EECS225d Audio Signal Processing in Humans and Machines **Spring 1995**

Lecture 17: Febuary 27, 1995

Lecturer: Dr. Hynek Hermansky *Scribe: Warner Warren*

Perceptual linear predictive analysis of speech

17.1 Speech Recognition Systems

A speech signal contains linguistic information along with a lot of other extraneous information such as the health of the speaker, his or her emotional state, identity etc. Speech recognition systems are designed to extract the linguistic information from the speech signal and discard everything else. Two methods exists for doing this. One can train the recognizer over all non-linguistic information in speech, or one can understand and suppress the non-linguistic information prior to processing the speech (See figure17.1). The second method is the preferred over the first, because the first method is expensive and may even be impossible.

17.2 Perceptual formant structure v.s. linear prediction

If one were to take a short segment of speech and compute its power spectrum one would see that the spectral envelope represents the formant structure of the speech. Hence, one might determine the formant structure of a segment of speech by extracting its power spectral envelope. The problem with this approach is that vocal tracts vary in shape and size from speaker to speaker, making the formants speaker dependent. For example, compare the male and female spectra shown in figure17.2. Notice that while the spectral envelopes have roughly the same shape, the positions and relative sizes of the spectral peaks are quite different.
The peaks of the spectral envelope convey much of the linguistic information in the speech signal. However, experiments with perceptual phonetic distance reveal the perceived phonetic variation in formant positions is much smaller than the Linear Prediction (LP) cepstral variation would lead us to expect. See table 17.1.

LP Cepstral Distance (Hermansky and Javkin, ASA 1986)

$\pm 10\%$	F1(700Hz)	F2(1000HZ)	F3(2500HZ)
	0.72	1.0	5.4

Perceptual Phonetic Distance (Klatt ICASSP 1982)

± 10	F1(700Hz)	F2(1000HZ)	F3(2500HZ)
	0.75	1.0	0.08

Table 17.1: Perceptual distance v.s. LP distance for 1st 3 formants

What accounts for this disparity between the perceived variation in formant frequencies among speakers and the variation of the lp spectra? One explanation is that most of what we perceive about the formant structure of speech is largely determined by the characteristics of the front of the speaker's vocal tract. This could explain why even though children have much smaller bodies than adults, and hence much different vocal tracts, their speech is still recognizable since their heads are approximately the same size as adult heads. Figure17.3 shows the spectrum of an adult male /i/ and that of a child's /i/. Notice the difference in the LP spectra between the child's /i/ and the adult male's /i/. The childs' spectral peaks occur at different frequencies than they do in the adult.

17.3 Speech perception

17.3.1 History

In 1665, Isaac Newton synthesized the succession of vowel sounds, /u/→/o/→/a/→/e/→/i/, with a single tube of varying length by gradually pouring beer out of bottle and blowing into it.

This experiment was repeated by Gunner Fant, in 1955, with harmonic tones (sinusoids). He asked people which vowel sound a particular tone represented. He found that there was only moderate correlation between their responses and the actual formant frequencies of the vowels. However, he did find that there was strong correlation between the mean resonant frequency of the front end of the vocal tract and the tones.

In 1960 Fant found that there was a perceptual correlation between F2 and F_{front}. It appeared that people were listening to something which was speaker independent. These observations suggest that the front end of the vocal tract determines many of the speaker independent properties of speech. Here are some properties of the front end of the vocal tract which make it useful for the speaker independent recognition task:

- easy to manipulate
- grows only about 10% during a lifetime
- front cavity resonance covers F2 and F3 of adult speech
- resonates in the frequency range where hearing is most sensitive

17.3.2 Perceptual spectral features

People integrate (smooth out) a lot of spectral information, particularly when perceiving speech. When listening to ordinary sounds people integrate about 1 bark of the spectrum; for discrimination of speech sounds, people people seem to integrate about 3.5 barks of spectrum. Therefore, from a speech perspective, people may not detect all spectral peaks just the gross shape of the spectral envelope. A little more history:

- In 1885, Helmholtz classified vowels as single peak vowels (/a/,/o/,/u/) and two peak vowels (/e/,/i/,/y/).
- In 1954,Delattre observed that two spectral peaks are sufficient to simulate front vowels and only one spectral peak is needed to simulate rear vowels.
- In 1960, Fant proposed that that the first spectral peak be kept at the first formant frequency F1. for the second spectral peak he used an effective second formant given by F2'= f(F1,F2,F3,F4 ..).
- In 1985, Chistovich observed that two spectral peaks are perceptually merged when they are closer than 3.5 Barks.

17.3.3 Basic psycho-acousic knowledge relevant to perception of speech

The following are key observations from psycho-acoustic experimentation:
a) frequency resolution is linear up to about 800HZ and logarithmic above 800HZ
b) Low frequencies mask higher ones (asymmetric critical-band filters)
c) Hearing most sensitive at mid-frequencies (equal loudness curves)
d) Loudness of sound proportional to cube root of sound intensity
e) Larger than 1 bark spectral integration hypothesized in speech perception

Perceptual Linear prediction (PLP) is method of spectral estimation which uses the above information to produce a model which has proven itself superior in many ways to the LP model for the speaker independent recognition task.

17.4 Perceptual linear prediction

Perceptual linear prediction is a technique for speech analysis that uses three psycho-acoustic concepts to estimate the auditory spectrum:

1. critical-band spectral analysis
2. the equal loudness curve
3. the intensity power law.

The PLP algorithm given below uses each of these concepts:

1. Compute power spectrum
2. Warp the frequency axis according to the bark scale
3. Convolve with critical-band masking curve and down-sample
4. Perform equal-loudness preemphasis
5. Perform intensity-loudness (cube root) warping

The Bark scale is linear at low frequencies and logarithmic at high frequencies. This effectively compresses the higher frequencies into a narrow band. The critical band masking, symmetric frequency domain convolution on the Bark-warped frequency scale, then allows low frequencies to mask the high frequencies while at the same time smoothing the spectrum, an effect consistent with the psycho-acoustic results cited above. Moreover, since the Bark-frequency spectrum is now smoother, and many of the rapid Bark-frequency variations have been filtered out, we can down sample in the frequency domain. The equal loudness preemphasis models the unequal sensitivity of of human hearing at different frequencies. The cube root warping is an approximation the which simulates the relation between intensity and loudness. The resulting spectrum is then modeled by a 5th order all pole filter. Experimentation reveals that the 5th order model is the optimal PLP model for cross-speaker word identification with a 4kHz BW. The PLP algorithm is diagramed in figure17.4. See Hermansky(1989) for a complete description of the PLP method.

The 5th order PLP model produces a spectrum with at most 2 spectral peaks. The first formant is approximate by the first PLP peak and the effective second formant is modelled by the other peak when the peaks are widely separated. As illustrated in figure17.5, when the peaks are close together (less than 3-4 barks), both peaks are merged into a single PLP peak somewhere between the two peaks. This behavior seems consistent with the experiments reported by Chistovich. See Hermansky(1989) p1746. In experiments on real articulatory data it was shown that there is a strong correlation between the perceptually estimated F2' and the F2' estimated by the PLP method.

17.4.1 PLP advantages over LP

- PLP approximates the speaker independent effective second formant
- PLP emphasizes F1 and F2 and deemphasizes high frequencies in contrast with LP which emphasizes high frequencies
- PLP reduces disparity between voiced and unvoiced speech
- PLP peaks are relatively insensitive to vocal tract length
- PLP estimates are highly correlated with the relatively speaker independent front of the vocal tract
- PLP reduces the sensitivity of ASR front ends to changes high frequencies and increases the sensitivity to changes in F2 and F1. While, contrary to the results of psycho-acoustic experimentation reported in by Flanagan, LP is actually more sensitive to changes in F3

17.5 Other Auditory-like warping methods

1. Mel LPC
2. Mel warped LPC
3. Mel cepstrum

17.6 Closing comments

- It pays to put some perceptual information into the system.
- Not much information is lost by smoothing if it is done in a way consistent with human hearing. It is believed that we would not be able to make sense of speech without ignoring some of the fine spectral detail.

17.7 Conclusions we would like to make

- The message is in the front part of the vocal tract which does not grow too much during one's lifetime and which we can easily manipulate
- Formants depend on both the front cavity length and shape and the length of the tract. This implies formants carry both the message and the speaker dependent information
- Human speech perception simplifies the formant structure by extracting the effective second formant F2', which is associated with the front part of the vocal tract

17.8 Conclusions we can make

- The child produces the vowels /a/ and /i/ by simulating the front part of the adult vocal tract

- In articulatory synthesis of the front to mid-vowels, the PLP esimated F2' correlates well with the resonance frequency of the uncoupled front cavity in spite of the radical changes in the vocal tract length and consequently also in the formant structure

- Movements of the front part of the tongue in the production of the vowel sentences by two male speakers correlate well with PLP estimated F2'

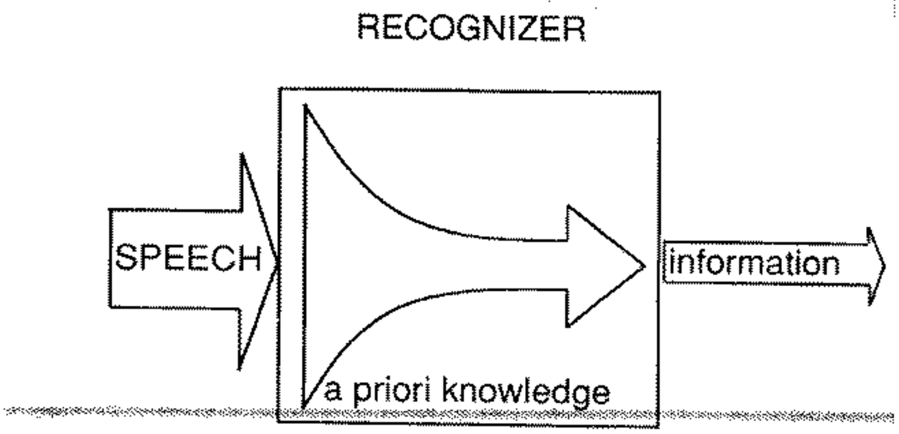

Figure 17.1: A speech recognizer

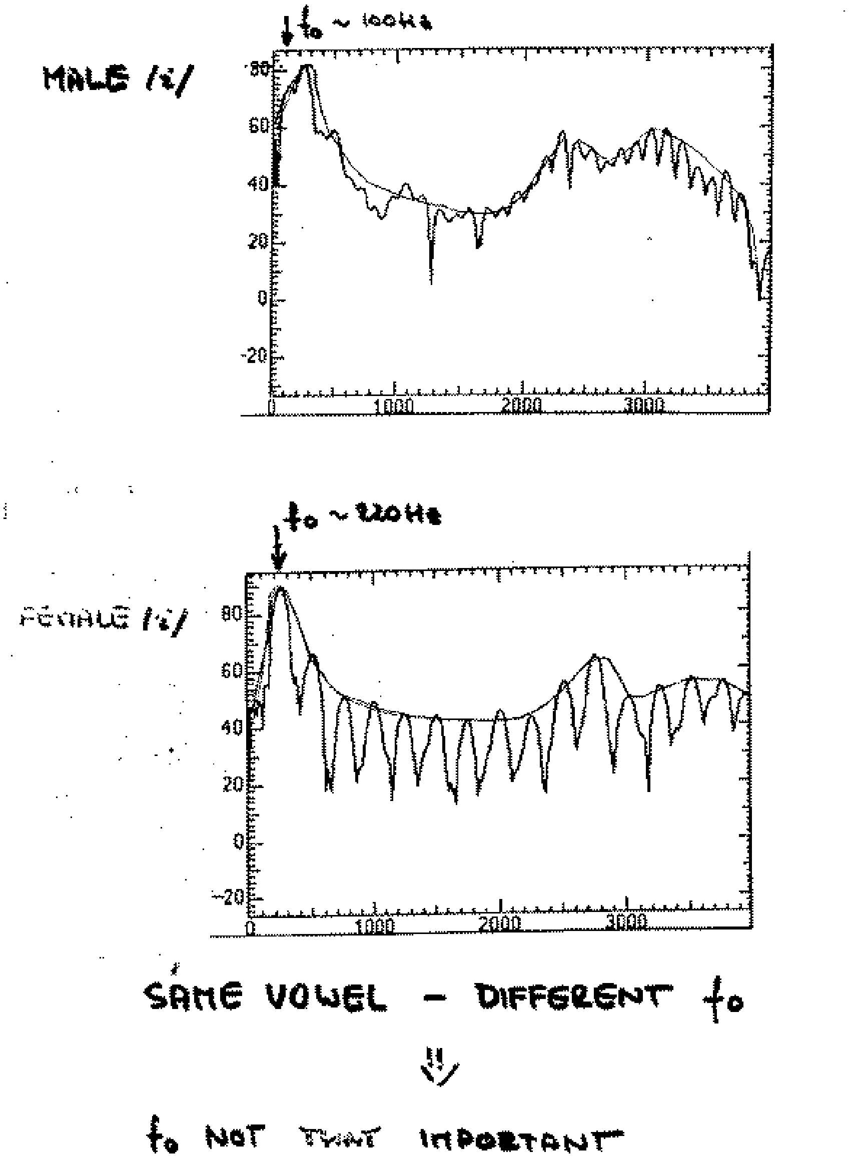

Figure 17.2: male and female spectra for /a/

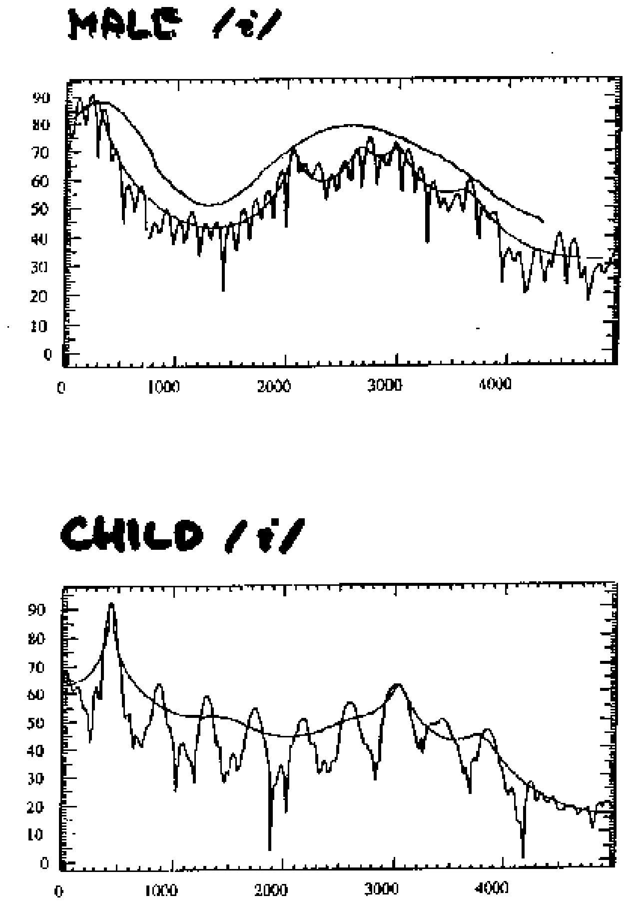

Figure 17.3: Child's /i/ and adult male /i/

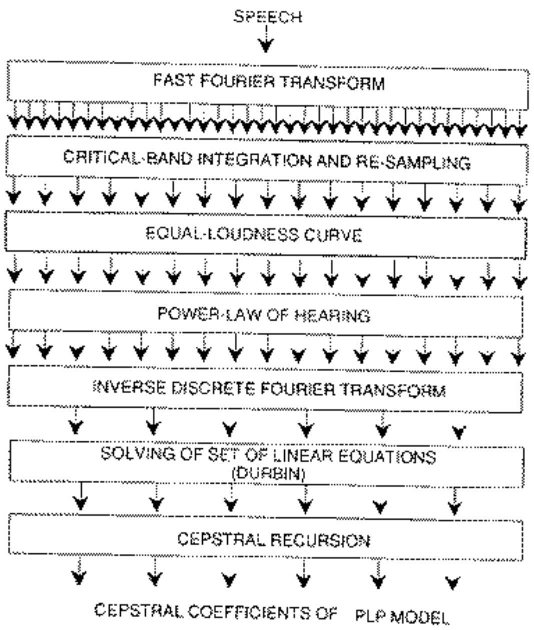

Figure 17.4: The PLP algorithm

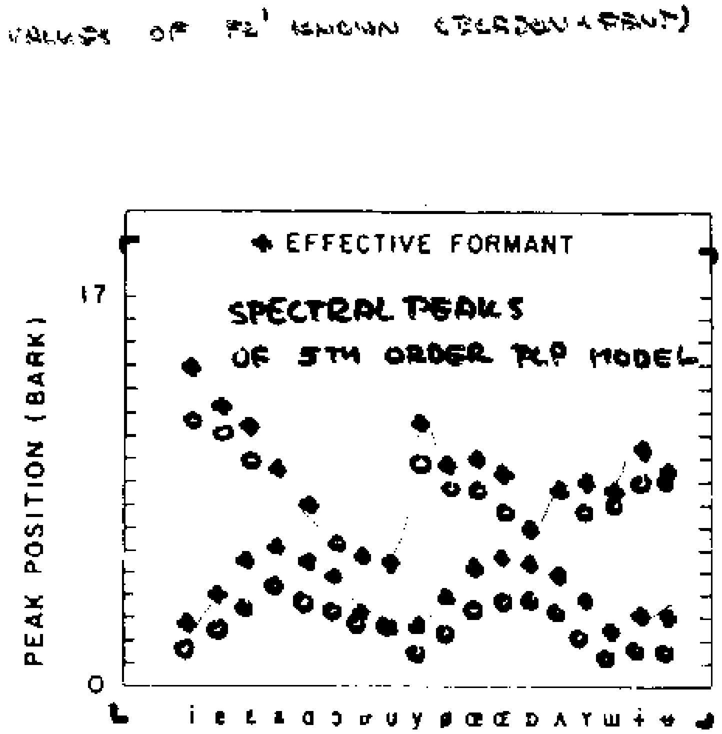

Figure 17.5: Fant's perceptually estimated peaks v.s. PLP peaks

17.9 References

Hermansky,H.(1989). "Perceptual linear predictive analysis of speech." J. Acoust. Soc. Am 87(4), April 1990. 1738-1752.