Data Reduction
Speech Data Rates

- **Waveform**
  - 8 bits/sample x 8000 samples/sec
    = 64000 bps (Telephone grade)

- **MelCep**
  - 12 coefficients x 32 bits = 384 bits/frame
  - 384 bits/frame x 100 frames/sec
    = 38400 bps

- **Acoustic Phonetic**
  - Assume 64 phonetic states
  - 6 bits/phone x 5 phones/sec = 30 bps!

NOTES

Our digital representations of speech are extremely redundant.

At minimum acceptable rates and sample resolutions, PCM waveform coding costs 64k Bits Per Sample (bps). This may be reduced by as much as 50% or so using careful encoding/compression techniques.

Using an analysis technique like the Mel-cepstrum, we can substantially reduce the data rates and produce better quality the the 8x8000 PCM. Just using raw floating point cepstral parameters @ 12 coefficients per frame and 100 frames per second, we are down to 38,400 bps. Again, careful encoding and compression techniques can bring the number down a good bit. Our present telephone grade speech encoders based on LPC analysis & resynthesis run at about 9600 bps. Low quality but intelligible LPC coded speech can get down to 2400 bps.

But, what is the real data rate for human speech communication? If we must identify about 64 allophones at an average of 5 per second, the rate looks like about 30 bps.

It would be nice to wring as much of the redundancy as possible out of the acoustic speech signal before we bother to do any pattern recognition on it. Acoustic analyses that isolate just the feature we need, whenever we need them seems like the way to go.
Formant Coding

- Formants seem to be the acoustic features of greatest perceptual importance.
- Speech is not a constant “frame rate”
- Assume we need 4 poles & 1 zero + F0 to adequately describe perceptually important information (encode in ~42 bits)
- Assume 3 snapshots per phone
- Data rate = 42 x 3 x 5 = 630 bps

NOTES

Lower frequency poles/zeros need to be coded with greater precision than higher frequency formants. Bandwidth can be coded with fairly coarse resolution. Source coding needs voiced/voiceless, pitch, amplitude, tilt.

From Flanagan (1972) *Speech Analysis, Synthesis, and Perception*:

- Formant frequency difference limen (JND) = 3 - 5% of F
- Bandwidth JND: F1 = 20%, F2 = 40% F3 = 40% (assumption)
- F0 JND = .3 - .5% of F0; other Not known.

F1 needs 6 bits; F2 & F3 need 4 bits = 14 bits
B1 needs 6 bits; B2 & B3 need 3 bits = 12 bits
F0 needs 8 bits; other source need 8 bits = 16 bits

Total = 42 bits
Problems with Formant Coding

• Difficult to track
  – Formant merging
  – additional/spurious formants
  – best for low pitch voices

• Costly to compute
  – Root Solving after LPC or related analysis
  – Analysis by Synthesis
Data-Based Approach

- Determine representation from data
- Record basic acoustic features like mcep
- Use unsupervised learning to
  - Characterize data using specific number of bits
  - Characterize data with specific error tolerance
Vector Quantification

- K-means algorithm
- LBG algorithm
  - Version 1
  - Version 2
- Suboptimal but faster
  - Binary tree
  - Unbalanced tree

NOTES
  • Arbitrary partition into K clusters
  • Compute cluster centroids
  • Reassign feature vectors
  • Repeat until nothing changes

LBG (Linde-Buzo-Gray) Version 1
  • Choose arbitrary set of K codes
  • Assign feature vectors per present codes
  • Compute centroids
  • Replace previous codes with these
  • If distortion still high, repeat

LBG Version 2
  • Calculate one centroid from all data
  • Split all centroids (old values +- epsilon)
  • Recalculate centroids
  • Reassign feature vectors
  • Repeat until max number of centroids reached or error low enough
VQ Training Implementation

\[ c_a = \frac{1}{N} \sum c_{ai} \]

Init primary Centroid

Split all centroids

Reassign vectors

Recompute Centroids

Compute Distortion

Y

Done

Done

k < K

Y

D > D_{crit}

N

N
Hierarchical VQ

Start

Perform Standard VQ

Done?

Start

Initialize New Codebook

Compute Residuals:
\[ R_{nm} = C_{mn} - V_{mn} \]
\[ 0 \leq n < N \]
\[ 0 \leq m < M \]
\[ 0 \leq k < K \]
\[ V_n \in C_k \]