Australian Language Technology Summer School

Introduction to

Speech Processing

Dr David B. Grayden
The Bionic Ear Institute
dgrayden@bionicear.org

Outline

1. The Speech Signal
2. Speech Perception
3. Analysis of Speech
4. Audio Coding
5. Audio Watermarking
6. Speech Recognition
7. Speech Synthesis

Some References

- O'Shaughnessy, Douglas (2000) *Speech Communications: Human and Machine*, IEEE Press, New York.
- Deller, J, Proakis, J, Hansen, J (1993) *Discrete-Time Processing of Speech Signals*, Macmillan, New York.
- Painter, T & Spanias, A. (2000) "Perceptual coding of digital audio," *Proc. IEEE* 88(4), 451-513.
- Swanson, MD, Kobayashi, M & Tewfik, AH (1998) "Multimedia Data-Embedding and Watermarking Technologies," *Proc. IEEE* 86(6), 1064-1087.

The Speech Signal

Speech communication is the transfer of information via speech, either between persons or between humans and machines.

Language is one of the most important of human capabilities. This makes it an ideal form of communication between humans and machines.

The Speech Chain

Production → Transmission → Perception

Human:		
Vocal tract	Pressure waves through air	Auditory system
Machine:		
Speech synthesis	Speech coding	Speech & speaker recognition

The Speech Organs

- The Lungs and Thorax
  - Generate the airflow that passes through the larynx and vocal tract.
- Larynx and Vocal Folds/Cords
  - Obstruct airflow from the lungs to create turbulent noise or pulses of air.
- Vocal Tract
  - Produces the many sounds of speech by:
    - Modifying the spectral distribution of energy and
    - Contributing to the generation of sound

The Vocal Folds

- Control the ***fundamental frequency*** (***F0***) of a speaker's voice by controlling the rate of vocal fold vibration when air passes between the folds.
- Sounds produced with vocal fold vibration are called ***voiced***.
  Sounds without vocal fold vibration are ***unvoiced***.
- Turbulence may also be created using the vocal folds for the production of sounds like /h/ and for whispered sounds.

Manner of Articulation

Manner of articulation describes the configuration of the articulators in the vocal tract.

Different manner of articulation categories are:

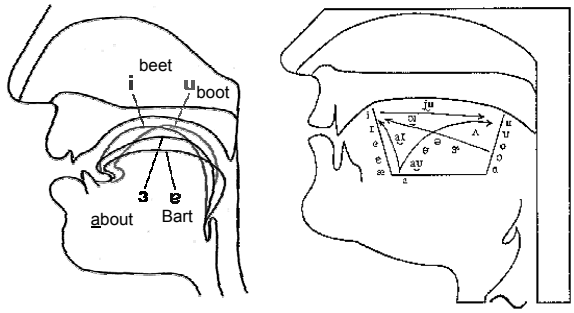
Category	Description	Example
Vowel	Little constriction of the vocal tract	'b**a**t'
Diphthong	Vowels with changing configuration	'b**ay**'
Glide	Transient sounds with fixed starting points	'**w**ay'
Liquid	Greater obstruction that vowels	'**r**ay'
Nasal	All the air passes through the nose	'**m**ay'
Fricative	Restricting airflow to create turbulence	'**s**ay'
Plosive	Closure of the air passage then release	'**b**ay'
Affricate	Plosive followed by a fricative sound	'**j**ay'

Place of Articulation

Place of articulation describes the configuration of the vocal tract that distinguishes between the ***phonemes*** within a manner of articulation group.

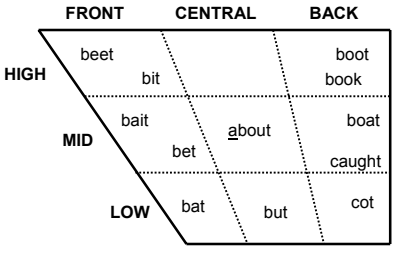
These are generally controlled by the position and shape of the tongue, though for some sounds the teeth and lips are also important articulators.

Vowel Place of Articulation



(Edwards H, 1997, "Applied Phonetics: The Sounds of American English," Singular Publishing Group)

Vowel Place of Articulation

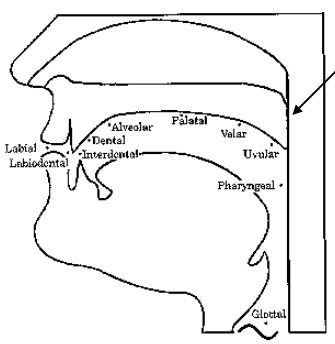


Position of maximum height of the tongue

Consonant Place of Articulation

- Consonant place of articulation describes the place where the tongue touches, or nearly touches, the vocal tract.
- If the tongue is not involved, then it describes the articulators that are most involved.
- Possible places of articulation are:
  - Labial, labiodental, dental, alveolar, palatal, velar, glottal.
  - Other terms used are front, back, retroflex, lateral.
  - For nasals the place of articulation is the position of constriction closing off the oral tract.

Consonant Places of Articulation

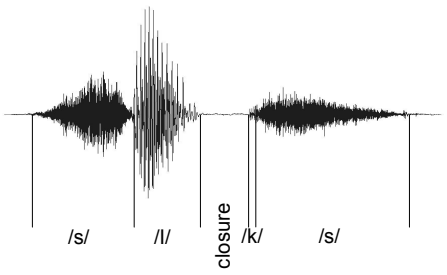


(Edwards, 1997)

Acoustic Phonetics

- The different articulations of speech result in different acoustic properties of the produced sounds. Acoustic phonetics describes the relationship between these acoustic properties and the phoneme that was uttered.
- The most significant differences between manners of articulation can be observed in the waveforms directly.
  - Plosives and affricates are characterised by a period of silence followed by a burst of activity.
  - Fricatives have very rapidly fluctuating waveforms because of their noisy production.
- Voiced phonemes show periodicity in the waveforms.

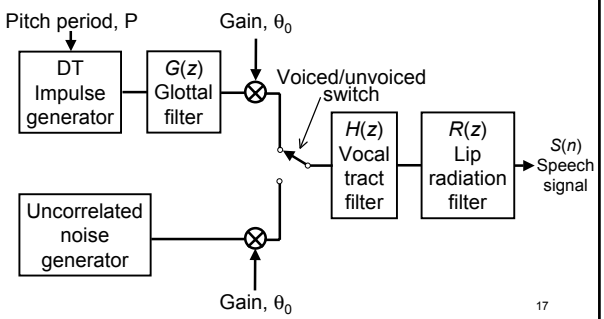
Waveform Example ('six')



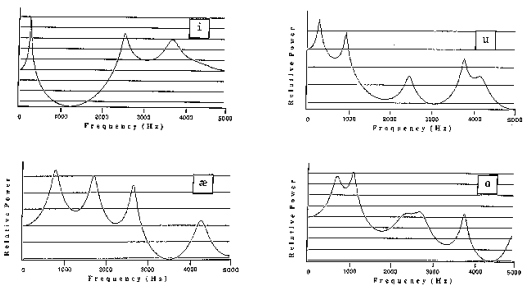
Acoustic Phonetics

- The main differences between phonemes can be seen in their spectra. The vocal tract creates resonance cavities with natural frequencies that are adjusted by changing the relative positions of the articulators.
- These resonance frequencies are called ***formants*** and are seen as regions of high energy in ***spectrograms***. The formants are the primary acoustic cues.
- Since the vocal apparatus during speech never stops moving, the formants vary continuously during an utterance.

Discrete-Time Speech Production Model

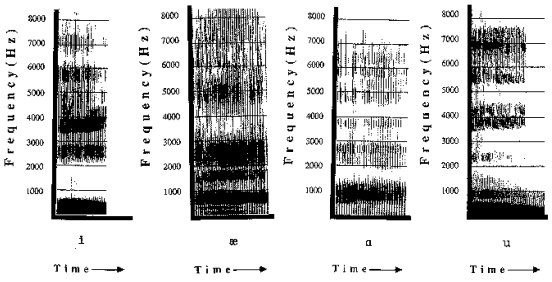


Vowel Spectra



(Edwards, 1997)

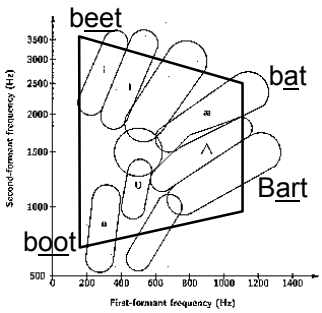
Vowel Spectrograms



e.g. beet bat Bart boot

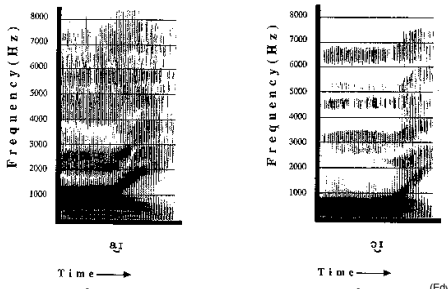
(Edwards, 1997)

Vowels: F1 & F2



(After Peterson G, Barney H, 1952, "Control methods in a study of vowels," *J. Acoust. Soc. Am.*, 24, 175-184)

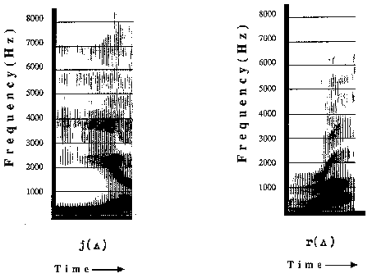
Diphthongs



e.g. bay boy

(Edwards, 1997)

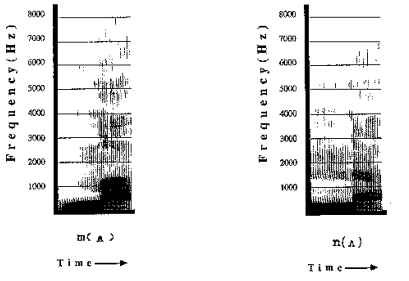
Glides and Liquids



e.g. yap rap

(Edwards, 1997)

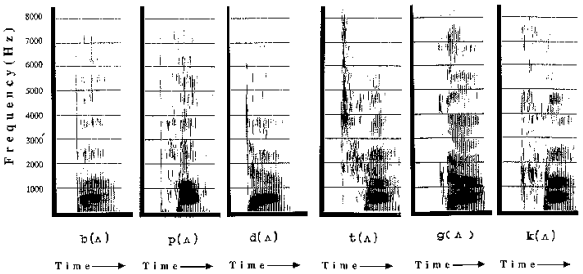
Nasals



e.g. met net

(Edwards, 1997)

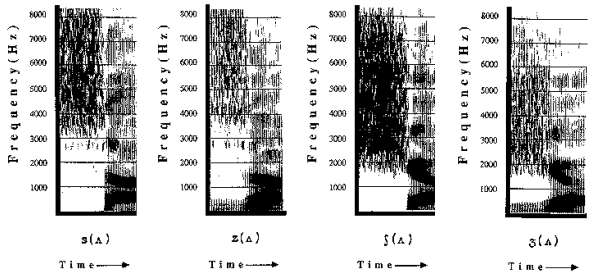
Plosives



e.g. but putt dot tot gut kut

(Edwards, 1997)

Fricatives



e.g. sue zoo shoe azure

(Edwards, 1997)

Coarticulation

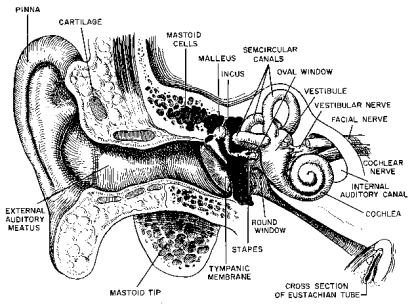
- ***Coarticulation*** is the mutual interaction of speech events that occurs because of the inability of the vocal apparatus to move instantaneously from one configuration to another.
- There is a spreading of spectral properties between sounds and stationary segments are virtually nonexistent. This is partly what makes speech recognition such a challenging task.

Prosody

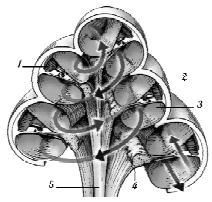
- The tonal and rhythmic aspects of speech are generally called ***prosodic features***.
- In English they normally extend over more than one phoneme and are therefore called ***suprasegmental***. They do not change the meaning of a sound but affect the meaning of what is said.
- Prosody concerns the relationships of duration, amplitude and F0 of sound sequences.
- The most obvious example is a statement versus a question.

Speech Perception

The Ear



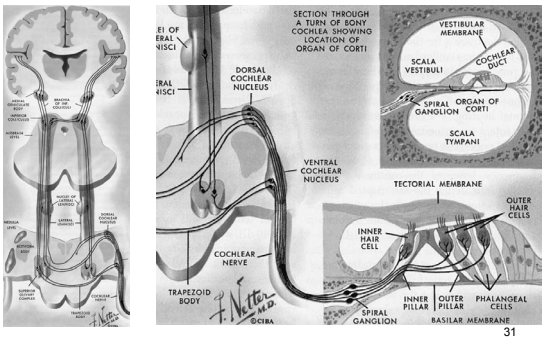
The Cochlea



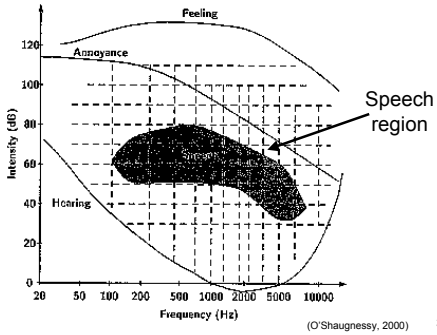
1. Organ of corti
2. Scala vestibuli
3. Scala tympani
4. Spiral ganglion
5. Auditory nerve fibres

(http://www.kameraarkasi.org/ses/kulak/cochlea.htm)

The Auditory Pathway



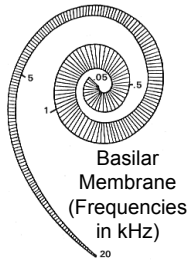
Thresholds



(O'Shaughnessy, 2000)

Pitch Perception

- The basilar membrane has increasing thickness with further distance from the oval window.
- Peak responses for sinusoidal signals are localised along the basilar membrane surface, with each peak occurring a particular distance from the oval window ("best frequency").



Basilar Membrane (Frequencies in kHz)

(http://www.kameraarkasi.org/ses/kulak/cochlea.htm)

Critical Bands

- ***Critical bandwidth*** is a function of frequency that quantifies the cochlear pass bands.
- Loudness (perceived intensity) remains the same for narrow-band noise as the noise bandwidth is increased (while keeping constant spectral level) up to the critical bandwidth. The loudness increases as the noise bandwidth increases beyond the critical bandwidth.

Critical Bands

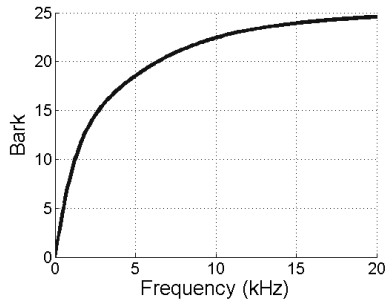
The *critical bandwidth* is approximated by

$$BW_c(f) = 25 + 75\left[1 + 1.4(f/1000)^2\right]^{0.69} \text{ (Hz)}$$

A distance of one critical band is called "one Bark"

$$Z(f) = 13\arctan(0.00076f) + 3.5\arctan\left[(f/7500)^2\right] \text{ (Bark)}$$

Bark Scale



Analysis of Speech

Time Domain Processing

- Windowing Signals
- Time Domain Parameters
  - Average Energy
  - Zero Crossing Rate (ZCR)
  - Autocorrelation

Windowing

Analysis of speech requires examination of small portions assumed to be pseudo-stationary.

Windowing yields a set of speech samples *x(n)* weighted by the shape of the window.

$$x(n) = s(n)w(n)$$

Generally, successive windows will overlap as *w(n)* tends to have a shape that will de-emphasise samples near it's edges. This breaks the speech down into a sequence of ***frames***.

Typical Windows

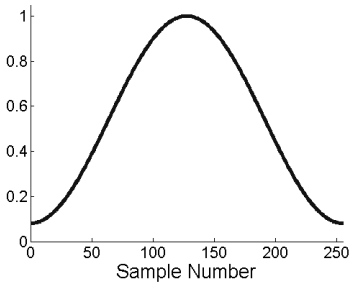
Rectangular:

$$w(n) = h(n) = \begin{cases} 1 & \text{for } 0 \le n \le N-1 \\ 0 & \text{otherwise} \end{cases}$$

Hamming:

$$w(n) = h(n) = \begin{cases} 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right) & \text{for } 0 \le n \le N-1 \\ 0 & \text{otherwise} \end{cases}$$

Hamming Window



Average Energy & Magnitude

$$E(n) = \sum_{m=-\infty}^{\infty} s^2(m)w(n-m) \qquad M(n) = \sum_{m=-\infty}^{\infty} |s|(m)w(n-m)$$

Energy emphasises high amplitudes more than magnitude.

Used to segment speech into smaller units such as:

- phonemes
- voiced / unvoiced portions
- words (isolated speech surrounded by pauses)
- speech boundaries (to reduce transmission size)

Zero-Crossing Rate (ZCR)

- Low-cost spectral information
- Simply calculated as the number of times the signal crosses the time axis within the specified window duration.

$$T[s(n)] = 0.5\,|\,\mathrm{sgn}(s(n)) - \mathrm{sgn}(s(n-1))\,|$$

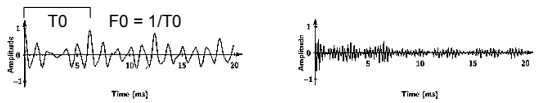
- Used to help in making voicing decisions (high ZCR indicates unvoiced speech)
- ZCR is highly sensitive to noise.

Autocorrelation

$$R_n(k) = \sum_{m=-\infty}^{\infty} s(m)w(n-m)s(m-k)w(n-m+k)$$

Sum the products of windowed speech *s(n)* with its delayed version *s(n-k).*

Used for F0 determination: $R_n(k)$ is a maximum when *k* is near the estimated pitch period.



voiced speech unvoiced speech

(O'Shaugnessy, 2000)

Frequency Domain Processing

- The Short-Term Fourier Transform

$$S_n(e^{j\omega}) = \sum_{m=-\infty}^{\infty} s(m)e^{-j\omega m}w(n-m)$$

- The Discrete Fourier Transform (DFT)

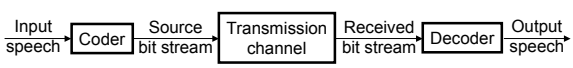
$$S_n(k) = \sum_{m=0}^{N-1} s(m)e^{-j2\pi km/N}w(n-m)$$

- Uses:
  - spectrograms
  - formant estimation

Audio Coding

Coding

- The major objective of audio coding is to compress the signal, i.e., to employ as few bits as possible in the digital representation of the signal.
- Considerations
  - bandwidth efficiency
  - level of signal distortion



Classes of Coders

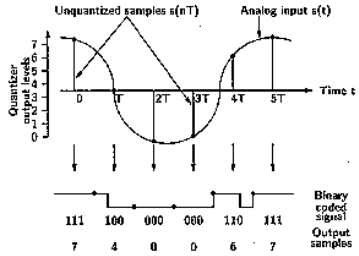
- Waveform coders
  - time-domain waveform coders take advantage of waveform redundancies.
  - spectral-domain waveform coders exploit the non-uniform distribution of speech info across frequencies.
- Source coders (vocoders)
  - follow a speech production model, encoding excitation and vocal tract excitation separately.
- Perceptual coders
  - make use of the properties of hearing especially simultaneous and nonsimultaneous masking.

Quantisation

- All digital speech coders use a form of pulse-code modulation (PCM) to convert an analog signal to a digital representation.
- Two parameters are required
  - the sampling rate, F_s, which is proportional to the bandwidth
  - the number of bits for each sample, B.

$$\text{Bit rate} = F_s B$$

Quantisation Example



(O'Shaugnessy, 2000)

Quantisation Error or Noise

- Quantisation noise is often assumed to be stationary white noise, uncorrelated with the input signal, with each error sample uniformly distributed in the range [- Δ/2, Δ/2] where Δ is the quantisation step.
- The signal to noise ratio is

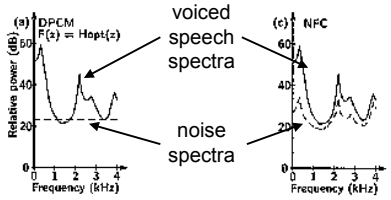
$$\text{SNR} = \frac{\sigma_x^2}{\sigma_e^2} = \frac{\sum_n x^2(n)}{\sum_n e^2(n)} \qquad \text{SNR(dB)} = 6.02B - 7.27$$

$$\text{for } \Delta = 2X_{\max} / 2^B$$

Coding Methods

- Non-uniform quantisation (companding)
  - A-law and μ-law
- Vector quantisation (VQ)
  - consider k samples together as a block or vector
- Adaptive-quantiser pulse-code modulation (APCM)
  - vary the quantiser step in proportion to short-time average speech amplitude
- Differential pulse-code modulation (DPCM)
  - exploit slowly varying spectral envelope structure
  - exploit the periodicity of voiced speech
- Exploit auditory limitations (Noise shaping)
  - minimise the subjective loudness of the noise by hiding quantisation noise at frequencies/times of high speech energy where it cannot be heard

Noise Shaping



Standard DPCM

DPCM with Noise Feedback Coding

(O'Shaugnessy, 2000)

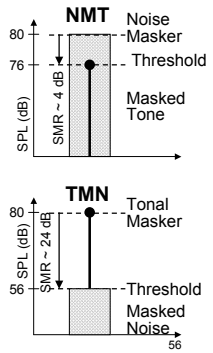
Perceptual Masking

Masking

- Masking refers to a process where one sound is rendered inaudible because of the presence of another sound. The presence of a strong noise or tone masker creates an excitation of sufficient strength on the BM at the critical band location to block detection of a weaker signal.
- Simultaneous sounds cause frequency masking.
- Sounds delayed with respect to each other can cause temporal masking.

Frequency Masking

- Noise-Masking-Tone (NMT): A narrow-band noise (1 Bark bandwidth) masks a tone within the same critical band if signal-to-mask ratio (SMR) is low (-5 to +5 dB min).
- Tone-Masking-Noise (TMN): A pure tone occurring at the centre of a critical band masks noise of an subcritical bandwidth or shape (threshold SMR 21-28 dB min).
- Noise-Masking-Noise (NMN): A narrow-band noise masks another narrow-band noise (threshold SMR 26 dB)



The Spread of Masking

- Inter-band masking also occurs, i.e., a masker centered within one critical band has a predictable effect on detection thresholds in other critical bands.
- This effect is modeled by an approx. triangular spreading function:

$$SF_{\mathrm{dB}}(x) = 15.81 + 7.5(x + 0.474) - 17.5\sqrt{1 + (x + 0.474)^2} \ \mathrm{dB}$$

where x has units of Barks and

$SF_{\mathrm{dB}}(x)$ is expressed in dB.

Masking Thresholds

After critical band analysis is done and spread of masking has been accounted for, masking thresholds may be expressed by the following dB relations:

TMN threshold: $TH_N = E_T - 14.5 - B$

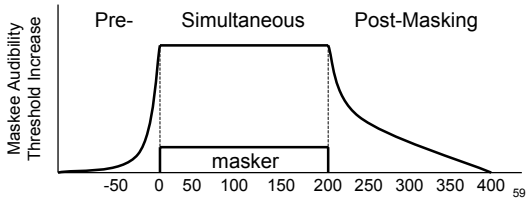
NMT threshold: $TH_T = E_N - K$

where

TH_N and TH_T	noise and tone masking thresholds
E_T and E_N	critical band noise and tone masker levels
B	critical band number
K	typically 3 – 5 dB

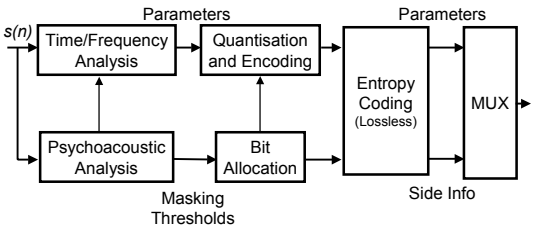
Temporal Masking

Masking also extends in time beyond the window of simultaneous stimulus presentation. For a masker of finite duration, non-simultaneous masking occurs both prior to masker onset and after masker removal.



Perceptual Coding of Sound

Generic Perceptual Audio Encoder



MPEG-1 Psychoacoustic Model 1

- Example Codec Perceptual Model that is used with MPEG-1 coding layers I and II.
- The standard (ISO/IEC 11172-3) recommends that model 2 be used for MPEG-1 layer III (MP3).
- The model determines the maximum allowable quantisation noise energy in each critical band such that the quantisation noise remains inaudible.

MPEG-1 Model 1 Steps

1. Perform spectral analysis using a 512-point FFT.
2. Estimate individual masking thresholds due to tone-like and noise-like maskers for each input frame.
3. Estimate a global masking threshold using a subset of the 256 freq. bins by adding tonal and nontonal masking thresholds.

1. Spectral Analysis

1. Normalize incoming audio samples *s(n)* according to FFT length N and bits per sample b, referencing the power spectrum to a 0-dB maximum
$$x(n) = \frac{s(n)}{N(2^{b-1})}$$
2. Segment into 12-ms (512 sample) frames using a 1/16th-overlapped Hanning window *w(n)* such that each frame contains 10.9 ms of new data.
3. Obtain a normalised PSD estimate using a 512-point FFT
$$P(k) = PN + 10\log_{10}\left|\sum_{n=0}^{N-1} w(n)x(n)e^{-j(2\pi kn/N)}\right|^2, 0 \le k \le N/2$$

PN = Power Normalisation = 90.302

2. Identification of Maskers (1)

Identify tonal and non-tonal masking components.

Tonal: local maxima that exceed neighbouring components by >7 dB within a certain Bark distance.

The "tonal" set S_T is defined as
$$S_T = \left\{ P(k) \middle| \begin{matrix} P(k) > P(k\pm 1), \\ P(k) > P(k\pm\Delta_k) + 7\,\text{dB} \end{matrix} \right\}$$

where
$$\Delta_k \in \begin{cases} 2 & 2 < k < 63 \quad (0.17 - 5.5\,\text{kHz}) \\ [2,3] & 63 \le k < 127 \quad (5.5 - 11\,\text{kHz}) \\ [2,6] & 127 \le k \le 256 \quad (11 - 20\,\text{kHz}) \end{cases}$$

Compute tonal maskers from the spectral peaks S_T as
$$P_{TM}(k) = 10\log_{10}\sum_{j=-1}^{1} 10^{0.1P(k+j)}\ (\text{dB})$$

2. Identification of Maskers (2)

Noise: a single noise masker for each critical band, $P_{NM}(k)$ is the computed from spectral lines not within the neighbourhood of a tonal masker using
$$P_{NM}(\bar{k}) = 10\log_{10}\sum_j 10^{0.1P(j)}\ (\text{dB})$$
$$\forall P(j) \notin \{P_{TM}(k, k\pm 1, k\pm\Delta_k)\}$$

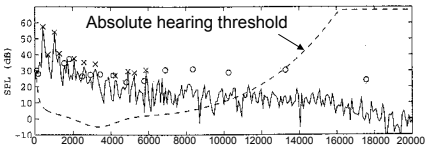
where k is defined as the geometric mean spectral line of the critical band, i.e.,
$$\bar{k} = \left(\prod_{j=l}^{u} j \right)^{1/(l-u+1)}$$

where l and u are the lower and upper spectral line boundaries of the critical band, respectively.

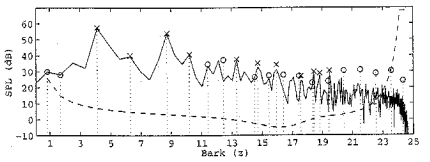
Steps 1 & 2 Example: PSD & Maskers

X – tonal maskers O – noise maskers

Linear frequency scale



Bark frequency scale



(Painter & Spanias, 2000)

3. Decimation and Reorganisation of Maskers

The number of maskers is reduced using two criteria.

1. Any tonal or noise maskers below the absolute threshold are discarded. Keep maskers that satisfy
$$P_{TM,NM}(k) \geq T_q(k)$$
where $T_q(k)$ is the SPL of the threshold in quiet at spectral line k.
2. A sliding 0.5-Bark-wide window is used to replace any pair of maskers occurring within a distance of 0.5 Bark by the stronger of the two.

Then reorganise masker frequency bins by the scheme
$$P_{TM,NM}(i) = P_{TM,NM}(k), \quad P_{TM,NM}(k) = 0$$

Where
$$i = \begin{cases} k & 1 \leq k \leq 48 \\ k + (k \bmod 2) & 49 \leq k \leq 96 \\ k + 3 - ((k-1) \bmod 4) & 97 \leq k \leq 232 \end{cases}$$

4. Calculate Individual Masking Thresholds (1)

Each threshold measurement represents a masking contribution at freq bin i due to the tone or noise masker in bin j. Tonal masker thresholds are

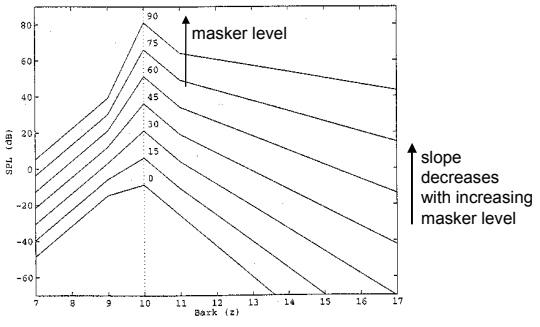
$$T_{TM}(i,j) = P_{TM}(j) - 0.275z(j) + SF(i,j) - 6.025 \text{ (dB SPL)}$$

where $P_{TM}(j)$ is the SPL of the tonal masker in bin j,
$z(j)$ denotes the Bark freq of bin j and
$SF(i,j)$ is the spread of masking from masker bin j to maskee bin i (in dB SPL)

$$SF(i,j) = \begin{cases} 17\Delta_z - 0.4P_{TM}(j) + 11 & -3 \leq \Delta_z < -1 \\ (0.4P_{TM}(j) + 6)\Delta_z & -1 \leq \Delta_z < 0 \\ -17\Delta_z & 0 \leq \Delta_z < 1 \\ (0.15P_{TM}(j) - 17)\Delta_z - 0.15P_{TM}(j) & 1 \leq \Delta_z < 8 \end{cases}$$

$$\Delta_z = z(i) - z(j)$$

Example Tonal Maskers at 10 Barks



(Painter & Spanias, 2000)

4. Calculate Individual Masking Thresholds (2)

Individual noise masker thresholds are given by

$$T_{NM}(i,j) = P_{NM}(j) - 0.175z(j) + SF(i,j) - 2.025 \text{ (dB SPL)}$$

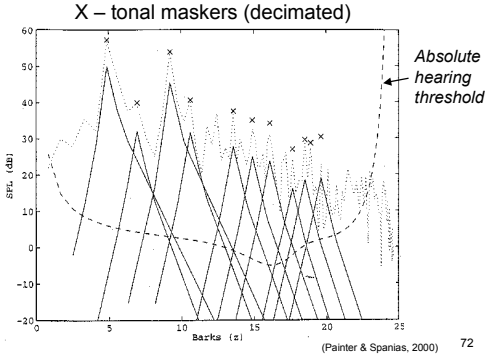
where $P_{NM}(j)$ is the SPL of the noise masker in bin j,
$z(j)$ denotes the Bark freq of bin j and
$SF(i,j)$ is

$$SF(i,j) = \begin{cases} 17\Delta_z - 0.4P_{NM}(j) + 11 & -3 \leq \Delta_z < -1 \\ (0.4P_{NM}(j) + 6)\Delta_z & -1 \leq \Delta_z < 0 \\ -17\Delta_z & 0 \leq \Delta_z < 1 \\ (0.15P_{NM}(j) - 17)\Delta_z - 0.15P_{NM}(j) & 1 \leq \Delta_z < 8 \end{cases}$$

$$\Delta_z = z(i) - z(j)$$

Step 4 Example: Tonal maskers' masking thresholds

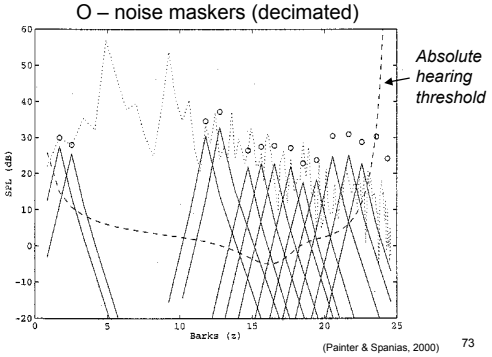
X – tonal maskers (decimated)



(Painter & Spanias, 2000)

Step 4 Example: Noise maskers' masking thresholds

O – noise maskers (decimated)



(Painter & Spanias, 2000)

5. Calculate Global Masking Thresholds

Combine individual noise masking thresholds to estimate a global masking threshold for each freq bin in the subset. The global masking threshold is

$$T_g(i) = 10\log_{10}\left(10^{0.1T_q(i)} + \sum_{l=1}^{L} 10^{0.1T_{TM}(i,l)} + \sum_{m=1}^{M} 10^{0.1T_{NM}(i,m)}\right) (\text{dB SPL})$$

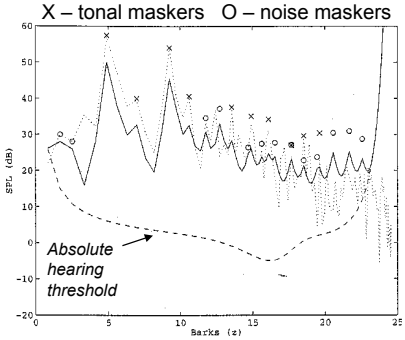
where

$T_q(i)$	absolute hearing threshold for freq bin i
$T_{TM}(i,l)$ and $T_{NM}(i,m)$	individual masking thresholds from step 4
L and M	tonal and noise maskers from step 3

Essentially, add the power of the tonal and noise maskers to the absolute threshold in quiet.

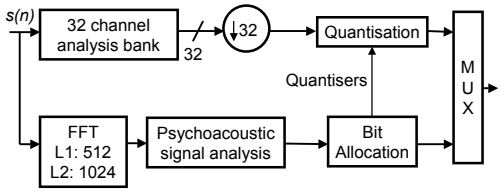
Step 5 Example: Global masking thresholds

X – tonal maskers O – noise maskers

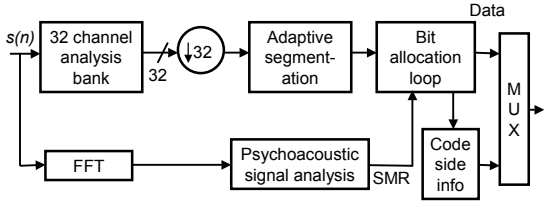


(Painter & Spanias, 2000)

MPEG-1 Layer I/II Encoder



MPEG-1 Layer III Encoder



Data Embedding & Watermarking

Steganography

- *stegano* (στεγανω) – covered
- *graphos* (γραος) – to write
- Steganography describes methods of concealing messages within other messages or objects.
- A human should not be able to differentiate between the original host signal and the signal with the inserted data.

Why steganography?

1. Passive and active copyright protection.
2. Embed control, descriptive or reference information in a signal.
   - pay-per-use applications
   - internet electronic commerce
   - track object creating, manipulation and modification
3. Covert communications and security.
4. Different levels of data
   - restrict what a person can see or hear
   - embed different sound tracks

Audio Data Embedding

- Use can be made of audio masking in the human auditory system to insert data without changing the perceptual quality.
- The same analysis used for MPEG-1 audio coding can be used to find masking thresholds for insertion of embedded information.
- Similarly, the human visual system also spatial masking patterns that may be used to embed data in images.

Audio Data Embedding Techniques using Masking

1. Phase-coding approach – data are embedded by modifying the phase values of Fourier transform (FT) coefficients of audio segments.
2. Embed data as spread-spectrum noise.
3. Echo coding – use multiple decaying echoes to place a peak in the cepstrum at a known location.
4. Replace FT coeffs. with data in mid-freq resgions.
5. Shape a pseudo-noise sequence according to the shape of the original sequence and code data in a pre-determined band.
6. Encode data in jittering of the timing between L and R audio channels.

Automatic Speech Recognition (ASR)

ASR Introduction

- Speech recognition is the process of transforming an incoming acoustic waveform into the sequence of linguistic words that it represents.
- Speech is the most natural means of communication for humans and is usually learned to a high degree of proficiency at a very early age. However, after years of research, the difficulties in implementing automatic speech recognition are still formidable.

ASR Applications

- Natural human-machine communication
- Interactive problem solving
- Telecommunications
  - cost reduction (replace human operators)
  - revenue generation (new services)
- Hands-free operation
- Dictation & automatic transcription
- Aids for the handicapped
- Automatic language translation

The ASR Problem

- Three primary considerations
  - vocabulary size
  - speaking mode
  - degree of speaker independence
- Other considerations
  - task and language constraints
  - environmental conditions
  - type of speech (eg. read vs. spontaneous)

ASR Systems

Four primary classes of pattern matchers have been developed for speech recognition:

- Knowledge/rule-based systems
- Template matchers
- Hidden Markov model (HMM) systems
- Artificial neural network (ANN) systems

In addition, many hybrid systems have been developed.

Knowledge-Based (KB) Systems

- Human spectrogram-reading experts have the ability to “read” speech from spectrograms with an accuracy of 80% to 90%.
- The rules used by these people have led to the development of knowledge-based systems that mimic the methods used by these experts.
- The basic method used is to create sets of rules that attempt to detect distinctive features and handle the variability of speech.

Limitations of KB Systems

- Integration of knowledge is laborious and difficult.
  - Rules used by spectrogram-reading experts are difficult to represent in programs.
  - Many exceptions and contextual references are required
- Existing knowledge about the acoustic-phonetic properties of speech is still incomplete.
- KB systems have difficulty in handling errors, especially those made at an early stage.
- KB systems that still not achieved sufficiently high performance for effective use.

Template Matchers

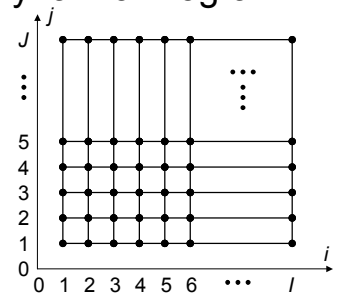
- Input speech is compared to stored reference tokens to find a closest match.
- Differences in duration and time alignment are accounted for by time normalisation using a process called *dynamic time warping* (*DTW*).
- DTW reduces the need to incorporate large amounts of acoustic-phonetic knowledge because the task is now one of pattern comparison.

DTW Assumptions

Basic DTW assumes that:

- global variations in speaking rate can be handled by linear time normalisation,
- local rate variations within each utterance are small and can be handled using distance penalties called *local continuity constraints,*
- each frame of the utterance contributes equally to ASR,
- a single distance measure applied uniformly across all frames is adequate.

Dynamic Programming



The problem is to find a "least-cost" path through the grid (going from dot to dot) from *origin node* (0,0) to *terminal node* (*I,J*).

DP – Cost Assignments

Costs are assigned to paths by cost assignments:

- Forward going *transition cost* (*Type T* case)

$$d_T[(i_k, j_k) \mid (i_{k-1}, j_{k-1})] = \text{cost from } (i_{k-1}, j_{k-1}) \text{ to } (i_k, j_k)$$

- Cost associated with a node (*Type N* case)

$$d_N(i, j) = \text{cost associated with node } (i, j)$$

- Both transitions and nodes have costs (*Type B* case). This is found by summing or multiplying d_T and d_N.

DP – Total Cost

The distance associated with a complete path is the sum of the costs of each transition and/or node along the path:

$$D = \sum_{k=1}^{K} d[(i_k, j_k) \mid (i_{k-1}, j_{k-1})]$$

$$\text{where } i_0 \equiv 0, j_0 \equiv 0, i_K \equiv I, j_K \equiv J$$

The objective is to find the path that minimises *D*.

Bellman Optimality Principle

The Bellman Optimality Principle states that the globally optimal path arriving at a node can be found by considering the set of best path segments arriving from all possible predecessor nodes and taking the one with the minimum distance. Therefore, the optimal path to a node has distance

$$D_{\min}(i_k, j_k) = \min_{(i_{k-1}, j_{k-1})} \{D_{\min}[(i_k, j_k) \mid (i_{k-1}, j_{k-1})]\}$$

$$= \min_{(i_{k-1}, j_{k-1})} \{D_{\min}(i_{k-1}, j_{k-1}) + d[(i_k, j_k) \mid (i_{k-1}, j_{k-1})]\}$$

DP Applied to Isolated Word Recognition

First, create a set of reference templates for each word by reducing the waveforms to strings of features.

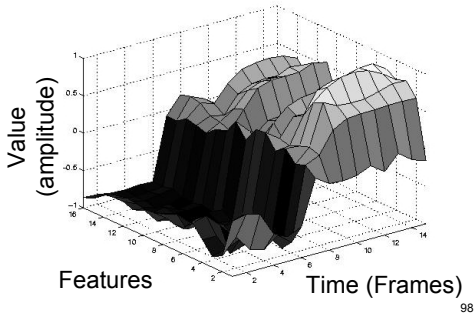
Recognition steps:

1. Determine the end points for the test utterance.
2. Create a string of features for the test utterance.
3. Match the test utterance with the reference templates using DP, obtaining the best global matching score for each pair.
4. The reference template that gave the best matching score is chosen as the recognised word.

Extracting Features

- Create a sequence of frames from the speech waveform by windowing the data.
- Extract features for each frame by some form of spectral analysis.
- A common set of features to use is filter bank outputs with filters separated according to critical bands. These are normally extracted with 5-10 ms separation between the start of each frame.

Example of a Template



Cost Functions

Say that the test features and reference features (for a particular word) are indexed by frame and are represented by:

$$\text{test features: } \mathbf{t}(1),\mathbf{t}(2),\mathbf{t}(3),...,\mathbf{t}(i),...,\mathbf{t}(I)$$
$$\text{reference features: } \mathbf{r}(1),\mathbf{r}(2),\mathbf{r}(3),...,\mathbf{r}(i),...,\mathbf{r}(J)$$

Type N cost is normally used and is commonly the Euclidean distance:

$$d_N(i_k, j_k) = d_2[\mathbf{t}(i_k), \mathbf{r}(j_k)] = \|\mathbf{t}(i_k) - \mathbf{r}(j_k)\|_2$$

The global cost of the entire match is:

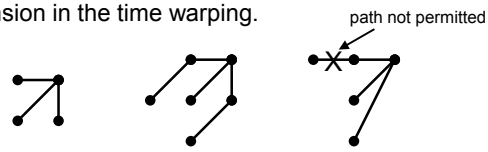
$$D = \sum_{k=1}^{K} d_N(i_k, j_k)$$

Path Constraints

The DP path for ASR should be monotonic. This means that

$$i_{k-1} \le i_k \text{ and } j_{k-1} \le j_k$$

To reduce the search space, path constraints specify the amount of allowable compression or expansion in the time warping.



Problems with DTW

1. Difficulty in adequately representing each speech unit with reference templates.
2. Heavy computational load.
3. Durational variations are always treated as noise to be eliminated by time warping.
4. Does not allow weighting different parts of an utterance by their information contribution to ASR.
5. Application to continuous speech.

Hidden Markov Models

The Hidden Markov Model (HMM) approach was developed as a stochastic extension of dynamic time warping.

The templates of DTW are instead stored as Markov chains where transition probabilities and output probabilities of the model states are derived from examples of the words or sub-word units that are being modelled.

Statistical Pattern Recognition

Speech is processed into feature vectors. The determination of the most likely word sequence is then obtained using Bayes' rule:

$$P(W \mid A) = \frac{P(W)P(A \mid W)}{P(A)}$$

This equation states that the probability that a particular word string, *W*, corresponds to an acoustic signal, *A*, is proportional to the *a priori* probability that the word string will occur times the probability of the acoustic signal occurring given that the word string was uttered.

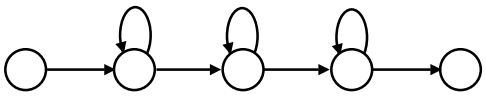
Bayes' Rule

The speech modelling is divided into two parts:

- acoustic modelling, which estimates $P(A|W) / P(A)$
- language modelling, which estimates $P(W)$

Hidden Markov modelling is one method of performing these tasks that is easily adaptable to large vocabulary continuous speech recognition.

The HMM



This is a typical HMM used in ASR. Each *state transition* has a probability of traversal and an *observation* is made upon entering each state. Each state has associated with it a probability distribution that describes the probability of each observation being made. Thus it is impossible to determine the exact state sequence traversed from the observation sequence alone (hidden).

HMM Assumptions

- The Markovian property: the probability of entering a state depends only on the previous state.
- The observation independence assumption: the probability of observing a symbol depends only on the current state and is independent of all other observations.
- Speech is a piecewise stationary process: this allows the use of frames in capturing speech characteristics.

HMM Training

- Typically in large vocabulary ASR each HMM represents a phoneme, though models may be used to represent words, sub-phonetic units or speech segments.
- The probability of observing the observation sequence (the string of features) of a phoneme for a particular HMM can be determined using the *Forward-Backward* (F-B) algorithm.
- A maximum likelihood training procedure is used to iteratively estimate the model parameters in order to maximise the probability that a model will produce the observed sequence of symbols. This is commonly the *Baum-Welsh re-estimation procedure*.

HMM Recognition

- Recognition is performed by determining which model has the highest probability of representing the input sequence of speech features.
- An efficient way of doing this is by the *Viterbi* algorithm, which approximates the likelihood of the most probable path through the models.
- If each model represents sub-units of the full utterance, they may be concatenated to cover sequences of phonemes and words.

HMM Strengths

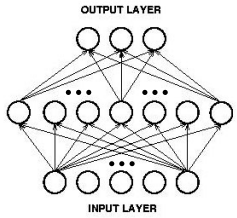
1. Ability to model any units of speech
2. Easily adapted to LVCSR
3. Training method is well established
4. Ease of adding language models
5. Ability to incorporate other methods
6. Generic noise and new word models
7. Different data representations
8. Adaptation to different speakers

HMM Weaknesses

1. Insufficient training data
2. Markovian assumption
3. Observations independence assumption
4. Representation of observation parameters
5. Poor discrimination ability
6. Poor duration modelling

Artificial Neural Networks

The motivation for artificial neural networks (ANNs) developed from a growing understanding of the human central nervous system.



ANN Strengths

1. Based on the human system
2. Nonparametric learning
3. Nonlinear universal approximators
4. Generalisation and interpolation
5. Discriminating ability
6. Efficient training procedures
7. Use of context
8. Application at different levels of ASR
9. Intrinsic robustness and fault tolerance

ANN Weaknesses

1. Long training times
2. Local minima
3. Many parameters to optimise
4. Largely static classifiers

The last weakness is the most serious. Static classifiers cannot adequately model the sequential, temporal nature of speech.

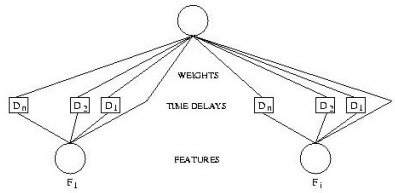
Example: Time-Delay Neural Network

- The Time-Delay Neural Network (TDNN) was developed to model context and to be robust to time alignment problems.
- The TDNN demonstrated very good performance and improved discrimination ability that could only be matched by very complicated HMM architectures.

Time Delays

- A set of features at a single instant of time is rarely sufficient to differentiate between phonemes. Therefore, it is important to encode the dynamic properties of the phonemes' spectra.
- One way is for the nodes to examine the input features at various time instances. Each connection has a time delta as well as a weight. Thus evidence is taken from each feature at $(n+1)$ time periods.

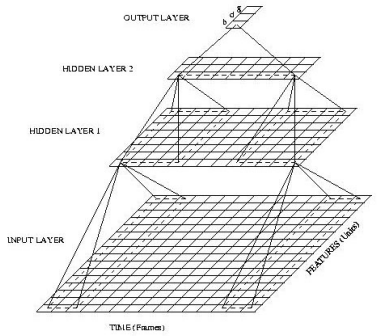
Introducing Time Delays



Complete TDNN Network

- The full architecture of a TDNN for classification of the phonemes /b/, /d/ and /g/ shows an expanded version of the basic TDNN component. There are now sufficient copies to cover all 15 input frames.
- The number of output nodes is made equal to the number of classes in the task. An extra layer is added to form the network output. It simply sums the outputs of rows of second layer nodes before applying the threshold function.

Complete TDNN ('bdg' task)



Temporal Invariance

- To overcome the problem of temporal dependence, each group of weights was made equal to all other groups on the same level of the TDNN.
- Each component was essentially the same so that each provided evidence for classification to the output layer that was independent of input alignment.
- Backpropagation training was performed as normal except that the average of the weight changes was used when updating the weights across a level.

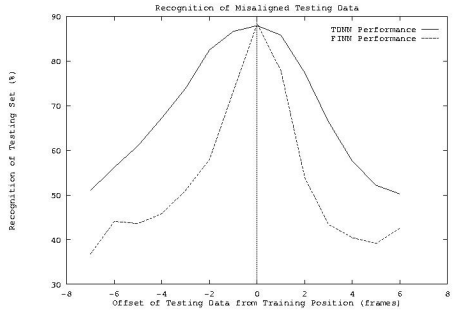
TDNN vs FINN Performance

- To evaluate the performance of the TDNN, it was compared to a fully interconnected neural network (FINN) architecture. The FINN had the same number of nodes on each layer, but each node on a layer was connected to all outputs of the layer below.
- The FINN had 5 times as many connections and 65 times more independent weights.

TDNN vs FINN

- The TDNN and FINN were trained using 3885 /b/, /d/ and /g/ utterances spoken by 306 male and female speakers in various phonetic contexts. Training was performed with the plosive onset at the 8th frame of the inputs.
- They were tested using 2197 utterances spoken by 168 speakers. Testing was performed with the plosive onset occurring at each frame of the network input.
- The TDNN showed greater tolerance to misaligned input data.

Time Invariance ('bdg' task)

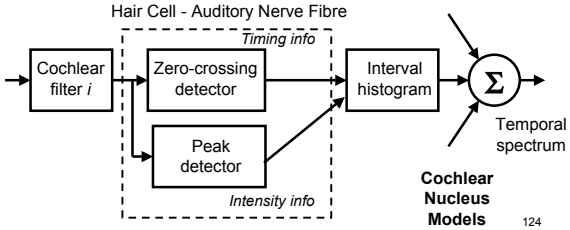


Hybrid Systems

- Hybrid systems attempt to combine different methods so that the strengths will overcome the weakness that each model presents.
- Some hybrid systems are:
  - TDNN plus spectrogram reading knowledge
  - TDNN integrated into a knowledge-based hierarchical framework.
  - ANN-HMM – ANN estimates HMM output probabilities
  - Pre-processing for HMM (LVQ)

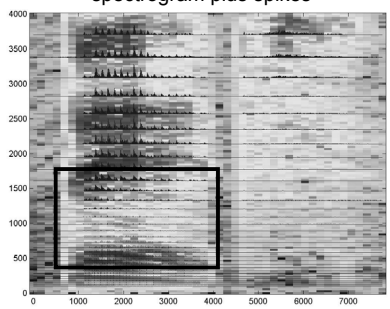
Auditory Models for ASR

More closely model human auditory processing:
Use an auditory-based model for feature extraction
- based upon spikes in the auditory nerve

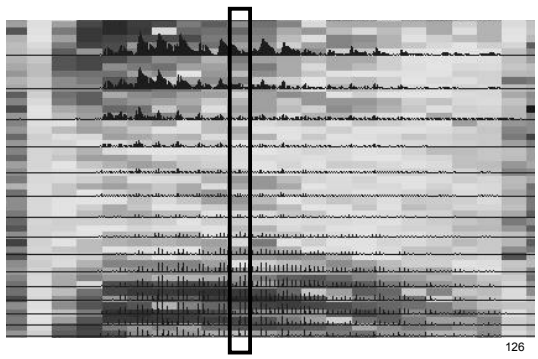


Auditory Model ("bet")

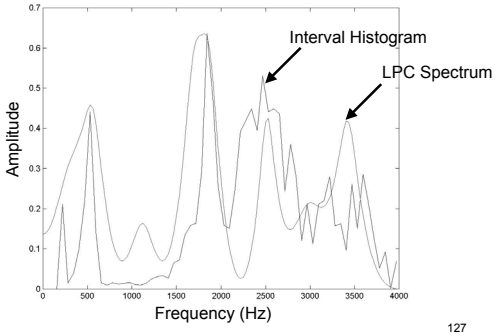
spectrogram plus spikes



Auditory Model ("bet")



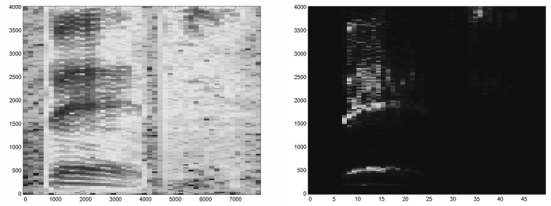
Interval Histogram



Zero Crossing Spectrogram

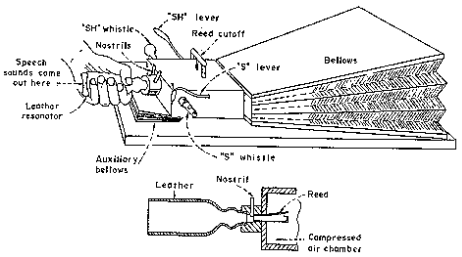
DFT Spectrogram

Zero Crossing Interval Spectrogram



Speech Synthesis

An Early Attempt



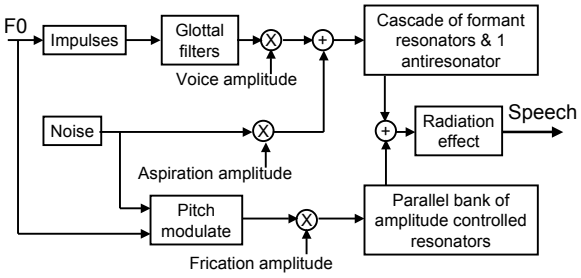
Text-to-Speech Synthesis

- Text-to-speech synthesis (TTS) is the automatic generation of a speech signal, starting from a normal text input and using previously analysed digital speech data.
- Real-time TTS is possible and is generally intelligible, but lacks naturalness. Poor naturalness is due to inadequate modelling of
  1. coarticulation
  2. intonation
  3. vocal tract excitation

Most Common Speech Synthesis Methods

- Formant synthesis – Speech waveform generation using a formant synthesiser and a set of rules.
- Waveform concatenation – message synthesis from stored waveforms.

Formant Synthesis



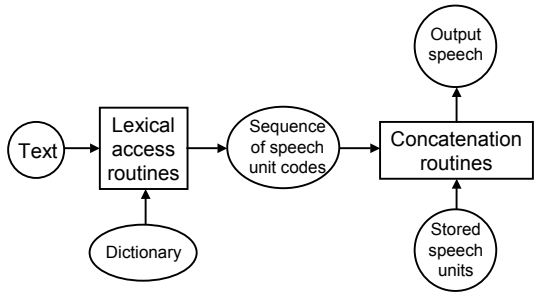
Formant Synthesis Components

- Excitation
  1. A periodic train of impulses for voiced speech.
  2. Pseudo-random noise for unvoiced speech.
  3. Periodically shaped noise for voiced fricatives.
- Vocal tract
  1. Cascade of digital resonators, one for each formant, especially good for vowels
  2. Parallel bank of filters for consonants and vowels, but each formant amplitude must be controlled explicitly.
  3. Anti-resonator for simulating nasals
- Radiation effect filter to simulate head effects and radiation from the lips.

Waveform Concatenation

- Waveform synthesis can yield very good synthetic speech at the cost of increased memory and speaker inflexibility.
- It is the most common method used today.
- Systems vary from simple systems that concatenate stored phrases to advanced systems that generate speech from sequences of basic sounds such as phonemes, though syllables, demisyllables, diphones and sub-phonemic units.

Steps in Waveform Synthesis



Phonetic Waveforms

- The spectral features of short concatenated sounds must be smoothed at their boundaries to avoid jumpy, discontinuous speech.
- However, the pronunciation of a phoneme in a phrase is heavily affected by coarticulation (the phonetic context) and intonation and speaking rate effects.
- This need to smooth at the sound boundaries gives these synthesisers less natural speech.

Areas of Speech Research

- Analysis and description of dynamic features.
- Extraction and normalisation of voice individuality.
- Adaptation to environmental variation.
- Basic units for speech processing.
- Advanced knowledge processing.
- Clarification of speech production mechanism.
- Clarification of speech perception mechanism.
- Improved modelling of speech at all levels.
- Aids for the handicapped.