11-751
Speech Recognition and Understanding

Phonetics and Phonology

Florian Metze

September 04, 2013

Carnegie Mellon

Agenda

- Introducing the TA – Yajie Miao, ymiao@cs.cmu.edu
- Phonetics, Phonology
- Words and beyond
- Units of classification
- Pattern Recognition

Phonetics: Phonemes and Phones

Phonetics vs. Phonology

Phonetics: Study of the individual production and classification and transcription of speech sounds

→ Focus is on the unique (since dependent on speaker, …) acoustic realization of speech sounds

Phonology: Study of the systematic distribution and patterning of speech sounds in a language, pronunciation

→ Focus is on finding gross characteristics of speech sounds that are adequate and required for description and classification of words (in a dictionary)

/phoneme/ vs. [phone]

Speech sounds do not have an inherent meaning – however we need to have a characteristics of speech sounds to describe and classify words and their pronunciation

Phoneme: a phoneme is the smallest speech unit which differentiates the meaning of a word pair (in a language)

- minimal pair like /bat/ vs. /pat/, thus
- /b/ /p/ are phonemes (linguistically distinct sounds)

Phone: a phoneme might have different acoustic realizations according to context, speaker, language, ...

- written as [b] and [p] …

Analogy to text coding

Analogy to the coding of text characters:

- The grapheme does not specify the size, shape, or orientation on the screen
- The phoneme does not specify the acoustic realization of a sound

Form	Genuine abstraction	Particular Realization
Text	Unicode U+0041 Grapheme	A,A,A,*A*, Glyphe ∀
Speech	/t/ Phoneme	[t] Phone

- The exact definition of a grapheme depends on the character code set
- The definition of a phoneme depends on the language, however even for the same language, linguists often disagree on the phoneme set
- We typically have phonetic dictionaries listing words and pronunciations

Not all variation is Phonetic

Phonology: linguistically discrete units

- May be a number of different ways to say them
- /r/ trill (Scottish or Spanish) vs US way

Phonetics vs Phonemics

- Phonetics: discrete units
- Phonemics: all sounds

/t/ in US English: becomes “flap”

- “water” / w ao t er /
- “water” / w ao dx er /

IPA Scheme for Consonants

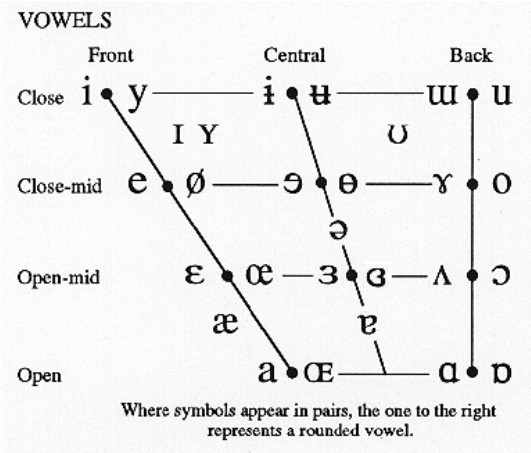
THE INTERNATIONAL PHONETIC ALPHABET (revised to 1993)

CONSONANTS (PULMONIC)

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k g	q ɢ		ʔ
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ		
Trill	ʙ			r					ʀ		
Tap or Flap				ɾ		ɽ					
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateral fricative				ɬ ɮ							
Approximant		ʋ		ɹ		ɻ	j	ɰ			
Lateral approximant				l		ɭ	ʎ	ʟ			

Where symbols appear in pairs, the one to the right represents a voiced consonant. Shaded areas denote articulations judged impossible.

IPA Scheme for Vowels



Vowels

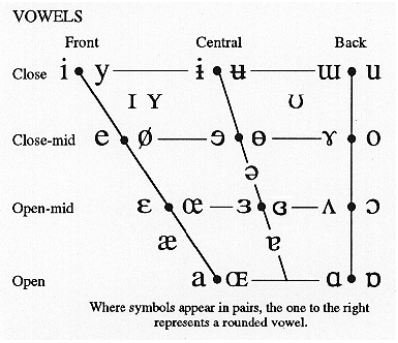
Vowel are characterized basically by three parameters, the tongue placement and shape and the lips shape

1. Vertical Position of the Tongue:
The higher the tongue is placed (vocal tract is more closed) the higher a vowel will sound.
E.g. [i:] in BEAT is higher (more closed) than [e] in BET.

2. Horizontal Position of the Tongue:
The more at the front the highest point of the tongue is, the "brighter" the vowel will sound.
E.g. [i:] in BEAT is a front vowel and [u:] in BOOT is a back vowel.

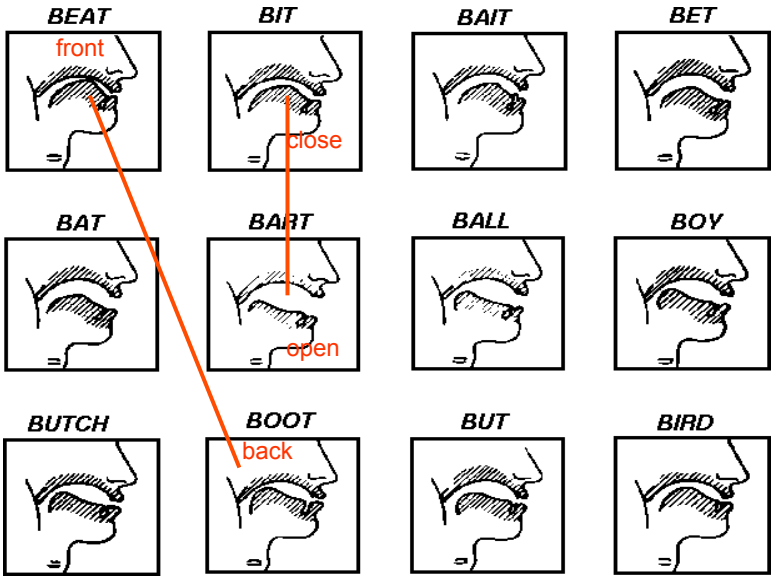
3. Shape of the Lips:
Depending on the shape of the lips, we call a vowel rounded or unrounded.
E.g. [æ] in BAT is unrounded and [o] in BOAT is rounded.

The Vowel-Quadrangle



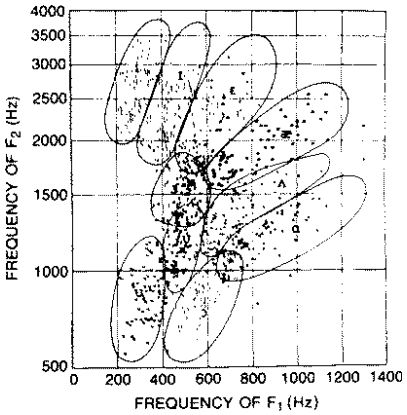
BEAT BIT BAIT BET BAT BOB BASIC ABOUT BIRD BOOT BOAT BUT BOUGHT BART

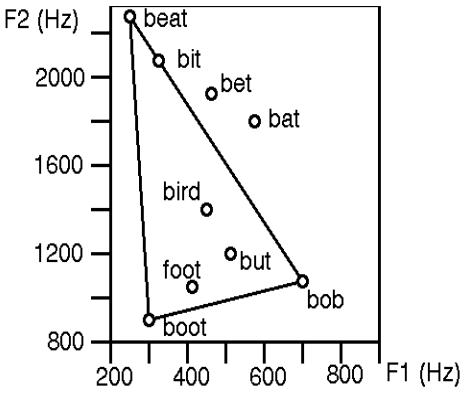
Different Shape of the Vocal Tract



Formants

The resonance frequencies of the vocal tract are called **formants**. In practice, only the first few formants are of interest.





Formants

F1: major resonance of the pharyngeal cavity

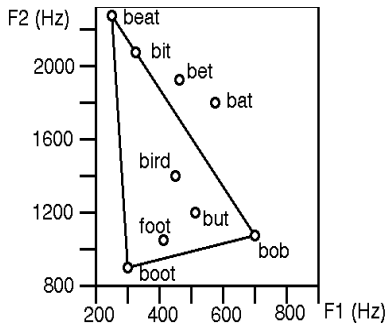
F2: major resonance of the oral cavity

The part from glottis to tongue (pharyngeal part) is longer than the forward part from the tongue (oral part), thus the resonance is lower

Tongue placement and oral tract shape determine F1 and F2

Example: *beat* tongue is far forward - longer pharyngeal part - lower F1, forward part extremely short - higher F2

Rounding the lips has the effect of extending the forward part thus lowering F2



Diphthongs

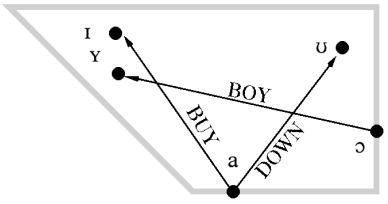
The characteristics F1 and F2 are sometimes called formant targets

- If a vowel has one specific target it is called mono-phthongs
- Vowels which combine two distinct sets of targets are called diphthongs
- Here the initial vowel target glides smoothly to the final configuration
- Some languages (like Mandarin) even have tri-phthongs

i.e.: /iy/ as in see /aa/ as in father: tongue moves from high to low

/iy/ /uw/ as in blue: movement from front to back

/iy/ /uw/: lip shape changes to round out



Consonants

As opposed to vowel, consonants are characterized by constriction or obstruction in the pharyngeal and/ or oral cavities

Consonants are classified by manner and place of articulation:

- **Manner of articulation:** refers to the articulation mechanism
- **Place:** refers to the location of the major constriction

Other characteristics:

- **Sonority:** continuous voicing - liquids (rat, lean), glides (yes) non-sonority requires (close to) complete obstruction
- **Voicing:** even non-sonorant consonants may have some voicing before the obstruction occurs in some consonants the vocal folds are vibrating thus making the sound voiced, e.g. Z/S, ZH/SH, B/P, D/T, G/K, V/F
- **Aspiration:** consonants can be aspirated (e.g. T in THOMAS), they can be strong (fortes) or weak (lenes)

English Consonants

	Labial	Labio-dental	Dental	Alveolar	Palatal	Velar	Glottal
Plosive	*p b*			*t d*		*k g*	*?*
Nasal	*m*			*n*		*ng*	
Fricative		*f v*	*th dh*	*s z*	*sh zh*		*h*
Retroflex Sonorant				*r*			
Lateral sonorant				*l*			
Glide	*w*				*y*		

English does not make full use of all possible mechanisms

Other languages require even more mechanisms

- Chinese: tonal language (Mandarin 4 tones + neutral)
- Japanese: vowel length is distinctive
- Spanish: trilled vs. implosive r

More about Phones, Syllables

Not all languages use the same set

- Aspirated stops (Korean, Hindi)
  - P vs PH
  - English uses both, but doesn't care
  - Pot vs sPot (place hand over mouth)
- L-R in Japanese not phonological
- US English dialects:
  - Mary, Merry, Marry
- Scottish English vs US English
- No distinction between "pull" and "pool"
- Distinction between: "for" and "four"

Not all variation is Phonetic

- Phonology: linguistically discrete units
  - May be a number of different ways to say them
  - /r/ trill (Scottish or Spanish) vs US way
- Phonetics vs Phonemics
  - Phonetics: discrete units
  - Phonemics: all sounds
- /t/ in US English: becomes "flap"
  - "water" / w ao t er /
  - "water" / w ao dx er /

Dialect and Idiolect

- Variation within language (and speakers)
- Phonetic
  - “Don” vs “Dawn”, “Cot” vs “Caught”
  - R deletion (Haavaad vs Harvard)
- Word choice:
  - Y’all, Yins
  - Politeness levels

Different language dimensions

- Vowel length
  - Bit vs beat
  - Japanese: shujin (husband) vs shuujin (prisoner)
- Tones
  - F0 (tune) used phonetically
  - Chinese, Thai, Burmese
- Clicks
  - Xhosa

Allophones

Phonemes are often modified in a systematic way by its phonetic neighborhood

This process is called coarticulation

- When the variation resulting from the coarticulatory process can be perceived, the modified phonemes are called allophones
- Allophonic differences are categorial, I.e. they can be understood and denoted by a small number of symbols

Example: /l/ in like (front part of tongue clearly touches the alveolar ridge) but /l/ in kill (tongue is often not touching any longer but stiffened in the mouth)

- Both are allophones of /l/ conditioned by the position in the syllable (initial vs final position)

Coarticulation

In continuously spoken speech (with varying speaking rate)

- Formant targets are less likely to be reached
- Stress patterns might be deleted
- Modification of sounds occur (assimilation)
- Sounds are completely deleted (elision)

Principle of efficiency

- Minimize the articulatory effort (but keep the information at its maximum)
- Increase speaking rate (speaker dependent)
- Reduce articulatory effort

BTW: Sounds within syllables influence one another‘s realization more than across syllable boundaries

Prosody

Besides regular intonation of each sound, a phrase can have its own melody.

The prosody carries information about:

- intention of the utterance (question, command, statement)
- stress (putting focus of attention to a specific part of it)
- resolving syntactic / semantic ambiguities
- describing the current mood / emotions of the speaker

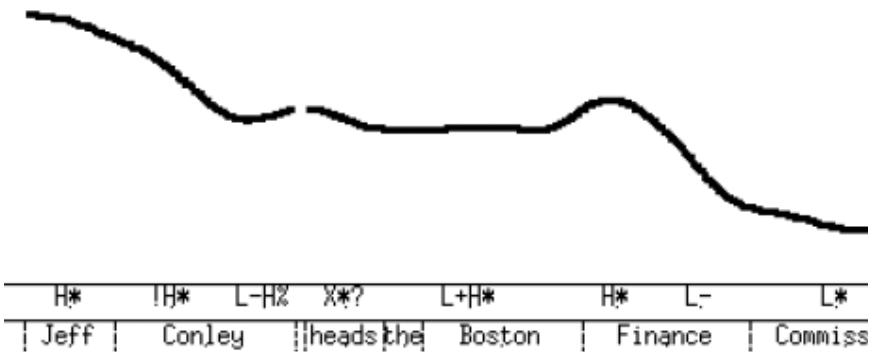
Enriching an utterance with prosodic information:

- intonation (pitch): produces a "melody"
- pauses: are used as markers for focus of attention or disambiguation
- stress: increase of loudness and pitch
- rhythm: the alternation of high power (sonorant) and low power sounds

Intonation (F0)

- Rate of vibration during voiced speech
  - Males: 80-140 times a second
  - Females: 130-220 times a second
  - Children: 180-320 times a second
- Used for:
  - Emphasis
  - Style: questions, statements, confidence etc

Intonation Contour



Spontaneously Spoken Speech

Used in every day live to communicate with other humans

- in opposite to planned speech
- speaking while thinking and vice versa

Example: "*I think we uhm we could meet maybe at hmm ah yeah maybe September 19th [pause] or uhm [laughter] in fact I have another mee- ah another meeting at this day so ah how about uhm how about September 21st.*"

- false starts, repetitions, hesitations, filled and non-filled pauses, non-verbal noises
- ill-formed grammar, sentences
- lots of coarticulation effects

Syllables and Words

- Phonemes are small blocks. They are easy to discriminate but they don‘t have a meaning by themselves
- In order to contribute to language meaning, they must be organized into longer cohesive spans
- The longer units must be combined in characteristic patterns to be meaningful
- The patterns might be different in structure and length depending on the language

Usually these patterns are: **Syllables and Words**

Syllables

In English syllables are centered around vowels (tom-cat); To split a word into syllables we have to judge about consonant affiliations

Done either by articulatory or perceptual criteria (unsolved!)

Syllable centers are thought of as peaks in sonority (high-amplitude, periodic sections of the speech waveform)

Ranking of decreasing sonority: stops, affricates, fricatives, approximants

Example: verbal: verb-al or ver-bal but not ve-rbal since approximant r - stop b is increasing, thus violates the decreasing criteria

As long as sonority conditions are met, affiliation of consonant is ambiguous

Unless higher order considerations of word structure like in beekeeper bee-keeper the structure blocks the affiliation but beaker could be either bea-ker or beak-er

Linguists define syllables as a unit with internal structure: It consists of an onset (initial consonants before the vowel peak - if any) and a rime. The rime consists of Nucleus (vowel peak) and Coda (trailing consonants)

Words, Syntax, and Semantics

Words

- In Indo-European languages the concept of word is intuitively obvious -- in the written form words are separated from each other by whitespaces
- Loosely defined as lexical item with a meaning (in a given community) that has the freedom of syntactic combination by its type (noun, verb, …)
- In languages like Japanese, Chinese or Thai no segmentation is given and the concept above is no longer unambiguous
- In spoken languages words are not marked by boundaries
- However, some phrases include pauses like
- ”Never give all the heart, for love“ = nevergiveallthe**heart // for**love“
- these units are intonation phrases

Lexical Part-of-Speech (POS)

Assigning of a word-type category to each word form in order to summarize syntactical or pragmatic facts

Typical set of POS categories:

- noun (refer to persons, places, things)
- verb (indicate relation between entities)
- adjective (specify noun references)
- adverb (specify verbal relations)
- interjection (express reaction)

content words

- conjunction (join phrases)
- determiner (narrow noun references)
- preposition (denote spatial and temporal relations)
- pronoun (substitute for introduced noun phrases)

function words

Word Classes: (Data-driven) process of grouping words together according to similarity of usage (semantic meaning) for LMing

Morphology

Morphology: patterns of word formation (inflection, derivation, compounds)

English morphology is relatively simple

- inflection: person and number agreement, tense marking
- derivation: productive pre- and suffixes re-, pre-, -ism, -ish, -ity ...
- compounds: usually max two roots are compounded

German (Compounds)

- Donau-dampf-schiffahrts-gesellschafts-kapitän
- The Captain of the Company that operates the Steamboats on the Donau River

Turkish (Inflection and Derivation)

- Osman-li-laç-tir-ama-yabil-ecek-ler-imiz-den-miş-siniz
- Behaving as if you were of those whom we might consider not converting into Ottoman

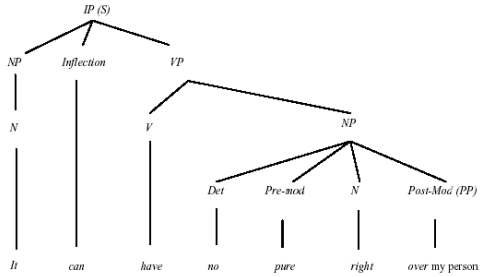
Syntax

Syntax is the study of the

- patterns of formation of sentences and phrases from words
- rules for the formation of grammatical sentences

Phrase schemata: create simple uniform template input from POS

Parse Tree Representation



Semantics

Semantic deals with the study of meaning (structure of meaning in language and changes in meaning over time)

- **Semantic Roles:** determine participants in an event to ask w-questions
- **Lexical Semantics:** is the level of meaning before words are composed into phrases and sentences

Practical problem Polysemy: context-dependent resolution of word sense; Example bank (river bank, money in the bank) [POS, mutual information, frequency analysis, a-priori]

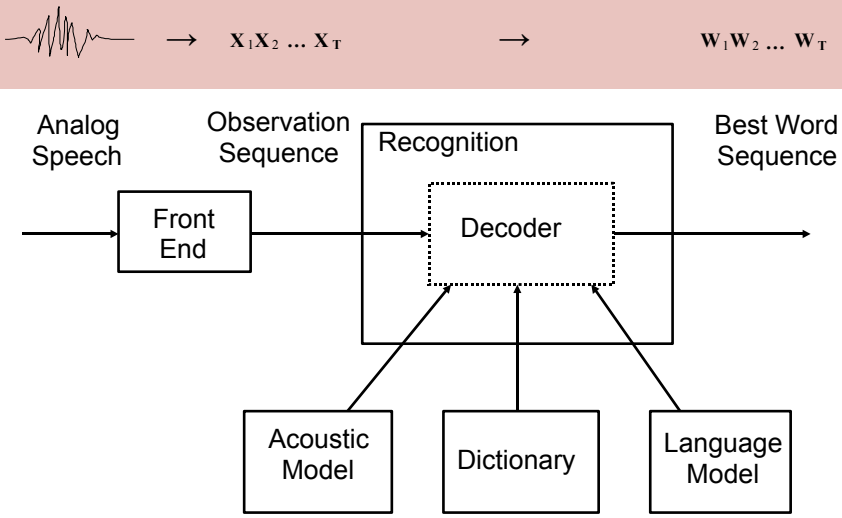
- **Logical Form:** To solve lexical, syntactic, and semantic ambiguities we need external context meta language, the like predicate logic are used to represent the logical form of language

Suggested Reading

- Jurafsky/ Martin: Chapter 7 “Phonetics” (Chapters 4, 5, 7 in 1st edition)
- Huang/ Acero/ Hon: Chapter 2.4 “Semantics”

Units of Modeling and Classification

Speech Recognition Components



Fundamental Problem of Speech Recognition

Given: an observation (ADC, FFT) $X = x_1, x_2, \ldots, x_T$

Wanted: the corresponding word sequence $W = w_1, w_2, \ldots, w_m$

Search: the most likely word sequence W'

$$W' = \arg\max_W P(W \mid X) = \arg\max_W \frac{p(X \mid W)P(W)}{p(X)} = \arg\max_W p(X \mid W)P(W)$$

(Bayes)

$p(X|W)$ = The **acoustic model**
(how likely is it to observe X when W is spoken)

$P(W)$ = The **language model**
(how likely is it that W is spoken a-priori)

Knowledge Sources

Fundamental Equation:

$$W' = \arg\max_W P(W \mid X) = \arg\max_W \frac{p(X \mid W)P(W)}{p(X)} = \arg\max_W p(X \mid W)P(W)$$

There exist several kinds of knowledge/ processing steps:

- **Signal Processing** deals with X
- A-priori knowledge in form of the **language model** $P(W)$
  - **Language Modeling** is dependent on domain, language, etc., but independent of concrete utterance
- Conditional probability in form of the **acoustic model** $p(X|W)$
  - **Acoustic Modeling** assumes know-how about domain, language, etc
- **Search** brings those two together
- **Dictionary** needs to be generated/ learned (as do the other models, see "phonetics and phonology" for more information)

Why Break Down Words into „Phones“

Need collection of reference patterns for each word

- High computational effort (esp. for large vocabularies), proportional to vocabulary size
- Large vocabulary also means: need huge amount of training data
- Difficult to train suitable references (or sets of references)
- Impossible to recognize untrained words

→ Replace whole words by **suitable sub-units**

Poor performance when the environment changes

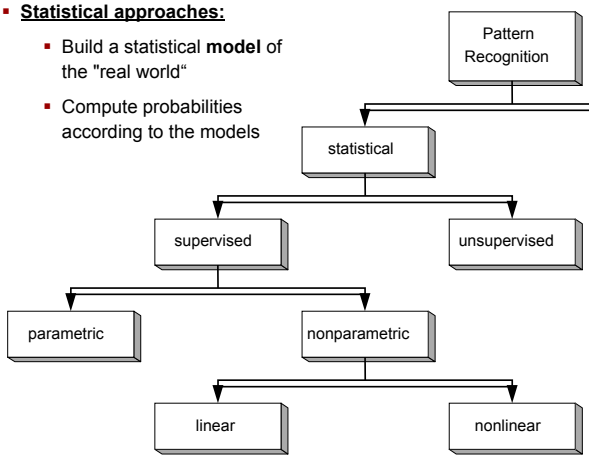
- Works only well for speaker-dependent recognition (variations)
- Unsuitable where speaker is unknown and no training is feasible
- Unsuitable for continuous speech (combinatorial explosion)
- Difficult to train/ recognize sub-word units

→ Replace the template approach by a better modeling process

Pattern Recognition

Pattern Recognition Approaches

- **Statistical approaches:**
  - Build a statistical **model** of the "real world“
  - Compute probabilities according to the models

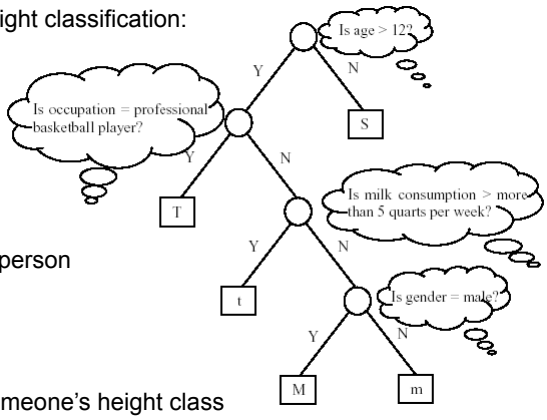


- Knowledge-based approaches:
  - Compile knowledge
  - Build decision trees
- Connectionist approaches:
  - Automatic knowledge acquisition, "black-box"
  - Simulation of biological processes

Knowledge Based: Decision Trees

Simple binary decision tree for height classification:

T =tall,
t =medium-tall,
M =medium,
m =medium-short,
S =short



Goal: Predict the height of a new person

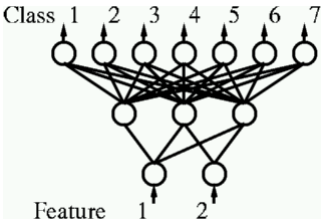
- Using decision tree to predict someone's height class by traversing the tree and answering the yes/ no questions
- Choice and order of questions is designed subjectively (knowledge-based)
- Classification and Regression Trees (CART) provide an automatic, data-driven framework to construct the decision process

Classification and Regression Trees (CART) Algorithm

1. Create a **set of questions Q** that consists of all possible questions
2. Pick a **splitting criterion** that can evaluate all possible questions
3. Create a tree with one root node consisting of all training samples
4. **Find the best composite question** for each terminal node
   (Goal is classification, so objective is to reduce uncertainty; use entropy H → find question which gives the greatest H reduction)
   1. generate a tree with several simple-question splits
   2. cluster leaf nodes into two classes according to the splitting criterion
   3. construct a corresponding composite question ($\wedge$,$\vee$)
5. **Split:** Take the split with the best criterion from step 4
6. **Stop criterion:** go to step 7 if all leaf nodes contain data samples from the same class or if the improvements of all potential splits fall below a defined threshold
7. Prune the tree into the optimal size using an independent test sample estimate or cross-validation to prevent the tree from over-modeling the training data (ensure generalization)

Connectionist: Neural Nets

- Parallel Supercomputers:
  ⇒ renewed interest in NN
- Appealing to ASR: parallel evaluation of many clues and facts
- Most common approach:
  Multi-layer Perceptron MLP



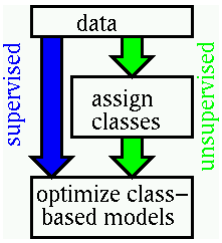
- NN: attempt real-time response and human-like performance
- Many simple processing elements operating in parallel
- Most common training procedure: **error back-propagation:**
  - Generalization of the MMSE (Minimum Mean Squared Error) algorithm
  - Gradient search: minimize difference between actual and wanted output
  - MLP approximates the a posteriori probabilities $P(\text{Class}|\text{Pattern})$
- Common problem: if an output of 0 0 0 ... 0 1 0 ... 0 0 0 is desired, the net tends to produce 0 0 0 ... 0 for all inputs

Statistical: Supervised vs Unsupervised

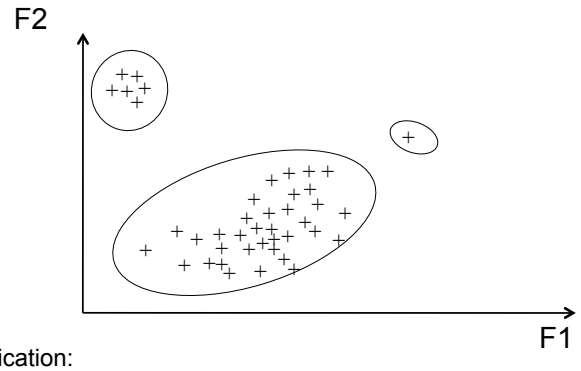
- **Supervised training:**
  Class to be recognized is known for each sample in training data.
  Requires a priori knowledge of useful features and knowledge/ labeling of each training token (cost!).

- **Unsupervised training:**
  Class is not known and structure is to be discovered automatically.
  Feature-space-reduction

  Example: clustering, auto-associative nets



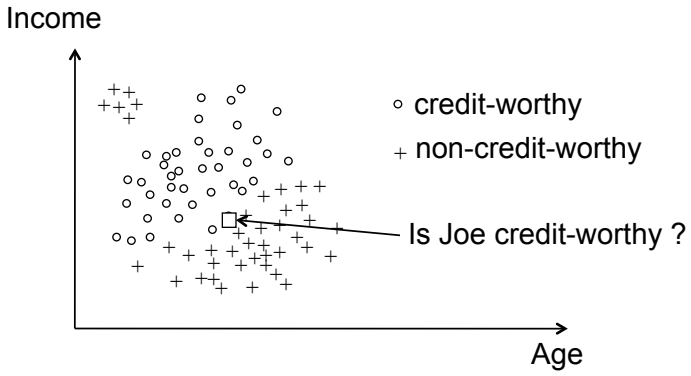
Unsupervised Training/ Classification



Classification:

- Classes not known, so find structure
- Clustering
- How to cluster? How many clusters?

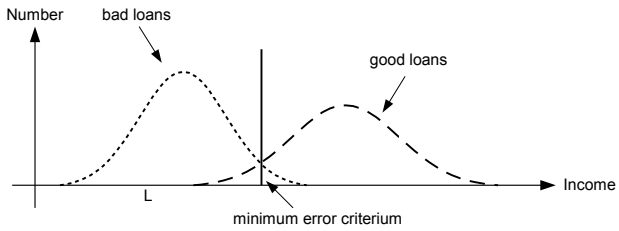
Supervised Classification



- Features: age, income $\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$
- Classes: creditworthy, non-creditworthy
- Problem: Given Joe's income and age, should a loan be made?
- Other Classification Problems: Fraud Detection, Customer Selection...

Parametric vs Non-Parametric

- **Parametric:**
  - Assume underlying probability distribution;
  - Estimate the parameters of this distribution.
  - Example: "Gaussian Classifier"



- Non-parametric:
  - Don't assume distribution.
  - Estimate probability of error or error criterion directly from training data.
  - Examples: Parzen Window, k-nearest neighbor, perceptron...

Bayes' Decision Theory

Bayes Rule: plays a central role for statistical pattern recognition

Concept of decision making based on:

1) prior knowledge of categories – prior probability: $P(\omega_j)$

AND

observation of x ↓

2) knowledge from observation data – posterior probability: $P(\omega_j / x)$

Bayes Rule: $$P(\omega_j / x) = \frac{p(x / \omega_j)P(\omega_j)}{p(x)}$$

where: $$p(x) = \sum_j p(x / \omega_j)P(\omega_j)$$

Class-conditional Probability Density function: $p(x / \omega_j)$
(referred to as the likelihood function – how likely is x generated)

Minimum Error Rate Decision Rule

$$P(\text{error} / x) = \begin{cases} P(\omega_1 / x) & \text{if we decide } \omega_2 \\ P(\omega_2 / x) & \text{else} \end{cases}$$

Classification error is minimized, if we:

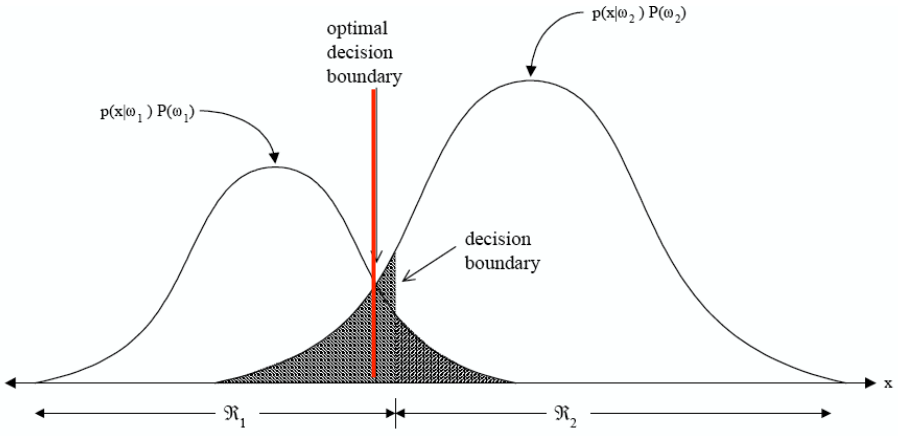
- decide ω_1 if $P(\omega_1 / x) > P(\omega_2 / x)$;
  ω_2 otherwise
- decide ω_1 if $p(x / \omega_1)P(\omega_1) > p(x / \omega_2)P(\omega_2)$;
  ω_2 otherwise

For the multi-category case:

Decide ω_i if $P(\omega_i / x) > P(\omega_j / x)$ for all $j \neq i$

Classification Error/ Bayes Decision Rule

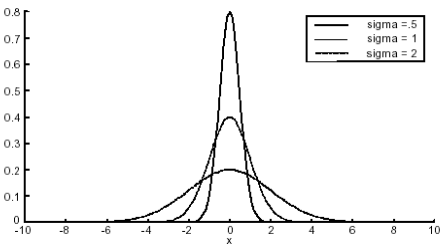
Move the decision boundary such that the decision is made to choose the class i based on the maximum value of $P(x|\omega_i)\, P(\omega_i)$. The tail integral area $P(\text{error})$ becomes minimum.



Classifier Design in Practice

- Need a priori probability $P(\varpi_i)$ (not too bad)
- Need class conditional PDF $p(x|\varpi_i)$
- Problems:
  - limited training data
  - limited computation
  - class-labeling potentially costly and prone to error
  - classes may not be known
  - good features not known
- Parametric Solution:
  - Assume that $p(x|\varpi_i)$ has a particular parametric form
  - Most common representative: multivariate normal density

Example: Gaussian Distributions



$$f(x\,|\,\mu,\sigma^2) = N(\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

Three Gaussian distributions with same mean but different variances (sigma)

Most important probability distribution, since random variables in physical experiments (incl. speech signals) have distributions which are approximately Gaussian:

"normal distribution"

X has a Gaussian distrib with mean μ and variance σ^2 if X has a continuous pdf of the form to the left

pdf = "probability density function"

Gaussian Densities

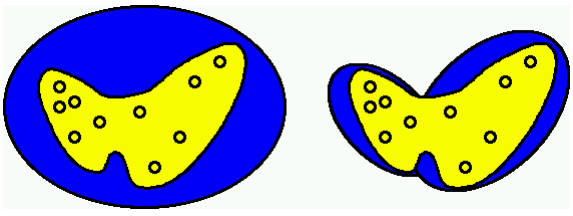
- The most often used model for speech signals are Gaussian densities.
- Often the "size" of the parameter spaces is measured in "number of densities“
- A multivariate Gaussian density looks like this:

$$\frac{1}{\sqrt{(2\pi)^d|\Sigma|}}e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}$$

- Its parameters are:
  - The **mean vector** μ (a vector with d coefficients)
  - The **covariance matrix** Σ (a symmetric $d*d$ matrix), if X indep., Σ is diagonal
  - The determinant of the cov. matrix | Σ | = $\frac{1}{\sqrt{(2\pi)^2 \cdot \left|\begin{pmatrix} a & b \\ b & d \end{pmatrix}\right|}} \cdot e^{-\frac{1}{2}\begin{pmatrix} x \\ y \end{pmatrix}^T \cdot \begin{pmatrix} a & b \\ b & d \end{pmatrix}^{-1} \cdot \begin{pmatrix} x \\ y \end{pmatrix}}$

Mixtures of Gaussian Densities

Often the shape of the set of vectors that belong to one class does not look like what can be modeled by a single Gaussian.
A (weighted) sum of Gaussians can approximate many more densities:



In general, a class can be modeled as a **mixture of Gaussians**:

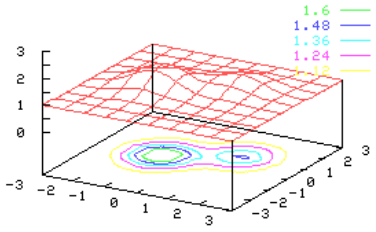
$$P(x_t|s_j) = \sum_{k=1}^{n_j} c_{jk} \cdot \frac{1}{\sqrt{(2\pi)^n|\Sigma_{jk}|}} e^{-\frac{1}{2}(x_t-\mu_{jk})^T\Sigma_{jk}^{-1}(x_t-\mu_{jk})}$$

Estimation of Parameters

- MLE, Maximum Likelihood Estimation, i.e. find the set of parameters that maximizes the likelihood of generating the observed data (see EM Algorithm)
- If $p(x|W)$ is assumed to be Gaussian, then W will be defined by the mean and the covariance matrix, which we need to estimate from data for each class i (or Gaussian j):

$$\hat{\mu} = \frac{1}{n}\sum_{k=1}^{n} \bar{x}_k$$

$$\hat{\Sigma} = \frac{1}{n}\sum_{k=1}^{n} (\bar{x}_k - \hat{\mu})(\bar{x}_k - \hat{\mu})^t$$

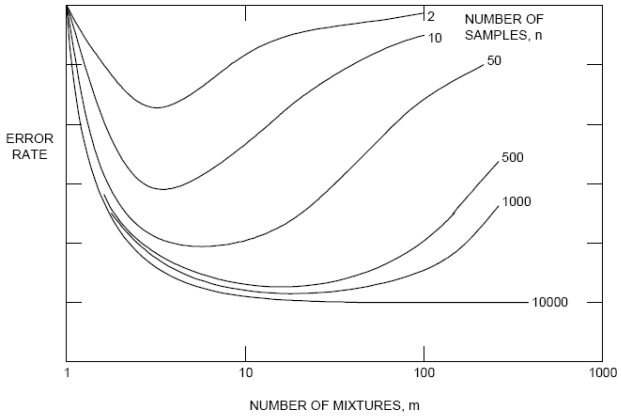


Problems of Parametric Classifier Design

- Speed of training? Evaluation?
- Database for Training and Development?
- Features:
  - What and how many features should be selected?
  - If additional features not useful (same mean and covariance), automatically ignore them?
- Adding more features
  - Adding independent features may help
  - BUT: adding indiscriminant features may lead to worse performance!
    - Training Data vs. Number of Parameters
    - Limited training data.
- Solution:
  - Select features carefully
  - Reduce dimensionality
  - Principle Component Analysis

Example of Trainability

- Two-phoneme classification example (Huang et al.), Phonemes modeled by Gaussian mixtures
- Parameters are trained with a varied set of training samples

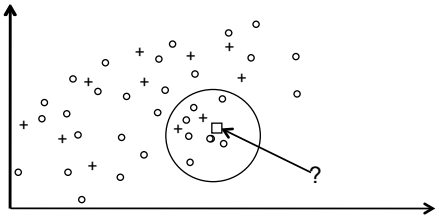


K-Nearest-Neighbour (KNN) Classifier

Idea: Let the volume be a function of the data. Include k-nearest neighbors in estimate and set $k = \sqrt{n}$; k-nearest neighbor rule for classification

To classify sample x:

- Find k-nearest neighbors of x.
- Determine the class most frequently represented among those k samples (take a vote)
- Assign x to that class.



k = 9
7 o
2 +
→ classify o

KNN Classifier Problem

For finite number of samples n, we want k to be:

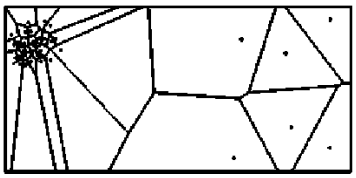
large for reliable estimate

small to guarantee that all k neighbors are reasonably close.

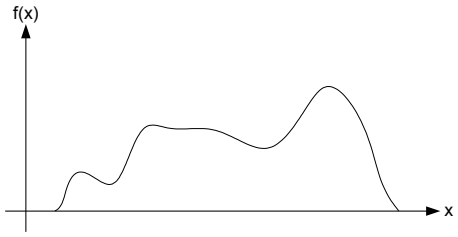
Need training database to be larger.

"There is no data like more data."

Also, classes have different variances:



Summary



- In Speech Processing, we usually assume a GMM (Gaussian Mixture Model) approximates the data distribution well enough
- Other densities may be mathematically intractable.
  → **non-parametric techniques**
- We use these models for classification based on Bayes criterion
- How do we get the models? → "Expectation Maximization Algorithm"

Extensions

- We've discussed (binary) classification – there's more
  - Evaluate probability density function (PDF) for each class on an input vector
  - Decide for class with biggest PDF (still Bayes)
- Often: probability expressed as cost $c = -\log(p)$
  - Advantage: numerical stability
    - High probability ➔ score near 0, low probability ➔ high score
  - Joint probability (product) of events ➔ sum of scores
- Has other consequences: can use distances between data points as cost

Extensions (2)

- Use distances/ cost, rather than probabilities
  - Take distances between vectors – sample and test in non-parametric algorithms
  - Can compare **data-model** (typically parametric), **data-data** (non-parametric), and **model-model** (KL divergence)
- "Curse of Dimensionality"
- Reading
  - Chap 3.1 "Probability Theory" in Huang/ Acero/ Hon
    - 3.3 "Significance Testing" often not done in speech, would be good, though
  - Chap. 4 "Pattern Recognition" in Huang/ Acero/ Hon
    - 4.3 "Discriminative Training" not needed yet

Backup

Suggested Reading so far

- Huang/ Acero/ Hon Book: (or equivalent)
  - Introduction (1.1, 1.2)
  - Spoken Language Structure (2.1, 2.2, 2.3)
  - Probability, Statistics, and Information Theory (3.1)
  - Pattern Recognition (4.1, 4.2)