ROBUST SPEECH RECOGNITION

Richard Stern

Robust Speech Recognition Group
Carnegie Mellon University

Telephone: (412) 268-2535
Fax: (412) 268-3890
rms@cs.cmu.edu
http://www.cs.cmu.edu/~rms

Short Course at Universidad Carlos III
July 12-15, 2005

ICSLP 2006 in Pittsburgh

September 17-21, 2006
www.interspeech2006.org

Robust speech recognition

- **As speech recognition is transferred from the laboratory to the marketplace robust recognition is becoming increasingly important**
- **“Robustness” in 1985:**
  – Recognition in a quiet room using desktop microphones
- **Robustness in 2005:**
  – Recognition ….
    » over a cell phone
    » in a car
    » with the windows down
    » and the radio playing
    » at highway speeds

Some of the hardest problems in speech recognition

- Speech in high noise (Navy F-18 flight line)
- Speech in background music
- Speech in background speech
- Transient dropouts and noise
- Spontaneous speech
- Reverberated speech
- Vocoded speech

Outline of discussion

- Summary of the state-of-the-art in speech technology at Carnegie Mellon and elsewhere
- Review of fundamentals of speech recognition
- Introduction to robust speech recognition: classical techniques
- Robust speech recognition using missing-feature techniques
- Use of multiple microphones for improved recognition accuracy
- The future of robust recognition:
  – Signal processing based on human auditory perception
  – Computational auditory scene analysis

Outline of discussion

- Summary of the state-of-the-art in speech technology at Carnegie Mellon and elsewhere
- Review of fundamentals of speech recognition
- Introduction to robust speech recognition: classical techniques
- Robust speech recognition using missing-feature techniques
- Use of multiple microphones for improved recognition accuracy
- The future of robust recognition:
  - Signal processing based on human auditory perception
  - Computational auditory scene analysis

Introduction

- **Background:**
  - The technologies of speech recognition and text-to-speech synthesis have advanced rapidly over the last decade
  - Nevertheless, there are relatively few commercially-practical speech-based applications being sold today
- **Goals of this talk:**
  - To summarize the present state of the art and future directions in speech technology
  - To discuss key unsolved problems in transitioning laboratory technology to practical systems
  - To describe and discuss several speech-based applications now under development at CMU and elsewhere

Speech and language research at Carnegie Mellon

- **Some facets of CMU's ongoing core research:**
  - Large-vocabulary speech recognition
  - Text-to-speech synthesis
  - Spoken language understanding
  - Conversational systems
  - Machine translation
  - Multi-modal integration

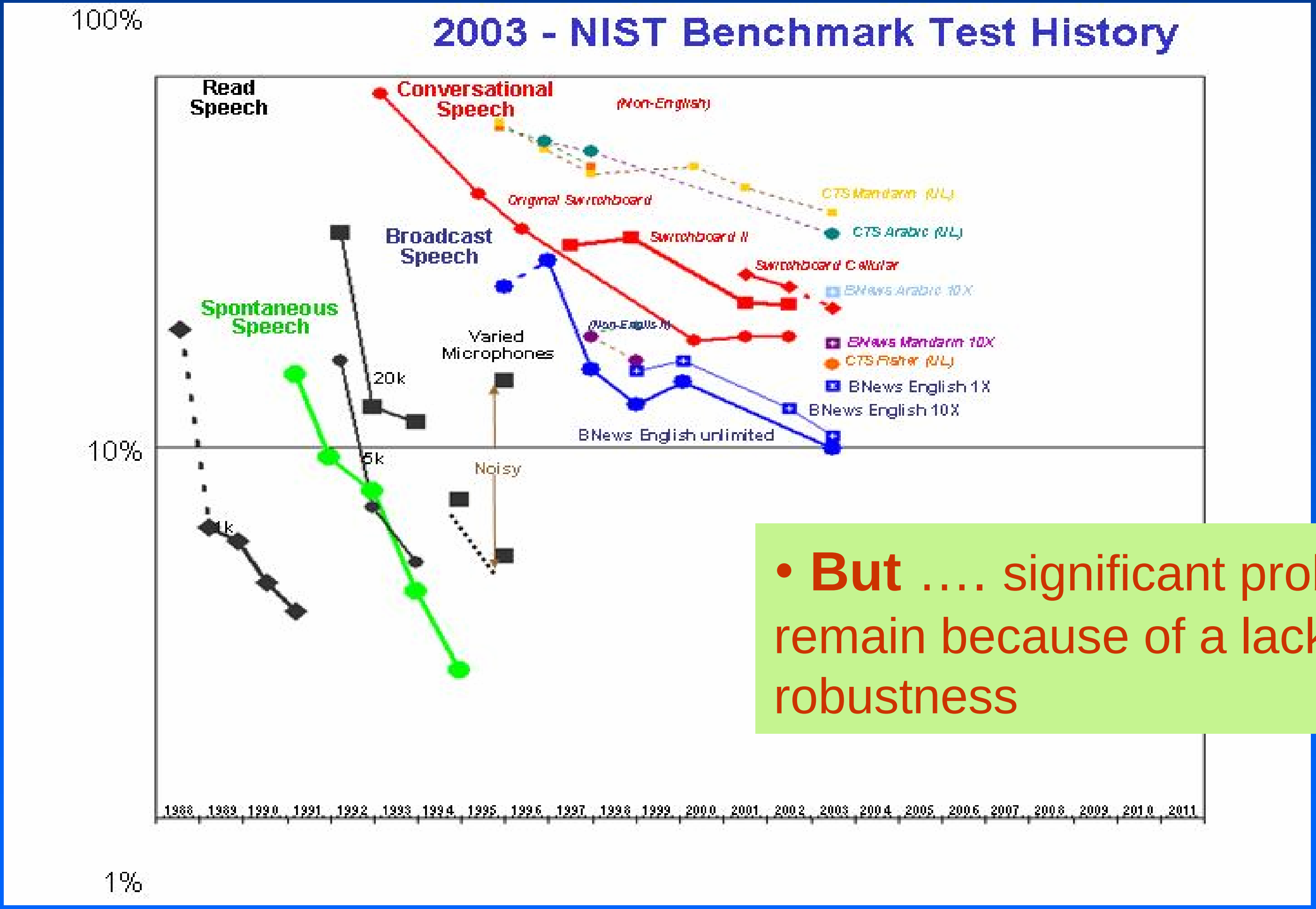
Speech and language research at Carnegie Mellon

- **Some application-focused efforts:**
  - The Communicator system (Alex Rudnicky)
  - Informedia group (Howard Wactlar)
    - Video on demand
  - LISTEN group (Jack Mostow):
    - Literacy training using speech input
  - CALL group (Maxine Eskenazi):
    - Foreign language training using speech input
  - Wearable computer group (Dan Sieworiek/Alex Rudnicky)

What we will discuss …

- **Core technology**
  - Automatic speech recognition
  - Text-to-speech synthesis
- **Introductory comments on commercial applications**
- **Information access through conversational systems**
  - CMU communicator and commercial information-access apps
- **Multi-media applications**
  - Informedia and LISTEN
- **User interface issues**
  - The Universal Speech Interface
- **Concluding remarks**

Speech recognition technology: accuracy is improving!



• **But** …. significant problems remain because of a lack of robustness

Carnegie Mellon

Speech recognition at CMU

- **The SPHINX-III system (1996-present):**
  - “Unlimited” vocabulary in English and Spanish; smaller versions in Serbo-Croatian, French, Korean, and Haitian Creole
  - ~60,000 words in unlimited-vocabulary language model
  - Continuous or semi-continuous hidden Markov models
  - Runs on Windows and Unix/Linux platforms
- **Sphinx-IV decoder in Java**
  - Funded by Sun, collaboration of CMU, Sun, MERL, HP, MIT
- **Code for both systems available in Open Source form**

Text-to-speech synthesis at Carnegie Mellon

- **Current TTS technology at CMU (and also AT&T, ATR, Microsoft, and elsewhere): synthesis based on concatention of selected recorded speech units**

- **Major research issues and problems:**
  – Recording natural domain-appropriate databases with good phonetic coverage
  – Joining units smoothly (currently units are selected based on F0, power, delta cepstra, with penalties for duration mismatch)
  – Prosody and naturalness

Personalized synthetic voices

- **Commercial voices from the Cepstral Corporation:**
  - David
  - Linda
  - Miguel
  - Marta
- **Cepstral voices are also presently available in Canadian French, British English, and German, with other languages to follow**
- **Sample voices developed at CMU:**
  - Rich

Open source code for ASR and TTS available from Carnegie Mellon

- **http://www.speech.cs.cmu.edu/hephaestus.html**
  - **ASR**: Sphinx and SphinxTrain
  - **TTS**: Festival, Festvox, FLITE
  - **Language factory**: QuickLM, Pronounce, Condition
  - **Spoken language**: CMU Communicator, SpeechLink, openvxi
- **http://mi.eng.cam.ac.uk/~prc14/toolkit.html**
  - **Language modeling**: CMU-Cambridge toolkit
- **http://speech.mty.itesm.mx/~jnolazco/proyectos.htm**
  - **Sphinx-III in (American) Spanish**

CMU TTS resources available in Open Source

- **Festival**
  - General multilingual speech synthesis engine (from the University of Edinburgh)
- **Festvox**
  - Tools for creating synthetic voices
- **FLITE**
  - Fast synthesis for embedded engines

What kinds of speech applications are available now?

- **Dictation systems:**
  - Large vocabulary and speaker-adaptive, with adaptable vocabularies and grammars
- **Command-and-control systems:**
  - Voice control of operating system and applications
  - Part of infrastructure of Windows XP and Mac OSX
- **Information-access systems:**
  - Frequently conversational in nature
  - Frequently involve telephone access (including cell phones)
- **Data entry using handheld terminals and simple wearable systems**
- **Primitive translation systems**

Command and control of operating systems and applications

- **Some attributes of current systems:**
  - Voice commands can begin to replace the mouse and keyboard
  - Limited vocabulary based on which window is in focus or based on user state
  - Probably will ultimately be a complement rather than a replacement for the keyboard and mouse

An (old) example of command and control in a commercial product

Dragon systems demo (circa 1998):

QuickTime™ and a
Video decompressor
are needed to see this picture.

Information access through spoken language systems

- **What is a spoken language system?**
- **Some attributes:**
  – Voice input and output
  – Intelligent interaction with a database to solve real problems
- **Some domains that have been studied:**
  – Travel planning, orientation, navigation
  – General information retrieval
  – General provision of advice
- **Comments:**
  – A "marriage" of speech recognition and natural language processing
  – Major goal: to develop voice systems that users will prefer over keyboard-driven systems

Conversational systems: the CMU Communicator

- **Mixed-initiative interaction**
  - Both the user and computer can initiate action and clarification
- **User and task modeling**
  - User preferences and defaults
  - Understanding of the semantics of the underlying task
- **Dialog scripting**
  - Knowledge of user goals and subgoals
  - Dynamic modification of lexicon and grammar based on dialog context
  - Guidance of user through planning procedures
- **Task analysis and domain knowledge needed for successful system development**

The CMU Communicator

QuickTime™ and a
DV/DVCPRO - NTSC decompressor
are needed to see this picture.

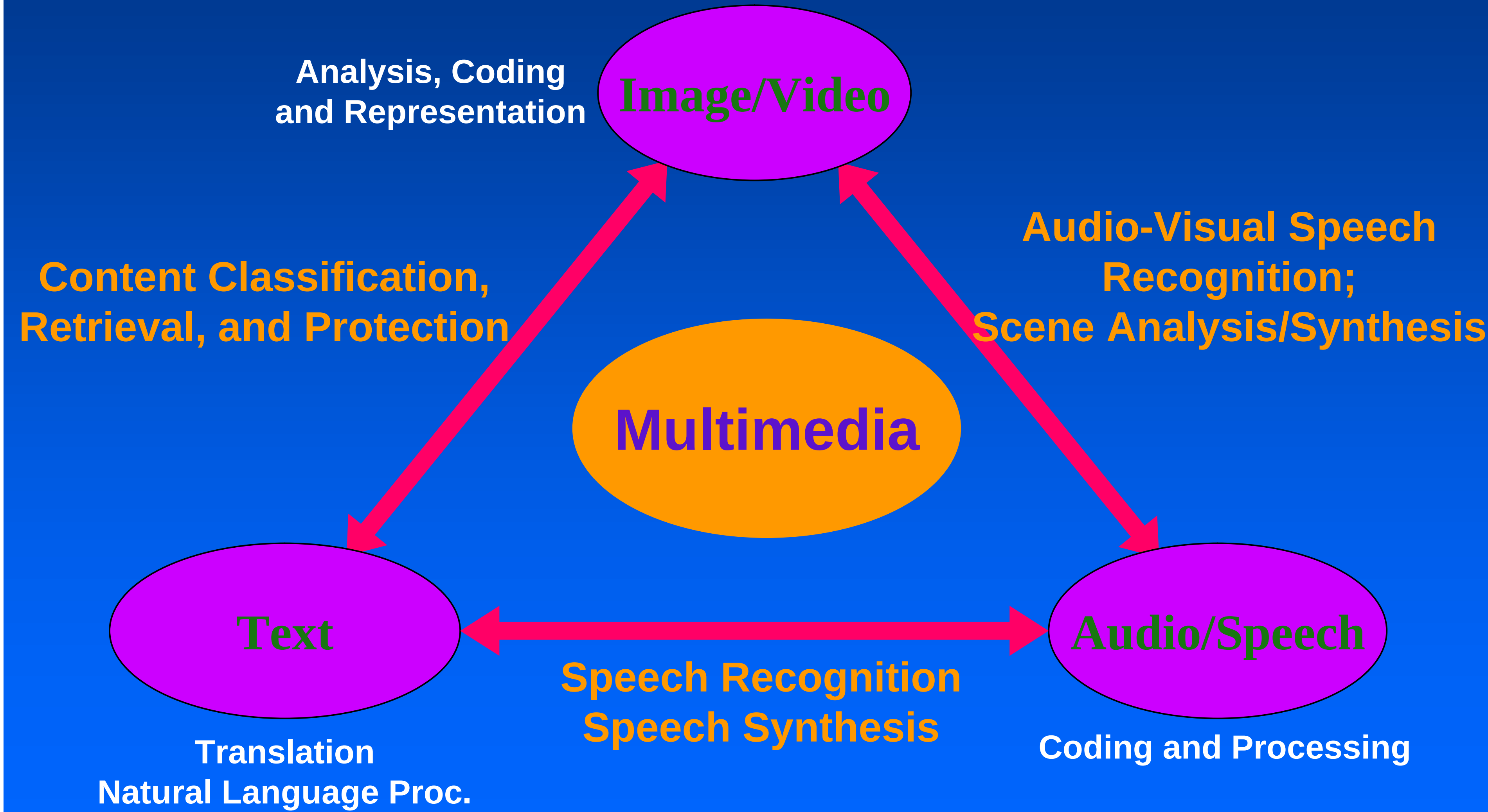
Examples of commercial spoken language systems

- Reservations on United Airlines (ScanSoft)
- Health care patient eligibility verification (Nuance)
- BeanTown Navigation on Nokia 3650 phone (ScanSoft)

CHALLENGES FOR CONVERSATIONAL SYSTEMS

- **Recognition of spontaneous speech**
- **Adaptation and learning at all levels**
  - Acoustic
  - Lexical
  - Semantic
  - Task domain
  - Environment
- **Domain awareness for both users and machines**
- **Training with very little data**
- **Establishing the right balance of initiative between user and system**
- **Development of toolkits for new applications**

THE CHALLENGE OF MULTIMEDIA



InformediaTM: News on Demand

Motivation:

- **Full-motion video is the most compelling presentation medium for display, training, and information access**
- **Video is the most difficult medium for browsing and searching**
- **Spoken language interface enables anyone to**
  - retrieve desired information ...
  - using natural fluent speech ...
  - with no special training

InformediaTM: News on Demand

The original Informedia system included:

- **Unlimited-vocabulary spoken language interface**
- **Real-time MPEG video playback**
- **Totally automatic indexing ...**
  - based on text captioning for television news
  - based on speech recognition for public radio broadcasts
- **Browsing capability**

Automatic indexing based on speech recognition ultimately could be extended to all digital video libraries.

The original Informedia system (~1997)

QuickTime™ and a
DV/DVCPRO - NTSC decompressor
are needed to see this picture.

- For more information: www.informedia.cs.cmu.edu

Informedia today

QuickTime™ and a
DV/DVCPRO - NTSC decompressor
are needed to see this picture.

- Speech is used to create transcripts and to align video to transcripts for indexing

ASR accuracy depends on speaking style and the environment

- CMU recognition error rates in transcription of Broadcast News TV and radio news broadcasts (1997 DARPA evaluations)

Prepared studio speech	15.5%
Spontaneous studio speech	22.8%
Telephone and similar channels	32.2%
Background music	33.4%
Background noise	30.8%
Non-native speakers	33.0%
OVERALL AVERAGE	24.0%

Another multimedia application: the LISTEN Reading Tutor

Using speech to help children and adults learn to read:

- Students read from prepared texts
- Computer listens, detects mistakes, and applies “helpful” feedback
- Many interesting issue in both speech recognition and application design

The CMU LISTEN Project

QuickTime™ and a
YUV420 codec decompressor
are needed to see this picture.

- For more information: http://www.cs.cmu.edu/~listen

Speech recognition on handheld teminals

- **Some characteristics:**
  - Noisy environment
  - Limited computation and memory
  - Terminals generally operated by single user

- **Some additional attributes of mobile phones:**
  - Power available only for limited periods of time
  - High cost sensitivity
  - Operate in multi-lingual environment and under coding

One approach to application design: The Universal Speech Interface

- **Goals of the Universal Speech Interface:**

- **Do for speech what Graffiti™ has done for mobile text entry**
  – semi-natural language: man, machine meet halfway
  – 5 minute training, via interactive tutorial
- **Do for speech what the Macintosh look-and-feel has done for GUIs**
  – a universal look-and-feel (rather, “sound-and-feel”) across all applications

Carnegie Mellon

The CMU
Universal Speech Interface

QuickTime™ and a
DV/DVCPRO - NTSC decompressor
are needed to see this picture.

- **For more info: http://www.cs.cmu.edu/~usi**

So why hasn't speech technology developed faster?

- **(Or why haven't we yet developed the "killer app" for speech input and output?)**
- **Even though core recognition has improved, we still need...**
- **Greater robustness ....**
  - To speakers and dialects
  - To the effects of unknown noise and filtering
  - To vocoded speech and telephone channels
- **Automatic adaptation to out-of-domain input:**
  - New words, syntax, and semantic concepts
- **Improved human-computer interfaces**
- **Lower cost?**

Summary: what's going on now?

- **Core speech recognition technology has improved greatly over the last decade and is now usable if deployed with care, but ……**
- **Current speech systems remain fragile to**
  - environmental degradation (including interfering sources, filtering and nonlinear distoration)
  - spontaneous and disfluent speech
  - out-of-vocabulary utterances, unusual syntax, and other unexpected types of input
- **Spoken language systems for information access has taken hold, but conversational systems are limited by recognition accuracy and application design**
- **Automatic detection and assimilation of new words and concepts remains extremely difficult**

Summary: what are some interesting trends to watch for?

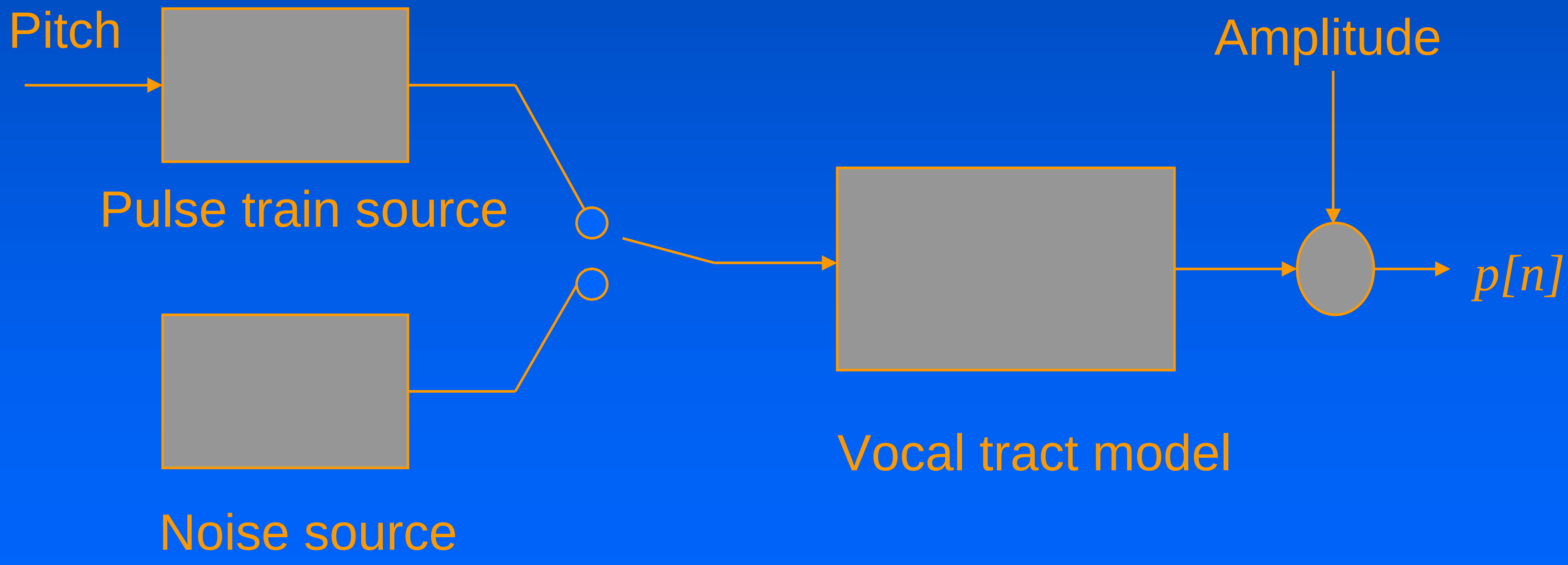
- **Greater commercial success as we conquer the major problems of**
  - robustness and adaptation for ASR
  - effective portable application design
- **Greater emphasis on multi-media applications in which speech is one of several input/output modalities**
- **Greater diffusion of speech-baased education and training applications**
- **Continued search for the right way to integrate speech, keyboard, and mouse in the OS**
- **An “interesting” period in which central servers and handsets both compete with board-level products as the site for recognition and related processing**

Outline of discussion

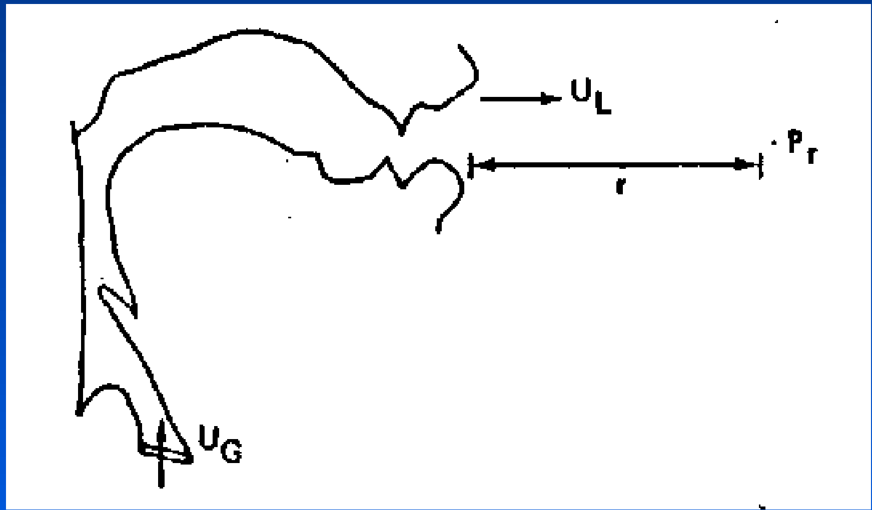
- Summary of the state-of-the-art in speech technology at Carnegie Mellon and elsewhere
- Review of fundamentals of speech recognition
- Introduction to robust speech recognition: classical techniques
- Robust speech recognition using missing-feature techniques
- Use of multiple microphones for improved recognition accuracy
- The future of robust recognition:
  - Signal processing based on human auditory perception
  - Computational auditory scene analysis

The source-filter model of speech production

A useful model for representing the generation of speech sounds:



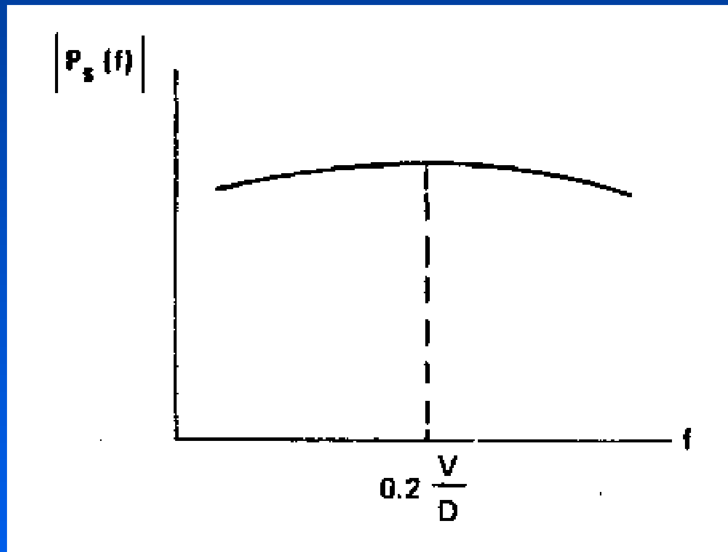
THE ACOUSTIC THEORY OF SPEECH PRODUCTION: MODELING THE VOCAL TRACT



- **The sound pressure at a distance *r* is determined by**
  – Spectrum of the excitation signal
  – Configuration of throat, jaw, tongue, lips, teeth, etc.
  – Loading effect of air

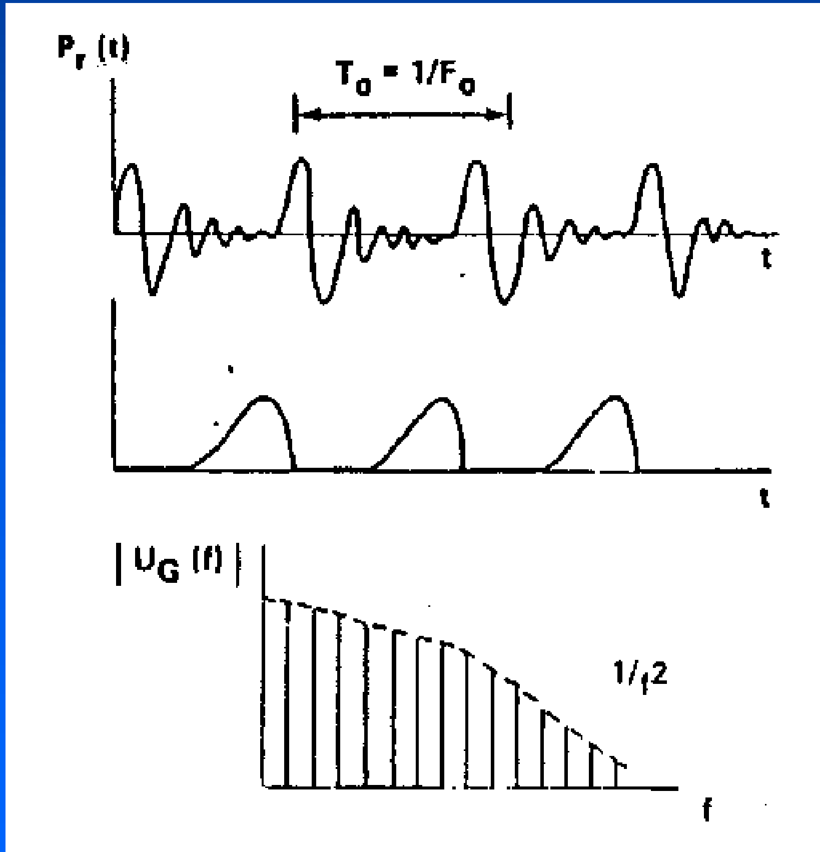
Unvoiced speech sources

- Turbulent voicing sources are approximately flat in frequency:

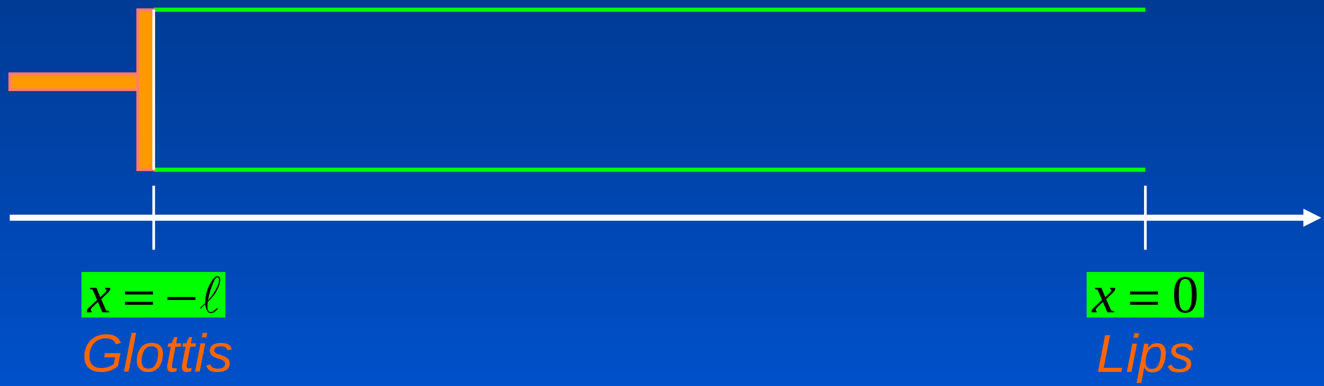


Voiced speech sources

- **Glottal pulses** have a spectrum that decreases with the square of frequency:

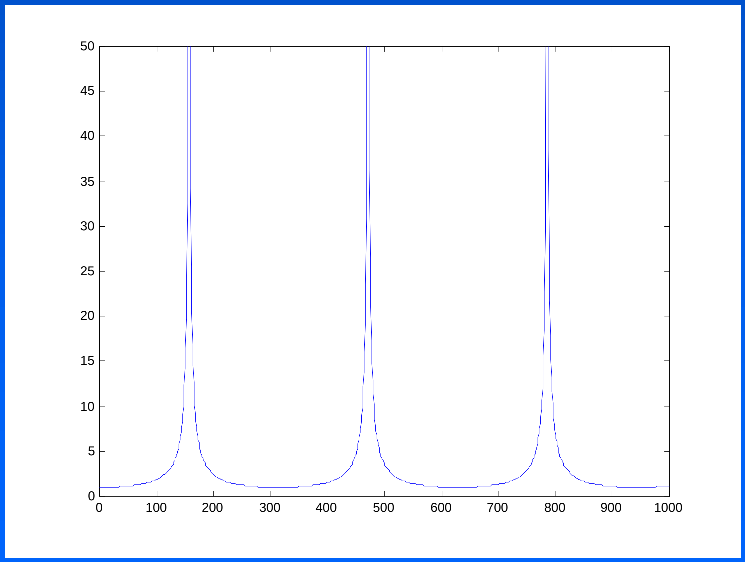


Sound propagation in a uniform tube

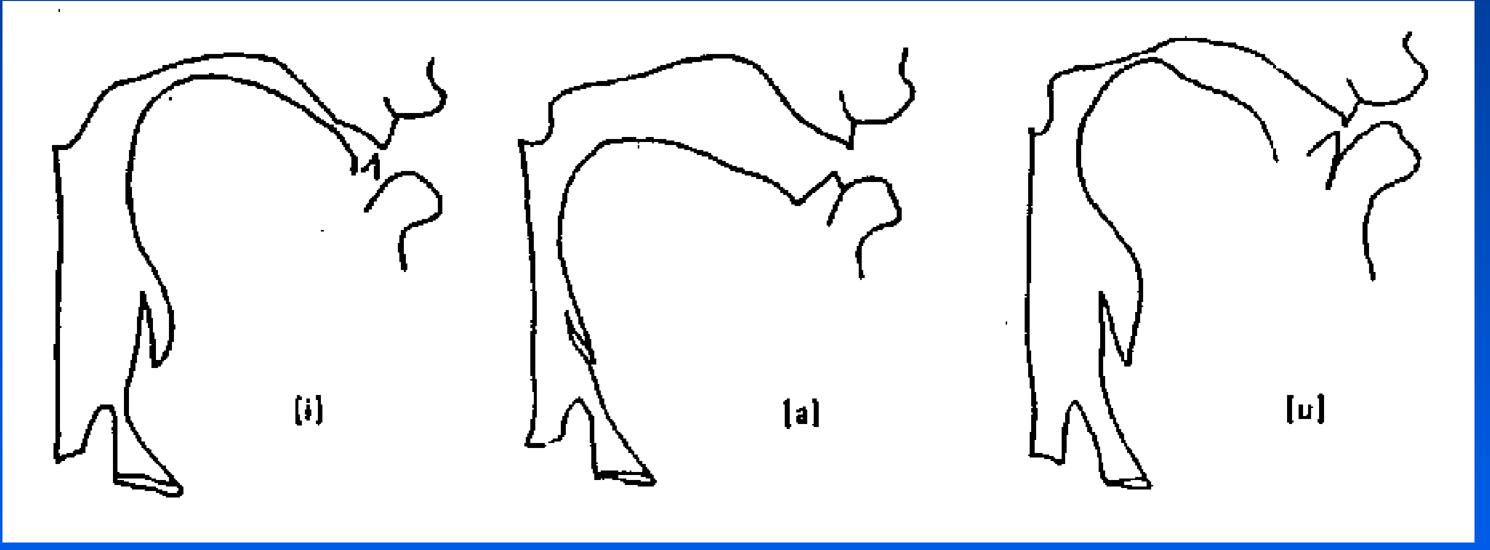


- **Frequency response:**

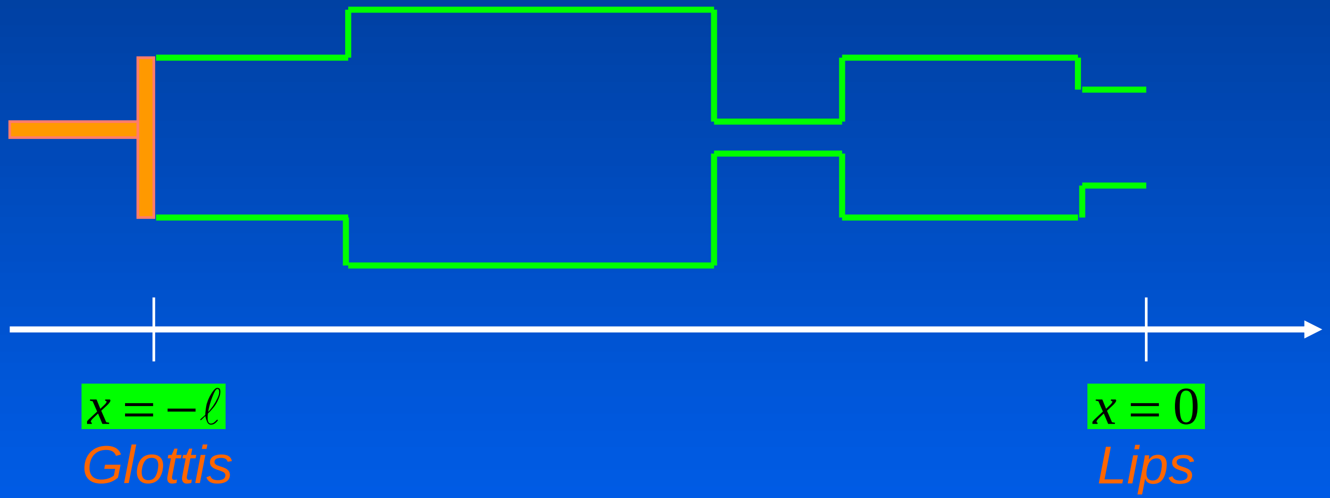
(assuming ideal perfectly reflective walls)



Vowel production in the vocal tract

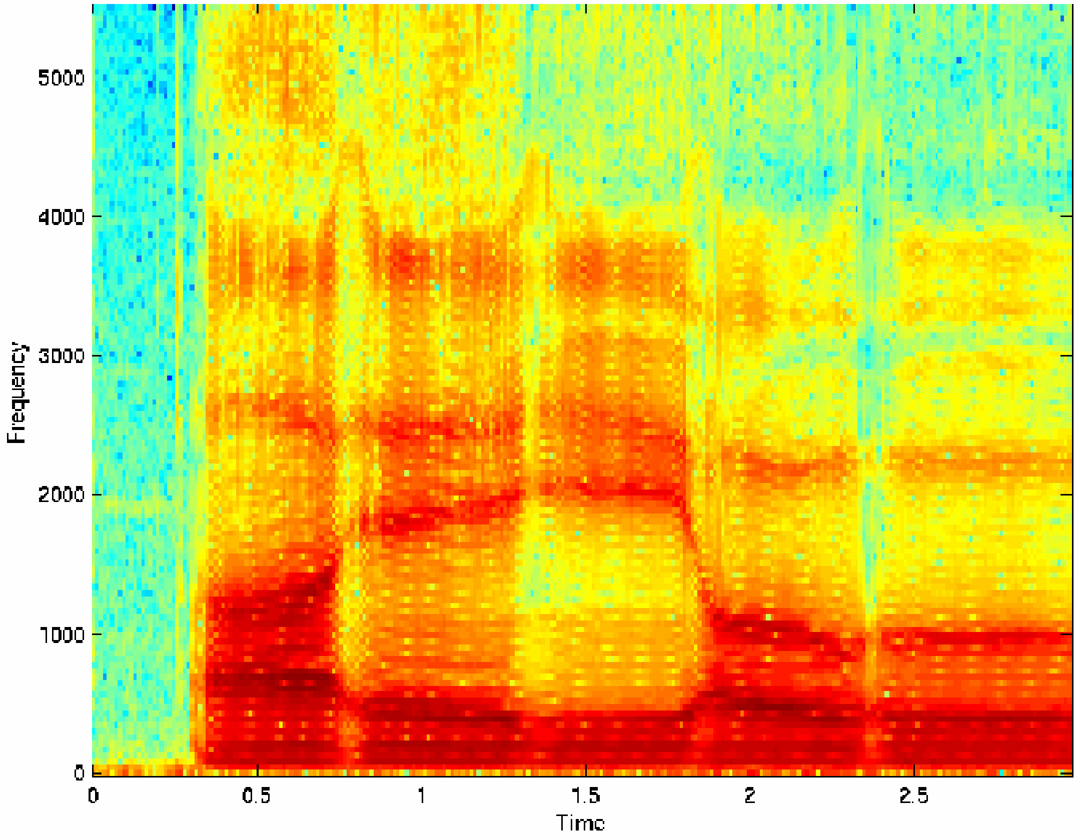


A more realistic model of sound production

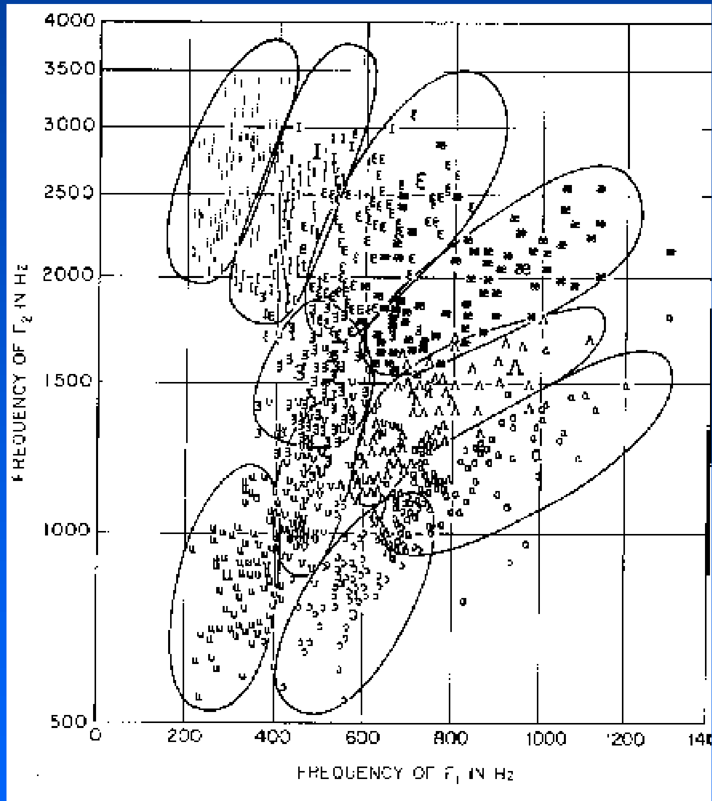


- **Comment:** Resonant frequencies now non-uniform

Some example vowels

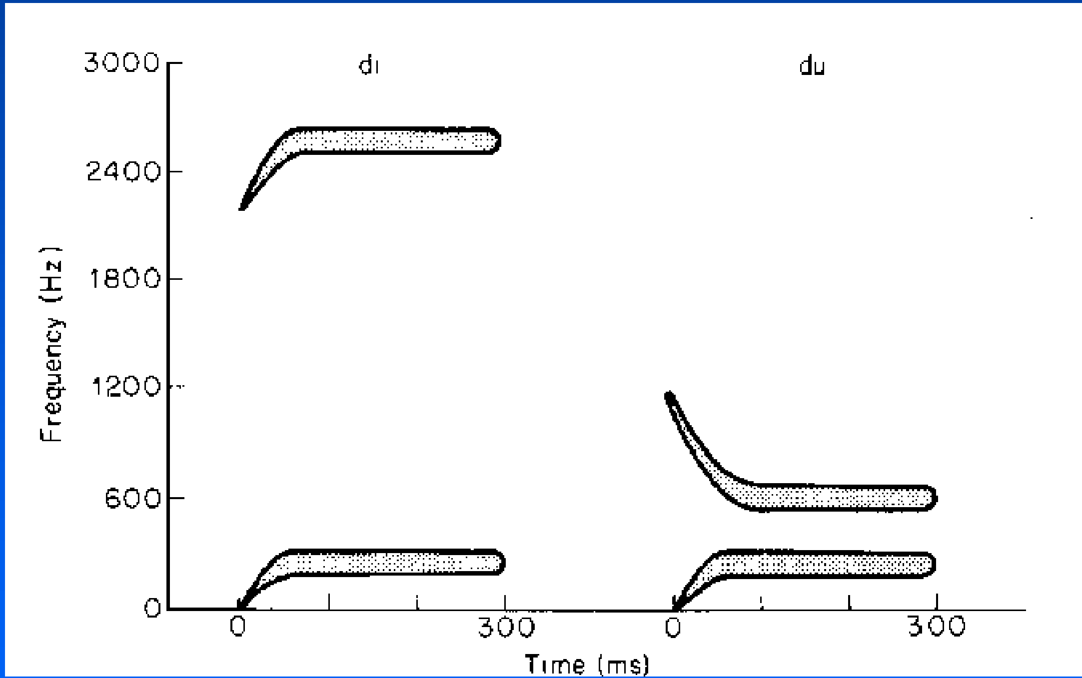


Vowel perception and formant frequencies



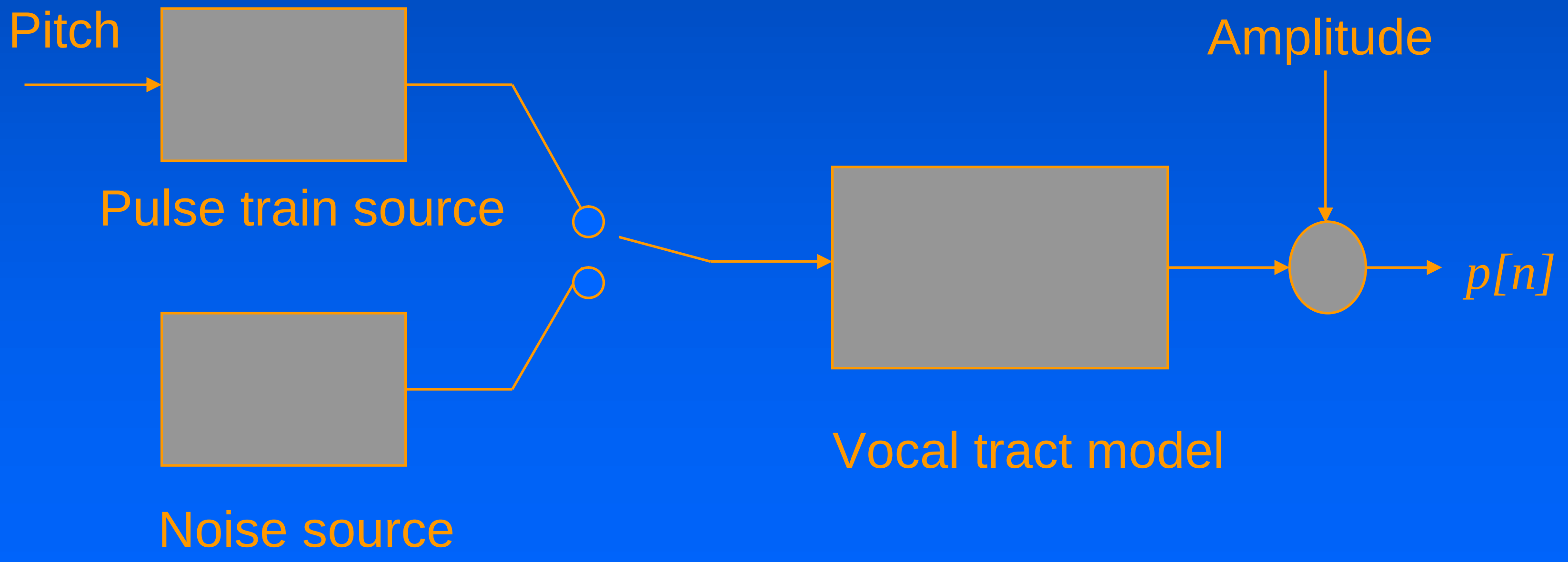
Context dependencies in speech production

- Spectral patterns that form /di/ and /du/:

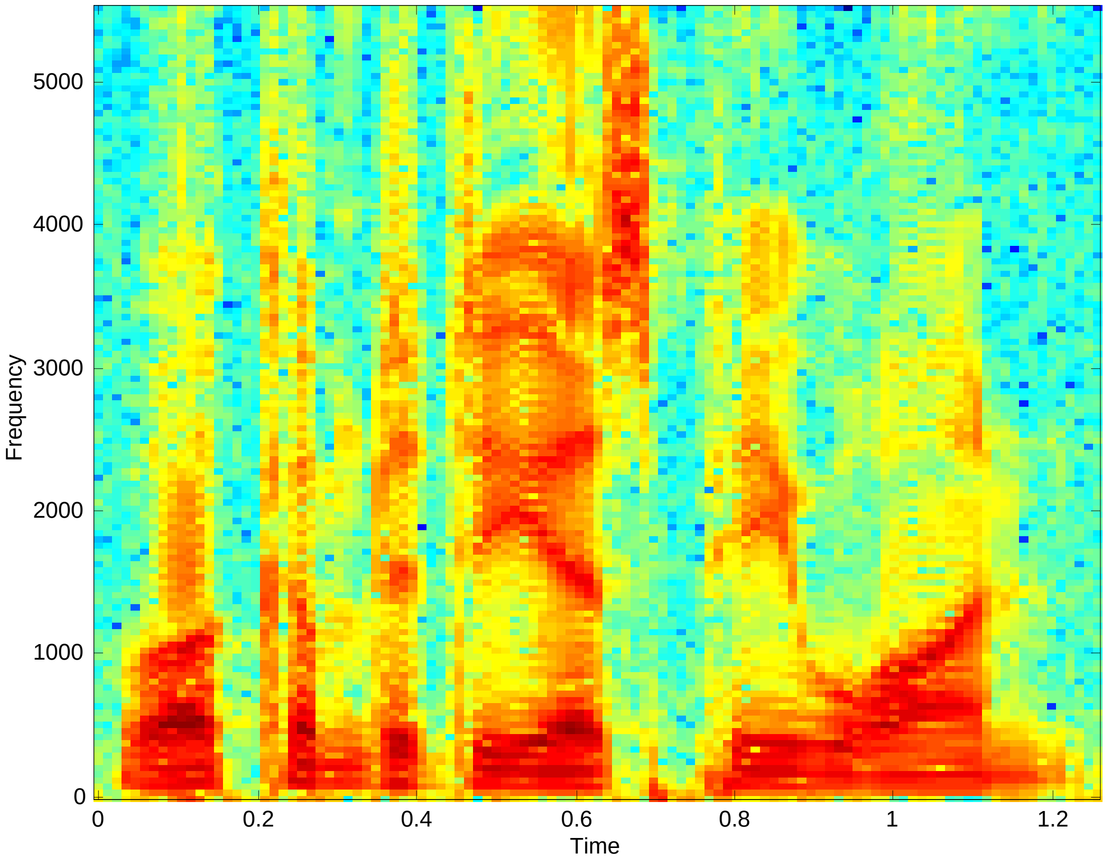


The source-filter model of speech production

A useful model **for representing the generation of speech sounds:**



The speech spectrogram



Separating the vocal-tract excitation from the filter

- Original speech:
- Speech with 75-Hz excitation:
- Speech with 150-Hz excitation:
- Speech with noise excitation:

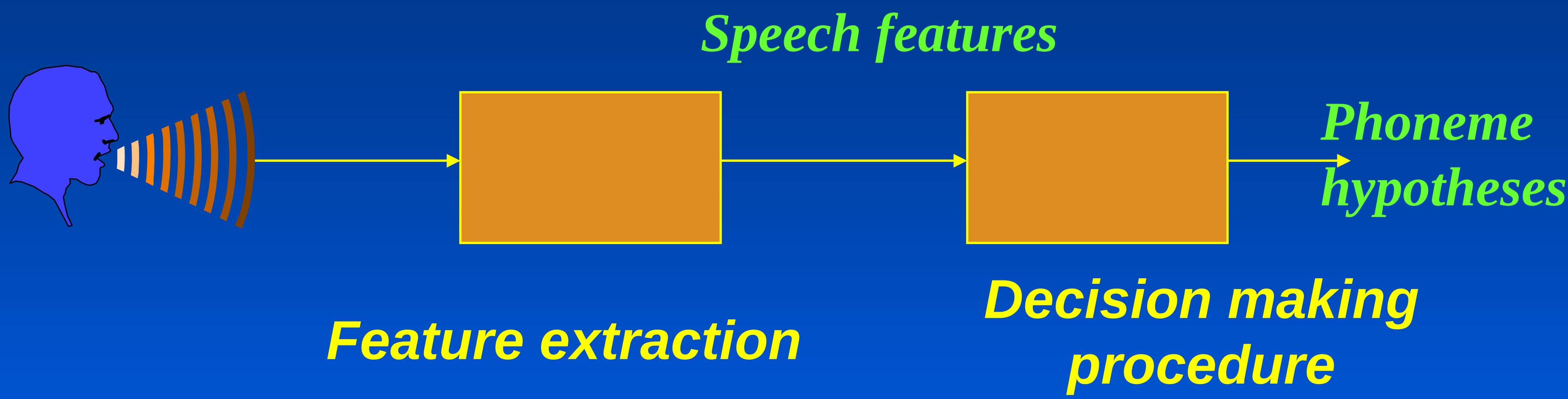
Summary: elements of speech production

- **We have discussed very superficially the production of speech sounds**
  - Source-filter model
  - Vocal tract transfer functions
  - Impact on perception
- **The source filter model is used**
  - As a way to model how we produce speech sounds
  - As a way to reduce the number of parameters needed to characterize speech sounds
  - As a way of extracting features that are used by speech recognition systems

Outline of discussion

- Basic mechanisms of speech production
- Basic mechanisms of auditory perception
- (Very!) basic review of automatic speech recognition
- Conventional signal processing for speech recognition
- Signal processing for improved speech recognition
- Signal processing for improved sound source separation

OVERVIEW OF SPEECH RECOGNITION



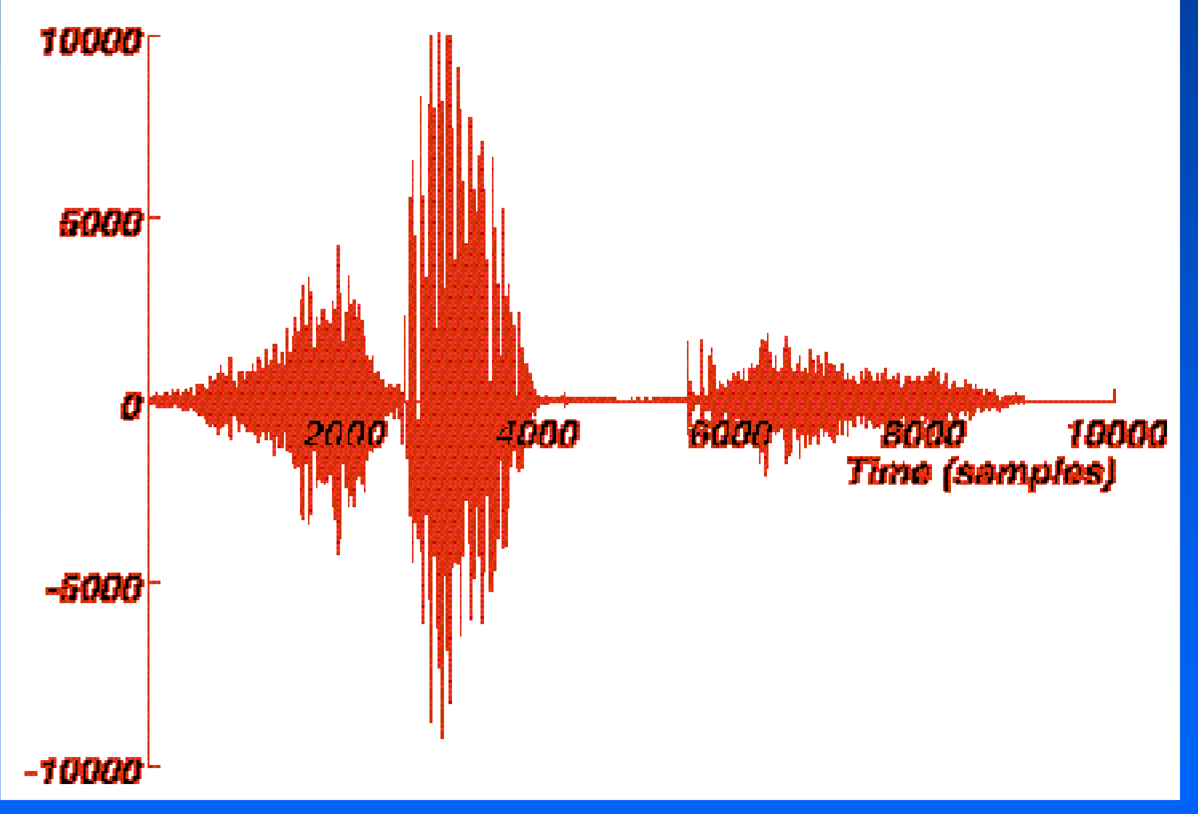
- **Major functional components:**
  - Signal processing to extract features from speech waveforms
  - Comparison of features to pre-stored templates
- **Important design choices:**
  - Choice of features
  - Specific method of comparing features to stored templates

GOALS OF SPEECH REPRESENTATIONS

- Capture important phonetic information in speech
- Computational efficiency
- Efficiency in storage requirements
- Optimize generalization

WHY PERFORM SIGNAL PROCESSING?

A look at the **time-domain waveform** of “six”:



It’s hard to infer much from the time-domain waveform

Carnegie Mellon

WHY PERFORM SIGNAL PROCESSING IN THE FREQUENCY DOMAIN?

- Human hearing is based on frequency analysis
- Use of frequency analysis often simplifies signal processing
- Use of frequency analysis often facilitates understanding

FEATURES FOR SPEECH RECOGNITION: CEPSTRAL COEFFICIENTS

- The cepstrum is the inverse transform of the log of the magnitude of the spectrum
- Useful for separating convolved signals (like the source and filter in the speech production model)
- Can be thought of as the Fourier series expansion of the magnitude of the Fourier transform
- Generally provides more efficient and robust coding of speech information than LPC coefficients
- Most common basic feature for speech recognition

THREE WAYS OF DERIVING CEPSTRAL COEFFIENTS

- **LPC-derived cepstral coefficients (LPCC):**
  - Compute “traditional” LPC coefficients
  - Convert to cepstra using linear transformation
  - Warp cepstra using bilinear transform

- **Mel-frequency cepstral coefficients (MFCC):**
  - Compute log magnitude of windowed signal
  - Multiply by triangular Mel weighting functions
  - Compute inverse discrete cosine transform

- **Perceptual linear prediction (PLP)**

COMPUTING CEPSTRAL COEFFICIENTS

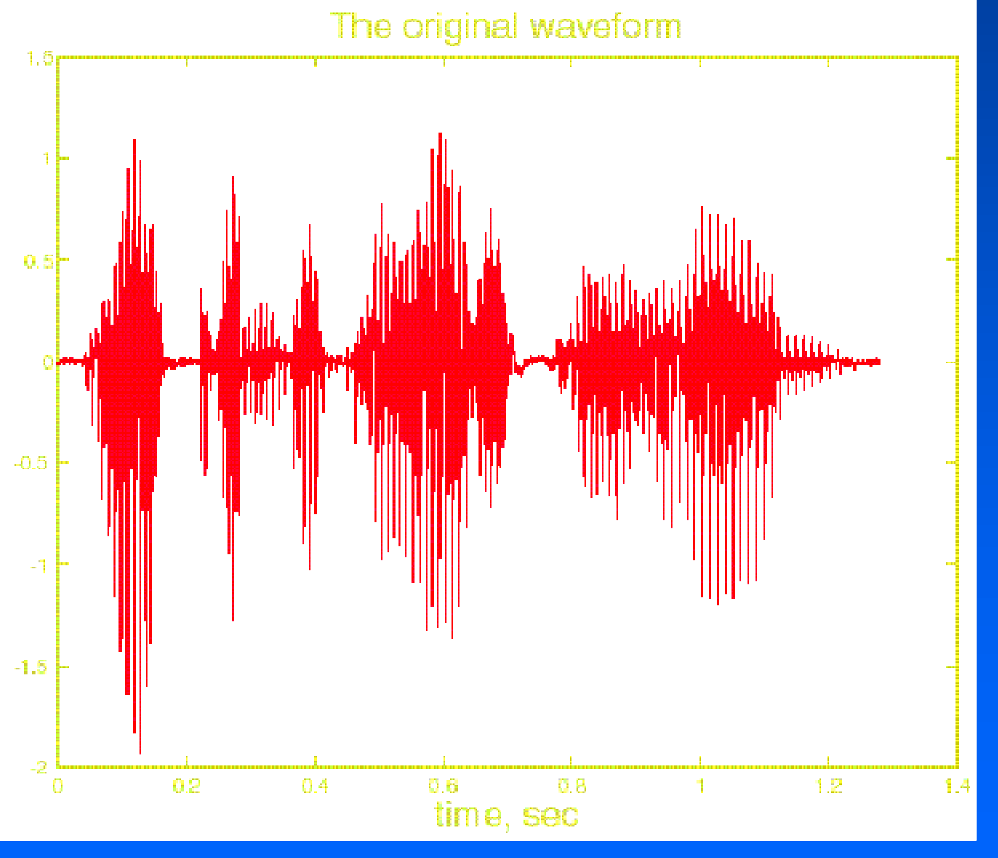
- **Comments:**
  - **MFCC** is currently the most popular representation.
  - Typical systems include a combination of
    - » MFCC coefficients
    - » “Delta” MFCC coefficients
    - » “Delta delta” MFCC coefficients
    - » Power and delta power coefficients

COMPUTING LPC CEPSTRAL COEFFICIENTS

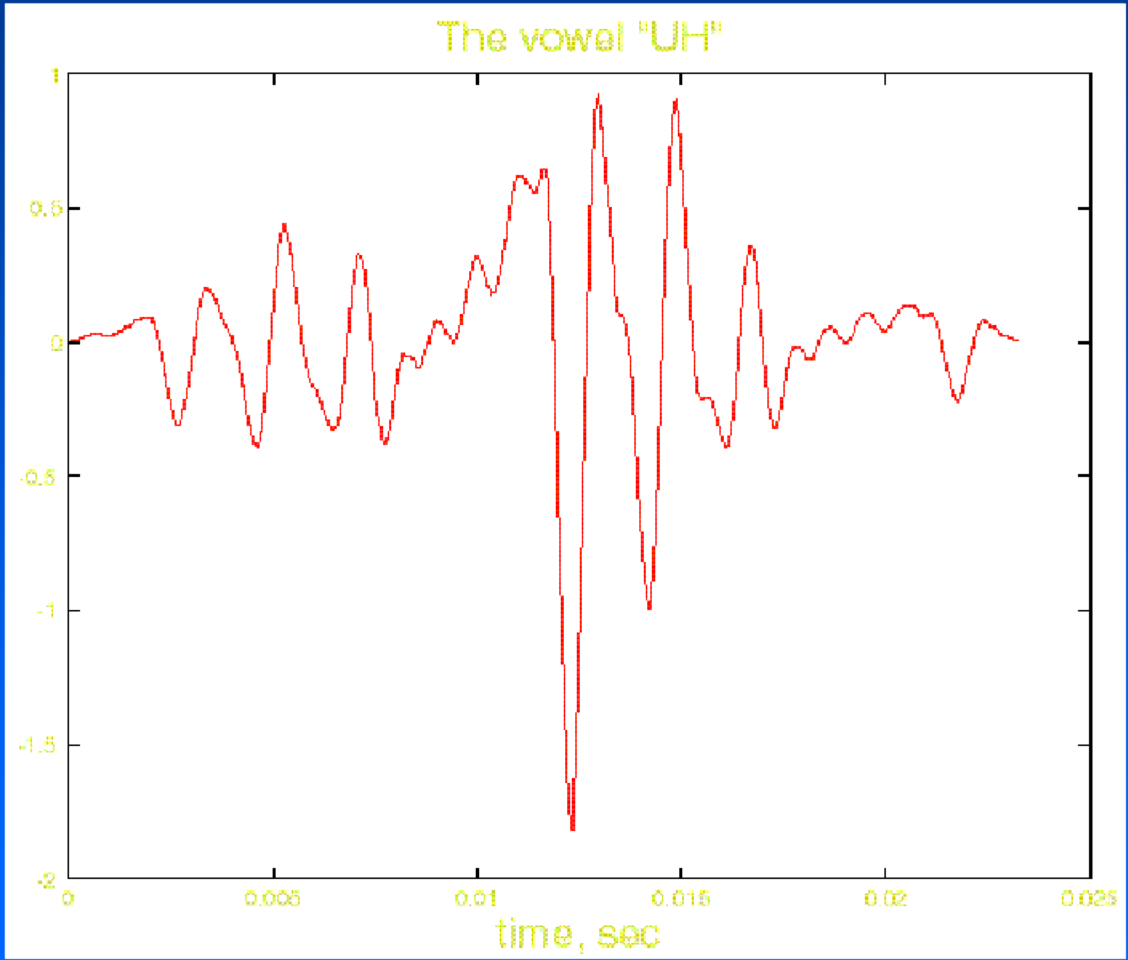
- **Procedure used in SPHINX-I:**
  - A/D conversion at 16-kHz sampling rate
  - Apply Hamming window, duration 320 samples (20 msec) with 50% overlap (100-Hz frame rate)
  - Pre-emphasize to boost high-frequency components
  - Compute first 14 auto-correlation coefficients
  - Perform Levinson-Durbin recursion to obtain 14 LPC coefficients
  - Convert LPC coefficients to cepstral coefficients
  - Perform frequency warping to spread low frequencies
  - Apply vector quantization to generate three codebooks

An example: the vowel in “welcome”

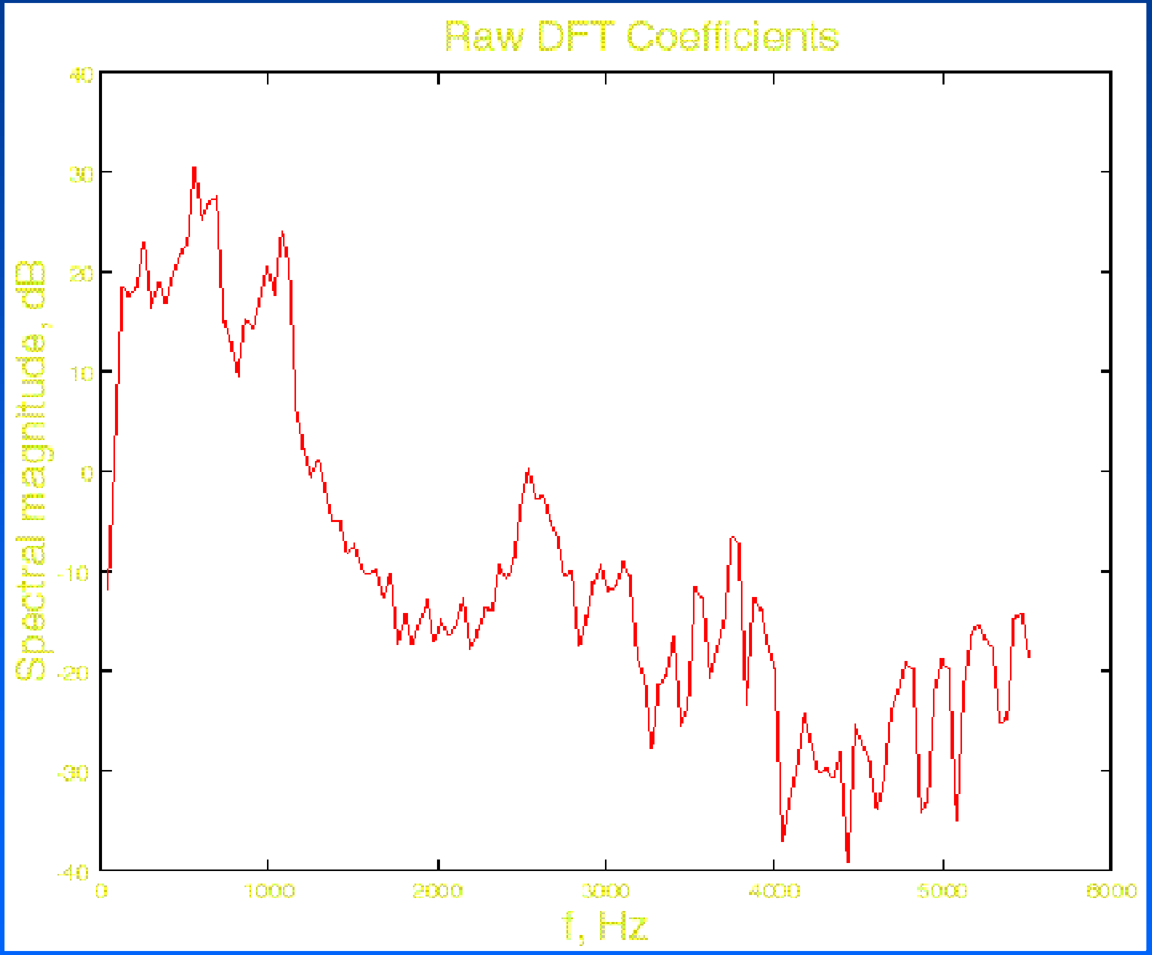
- **The original time function:**



THE TIME FUNCTION AFTER WINDOWING



THE RAW SPECTRUM



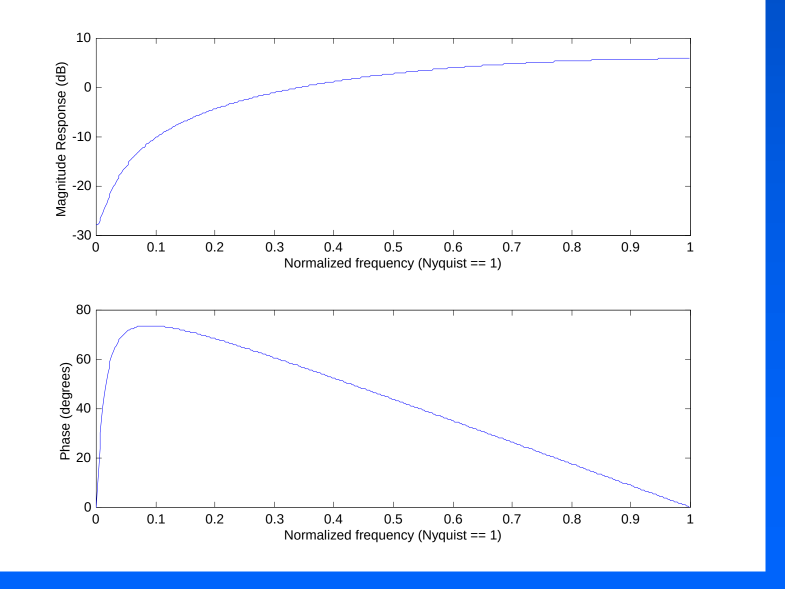
Carnegie Mellon

PRE-EMPHASIZING THE SIGNAL

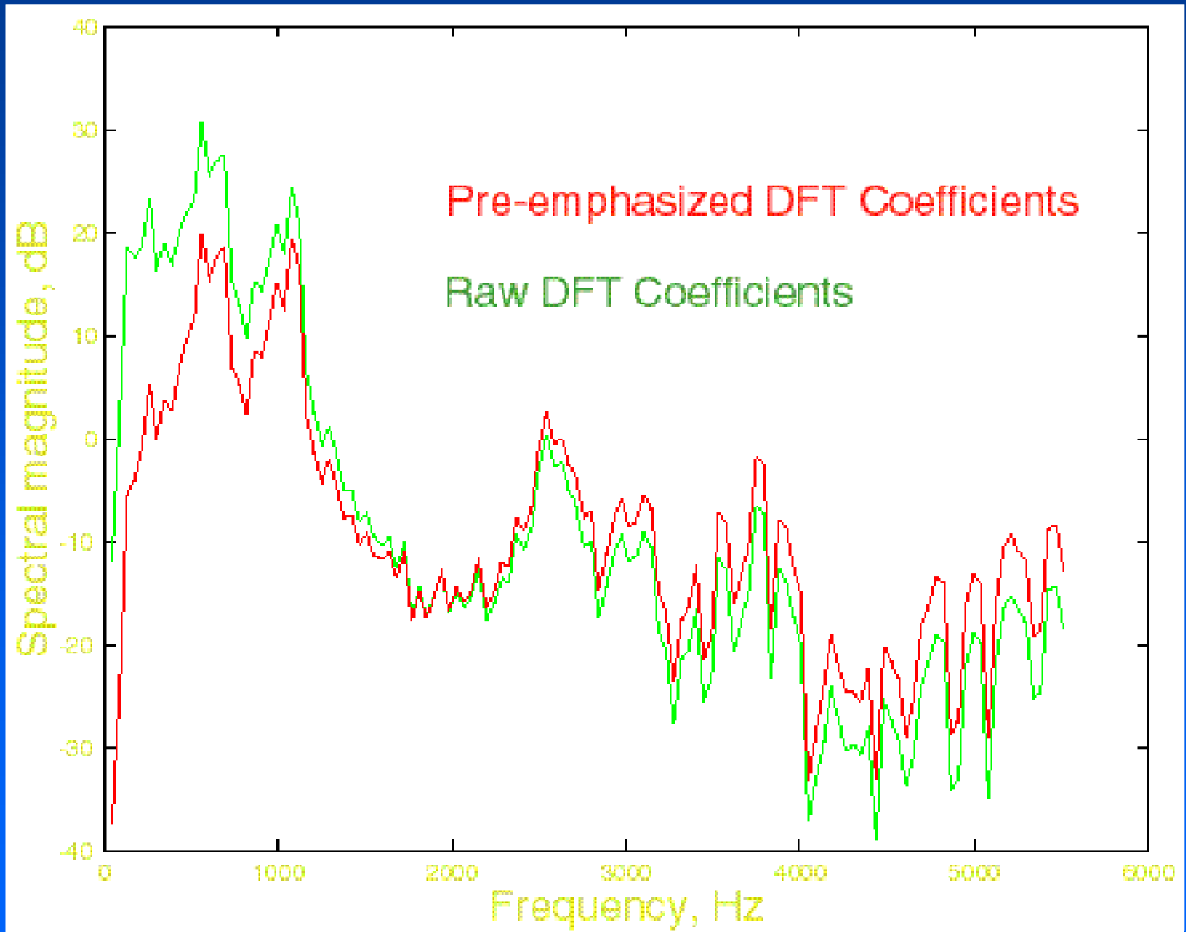
- Typical pre-emphasis filter:

$$y[n] = x[n] - .96x[n-1]$$

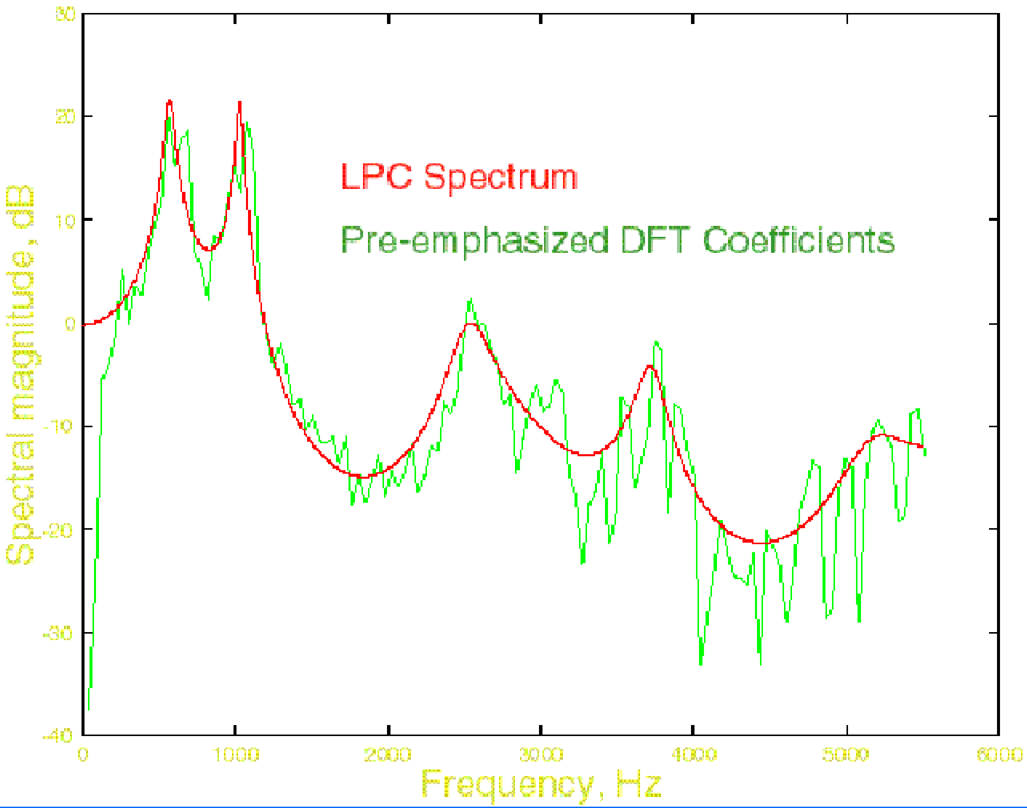
- Its frequency response:



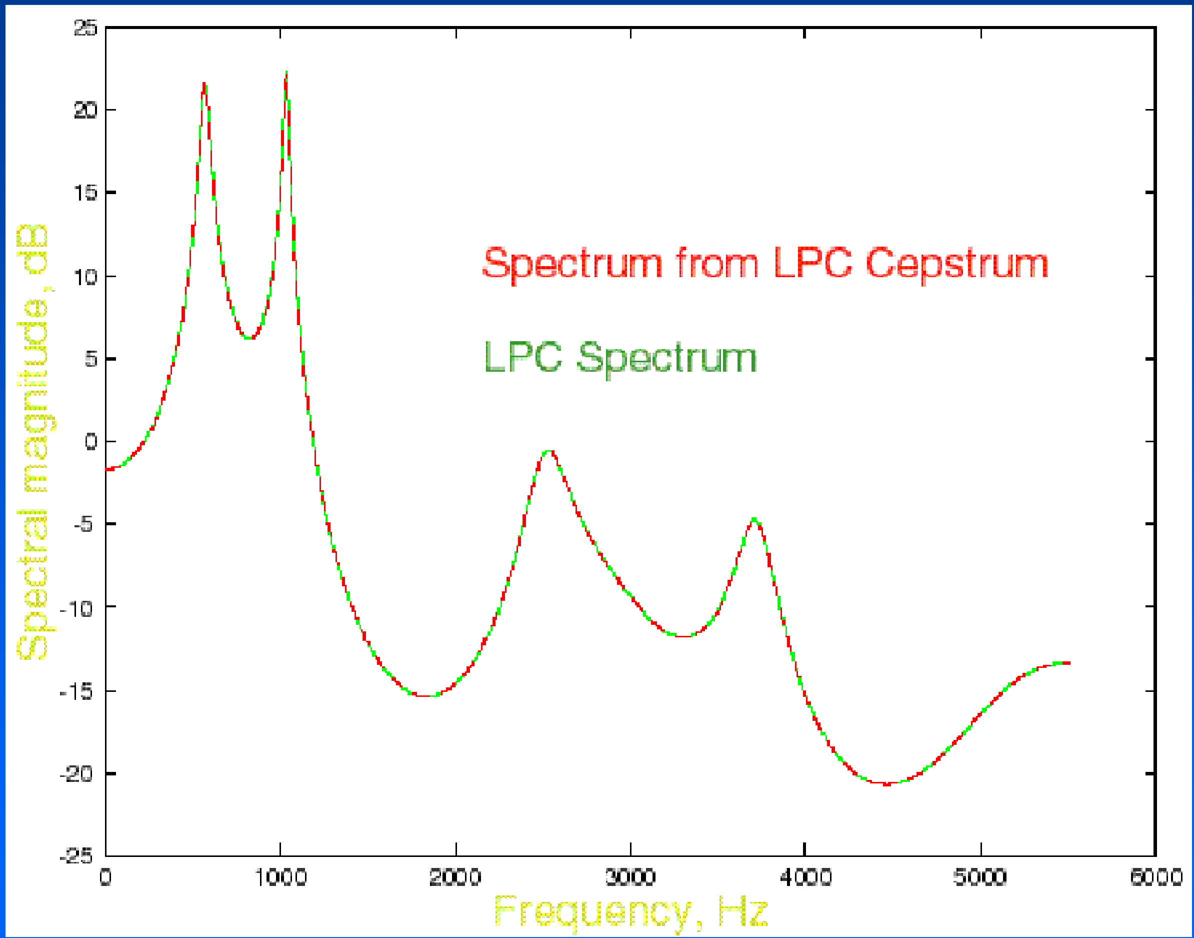
THE SPECTRUM OF THE PRE-EMPHASIZED SIGNAL



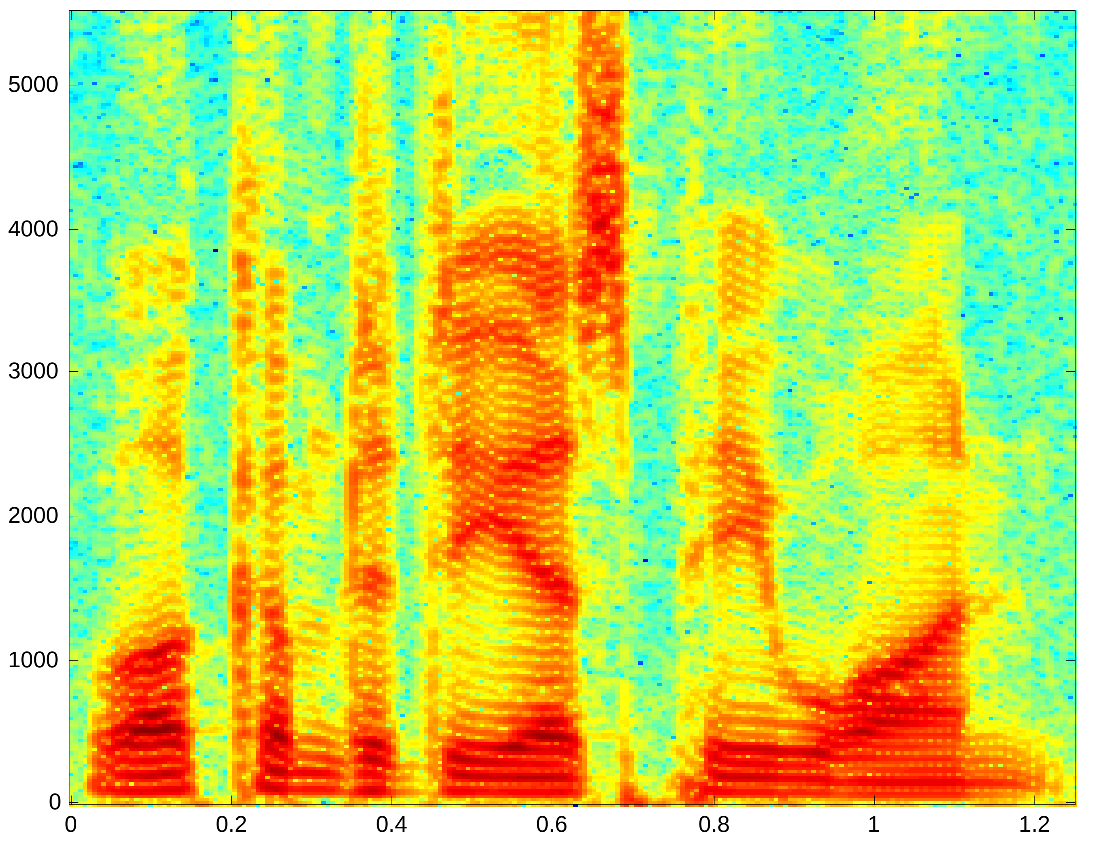
THE LPC SPECTRUM



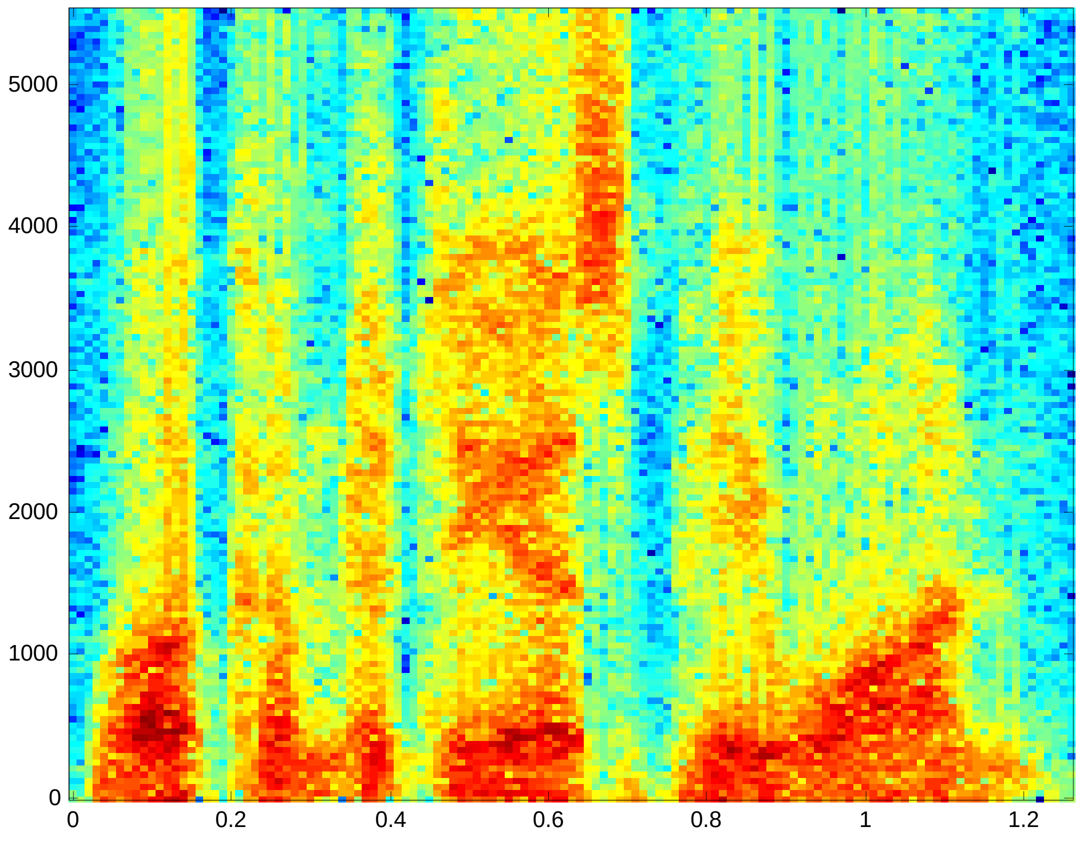
THE TRANSFORM OF THE CEPSTRAL COEFFICIENTS



THE BIG PICTURE:
THE ORIGINAL SPECTROGRAM

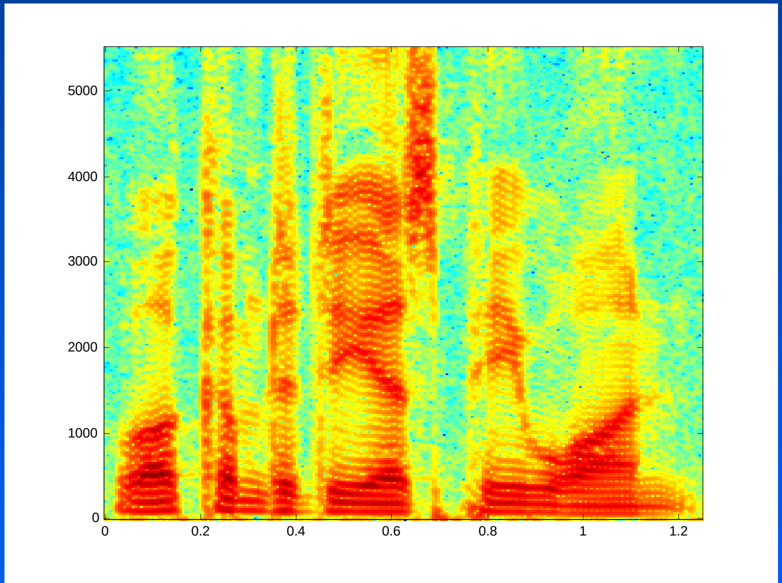


EFFECTS OF LPC PROCESSING

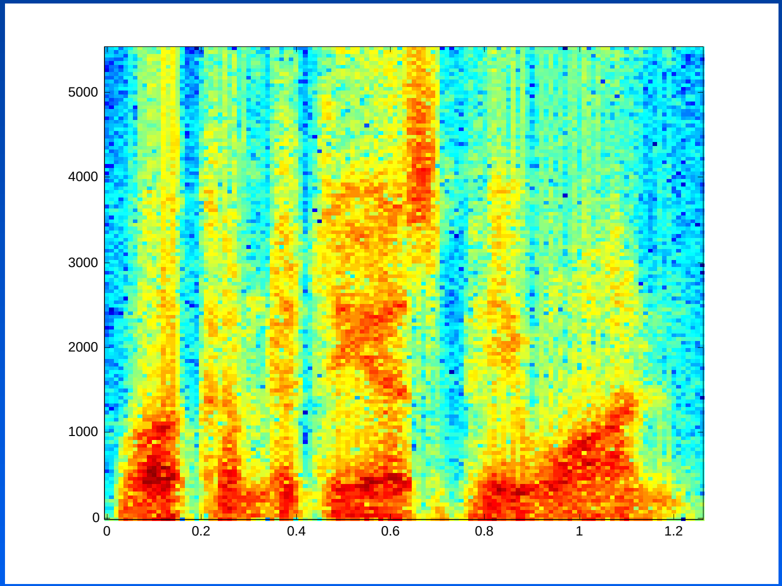


COMPARING REPRESENTATIONS

ORIGINAL SPEECH



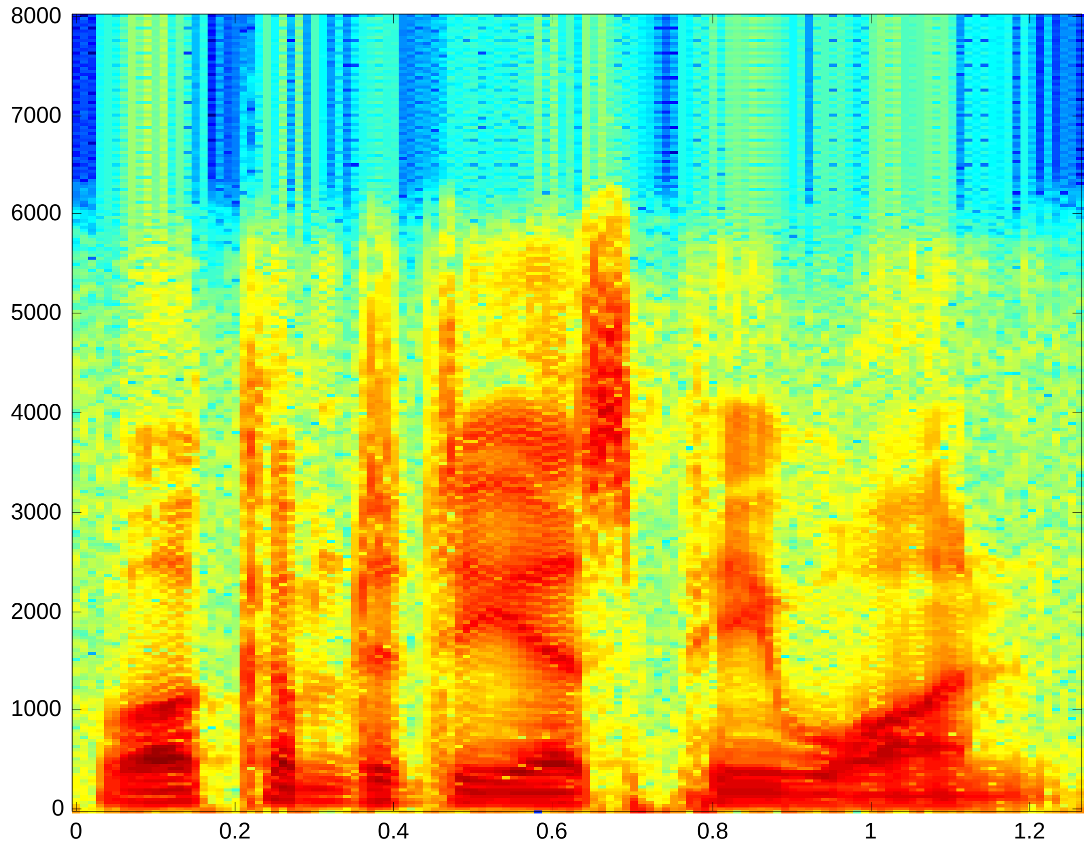
LPCC CEPSTRA (unwarped)



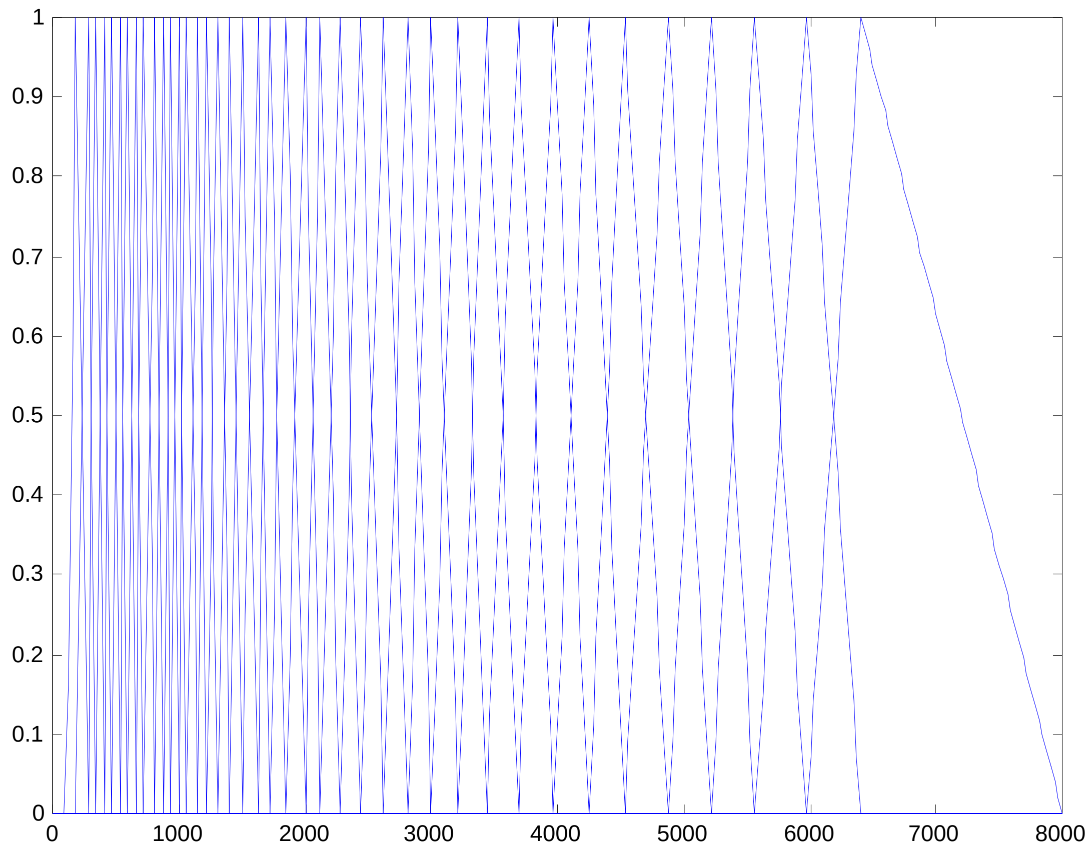
COMPUTING MEL FREQUENCY CEPSTRAL COEFFICIENTS

- Segment incoming waveform into frames
- Compute frequency response for each frame using DFTs
- Multiply magnitude of frequency response by triangular weighting functions to produce 25-40 channels
- Compute log of weighted magnitudes for each channel
- Take inverse discrete cosine transform (DCT) of weighted magnitudes for each channel, producing ~14 cepstral coefficients for each frame
- Calculate delta and double-delta coefficients

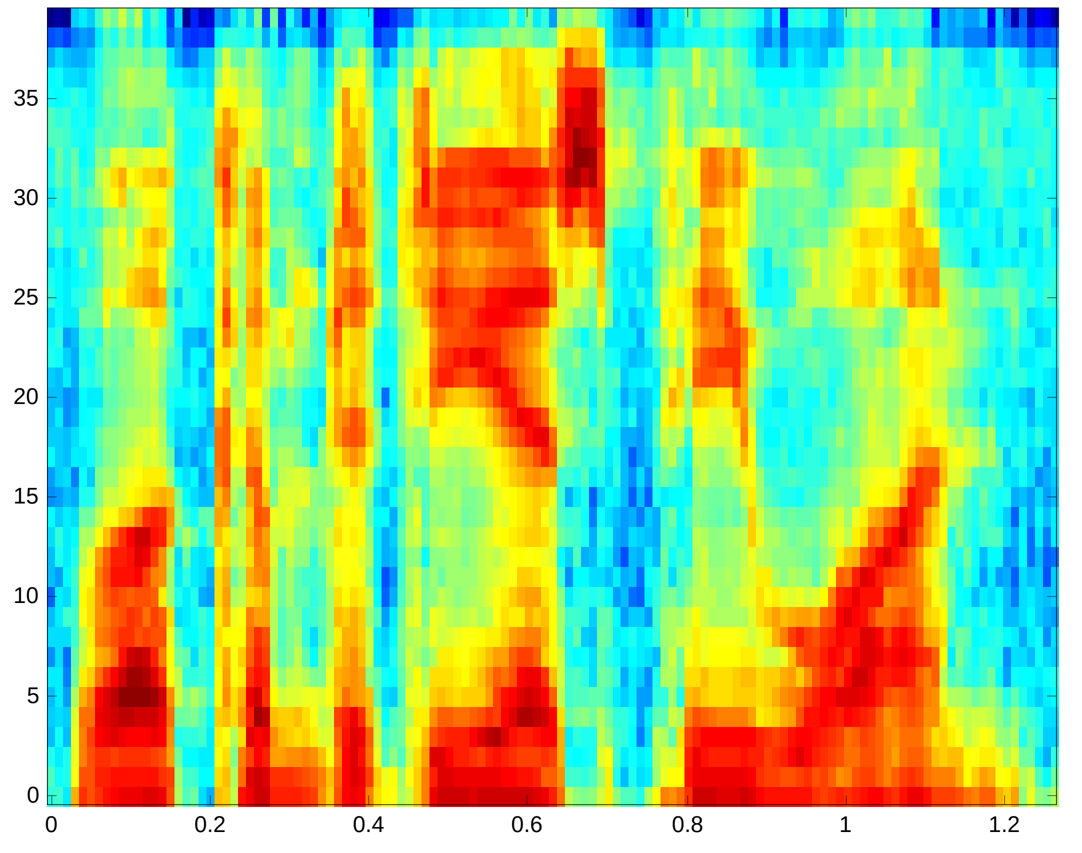
AN EXAMPLE: DERIVING MFCC coefficients



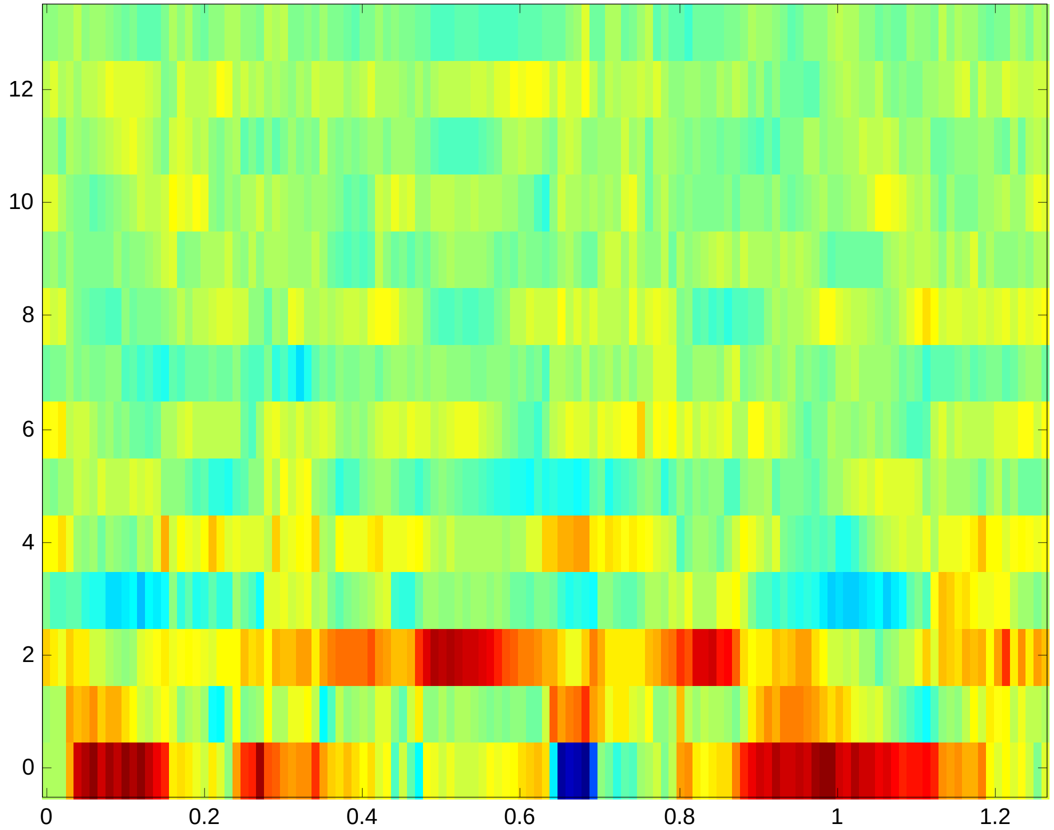
THE MEL WEIGHTING FUNCTIONS



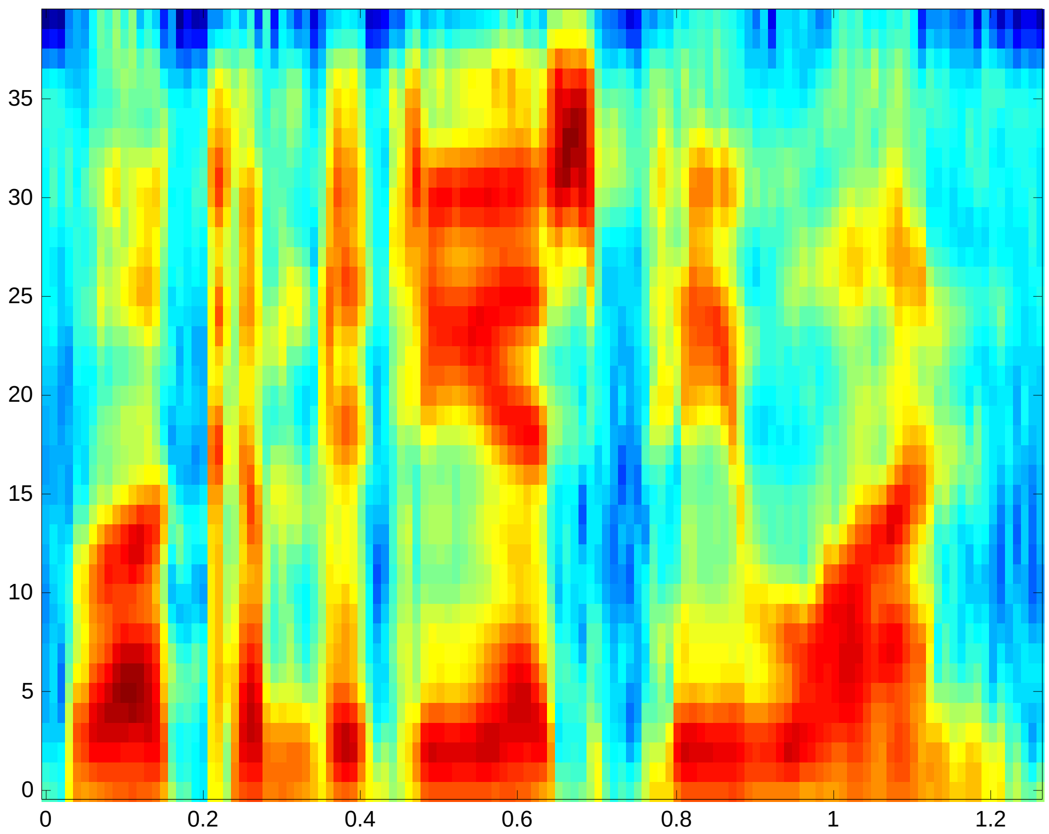
THE LOG ENERGIES OF THE MEL FILTER OUTPUTS



THE CEPSTRAL COEFFICIENTS

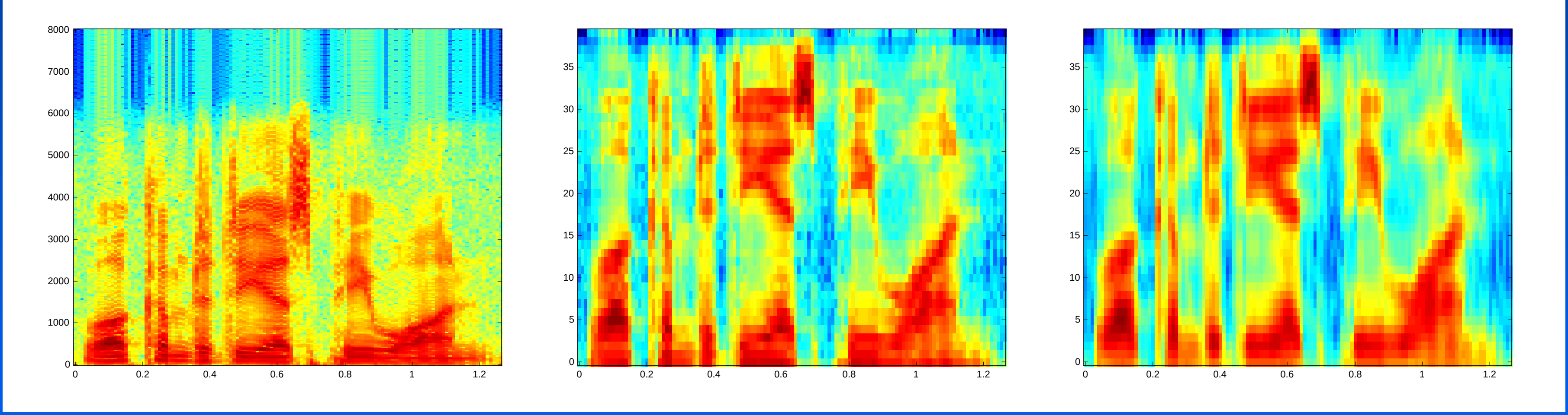


LOGSPECTRA RECOVERED FROM CEPSTRA



COMPARING SPECTRAL REPRESENTATIONS

ORIGINAL SPEECH | MEL LOG MAGS | AFTER CEPSTRA



Comments on the MFCC representation

- **It's very "blurry" compared to a wideband spectrogram!**
- **Aspects of auditory processing represented:**
  – Frequency selectivity and spectral bandwidth (but using a constant analysis window duration!)
    » Wavelet schemes exploit time-frequency resolution better
  – Nonlinear amplitude response
- **Aspects of auditory processing NOT represented:**
  – Detailed timing structure
  – Lateral suppression
  – Enhancement of temporal contrast
  – Other auditory nonlinearities

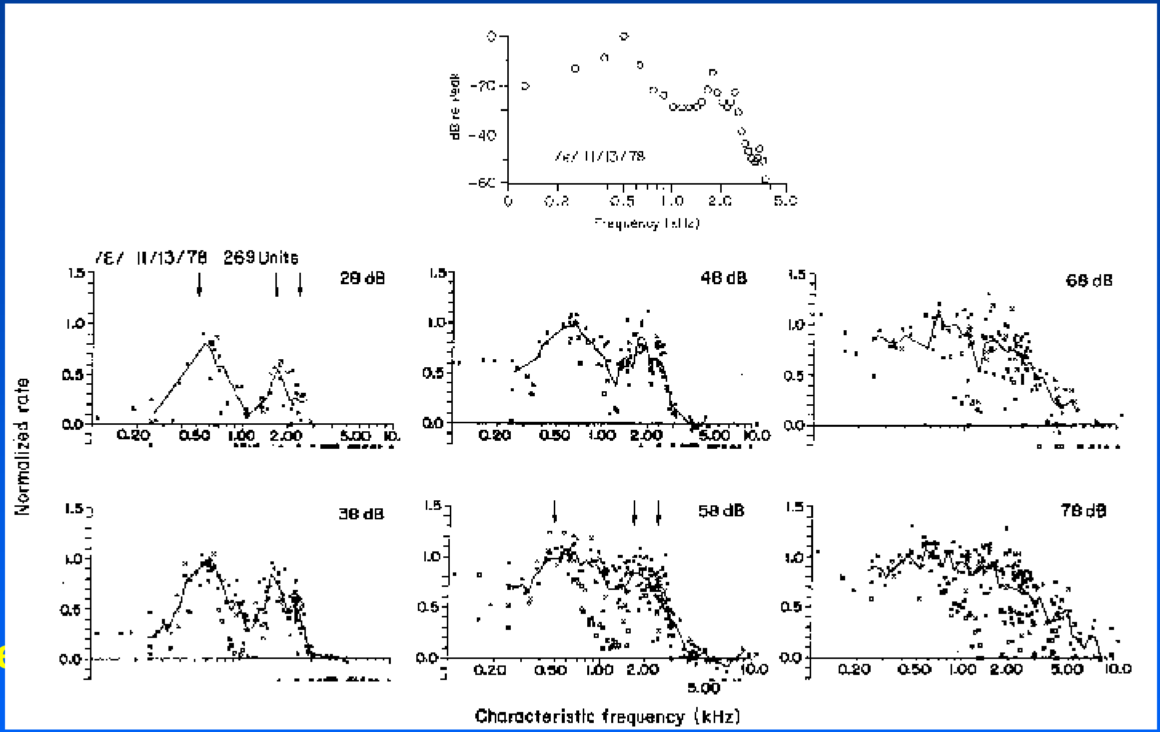
Carnegie Mellon

Outline of discussion

- Basic mechanisms of speech production
- Basic mechanisms of auditory perception
- (Very!) basic review of automatic speech recognition
- Conventional signal processing for speech recognition
- Signal processing for improved speech recognition
- Signal processing for improved sound source separation

Speech representation using mean rate

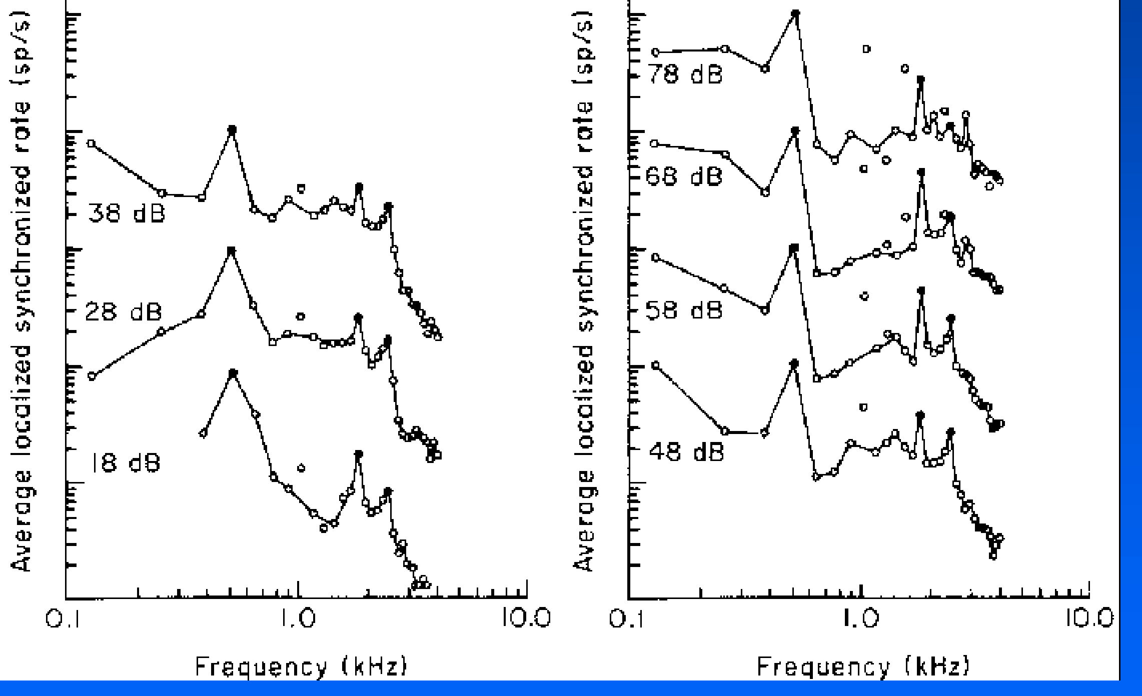
- Representation of vowels by Young and Sachs using mean rate:



- Mean rate

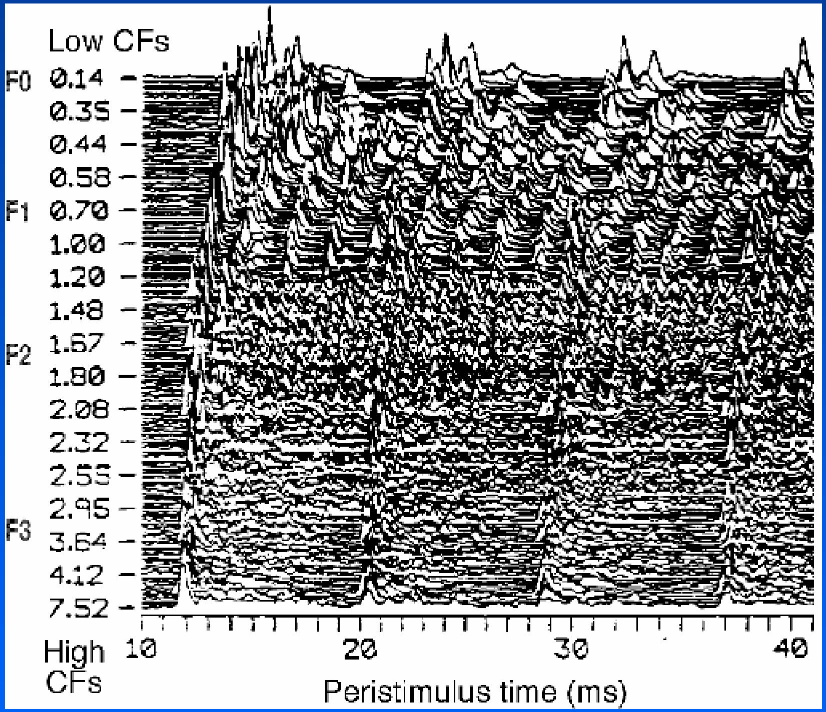
Speech representation using average localized synchrony measure

- Representation of vowels by Young and Sachs using ALSR:



The importance of timing information

- Re-analysis of Young-Sachs data by Searle:



- Temporal processing captures dominant formants in a spectral region

Paths to the realization of temporal fine structure in speech

- **Correlograms (Slaney and Lyon)**
- **Computations based on interval processing**
  - Ghitza's Ensemble Interval Histogram (EIH) model
  - Kim's Zero Crossing Peak Analysis (ZCPA) model