KBM – Audio

Frank Nack

Outline

- Last week
- Audio – a sonic sign system
- Sound and emotions

Video Application – summary

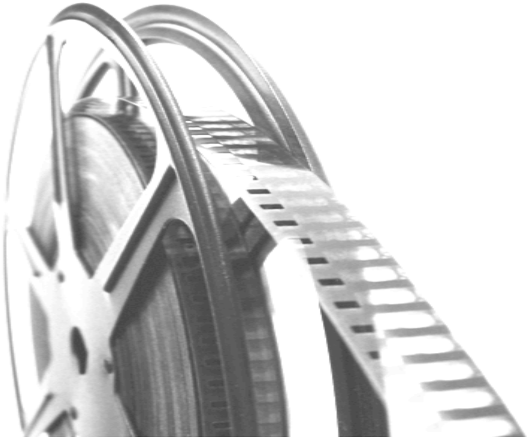
Investigated

AUTEUR

Findings

- The content determines the application
- Content description
  - Application dependent
  - Complex
  - Recourse demanding
  - Time critical
  - Incomplete
- Modular Schemata
- Description environment
  - Production supportive
  - Archive supportive

Change



Vision is the ability of the brain and eye to detect electromagnetic waves within the visible range (light) and to interpret the image as "sight."

Audition is the sense of sound perception in response to changes in the pressure exerted by atmospheric particles within a range of 20 to 20000 Hz.

Audio – a sonic sign system

Audio – Listen

Sit for a minute and listen with closed eyes to the sound of the environment.

Observe how you perceive the sound(s).

Audio – Listen

Sound listening experiment results:

- Need to focus on focus on sound signal for identification purposes
- Sound identification also determines location (of sound and listener)
- Sound signals trigger thought associations

Audio – Produce

Talk for a minute to the person the furthest away from you in the room, without looking at him or her.

Observe how you generate sound and which cues you use to interpret the input of your conversation partner.

Audio – Produce

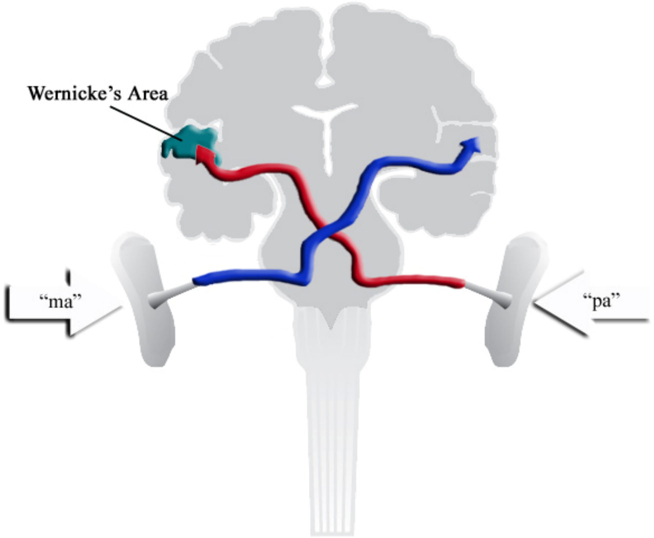
Sound producing experiment results:

- The higher the interference the louder the voice.
- At a certain sound level one has to look at the conversation partner to focus on sound stream
- The noise level determines the use of additional clues, e.g. eye contact, use of hands to emphasise importance, etc.

Audio – Sound

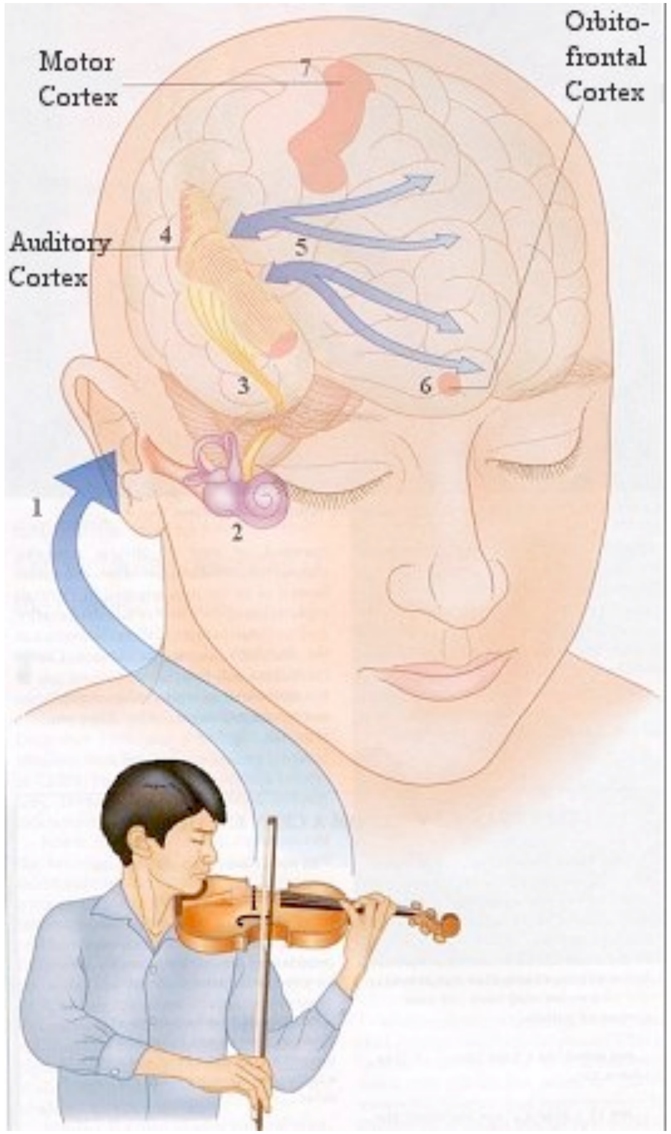
Generating

Hearing



Listening

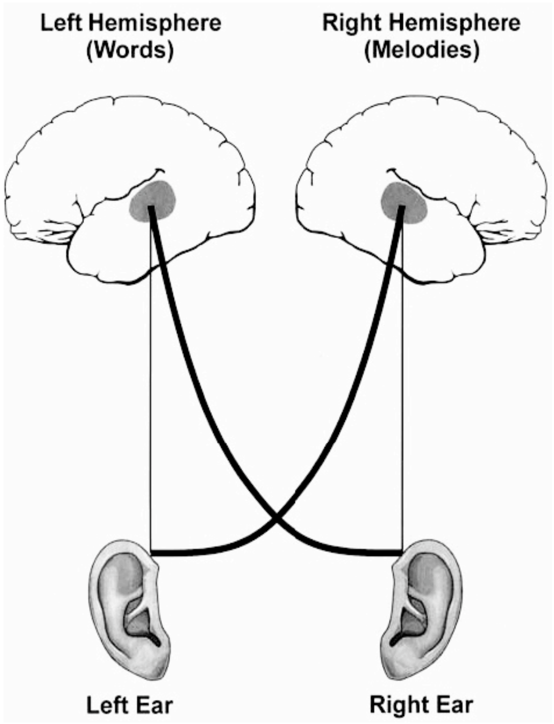
Audio – Sound



http://universe-review.ca/R12-03-wave.htm

1. Sound waves travel to the outer ear.
2. The sound waves are transduced into neural impulses by the inner ear.
3. The information travels through several waystations in the brainstem and midbrain to reach the auditory cortex.
4. The auditory cortex analyses and interprets the various aspects of the sound (comprehension, naming, verbal).
5. Information from this region interacts with many other brain areas, especially the frontal lobe, for memory formation and interpretation (episodic and declarative memory, transfer from short to long term memory).
6. The orbitofrontal region is one of many involved in emotional evaluation.
7. The motor cortex is involved in sensory-motor feedback circuits, and in controlling the movements needed to produce music using an instrument.

Audio – Sound listening



Listening Factors

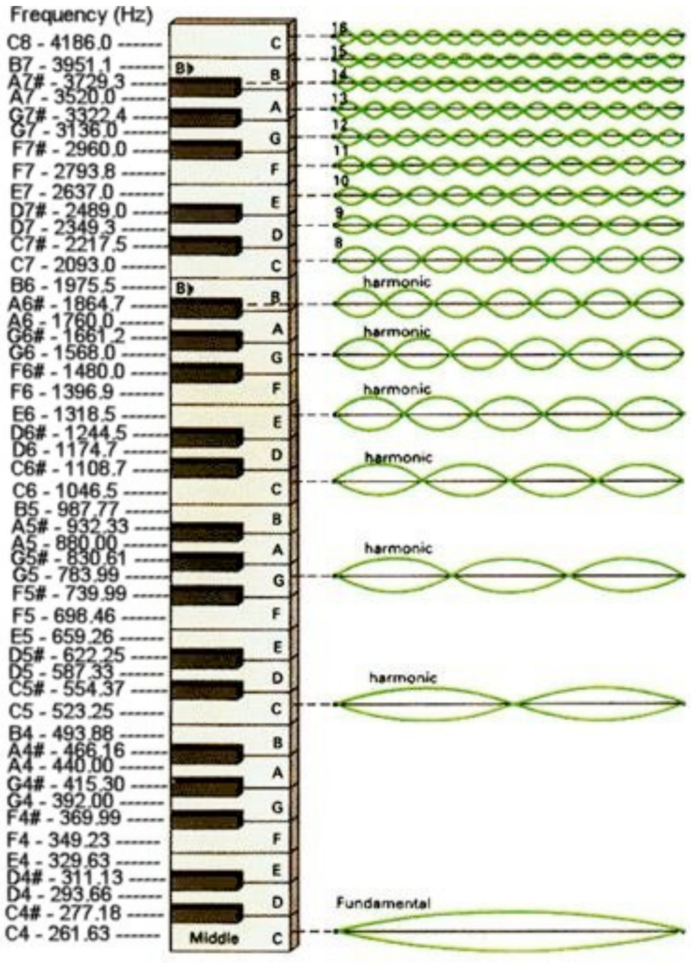
- Physical entities

  Frequency, wavelength, amplitude, intensity, speed, direction

- Context
  - danger, navigation, communication
  - speech, music, noise
  - cues (e.g. facial expressions, gestures)
  - personal experience and feelings

- Sound quality
  - pitch (melody and harmony),
  - rhythm (tempo, meter, and articulation),
  - dynamics,
  - structure,
  - sonic qualities (timbre and texture).

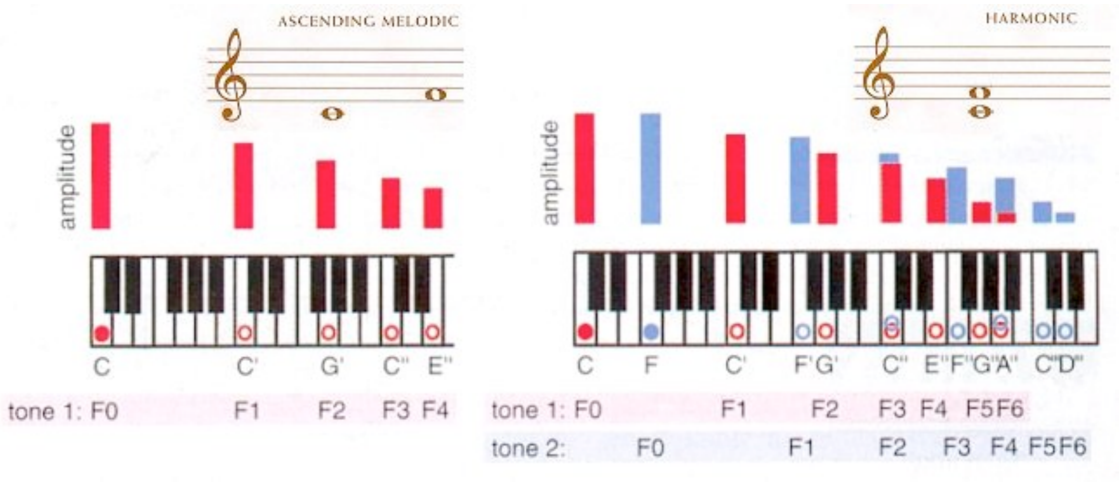
Audio – Sound listening



Pitch is the perceived fundamental frequency of a sound.

Melody is a series of linear events, not a simultaneity as in a chord.

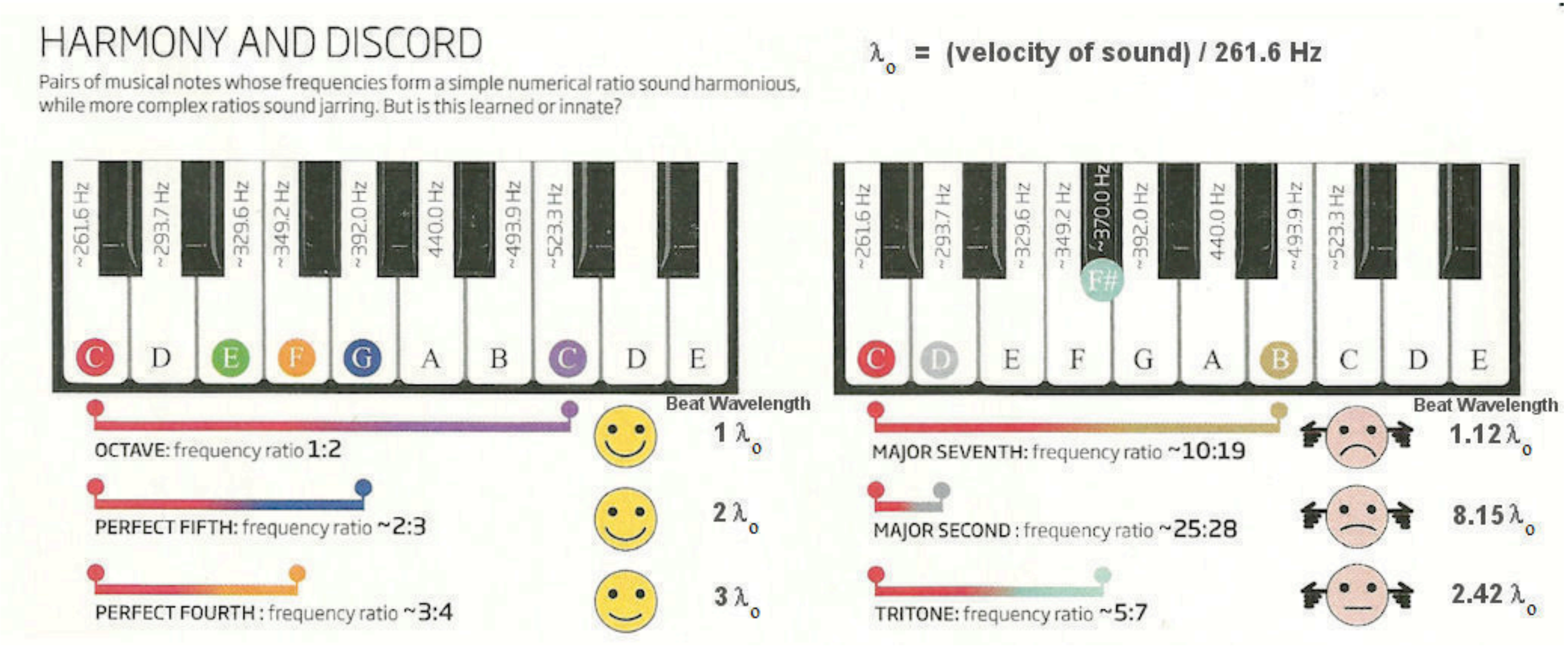
Harmony is the use of different pitches simultaneously.



All images from http://universe-review.ca/R12-03-wave.htm

Audio – Sound listening

Harmony is the use of different pitches simultaneously.



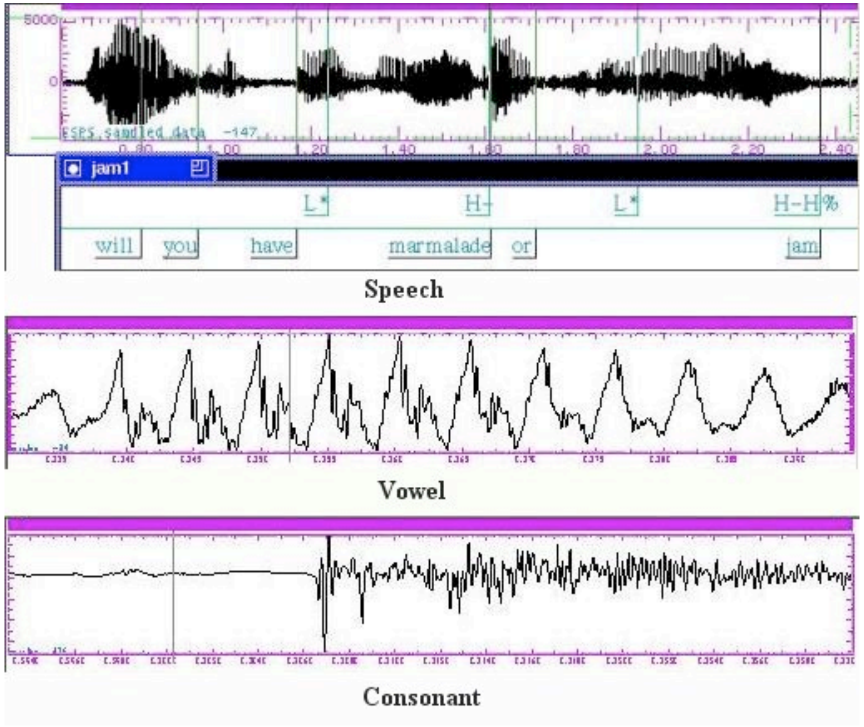
HARMONY AND DISCORD

Pairs of musical notes whose frequencies form a simple numerical ratio sound harmonious, while more complex ratios sound jarring. But is this learned or innate?

λ_o = (velocity of sound) / 261.6 Hz

http://universe-review.ca/R12-03-wave.htm

Audio – Sound listening



Speech

Vowel

Consonant

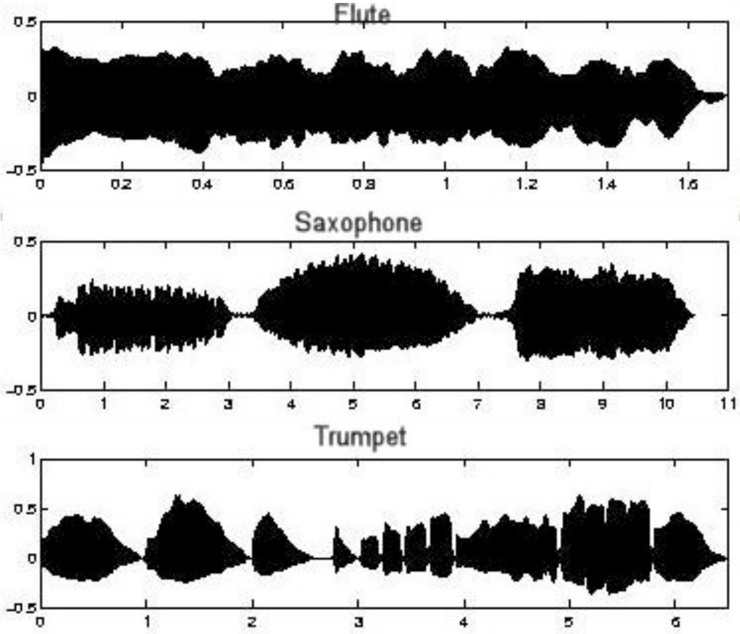
http://universe-review.ca/R12-03-wave.htm

Rhythm is the variation of the length and accentuation of a series of sounds.

Tempo is the speed or pace of a given piece.

Articulation refers to the performance technique which affects the transition or continuity on a single note or between multiple notes or sounds.

Audio – Sound listening



http://universe-review.ca/R12-03-wave.htm

Dynamics normally refers to the softness or loudness of a sound (Amplitude).

Timbre is the quality of a sound that distinguishes different types of sound production, such as voices or musical instruments.

Texture is the overall quality of sound of a piece, most often indicated by the number of voices in the music and by the relationship between these voices, using terms such as "thick" and "light", "rough" or "smooth". The perceived texture of a piece can be affected by the timbre of the instruments or voices playing these parts and the harmony, tempo, and rhythms used.

Audio – Interpretation

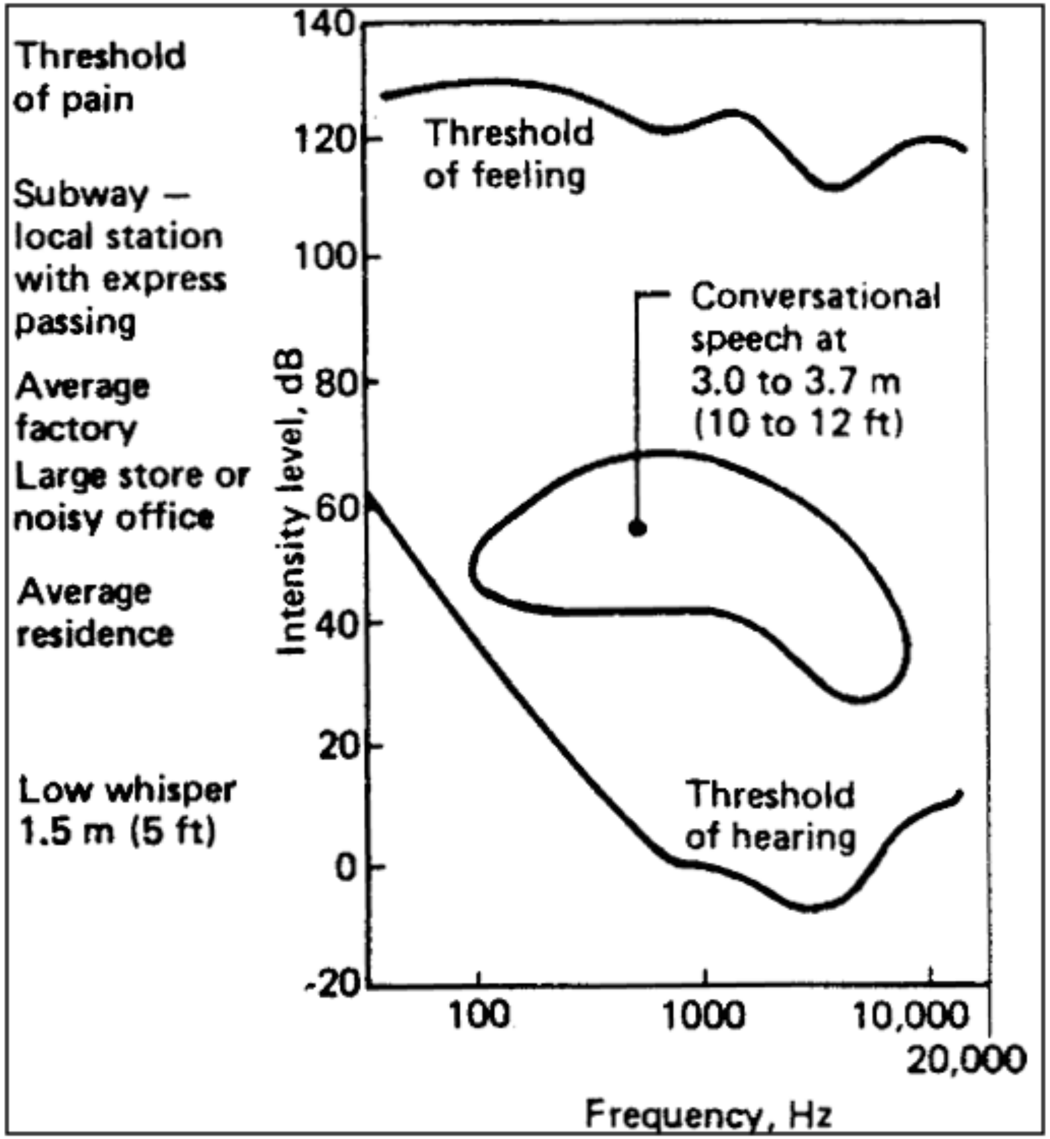
Prosody is the rhythm, stress, and intonation of speech or singing.

Prosody may reflect the emotional state of a speaker

- whether an utterance is a statement, a question, or a command
- whether the speaker is being ironic or sarcastic; emphasis, contrast and focus
- in short everything in language that is not encoded by grammar.

Eaxmple: http://vladlen.info/publications/real-time-prosody-driven-synthesis-of-body-language/

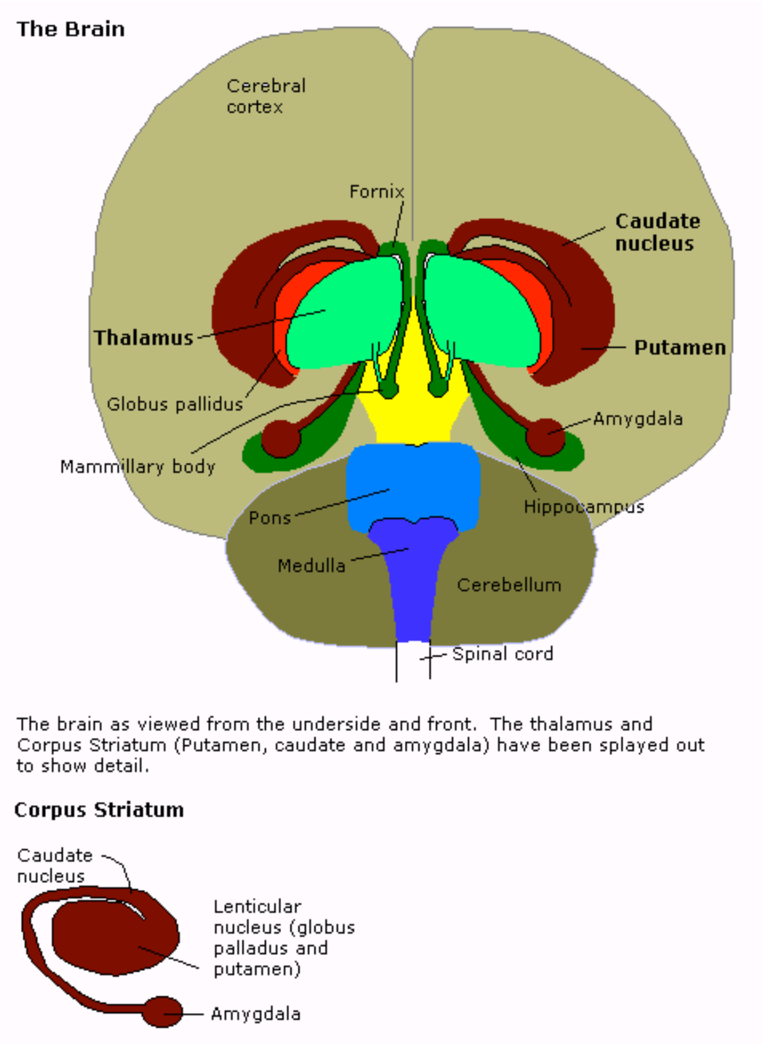
Audio – Thresholds of hearing



Human Auditory Response	Decibels (db)
Threshold of pain	140
Subway - local station with express passing	120
	100
Average factory, large store, or noisy office	80
	60
Average residence	40
Low whisper 1.5 m (5 ft)	20
	0

http://msis.jsc.nasa.gov/sections/section04.htm#_4.3_AUDITORY_SYSTEM

Audio – Interpretation



The brain as viewed from the underside and front. The thalamus and Corpus Striatum (Putamen, caudate and amygdala) have been splayed out to show detail.

The **amygdalae** are almond-shaped groups of nuclei located deep within the medial temporal lobes of the brain in complex vertebrates, including humans.

Shown in research to perform a primary role in the processing and memory of emotional reactions, i.e. **physiological arousal**, **expressive behaviors**, and **conscious experience** the amygdalae are considered part of the limbic system.

Audio – Interpretation

Sound in music or human speech is about : **what** and **how**.

The how expresses the emotional value.

Primary emotions (i.e., innate emotions, such as fear) depend on the limbic system circuitry, with the amygdala and anterior cingulate gyrus being "key players".

Secondary emotions (i.e., feelings attached to objects [e.g., to dental drills], events, and situations through learning) require additional input, based largely on memory, from the prefrontal and somatosensory cortices. The stimulus may still be processed directly via the amygdala but is now also analyzed in the thought process.

Audio – Interpretation

Friends https://www.youtube.com/watch?v=oLvB_ybcKt0

Two and a half men https://www.youtube.com/watch?v=TbZCzu36ucc

No country for old men https://www.youtube.com/watch?v=LluIBu8HdV4

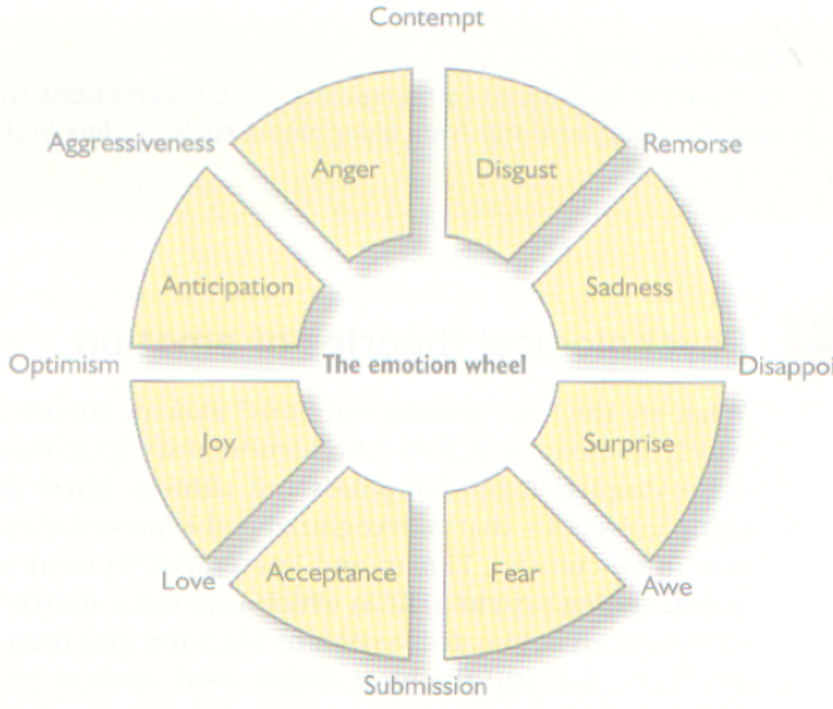
The dark knight https://www.youtube.com/watch?v=6p1kjOU1O_g

Audio – a sonic sign system - summary

- Audition is the sense of sound perception in response to changes in the pressure exerted by atmospheric particles within a range of 20 to 20000 Hz.
- Audio interpretation of speech, music, noise is context dependent (space, semantic, emotion)
- Individual characteristics are important (thresholds of frequency spaces)
- Audio stimulates primary emotions

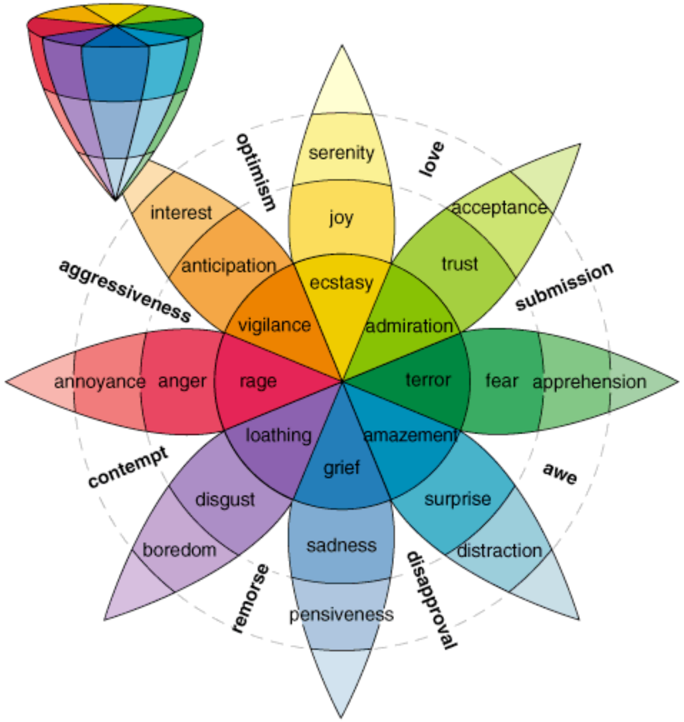
Audio – emotional representation

Audio – emotion in human speech I

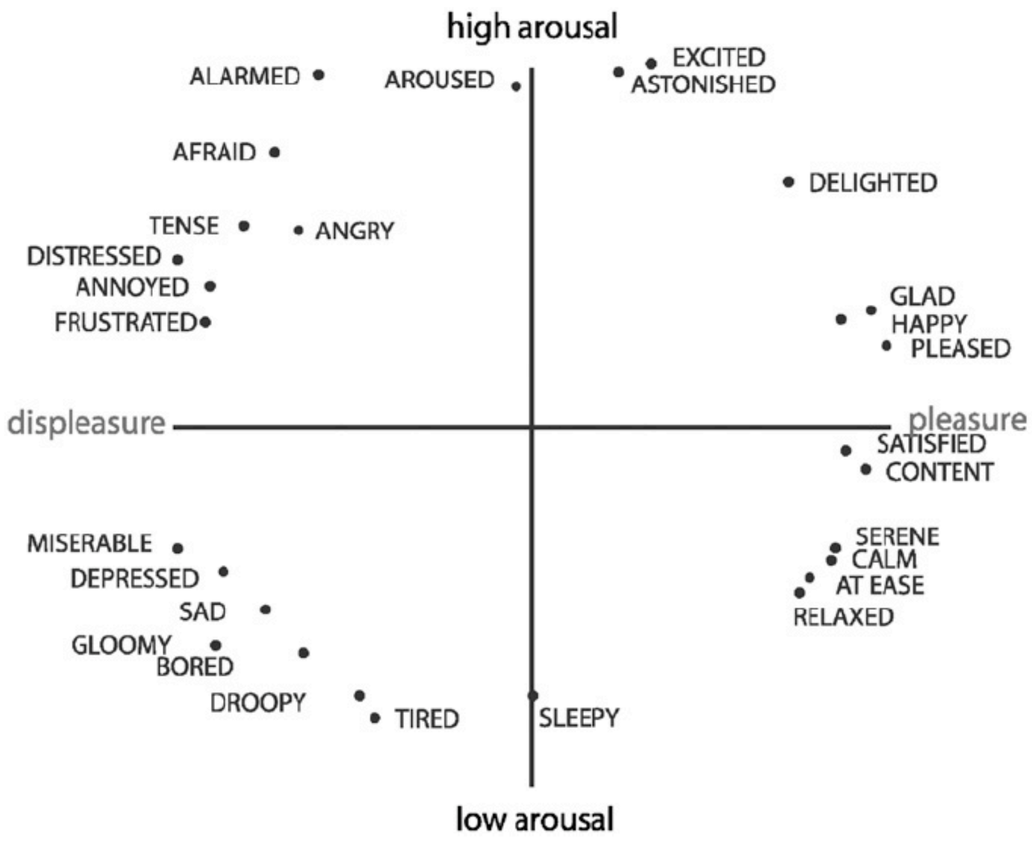


Robert Plutchik (1980)

Joy vs Sadness
Trust vs Disgust
Fear vs Anger
Surprise vs Anticipation



Audio – emotion in human speech II



Circumplex Model [Russel, 1980]

Arousal is the state of being awake (leading to increased heart rate and blood pressure and a condition of sensory alertness, mobility and readiness to respond).

Valence is the intrinsic attractiveness (positive valence) or aversiveness (negative valence) of an event, object, or situation.

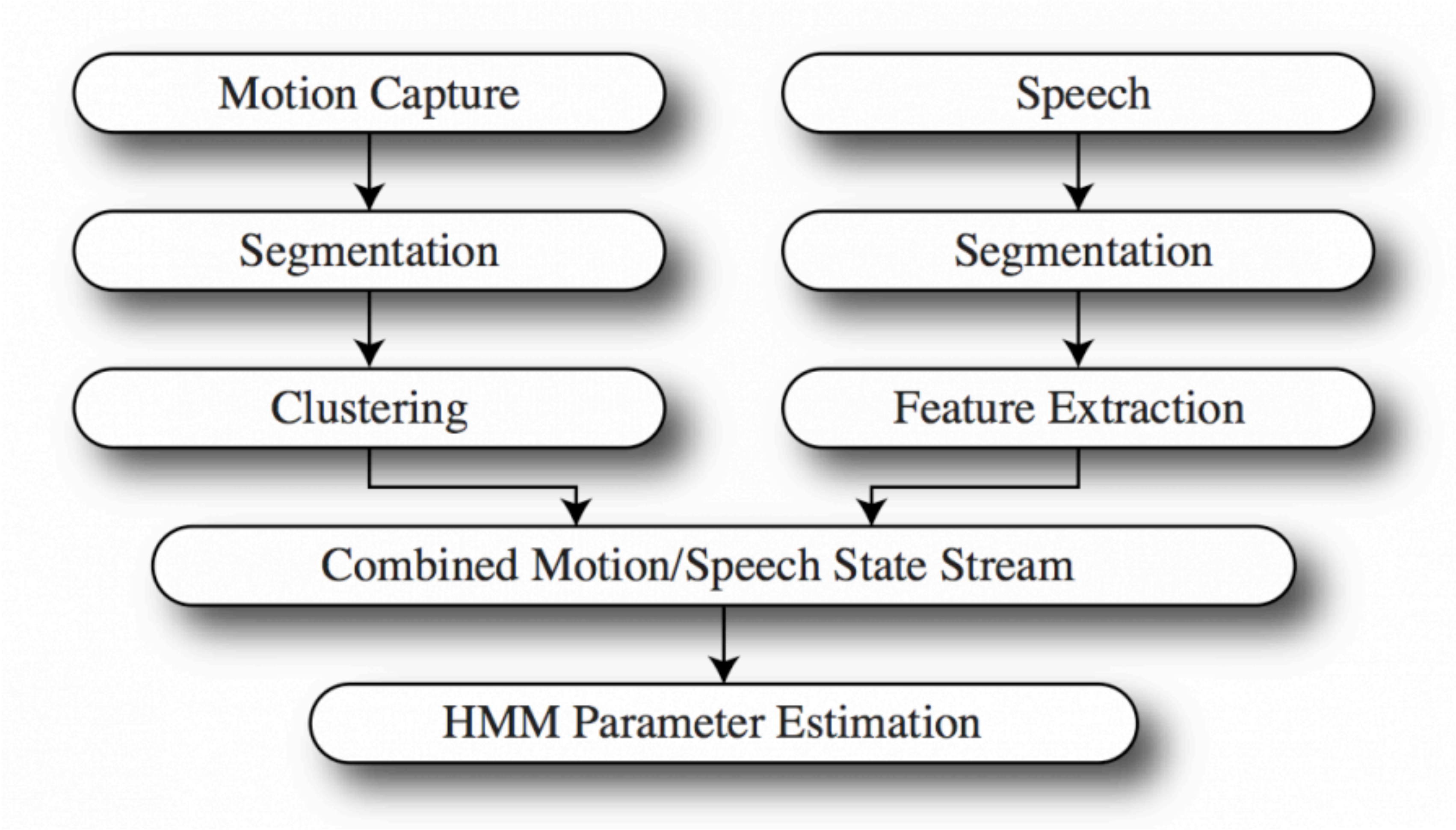
Audio – emotion in human speech III

Variations of accoustic variables Murray & Arnott (1993)

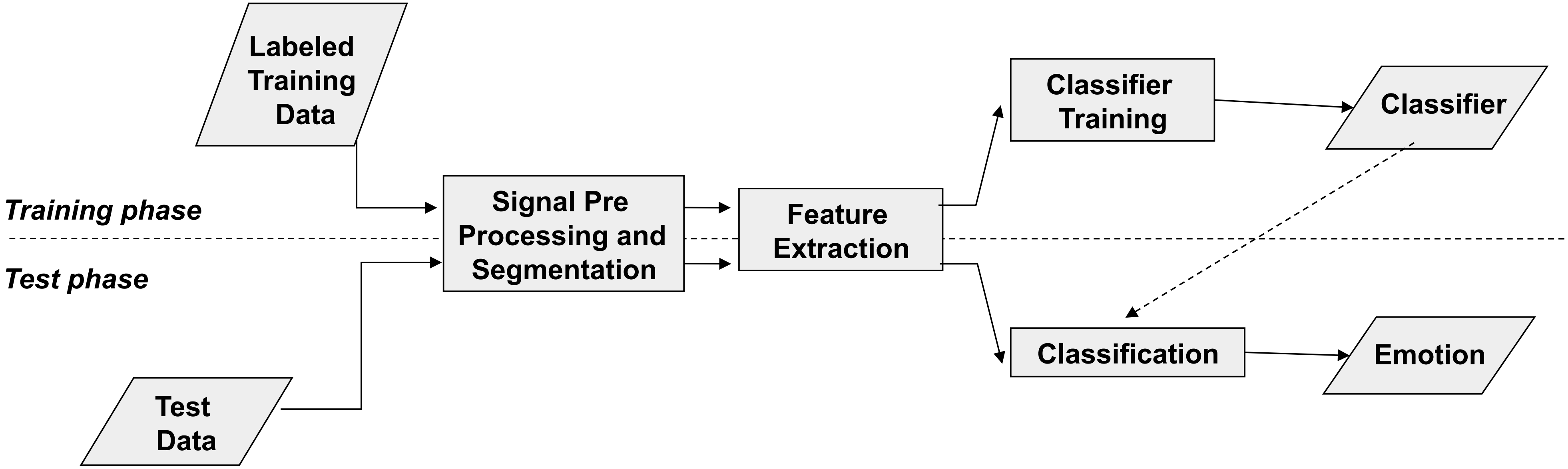
Emotion	**Pitch**	**Intensity**	**Speaking Rate**	**Voice Quality**
Anger	high mean, wide range	increased	increased	breathy
Joy	Increased mean and range	increased	slow tempo; increased rate	sometimes breathy; moderately blaring timbre
Sadness	normal or lower than normal mean, narrow range	decreased	slow	resonant timbre

Audio – Measurement

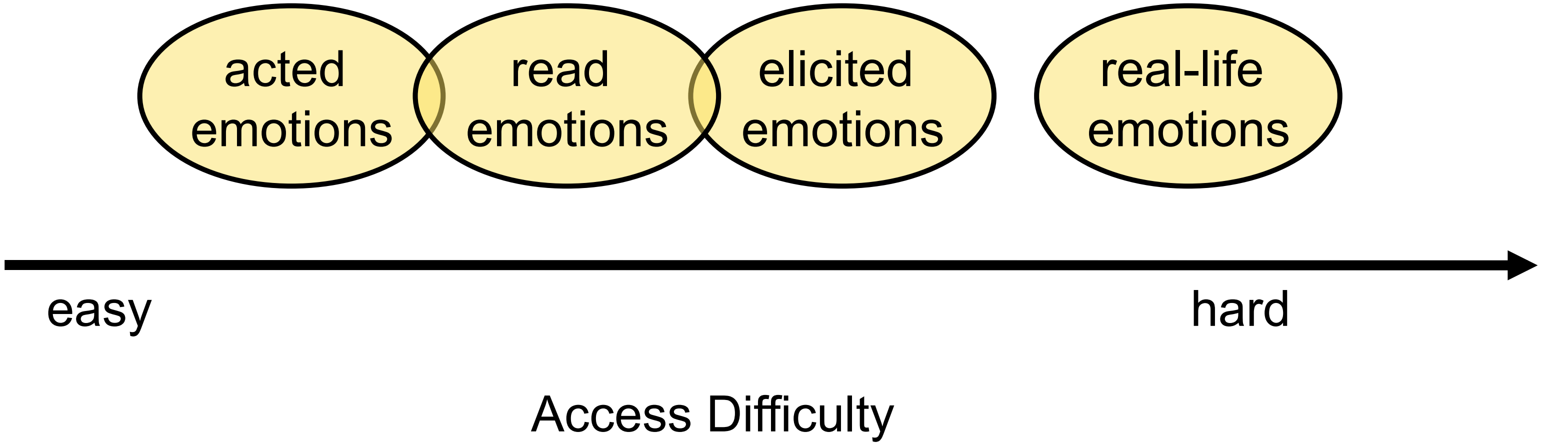
Audio - Emotion Recognition System



Audio - Emotion Recognition System



Audio – speech material problem



Audio – Segmentation

Goal: segment a speech signal into units that are representative for emotions.

Target: words, utterances, words in context, fixed time intervals

Requirement:

- long enough to reliably calculate features by means of statistical functions
- short enough to guarantee stable acoustic properties with respect to emotions within the segment

Audio – Feature extraction

Goal: find those properties of the digitised and pre-processed acoustic signal that are characteristic for emotions and to represent them in a n-dimensional feature vector.

Features: Pitch and energy related features (e.g. short time energy, maximum amplitude, average of the n largest amplitudes, ADSR model, etc.)
Durational and pause related features
Types of voice quality features.

Problem: A high number of features is not beneficial because most classifiers are negatively influenced by redundant, correlated or irrelevant features.

Solution: Principal component analysis (PCA) to reduce multidimensional data sets to lower dimensions for analysis.

Compute a high number of features and then apply a feature selection algorithm that chooses the most significant features of the training data for the given task.

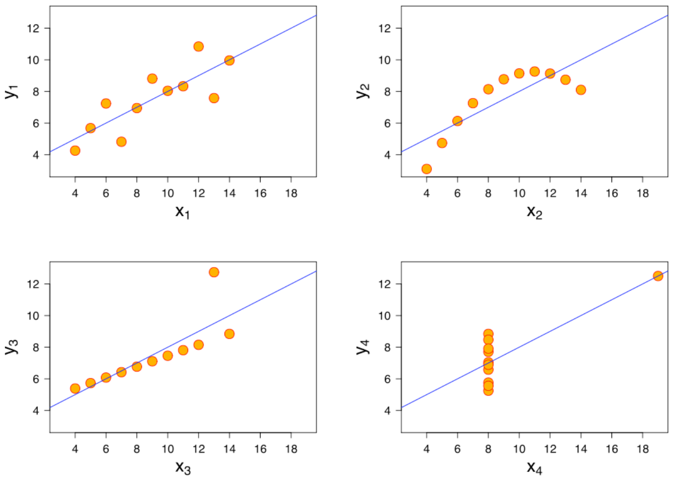
Audio – Feature extraction II

The raw pitch, energy, etc. contours can be used as is, and are then called short-term features, or more often, the actual features are derived from these acoustic variables by applying (statistic) functions over the sequence of values within an emotion segment, thus called global statistics features.

Example:

- mean pitch of a word or an utterance
- maximum, or minimum, etc. of the segment,
- Regression : estimating the relationships among variables.

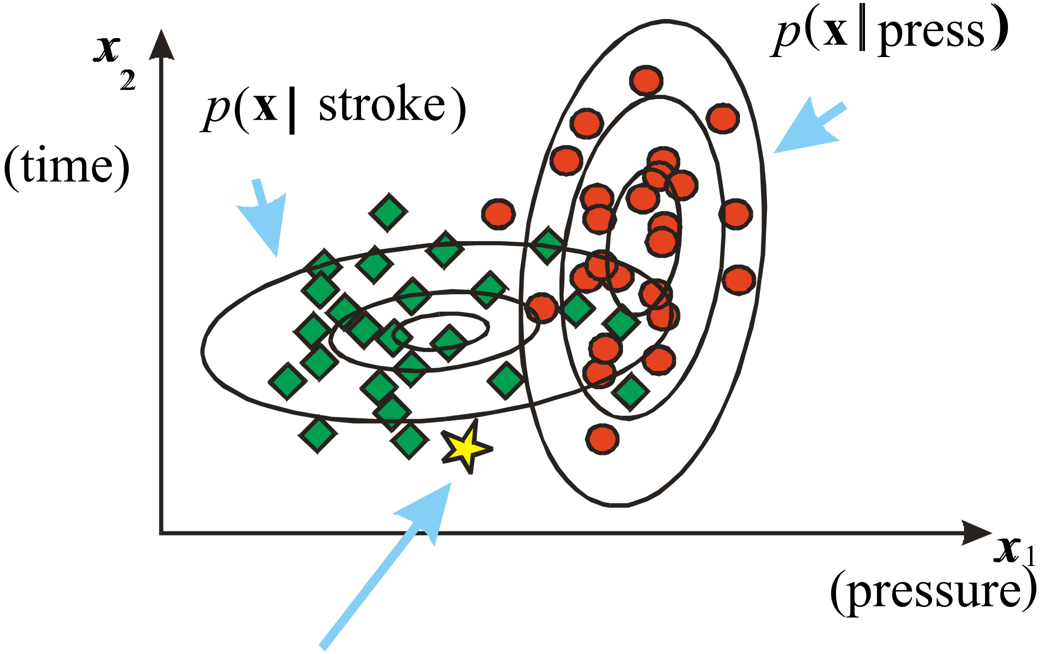
The focus is on the relationship between a **Dependent variable** and one or more **independent variables**. Regression analysis helps one understand how the typical value of the dependent variable changes when any one of the independent variables is varied, while the other independent variables are held fixed.



Audio – Classification

Each input unit is represented by a feature vector.

The problem of recognition is now a general data mining problem.
Any statistical classifier that can deal with high-dimensional data can be used.
as long as it achieves the following:



if $p(\text{stroke} \mid \mathbf{x}) > p(\text{press} \mid \mathbf{x})$ $\mathbf{x} \rightarrow$ stroke

else $\mathbf{x} \rightarrow$ press

Audio – Classification

4 principal approaches:

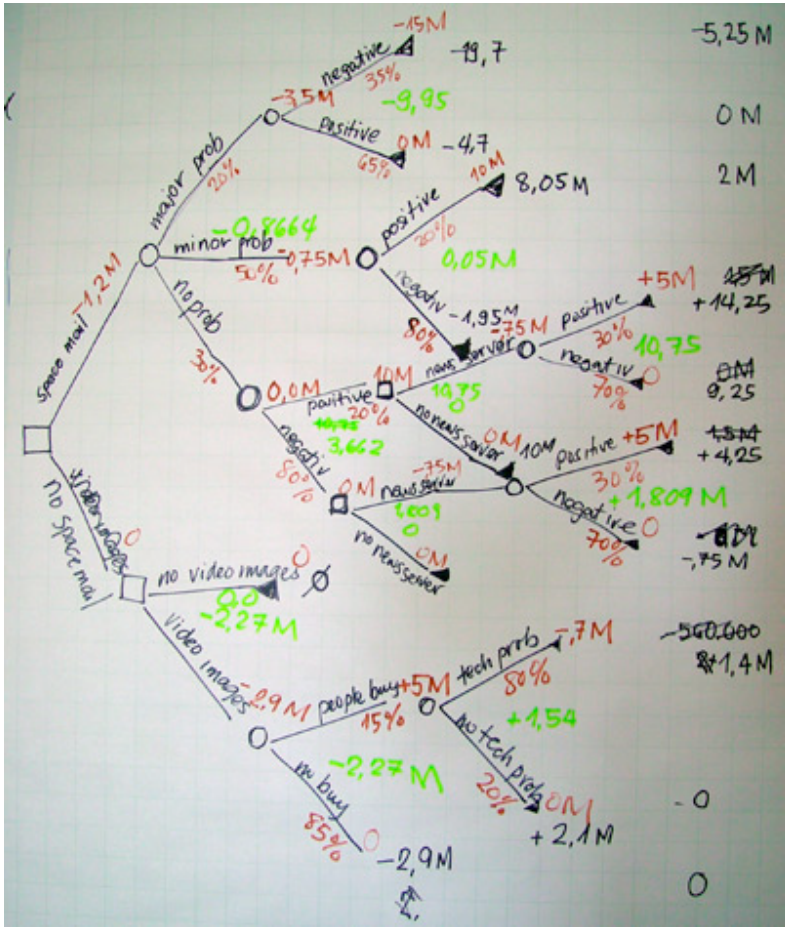
- Rule-base differentiation
- Similarity-based grouping
- Probability-based grouping
- Neural discrimination

Audio – Rule-base Categorisation

A **decision tree** is a support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility.

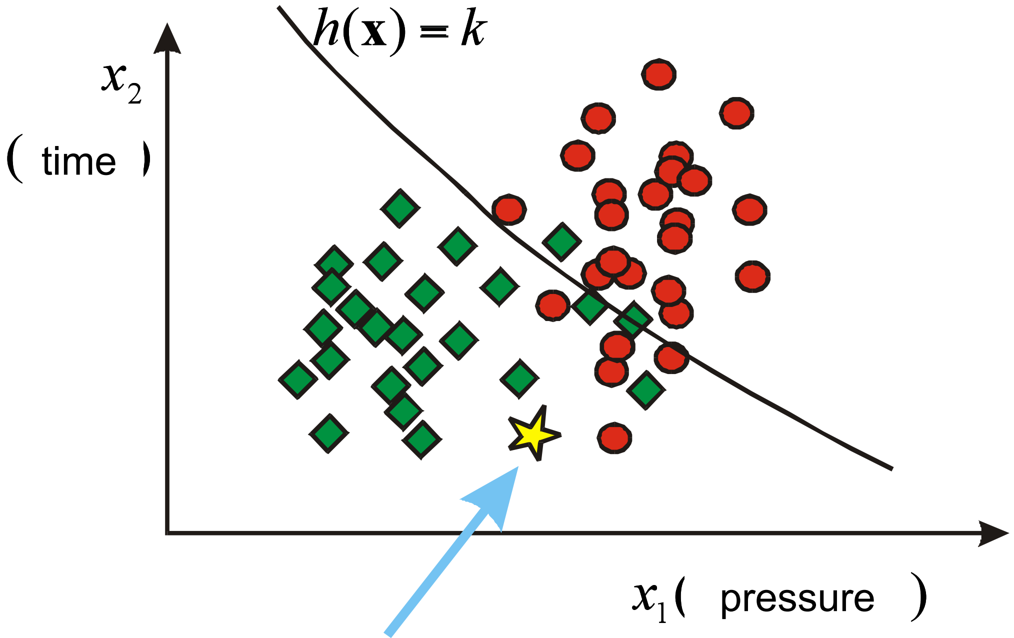
```
If f < e then
        follow left branch
else
        follow right branch
endif
```

f is a description element
e is the threshold



Audio – Distance-based classification

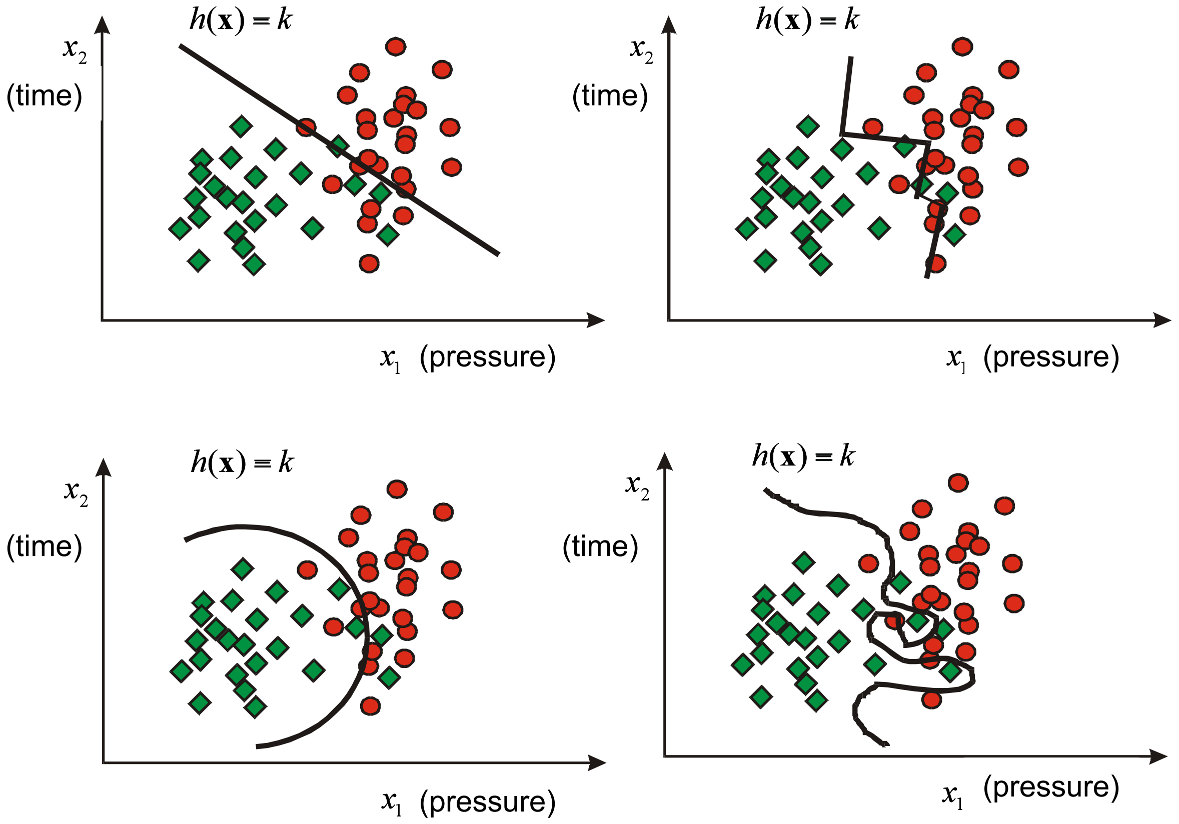
- Choose decision function in feature space.
- Estimate the function from the training set.
- Classify a new pattern on the basis of this decision rule.



$$\text{if } h(\mathbf{x}) > k \quad \mathbf{x} \Rightarrow \text{stroke}$$

$$< k \quad \mathbf{x} \rightarrow \text{press}$$

Audio – Distance-based classification



Linear

Piecewise linear

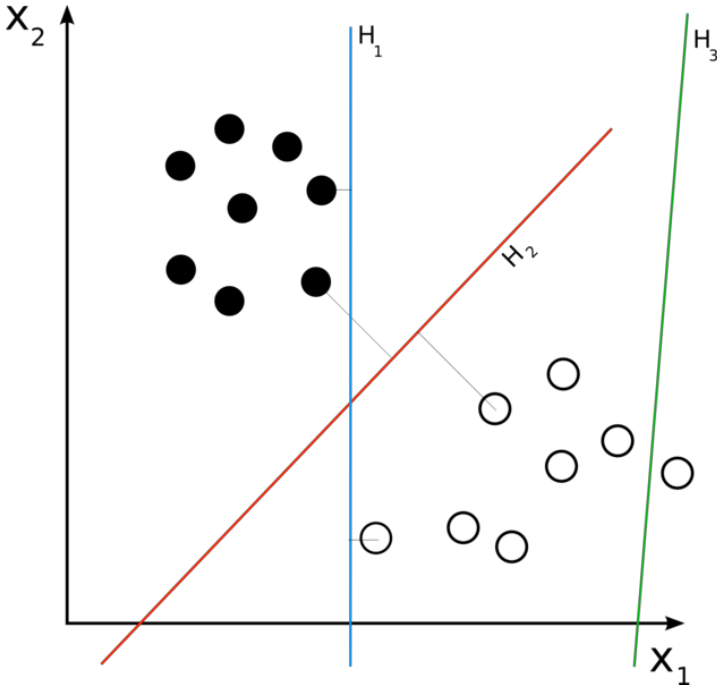
Polynomial

Feed-forward neural network

Audio – Distance-based classification

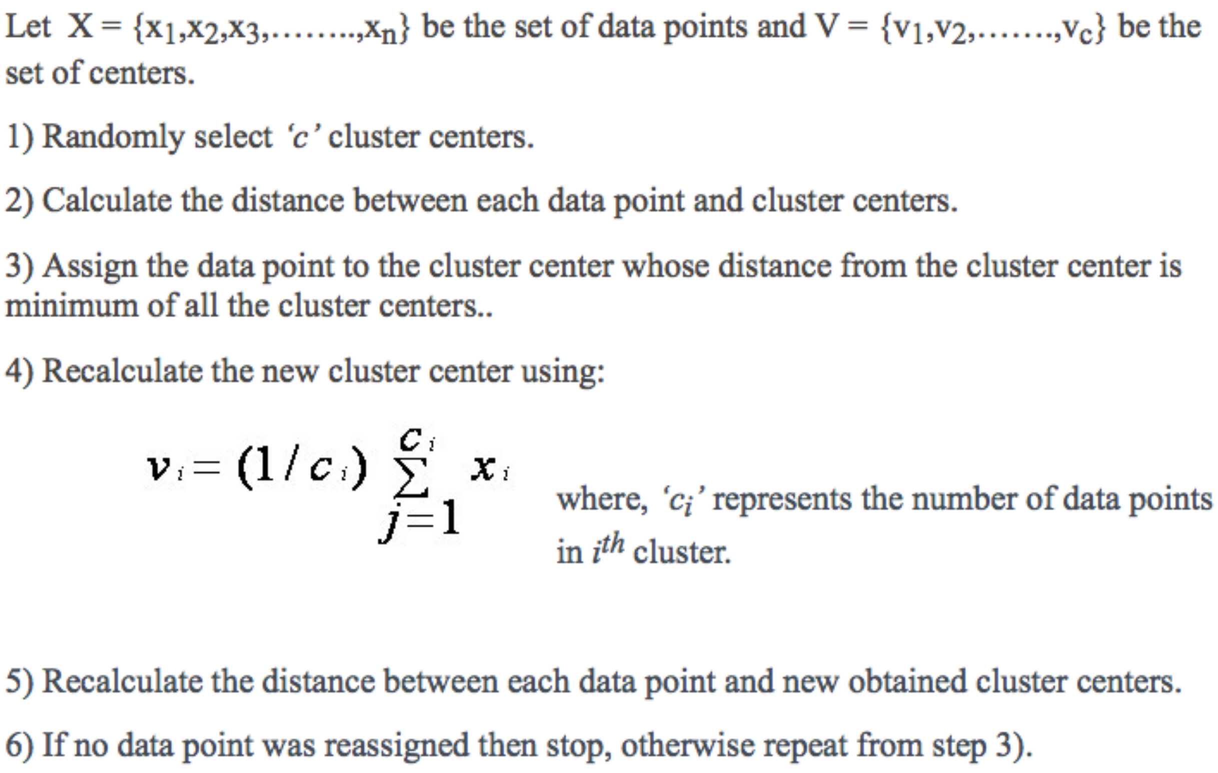
Support vector machines form a set of related supervised learning methods that analyze data and recognize patterns, used for classification and regression analysis.

H3 (green) doesn't separate the 2 classes. H1 (blue) does, with a small margin and H2 (red) with the maximum margin.



Audio – Distance-based classification

K-means clustering k-means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed apriori.



Let $X = \{x_1, x_2, x_3, \ldots\ldots.., x_n\}$ be the set of data points and $V = \{v_1, v_2, \ldots\ldots., v_c\}$ be the set of centers.

1) Randomly select *'c'* cluster centers.

2) Calculate the distance between each data point and cluster centers.

3) Assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers..

4) Recalculate the new cluster center using:

$$v_i = (1/c_i) \sum_{j=1}^{c_i} x_i$$

where, *'c_i'* represents the number of data points in i^{th} cluster.

5) Recalculate the distance between each data point and new obtained cluster centers.

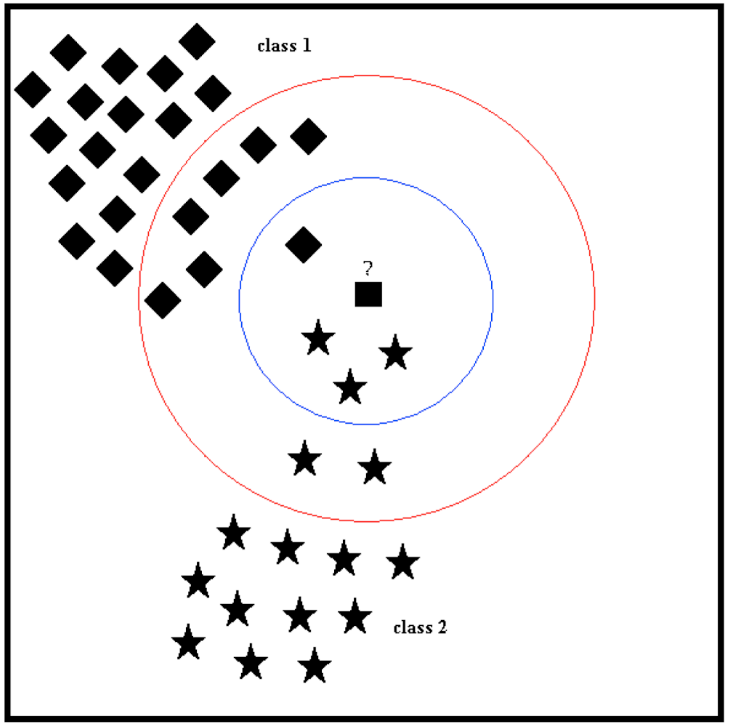
6) If no data point was reassigned then stop, otherwise repeat from step 3).

Audio – Distance-based classification

K-nearest neighbor the input consists of the *k* closest training examples in the feature space. The output depends on whether *k*-NN is used for classification or regression:

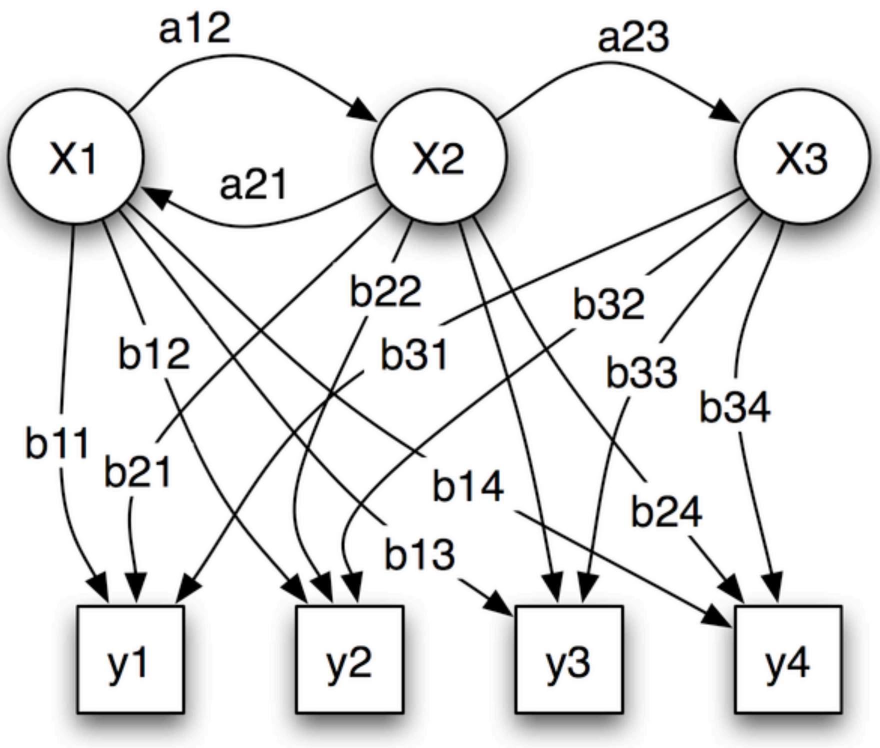
In *k-NN classification*, the output is a class membership. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its *k* nearest neighbors (*k* is a positive integer, typically small). If *k* = 1, then the object is simply assigned to the class of that single nearest neighbor.

In *k-NN regression*, the output is the property value for the object. This value is the average of the values of its *k* nearest neighbors.



Audio – Probability-based grouping

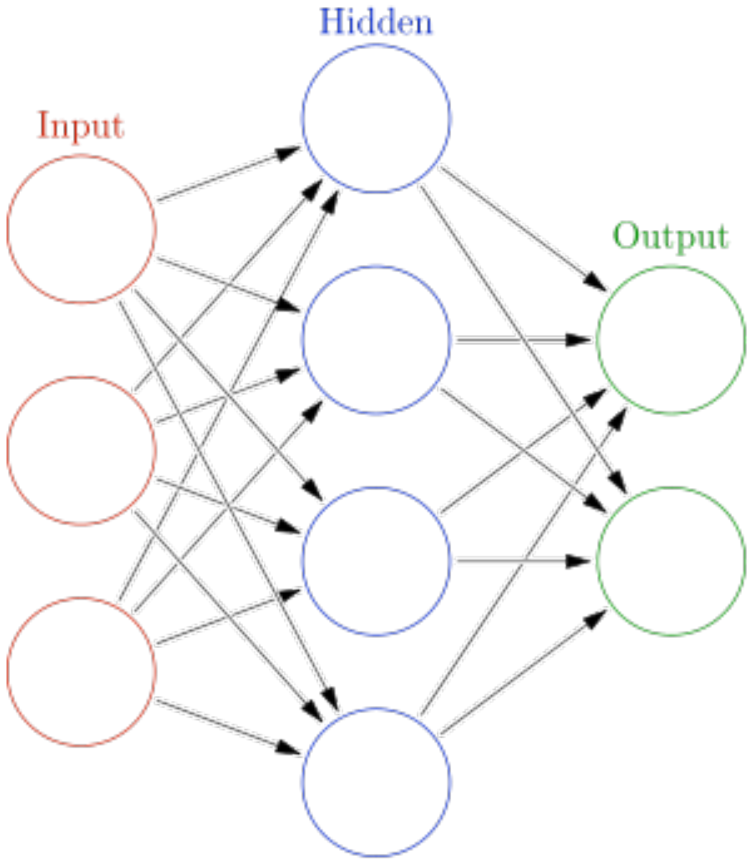
HMM is a statistical Markov model in which the system being modeled is assumed to be a Markov process with unobserved (*hidden*) states. An HMM can be considered as the simplest dynamic Bayesian network.

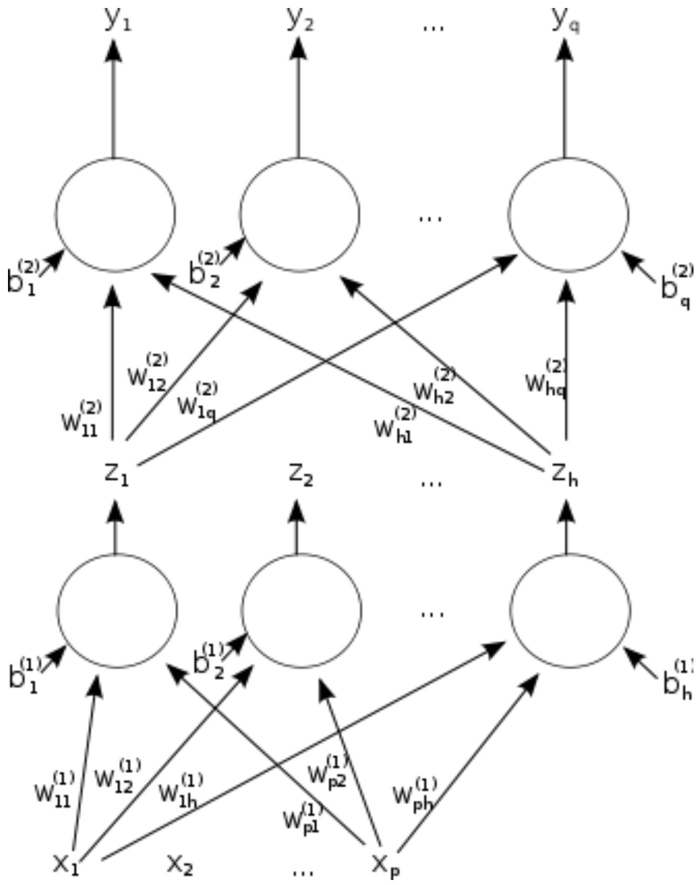


Probabilistic parameters of a hidden Markov model (example)
x — states
y — possible observations
a — state transition probabilities
b — output probabilities

Audio – Common statistical methods

Neural networks are adaptive systems that change their structure based on external or internal information that flows through the network during the learning phase.





Audio – Measurement - summary

- No standard database for benchmarking
- It is rarely possible to compare features across published work, since conditions vary a lot and even slight changes in the general set-up can make results incomparable.
- Audio segmentation can be performed by voice activity detection which is a fast segmentation method not requiring high-level linguistic knowledge.
- Feature extraction may only rely on automatically computable properties of the acoustic signal, but this is not a major limitation.
- For classification, in principle any statistical classifier can be used with sophisticated classifiers being superior in accuracy, but simple and thus fast classifiers are often sufficient.
- The speech database used to train the classifier should be adjusted to the particular application as much as possible.

Audio – summary

- Audition is a sign system but it is more complicated to identify the smallest sign unit, as the temporal nature of sound is essential.
- The properties of the acoustic signal are enough to identify a sound BUT individual characteristics are important (thresholds of frequency spaces)
- The interpretation of sound relies on context (other clues, such as facial expression or gestures for spoken language, are of importance)
- Audio stimulates primary emotions, though the identification for each individual depends on its thresholds for arousal and valence.

Audio – References

Audio – References

Murray, I., Arnott, J.: Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. Journal of the Acoustical Society of America 93(2) (1993,) 1097–1108

Plutchik, R. (1980) A general psychoevolutionary theory of emotion. In R. Plutchik & H. Kellermann (eds.), Emotion: Theory, research, and experience, Volume I: Theories of emotions, pp. 3 – 31, New york: Accademic Press.

Russell, J.A., 1980. A circumplex model of affect. Journal of Personality and Social Psychology 39(6), pp. 1161–1178, American Psychological Association.