Comparison: 1024 segment templates unrestricted *vs* Restricted to 256 segment templates. Almost no difference in quantization error

A Phonetic Vocoder - Picone et al (1989)

- Hidden Markov Model (HMM) based system
- A large acoustic phonetically transcribed database (60 phones)
- Phone recognition system is implemented using HMMs
- Premise: In spite of phone recognition errors, acoustic match is usually good
- 480 clusters used (8 per phone) - Approx 170 bits/s

Concatenative synthesis of waveforms

- The original waveform corresponding to the nearest codebook template is used for synthesis
- Instead of speech generation framework, concatenate short speech waveforms
- Better perceptive quality than LPC synthesized speech

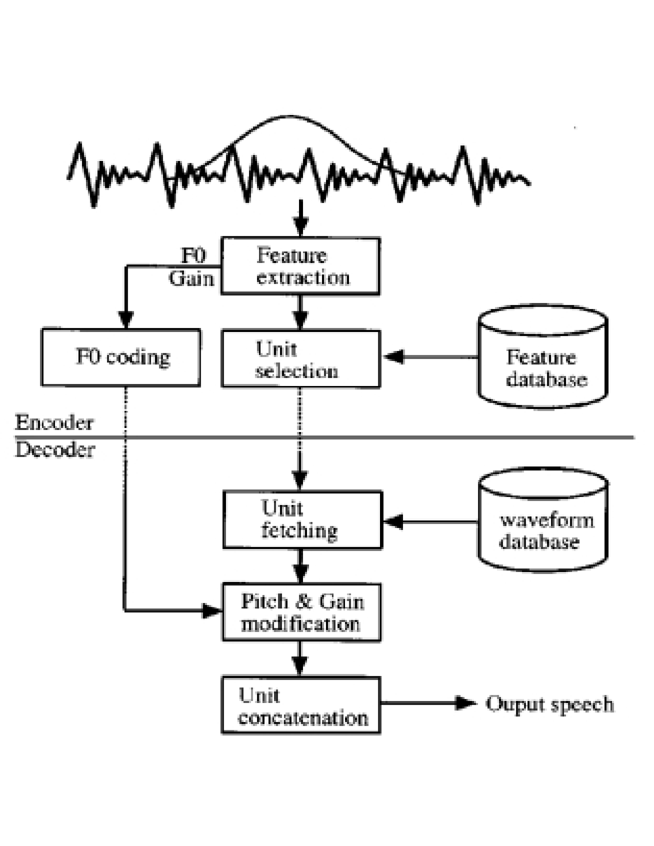
The waveform segment vocoder - Roucos et al (1985)

- Modification to the segment vocoder
- Pitch, energy and duration of the template waveforms are independently modified to match the input segment
- Speaker dependent
- Synthesized speech has a *choppy* quality due to segment boundary discontinuities
- Potential for transmitting speech at 300 bits/s

A very low bit rate Speech Coder based on a recognition/synthesis paradigm - Lee and Cox (2001)

- Frames are used as units of selection
- Mel-frequency cepstrum coefficients (MFCCs) used
- 460,000 frames of 10ms each
- Feature database and Waveform database
- Pitch and gain are also transmitted

Block diagram of the coder



Unit Selection

- State transition network (similar to HMMs)
- State ocupancy cost - distance between a database unit and a target
- Transition cost - estimate of the quality of concatenation of two consecutive units
- Total cost for a sequence of n units:
  $C(t_1^n, u_1^n) = \sum_{i=1}^{n} C^t(t_i, u_i) + \sum_{i=2}^{n} C^c(u_{i-1}, u_i)$
- Minimization using Viterbi's algorithm

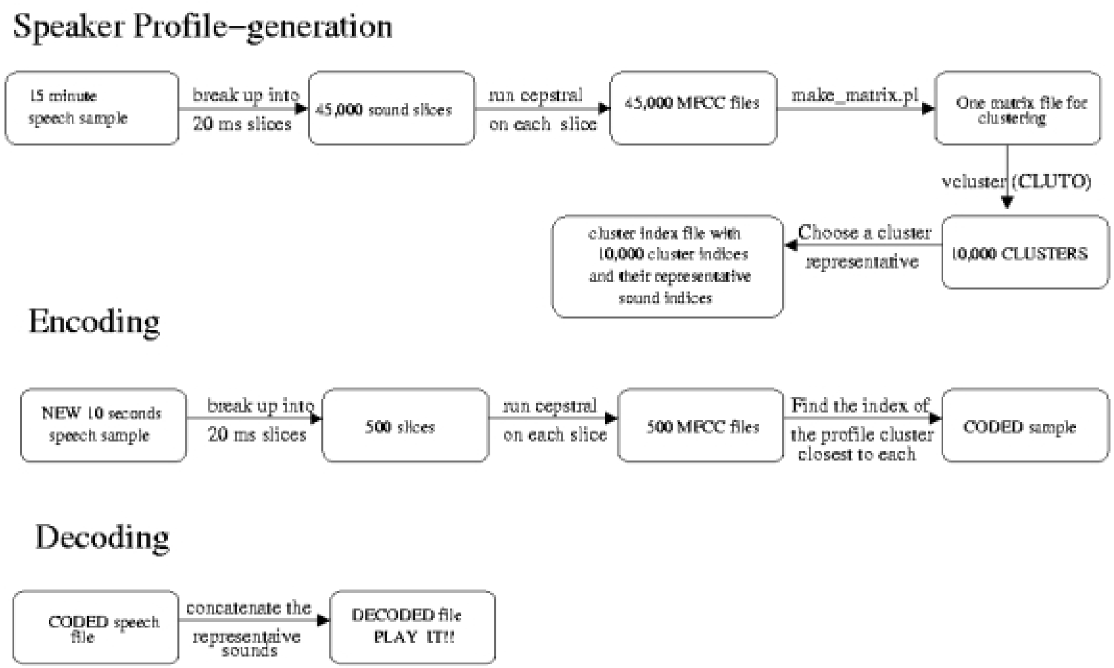
Pruning the search space for Viterbi's algorithm

- To reduce computational time for search in 460,000 units
- Choose only those units whose spectral envelopes are close to that of the input frame
- Vector quantization using clustering - Units quantized into the same code vector are selected as candidate frames. Codebook size: 10 or 11 bits
- Phonetically labelled speech database used - Illegal transistions eliminated

Coding the selected unit sequence

- Concatenation cost is set to zero if two units are consecutive in the database
- The resulting unit sequence has many sequential frames
- *Run-length coding* is used. Only transmit start frame and number of following consecutive frames
- About 800 bits/s

IT608 course project



Thank you!