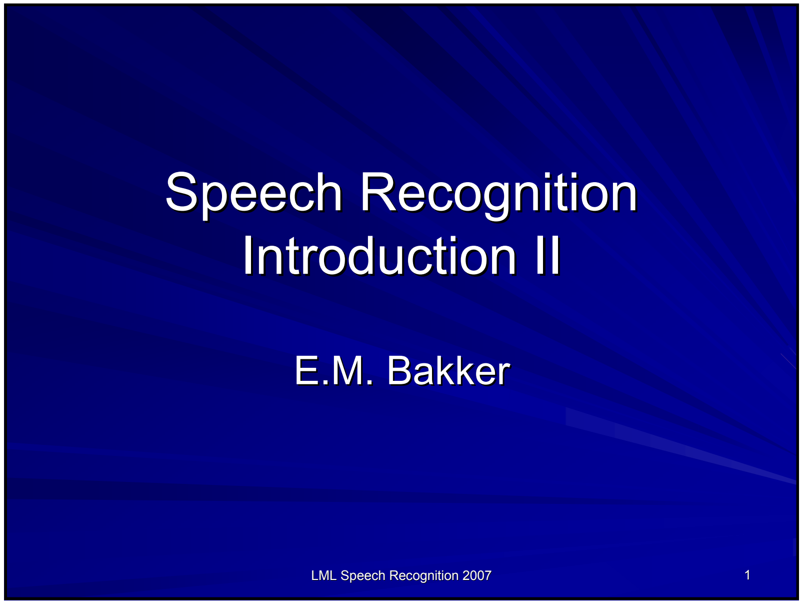
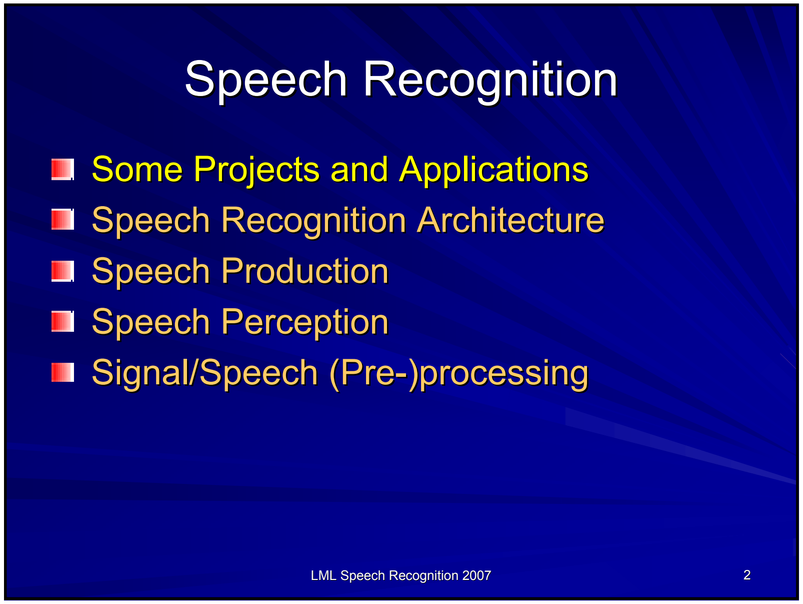


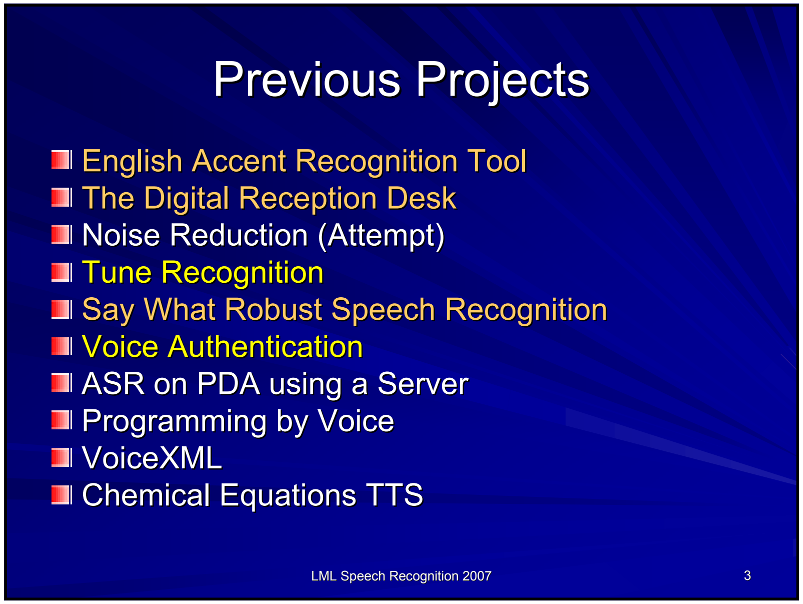
Speech Recognition Introduction II

E.M. Bakker



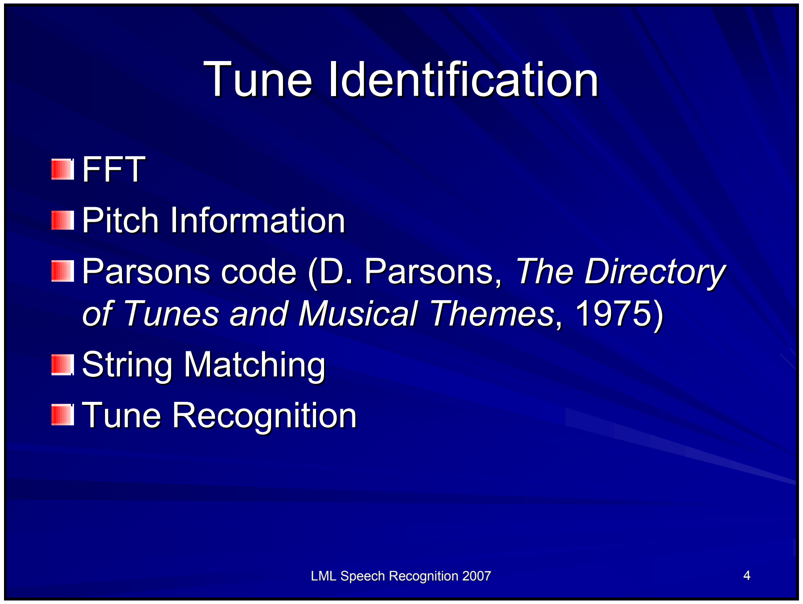
Speech Recognition

- Some Projects and Applications
- Speech Recognition Architecture
- Speech Production
- Speech Perception
- Signal/Speech (Pre-)processing



Previous Projects

- English Accent Recognition Tool
- The Digital Reception Desk
- Noise Reduction (Attempt)
- Tune Recognition
- Say What Robust Speech Recognition
- Voice Authentication
- ASR on PDA using a Server
- Programming by Voice
- VoiceXML
- Chemical Equations TTS



Tune Identification

- FFT
- Pitch Information
- Parsons code (D. Parsons, *The Directory of Tunes and Musical Themes*, 1975)
- String Matching
- Tune Recognition

Speaker Identification

- Learning Phase
  - Special text read by subjects
  - FFT-Energy features stored as template
- Recognition by dynamic time warping to match the stored templates (Euclidean distance with threshold)
- Second method proposed uses Vector Quantization

Audio Indexing

- Several Audio Classes
  - Car, Explosion, Voices, Wind, Sea, Crowd, etc.
  - Classical Music, Pop Music, R&B, Trance, Romantic, etc.
- Determine features capturing, pitch, rhythm, loudness, etc.
  - Short time Energy
  - Zero crossing rates
  - Level crossing rates
  - Spectral energy, formant analysis, etc.
- Use Vector Quantization for learning and recognizing the different classes

Bimodal Emotion Recognition

Nicu Sebe[1], Erwin Bakker[2], Ira Cohen[3], Theo Gevers[1], Thomas S. Huang[4]

[1]University of Amsterdam, The Netherlands
[2]Leiden University, The Netherlands
[3]HP Labs, USA
[4]University of Illinois at Urbana-Champaign, USA

(Sept, 2005)

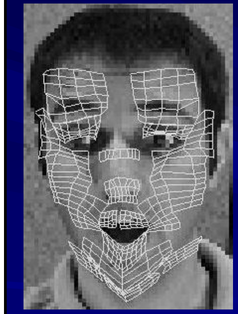
Emotion from Auditory Cues: Prosody

- Prosody is the melody or musical nature of the spoken voice
- We are able to differentiate many emotions from prosody alone e.g. anger, sadness, happiness
- Universal and early skill
- Are the neural bases for this ability the same as for differentiating emotion from visual cues?

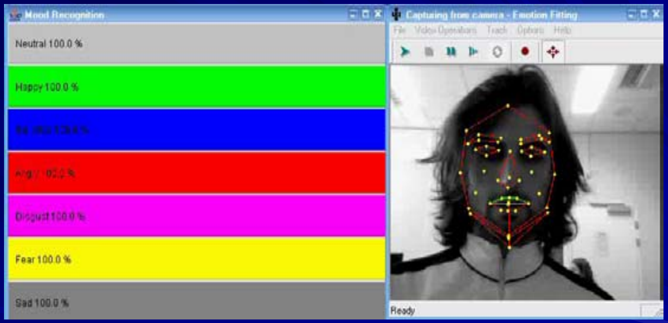
Bimodal Emotion Recognition: Experiments

- Video features
  - "Connected Vibration" video tracking
  - Eyebrow position, cheek lifting, mouth opening, etc.
- Audio features
  - "Prosodic features": 'Prosody' ~ the melody of the voice.
    - logarithm of energy
    - syllable rate
    - pitch

Face Tracking



- 2D image motions are measured using template matching between frames at different resolutions
- 3D motion can be estimated from the 2D motions of many points of the mesh
- The recovered motions are represented in terms of magnitudes of facial features
- Each feature motion corresponds to a simple deformation of the face



Bimodal Database

Applications
Audio Indexing of Broadcast News

Broadcast news offers some unique challenges:

- Lexicon: important information in infrequently occurring words
- Acoustic Modeling: variations in channel, particularly within the same segment (“ in the studio” vs. “on location”)
- Language Model: must adapt (“ Bush,” “Clinton,” “Bush,” “McCain,” “???”)
- Language: multilingual systems? language-independent acoustic modeling?

Content Based Indexing

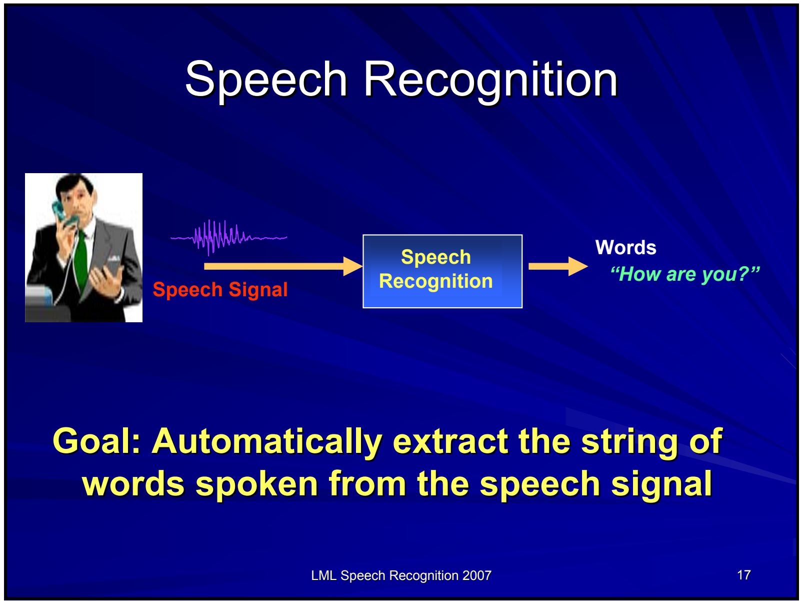
- Language identification
- Speech Recognition
- Speaker Recognition
- Emotion Recognition
- Environment Recognition: indoor, outdoor, etc.
- Object Recognition: car, plane, gun, footsteps, etc.
- …

Meta Data Extraction

- Relative location of the speaker?
- Who is speaking?
- What emotions are expressed?
- Which language is spoken?
- What is spoken?
- What are the keywords? (Indexing)
- What is the meaning of the spoken text?
- Etc.

Speech Recognition

- Some Projects and Applications
- Speech Recognition Architecture
- Speech Production
- Speech Perception
- Signal/Speech (Pre-)processing



Speech Recognition

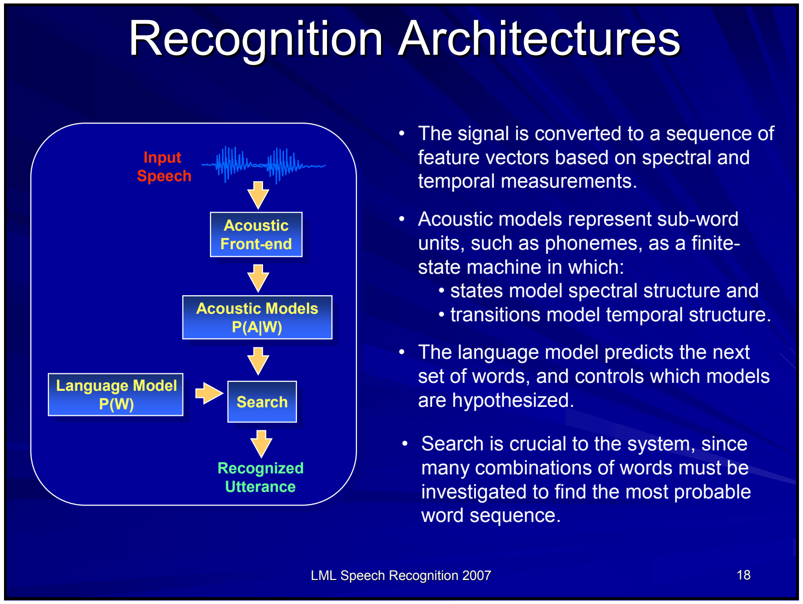
Speech Signal

Speech Recognition

Words

"How are you?"

Goal: Automatically extract the string of words spoken from the speech signal



Recognition Architectures

Input Speech

Acoustic Front-end

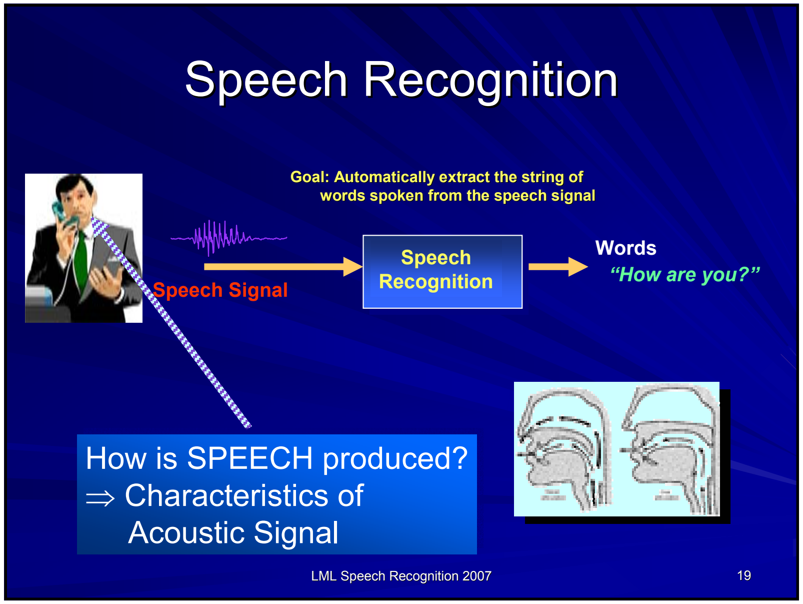
Acoustic Models P(A|W)

Language Model P(W)

Search

Recognized Utterance

- The signal is converted to a sequence of feature vectors based on spectral and temporal measurements.
- Acoustic models represent sub-word units, such as phonemes, as a finite-state machine in which:
  - states model spectral structure and
  - transitions model temporal structure.
- The language model predicts the next set of words, and controls which models are hypothesized.
- Search is crucial to the system, since many combinations of words must be investigated to find the most probable word sequence.



Speech Recognition

Goal: Automatically extract the string of words spoken from the speech signal

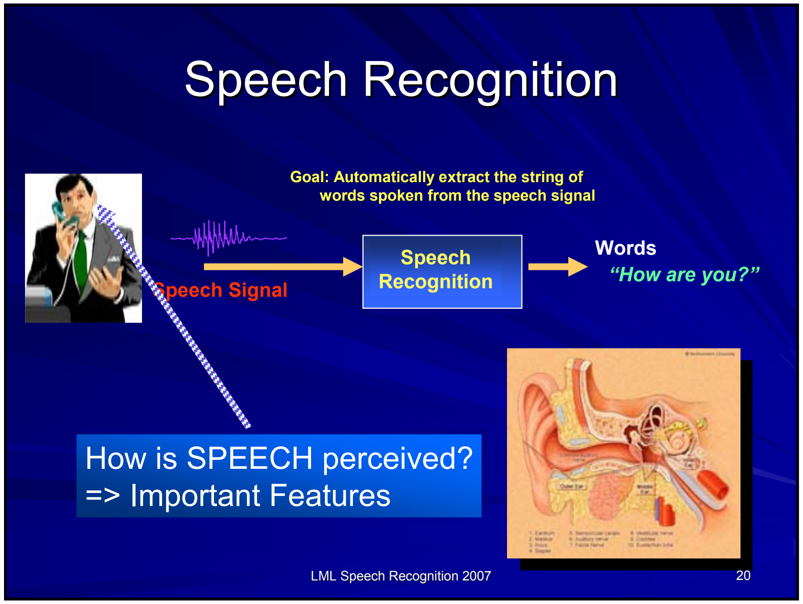
Speech Signal

Speech Recognition

Words

"How are you?"

How is SPEECH produced?
⇒ Characteristics of
Acoustic Signal



Speech Recognition

Goal: Automatically extract the string of words spoken from the speech signal

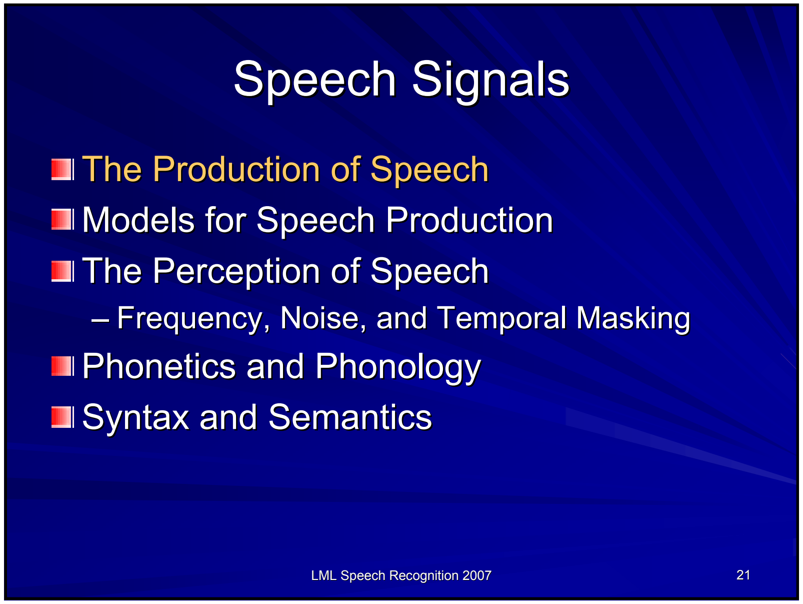
Speech Signal

Speech Recognition

Words

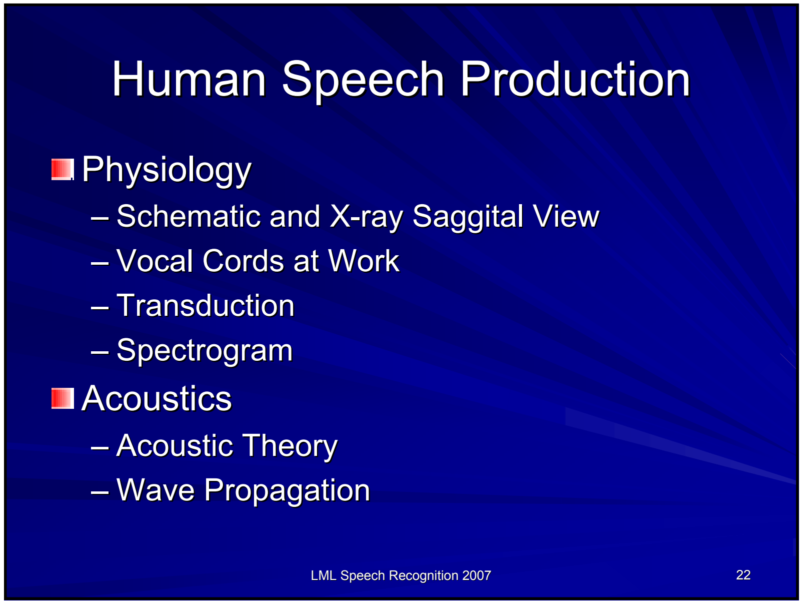
"How are you?"

How is SPEECH perceived?
=> Important Features



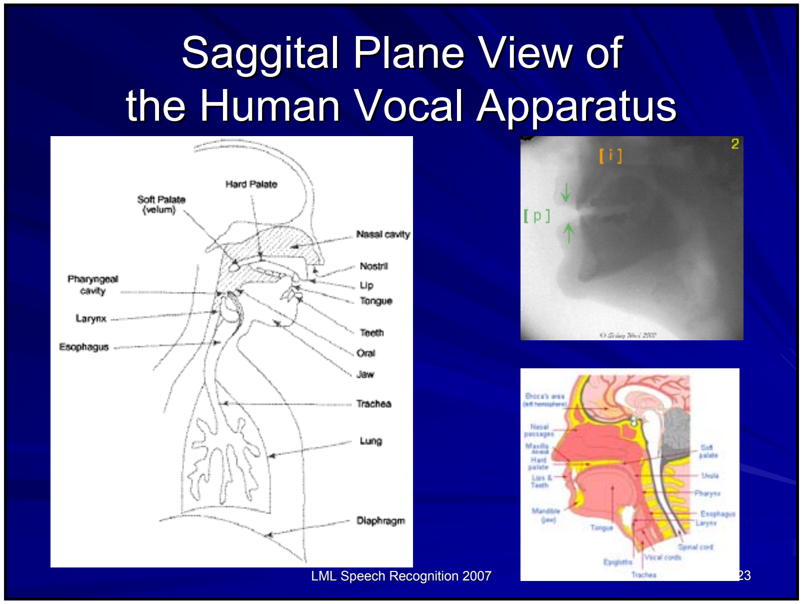
Speech Signals

- The Production of Speech
- Models for Speech Production
- The Perception of Speech
  - – Frequency, Noise, and Temporal Masking
- Phonetics and Phonology
- Syntax and Semantics

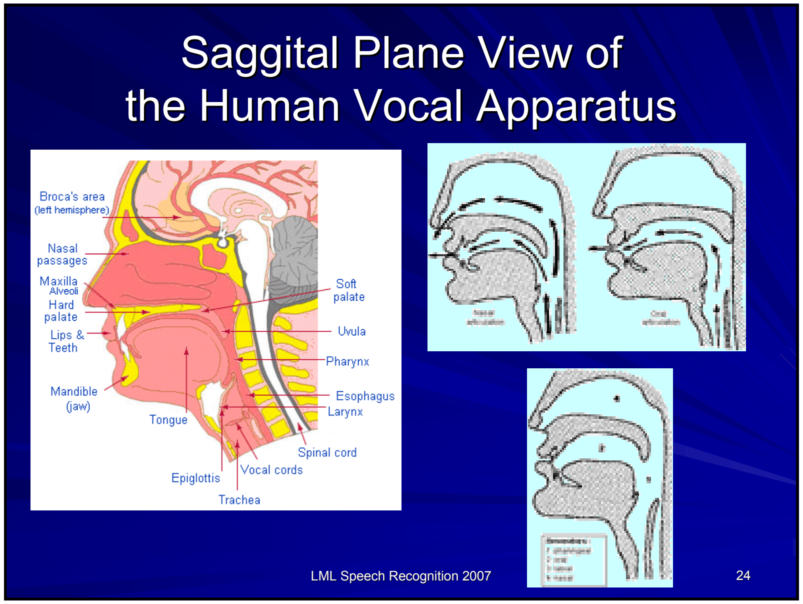


Human Speech Production

- Physiology
  - – Schematic and X-ray Saggital View
  - – Vocal Cords at Work
  - – Transduction
  - – Spectrogram
- Acoustics
  - – Acoustic Theory
  - – Wave Propagation



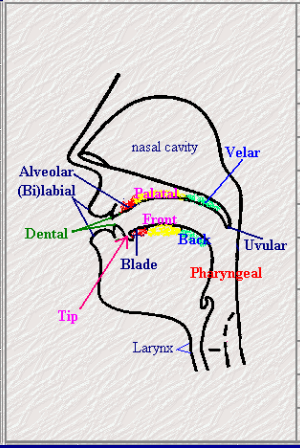
Saggital Plane View of the Human Vocal Apparatus



Saggital Plane View of the Human Vocal Apparatus

Characterization of English Phonemes

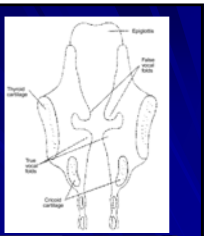
/f/	Labio-dental
/v/	Labio-dental
/θ/	Tip-dental
/ð/	Tip-dental
/s/	Blade-alveolar
/z/	Blade-alveolar
/ʃ/	Blade/front –palato-alveolar
/ʒ/	Blade/front –palato-alveolar
/h/	Glottal
/l/	Tip-alveolar
/r/	Blade-postalveolar
/w/	Bilabial back-velar
/j/	Front-palatal



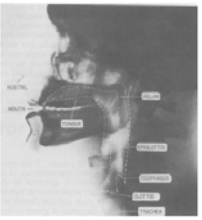
Consonants	Place
/p/	Bilabial
/b/	Bilabial
/t/	Tip-alveolar
/d/	Tip-alveolar
/k/	Back-velar
/g/	Back-velar
/tʃ/	Blade/front –palato-alveolar
/dʒ/	Blade/front –palato-alveolar
/m/	Bilabial
/n/	Tip-alveolar
/ŋ/	Back-velar

Vocal Chords

- The Source of Sound



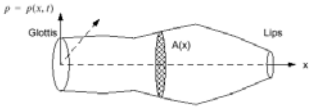
Models for Speech Production



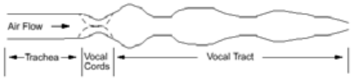
A detailed acoustic theory must consider the effects of the following:

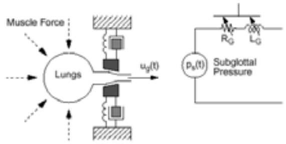
- Time variation of the vocal tract shape
- Losses due to heat conduction and viscous friction at the vocal tract walls
- Softness of the vocal tract walls
- Radiation of sound at the lips
- Nasal coupling
- Excitation of sound in the vocal tract

Let us begin by considering a simple case of a lossless tube:



How do we couple energy into the vocal tract?





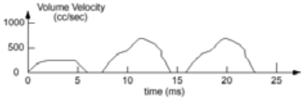
The glottal impedance can be approximated by:

$$Z_G = R_G + j\Omega L_G$$

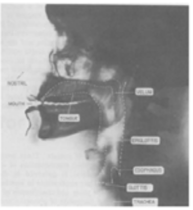
The boundary condition for the volume velocity is:

$$U(0, \Omega) = U_G(\Omega) - P(0, \Omega)/Z_G(\Omega)$$

For voiced sounds, the glottal volume velocity looks something like this:



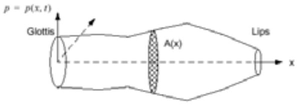
Models for Speech Production

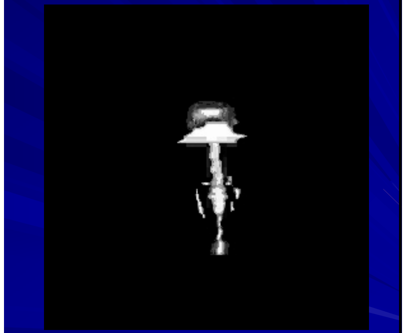


A detailed acoustic theory must consider the effects of the following:

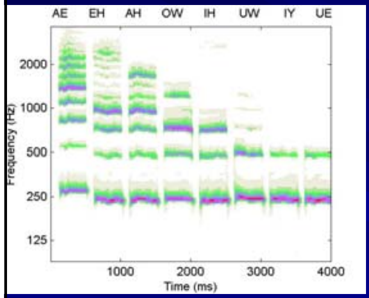
- Time variation of the vocal tract shape
- Losses due to heat conduction and viscous friction at the vocal tract walls
- Softness of the vocal tract walls
- Radiation of sound at the lips
- Nasal coupling
- Excitation of sound in the vocal tract

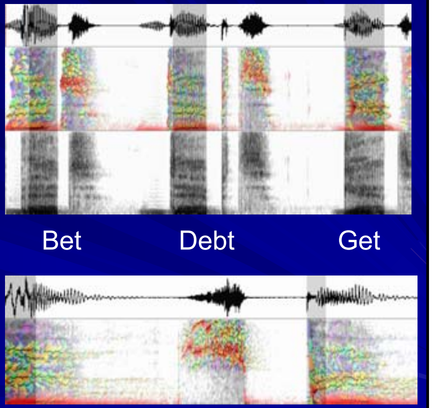
Let us begin by considering a simple case of a lossless tube:





English Phonemes





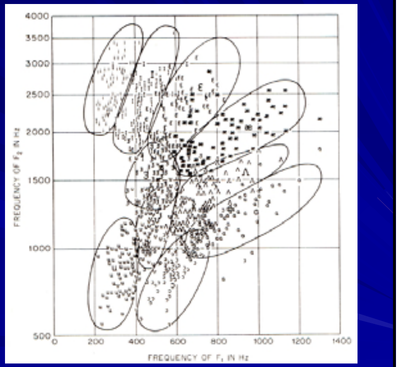
Bet Debt Get

Pin Sp i n

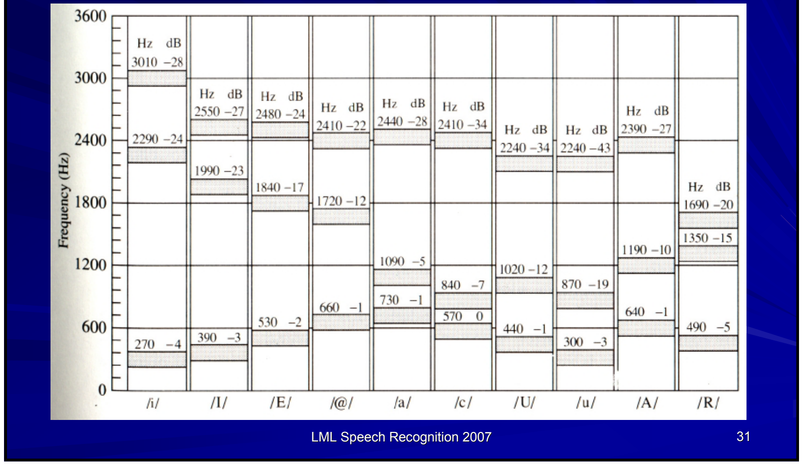
Allophone

The Vowel Space

- We can characterize a vowel sound by the locations of the first and second spectral resonances, known as formant frequencies:
- Some voiced sounds, such as diphthongs, are transitional sounds that move from one vowel location to another.



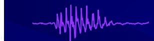
Phonetics
Formant Frequency Ranges



	/i/	/I/	/E/	/@/	/a/	/c/	/U/	/u/	/A/	/R/
Hz dB	3010 −28	2550 −27	2480 −24	2410 −22	2440 −28	2410 −34	2240 −34	2240 −43	2390 −27	1690 −20
	2290 −24	1990 −23	1840 −17	1720 −12	1090 −5	840 −7	1020 −12	870 −19	1190 −10	1350 −15
	270 −4	390 −3	530 −2	660 −1	730 −1	570 0	440 −1	300 −3	640 −1	490 −5

Speech Recognition

Goal: Automatically extract the string of words spoken from the speech signal



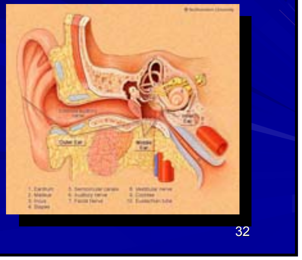
Speech Signal

Speech Recognition

Words

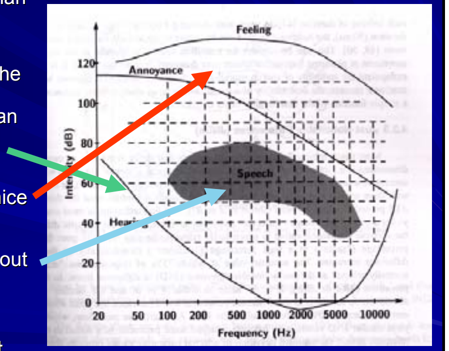
"How are you?"

How is SPEECH perceived?



The Perception of Speech Sound Pressure

- The ear is the most sensitive human organ. Vibrations on the order of angstroms are used to transduce sound. It has the largest dynamic range (~140 dB) of any organ in the human body.
- The lower portion of the curve is an audiogram - hearing sensitivity. It can vary up to 20 dB across listeners.
- Above 120 dB corresponds to a nice pop-concert (or standing under a Boeing 747 when it takes off).
- Typical ambient office noise is about 55 dB.

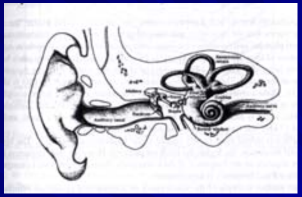


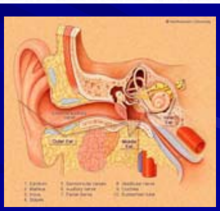
x dB = 10 $\log_{10}(x/x_0)$, x_0 = 1kHz signal with intensity that is just hearable.

dB (SPL)	Source (with distance)
194	Theoretical limit for a sound wave at 1 atmosphere environmental pressure; pressure waves with a greater intensity behave as shock waves.
188	Space Shuttle liftoff as heard from launch tower (less than 100 feet) (source: acoustical studies [1] [2]).
180	Krakatoa volcano explosion at 1 mile (1.6 km) in air [3]
160	M1 Garand being fired at 1 meter (3 ft); Space Shuttle liftoff as heard from launch pad perimeter (approx. 1500 feet) (source: acoustical studies [4] [5]).
150	Jet engine at 30 m (100 ft)
140	Low Calibre Rifle being fired at 1m (3 ft); the engine of a Formula One car at 1 meter (3 ft)
130	Threshold of pain; civil defense siren at 100 ft (30 m)
120	Space Shuttle from three mile mark, closest one can view launch. (Source: acoustical studies) [6] [7]. [Train horn]] at 1 m (3 ft). Many foghorns produce around this volume.
110	Football stadium during kickoff at 50 yard line; chainsaw at 1 m (3 ft)
100	Jackhammer at 2 m (7 ft); inside discothèque
90	Loud factory, heavy truck at 1 m (3 ft), kitchen blender
80	Vacuum cleaner at 1 m (3 ft), curbside of busy street, PLVI of city
70	Busy traffic at 5 m (16 ft)
60	Office or restaurant inside
50	Quiet restaurant inside
40	Residential area at night
30	Theatre, no talking
20	Whispering
10	Human breathing at 3 m (10 ft)
0	Threshold of human hearing (with healthy ears); sound of mosquito flying 3 m (10 ft) away

The Perception of Speech The Ear

- Three main sections, outer, middle, and inner:
  - The outer and middle ears reproduce the analog signal (impedance matching)
  - the inner ear transduces the pressure wave into an electrical signal.
- The outer ear consists of the external visible part and the auditory canal. The tube is about 2.5 cm long.
- The middle ear consists of the eardrum and three bones (malleus, incus, and stapes). It converts the sound pressure wave to displacement of the oval window (entrance to the inner ear).





The Perception of Speech The Ear

- The inner ear primarily consists of a fluid-filled tube (cochlea) which contains the basilar membrane. Fluid movement along the basilar membrane displaces hair cells, which generate electrical signals.
- There are a discrete number of hair cells (30,000). Each hair cell is tuned to a different frequency.
- Place vs. Temporal Theory: firings of hair cells are processed by two types of neurons (onset chopper units for temporal features and transient chopper units for spectral features).

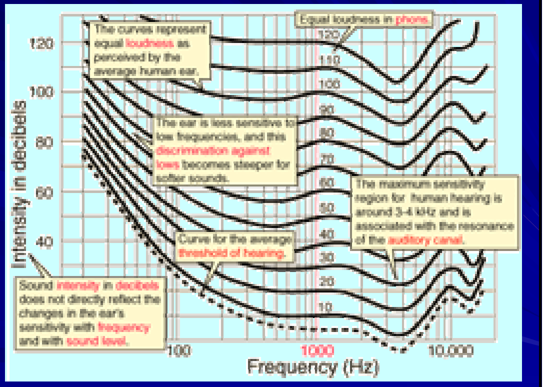
Perception
Psychoacoustics

- Psychoacoustics: a branch of science dealing with hearing, the sensations produced by sounds.
- A basic distinction must be made between the perceptual attributes of a sound vs measurable physical quantities:
- Many physical quantities are perceived on a logarithmic scale (e.g. loudness). Our perception is often a nonlinear function of the absolute value of the physical quantity being measured (e.g. equal loudness).
- Timbre can be used to describe why musical instruments sound different.
- What factors contribute to speaker identity?

Physical Quantity	Perceptual Quality
Intensity	Loudness
Fundamental Frequency	Pitch
Spectral Shape	Timbre
Onset/Offset Time	Timing
Phase Difference (Binaural Hearing)	Location

Perception
Equal Loudness

- **Just Noticeable Difference (JND)**: The acoustic value at which 75% of responses judge stimuli to be different (limen)
- The perceptual loudness of a sound is specified via its relative intensity above the threshold. A sound's loudness is often defined in terms of how intense a reference 1 kHz tone must be heard to sound as loud.



Perception
Non-Linear Frequency Warping: Bark and Mel Scale

- **Critical Bandwidths**: correspond to approximately 1.5 mm spacings along the basilar membrane, suggesting a set of 24 bandpass filters.
- **Critical Band**: can be related to a bandpass filter whose frequency response corresponds to the tuning curves of auditory neurons. A frequency range over which two sounds will sound like they are fusing into one.
- **Bark Scale**:

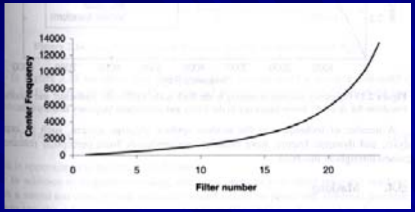
$$Bark = 13\,\mathrm{atan}\left(\frac{0.76f}{1000}\right) + 3.5\,\mathrm{atan}\left(\frac{f^2}{(7500)^2}\right)$$

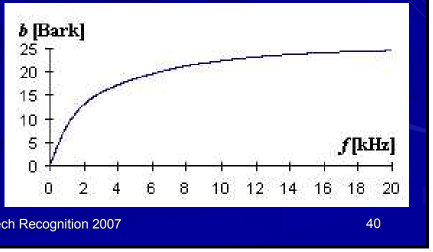
- **Mel Scale**:

$$mel\ frequency = 2595\ \log 10\,(1 + f/700.0)$$

Perception
Bark and Mel Scale

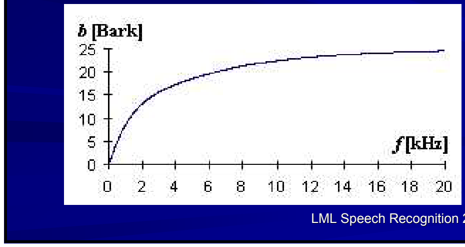
- **The Bark scale implies a nonlinear frequency mapping**

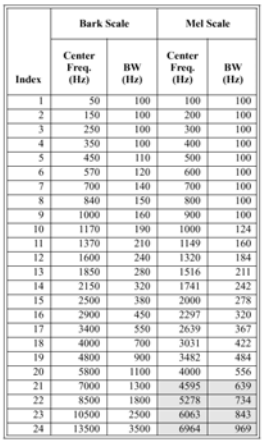




Perception
Bark and Mel Scale

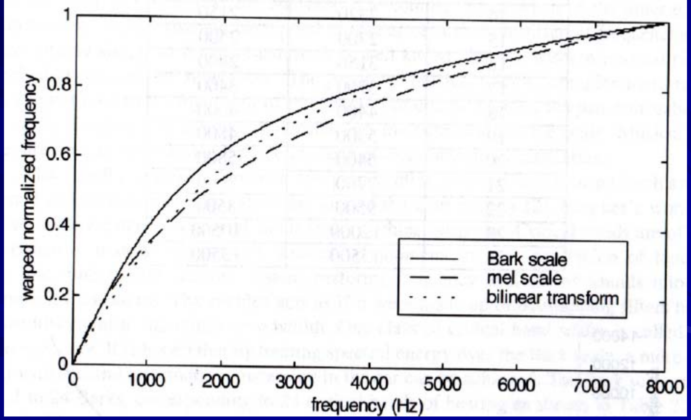
- **Filter Banks used in ASR:**
- **The Bark scale implies a nonlinear frequency mapping**





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

Comparison of
Bark and Mel Space Scales



Perception
Tone-Masking Noise

- **Frequency masking**: one sound cannot be perceived if another sound close in frequency has a high enough level. The first sound *masks* the second.
- **Tone-masking noise**: noise with energy EN (dB) at Bark frequency *g* masks a tone at Bark frequency *b* if the tone's energy is below the threshold:

  TT(*b*) = EN - 6.025 - 0.275*g* + Sm(*b-g*) (dB SPL)

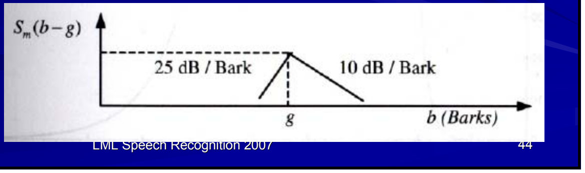
  where the *spread-of-masking* function Sm(*b*) is given by:

  Sm(*b*) = 15.81 + 7.5(*b*+0.474)-17.5*
  sqrt(1 + (*b*+0.474)2) (dB)
- **Temporal Masking**: onsets of sounds are masked in the time domain through a similar masking process.
- Thresholds are frequency and energy dependent.
- Thresholds depend on the nature of the sound as well.

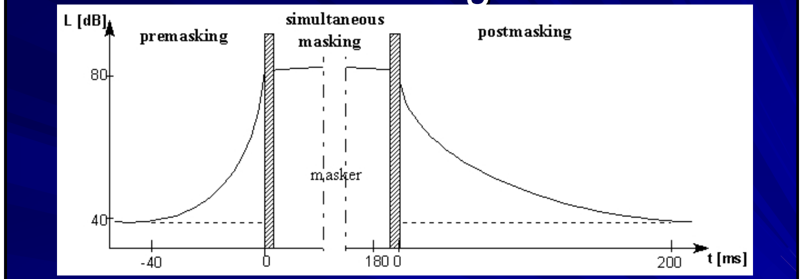
Perception
Noise-Masking Tone

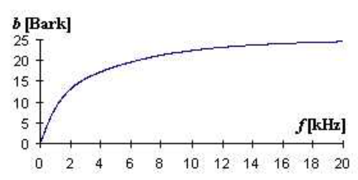
- **Noise-masking tone**: a tone at Bark frequency *g* with energy ET (dB) masks noise at Bark frequency *b* if the noise energy is below the threshold:

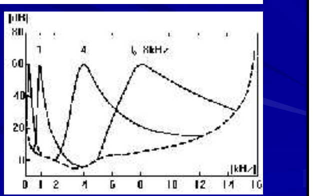
  TN(*b*) = ET - 2.025 - 0.17*g* + Sm(*b-g*) (dB SPL)
- Masking thresholds are commonly referred to as Bark scale functions of *just noticeable differences* (JND).
- Thresholds are not symmetric.
- Thresholds depend on the nature of the noise and the sound.



Masking

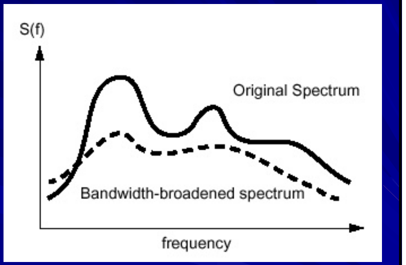






Perceptual Noise Weighting

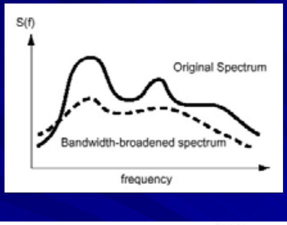
- **Noise-weighting**: shaping the spectrum to hide noise introduced by imperfect analysis and modeling techniques (essential in speech coding).
- Humans are sensitive to noise introduced in low-energy areas of the spectrum.
- Humans tolerate more additive noise when it falls under high energy areas of the spectrum. The amount of noise tolerated is greater if it is spectrally shaped to match perception.
- We can simulate this phenomena using "bandwidth-broadening":

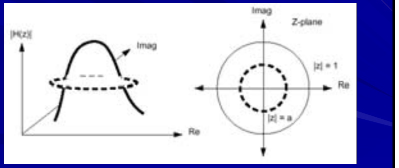


Perceptual Noise Weighting

Simple Z-Transform interpretation:

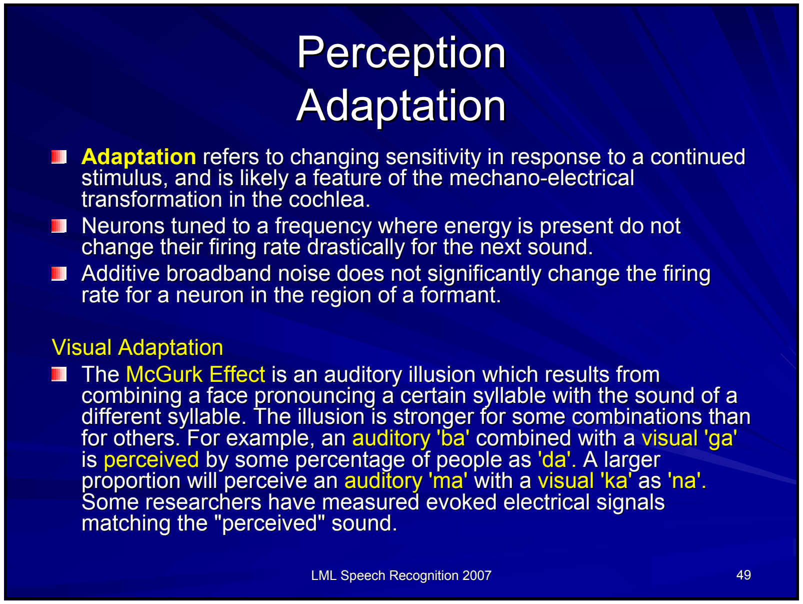
- can be implemented by evaluating the Z-Transform around a contour closer to the origin in the z-plane:
  Hnw(z) = H(*a*z).
- Used in many speech compression systems (Code Excited Linear Prediction).
- Analysis performed on bandwidth-broadened speech; synthesis performed using normal speech. Effectively shapes noise to fall under the formants.





Perception
Echo and Delay

- Humans are used to hearing their voice while they speak - real-time feedback (side tone).
- When we place headphones over our ears, which dampens this feedback, we tend to speak louder.
- **Lombard Effect**: Humans speak louder in the presence of ambient noise.
- When this side-tone is delayed, it interrupts our cognitive processes, and degrades our speech.
- This effect begins at delays of approximately 250 ms.
- Modern telephony systems have been designed to maintain delays lower than this value (long distance phone calls routed over satellites).
- Digital speech processing systems can introduce large amounts of delay due to non-real-time processing.

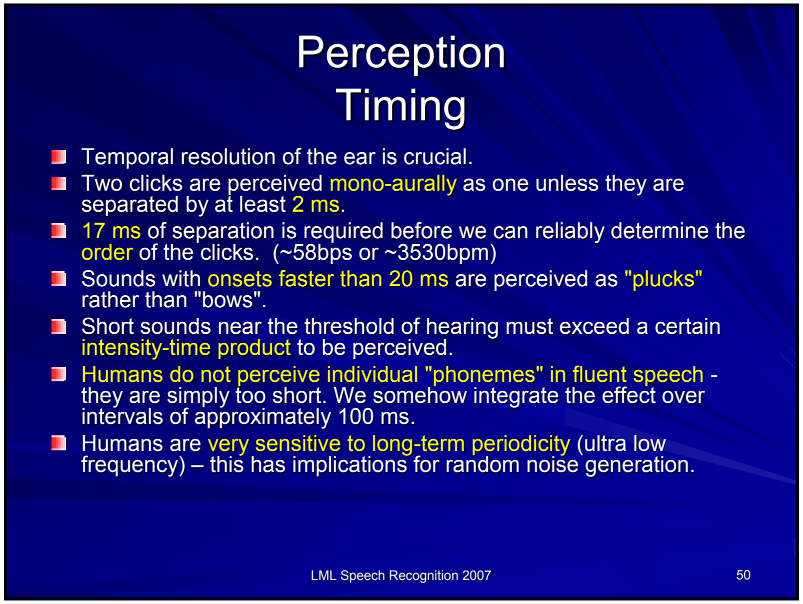


Perception Adaptation

- **Adaptation** refers to changing sensitivity in response to a continued stimulus, and is likely a feature of the mechano-electrical transformation in the cochlea.
- Neurons tuned to a frequency where energy is present do not change their firing rate drastically for the next sound.
- Additive broadband noise does not significantly change the firing rate for a neuron in the region of a formant.

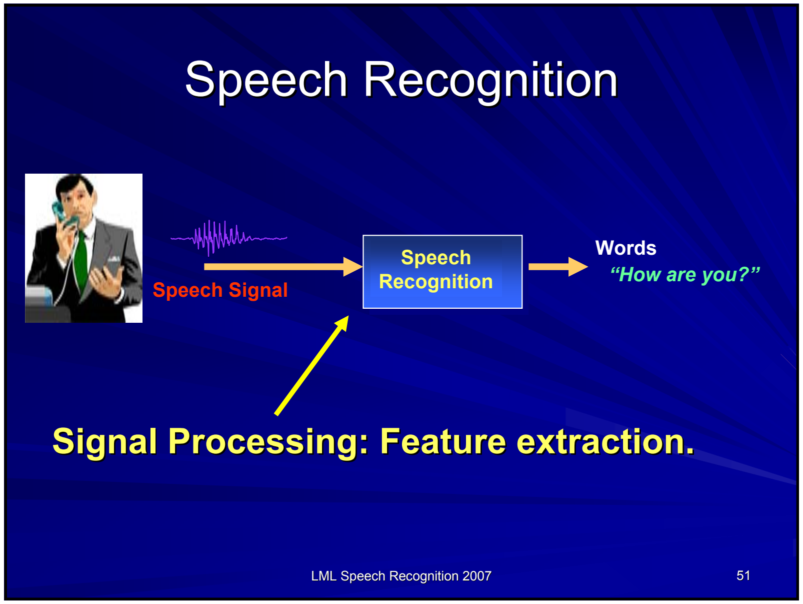
Visual Adaptation

- The McGurk Effect is an auditory illusion which results from combining a face pronouncing a certain syllable with the sound of a different syllable. The illusion is stronger for some combinations than for others. For example, an auditory 'ba' combined with a visual 'ga' is perceived by some percentage of people as 'da'. A larger proportion will perceive an auditory 'ma' with a visual 'ka' as 'na'. Some researchers have measured evoked electrical signals matching the "perceived" sound.



Perception Timing

- Temporal resolution of the ear is crucial.
- Two clicks are perceived mono-aurally as one unless they are separated by at least 2 ms.
- 17 ms of separation is required before we can reliably determine the order of the clicks. (~58bps or ~3530bpm)
- Sounds with onsets faster than 20 ms are perceived as "plucks" rather than "bows".
- Short sounds near the threshold of hearing must exceed a certain intensity-time product to be perceived.
- Humans do not perceive individual "phonemes" in fluent speech - they are simply too short. We somehow integrate the effect over intervals of approximately 100 ms.
- Humans are very sensitive to long-term periodicity (ultra low frequency) – this has implications for random noise generation.



Speech Recognition

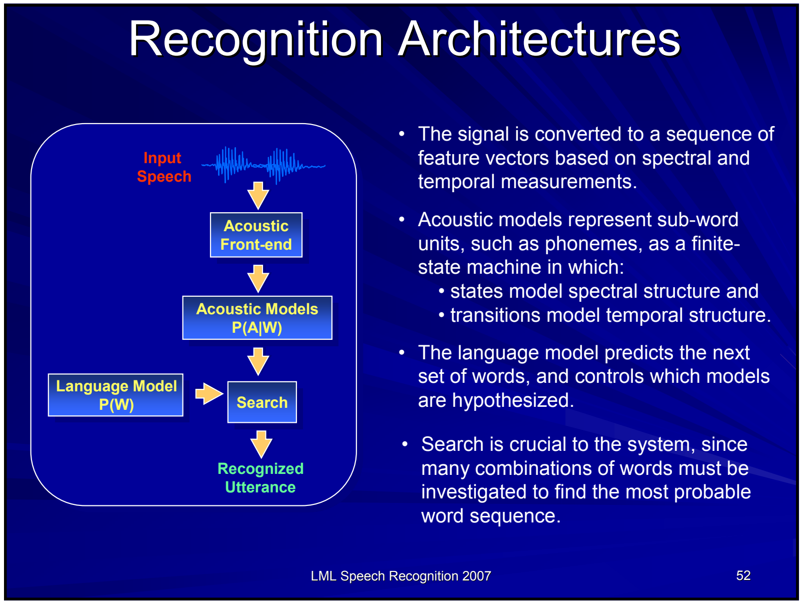
Speech Signal

Speech Recognition

Words

"How are you?"

Signal Processing: Feature extraction.



Recognition Architectures

Input Speech

Acoustic Front-end

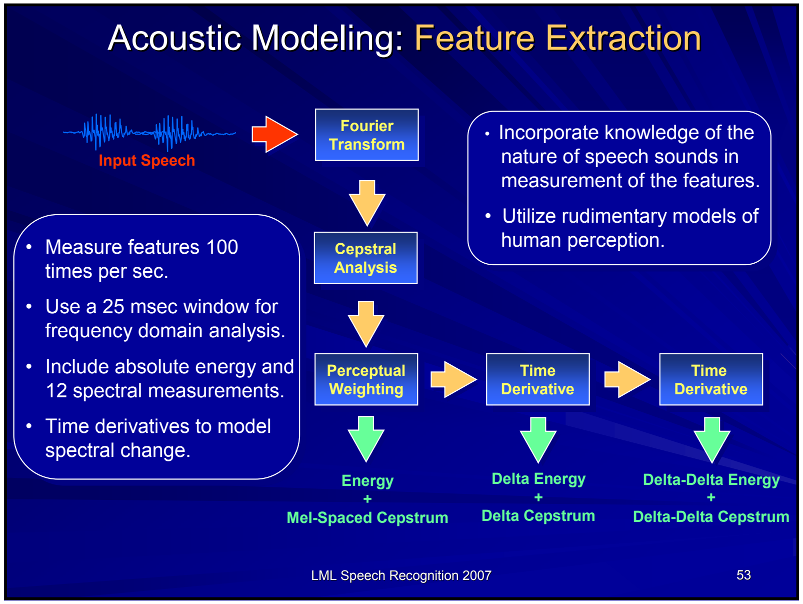
Acoustic Models P(A|W)

Language Model P(W)

Search

Recognized Utterance

- The signal is converted to a sequence of feature vectors based on spectral and temporal measurements.
- Acoustic models represent sub-word units, such as phonemes, as a finite-state machine in which:
  - states model spectral structure and
  - transitions model temporal structure.
- The language model predicts the next set of words, and controls which models are hypothesized.
- Search is crucial to the system, since many combinations of words must be investigated to find the most probable word sequence.



Acoustic Modeling: Feature Extraction

Input Speech → Fourier Transform → Cepstral Analysis → Perceptual Weighting → Time Derivative → Time Derivative

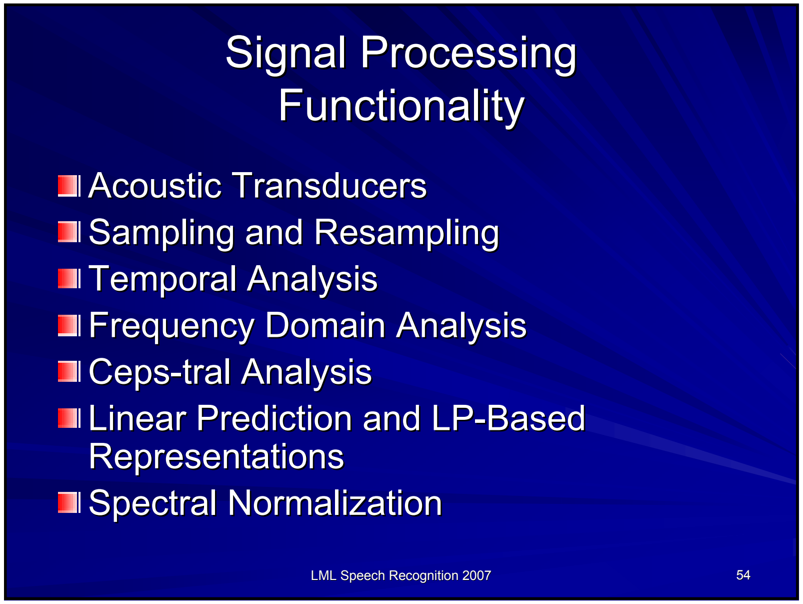
- Incorporate knowledge of the nature of speech sounds in measurement of the features.
- Utilize rudimentary models of human perception.

- Measure features 100 times per sec.
- Use a 25 msec window for frequency domain analysis.
- Include absolute energy and 12 spectral measurements.
- Time derivatives to model spectral change.

Energy + Mel-Spaced Cepstrum

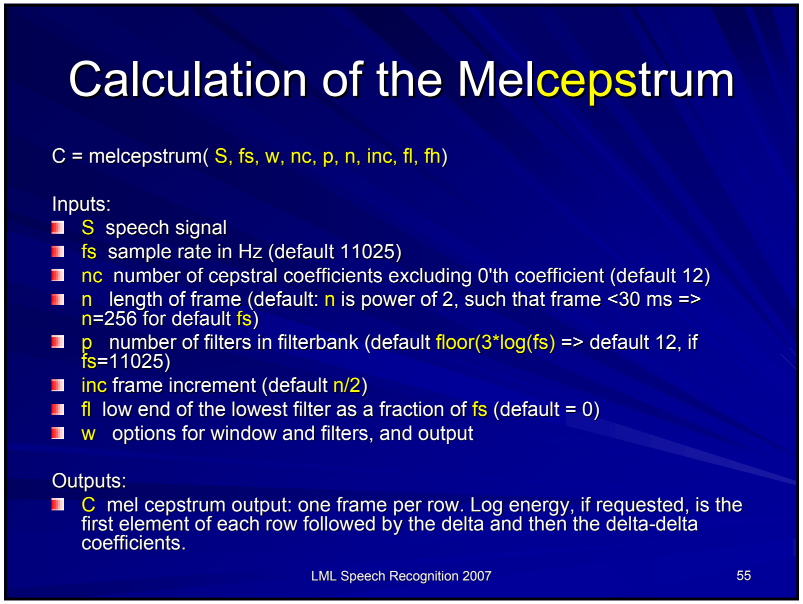
Delta Energy + Delta Cepstrum

Delta-Delta Energy + Delta-Delta Cepstrum



Signal Processing Functionality

- Acoustic Transducers
- Sampling and Resampling
- Temporal Analysis
- Frequency Domain Analysis
- Ceps-tral Analysis
- Linear Prediction and LP-Based Representations
- Spectral Normalization



Calculation of the Melcepstrum

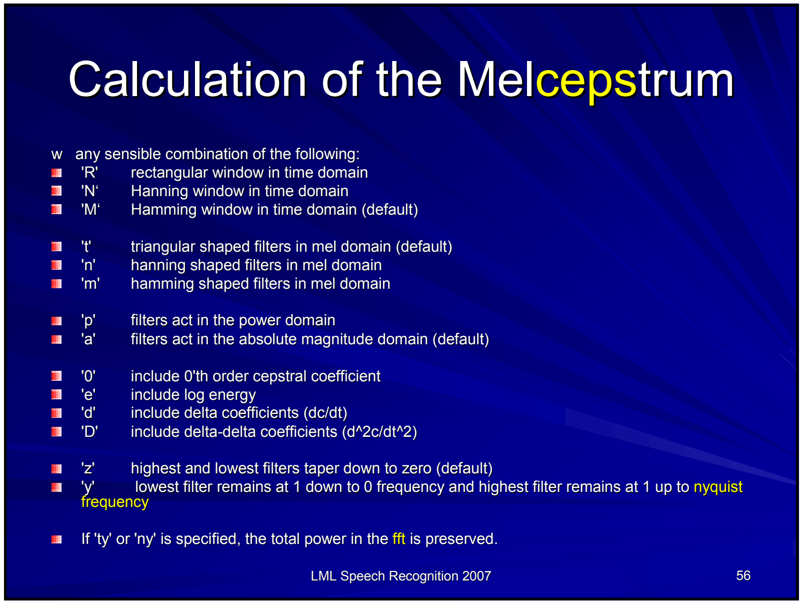
C = melcepstrum(S, fs, w, nc, p, n, inc, fl, fh)

Inputs:

- S speech signal
- fs sample rate in Hz (default 11025)
- nc number of cepstral coefficients excluding 0'th coefficient (default 12)
- n length of frame (default: n is power of 2, such that frame <30 ms => n=256 for default fs)
- p number of filters in filterbank (default floor(3*log(fs)) => default 12, if fs=11025)
- inc frame increment (default n/2)
- fl low end of the lowest filter as a fraction of fs (default = 0)
- w options for window and filters, and output

Outputs:

- C mel cepstrum output: one frame per row. Log energy, if requested, is the first element of each row followed by the delta and then the delta-delta coefficients.



Calculation of the Melcepstrum

w any sensible combination of the following:

- 'R' rectangular window in time domain
- 'N' Hanning window in time domain
- 'M' Hamming window in time domain (default)

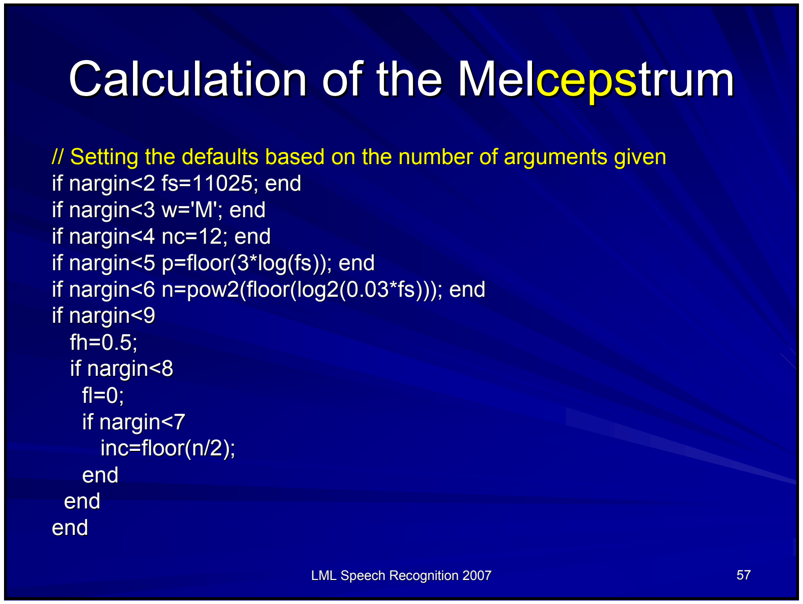
- 't' triangular shaped filters in mel domain (default)
- 'n' hanning shaped filters in mel domain
- 'm' hamming shaped filters in mel domain

- 'p' filters act in the power domain
- 'a' filters act in the absolute magnitude domain (default)

- '0' include 0'th order cepstral coefficient
- 'e' include log energy
- 'd' include delta coefficients (dc/dt)
- 'D' include delta-delta coefficients (d^2c/dt^2)

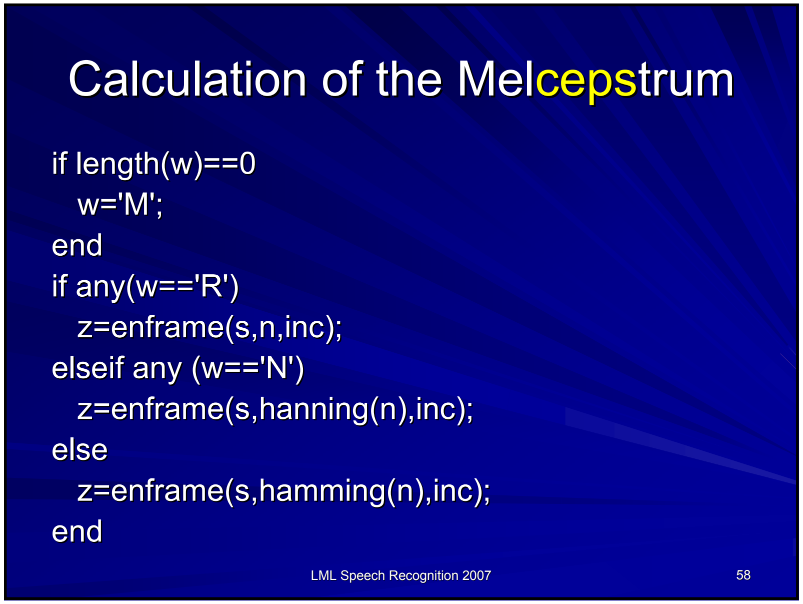
- 'z' highest and lowest filters taper down to zero (default)
- 'y' lowest filter remains at 1 down to 0 frequency and highest filter remains at 1 up to nyquist frequency

- If 'ty' or 'ny' is specified, the total power in the fft is preserved.



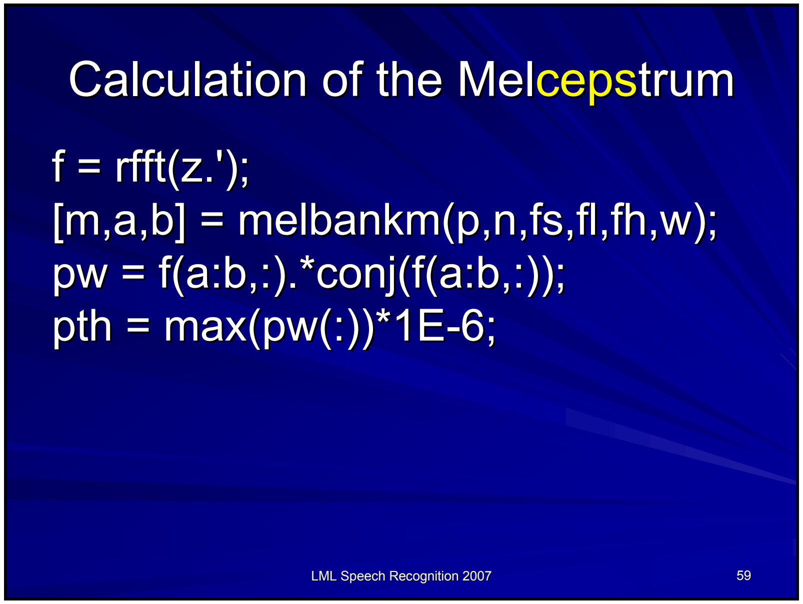
Calculation of the Melcepstrum

```
// Setting the defaults based on the number of arguments given
if nargin<2 fs=11025; end
if nargin<3 w='M'; end
if nargin<4 nc=12; end
if nargin<5 p=floor(3*log(fs)); end
if nargin<6 n=pow2(floor(log2(0.03*fs))); end
if nargin<9
   fh=0.5;
   if nargin<8
     fl=0;
     if nargin<7
        inc=floor(n/2);
     end
   end
end
```



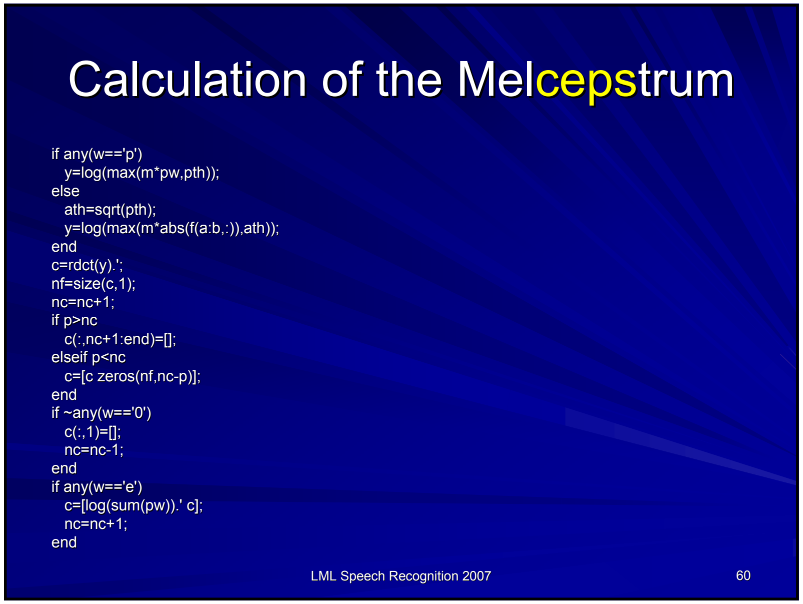
Calculation of the Melcepstrum

```
if length(w)==0
   w='M';
end
if any(w=='R')
   z=enframe(s,n,inc);
elseif any (w=='N')
   z=enframe(s,hanning(n),inc);
else
   z=enframe(s,hamming(n),inc);
end
```



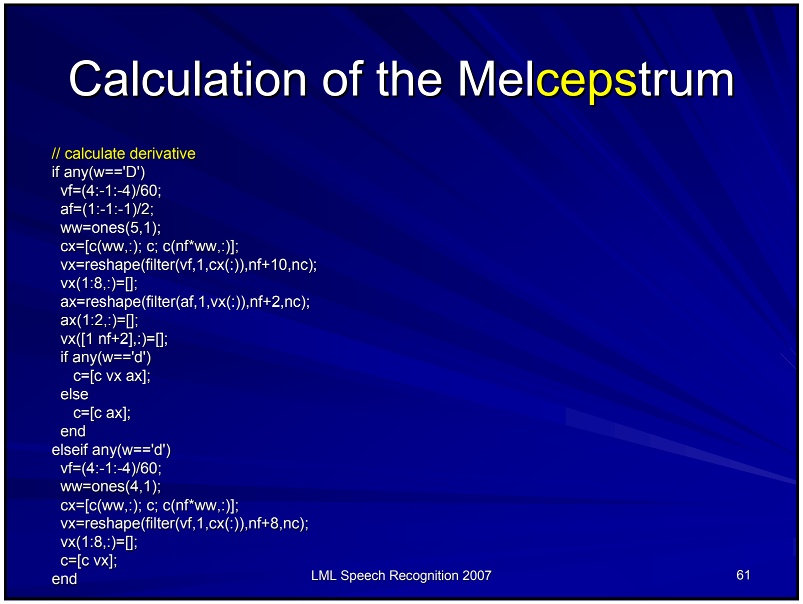
Calculation of the Melcepstrum

```
f = rfft(z.');
[m,a,b] = melbankm(p,n,fs,fl,fh,w);
pw = f(a:b,:).*conj(f(a:b,:));
pth = max(pw(:))*1E-6;
```



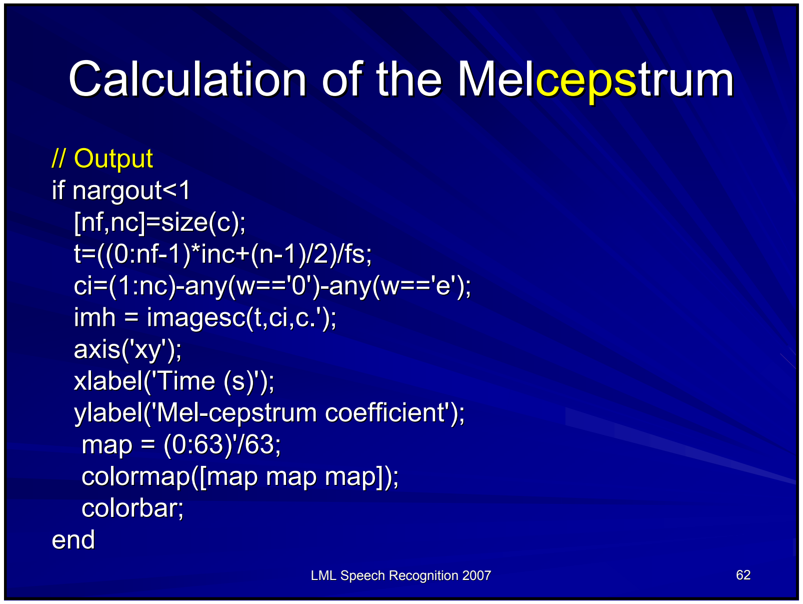
Calculation of the Melcepstrum

```
if any(w=='p')
   y=log(max(m*pw,pth));
else
   ath=sqrt(pth);
   y=log(max(m*abs(f(a:b,:)),ath));
end
c=rdct(y).';
nf=size(c,1);
nc=nc+1;
if p>nc
   c(:,nc+1:end)=[];
elseif p<nc
   c=[c zeros(nf,nc-p)];
end
if ~any(w=='0')
   c(:,1)=[];
   nc=nc-1;
end
if any(w=='e')
   c=[log(sum(pw)).' c];
   nc=nc+1;
end
```



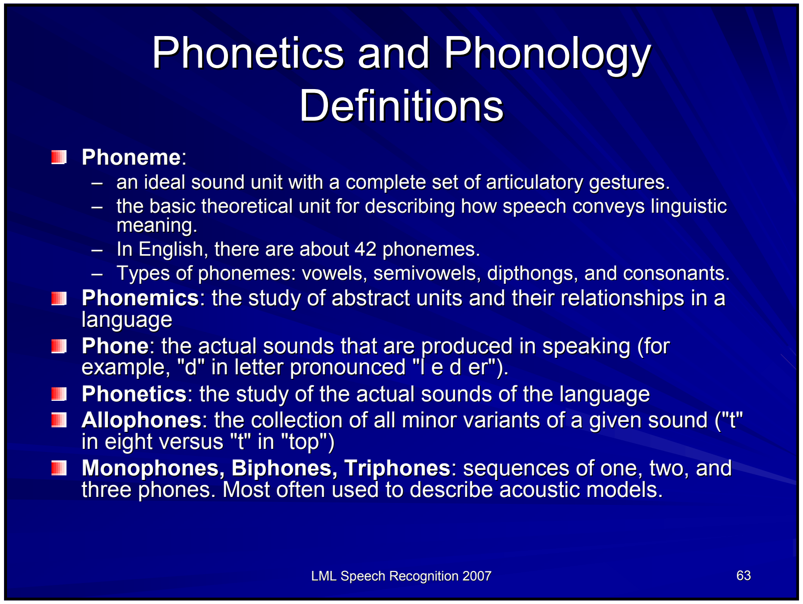
Calculation of the Melcepstrum

```
// calculate derivative
if any(w=='D')
  vf=(4:-1:-4)/60;
  af=(1:-1:-1)/2;
  ww=ones(5,1);
  cx=[c(ww,:); c; c(nf*ww,:)];
  vx=reshape(filter(vf,1,cx(:)),nf+10,nc);
  vx(1:8,:)=[];
  ax=reshape(filter(af,1,vx(:)),nf+2,nc);
  ax(1:2,:)=[];
  vx([1 nf+2],:)=[];
  if any(w=='d')
     c=[c vx ax];
  else
     c=[c ax];
  end
elseif any(w=='d')
  vf=(4:-1:-4)/60;
  ww=ones(4,1);
  cx=[c(ww,:); c; c(nf*ww,:)];
  vx=reshape(filter(vf,1,cx(:)),nf+8,nc);
  vx(1:8,:)=[];
  c=[c vx];
end
```



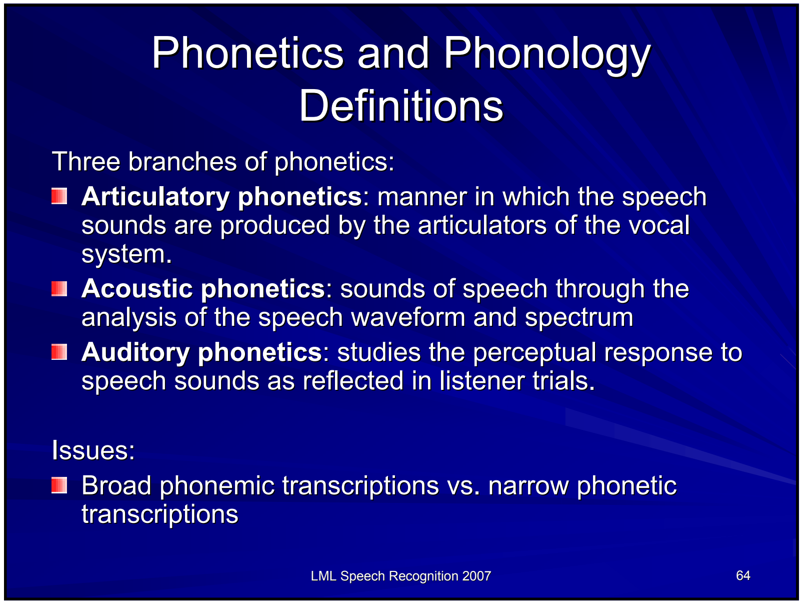
Calculation of the Melcepstrum

```
// Output
if nargout<1
  [nf,nc]=size(c);
  t=((0:nf-1)*inc+(n-1)/2)/fs;
  ci=(1:nc)-any(w=='0')-any(w=='e');
  imh = imagesc(t,ci,c.');
  axis('xy');
  xlabel('Time (s)');
  ylabel('Mel-cepstrum coefficient');
   map = (0:63)'/63;
   colormap([map map map]);
   colorbar;
end
```



Phonetics and Phonology Definitions

- **Phoneme**:
  - an ideal sound unit with a complete set of articulatory gestures.
  - the basic theoretical unit for describing how speech conveys linguistic meaning.
  - In English, there are about 42 phonemes.
  - Types of phonemes: vowels, semivowels, dipthongs, and consonants.
- **Phonemics**: the study of abstract units and their relationships in a language
- **Phone**: the actual sounds that are produced in speaking (for example, "d" in letter pronounced "l e d er").
- **Phonetics**: the study of the actual sounds of the language
- **Allophones**: the collection of all minor variants of a given sound ("t" in eight versus "t" in "top")
- **Monophones, Biphones, Triphones**: sequences of one, two, and three phones. Most often used to describe acoustic models.



Phonetics and Phonology Definitions

Three branches of phonetics:

- **Articulatory phonetics**: manner in which the speech sounds are produced by the articulators of the vocal system.
- **Acoustic phonetics**: sounds of speech through the analysis of the speech waveform and spectrum
- **Auditory phonetics**: studies the perceptual response to speech sounds as reflected in listener trials.

Issues:

- Broad phonemic transcriptions vs. narrow phonetic transcriptions

English Phonemes

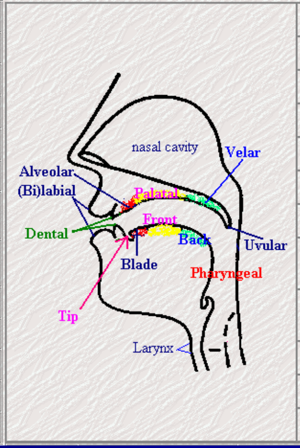
Vowels and Diphthongs		
Phonemes	**Word Examples**	**Description**
iy	**fee**l, **e**v**e**, m**e**	front close unrounded
ih	f**i**ll, h**i**t, l**i**d	front close unrounded (lax)
ae	**a**t, c**a**rry, g**a**s	front open unrounded (tense)
aa	f**a**ther, **ah**, c**ar**	back open rounded
ah	c**u**t, b**u**d, **u**p	open mid-back rounded
ao	d**o**g, l**aw**n, c**au**ght	open-mid back round
ay	t**ie**, **i**ce, b**i**te	diphthong with quality: aa + ih
ax	ag**o**, comply	central close mid (schwa)
ey	**a**te, d**ay**, t**a**pe	front close-mid unrounded (tense)
eh	p**e**t, b**e**rry, t**e**n	front open-mid unrounded
er	t**ur**n, f**ur**, met**er**	central open-mid unrounded
ow	g**o**, **ow**n, t**ow**n	back close-mid rounded
aw	f**ou**l, h**ow**, **our**	diphthong with quality: aa + uh
oy	t**oy**, c**oi**n, **oi**l	diphthong with quality: ao + ih
uh	b**oo**k, p**u**ll, g**oo**d	back close-mid unrounded (lax)
uw	t**oo**l, cr**ew**, m**oo**	back close round

English Phonemes

Consonants and Liquids		
Phonemes	**Word Examples**	**Description**
b	**b**ig, a**b**le, ta**b**	voiced bilabial plosive
p	**p**ut, o**p**en, ta**p**	voiceless bilabial plosive
d	**d**ig, i**d**ea, wa**d**	voiced alveolar plosive
t	**t**alk, sa**t**	voiceless alveolar plosive
g	**g**ut, an**g**le, ta**g**	voiced velar plosive
t	me**t**er	alveolar flap
g	**g**ut, an**g**le, ta**g**	voiced velar plosive
k	**c**ut, **k**en, ta**k**e	voiceless velar plosive
f	**f**ork, a**f**ter, i**f**	voiceless labiodental fricative
v	**v**at, o**v**er, ha**v**e	voiced labiodental fricative
s	**s**it, ca**s**t, to**ss**	voiceless alveolar fricative
z	**z**ap, la**z**y, ha**z**e	voiced alveolar fricative

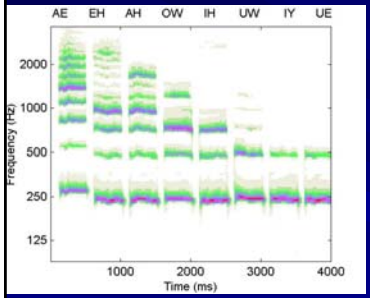
English Phonemes

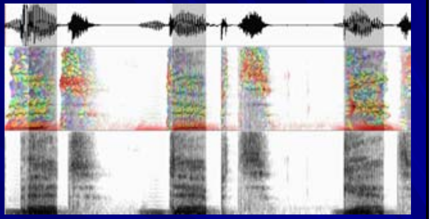
/f/	Labio-dental
/v/	Labio-dental
/θ/	Tip-dental
/ð/	Tip-dental
/s/	Blade-alveolar
/z/	Blade-alveolar
/ʃ/	Blade/front –palato-alveolar
/ʒ/	Blade/front –palato-alveolar
/h/	Glottal
/l/	Tip-alveolar
/r/	Blade-postalveolar
/w/	Bilabial back-velar
/j/	Front-palatal



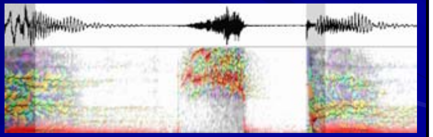
Consonants	Place
/p/	Bilabial
/b/	Bilabial
/t/	Tip-alveolar
/d/	Tip-alveolar
/k/	Back-velar
/g/	Back-velar
/tʃ/	Blade/front –palato-alveolar
/dʒ/	Blade/front –palato-alveolar
/m/	Bilabial
/n/	Tip-alveolar
/ŋ/	Back-velar

English Phonemes





Bet Debt Get



Pin Sp i n

Allophone

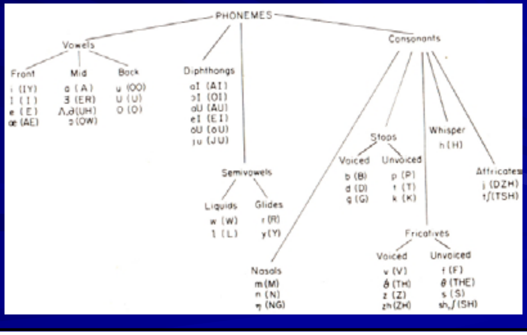
Transcription

Major governing bodies for phonetic alphabets:

- **International Phonetic Alphabet (IPA)**: over 100 years of history
- **ARPAbet**: developed in the late 1970's to support ARPA research
- **TIMIT**: TI/MIT variant of ARPAbet used for the TIMIT corpus
- **Worldbet**: developed by Hieronymous (AT&T) to deal with multiple languages within a single ASCII system
- **Unicode**: character encoding system that includes IPA phonetic symbols.

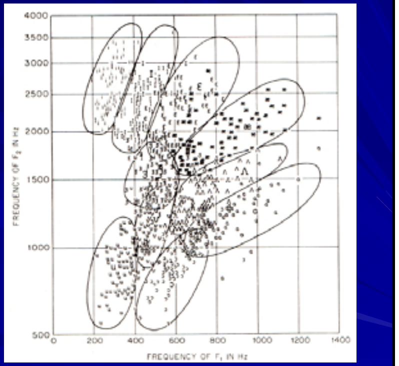
Phonetics
The Vowel Space

- Each fundamental speech sound can be categorized according to the position of the articulators. (Acoustic Phonetics.)



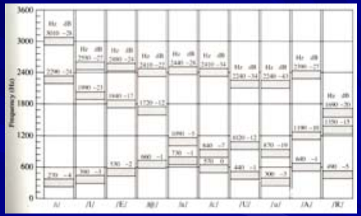
The Vowel Space

- We can characterize a vowel sound by the locations of the first and second spectral resonances, known as formant frequencies:
- Some voiced sounds, such as diphthongs, are transitional sounds that move from one vowel location to another.

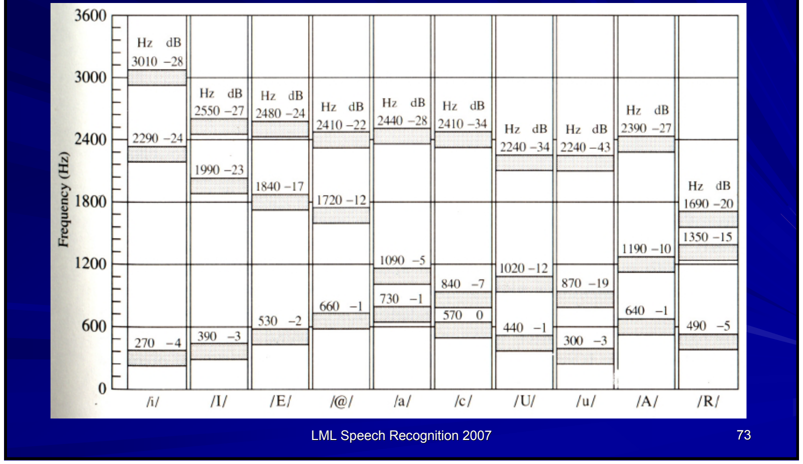


Phonetics
The Vowel Space

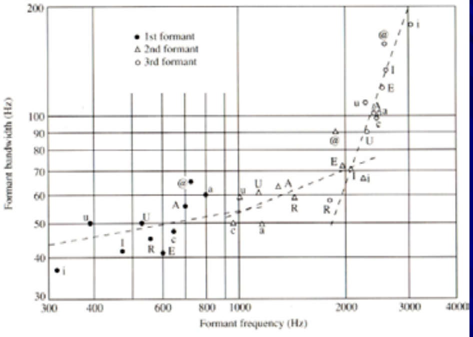
- Some voiced sounds, such as diphthongs, are transitional sounds that move from one vowel location to another.



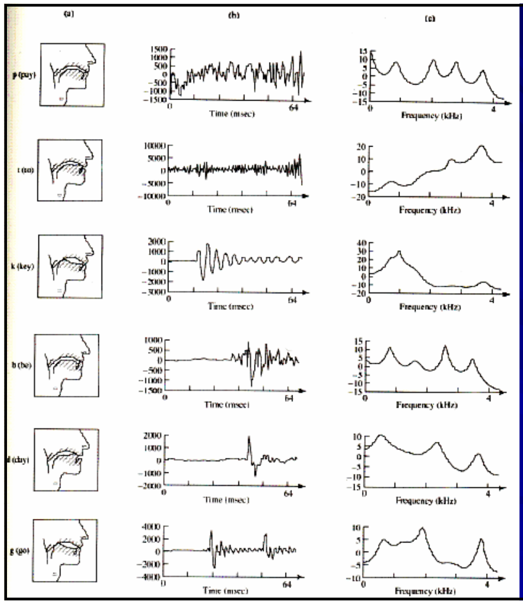
Phonetics Formant Frequency Ranges



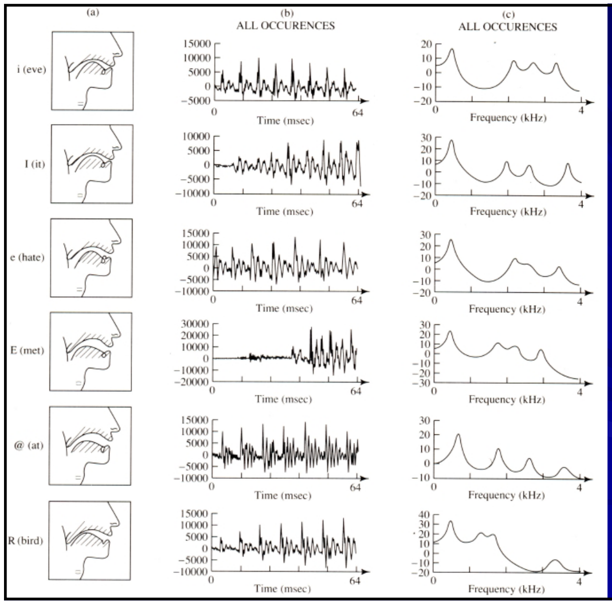
Bandwidth and Formant Frequencies



Vowels	F_1 Avg.	F_1 Extremes		F_2 Avg.	F_2 Extremes		F_3 Avg.	F_3 Extremes	
i	38	30	80	66	30	120	171	60	300
I	42	30	100	71	40	120	142	60	300
E	42	30	120	72	30	140	126	50	300
@	65	30	140	90	40	200	156	50	300
a	60	30	160	50	30	80	102	40	300
c	47	30	120	50	30	200	98	40	240
u	50	30	120	58	30	200	107	50	200
U	51	30	100	61	30	140	90	40	200
A	56	30	140	63	30	140	102	50	300
R	46	30	80	59	30	120	58	40	120
Avg.	49.7			64.0			115.2		



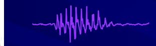
Acoustic Theory: Vowel Production



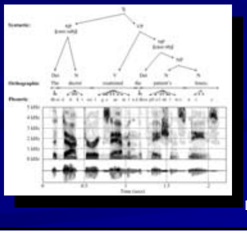
Acoustic Theory: Consonants

Speech Recognition
Syntax and Semantics

Goal: Automatically extract the string of words spoken from the speech signal



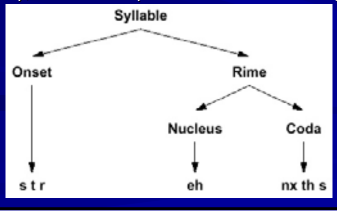
Speech Signal → **Speech Recognition** → Words *"How are you?"*

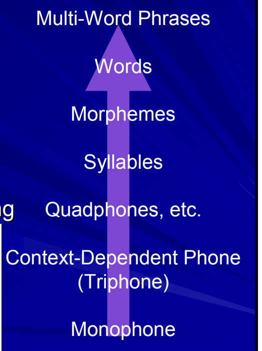


What LANGUAGE is spoken?

Syntax and Semantics
Syllables: Coarticulation

- Acoustically distinct.
- There are over 10,000 syllables in English.
- There is no universal definition of a syllable.
- Can be defined from both a production and perception viewpoint.
- Centered around vowels in English.
- Consonants often span two syllables ("ambisyllabic" - "bottle").
- Three basic parts: onset (initial consonants), nucleus (vowel), and coda (consonants following the nucleus).





Words

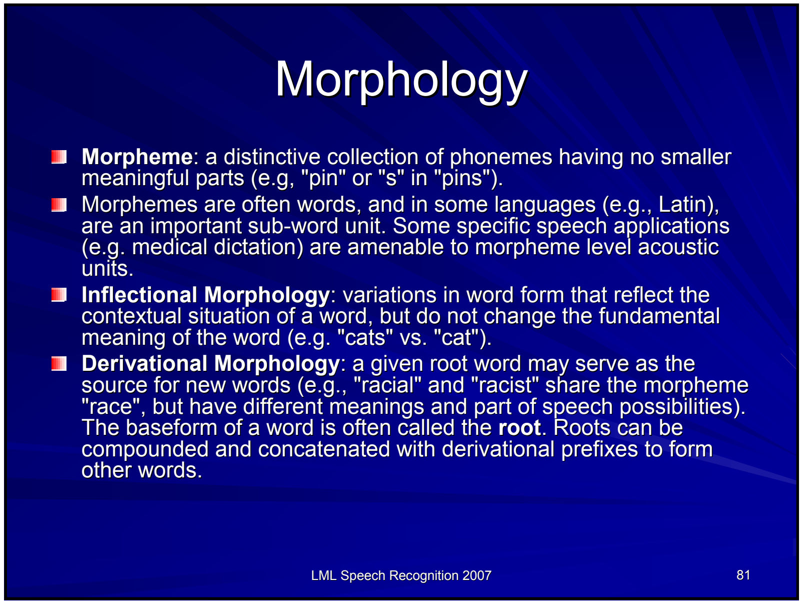
- Loosely defined as a lexical unit - there is an agreed upon meaning in a given community.
- In many languages (e.g., Indo-European), easily observed in the orthographic (writing) system since it is separated by white space.
- In spoken language, however, there is a segmentation problem: words run together.
- **Syntax**: certain facts about word structure and combinatorial possibilities are evident to most native speakers.
- **Paradigmatic**: properties related to meaning.
- **Syntagmatic**: properties related to constraints imposed by word combinations (grammar).
- Word-level constraints are the most common form of "domain knowledge" in a speech recognition system.
- **N-gram** models are the most common way to implement word-level constraints.
- N-gram distributions are very interesting!

Lexical Part of Speech

- **Lexicon**: alphabetic arrangement of words and their definitions.
- **Lexical Part of Speech**: A restricted inventory of word-type categories which capture generalizations of word forms and distributions
- **Part of Speech (POS)**: noun, verb, adjective, adverb, interjection, conjunction, determiner, preposition, and pronoun.
- **Proper Noun**: names such as "Velcro" or "Spandex".
- **Open POS Categories**:

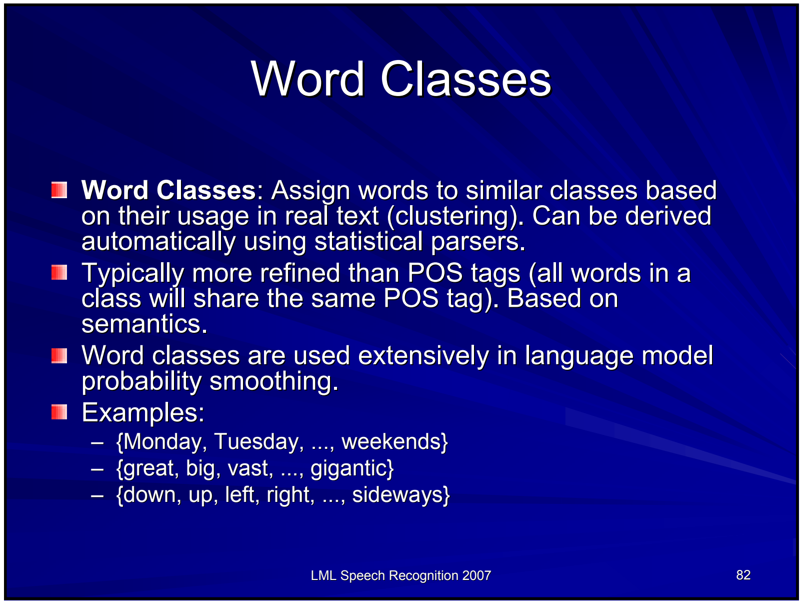
Tag	Description	Function	Example
N	Noun	Named entity	*cat*
V	Verb	Event or condition	*forget*
Adj	Adjective	Descriptive	*yellow*
Adv	Adverb	Manner of action	*quickly*
Interj	Interjection	Reaction	*Oh!*

- **Closed POS Categories**: some level of universal agreement on the categories
- **Lexical reference systems: Penn Treebank**, **Wordnet**



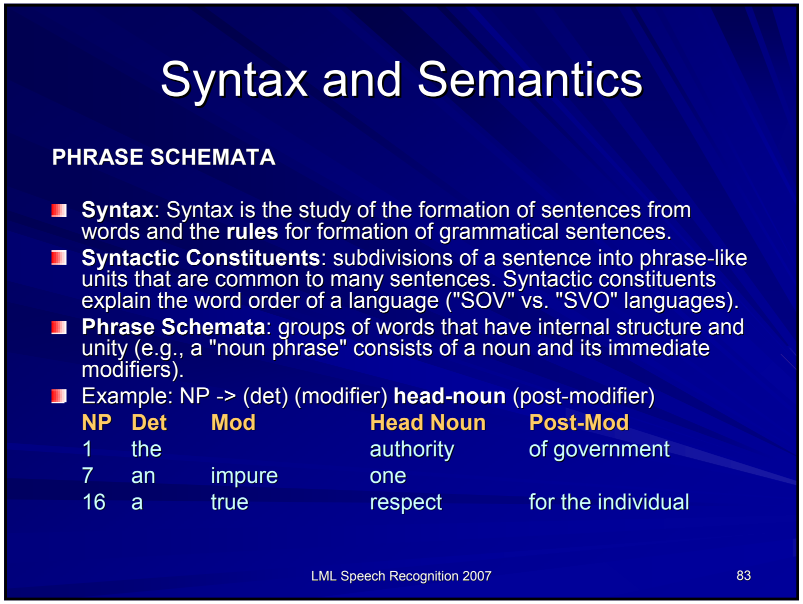
Morphology

- **Morpheme**: a distinctive collection of phonemes having no smaller meaningful parts (e.g, "pin" or "s" in "pins").
- Morphemes are often words, and in some languages (e.g., Latin), are an important sub-word unit. Some specific speech applications (e.g. medical dictation) are amenable to morpheme level acoustic units.
- **Inflectional Morphology**: variations in word form that reflect the contextual situation of a word, but do not change the fundamental meaning of the word (e.g. "cats" vs. "cat").
- **Derivational Morphology**: a given root word may serve as the source for new words (e.g., "racial" and "racist" share the morpheme "race", but have different meanings and part of speech possibilities). The baseform of a word is often called the **root**. Roots can be compounded and concatenated with derivational prefixes to form other words.



Word Classes

- **Word Classes**: Assign words to similar classes based on their usage in real text (clustering). Can be derived automatically using statistical parsers.
- Typically more refined than POS tags (all words in a class will share the same POS tag). Based on semantics.
- Word classes are used extensively in language model probability smoothing.
- Examples:
  - {Monday, Tuesday, ..., weekends}
  - {great, big, vast, ..., gigantic}
  - {down, up, left, right, ..., sideways}

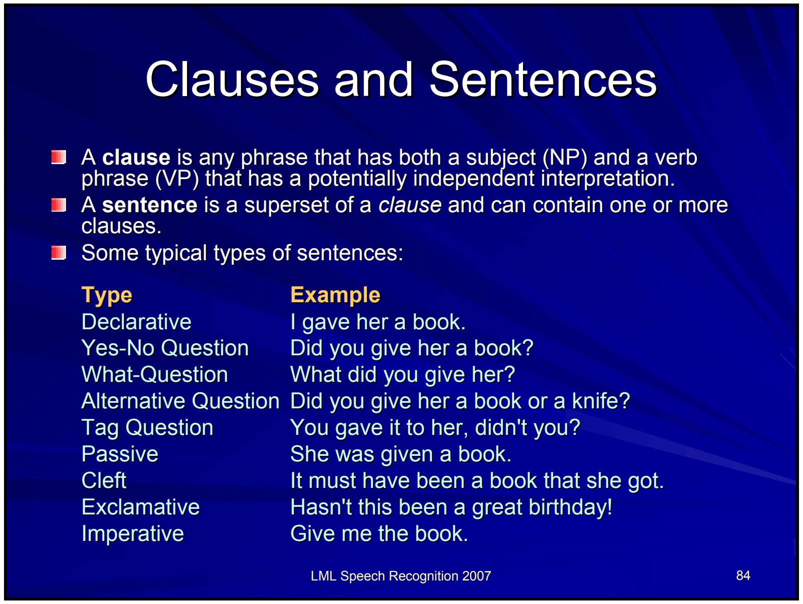


Syntax and Semantics

PHRASE SCHEMATA

- **Syntax**: Syntax is the study of the formation of sentences from words and the **rules** for formation of grammatical sentences.
- **Syntactic Constituents**: subdivisions of a sentence into phrase-like units that are common to many sentences. Syntactic constituents explain the word order of a language ("SOV" vs. "SVO" languages).
- **Phrase Schemata**: groups of words that have internal structure and unity (e.g., a "noun phrase" consists of a noun and its immediate modifiers).
- Example: NP -> (det) (modifier) **head-noun** (post-modifier)

NP	Det	Mod	Head Noun	Post-Mod
1	the		authority	of government
7	an	impure	one	
16	a	true	respect	for the individual



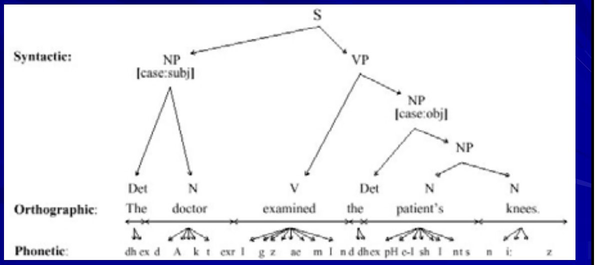
Clauses and Sentences

- A **clause** is any phrase that has both a subject (NP) and a verb phrase (VP) that has a potentially independent interpretation.
- A **sentence** is a superset of a *clause* and can contain one or more clauses.
- Some typical types of sentences:

Type	Example
Declarative	I gave her a book.
Yes-No Question	Did you give her a book?
What-Question	What did you give her?
Alternative Question	Did you give her a book or a knife?
Tag Question	You gave it to her, didn't you?
Passive	She was given a book.
Cleft	It must have been a book that she got.
Exclamative	Hasn't this been a great birthday!
Imperative	Give me the book.

Parse Tree

- **Parse Tree**: used to represent the structure of a sentence and the relationship between its constituents.
- Markup languages such as the standard generalized markup language (**SGML**) are often used to represent a parse tree in a textual form.
- Example:



Semantic Roles

- **Grammatical roles** are often used to describe the direction of action (e.g., subject, object, indirect object).
- **Semantic roles**, also known as **case relations**, are used to make sense of the participants in an event (e.g., "who did what to whom").
- Example: "The doctor examined the patient's knees“

Role	Description
Agent	cause or inhibitor of action
Patient/Theme	undergoer of the action
Instrument	how the action is accomplished
Goal	to whom the action is directed
Result	result or outcome of the action
Location	location or place of the action

Lexical Semantics

- **Lexical Semantics**: the semantic structure associated with a word, as represented in the lexicon.
- **Taxonomy**: orderly classification of words according to their presumed natural relationships.
- Examples:
  - **Is-A Taxonomy**: a *crow* is a bird.
  - **Has-a Taxonomy**: a *car* has a windshield.
  - **Action-Instrument**: a *knife* can cut.
- Words can appear in many relations and have multiple meanings and uses.

Lexical Semantics

- There are no universally-accepted taxonomies:

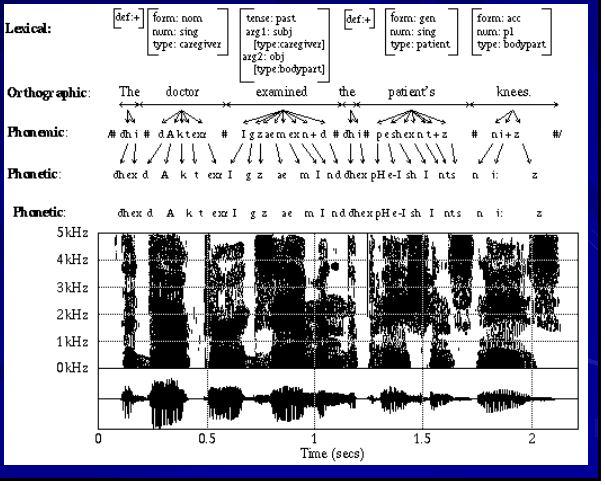
Family	Subtype	Example
Contrasts	Contrary	old-young
	Contradictory	alive-dead
	Reverse	buy-sell
	Directional	front-back
	Incompatible	happy-morbid
	Asymmetric	contrary hot-cool
	Attribute similar	rake-fork
Case Relations	Agent-action	artist-paint
	Agent-instrument	farmer-tractor
	Agent-object	baker-bread
	Action-recipient	sit-chair
	Action-instrument	cut-knife

Logical Form

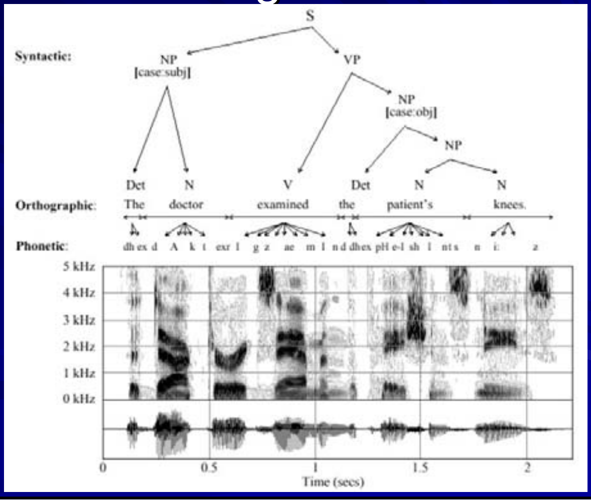
- **Logical form**: a metalanguage in which we can concretely and succinctly express all linguistically possible meanings of an utterance.
- Typically used as a representation to which we can apply discourse and world knowledge to select the single-best (or N-best) alternatives.
- An attempt to bring formal logic to bear on the language understanding problem (**predicate logic**).
- Example:
  - If Romeo is happy, Juliet is happy:
    Happy(Romeo) -> Happy(Juliet)
  - "The doctor examined the patient's knees"

Logical Form

- "The doctor examined the patient's knee"

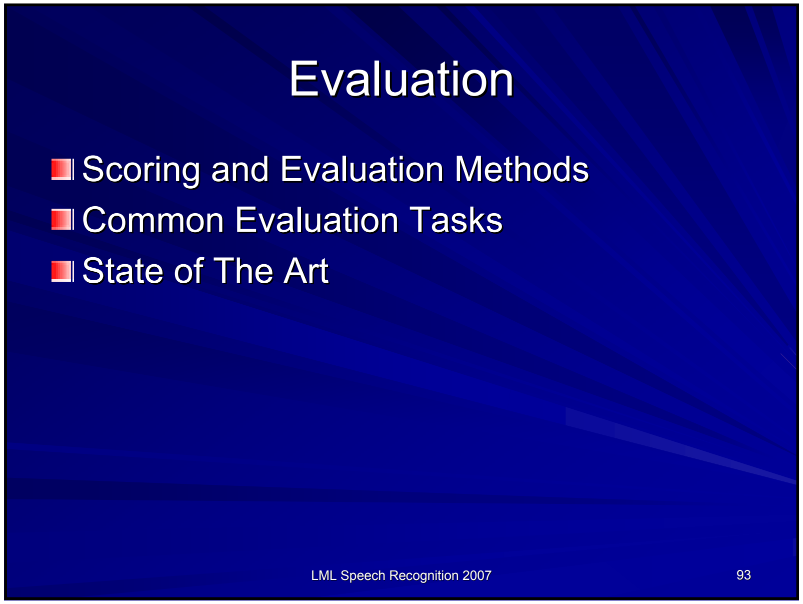


Integration



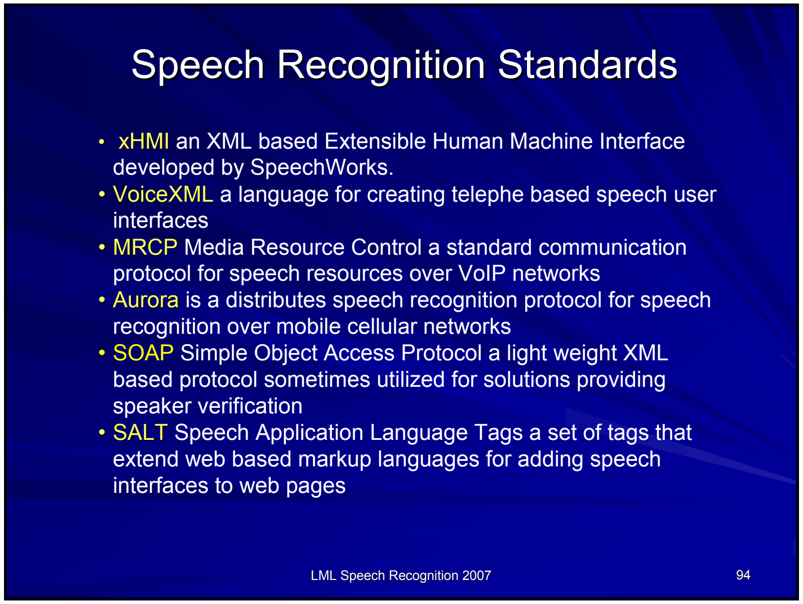
Parametrization

- Differentiation
- Principal Components
- Linear Discriminant Analysis



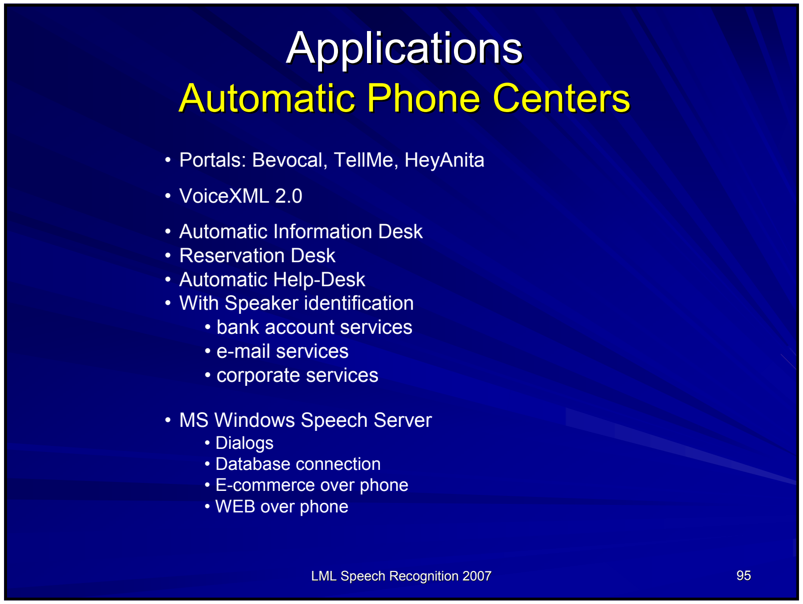
Evaluation

- Scoring and Evaluation Methods
- Common Evaluation Tasks
- State of The Art



Speech Recognition Standards

- xHMI an XML based Extensible Human Machine Interface developed by SpeechWorks.
- VoiceXML a language for creating telephe based speech user interfaces
- MRCP Media Resource Control a standard communication protocol for speech resources over VoIP networks
- Aurora is a distributes speech recognition protocol for speech recognition over mobile cellular networks
- SOAP Simple Object Access Protocol a light weight XML based protocol sometimes utilized for solutions providing speaker verification
- SALT Speech Application Language Tags a set of tags that extend web based markup languages for adding speech interfaces to web pages



Applications
Automatic Phone Centers

- Portals: Bevocal, TellMe, HeyAnita
- VoiceXML 2.0
- Automatic Information Desk
- Reservation Desk
- Automatic Help-Desk
- With Speaker identification
  - bank account services
  - e-mail services
  - corporate services
- MS Windows Speech Server
  - Dialogs
  - Database connection
  - E-commerce over phone
  - WEB over phone

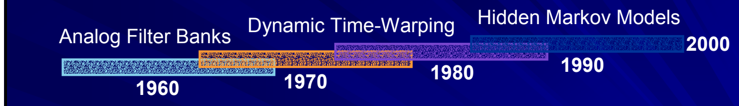
Applications
Real-Time Translation

- From President Clinton’s State of the Union address (January 27, 2000):

> “These kinds of innovations are also propelling our remarkable prosperity... Soon researchers will bring us devices that can translate foreign languages as fast as you can talk... molecular computers the size of a tear drop with the power of today’s fastest supercomputers.”

- Imagine a world where:
  - You book a travel reservation from your cellular phone while driving in your car without ever talking to a human (database query)
  - You converse with someone in a foreign country and neither speaker speaks a common language (universal translator)
  - You place a call to your bank to inquire about your bank account and never have to remember a password (transparent telephony)
  - You can ask questions by voice and your Internet browser returns answers to your questions (intelligent query)
- Human Language Engineering: a sophisticated integration of many speech and language related technologies... a science for the next millennium.

Technology: Future Directions



Conclusions:

- supervised training is a good machine learning technique
- large databases are essential for the development of robust statistics

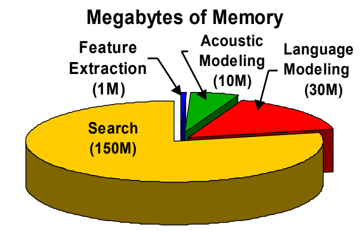
Challenges:

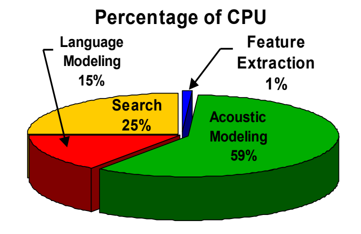
- discrimination vs. representation
- generalization vs. memorization
- pronunciation modeling
- human-centered language modeling

The algorithmic issues for the next decade:

- Better features by extracting articulatory information?
- Bayesian statistics? Bayesian networks?
- Decision Trees? Information-theoretic measures?
- Nonlinear dynamics? Chaos?

Implementation Issues
Search Is Resource Intensive


Megabytes of Memory

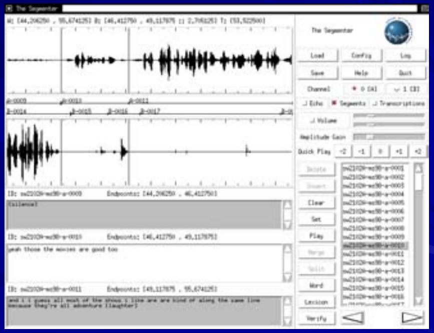

Percentage of CPU

- **Typical LVCSR systems have about 10M free parameters, which makes training a challenge.**
- **Large speech databases are required (several hundred hours of speech).**
- **Tying, smoothing, and interpolation are required.**

ASR: Conversational Speech

- **Conversational speech collected over the telephone contains background noise, music, fluctuations in the speech rate, laughter, partial words, hesitations, mouth noises, etc.**
- **WER (Word Error Rate) has decreased from 100% to 30% in six years.**

- **Laughter**
- **Singing**
- **Unintelligible**
- **Spoonerism**
- **Background Speech**
- **No pauses**
- **Restarts**
- **Vocalized Noise**
- **Coinage**



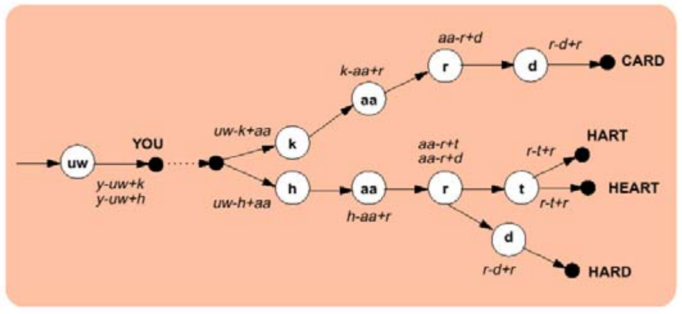
[silence]

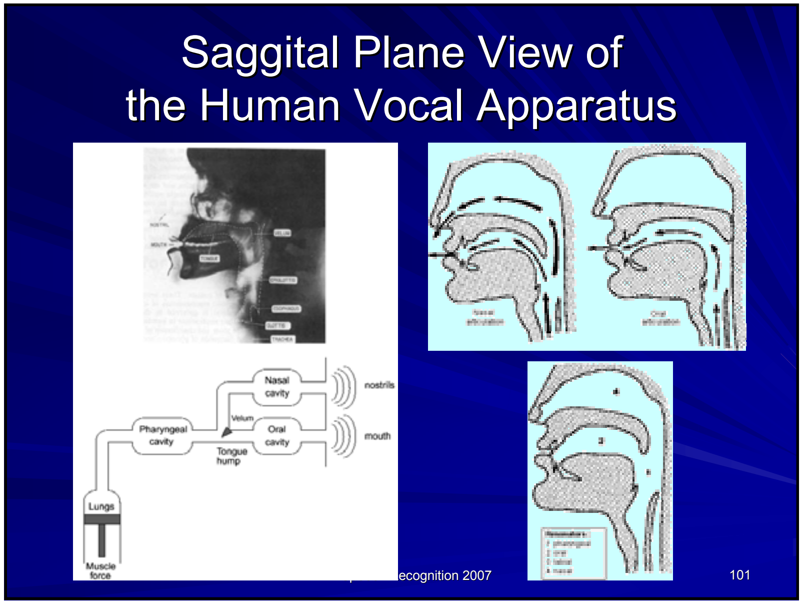
yeah those the movies are good too

and i i guess all each of the shows i like are are kind of along the same line because they're all adventure [laughter]

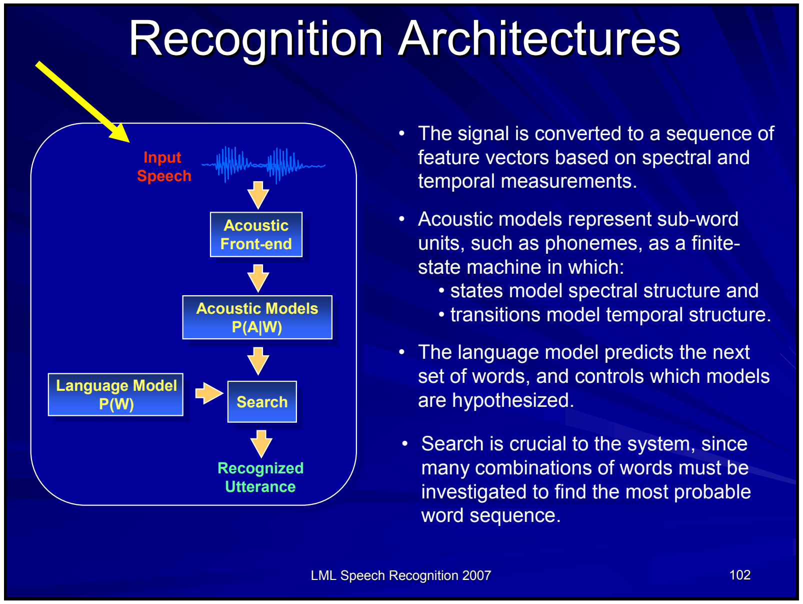
Implementation Issues
Cross-Word Decoding Is Expensive

- Cross-word Decoding: since word boundaries don't occur in spontaneous speech, we must allow for sequences of sounds that span word boundaries.
- Cross-word decoding significantly increases memory requirements.





Saggital Plane View of the Human Vocal Apparatus



Recognition Architectures

Acoustic Front-end

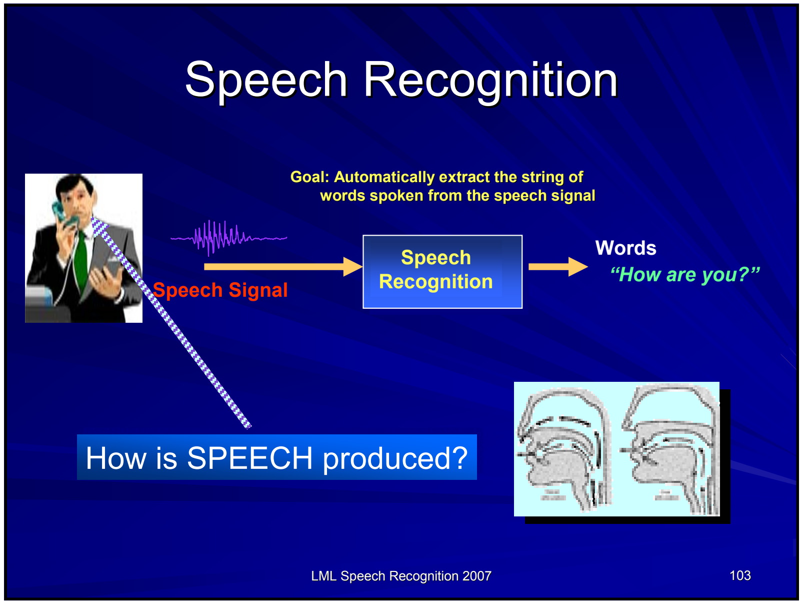
Acoustic Models P(A|W)

Language Model P(W)

Search

Recognized Utterance

- The signal is converted to a sequence of feature vectors based on spectral and temporal measurements.
- Acoustic models represent sub-word units, such as phonemes, as a finite-state machine in which:
  - states model spectral structure and
  - transitions model temporal structure.
- The language model predicts the next set of words, and controls which models are hypothesized.
- Search is crucial to the system, since many combinations of words must be investigated to find the most probable word sequence.



Speech Recognition

Goal: Automatically extract the string of words spoken from the speech signal

Speech Signal

Speech Recognition

Words

“How are you?”

How is SPEECH produced?