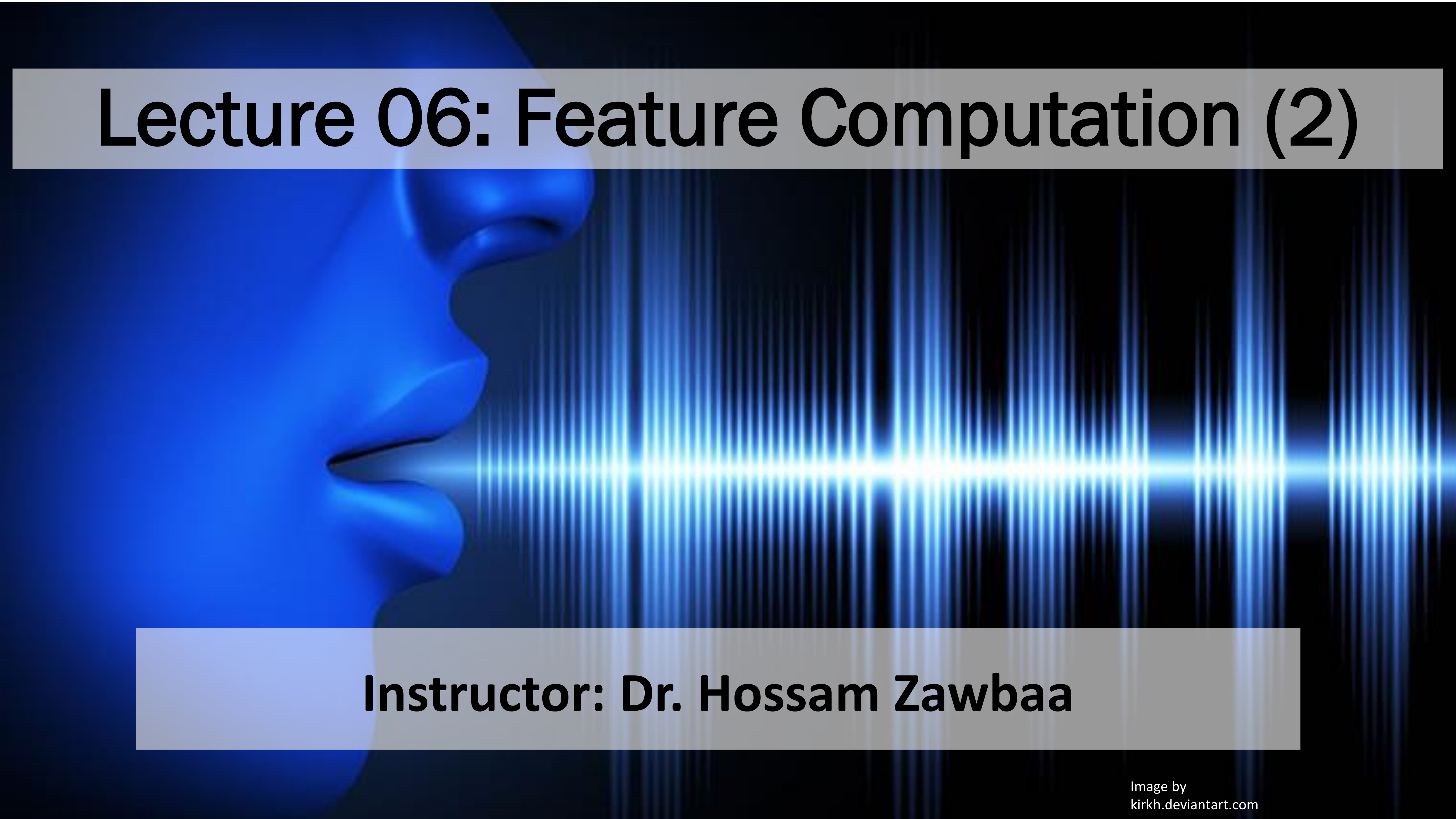
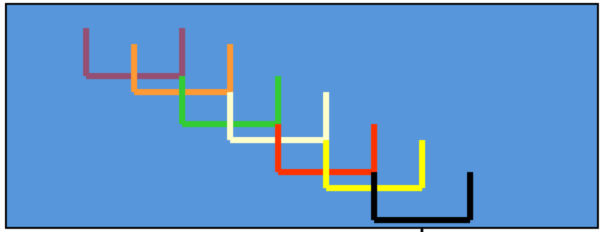
Lecture 06: Feature Computation (2)



Instructor: Dr. Hossam Zawbaa

Image by
kirkh.deviantart.com

The process of parametrization



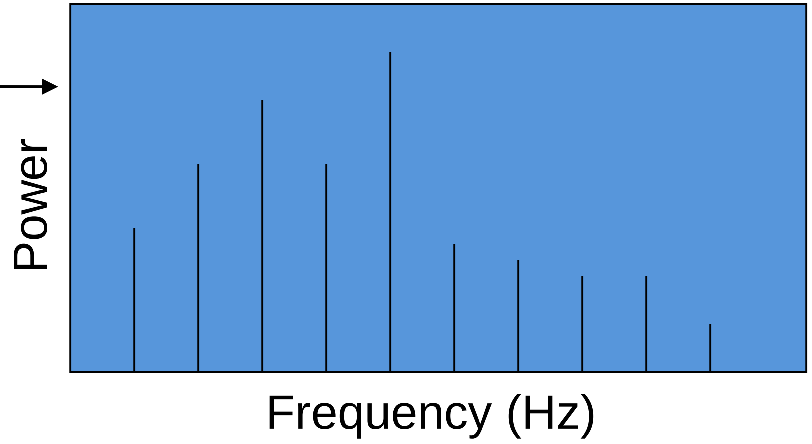
Each segment is pre-emphasized

Pre-emphasized segment

The pre-emphasized segment is windowed

Pre-emphasized and windowed segment

The DFT of the segment, and from it the power spectrum of the segment is computed



= power spectrum

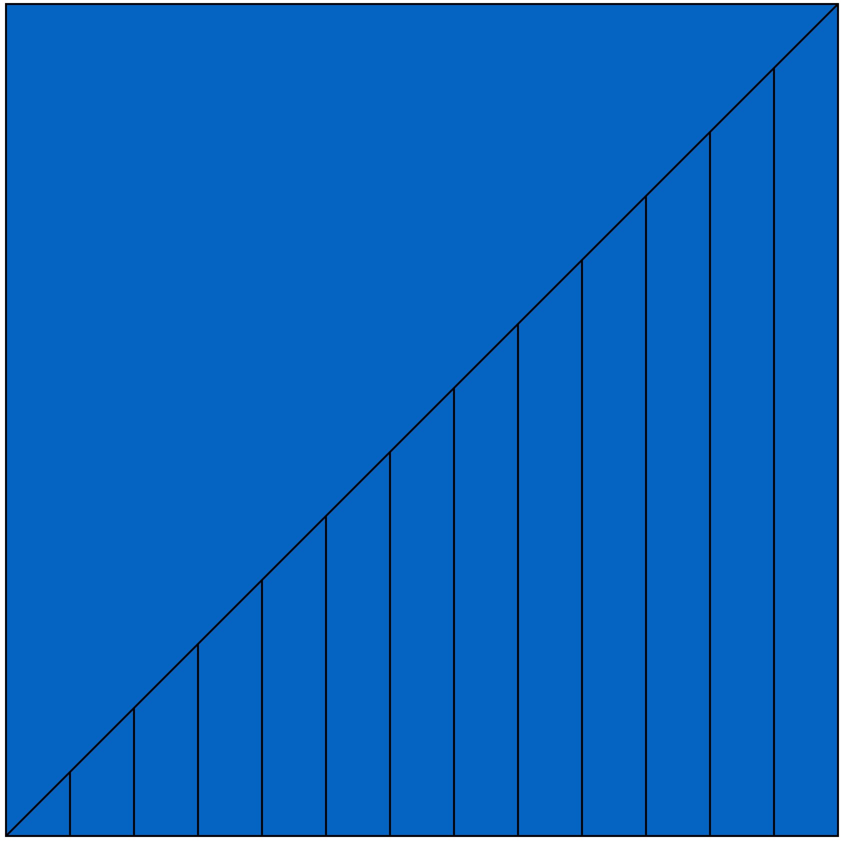
Auditory Perception

- Conventional Spectral analysis decomposes the signal into a number of linearly spaced frequencies
  - The resolution (differences between adjacent frequencies) is the same at all frequencies
- The human ear, on the other hand, has non-uniform resolution
  - At low frequencies we can detect small changes in frequency
  - At high frequencies, only gross differences can be detected
- Feature computation must be performed with similar resolution
  - Since the information in the speech signal is also distributed in a manner matched to human perception

Matching Human Auditory Response

- Modify the spectrum to model the frequency resolution of the human ear
- *Warp* the frequency axis such that small differences between frequencies at lower frequencies are given the same importance as larger differences at higher frequencies

Warping the frequency axis



Linear frequency axis: equal increments of frequency at equal intervals

Filter Bank

- Each hair cell in the human ear actually responds to a *band* of frequencies, with a peak response at a particular frequency

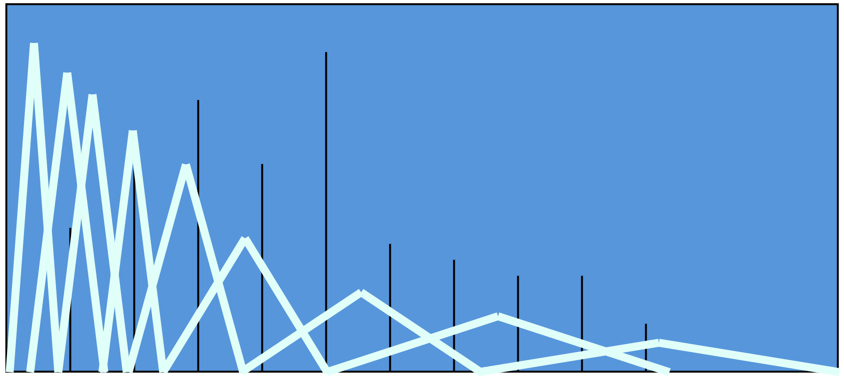
- To mimic this, we apply a bank of “auditory” filters
    - Filters are triangular
        - An approximation: hair cell response is not triangular
    - A small number of filters (40)
        - Far fewer than hair cells (~3000)

The process of parametrization

For each filter:

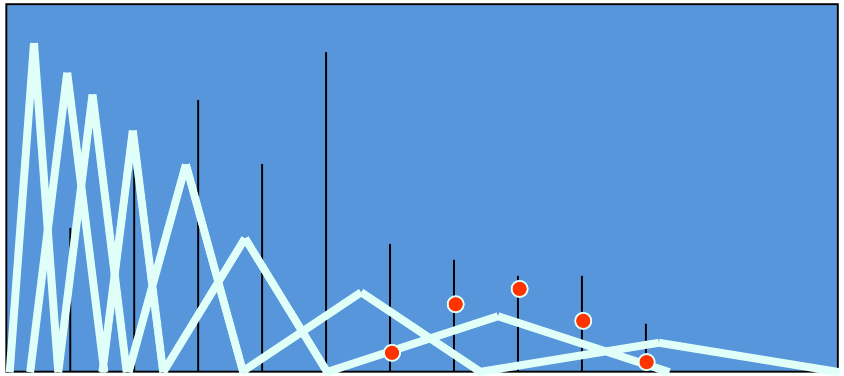
Each power spectral value is weighted by the value of the filter at that frequency.

This picture shows a bank or collection of triangular filters that overlap by 50%



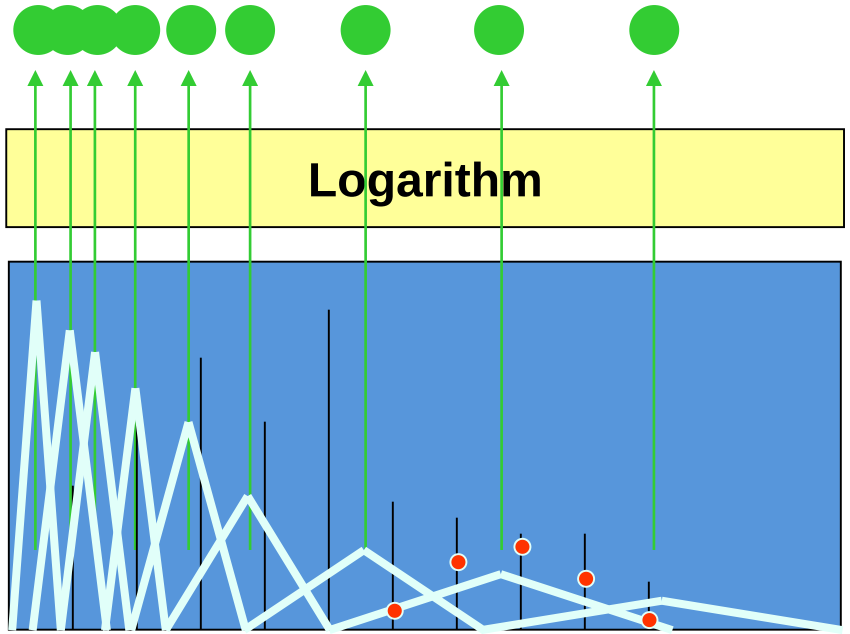
The process of parametrization

For each filter:
All weighted spectral values are integrated (added), giving one value for the filter



The process of parametrization

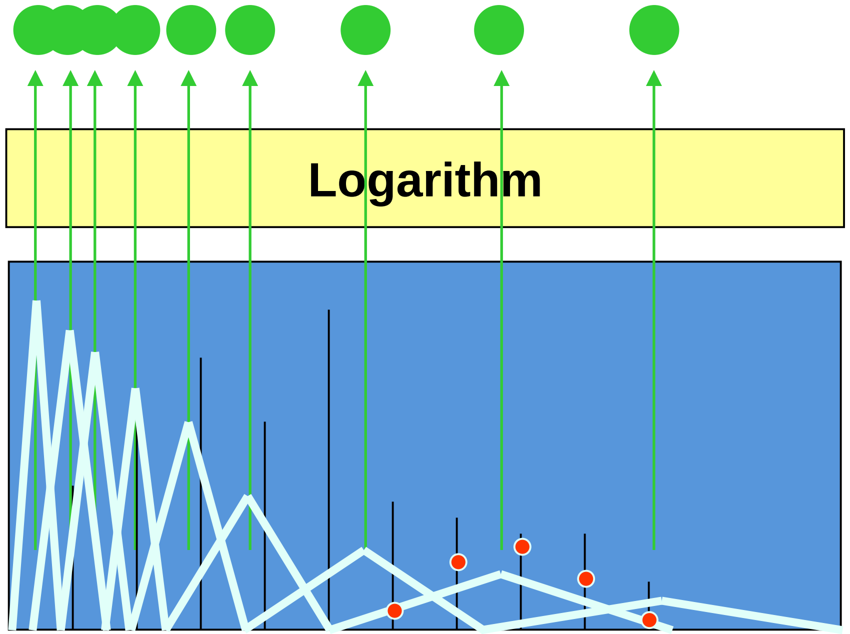
All weighted spectral values for each filter are integrated (added), giving one value per filter



Additional Processing

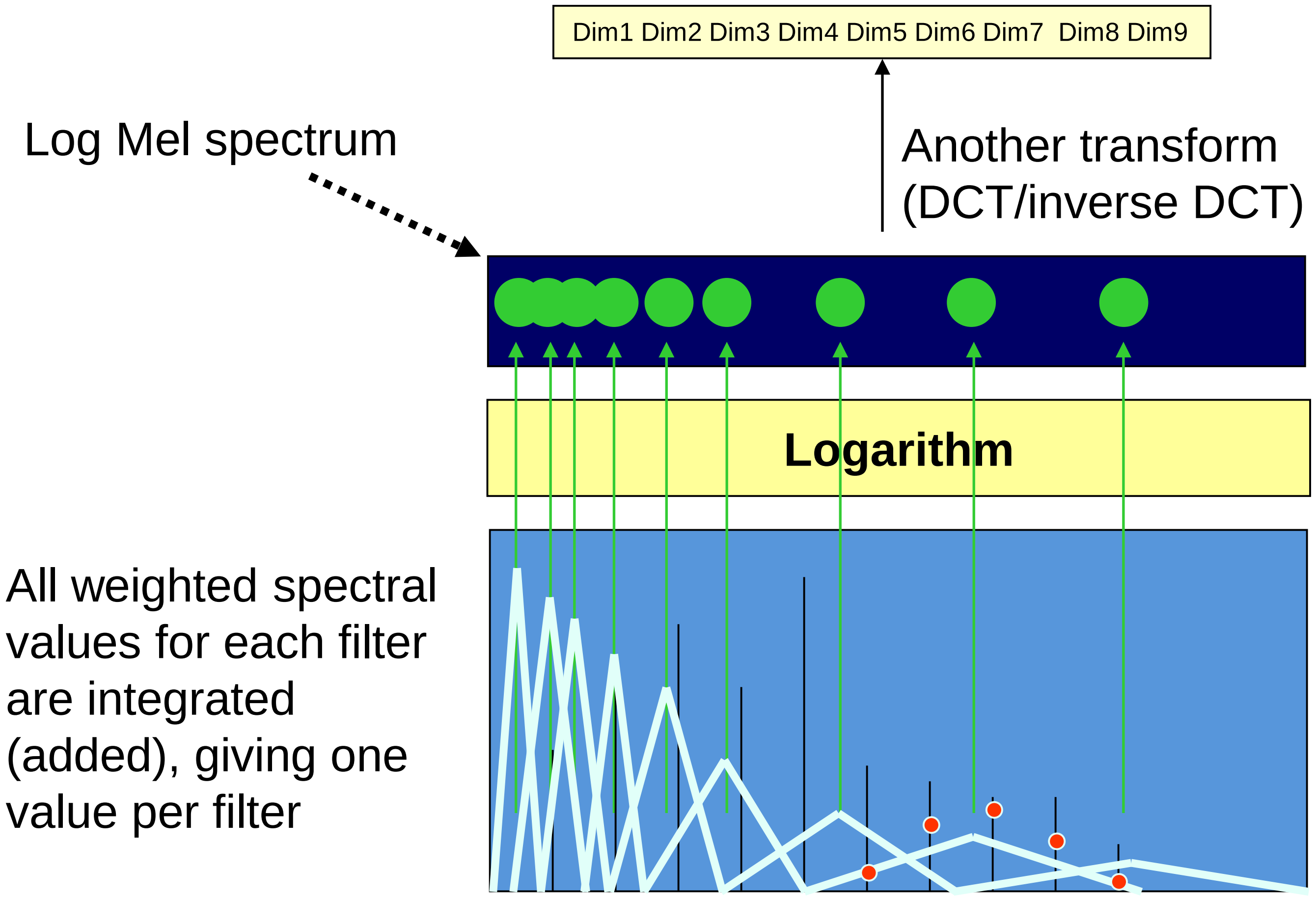
- The **Mel spectrum** represents energies in frequency bands
  - Highly unequal in different bands
    - Energy and variations in energy are both much greater at lower frequencies
    - May dominate any pattern classification or template matching scores
  - High-dimensional representation: many filters
- Compress the energy values to reduce imbalance
- Reduce dimensions for computational tractability
  - Also, for generalization: reduced dimensional representations have lower variations across speakers for any sound

The process of parametrization

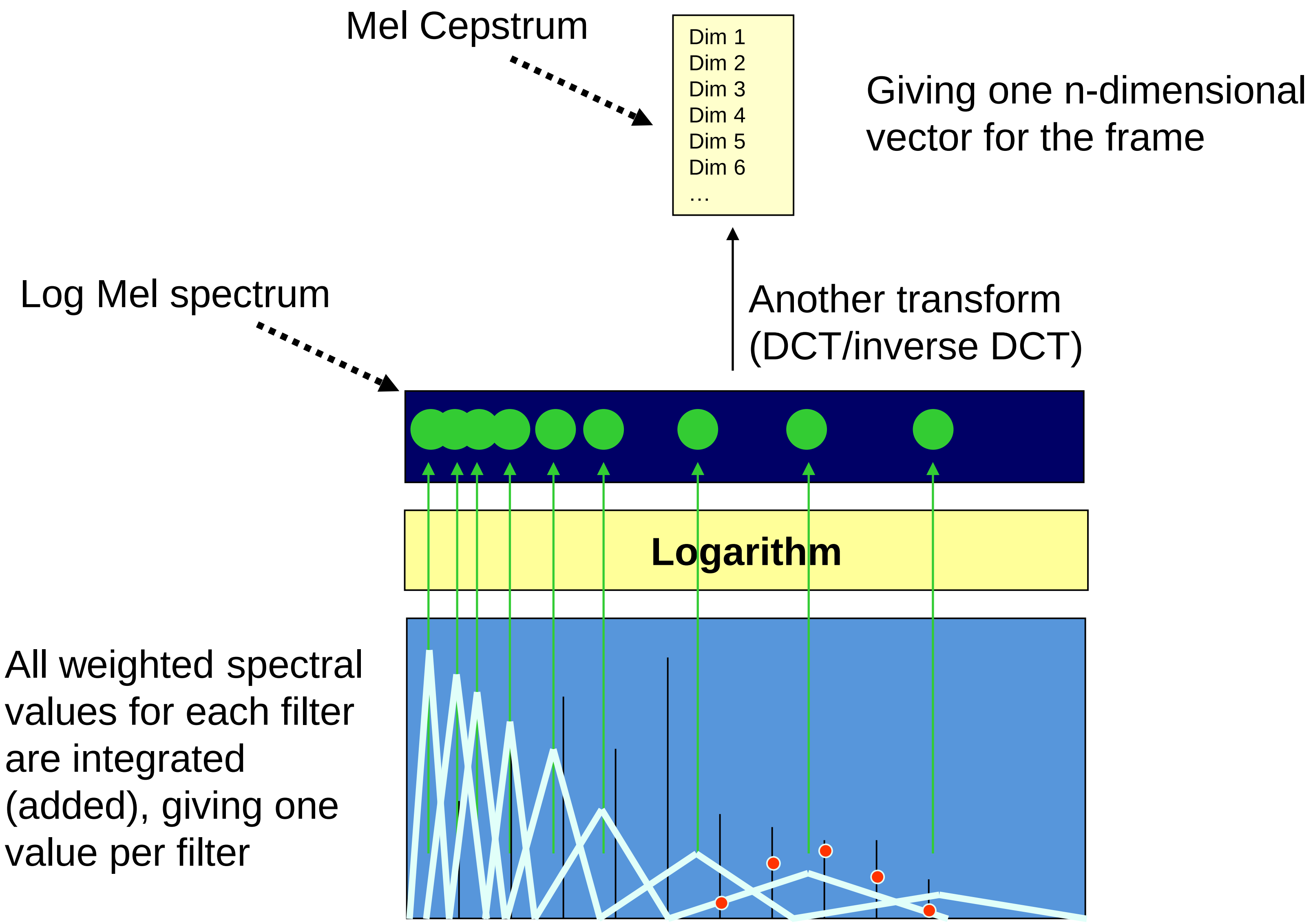


All weighted spectral values for each filter are integrated (added), giving one value per filter

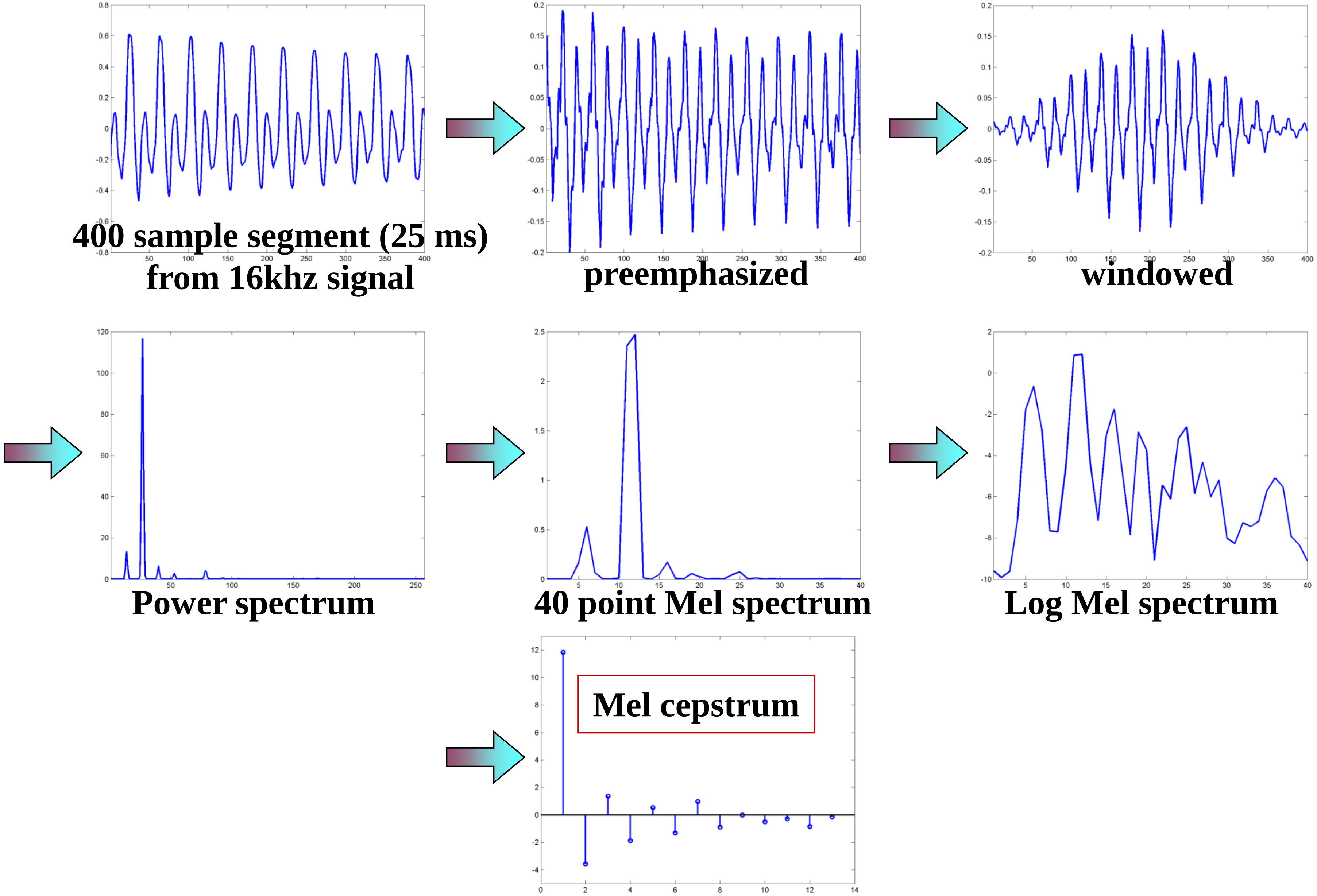
The process of parametrization



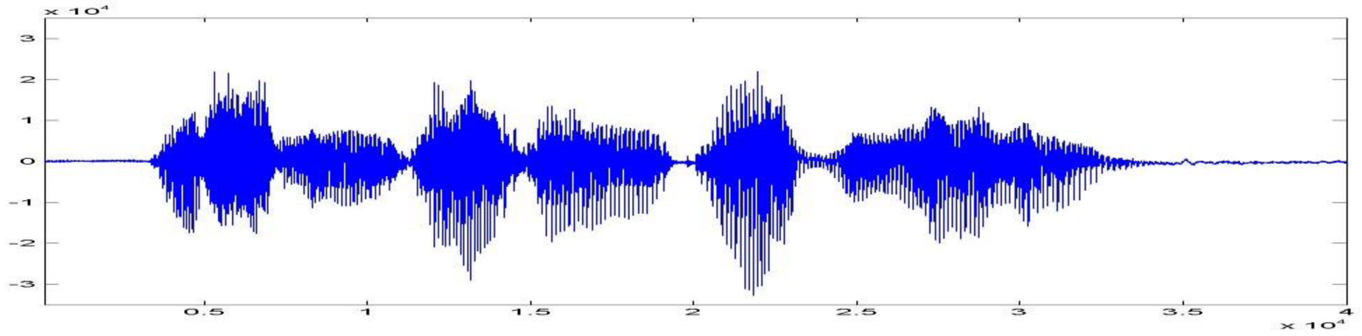
The process of parametrization

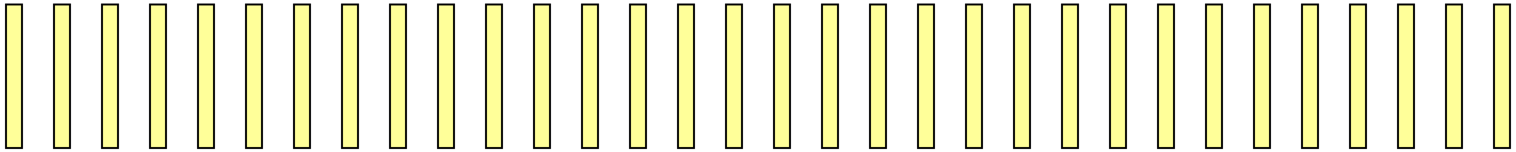


An example segment



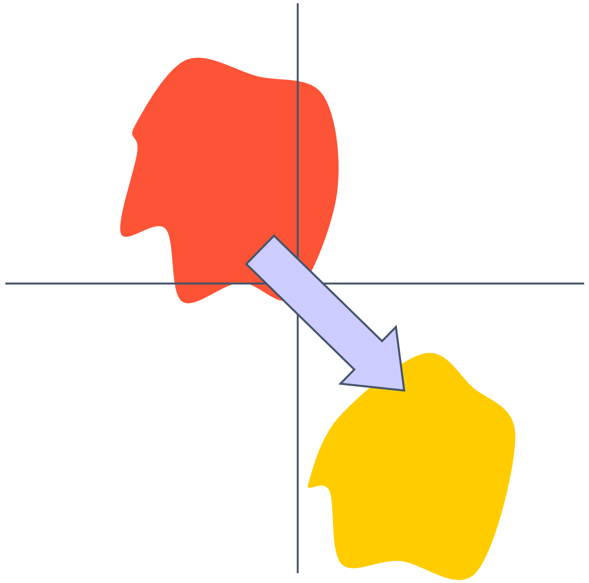
The process of feature extraction





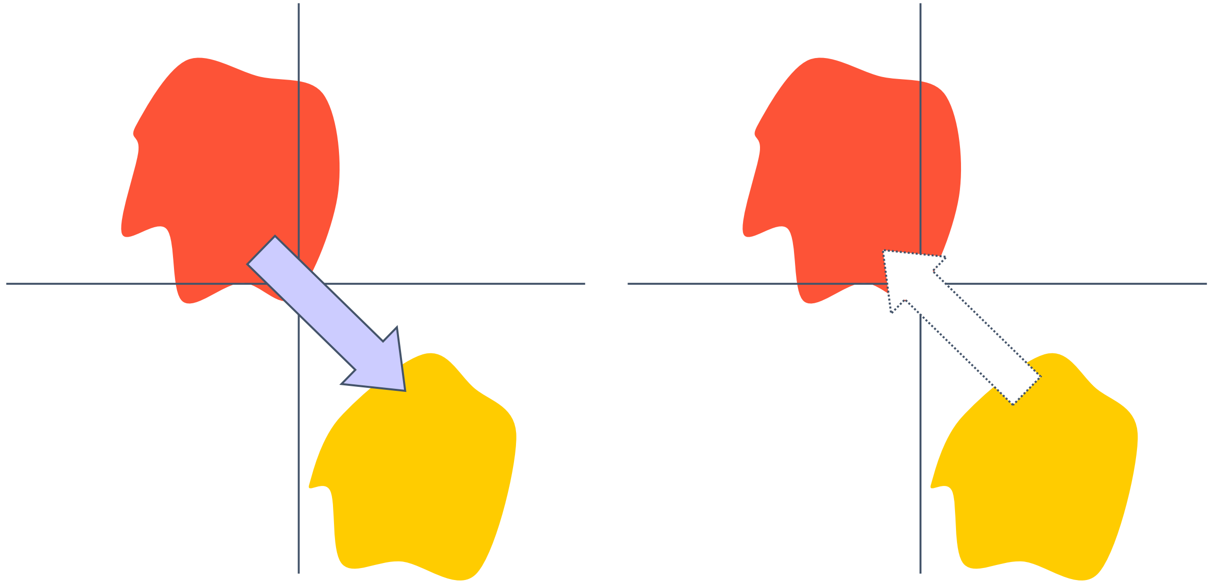
The entire speech signal is thus converted into a sequence of vectors. <u>These are cepstral vectors</u>.
There are other ways of converting the speech signal into a sequence of vectors

Effect of Speaker Variations, Microphone Variations, Noise etc.



- Noise, channel and speaker variations change the *distribution* of cepstral values

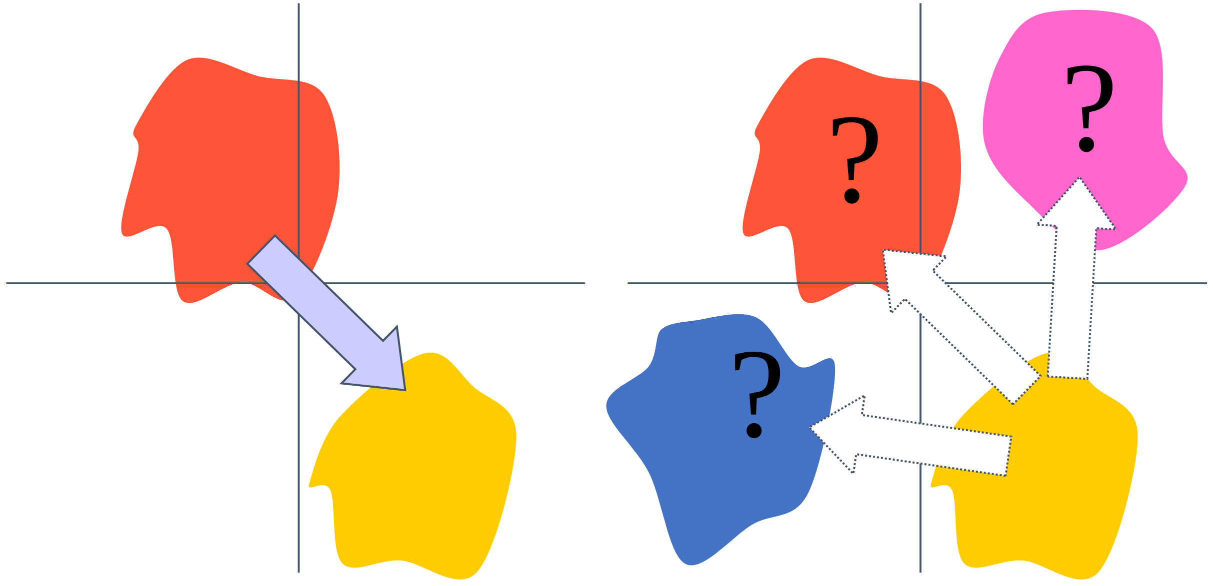
Ideal Correction for Variations



- Noise, channel and speaker variations change the *distribution* of cepstral values

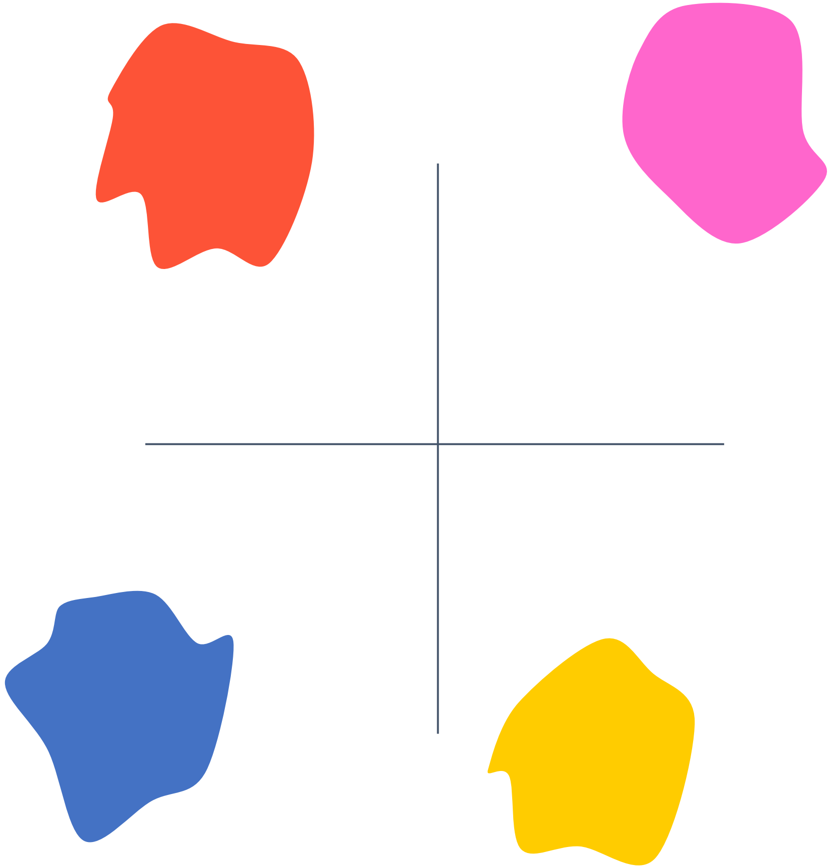
- To compensate for these, we would like to undo these changes to the distribution

Effect of Noise Etc.



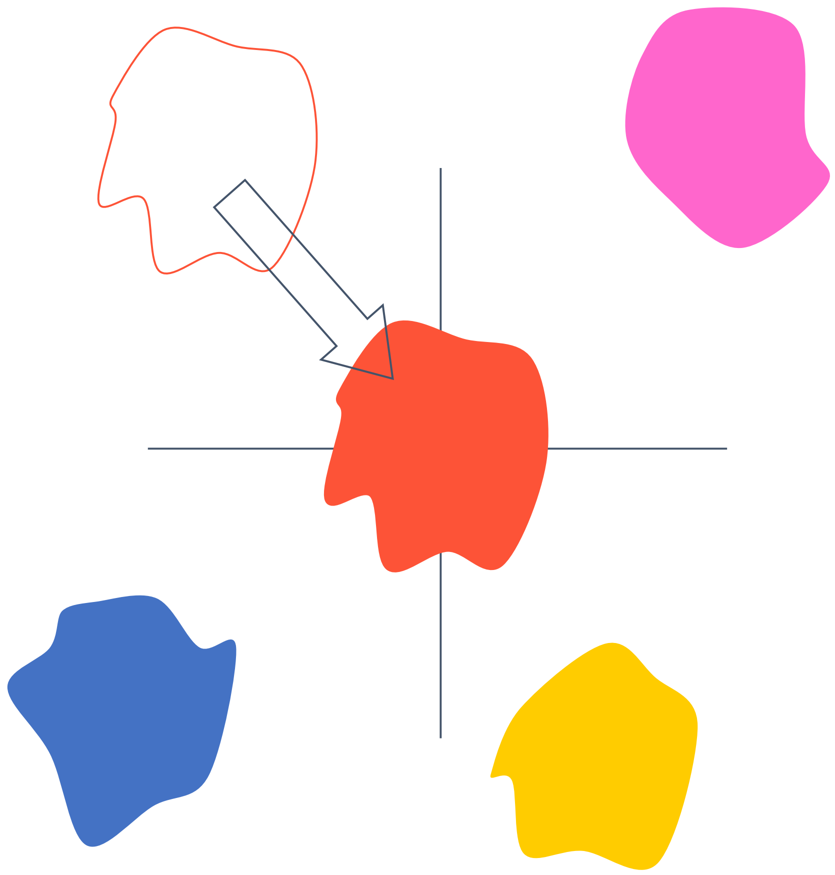
- Noise, channel and speaker variations change the *distribution* of cepstral values
- To compensate for these, we would like to undo these changes to the distribution
- Unfortunately, the precise position of the distributions of the "good" speech is hard to know

Solution: Move all distributions to a "standard" location



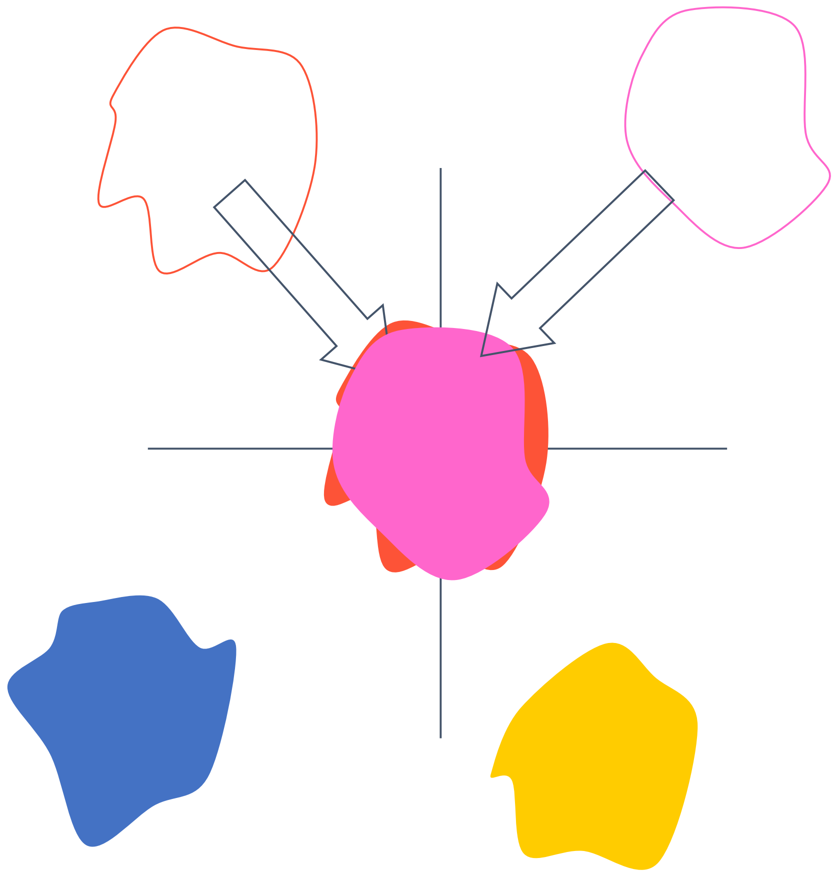
- "Move" all utterances to have a mean of 0
- This ensures that all the data is centered at 0
  - Thereby eliminating *some* of the mismatch

Solution: Move all distributions to a "standard" location



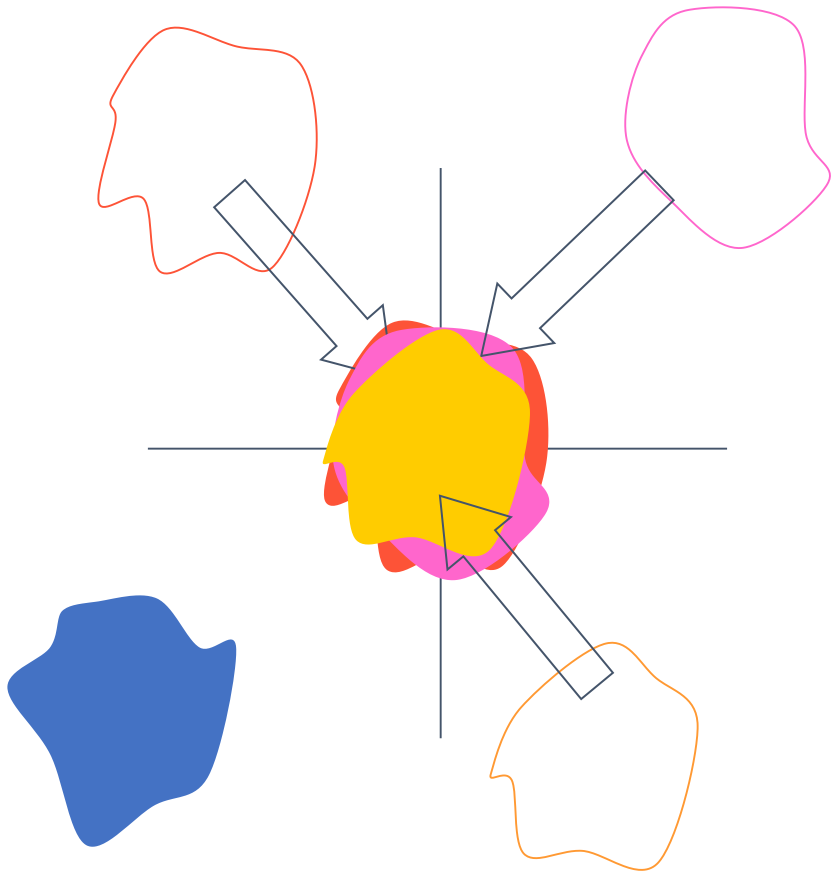
- "Move" all utterances to have a mean of 0
- This ensures that all the data is centered at 0
    - Thereby eliminating *some* of the mismatch

Solution: Move all distributions to a "standard" location



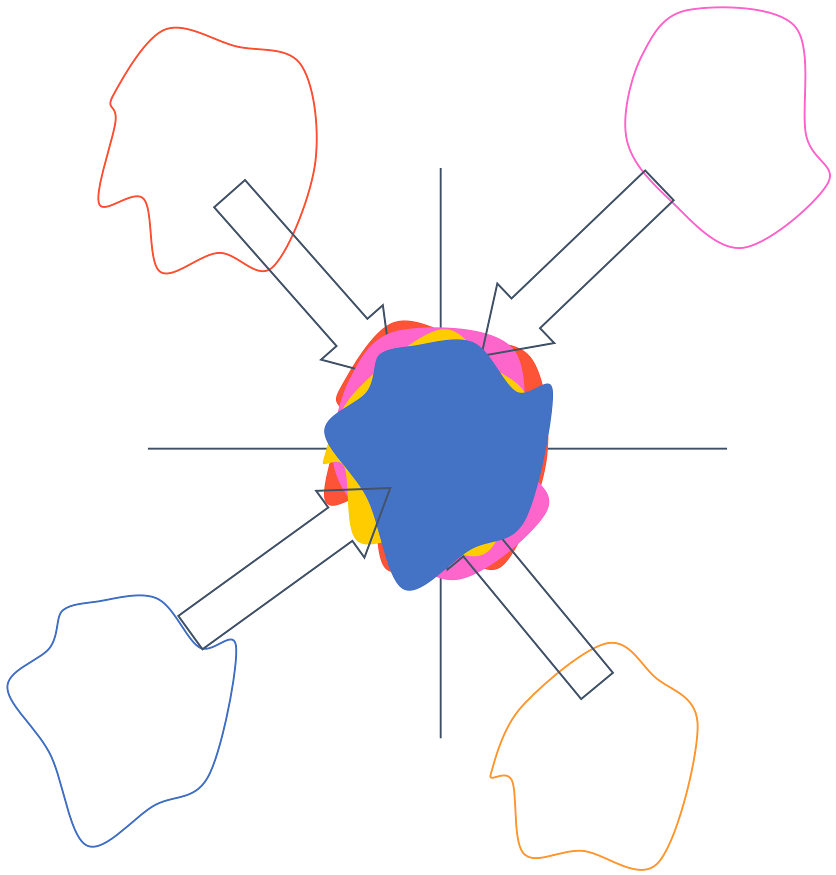
- "Move" all utterances to have a mean of 0
- This ensures that all the data is centered at 0
  - Thereby eliminating *some* of the mismatch

Solution: Move all distributions to a "standard" location



- "Move" all utterances to have a mean of 0
- This ensures that all the data is centered at 0
  - Thereby eliminating *some* of the mismatch

Solution: Move all distributions to a "standard" location



- "Move" all utterances to have a mean of 0
- This ensures that all the data is centered at 0
    - Thereby eliminating *some* of the mismatch

Cepstra Mean Normalization

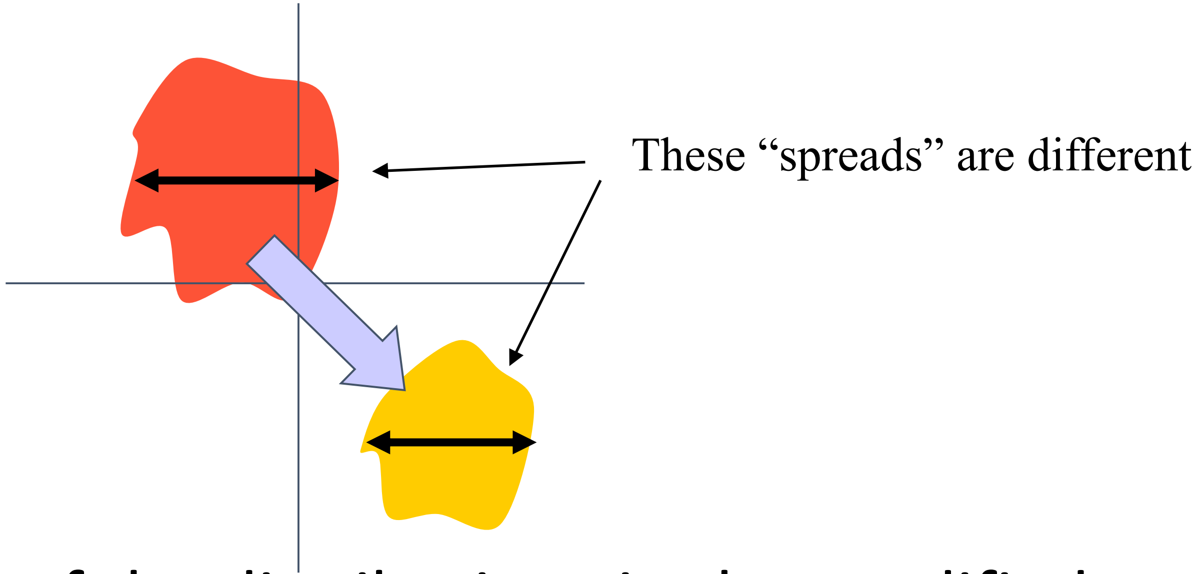
- For each utterance encountered (both in “training” and in “testing”)
- Compute the mean of all cepstral vectors

$$M_{recording} = \frac{1}{Nframes} \sum_t c_{recording}(t)$$

- Subtract the mean out of all cepstral vectors

$$c_{normalized}(t) = c_{recording}(t) - M_{recording}$$

Variance



- The *variance* of the distributions is also modified by the corrupting factors

- This can also be accounted for by variance normalization

Variance Normalization

- Compute the standard deviation of the mean-normalized cepstra

$$sd_{recording} = \sqrt{\frac{1}{Nframes}\sum_t c_{normalized}(t)}$$

- Divide all mean-normalized cepstra by this standard deviation

$$c_{\mathrm{var}\,normalized}(t) = \frac{1}{sd_{recording}} c_{normalized}(t)$$

- The resultant cepstra for any recording have 0 mean and a variance of 1.

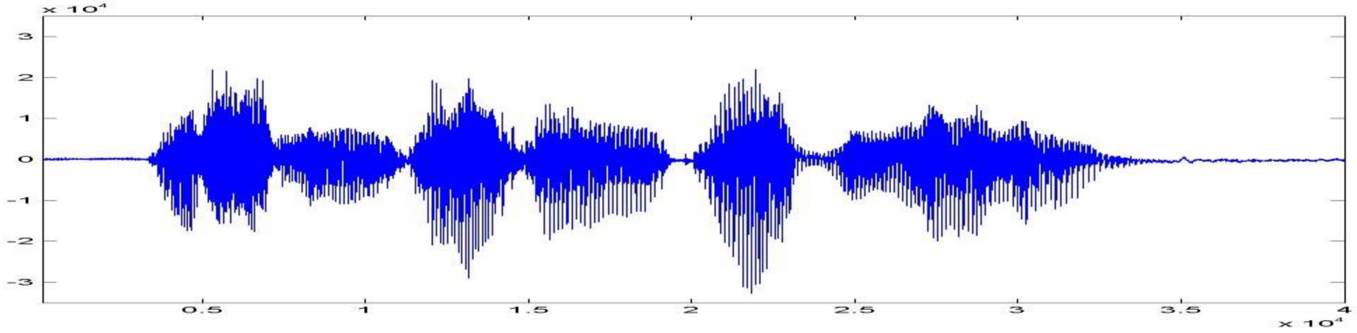
Temporal Variations

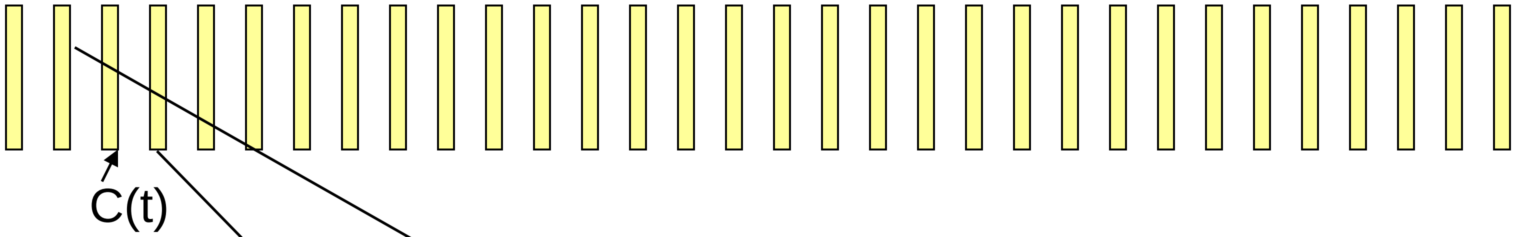
- The cepstral vectors capture instant information only
  - Or, more precisely, current spectral structure within the analysis window
- Phoneme identity resides not just in the snapshot information, but also in the temporal structure
  - Manner in which these values change with time
  - Most characteristic features
    - Velocity: rate of change of value with time
    - Acceleration: rate with which the velocity changes
- These must also be represented in the feature

Velocity Features

- For every component in the cepstrum for any frame
  - compute the difference between the corresponding feature value for the next frame and the value for the previous frame
  - For 13 cepstral values, we obtain 13 “delta” values
- The set of all delta values gives us a “delta feature”

The process of feature extraction





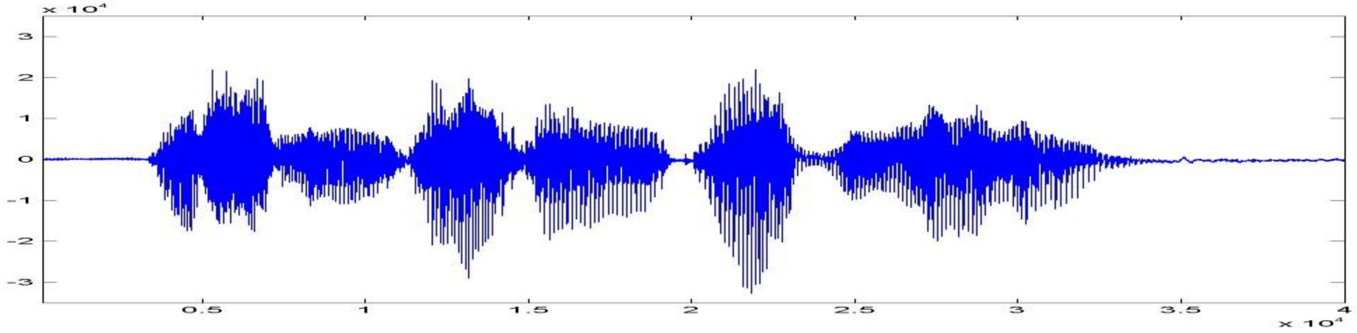
$$\Delta c(t)=c(t+\tau)-c(t-\tau)$$

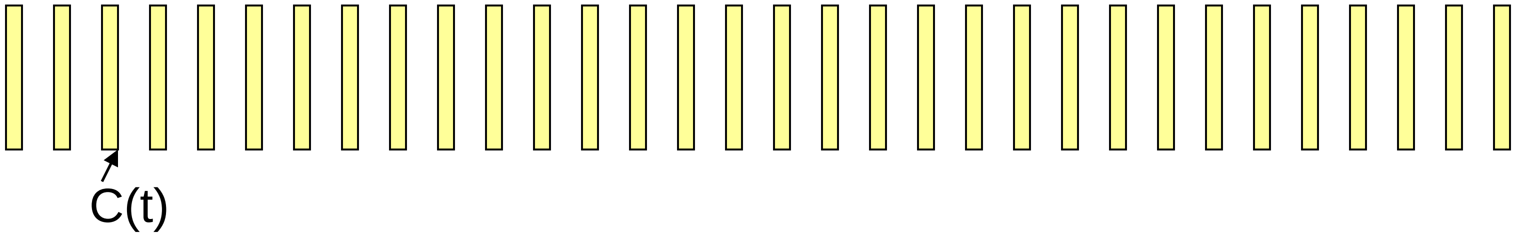
Representing Acceleration

- The *acceleration* represents the manner in which the velocity changes
- Represented as the derivative of velocity
- The DOUBLE-delta or Acceleration Feature captures this
- For every component in the cepstrum for any frame
  - compute the difference between the corresponding *delta* feature value for the next frame and the *delta* value for the previous frame
  - For 13 cepstral values, we obtain 13 “double-delta” values

- The set of all double-delta values gives us an “acceleration feature”

The process of feature extraction

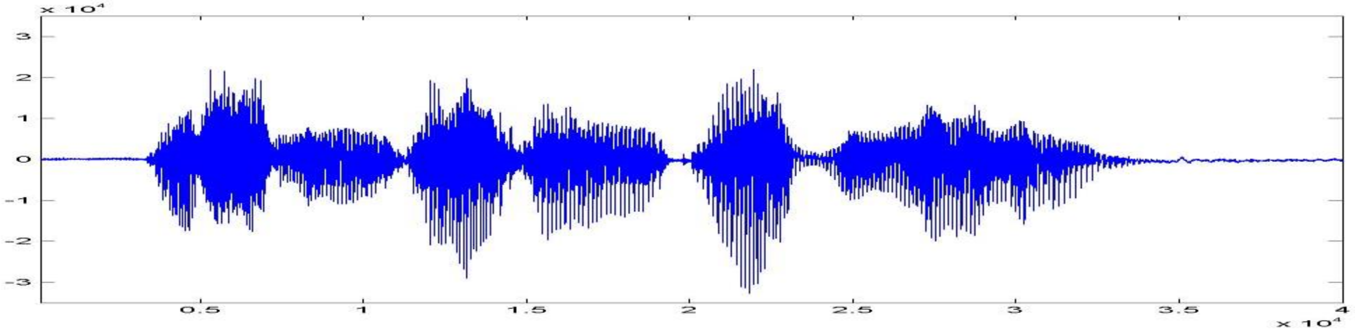


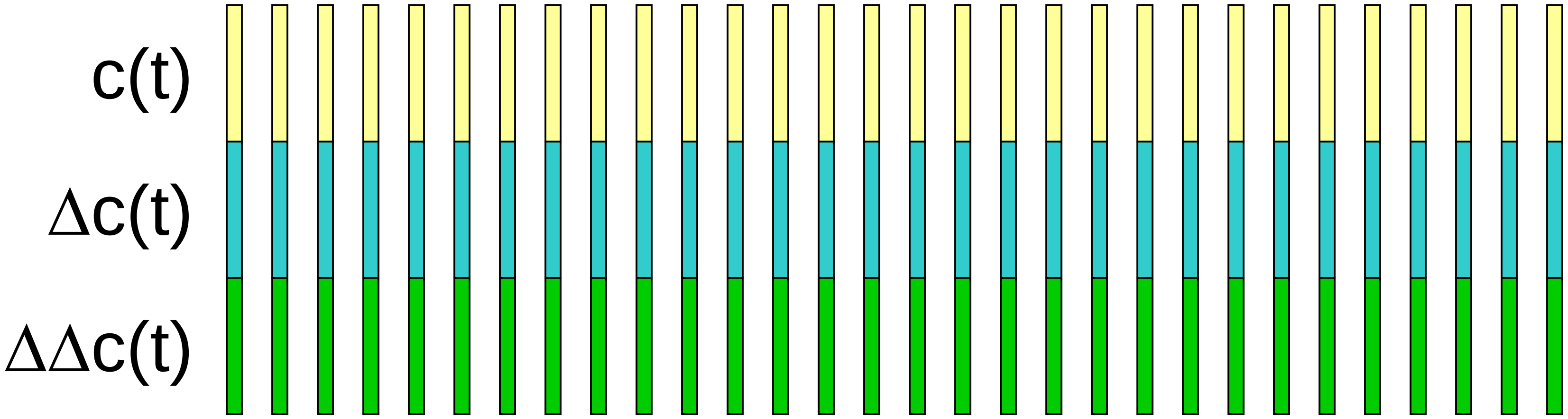


$$\Delta c(t)=c(t+\tau)-c(t-\tau)$$

$$\Delta\Delta c(t)=\Delta c(t+\tau)-\Delta c(t-\tau)$$

Feature extraction

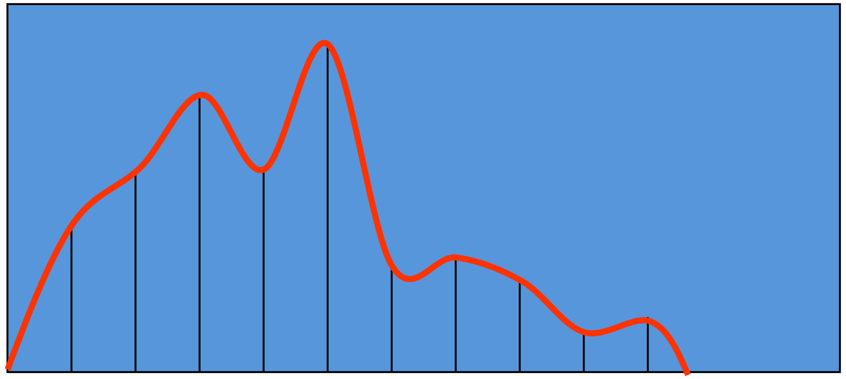




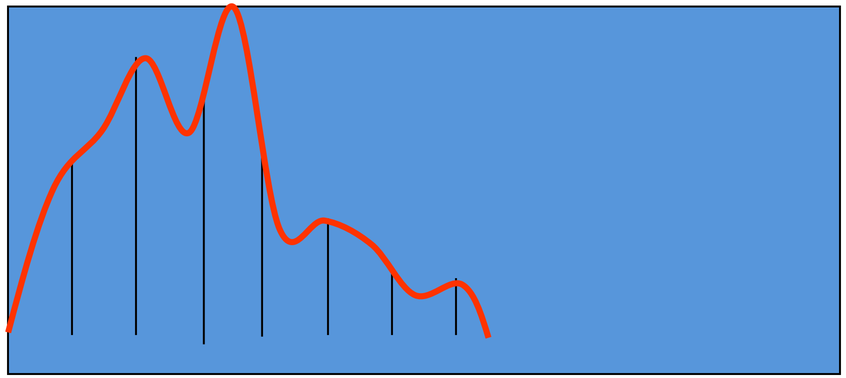
Normalization

- Vocal tracts of different people are different in length
- A longer vocal tract has lower resonant frequencies
- The overall spectral structure changes with the length of the vocal tract

Effect of vocal tract length



- A spectrum for a sound produced by a person with a short vocal tract length

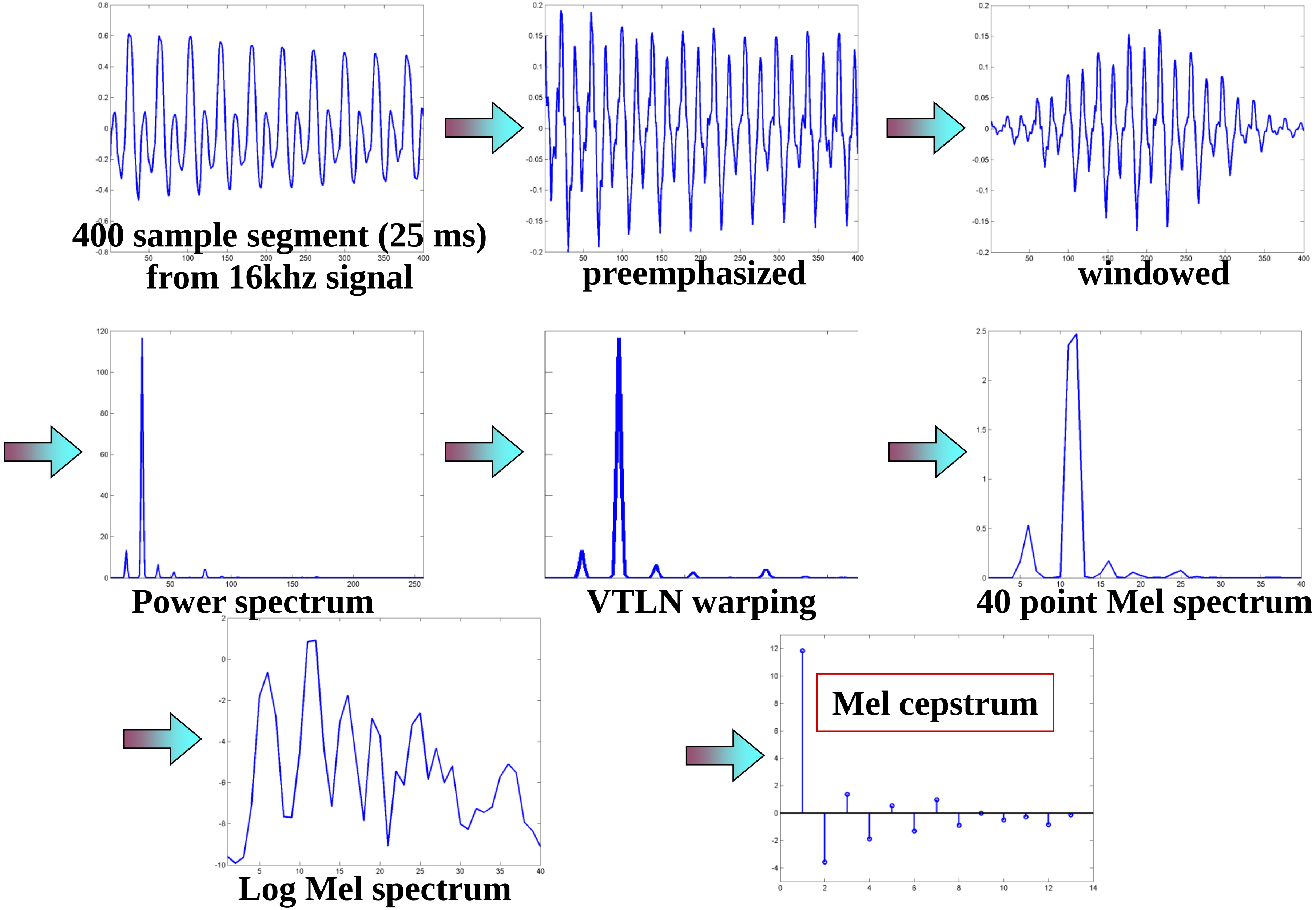


- The same sound produced by someone with a longer vocal tract

Accounting for Vocal Tract Length Variation

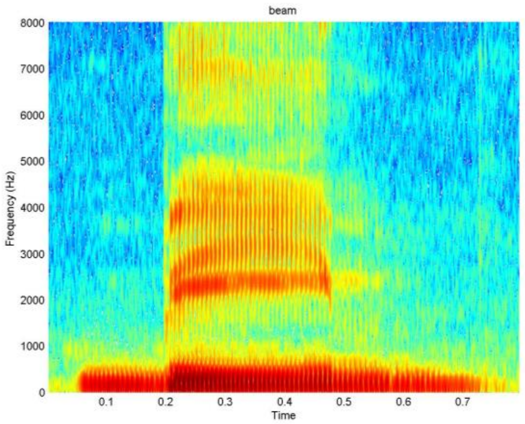
- Recognition performance can be improved if the variation in spectrum due to differences in vocal tract length are reduced
  - Reduces variance of each sound class
- Way to reduce spectral variation:
  - Linearly “warp” the spectrum of every speaker to a canonical speaker
  - The canonical speaker may be any speaker in the data
  - The canonical speaker may even be a “virtual” speaker

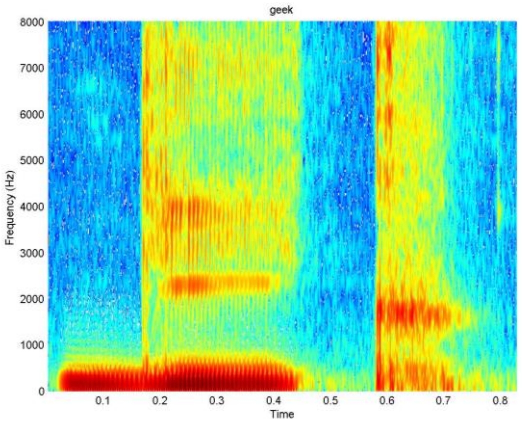
Frequency-warped Feature Comptuation



Spectral-Characteristic-based Estimation

- Formants are distinctive spectral characteristics
  - Trajectories of peaks in the envelope

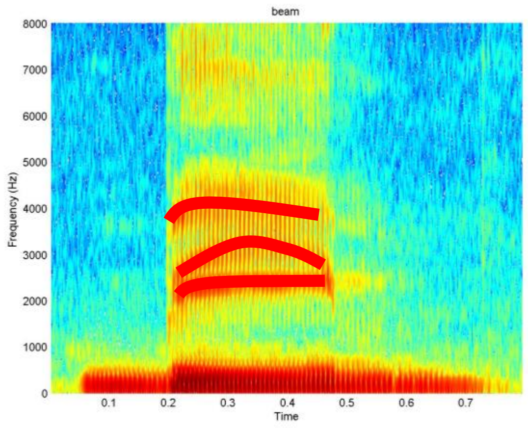


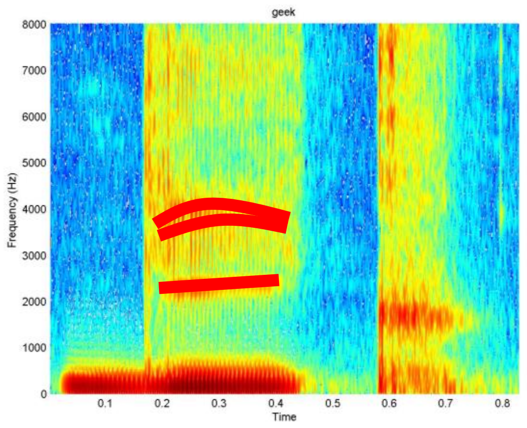


  - These trajectories are similar for different instances of the phoneme
  - But vary in a absolute frequency due to vocal tract length variations

Spectral-Characteristic-based Estimation

- Formants are distinctive spectral characteristics
  - Trajectories of peaks in the envelope





  - These trajectories are similar for different instances of the phoneme
  - But vary in a absolute frequency due to vocal tract length variations

Formants

- Formants are visually identifiable characteristics of speech spectra
- Formants typically identified as F1, F2, etc. for the first formant, second formant, etc.
  - F0 typically refers to the fundamental frequency – pitch
- The characteristics of phonemes are largely encoded in formant positions

Length Normalization

- To warp a speaker's frequency axis to the canonical speaker, it is sufficient to match formant frequencies for the two
  - i.e. warp the frequency so that F1(speaker) = F1(canonical), F2(speaker) = F2(canonical) etc. on average
- i.e. compute α such that αF1(speaker) = F1(canonical) (and so on) on average

Spectrum-based Vocal Tract Length Normalization

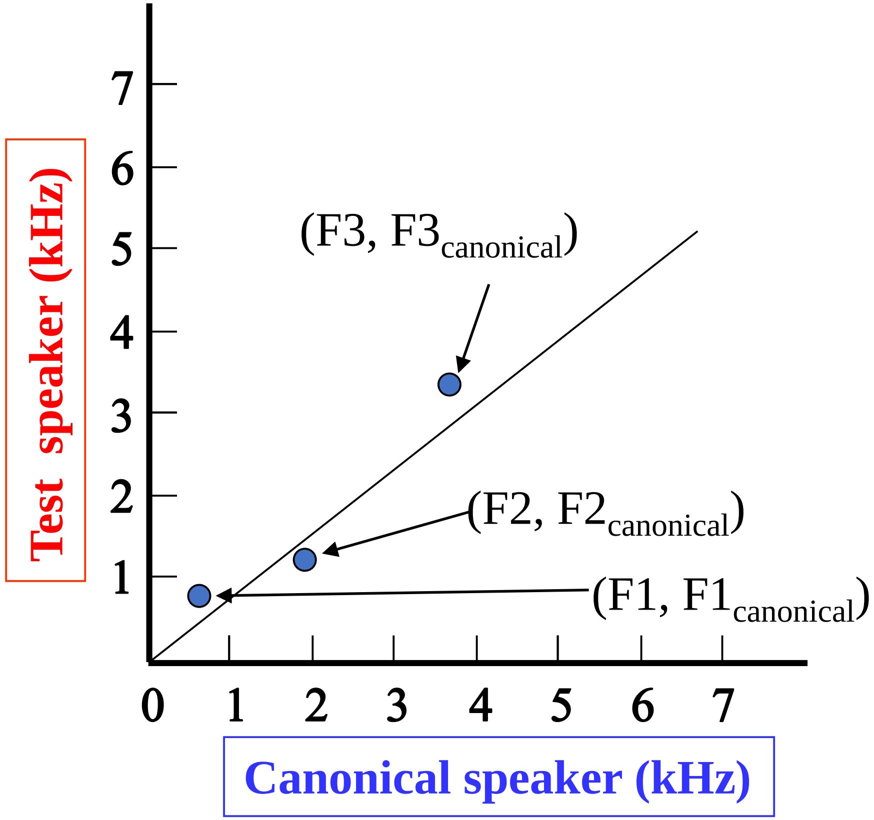
- Compute average F1, F2, F3 for the speaker’s speech
  - Run a formant tracker on the speech
    - Returns formants F1, F2, F3.. for each analysis frame
  - Average F1 values for all frames
    - Similarly compute average F2 and F3.
  - Three formants are sufficient

- Minimize the error:

$$(\alpha F1 - F1_{canonical})^2 + (\alpha F2 - F2_{canonical})^2 + (\alpha F3 - F3_{canonical})^2$$

  - The variables in the above equation are all average formant values
  - This computes a regression between the average formant values for the canonical speaker and those for the test speaker

Spectrum-Based Warping Function



- A is the slope of the regression between (F1, $F1_{canonical}$), (F2, $F2_{canonical}$) and (F3, $F3_{canonical}$)

But WHO is this canonical speaker?

- Simply an average speaker
  - Compute average F1 for all utterances of all speakers
  - Compute average F2 for all utterances of all speakers
  - Compute average F3 for all utterances of all speakers

Overall procedure

- Training:
  - Compute average formant values for all speakers
  - Compute speaker specific frequency warps for each speaker
  - Frequency warp all spectra for the speaker
- Testing:
  - Compute average formant values for the test utterance (or speaker)
  - Compute utterance (or speaker) specific frequency warps
  - Frequency warp all spectra prior to additional processing

Other Processing: Dealing with Noise

- The incoming speech signal is often corrupted by noise
- Noise may be reduced through spectral subtraction
- Theory:
  - Noise is uncorrelated to speech
  - The power spectrum of noise adds to that of speech, to result in the power spectrum of noisy speech
  - If the power spectrum of noise were known, it could simply be subtracted out from the power spectrum of noisy speech
    - To obtain clean speech