Signal Modeling Techniques in Speech Recognition

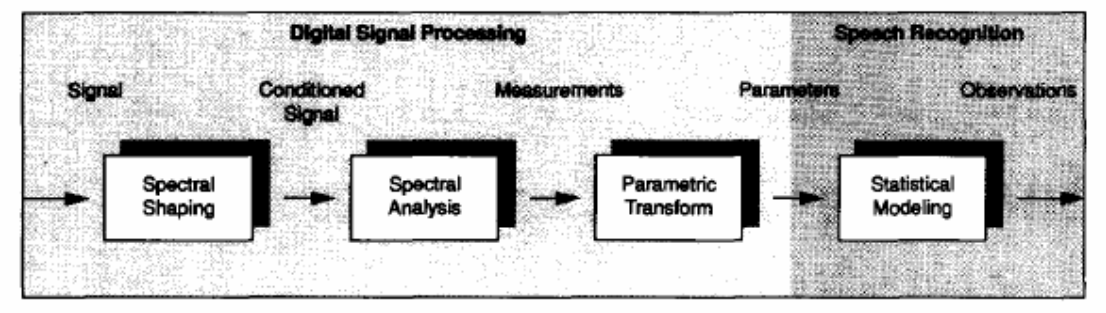
Hassan A. Kingravi

Outline

- Introduction
- Spectral Shaping
- Spectral Analysis
- Parameter Transforms
- Statistical Modeling
- Discussion
- Conclusions

1: Introduction

- Signal Modeling:
  The main focus of the paper. Signal modeling is the process of converting a speech sample to an observation vector in a probability space.
- Approach:



- Goals:
  Desire parameters that are **perceptually meaningful**, **invariant**, and that capture **temporal correlation**.

2: Spectral Shaping

- Analog-to-Digital Conversion
- Digital Filtering

  Create a data representation of the signal with as high an SNR as is possible.

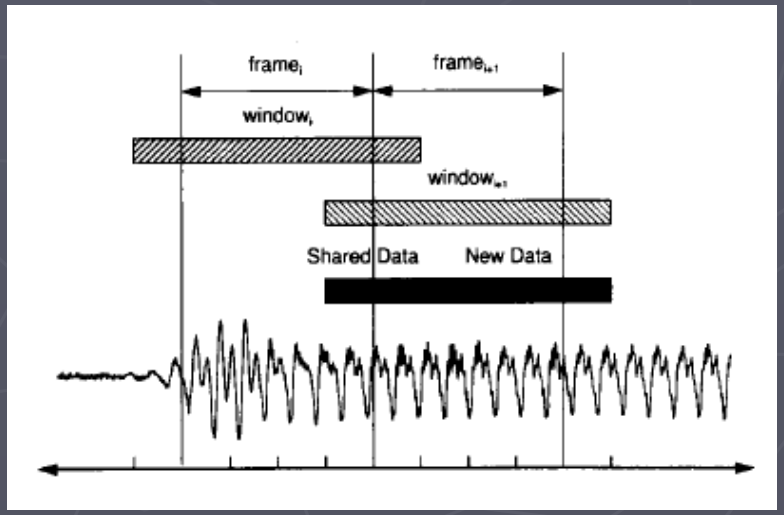
  Once conversion is finished, the signal is filtered to emphasize important frequencies in the signal. This is performed via the use of Finite Impulse Response filters. For e.g. hearing is more sensitive in the 1-KHz region of the spectrum, so this is emphasized.

3: Spectral Analysis

- Step 1: Extraction of fundamental frequency and power.
- Fundamental frequency:
  Measured via different methods;
  a) Gold-Rabiner: multiple measures of periodicity in the signal and votes among them.
  b) NSA: average magnitude difference function and DA of multiple voicing measures
  c) Dynamic Programming: evaluate several measures of correlation and spectral change in the signal, and compute an optimal fundamental frequency and voicing pattern
  d) Cepstrum

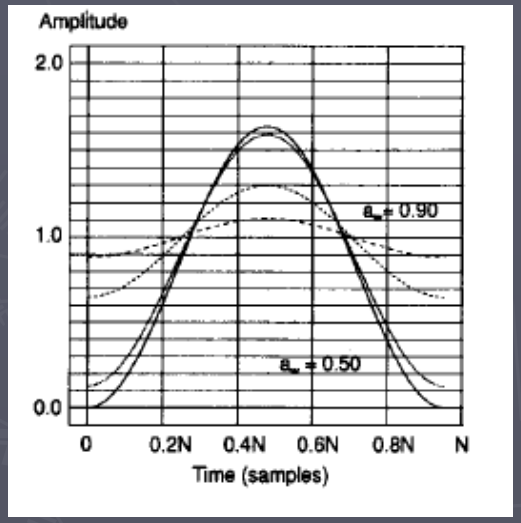
3: Spectral Analysis

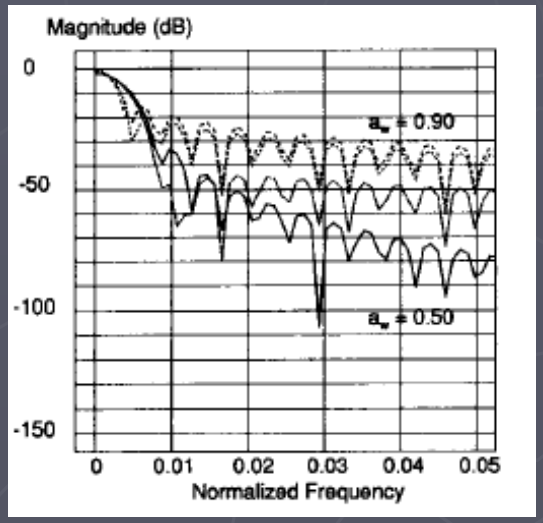
- Signal Power:
  Average of its energy (energy can be thought of as the area underneath the curve.) The formulation of power in signals uses a weighting function called a window, to favor samples in the center of the window in a frame-by-frame traversal. This, combined with redundancy, creates a smoothing effect.



3: Spectral Analysis

- Hanning window example:
  Smoothing is important because in some areas, the power of a signal varies rapidly.





3: Spectral Analysis

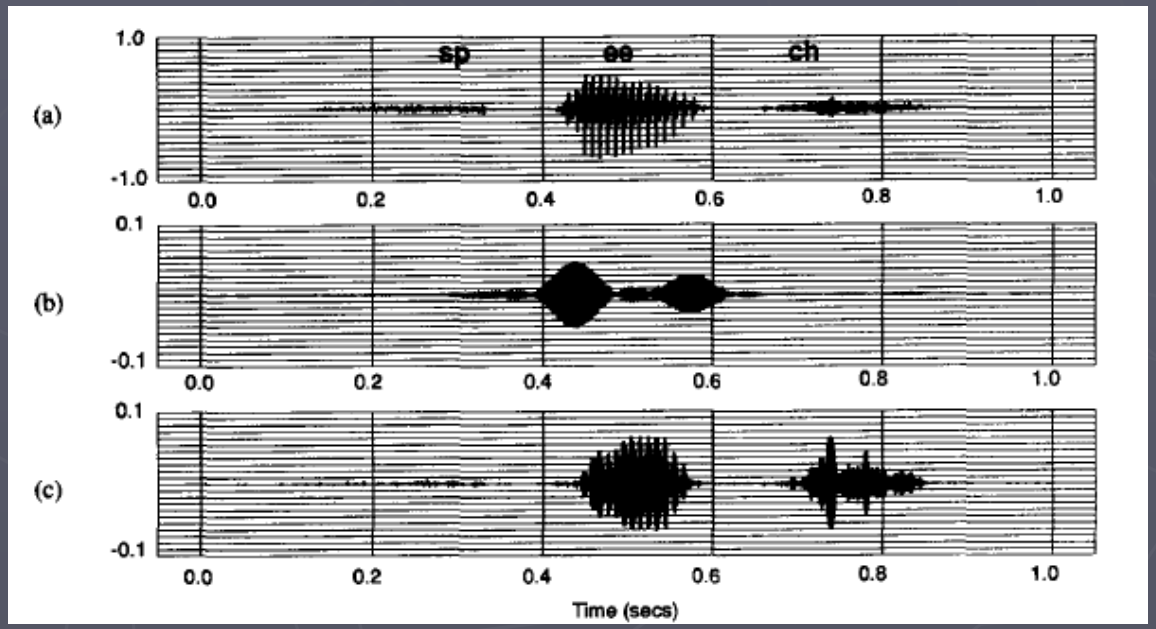
- Step 2: Spectral Analysis, proper.
- Digital Filter Bank
- Fourier Transform Filter Bank
- Cepstral Coefficients
- Linear Prediction Coefficients
- LP-Derived Filter Bank Amplitudes
- LP-Derived Cepstral Coefficients

Digital Filter Bank

- Motivations:
  Physiological, i.e. the "place theory" of hearing, and **perceptual**, i.e. frequencies of complex sounds with a certain bandwidth cannot be individually identified.
- Critical Bandwidth
  Human auditory system cannot distinguish between frequencies close to one another; the higher the frequency, the larger the interval; this is the critical band for a frequency.
- Filter Bank
  Collection of linear phase FIR bandpass filters arranged linearly across the *Bark* or *mel* scales. Used in systems that attempt to emulate human auditory processing. Merit: certain filter outputs correlated with certain speech sounds.

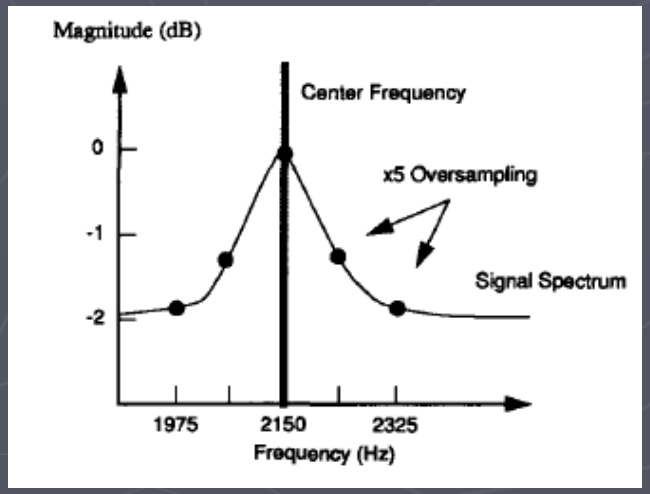
Digital Filter Bank

Index	Bark Scale		Mel Scale	
	Center Freq. (Hz)	BW (Hz)	Center Freq. (Hz)	BW (Hz)
1	50	100	100	100
2	150	100	200	100
3	250	100	300	100
4	350	100	400	100
5	450	110	500	100
6	570	120	600	100
7	700	140	700	100
8	840	150	800	100
9	1000	160	900	100
10	1170	190	1000	124
11	1370	210	1149	160
12	1600	240	1320	184
13	1850	280	1516	211
14	2150	320	1741	242
15	2500	380	2000	278
16	2900	450	2297	320
17	3400	550	2639	367
18	4000	700	3031	422
19	4800	900	3482	484
20	5800	1100	4000	556
21	7000	1300	4595	639
22	8500	1800	5278	734
23	10500	2500	6063	843
24	13500	3500	6964	969

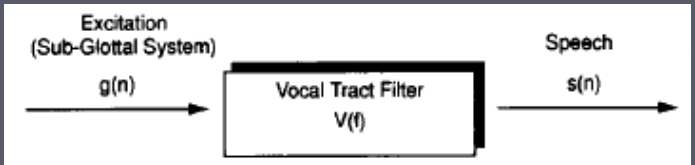


Fourier Transform Filter Bank

- Motivations:
  FBs have non-uniformly spaced frequency samples. If we perform a DFT of the signal, we go to the frequency domain, and can get the instance at any frequency. The DFT (or the FFT) also oversamples the spectrum; the value for each filter in the bank is calculated as an average of several frequencies, resulting in smoother estimates.
- Method:
  Sample spectrum at frequencies shown in last table. Often discard low-amplitude regions using a dynamic range threshold.



Cepstral Coefficients



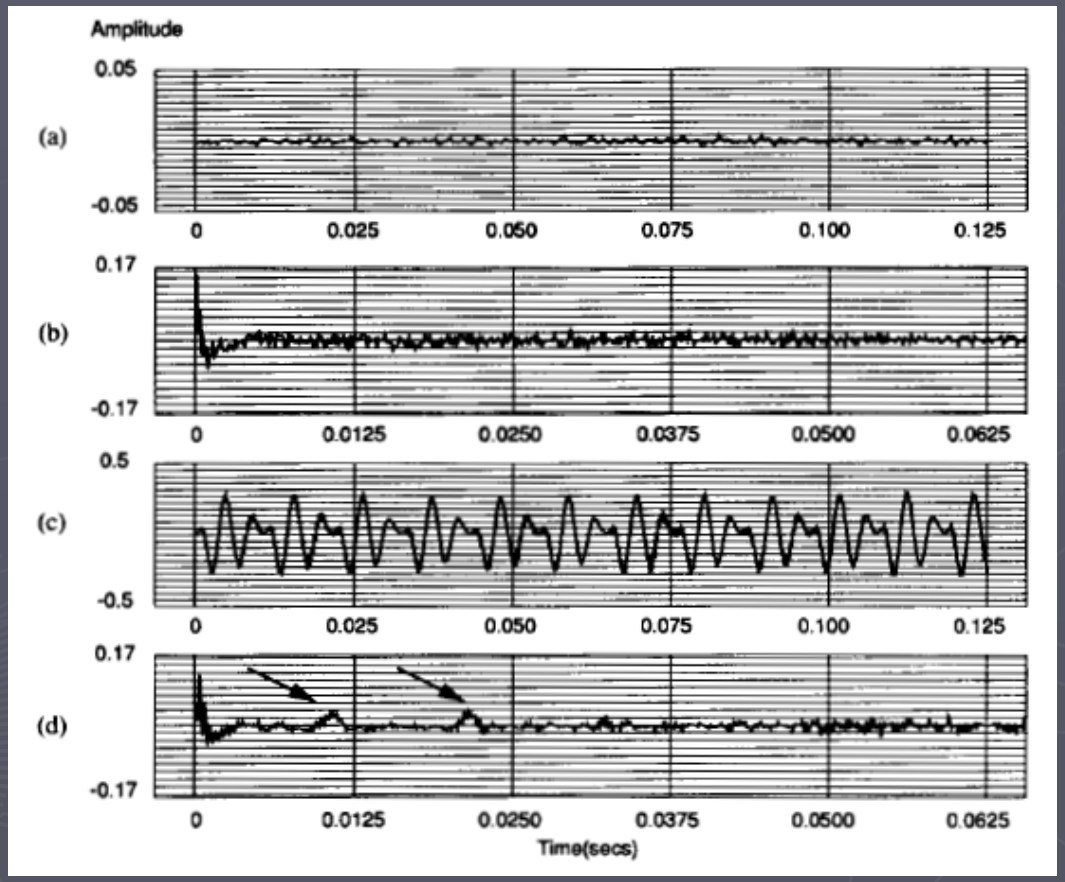
- Motivations:

  Process is nonlinear. Theory; there are waves with high frequency (excitation) and waves with low frequency (vocal tract) that are superimposed in the spectrum. Need holomorphic techniques obeying generalized rule of superposition; logarithmic in nature. This separates the excitation and vocal tract shape components.

- Method:

  We take the inverse DFT of the log spectral magnitudes, calling this the cepstrum. Low order terms of the cepstrum represent vocal tract shape and high order terms represent excitation.

Cepstral Coefficients



Linear Prediction Coefficients

- Method:
  An autoregressive process that parametrically fits the spectrum; you model the signal as a linear combination of its previous samples.

$$s(n) = -\sum_{i=1}^{N_{\mathrm{LP}}} a_{\mathrm{LP}}(i)s(n-i) + e(n)$$

  Minimize the error of this equation to find the *p a*'s, the LPC coefficients. The greater the value of *p*, the better the model.

- Ignoring Low-Amplitude Regions:
  Dynamic thresholding using autocorrelation is used. We add a small amount of noise to the signal, which prevents the LP model from modelling sharp nulls in the spectrum.

LP-Derived Filter Banks

- Motivations:
  Similar to the other filter banks, except we sample the LP model at the given frequencies, instead of the spectrum. This is done because the LP model gives more robust spectral estimates.
- Caveat:
  As DSP technology has developed further, difference between approaches not so great any more.

LP-Derived Cepstral Coefficients

- Method:
  Recall that the cepstral coefficients are calculated from the FFT of the log amplitudes. The LP-derived cepstral coefficents are the LP equivalent, where we take the logarithm of the inverse filter (remember, the LP equation is a filter). The number of cepstral coefficients usually equivalent to the number of LP coefficients.
- Difficulties:
  The coefficients calculated reflect a linear frequency scale; work needs to be done to make it nonlinear.

4: Parameter Transforms

- Motivations:
  The previous process gives us absolute measurements. Now we generate signal parameters from the measurements via differentiation and concatenation. Output; parameter vector with raw estimates of signal.
- Differentiation:
  Using the concept of a gradient, we take approximate derivatives of the signal to highlight changes in variation.
- Concatenation:
  Take as input a large matrix of all the measurements called X. Using auxiliary matrices, we perform all the necessary filtering operations discussed, including weighting, differentiation and averaging; results in creation of single parameter vector per frame that contains all desired signal parameter.

5: Statistical Modeling

- Motivations:
  Assumption; signal parameters generated via some underlying multivariate probability distribution. Job is to find the model; treat the data as signal observations. Uses topics from:
- Multivariate Statistical Models
- Distance Measures

Multivariate Statistical Models

- Motivations:
  Quantities with different numerical scales and that are correlated are mixed with each other; lot of redundancy. Scales can be re-adjusted; correlation is more difficult to deal with.
- The Whitening Transform:
  Assuming that the process we are modeling is Gaussian in nature, we can use a prewhitening transform to decorrelate the parameters.

$$\overline{y} = \boldsymbol{\Psi}(\bar{\nu} - \bar{\mu}_{\nu}) \qquad \boldsymbol{\Psi} = \boldsymbol{\Lambda}^{-1/2}\boldsymbol{\Phi}^{\dagger}$$

  This involves computing the eigenvalues and eigenvectors of the covariance matrix. The eigenvalues can be further used to determine which eigenvectors are the most significant, leading to a reduction in features.

  Note: need data for this for mean and variance calculations.

Multivariate Statistical Models

- Vector Quantization:
  If the Gaussian model is not appropriate, need a non-parametric representation. Hypothesize a discrete distribution and try to fit it. Vector quantizers compress the distribution; physiological precedents include the assumption that a finite set of stationary vocal tract shapes exists.
  Need to estimate a vector quantizing codebook *Q,* which contains a finite set of 'ideal' vectors, which will replace an input vector. Secondly, must maximize P(y|*Q*).
  Latter is simple. Former is done by use of the K-Means algorithm.

Multivariate Statistical Models

- K-Means Algorithm:
  Given the training data, and a number, say *k*, of clusters, place the *k* points in the space as far apart as possible. For each element in the data, assign it to the cluster closest to it. Once this is done, recompute the centroids of the clusters. Repeat this until the centroids no longer move. This results in the *k* groups we need, and the centroids become the components of *Q*.

Distance Measures

- Motivation:
  The k-means algorithm requires the definition of a distance measure in the vector space, without which it cannot proceed.
- Properties:
  Distance measures must satisfy the properties of nonnegativity, symmetry, and the triangle inequality.
- Examples:
  Most famous form is Euclidean. An interesting approximation that's presented is the dot product form of the Euclidean distance, which is obviously very cheap to evaluate.

Discussion

- Overview:
  Author gives a table of methods used in the community as of 1993; not really important.
- Comments:
  Neural Network based classification systems employ filter banks (tying in with cognition and the auditory system etc.)

  Cepstral coefficients dominant acoustic measurement.

  FFT-derived cepstral coefficients are more common than LP-derived ones. FFT is still popular because of its immunity to noise.

Summary

- Overview:
  Presented the various processes involved in the modeling of a signal and its conversion to a meaningful representation, one necessary for analysis.