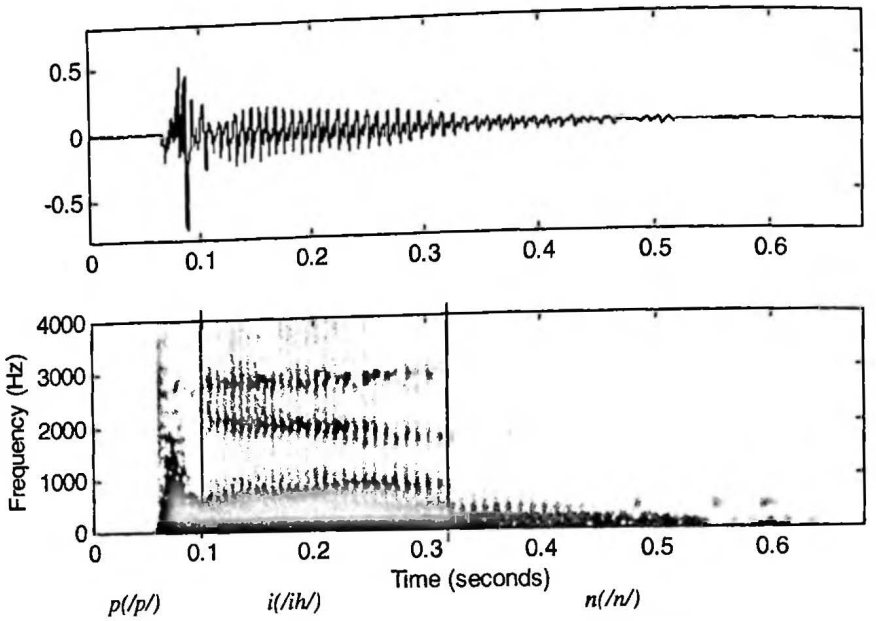

Figure 2.20 Spectrogram: stop release *burst* of */p/* in the word *pin*.

parison, the voicing of voiced plosive consonants may not always be obvious in a spectrogram.

A consonant that involves nearly complete blockage of some position in the oral cavity creates a narrow stream of turbulent air. The friction of this air stream creates a nonperiodic hiss-like effect. Sounds with this property are called fricatives and include */s, z/*. There is no voicing during the production of *s*, while there can be voicing (in addition to the frication noise), during the production of *z*, as discussed above. */s, z/* have a common place of articulation, as explained below, and thus form a natural similarity class. Though controversial, */h/* can also be thought of as a (glottal) fricative. */s/* in word-initial position and */z/* in word-final position are exemplified in Figure 2.5.

Some sounds are complex combinations of manners of articulation. For example, the *affricates* consist of a stop (e.g., */t/*), followed by a fricative [e.g., */sh/*) combining to make a unified sound with rapid phases of closure and continuancy (e.g., $\{t + sh\} = ch$ as in *church*). The affricates in English are the voiced/unvoiced pairs: */j/* $(d + zh)$ and */ch/* $(t + sh)$. The complete consonant inventory of English is shown in Table 2.9.

Consider the set */m/*, */n/*, */ng/* from Table 2.9. They are all voiced nasal consonants, yet they sound distinct to us. The difference lies in the location of the major constriction along the top of the oral cavity (from lips to velar area) that gives each consonant its unique quality. The articulator used to touch or approximate the given location is usually some spot along the length of the tongue. As shown in Figure 2.21, the combination of articulator and place of articulation gives each consonant its characteristic sound:

Table 2.9 Manner of articulation of English consonants.

Consonant Labels	Consonant Examples	Voiced?	Manner
b	**b**ig, able, ta**b**	+	plosive
p	**p**ut, open, ta**p**	-	plosive
d	**d**ig, idea, wa**d**	+	plosive
t	talk, sa**t**	-	plosive
g	**g**ut, angle, ta**g**	+	plosive
k	**c**ut, oa**k**en, ta**k**e	-	plosive
v	**v**at, over, ha**v**e	+	fricative
f	**f**ork, after, i**f**	-	fricative
z	za**z**, lazy, ha**z**e	+	fricative
s	sit, cast, toss	-	fricative
dh	**th**en, fa**th**er, scy**th**e	+	fricative
th	**th**in, no**th**ing, tru**th**	-	fricative
zh	genre, azure, beige	+	fricative
sh	**sh**e, cu**sh**ion, wa**sh**	-	fricative
jh	**j**oy, agile, e**dg**e	+	affricate
ch	**ch**in, ar**ch**er, ma**rch**	-	affricate
l	**l**id, el**b**ow, sa**il**	+	lateral
r	**r**ed, pa**r**t, fa**r**	+	retroflex
y	**y**acht, onion, **y**ard	+	glide
w	**w**ith, a**w**ay	+	glide
hh	**h**elp, a**h**ead, **h**otel	+	fricative
m	**m**at, a**m**id, a**im**	+	nasal
n	**n**o, e**n**d, pa**n**	+	nasal
ng	si**ng**, ang**er**, dri**nk**	+	nasal

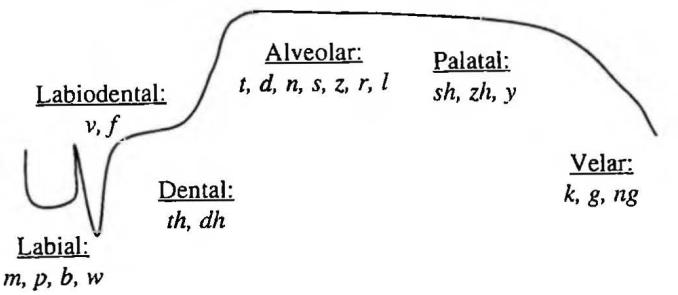

Figure 2.21 The major places of consonant articulation with respect to the human mouth.

- The *labial* consonants have their major constriction at the lips. This includes */p/, /b/* (these two differ only by manner of articulation) and */m/* and */w/*.
- The class of *dental or labio-dental* consonants includes */f, v/* and */th, dh/* (the members of these groups differ in manner, not place).
- *Alveolar* consonants bring the front part of the tongue, called the tip or the part behind the tip called the blade, into contact or approximation to the alveolar ridge, rising semi-vertically above and behind the teeth. These include */t, d, n, s, z, r, l/*. The members of this set again differ in manner of articulation (voicing, continuity, nasality), rather than place.
- *Palatal* consonants have approximation or constriction on or near the roof of the mouth, called the palate. The members include */sh, zh, y/*.
- *Velar* consonants bring the articulator (generally the back of the tongue), up to the rearmost top area of the oral cavity, near the velar flap. Velar consonants in English include */k, g/* (differing by voicing) and the nasal continuant */ng/*.

With the place terminology, we can complete the descriptive inventory of English consonants, arranged by manner (rows), place (columns), and voiceless/voiced (pairs in cells) as illustrated in Table 2.10.

Table 2.10 The consonants of English arranged by place (columns) and manner (rows).

	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*		

2.2.1.3. Phonetic Typology

The oral, nasal, pharyngeal, and glottal mechanisms actually make available a much wider range of effects than English happens to use. So, it is expected that other languages would utilize other vocal mechanisms, in an internally consistent but essentially arbitrary fashion, to represent their lexicons. In addition, often a vocal effect that is part of the systematic linguistic phonetics of one language is present in others in a less codified, but still perceptible, form. For example, Japanese vowels have a characteristic distinction of length that can be hard for non-natives to perceive and to use when learning the language. The words *kado* (*corner*) and *kaado* (*card*) are spectrally identical, differing only in that *kado* is much shorter

in all contexts. The existence of such minimally-contrasting pairs is taken as conclusive evidence that *length* is phonemically distinctive for Japanese. As noted above, what is linguistically distinctive in any one language is generally present as a less *meaningful* signaling dimension in other languages. Thus, vowel length can be manipulated in any English word as well, but this occurs either consciously for emphasis or humorous effect, or unconsciously and very predictably at clause and sentence end positions, rather than to signal lexical identity in all contexts, as in Japanese.

Other interesting sounds that the English language makes no linguistic use of include the trilled *r* sound and the implosive. The trilled *r* sound is found in Spanish, distinguishing (for example) the words *pero* (*but*) and *perro* (*dog*). This trill could be found in times past as a non-lexical sound used for emphasis and interest by American circus ringmasters and other showpersons.

While the world's languages have all the variety of manner of articulation exemplified above and a great deal more, the primary dimension lacking in English that is exploited by a large subset of the world's languages is pitch variation. Many of the huge language families of Asia and Africa are tonal, including all varieties of Chinese. A large number of other languages are not considered strictly tonal by linguistics, yet they make systematic use of pitch contrasts. These include Japanese and Swedish. To be considered tonal, a language should have lexical meaning contrasts cued by pitch, just as the lexical meaning contrast between *pig* and *big* is cued by a voicing distinction in English. For example, Mandarin Chinese has four primary tones (tones can have minor context-dependent variants just like ordinary phones, as well) as shown in Table 2.11.

Table 2.11 The contrastive tones of Mandarin Chinese.

Tone	Shape	Example	Chinese	Meaning
1	High level	*ma*	妈	mother
2	High rising	*ma*	麻	numb
3	Low rising	*ma*	马	horse
4	High falling	*ma*	骂	to scold

Though English does not make systematic use of pitch in its inventory of word contrasts, nevertheless, as we always see with any possible phonetic effect, pitch is systematically varied in English to signal a speaker's emotions, intentions, and attitudes, and it has some linguistic function in signaling grammatical structure as well. Pitch variation in English will be considered in more detail in Chapter 15.

2.2.2. The Allophone: Sound and Context

The vowel and consonant charts provide abstract symbols for the phonemes – major sound distinctions. Phonemic units should be correlated with potential meaning distinctions. For example, the change created by holding the tongue high and front (*/iy/*) vs. directly down

from the (frontal) position for */eh/*, in the consonant context */m _ n/*, corresponds to an important meaning distinction in the lexicon of English: *mean /m iy n/* vs. *men /m eh n/*. This meaning contrast, conditioned by a pair of rather similar sounds, in an identical context, justifies the inclusion of */iy/* and */eh/* as logically separate distinctions.

However, one of the fundamental, meaning-distinguishing sounds is often modified in some systematic way by its phonetic neighbors. The process by which neighboring sounds influence one another is called *coarticulation*. Sometimes, when the variations resulting from coarticulatory processes can be consciously perceived, the modified phonemes are called *allophones*. Allophonic differences are always *categorical*, that is, they can be understood and denoted by means of a small, bounded number of symbols or diacritics on the basic phoneme symbols.

As an experiment, say the word *like* to yourself. Feel the front of the tongue touching the alveolar ridge (cf. Figure 2.21) when realizing the initial phoneme */l/*. This is one allophone of */l/*, the so-called *light* or *clear /l/*. Now say *kill*. In this word, most English speakers will no longer feel the front part of the tongue touch the alveolar ridge. Rather, the */l/* is realized by stiffening the broad midsection of the tongue in the rear part of the mouth while the continuant airstream escapes laterally. This is another allophone of */l/*, conditioned by its syllable-final position, called the *dark /l/*. Predictable contextual effects on the realization of phones can be viewed as a nuisance for speech recognition, as will be discussed in Chapter 9. On the other hand, such variation, because it is systematic, could also serve as a cue to the syllable, word, and prosodic structure of speech.

Now experiment with the sound */p/* by holding a piece of tissue in front of your mouth while saying the word *pin* in a normal voice. Now repeat this experiment with *spin*. For most English speakers, the word *pin* produces a noticeable puff of air, called aspiration. But the same phoneme, */p/*, embedded in the consonant cluster */sp/* loses its aspiration (burst, see the lines bracketing the */p/* release in *pin* and *spin* in Figure 2.22), and because these two types of */p/* are in complementary distribution (completely determined by phonetic and syllabic context), the difference is considered allophonic.

Try to speak the word *bat* in a framing phrase *say bat again*. Now speak *say bad again*. Can you feel the length difference in the vowel */ae/*? A vowel before a voiced consonant, e.g., */d/*, seems typically longer than the same vowel before the unvoiced counterpart, in this case */t/*.

A sound phonemicized as */t/* or */d/*, that is, a stop made with the front part of the tongue, may be reduced to a quick tongue tap that has a different sound than either */t/* or */d/* in fuller contexts. This process is called flapping. It occurs when */t/* or */d/* closes a stressed vowel (coda position) followed by an unstressed vowel, as in: *bitter, batter, murder, quarter, humidity,* and can even occur across words as long as the preconditions are met, as in *you can say that again*. Sometimes the velar flap opens too soon (anticipation), giving a characteristically nasal quality to some pre-nasal vowels such as */ae/* in *ham* vs. *had*. We have a more detailed discussion on allophones in Chapter 9.

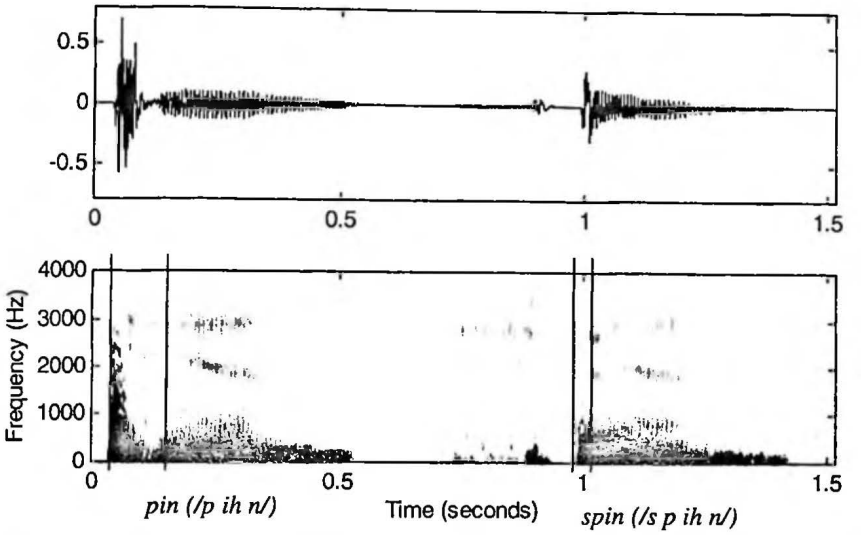

Figure 2.22 Spectrogram: bursts of *pin* and *spin*. The relative duration of a *p*-burst in different phonetic contexts is shown by the differing width of the area between the vertical lines.

2.2.3. Speech Rate and Coarticulation

In addition to allophones, there are other variations in speech for which no small set of established categories of variation can be established. These are *gradient*, existing along a scale for each relevant dimension, with speakers scattered widely. In general, it is harder to become consciously aware of coarticulation effects than of allophonic alternatives.

Individual speakers may vary their rates according to the content and setting of their speech, and there may be great inter-speaker differences as well. Some speakers may pause between every word, while others may speak hundreds of words per minute with barely a pause between sentences. At the faster rates, formant targets are less likely to be fully achieved. In addition, individual allophones may merge.

For example [20], consider the utterance *Did you hit it to Tom*? The pronunciation of this utterance is */d ih d y uw h ih t ih t t uw t aa m/*. However, a realistic, casual rendition of this sentence would appear as */d ih jh ax hh ih dx ih t ix t aa m/*, where */ix/* is a reduced schwa */ax/* that is short and often unvoiced, and */dx/* is a kind of shortened, indistinct stop, intermediate between */d/* and */t/*. The following five phonologic rules have operated on altering the pronunciation in the example:

- Palatalization of */d/* before */y/* in *did you*
- Reduction of unstressed */u/* to schwa in *you*

- Flapping of intervocalic /t/ in *hit it*
- Reduction of schwa and devoicing of /ʉ/ in *to*
- Reduction of geminate (double consonant) /t/ in *it to*

There are also coarticulatory influences in the spectral appearance of speech sounds, which can only be understood at the level of spectral analysis. For example, in vowels, consonant neighbors can have a big effect on formant trajectories near the boundary. Consider the differences in *F1* and *F2* in the vowel */eh/* as realized in words with different initial consonants *bet*, *debt*, and *get*, corresponding to the three major places of articulation (labial, alveolar, and velar), illustrated in Figure 2.23. You can see the different relative spreads of *F1* and *F2* following the initial stop consonants.

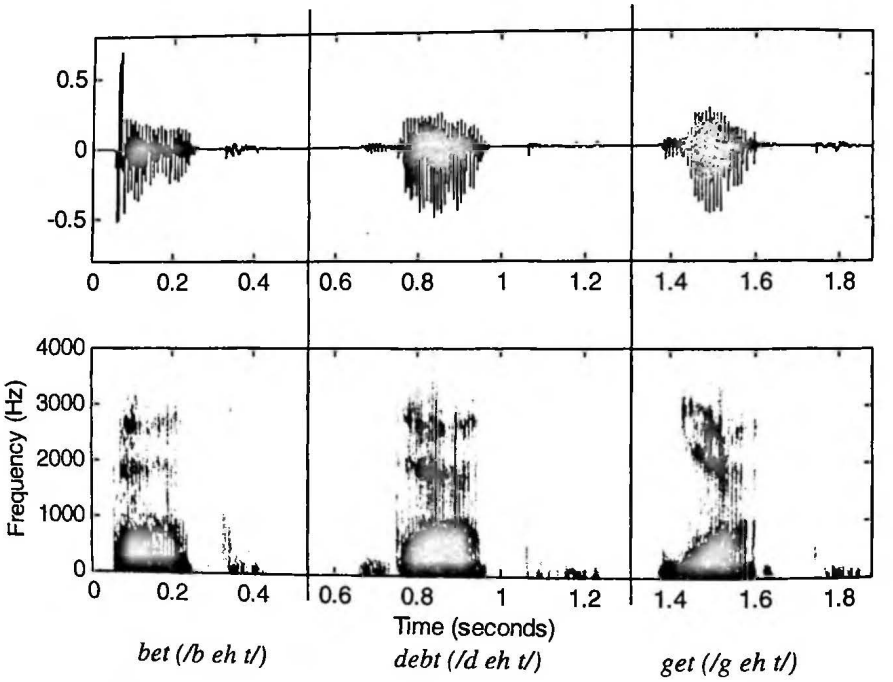

Figure 2.23 Spectrogram: *bet*, *debt*, and *get* (separated by vertical lines). Note different relative spreads of *F1* and *F2* following the initial stop consonants in each word.

Now let's see different consonants following the same vowel, *ebb*, *head*, and *egg*. In Figure 2.23, the coarticulatory effect is *perseverance*; i.e., in the early part of the vowel the articulators are still somewhat set from realization of the initial consonant. In the *ebb*, *head*, and *egg* examples shown in Figure 2.24, the coarticulatory effect is *anticipation*; i.e., in the latter part of the vowel the articulators are moving to prepare for the upcoming consonant articulation. You can see the increasing relative spread of *F1* and *F2* at the final vowel-consonant transition in each word.

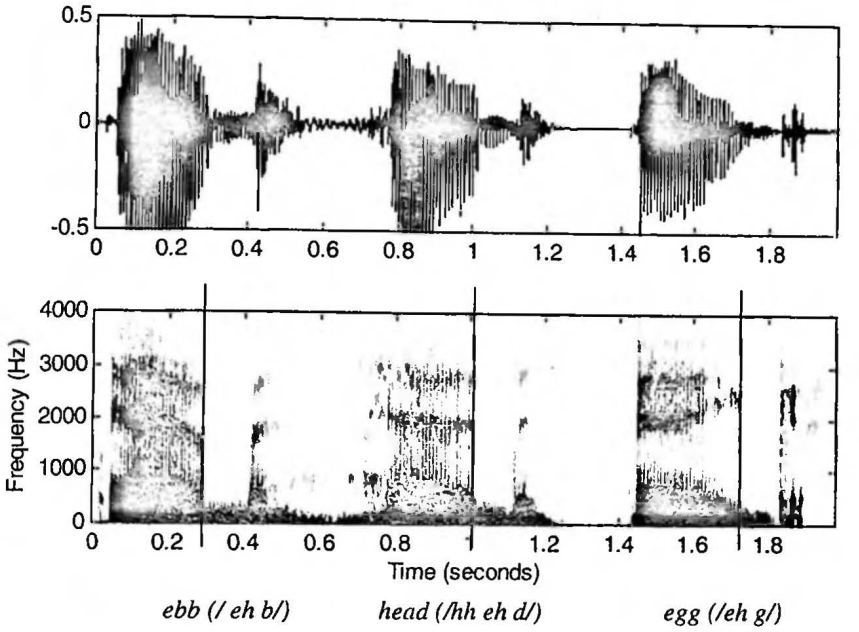

Figure 2.24 Spectrogram: *ebb*, *head*, and *egg*. Note the increasing relative spread of *F1* and *F2* at the final vowel-consonant transition in each word.

2.3. SYLLABLES AND WORDS

Phonemes are small building blocks. To contribute to language meaning, they must be organized into longer cohesive spans, and the units so formed must be combined in characteristic patterns to be meaningful, such as syllables and words in the English language.

2.3.1. Syllables

An intermediate unit, the *syllable*, is sometimes thought to interpose between the phones and the word level. The syllable is a slippery concept, with implications for both production and perception. Here we will treat it as a perceptual unit. Syllables are generally centered around vowels in English, giving two perceived syllables in a word like *tomcat*: */tOm-cAt/*. To completely parse a word into syllables requires making judgments of consonant affiliation (with the syllable peak vowels). The question of whether such judgments should be based on articulatory or perceptual criteria, and how they can be rigorously applied, remains unresolved.

Syllable centers can be thought of as *peaks* in sonority (high-amplitude, periodic sections of the speech waveform). These sonority peaks have affiliated *shoulders* of strictly non-increasing sonority. A scale of sonority can be used, ranking consonants along a continuum of stops, affricates, fricatives, and approximants. So, in a word like *verbal*, the syllabification would be *ver-bal*, or *verb-al*, but not *ve-rbal*, because putting the approximant /r/ before the stop /b/ in the second syllable would violate the non-decreasing sonority requirement heading into the syllable.

As long as the sonority conditions are met, the exact affiliation of a given consonant that could theoretically affiliate on either side can be ambiguous, unless determined by higher-order considerations of word structure, which may block affiliation. For example, in a word like *beekeeper*, an abstract boundary in the compound between the component words *bee* and *keeper* keeps us from accepting the syllable parse: *beek-eeper*, based on lexical interpretation. However, the same phonetic sequence in *beaker* could, depending on one's theory of syllabicity, permit affiliation of the *k: beak-er*. In general, the syllable is a unit that has intuitive plausibility but remains difficult to pin down precisely.

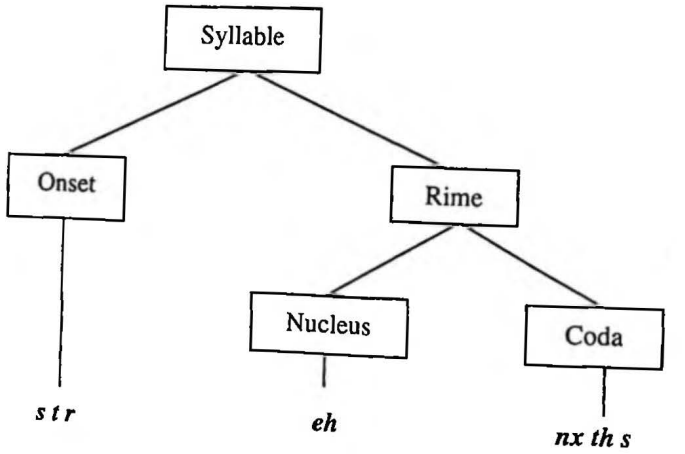

Figure 2.25 The word/syllable *strengths* (*/s t r eh nx th s/*) is the longest syllable of English.

Syllables are thought (by linguistic theorists) to have internal structure, and the terms used are worth knowing. Consider a big syllable such as *strengths /s t r eh nx th s/*. This consists of a vowel peak, called the *nucleus*, surrounded by the other sounds in characteristic positions. The *onset* consists of initial consonants if any, and the rime is the nucleus with trailing consonants (the part of the syllable that matters in determining poetic rhyme). The coda consists of consonants in the rime following the nucleus (in some treatments, the last consonant in a final cluster would belong to an *appendix*). This can be diagrammed as a syllable parse tree as shown in Figure 2.25. The syllable is sometimes thought to be the primary domain of coarticulation, that is, sounds within a syllable influence one another's realization more than the same sounds separated by a syllable boundary.

2.3.2. Words

The concept of words seems intuitively obvious to most speakers of Indo-European languages. It can be loosely defined as a lexical item, with an agreed-upon meaning in a given speech community, that has the freedom of syntactic combination allowed by its type (noun, verb, etc.).

In spoken language, there is a segmentation problem: words *run together* unless affected by a disfluency (unintended speech production problem) or by the deliberate placement of a pause (silence) for some structural or communicative reason. This is surprising to many people, because literacy has conditioned speakers/readers of Indo-European languages to expect a *blank space* between words on the printed page. But in speech, only a few true pauses (the aural equivalent of a blank space) may be present. So, what appears to the reading eye as *never give all the heart, for love* would appear to the ear, if we simply use letters to stand for their corresponding English speech sounds, as *nevergivealltheheart forlove* or, in phonemes, as *n eh v er g ih v ah l dh ax h aa r t \\ f ao r l ah v*. The \\ symbol marks a linguistically motivated pause, and the units so formed are sometimes called *intonation phrases*, as explained in Chapter 15.

Certain facts about word structure and combinatorial possibilities are evident to most native speakers and have been confirmed by decades of linguistic research. Some of these facts describe relations among words when considered in isolation, or concern groups of related words that seem intuitively similar along some dimension of form or meaning – these properties are *paradigmatic*. Paradigmatic properties of words include part-of-speech, inflectional and derivational morphology, and compound structure. Other properties of words concern their behavior and distribution when combined for communicative purposes in fully functioning utterances – these properties are *syntagmatic*.

2.3.2.1. Lexical Part-of-Speech

Lexical part-of-speech (POS) is a primitive form of linguistic theory that posits a restricted inventory of word-type categories, which capture generalizations of word forms and distributions. Assignment of a given POS specification to a word is a way of summarizing certain facts about its potential for syntagmatic combination. Additionally, paradigms of word formation processes are often similar within POS types and subtypes as well. The word properties upon which POS category assignments are based may include affixation behavior, very abstract semantic typologies, distributional patterns, compounding behavior, historical development, productivity and generalizabilty, and others.

A typical set of POS categories would include *noun*, *verb*, *adjective*, *adverb*, *interjection*, *conjunction*, *determiner*, *preposition*, and *pronoun*. Of these, we can observe that certain classes of words consist of infinitely large membership. This means new members can be added at any time. For example, the category of noun is constantly expanded to accommodate new inventions, such as Velcro or Spandex. New individuals are constantly being born, and their names are a type of noun called *proper noun*. The proliferation of words us-

ing the descriptive prefix *cyber* is another recent set of examples: *cyberscofflaw, cybersex,* and even *cyberia* illustrate the infinite creativity of humans in manipulating word structure to express new shades of meaning, frequently by analogy with, and using fragments of, existing vocabulary. Another example is the neologism *sheeple*, a noun combining the forms and meanings of *sheep* and *people* to refer to large masses of people who lack the capacity or willingness to take independent action. We can create new words whenever we like, but they had best fall within the predictable paradigmatic and syntagmatic patterns of use summarized by the existing POS generalizations, or there will be little hope of their adoption by any other speaker. These open POS categories are listed in Table 2.12. Nouns are inherently referential. They refer to persons, places, and things. Verbs are predicative; they indicate relations between entities and properties of entities, including participation in events. Adjectives typically describe and more completely specify noun reference, while adverbs describe, intensify, and more completely specify verbal relations. Open-class words are sometimes called *content* words, for their referential properties.

Table 2.12 Open POS categories.

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

In contrast to the open-class categories, certain other categories of words only rarely and very slowly admit new members over the history of English development. These closed POS categories are shown in Table 2.13. The closed-category words are fairly stable over time. Conjunctions are used to join larger syntactically complete phrases. Determiners help to narrow noun reference possibilities. Prepositions denote common spatial and temporal relations of objects and actions to one another. Pronouns provide a convenient substitute for noun phrases that are fully understood from context. These words denote grammatical relations of other words to one another and fundamental properties of the world and how humans understand it. They can, of course, change slowly; for example, the Middle English pronoun *thee* is no longer in common use. The closed-class words are sometimes called *function* words.

Table 2.13 Closed POS categories.

Tag	Description	Function	Example
Conj	Conjunction	Coordinates phrases	*and*
Det	Determiner	Indicates definiteness	*the*
Prep	Preposition	Relations of time, space, direction	*from*
Pron	Pronoun	Simplified reference	*she*

The set of POS categories can be extended indefinitely. Examples can be drawn from the Penn Treebank project (http://www.cis.upenn.edu/ldc) as shown in Table 2.14, where you can find the proliferation of sub-categories, such as *Verb, base form* and *Verb, past tense*. These categories incorporate *morphological* attributes of words into the POS label system discussed in Section 2.3.2.2.

Table 2.14 Treebank POS categories – an expanded inventory.

String	Description	Example
CC	Coordinating conjunction	and
CD	Cardinal number	two
DT	Determiner	the
EX	Existential *there*	there (*There was an old lady*)
FW	Foreign word	omerta
IN	Preposition, subord. conjunction	over, but
JJ	Adjective	yellow
JJR	Adjective, comparative	better
JJS	Adjective, superlative	best
LS	List item marker	
MD	Modal	might
NN	Noun, singular or mass	rock, water
NNS	Noun, plural	rocks
NNP	Proper noun, singular	Joe
NNPS	Proper noun, plural	Red Guards
PDT	Predeterminer	all (*all the girls*)
POS	Possessive ending	's
PRP	Personal pronoun	I
PRP$	Possessive pronoun	mine
RB	Adverb	quickly
RBR	Adverb, comparative	higher (*shares closed higher.*)
RBS	Adverb, superlative	highest (*he jumped highest of all.*)
RP	Particle	up (*take up the cause*)
TO	*to*	to
UH	Interjection	hey!
VB	Verb, base form	choose
VBD	Verb, past tense	chose
VBG	Verb, gerund, or present participle	choosing
VBN	Verb, past participle	chosen
VBP	Verb, non-third person sing. present	jump
VBZ	Verb, third person singular present	jumps
WDT	Wh-determiner	which
WP	Wh-pronoun	who
WP$	Possessive wh-pronoun	whose
WRB	Wh-adverb	when (*When he came, it was late.*)

POS tagging is the process of assigning a part-of-speech or other lexical class marker to each word in a corpus. There are many algorithms to automatically tag input sentences into a set of tags. Rule-based methods [45], hidden Markov models (see Chapter 8) [23, 29, 46], and machine-learning methods [6] are used for this purpose.

2.3.2.2. Morphology

Morphology is about the subparts of words, i.e., the patterns of word formation including inflection, derivation, and the formation of compounds. English mainly uses prefixes and suffixes to express *inflection* and *derivational* morphology.

Inflectional morphology deals with variations in word form that reflect the contextual situation of a word in phrase or sentence syntax, and that rarely have direct effect on interpretation of the fundamental meaning expressed by the word. English inflectional morphology is relatively simple and includes person and number agreement and tense markings only. The variation in *cats* (vs. *cat*) is an example. The plural form is used to refer to an indefinite number of cats greater than one, depending on a particular situation. But the basic POS category (*noun*) and the basic meaning (*felis domesticus*) are not substantially affected. Words related to a common lemma via inflectional morphology are said to belong to a common paradigm, with a single POS category assignment. In English, common paradigm types include the verbal set of affixes (pieces of words): *-s, -ed, -ing*; the noun set: *-s*; and the adjectival *-er, -est*. Note that sometimes the base form may change spelling under affixation, complicating the job of automatic textual analysis methods. For historical reasons, certain paradigms may consist of highly idiosyncratic irregular variation as well, e.g., *go, going, went, gone* or *child, children*. Furthermore, some words may belong to defective paradigms, where only the singular (noun: *equipment*) or the plural (noun: *scissors*) is provided for.

In *derivational morphology*, a given root word may serve as the source for wholly new words, often with POS changes as illustrated in Table 2.15. For example, the terms *racial* and *racist*, though presumably based on a single root word *race*, have different POS possibilities (*adjective* vs. *noun-adjective*) and meanings. Derivational processes may induce pronunciation change or stress shift (e.g., *electric* vs. *electricity*). In English, typical derivational affixes (pieces of words) that are highly productive include prefixes and suffixes: *re-, pre-, -ial, -ism, -ish, -ity, -tion, -ness, -ment, -ious, -ify, -ize*, and others. In many cases, these can be added successively to create a complex layered form.

Table 2.15 Examples of stems and their related forms across POS categories.

Noun	Verb	Adjective	Adverb
criticism	*criticize*	*critical*	*critically*
fool	*fool*	*foolish*	*foolishly*
industry, industrialization	*industrialize*	*industrial,industrious*	*industriously*
employ, employee, employer	*employ*	*employable*	*employably*
certification	*certify*	*certifiable*	*certifiably*

Generally, word formation operates in layers, according to a kind of word syntax: *(deriv-prefix)* **root** (root)* (deriv-suffix)* (infl-suffix).* This means that one or more *roots* can be compounded in the inner layer, with one or more optional *derivational prefixes*, followed by any number of optional *derivational suffixes*, capped off with no more than one *inflectional suffix*. There are, of course, limits on word formation, deriving both from semantics of the component words and simple lack of imagination. An example of a nearly maximal word in English might be *autocyberconceptualizations*, meaning (perhaps!) multiple instances of automatically creating computer-related concepts. This word lacks only compounding to be truly maximal. This word has a derivational prefix *auto-*, two root forms compounded (*cyber* and *concept*, though some may prefer to analyze *cyber-* as a prefix), three derivational suffixes (*-ual, -ize, -ation*), and is capped off with the plural inflectional suffix for nouns, *-s*.

2.3.2.3. Word Classes

POS classes are based on traditional grammatical and lexical analysis. With improved computational resources, it has become possible to examine words in context and assign words to groups according to their actual behavior in real text and speech from a statistical point of view. These kinds of classifications can be used in language modeling experiments for speech recognition, text analysis for text-to-speech synthesis, and other purposes.

One of the main advantages of word classification is its potential to derive more refined classes than traditional POS, while only rarely actually crossing traditional POS group boundaries. Such a system may group words automatically according to the similarity of usage with respect to their word neighbors. Consider classes automatically found by the classification algorithms of Brown *et al.* [7]:

```
{Friday Monday Thursday Wednesday Tuesday Saturday Sunday weekends}
{great big vast sudden mere sheer gigantic lifelong scant colossal}
{down backwards ashore sideways southward northward overboard aloft adrift}
{mother wife father son husband brother daughter sister boss uncle}
{John George James Bob Robert Paul William Jim David Mike}
{feet miles pounds degrees inches barrels tons acres meters bytes}
```

You can see that words are grouped together based on the semantic meaning, which is different from word classes created purely from syntactic point of view. Other types of classification are also possible, some of which can identify semantic relatedness across traditional POS categories. Some of the groups derived from this approach may include follows:

```
{problems problem solution solve analyzed solved solving}
{write writes writing written wrote pen}
{question questions asking answer answers answering}
{published publication author publish writer titled}
```

2.4. SYNTAX AND SEMANTICS

Syntax is the study of the patterns of formation of sentences and phrases from words and the rules for the formation of grammatical sentences. Semantics is another branch of linguistics dealing with the study of meaning, including the ways meaning is structured in language and changes in meaning and form over time.

2.4.1. Syntactic Constituents

Constituents represent the way a sentence can be divided into its grammatical subparts as constrained by common grammatical patterns (which implicitly incorporate normative judgments on acceptability). Syntactic constituents at least respect, and at best explain, the linear order of words in utterances and text. In this discussion, we will not strictly follow any of the many theories of syntax but will instead bring out a few basic ideas common to many approaches. We will not attempt anything like a complete presentation of the grammar of English but instead focus on a few simple phenomena.

Most work in syntactic theory has adopted machinery from traditional grammatical work on written language. Rather than analyze toy sentences, let's consider what kinds of superficial syntactic patterns are lurking in a random chunk of serious English text, excerpted from David Thoreau's essay *Civil Disobedience* [43]:

The authority of government, even such as I am willing to submit to – for I will cheerfully obey those who know and can do better than I, and in many things even those who neither know nor can do so well – is still an impure one: to be strictly just, it must have the sanction and consent of the governed. It can have no pure right over my person and property but what I concede to it. The progress from an absolute to a limited monarchy, from a limited monarchy to a democracy, is a progress toward a true respect for the individual.

2.4.1.1. Phrase Schemata

Words may be combined to form phrases that have internal structure and unity. We use generalized schemata to describe the phrase structure. The goal is to create a simple, uniform template that is independent of POS category.

Let's first consider nouns, a fundamental category referring to persons, places, and things in the world. The noun and its immediate modifiers form a constituent called the noun phrase (*NP*). To generalize this, we consider a word of arbitrary category, say category *X* (which could be a noun *N* or a verb *V*). The generalized rule for a phrase *XP* is $XP \Rightarrow$ *(modifiers) X-head (post-modifiers),* where *X* is the head, since it dominates the configuration and names the phrase. Elements preceding the head in its phrase are *premodifiers* and elements following the head are *postmodifiers*. *XP*, the culminating phrase node, is called a *maximal projection* of category *X*. We call the whole structure an *x-template*. Maximal projections, *XP*, are the primary currency of basic syntactic processes. The post-modifiers are usually maximal projections (another head, with its own post-modifiers forming an *XP* on its own) and are sometimes termed *complements*, because they are often required by the lexical properties of the head for a complete meaning to be expressed (e.g., when *X* is a preposition

or verb). Complements are typically noun phrases (*NP*), prepositional phrases (*PP*), verb phrases (*VP*), or sentence/clause (*S*), which make an essential contribution to the head's reference or meaning, and which the head requires for semantic completeness. Premodifiers are likely to be adverbs, adjectives, quantifiers, and determiners, i.e., words that help to specify the meaning of the head but may not be essential for completing the meaning. With minor variations, the *XP* template serves for most phrasal types, based on the POS of the head (*N, V, ADJ*, etc.).

For *NP*, we thus have *NP* ⇒ *(det) (modifier)* ***head-noun*** *(post-modifier).* This rule describes an *NP* (noun phrase – left side of arrow) in terms of its optional and required internal contents (right side of the arrow). *Det* is a word like *the* or *a* that helps to resolve the reference to a specific or an unknown instance of the noun. The *modifier* gives further information about the noun. The *head* of the phrase, and the only mandatory element, is the noun itself. *Post-modifiers* also give further information, usually in a more elaborate syntactic form than the simpler pre-modifiers, such as a relative clause or a prepositional phrase (covered below). The noun phrases of the passage above can be parsed as shown in Table 2.16. The head nouns may be personal pronouns (*I, it*), demonstrative and relative pronouns (*those*), coordinated nouns (*sanction and consent*), or common nouns (*individual*). The modifiers are mostly adjectives (*impure, pure*) or verbal forms functioning as adjectives (*limited*). The post-modifiers are interesting, in that, unlike the (pre-)modifiers, they are typically full phrases themselves, rather than isolated words. They include relative clauses (which are a kind of dependent sentence, e.g., *[those] who know and can do better than I*), as well as prepositional phrases (*of the governed*).

Table 2.16 *NP*s of the sample passage.

NP	Det	Mod	Head Noun	Post-Mod
1	the		**authority**	of government
2		even	**such**	as I am willing to submit to
3			**I**	
4			**those**	who know and can do better than I
5		many	**things**	
6		even	**those**	who neither know nor can do so well
7	an	impure	**one**	
8			**it**	
9	the		**sanction and consent**	of the governed
10	no	pure	**right**	over my person ... concede to it.
11	the		**progress**	from an absolute to a limited monarchy
12	an	absolute	**[monarchy]**	
13	a	limited	**monarchy**	
14	a		**democracy**	
15	a		**progress**	
16	a	true	**respect**	for the individual
17	the		**individual**	

Table 2.17 *PPs* of the sample passage.

Head Prep	Complement (Postmodifier)
of	Government
as	I am willing to submit to
than	I
in	many things
of	the governed
over	my person and property
to	it
from	an absolute [monarchy]
to	a limited monarchy
to	a democracy
toward	a true respect [for the individual]
for	the individual

Prepositions express spatial and temporal relations, among others. These are also said to project according to the *X*-template, but usually lack a pre-modifier. Some examples from the sample passage are listed in Table 2.17. The complements of *PP* are generally *NPs*, which may be simple head nouns like *government*. However, other complement types, such as the verb phrase in *after discussing it with Jo*, are also possible.

For verb phrases, the postmodifier (or complement) of a head verb would typically be one or more *NP* (noun phrase) maximal projections, which might, for example, function as a direct object in a *VP* like *pet the cat*. The complement may or may not be optional, depending on characteristics of the head. We can now make some language-specific generalizations about English. Some verbs, such as *give*, may take more than one kind of complement. So an appropriate template for a *VP* maximal projection in English would appear abstractly as *VP* $\Rightarrow$ *(modifier) verb (modifier) (Complement1, Complement2 ComplementN)*. Complements are usually regarded as maximal projections, such as *NP*, *ADJP*, etc., and are enumerated in the template above, to cover possible multi-object verbs, such as *give*, which take both direct and indirect objects. Certain types of adverbs (*really*, *quickly*, *smoothly*, etc.) could be considered fillers for the *VP* modifier slots (before and after the head). In the sample passage, we find the following verb phrases as shown in Table 2.18.

VP presents some interesting issues. First, notice the multi-word verb *submit to*. Multi-word verbs such as *look after* and *put up with* are common. We also observe a number of auxiliary elements clustering before the verb in sentences of the sample passage: *am willing to submit to*, *will cheerfully obey*, and *can do better*. Rather than considering these as simple modifiers of the verbal head, they can be taken to have *scope* over the *VP* as a whole, which implies they are outside the *VP*. Since they are outside the *VP*, we can assume them to be heads in their own right, of phrases which require a *VP* as their complement. These elements mainly express tense (time or duration of verbal action) and modality (likelihood or probability of verbal action). In a full sentence, the *VP* has explicit or implicit inflection (projected from its verbal head) and indicates the person, number, and other context-dependent

features of the verb in relation to its arguments. In English, the person (first, second, third) and number (singular, plural) attributes, collectively called agreement features, of subject and verb must match. For simplicity, we will lump all these considerations together as inflectional elements, and posit yet another phrase type, the Inflectional Phrase (*IP*): *IP* ⇒ *premodifier head VP-complement.*

Table 2.18 *VP*s of the sample passage.

Pre-mod	Verb Head	Post-mod	Complement
	submit to		[the authority of government]
cheerfully	**obey**		those who know and can do better than I
	is	still	an impure one
	be		strictly just
	have		the sanction
	have		no pure right
	concede		to it
	is		a progress

The premodifier slot (sometimes called the *specifier* position in linguistic theory) of an *IP* is often filled by the subject of the sentence (typically a noun or *NP*). Since the *IP* unites the subject of a sentence with a *VP*, *IP* can also be considered simply as the sentence category, often written as *S* in speech grammars.

2.4.1.2. Clauses and Sentences

The *subject* of a sentence is what the sentence is mainly about. A *clause* is any phrase with both a subject and a *VP* (*predicate* in traditional grammars) that has potentially independent interpretation – thus, for us, a clause is an *IP*, a kind of sentence. A phrase is a constituent lacking either subject, predicate, or both. We have reviewed a number of phrase types above. There are also various types of clauses and sentences.

Even though clauses are sentences from an internal point of view (having subject and predicate), they often function as simpler phrases or words would, e.g., as modifiers (adjective and adverbs) or nouns and noun phrases. Clauses may appear as post-modifiers for nouns (so-called *relative clauses*), basically a kind of adjective clause, sharing their subjects with the containing sentence. Some clauses function as *NP*s in their own right. One common clause type substitutes a *wh-word* like *who* or *what* for a direct object of a verb in the embedded clause, to create a *questioned noun phrase* or indirect question: (*I don't know who Jo saw.*). In these clauses, it appears to syntacticians that the *questioned* object of the verb [*VP* saw **who**] has been extracted or moved to a new surface position (following the main clause verb *know*). This is sometimes shown in the phrase-structure diagram by co-indexing an empty *ghost* or trace constituent at the original position of the question pronoun with the question-*NP* appearing at the surface site:

I don't know $[_{NPobj}\,[_{IP}\,[_{NPi}$ *who] Jo saw* $[_{NPi}\,\_\,]]]$
$[_{NPsubj}\,[_{IP}$ *Whoever wins the game]] is our hero.*

There are various characteristic types of sentences. Some typical types include:

- Declarative: *I gave her a book.*
- Yes-no question: *Did you give her a book?*
- Wh-question: *What did you give her?*
- Alternatives question: *Did you give her a book, a scarf, or a knife?*
- Tag question: *You gave it to her, didn't you?*
- Passive: *She was given a book.*
- Cleft: *It must have been a book that she got.*
- Exclamative: *Hasn't this been a great birthday!*
- Imperative: *Give me the book.*

2.4.1.3. Parse Tree Representations

Sentences can be diagrammed in parse trees to indicate phrase-internal structure and linear precedence and immediate dominance among phrases. A typical phrase-structure tree for part of an embedded sentence is illustrated in Figure 2.26.

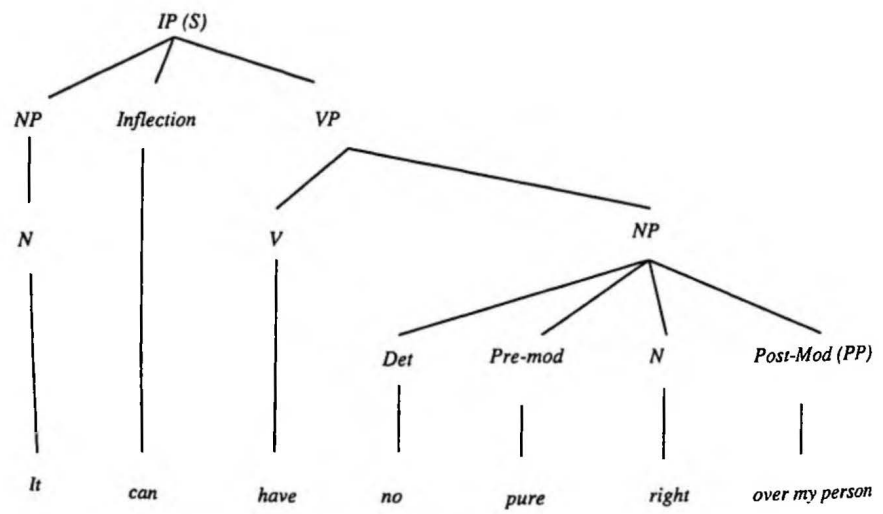

Figure 2.26 A simplified phrase-structure diagram.

For brevity, the same information illustrated in the tree can be represented as a bracketed string as follows:

$[_{IP}\,[_{NP}\,[_{N}\,It\,]_{N}]_{NP}\,[_{I}\,can\,]_{I}\,[_{VP}[_{V}have\,]_{V}\,[_{NP}\,no\ pure\ right\ [_{PP}\,over\ my\ person\,]_{PP}]_{NP}]_{VP}]_{IP}$

With such a bracketed representation, almost every type of syntactic constituent can be coordinated or joined with another of its type, and usually a new phrase node of the common type is added to subsume the constituents such as *NP: We have* $[_{NP}\,[_{NP}$ *tasty berries] and* $[_{NP}$ *tart juices]], IP/S:* $[_{IP}\,[_{IP}$ *Many have come] and* $[_{IP}$ *most have remained]], PP: We went* $[_{PP}\,[_{PP}$ *over the river] and* $[_{PP}$ *into the trees]],* and *VP: We want to* $[_{VP}\,[_{VP}$ *climb the mountains] and* $[_{VP}$ *sail the seas]].*

2.4.2. Semantic Roles

In traditional syntax, grammatical roles are used to describe the direction or control of action relative to the verb in a sentence. Examples include the ideas of *subject*, *object*, *indirect object*, etc. Semantic roles, sometimes called case relations, seem similar but dig deeper. They are used to make sense of the participants in an event, and they provide a vocabulary for us to answer the basic question *who did what to whom.* As developed by [13] and others, the theory of semantic roles posits a limited number of universal roles. Each basic meaning of each verb in our mental dictionary is tagged for the obligatory and optional semantic roles used to convey the particular meaning. A typical inventory of case roles is given below:

Agent	*cause or initiator of action, often intentional*
Patient/Theme	*undergoer of the action*
Instrument	*how action is accomplished*
Goal	*to whom action is directed*
Result	*result of action*
Location	*location of action*

These can be realized under various syntactic identities, and can be assigned to both required complement and optional adjuncts. A noun phrase in the Agentive role might be the surface subject of a sentence, or the object of the preposition *by* in a passive. For example, the verb *put* can be considered a process that has, in one of its senses, the case role specifications shown in Table 2.19.

Table 2.19 Analysis of a sentence with *put.*

Analysis	**Example**			
	Kim	*put*	*the book*	*on the table.*
Grammatical functions	*Subject (NP)*	*Predicate (VP)*	*Object (NP)*	*Adverbial (ADVP)*
Semantic roles	*Agent*	*Instrument*	*Theme*	*Location*

Now consider this passive-tense example, where the semantic roles align with different grammatical roles shown in Table 2.20. Words that look and sound identical can have different meaning or different *senses* as shown in Table 2.21. The sporting sense of *put* (as in the sport of shot-put) illustrates the meaning/sense-dependent nature of the role patterns, because in this sense the Locative case is no longer obligatory, as it is in the original sense illustrated in Table 2.19 and Table 2.20.

Table 2.20 Analysis of passive sentence with *put*.

Analysis	Example		
	The book	*was put*	*on the table.*
Grammatical functions	*Subject (NP)*	*Predicate (VP)*	*Adverbial (ADVP)*
Semantic roles	*Agent*	*Instrument*	*Location*

Table 2.21 Analysis of a different pattern of *put*.

Analysis	Example		
	Kim	*put*	*the shot.*
Grammatical functions	*Subject (NP)*	*Predicate (VP)*	*Object (NP)*
Semantic roles	*Agent*	*Instrument*	*Theme*

The lexical meaning of a verb can be further decomposed into primitive semantic relations such as CAUSE, CHANGE, and BE. The verb *open* might appear as *CAUSE(NP1,PHYSICAL-CHANGE(NP2,NOT-OPEN,OPEN))*. This says that for an agent (*NP1*) to *open* a theme (*NP2*) is to cause the patient to change from a not-opened state to an opened state. Such systems can be arbitrarily detailed and exhaustive, as the application requires.

2.4.3. Lexical Semantics

The specification of particular meaning templates for individual senses of particular words is called *lexical semantics*. When words combine, they may take on propositional meanings resulting from the composition of their meanings in isolation. We could imagine that a speaker starts with a proposition in mind (logical form as will be discussed in the next section), creating a need for particular words to express the idea (lexical semantics); the proposition is then linearized (syntactic form) and spoken (phonological/phonetic form). Lexical semantics is the level of meaning before words are composed into phrases and sentences, and it may heavily influence the possibilities for combination.

Words can be defined in a large number of ways including by relations to other words, in terms of decomposition semantic primitives, and in terms of non-linguistic cognitive constructs, such as perception, action, and emotion. There are hierarchical and non-hierarchical relations. The main hierarchical relations would be familiar to most object-oriented programmers. One is *is-a* taxonomies (a *crow* is-a *bird*), which have transitivity of properties

from type to subtype (inheritance). Another is *has-a* relations (a *car* has-a *windshield*), which are of several differing qualities, including process/subprocess (*teaching* has-a subprocess *giving exams*), and arbitrary or natural subdivisions of part-whole relations (*bread* has-a division into *slices, meter* has-a division into *centimeters*). Then there are non-branching hierarchies (no fancy name) that essentially form scales of degree, such as *frozen* ⇒ *cold* ⇒ *lukewarm* ⇒ *hot* ⇒ *burning*. Non-hierarchical relations include synonyms, such as *big/large*, and antonyms such as *good/bad*.

Words seem to have natural affinities and disaffinities in the semantic relations among the concepts they express. Because these affinities could potentially be exploited by future language understanding systems, researchers have used the generalizations above in an attempt to tease out a parsimonious and specific set of basic relations under which to group entire lexicons of words. A comprehensive listing of the families and subtypes of possible semantic relations has been presented in [10]. In Table 2.22, the leftmost column shows names for families of proposed relations, the middle column differentiates subtypes within each family, and the rightmost column provides examples of word pairs that participate in the proposed relation. Note that case roles have been modified for inclusion as a type of semantic relation within the lexicon.

We can see from Table 2.22 that a single word could participate in multiple relations of different kinds. For example, *knife* appears in the examples for *Similars: invited attribute* (i.e., a desired and expected property) as: *knife-sharp*, and also under *Case Relations*: *action-instrument*, which would label the relation of *knife* to the action *cut* in *He cut the bread with a knife.* This suggests that an entire lexicon could be viewed as a graph of semantic relations, with words or idioms as nodes and connecting edges between them representing semantic relations as listed above. There is a rich tradition of research in this vein.

The biggest practical problem of lexical semantics is the context-dependent resolution of senses of words – so-called polysemy. A classic example is *bank – bank of the stream* as opposed to *money in the bank.* While lexicographers try to identify distinct senses when they write dictionary entries, it has been generally difficult to rigorously quantify exactly what counts as a discrete sense of a word and to disambiguate the senses in practical contexts. Therefore, designers of practical speech understanding systems generally avoid the problem by limiting the domain of discourse. For example, in a financial application, generally only the sense of *bank* as a fiduciary institution is accessible, and others are assumed not to exist. It is sometimes difficult to make a principled argument as to how many distinct senses a word has, because at some level of depth and abstraction, what might appears as separate senses seem to be similar or related, as *face* could be *face of a clock* or *face of person.*

Senses are usually distinguished within a given part-of-speech (POS) category. Thus, when an occurrence of *bank* has been identified as a verb, the *shore* sense might be automatically eliminated, though depending on the sophistication of the system's lexicon and goals, there can be sense differences for many English verbs as well. Within a POS category, often the words that occur near a given ambiguous form in the utterance or discourse are clues to interpretation, where links can be established using semantic relations as described above. Mutual information measures as discussed in Chapter 3 can sometimes provide hints. In a context of dialog where other, less ambiguous financial terms come up

frequently, the sense of *bank* as fiduciary institution is more likely. Finally, when all else fails, often senses can be ranked in terms of their a priori likelihood of occurrence. It should always be borne in mind that language is not static; it can change form under a given analysis at any time. For example, the stable English form *spinster*, a somewhat pejorative term for an older, never-married female, has recently taken on a new morphologically complex form, with the new sense of a high political official, or media spokesperson, employed to provide bland disinformation (*spin*) on a given topic.

Table 2.22 Semantic relations.

Family	Subtype	Example
Contrasts	Contrary	*old-young*
	Contradictory	*alive-dead*
	Reverse	*buy-sell*
	Directional	*front-back*
	Incompatible	*happy-morbid*
	Asymmetric contrary	*hot-cool*
	Attribute similar	*rake-fork*
Similars	Synonymity	*car-auto*
	Dimensional similar	*smile-laugh*
	Necessary attribute	*bachelor-unmarried*
	Invited attribute	*knife-sharp*
	Action subordinate	*talk-lecture*
Class Inclusion	Perceptual subord.	*animal-horse*
	Functional subord.	*furniture-chair*
	State subord.	*disease-polio*
	Activity subord.	*game-chess*
	Geographic subord.	*country-Russia*
	Place	*Germany-Hamburg*
Case Relations	Agent-action	*artist-paint*
	Agent-instrument	*farmer-tractor*
	Agent-object	*baker-bread*
	Action-recipient	*sit-chair*
	Action-instrument	*cut-knife*
Part-Whole	Functional object	*engine-car*
	Collection	*forest-tree*
	Group	*choir-singer*
	Ingredient	*table-wood*
	Functional location	*kitchen-stove*
	Organization	*college-admissions*
	Measure	*mile-yard*

2.4.4. Logical Form

Because of all the lexical, syntactic, and semantic ambiguity in language, some of which requires external context for resolution, it is desirable to have a metalanguage in which to concretely and succinctly express all linguistically possible meanings of an utterance before discourse and world knowledge are applied to choose the most likely interpretation. The favored metalanguage for this purpose is called the predicate logic, used to represent the logical form, or context-independent meaning, of an utterance. The semantic component of many SLU architectures builds on a substrate of two-valued, first-order, logic. To distinguish *shades of meaning* beyond truth and falsity requires more powerful formalisms for knowledge representation.

In a typical first-order system, predicates correspond to events or conditions denoted by verbs (such as *Believe* or *Like*), states of identity (such as being a *Dog* or *Cat*), and properties of varying degrees of permanence (*Happy*). In this form of logical notation, predicates have open places, filled by arguments, as in a programming language subroutine definition. Since individuals may have identical names, subscripting can be used to preserve unique reference. In the simplest systems, predication ranges over individuals rather than higher-order entities such as properties and relations.

Predicates with filled argument slots map onto sets of individuals (constants) in the universe of discourse, in particular those individuals possessing the properties, or participating in the relation, named by the predicate. One-place predicates like *Soldier*, *Happy*, or *Sleeps* range over sets of individuals from the universe of discourse. Two-place predicates, like transitive verbs such as *loves*, range over a set consisting of ordered pairs of individual members (constants) of the universe of discourse. For example, we can consider the universe of discourse to be U = {*Romeo, Juliet, Paris, Rosaline, Tybalt*}, people as characters in a play. They do things with and to one another, such as loving and killing. Then we could imagine the relation *Loves* interpreted as the set of ordered pairs: {<*Romeo, Juliet*>, <*Juliet, Romeo*>, <*Tybalt, Tybalt*>, <*Paris, Juliet*>}, a subset of the Cartesian product of theoretically possible love matches $U \times U$. So, for any ordered pair x, y in U, *Loves*(x, y) is true if the ordered pair $\langle x,y \rangle$ is a member of the extension of the *Loves* predicate as defined, e.g., *Romeo loves Juliet*, *Juliet loves Romeo*, etc.. Typical formal properties of relations are sometimes specially marked by grammar, such as the reflexive relation *Loves*(*Tybalt*, *Tybalt*), which can rendered in natural language as *Tybalt loves himself*. Not every possibility is present; for instance in our example, the individual Rosaline does not happen to participate at all in this extensional definition of *Loves* over U, as her omission from the pairs list indicates. Notice that the subset of *Loves*(x, y) of ordered pairs involving both *Romeo* and *Juliet* is symmetric, also marked by grammar, as in *Romeo and Juliet love each other.* This general approach extends to predicates with any arbitrary number of arguments, such as intransitive verbs like *give*.

Just as in ordinary propositional logic, connectives such as negation, conjunction, disjunction, and entailment are admitted, and can be used with predicates to denote common natural language meanings:

Romeo isn't happy = ¬Happy(Romeo)
Romeo isn't happy, but Tybalt is (happy) = ¬Happy(Romeo) ∧ Happy(Tybalt)
Either Romeo or Tybalt is happy = Happy(Romeo) ∨ Happy(Tybalt)
If Romeo is happy, Juliet is happy = Happy(Romeo) → Happy(Juliet)

Formulae, such as those above, are also said to bear a binary truth value, true or false, with respect to a world of individuals and relations. The determination of the truth value is compositional, in the sense that the truth value of the whole depends on the truth value of the parts. This is a simplistic but formally tractable view of the relation between language and meaning.

Predicate logic can also be used to denote quantified noun phrases. Consider a simple case such as *Someone killed Tybalt*, predicated over our same U = {*Romeo, Juliet, Paris, Rosaline, Tybalt*}. We can now add an *existential* quantifier, ∃, standing for *there exists* or *there is at least one*. This quantifier will bind a variable over individuals in U, and will attach to a proposition to create a new, quantified proposition in logical form. The use of variables in propositions such as *killed(x, y)* creates open propositions. Binding the variables with a quantifier over them closes the proposition. The quantifier is prefixed to the original proposition: ∃*x Killed(x, Tybalt)*.

To establish a truth (semantic) value for the quantified proposition, we have to satisfy the disjunction of propositions in U: *Killed*(*Romeo*, *Tybalt*) ∨ *Killed*(*Juliet, Tybalt*) ∨ *Killed*(*Paris, Tybalt*) ∨ *Killed*(*Rosaline, Tybalt*) ∨ *Killed*(*Tybalt, Tybalt*). The set of all such bindings of the variable x is the space that determines the truth or falsity of the proposition. In this case, the binding of x = *Romeo* is sufficient to assign a value true to the existential proposition.

2.5. HISTORICAL PERSPECTIVE AND FURTHER READING

Motivated to improve speech quality over the telephone, AT&T Bell Labs has contributed many influential discoveries in speech hearing, including the critical band and articulation index [2, 3]. The *Auditory Demonstration* CD prepared by Houtsma, Rossing, and Wagenaars [18] has a number of very interesting examples on psychoacoustics and its explanations. *Speech, Language, and Communication* [30] and *Speech Communication – Human and Machine* [32] are two good books that provide modern introductions to the structure of spoken language. Many speech perception experiments were conducted by exploring how phonetic information is distributed in the time or frequency domain. In addition to the formant structures for vowels, frequency importance function [12] has been developed to study how features related to phonetic categories are stored at various frequencies. In the time domain, it has been observed [16, 19, 42] that salient perceptual cues may not be evenly distributed over the speech segments and that certain perceptual critical points exist.

As intimate as speech and acoustic perception may be, there are also strong evidences that lexical and linguistic effects on speech perception are not always consistent with acoustic ones. For instance, it has long been observed that humans exhibit difficulties in distinguishing non-native phonemes. Human subjects also carry out categorical goodness

difference assimilation based on their mother tongue [34], and such perceptual mechanism can be observed as early as in six-month-old infants [22]. On the other hand, hearing-impaired listeners are able to effortlessly overcome their acoustical disabilities for speech perception [8]. Speech perception is not simply an auditory matter. McGurk and MacDonald (1976) [27, 28] dramatically demonstrated this when they created a videotape on which the auditory information (phonemes) did not match the visual speech information. The effect of this mismatch between the auditory signal and the visual signal was to create a third phoneme different from both the original auditory and visual speech signals. An example is dubbing the phoneme */ba/* to the visual speech movements */ga/*. This mismatch results in hearing the phoneme */da/*. Even when subjects know of the effect, they report the McGurk effect percept. The McGurk effect has been demonstrated for consonants. vowels, words, and sentences.

The earliest scientific work on phonology and grammars goes back to Panini, a Sanskrit grammarian of the fifth century B.C. (estimated), who created a comprehensive and scientific theory of phonetics, phonology, and morphology, based on data from Sanskrit (the classical literary language of the ancient Hindus). Panini created formal production rules and definitions to describe Sanskrit grammar, including phenomena such as construction of sentences, compound nouns, etc. Panini's formalisms function as ordered rules operating on underlying structures in a manner analogous to modern linguistic theory. Panini's phonological rules are equivalent in formal power to Backus-Nauer form (BNF). A general introduction to this pioneering scientist is Cardona [9].

An excellent introduction to all aspects of phonetics is *A Course in Phonetics* [24]. A good treatment of the acoustic structure of English speech sounds and a through introduction and comparison of theories of speech perception is to be found in [33]. The basics of phonology as part of linguistic theory are treated in *Understanding Phonology* [17]. An interesting treatment of word structure (morphology) from a computational point of view can be found in *Morphology and Computation* [40]. A comprehensive yet readable treatment of English syntax and grammar can be found in *English Syntax* [4] and *A Comprehensive Grammar of the English Language* [36]. Syntactic theory has traditionally been the heart of linguistics, and has been an exciting and controversial area of research since the 1950s. Be aware that almost any work in this area will adopt and promote a particular viewpoint, often to the exclusion or minimization of others. A reasonable place to begin with syntactic theory is *Syntax: A Minimalist Introduction* [37]. An introductory textbook on syntactic and semantic theory that smoothly introduces computational issues is *Syntactic Theory: A Formal Introduction* [39]. For a philosophical and entertaining overview of various aspects of linguistic theory, see *Rhyme and Reason: An Introduction to Minimalist Syntax* [44]. A good and fairly concise treatment of basic semantics is *Introduction to Natural Language Semantics* [11]. Deeper issues are covered in greater detail and at a more advanced level in *The Handbook of Contemporary Semantic Theory* [25]. The intriguing area of lexical semantics (theory of word meanings) is comprehensively presented in *The Generative Lexicon* [35]. *Concise History of the Language Sciences* [21] is a good edited book if you are interested in the history of linguistics.

REFERENCES

[1] Aliprand, J., *et al.*, *The Unicode Standard, Version 2.0*, 1996, Addison Wesley.

[2] Allen, J.B., "How Do Humans Process and Recognize Speech?," *IEEE Trans. on Speech and Audio Processing*, 1994, **2**(4), pp. 567-577.

[3] Allen, J.B., "Harvey Fletcher 1884–1981" in *The ASA Edition of Speech and Hearing Communication* 1995, Woodbury, New York, pp. A1-A34, Acoustical Society of America.

[4] Baker, C.L., *English Syntax*, 1995, Cambridge, MA, MIT Press.

[5] Blauert, J., *Spatial Hearing*, 1983, MIT Press.

[6] Brill, E., "Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part-of-Speech Tagging," *Computational Linguistics*, 1995, **21**(4), pp. 543-566.

[7] Brown, P., *et al.*, "Class-Based N-gram Models of Natural Language," *Computational Linguistics*, 1992, **18**(4).

[8] Caplan, D. and J. Utman, "Selective Acoustic Phonetic Impairment and Lexical Access in an Aphasic Patient," *Journal of the Acoustical Society of America*, 1994, **95**(1), pp. 512-517.

[9] Cardona, G., *Panini: His Work and Its Traditions: Background and Introduction*, 1988, Motilal Banarsidass.

[10] Chaffin, R., and Herrmann, D., "The Nature of Semantic Relations: A Comparison of Two Approaches" in *Representing Knowledge in Semantic Networks*, M. Evens, ed., 1988, Cambridge, UK, Cambridge University Press.

[11] de Swart, H., *Introduction to Natural Language Semantics*, 1998, Stanford, CA, Center for the Study of Language and Information Publications.

[12] Duggirala, V., *et al.*, "Frequency Importance Function for a Feature Recognition Test Material," *Journal of the Acoustical Society of America*, 1988, **83**(9), pp. 2372-2382.

[13] Fillmore, C.J., "The Case for Case" in *Universals in Linguistic Theory*, E. Bach and R. Harms, eds. 1968, New York, NY, Holt, Rinehart and Winston.

[14] Fletcher, H., "Auditory Patterns," *Rev. Mod. Phys.*, 1940, **12**, pp. 47-65.

[15] Fry, D.B., *The Physics of Speech*, Cambridge Textbooks in Linguistics, 1979, Cambridge, U.K., Cambridge University Press.

[16] Furui, S., "On The Role of Spectral Transition for Speech Perception," *Journal of the Acoustical Society of America*, 1986, **80**(4), pp. 1016-1025.

[17] Gussenhoven, C., and Jacobs, H., *Understanding Phonology*, Understanding Language Series, 1998, Edward Arnold.

[18] Houtsma, A., T. Rossing, and W. Wagenaars, *Auditory Demonstrations*, 1987, Institute for Perception Research, Eindhovern, The Netherlands, Acoustic Society of America.

[19] Jenkins, J., W. Strange, and S. Miranda, "Vowel Identification in Mixed-Speaker Silent-Center Syllables," *Journal of the Acoustical Society of America*, 1994, **95**(2), pp. 1030-1041.

[20] Klatt, D., "Review of the ARPA Speech Understanding Project," *Journal of Acoustical Society of America*, 1977, **62**(6), pp. 1324-1366.

[21] Koerner, E. and E. Asher, eds. *Concise History of the Language Sciences*, 1995, Oxford, Elsevier Science.

[22] Kuhl, P., "Infant's Perception and Representation of Speech: Development of a New Theory," *Int. Conf. on Spoken Language Processing*, 1992, Alberta, Canada, pp. 449-452.

[23] Kupeic, J., "Robust Part-of-Speech Tagging Using a Hidden Markov Model," *Computer Speech and Language*, 1992, **6**, pp. 225-242.

[24] Ladefoged, P., *A Course in Phonetics*, 1993, Harcourt Brace Johanovich.

[25] Lappin, S., *The Handbook of Contemporary Semantic Theory*, Blackwell Handbooks in Linguistics, 1997, Oxford, UK, Blackwell Publishsers Inc.

[26] Lindsey, P. and D. Norman, *Human Information Processing*, 1972, New York and London, Academic Press.

[27] MacDonald, J. and H. McGurk, "Visual Influence on Speech Perception Process," *Perception and Psychophysics*, 1978, **24**(3), pp. 253-257.

[28] McGurk, H. and J. MacDonald, "Hearing Lips and Seeing Voices," *Nature*, 1976, **264**, pp. 746-748.

[29] Merialdo, B., "Tagging English Text with a Probabilistic Model," *Computational Linguistics*, 1994, **20**(2), pp. 155-172.

[30] Miller, J. and P. Eimas, *Speech, Language and Communication*, Handbook of Perception and Cognition, eds. E. Carterette and M. Friedman, 1995, Academic Press.

[31] Moore, B.C., *An Introduction to the Psychology of Hearing*, 1982, London, Academic Press.

[32] O'Shaughnessy, D., *Speech Communication – Human and Machine*, 1987, Addison-Wesley.

[33] Pickett, J.M., *The Acoustics of Speech Communication*, 1999, Needham Heights, MA, Allyn & Bacon.

[34] Polka, L., "Linguistic Influences in Adult Perception of Non-native Vowel Contrast," *Journal of the Acoustical Society of America*, 1995, **97**(2), pp. 1286-1296.

[35] Pustejovsky, J., *The Generative Lexicon*, 1998, Bradford Books.

[36] Quirk, R., Svartvik, J., Leech, G., *A Comprehensive Grammar of the English Language*, 1985, Addison-Wesley Pub. Co.

[37] Radford, A., *Syntax: A Minimalist Introduction*, 1997, Cambridge, U.K., Cambridge Univ. Press.

[38] Rossing, T.D., *The Science of Sound*, 1982, Reading, MA, Addison-Wesley.

[39] Sag, I., Wasow, T., *Syntactic Theory: A Formal Introduction*, 1999, Cambridge, UK, Cambridge University Press.

[40] Sproat, R., *Morphology and Computation*, ACL-MIT Press Series in Natural Language Processing, 1992, Cambridge, MA, MIT Press.

[41] Stevens, S.S. and J. Volkman, "The Relation of Pitch to Frequency," *Journal of Psychology*, 1940, **53**, pp. 329.

[42] Strange, W., J. Jenkins, and T. Johnson, "Dynamic Specification of Coarticulated Vowels," *Journal of the Acoustical Society of America*, 1983, 74(3), pp. 695-705.

[43] Thoreau, H.D., *Civil Disobedience, Solitude and Life Without Principle*, 1998, Prometheus Books.

[44] Uriagereka, J., *Rhyme and Reason: An Introduction to Minimalist Syntax*, 1998, Cambridge, MA, MIT Press.

[45] Voutilainen, A., "Morphological Disambiguation" in *Constraint Grammar: A Language-Independent System for Parsing Unrestricted Text* 1995, Berlin, Mouton de Gruyter.

[46] Weischedel, R., "BBN: Description of the PLUM System as Used for MUC-6," *The 6th Message Understanding Conferences (MUC-6)*, 1995, San Francisco, Morgan Kaufmann, pp. 55-70.

CHAPTER 3

Probability, Statistics, and Information Theory

Randomness and uncertainty play an important role in science and engineering. Most spoken language processing problems can be characterized in a probabilistic framework. Probability theory and statistics provide the mathematical language to describe and analyze such systems.

The criteria and methods used to estimate the unknown probabilities and probability densities form the basis for estimation theory. Estimation theory is critical to parameter learning in pattern recognition. In this chapter, three widely used estimation methods are discussed. They are *minimum mean squared error estimation* (MMSE), *maximum likelihood estimation* (MLE), and *maximum posterior probability estimation* (MAP).

Significance testing deals with the confidence of statistical inference, such as knowing whether the estimation of some parameter can be accepted with confidence. In pattern recognition, significance testing is important for determining whether the observed difference between two different classifiers is real. In our coverage of significance testing, we describe various methods that are used in pattern recognition, discussed in Chapter 4.

Information theory was originally developed for efficient and reliable communication systems. It has evolved into a mathematical theory concerned with the very essence of the communication process. It provides a framework for the study of fundamental issues, such as the efficiency of information representation and the limitations in reliable transmission of information over a communication channel. Many of these problems are fundamental to spoken language processing.

3.1. PROBABILITY THEORY

Probability theory deals with the averages of mass phenomena occurring sequentially or simultaneously. We often use probabilistic expressions in our day-to-day lives, such as when saying, *It is very likely that the Dow (Dow Jones Industrial index) will hit 12,000 points next month*, or, *The chance of scattered showers in Seattle this weekend is high.* Each of these expressions is based upon the concept of the probability, or the likelihood, that some specific event will occur.

Probability can be used to represent the degree of confidence in the outcome of some actions (observations), which are not definite. In probability theory, the term *sample space, S,* is used to refer to the collection (set) of all possible outcomes. An *event* refers to a subset of the sample space or a collection of outcomes. The *probability of event A,* denoted as $P(A)$, can be interpreted as the *relative frequency* with which event A would occur if the process were repeated a large number of times under similar conditions. Based on this interpretation, $P(A)$ can be computed simply by counting the total number, N_S, of all observations and the number of observations N_A whose outcome belongs to event A. That is,

$$P(A) = \frac{N_A}{N_S} \tag{3.1}$$

$P(A)$ is bounded between zero and one, i.e.,

$$0 \le P(A) \le 1 \text{ for all } A \tag{3.2}$$

The lower bound of probability $P(A)$ is zero when the event set A is an empty set. On the other hand, the upper bound of probability $P(A)$ is one when the event set A happens to be S.

If there are n events $A_1, A_2, \cdots, A_n$ in S such that $A_1, A_2, \cdots, A_n$ are disjoint and $\bigcup_{i=1}^{n} A_i = S$, events $A_1, A_2, \cdots, A_n$ are said to form a *partition* of S. The following obvious equation forms a fundamental axiom for probability theory.

$$P(A_1 \cup A_2 \cup \ldots A_n) = \sum_{i=1}^{n} P(A_i) = 1 \tag{3.3}$$

Based on the definition in Eq. (3.1), the *joint probability* of event A and event B occurring concurrently is denoted as $P(AB)$ and can be calculated as:

$$P(AB) = \frac{N_{AB}}{N_S} \tag{3.4}$$

3.1.1. Conditional Probability and Bayes' Rule

It is useful to study the way in which the probability of an event A changes after it has been learned that some other event B has occurred. This new probability denoted as $P(A \mid B)$ is called the *conditional probability* of event A given that event B has occurred. Since the set of those outcomes in B that also result in the occurrence of A is exactly the set AB as illustrated in Figure 3.1, it is natural to define the conditional probability as the proportion of the total probability $P(B)$ that is represented by the joint probability $P(AB)$. This leads to the following definition:

$$P(A \mid B) = \frac{P(AB)}{P(B)} = \frac{N_{AB}/N_S}{N_B/N_S} \tag{3.5}$$

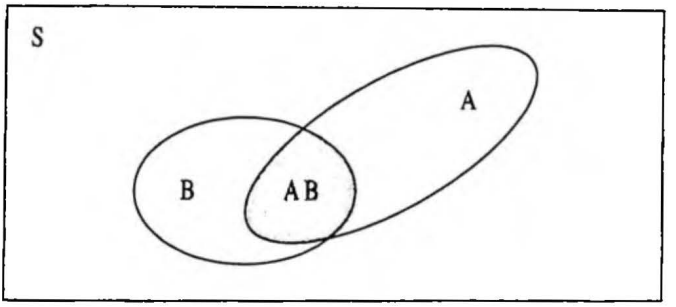

Figure 3.1 The intersection AB represents where the joint event A and B occurs concurrently.

Based on the definition of conditional probability, the following expressions can be easily derived.

$$P(AB) = P(A \mid B)P(B) = P(B \mid A)P(A) \tag{3.6}$$

Equation (3.6) is the simple version of the *chain rule*. The chain rule, which can specify a joint probability in terms of multiplication of several cascaded conditional probabilities, is often used to decompose a complicated joint probabilistic problem into a sequence of stepwise conditional probabilistic problems. Equation (3.6) can be converted to such a general chain:

$$P(A_1A_2 \cdots A_n) = P(A_n \mid A_1 \cdots A_{n-1}) \cdots P(A_2 \mid A_1)P(A_1) \tag{3.7}$$

When two events, A and B, are independent of each other, in the sense that the occurrence or of either of them has no relation to and no influence on the occurrence of the other, it is obvious that the conditional probability $P(B \mid A)$ equals to the unconditional probability

$P(B)$. It follows that the joint probability $P(AB)$ is simply the product of $P(A)$ and $P(B)$ if A and B, are independent.

If the n events $A_1, A_2, \cdots, A_n$ form a partition of S and B is any event in S as illustrated in Figure 3.2, the events $A_1B, A_2B, \cdots, A_nB$ form a partition of B. Thus, we can rewrite:

$$B = A_1B \cup A_2B \cup \cdots \cup A_nB \tag{3.8}$$

Since $A_1B, A_2B, \cdots, A_nB$ are disjoint,

$$P(B) = \sum_{k=1}^{n} P(A_kB) \tag{3.9}$$

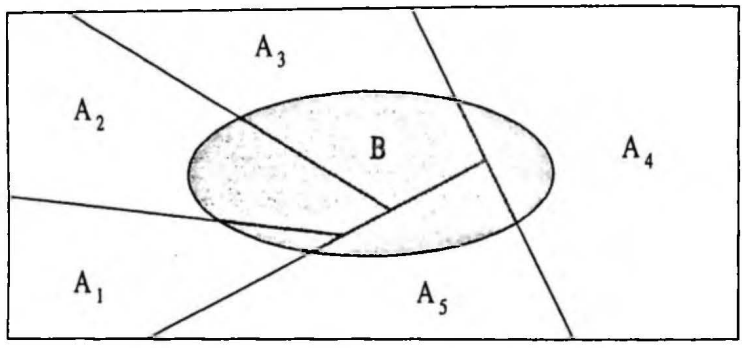

Figure 3.2 The intersections of B with partition events $A_1, A_2, \cdots, A_n$.

Equation (3.9) is called the *marginal probability* of event B, where the probability of event B is computed from the sum of joint probabilities.

According to the chain rule, Eq. (3.6), $P(A_iB) = P(A_i)P(B \mid A_i)$, it follows that

$$P(B) = \sum_{k=1}^{n} P(A_k)P(B \mid A_k) \tag{3.10}$$

Combining Eqs. (3.5) and (3.10), we get the well-known *Bayes' rule*:

$$P(A_i \mid B) = \frac{P(A_iB)}{P(B)} = \frac{P(B \mid A_i)P(A_i)}{\sum_{k=1}^{n} P(B \mid A_k)P(A_k)} \tag{3.11}$$

Bayes' rule is the basis for pattern recognition that is described in Chapter 4.

3.1.2. Random Variables

Elements in a sample space may be numbered and referred to by the numbers given. A variable X that specifies the numerical quantity in a sample space is called a *random variable*. Therefore, a random variable X is a function that maps each possible outcome s in the sample space S onto real numbers $X(s)$. Since each event is a subset of the sample space, an event is represented as a set of $\{s\}$ which satisfies $\{s \mid X(s)=x\}$. We use capital letters to denote random variables and lower-case letters to denote fixed values of the random variable. Thus, the probability that $X=x$ is denoted as:

$$P(X=x)=P(s \mid X(s)=x) \tag{3.12}$$

A random variable X is a *discrete* random variable, or X has a *discrete distribution*, if X can take only a finite number n of different values $x_1,x_2,\cdots,x_n$, or at most, an infinite sequence of different values $x_1,x_2,\cdots$. If the random variable X is a discrete random variable, the *probability function* (pf) or *probability mass function* (pmf) of X is defined to be the function p such that for any real number x,

$$p_X(x)=P(X=x) \tag{3.13}$$

For the cases in which there is no confusion, we drop the subscription X for $p_X(x)$. The sum of probability mass over all values of the random variable is equal to unity.

$$\sum_{i=1}^{n} p(x_i)=\sum_{i=1}^{n} P(X=x_i)=1 \tag{3.14}$$

The marginal probability, chain rule and Bayes' rule can also be rewritten with respect to random variables:

$$p_X(x_i)=P(X=x_i)=\sum_{k=1}^{m} P(X=x_i,Y=y_k)=\sum_{k=1}^{m} P(X=x_i \mid Y=y_k)P(Y=y_k) \tag{3.15}$$

$$\begin{aligned}&P(X_1=x_1,\cdots,X_n=x_n)=\\&P(X_n=x_n \mid X_1=x_1,\cdots,X_{n-1}=x_{n-1})\cdots P(X_2=x_2 \mid X_1=x_1)P(X_1=x_1)\end{aligned} \tag{3.16}$$

$$P(X=x_i \mid Y=y)=\frac{P(X=x_i,Y=y)}{P(Y=y)}=\frac{P(Y=y \mid X=x_i)P(X=x_i)}{\sum_{k=1}^{n} P(Y=y \mid X=x_k)P(X=x_k)} \tag{3.17}$$

In a similar manner, if the random variables X and Y are statistically independent, they can be represented as:

$$P(X=x_i,Y=y_j)=P(X=x_i)P(Y=y_j)=p_X(x_i)p_Y(y_j) \ \forall \text{ all } i \text{ and } j \tag{3.18}$$

A random variable X is a *continuous* random variable, or X has a *continuous distribution*, if there exists a nonnegative function f, defined on the real line, such that for an interval A,

$$P(X \in A) = \int_A f_X(x)dx \tag{3.19}$$

The function f_X is called the *probability density function* (abbreviated pdf) of X. We drop the subscript X for f_X if there is no ambiguity. As illustrated in Figure 3.3, the area of shaded region is equal to the value of $P(a \le X \le b)$.

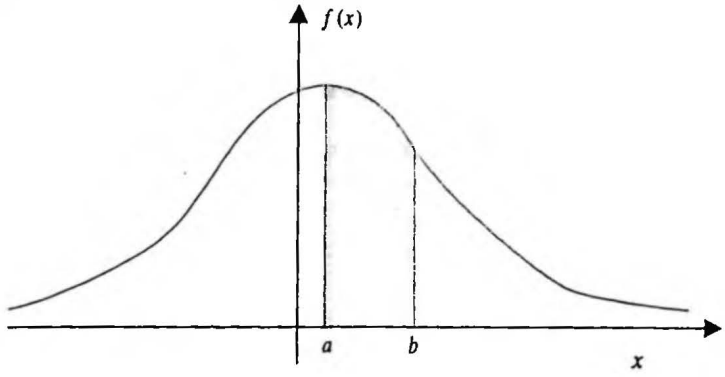

Figure 3.3 An example of pdf. The area of the shaded region is equal to the value of $P(a \le X \le b)$.

Every pdf must satisfy the following two requirements:

$$\begin{aligned} &f(x) \ge 0 \text{ for } -\infty \le x \le \infty \text{ and} \\ &\int_{-\infty}^{\infty} f(x)dx = 1 \end{aligned} \tag{3.20}$$

The marginal probability, chain rule, and Bayes' rule can also be rewritten with respect to continuous random variables:

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy = \int_{-\infty}^{\infty} f_{X|Y}(x \mid y) f_Y(y)dy \tag{3.21}$$

$$f_{X_1,\cdots,X_n}(x_1,\cdots,x_n) = f_{X_n|X_1,\cdots,X_{n-1}}(x_n \mid x_1,\cdots,x_{n-1}) \cdots f_{X_2|X_1}(x_2 \mid x_1) f_{X_1}(x_1) \tag{3.22}$$

$$f_{X|Y}(x \mid y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \frac{f_{Y|X}(y \mid x) f_X(x)}{\int_{-\infty}^{\infty} f_{Y|X}(y \mid x) f_X(x)dx} \tag{3.23}$$

The *distribution function* or *cumulative distribution function* F of a discrete or continuous random variable X is a function defined for every real number x as follows:

$$F(x) = P(X \le x) \quad \text{for } -\infty \le x \le \infty \tag{3.24}$$

For continuous random variables, it follows that:

$$F(x) = \int_{-\infty}^{x} f_X(x)dx \tag{3.25}$$

$$f_X(x) = \frac{dF(x)}{dx} \tag{3.26}$$

3.1.3. Mean and Variance

Suppose that a discrete random variable X has a pf $f(x)$; the *expectation* or *mean* of X is defined as follows:

$$E(X) = \sum_{x} xf(x) \tag{3.27}$$

Similarly, if a continuous random variable X has a pdf f, the *expectation* or *mean* of X is defined as follows:

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx \tag{3.28}$$

In physics, the mean is regarded as the center of mass of the probability distribution. The expectation can also be defined for any function of the random variable X. If X is a continuous random variable with pdf f, then the expectation of any function $g(X)$ can be defined as follows:

$$E\left[g(X)\right] = \int_{-\infty}^{\infty} g(x)f(x)dx \tag{3.29}$$

The expectation of a random variable is a linear operator. That is, it satisfies both additivity and homogeneity properties:

$$E(a_1X_1 + \cdots + a_nX_n + b) = a_1E(X_1) + \cdots + a_nE(X_n) + b \tag{3.30}$$

where $a_1, \cdots, a_n, b$ are constants.

Equation (3.30) is valid regardless of whether or not the random variables $X_1, \cdots, X_n$ are independent.

Suppose that X is a random variable with mean $\mu = E(X)$. The *variance* of X denoted as $Var(X)$ is defined as follows:

$$Var(X) = \sigma^2 = E\left[(X - \mu)^2\right] \tag{3.31}$$

where σ, the nonnegative square root of the variance is known as the *standard deviation* of random variable X. Therefore, the variance is also often denoted as σ^2.

The variance of a distribution provides a measure of the spread or dispersion of the distribution around its mean μ. A small value of the variance indicates that the probability distribution is tightly concentrated around μ, and a large value of the variance typically indicates the probability distribution has a wide spread around μ. Figure 3.4 illustrates three different Gaussian distributions[1] with the same mean, but different variances.

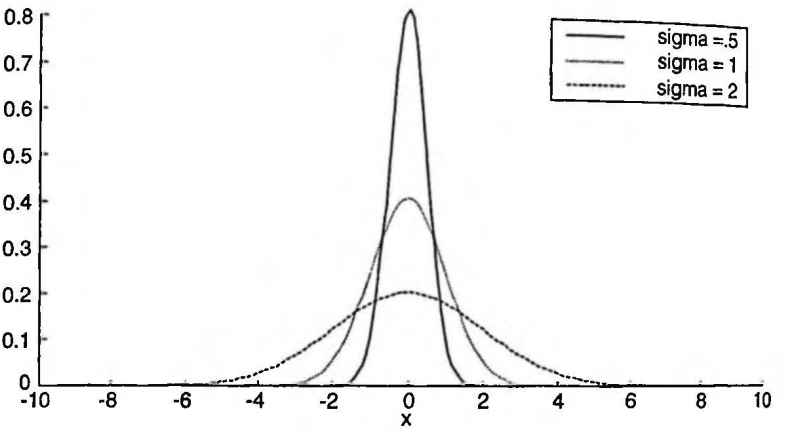

Figure 3.4 Three Gaussian distributions with same mean μ, but different variances, 0.5, 1.0, and 2.0, respectively. The distribution with a large value of the variance has a wide spread around the mean μ.

The variance of random variable X can be computed in the following way:

$$Var(X) = E(X^2) - \left[E(X)\right]^2 \tag{3.32}$$

In physics, the expectation $E(X^k)$ is called the k^{th} moment of X for any random variable X and any positive integer k. Therefore, the variance is simply the difference between the second moment and the square of the first moment.

The variance satisfies the following additivity property, if random variables X and Y are independent:

$$Var(X+Y) = Var(X) + Var(Y) \tag{3.33}$$

However, it does not satisfy the homogeneity property. Instead for constant a,

$$Var(aX) = a^2 Var(X) \tag{3.34}$$

[1] We describe Gaussian distributions in Section 3.1.7.

Since it is clear that $Var(b) = 0$ for any constant b, we have an equation similar to Eq. (3.30) if random variables $X_1, \cdots, X_n$ are independent.

$$Var(a_1X_1 + \cdots + a_nX_n + b) = a_1^2 Var(X_1) + \cdots + a_n^2 Var(X_n) \tag{3.35}$$

Conditional expectation can also be defined in a similar way. Suppose that X and Y are discrete random variables and let $f(y \mid x)$ denote the conditional pf of Y given $X = x$, then the conditional expectation $E(Y \mid X)$ is defined to be the function of X whose value $E(Y \mid x)$ when $X = x$ is

$$E_{Y|X}(Y \mid X = x) = \sum_y y f_{Y|X}(y \mid x) \tag{3.36}$$

For continuous random variables X and Y with $f_{Y|X}(y \mid x)$ as the conditional pdf. of Y given $X = x$, the conditional expectation $E(Y \mid X)$ is defined to be the function of X whose value $E(Y \mid x)$ when $X = x$ is

$$E_{Y|X}(Y \mid X = x) = \int_{-\infty}^{\infty} y f_{Y|X}(y \mid x) dy \tag{3.37}$$

Since $E(Y \mid X)$ is a function of random variable X, it itself is a random variable whose probability distribution can be derived from the distribution of X. It can be shown that

$$E_X\left[E_{Y|X}(Y \mid X)\right] = E_{X,Y}(Y) \tag{3.38}$$

More generally, suppose that X and Y have a continuous joint distribution and that $g(x, y)$ is any arbitrary function of X and Y. The conditional expectation $E[g(X, Y) \mid X]$ is defined to be the function of X whose value $E[g(X, Y) \mid x]$ when $X = x$ is

$$E_{Y|X}\left[g(X,Y) \mid X = x\right] = \int_{-\infty}^{\infty} g(x, y) f_{Y|X}(y \mid x) dy \tag{3.39}$$

Equation (3.38) can also be generalized into the following equation:

$$E_X\left\{E_{Y|X}\left[g(X,Y) \mid X\right]\right\} = E_{X,Y}\left[g(X,Y)\right] \tag{3.40}$$

Finally, it is worthwhile to introduce *median* and *mode*. The median of a distribution of X is defined to be a point m, such that $P(X \le m) \ge 1/2$ and $P(X \ge m) \ge 1/2$. Thus, the median m divides the total probability into two equal parts, i.e., the probability to the left of m and the probability to the right of m are exactly $1/2$.

Suppose a random variable X has either a discrete distribution with pf $p(x)$ or continuous pdf $f(x)$; a point ϖ is called the mode of the distribution if $p(x)$ or $f(x)$ attains the maximum value at the point ϖ. A distribution can have more than one mode.

3.1.3.1. The Law of Large Numbers

The concept of *sample mean* and *sample variance* is important in statistics because most statistical experiments involve sampling. Suppose that the random variables $X_1,\cdots,X_n$ form a random sample of size n from some distribution for which the mean is μ and the variance is σ^2. In other words, the random variables $X_1,\cdots,X_n$ are *independent identically distributed* (often abbreviated by iid) and each has mean μ and variance σ^2. Now if we denote $\bar{X}_n$ as the arithmetic average of the n observations in the sample, then

$$\bar{X}_n = \frac{1}{n}(X_1 + \cdots + X_n) \tag{3.41}$$

$\bar{X}_n$ is a random variable and is referred to as *sample mean*. The mean and variance of $\bar{X}_n$ can be easily derived based on the definition.

$$E(\bar{X}_n) = \mu \quad \text{and} \quad \text{Var}(\bar{X}_n) = \frac{\sigma^2}{n} \tag{3.42}$$

Equation (3.42) states that the mean of sample mean is equal to mean of the distribution, while the variance of sample mean is only $1/n$ times the variance of the distribution. In other words, the distribution of $\bar{X}_n$ will be more concentrated around the mean μ than was the original distribution. Thus, the sample mean is closer to μ than is the value of just a single observation X_i from the given distribution.

The *law of large numbers* is one of most important theorems in probability theory. Formally, it states that the sample mean $\bar{X}_n$ converges to the mean μ in probability, that is,

$$\lim_{n\to\infty} P\left(|\bar{X}_n - \mu| < \varepsilon\right) = 1 \quad \text{for any given number } \varepsilon > 0 \tag{3.43}$$

The law of large numbers basically implies that the sample mean is an excellent estimate of the unknown mean of the distribution when the sample size n is large.

3.1.4. Covariance and Correlation

Let X and Y be random variables having a specific joint distribution, and $E(X) = \mu_X$, $E(Y) = \mu_Y$, $Var(X) = \sigma_X^2$, and $Var(Y) = \sigma_Y^2$. The *covariance* of X and Y, denoted as $Cov(X,Y)$, is defined as follows:

$$Cov(X,Y) = E\left[(X-\mu_X)(Y-\mu_Y)\right] = Cov(Y,X) \tag{3.44}$$

In addition, the *correlation coefficient* of X and Y, denoted as ρ_{XY}, is defined as follows:

$$\rho_{XY} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} \tag{3.45}$$

It can be shown that $\rho(X,Y)$ should be bound within $[-1...1]$, that is,

$$-1 \le \rho(X,Y) \le 1 \tag{3.46}$$

X and Y are said to be *positively correlated* if $\rho_{XY} > 0$, *negatively correlated* if $\rho_{XY} < 0$, and *uncorrelated* if $\rho_{XY} = 0$. It can also be shown that $Cov(X,Y)$ and ρ_{XY} must have the same sign; that is, both are positive, negative, or zero at the same time. When $E(XY) = 0$, the two random variables are called *orthogonal.*

There are several theorems pertaining to the basic properties of covariance and correlation. We list here the most important ones:

Theorem 1 For any random variables X and Y

$$Cov(X,Y) = E(XY) - E(X)E(Y) \tag{3.47}$$

Theorem 2 If X and Y are independent random variables, then

$$Cov(X,Y) = \rho_{XY} = 0$$

Theorem 3 Suppose X is a random variable and Y is a linear function of X in the form of $Y = aX + b$ for some constant a and b, where $a \neq 0$. If $a > 0$, then $\rho_{XY} = 1$. If $a < 0$, then $\rho_{XY} = -1$. Sometimes, ρ_{XY} is referred to as the amount of linear dependency between random variables X and Y.

Theorem 4 For any random variables X and Y,

$$Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y) \tag{3.48}$$

Theorem 5 If $X_1, \cdots, X_n$ are random variables, then

$$Var(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} Var(X_i) + 2\sum_{i=1}^{n}\sum_{j=1}^{i-1} Cov(X_i, X_j) \tag{3.49}$$

3.1.5. Random Vectors and Multivariate Distributions

When a random variable is a vector rather than a scalar, it is called a random vector and we often use boldface variable like $\mathbf{X} = (X_1, \cdots, X_n)$ to indicate that it is a random vector. It is said that n random variables $X_1, \cdots, X_n$ have a *discrete joint distribution* if the random vector $\mathbf{X} = (X_1, \cdots, X_n)$ can have only a finite number or an infinite sequence of different

values $(x_1,\cdots,x_n)$ in R^n. The joint pf of $X_1,\cdots,X_n$ is defined to be the function f_X such that for any point $(x_1,\cdots,x_n)\in R^n$,

$$f_X(x_1,\cdots,x_n)=P(X_1=x_1,\cdots,X_n=x_n) \tag{3.50}$$

Similarly, it is said that n random variables $X_1,\cdots,X_n$ have a *continuous joint distribution* if there is a nonnegative function f defined on R^n such that for any subset $A\subset R^n$,

$$P[(X_1,\cdots,X_n)\in A]=\int_A\cdots\int f_X(x_1,\cdots,x_n)dx_1\cdots dx_n \tag{3.51}$$

The *joint distribution function* can also be defined similarly for n random variables $X_1,\cdots,X_n$ as follows:

$$F_X(x_1,\cdots,x_n)=P(X_1\le x_1,\cdots,X_n\le x_n) \tag{3.52}$$

The concept of mean and variance for a random vector can be generalized into *mean vector* and *covariance matrix*. Supposed that $\mathbf{X}$ is an n-dimensional random vector with components $X_1,\cdots,X_n$, under matrix representation, we have

$$\mathbf{X}=\begin{bmatrix}X_1\\ \vdots\\ X_n\end{bmatrix} \tag{3.53}$$

The *expectation (mean) vector* $E(\mathbf{X})$ of random vector $\mathbf{X}$ is an n-dimensional vector whose components are the expectations of the individual components of $\mathbf{X}$, that is,

$$E(\mathbf{X})=\begin{bmatrix}E(X_1)\\ \vdots\\ E(X_n)\end{bmatrix} \tag{3.54}$$

The *covariance matrix* $Cov(\mathbf{X})$ of random vector $\mathbf{X}$ is defined to be an $n\times n$ matrix such that the element in the i^{th} row and j^{th} column is $Cov(X_i,Y_j)$, that is,

$$Cov(\mathbf{X})=\begin{bmatrix}Cov(X_1,X_1) & \cdots & Cov(X_1,X_n)\\ \vdots & & \vdots\\ Cov(X_n,X_1) & \cdots & Cov(X_n,X_n)\end{bmatrix}=E\left[[X-E(X)][X-E(X)]^t\right] \tag{3.55}$$

It should be emphasized that the n diagonal elements of the covariance matrix $Cov(\mathbf{X})$ are actually the variances of $X_1,\cdots,X_n$. Furthermore, since the covariance is symmetric, i.e., $Cov(X_i,X_j)=Cov(X_j,X_i)$, the covariance matrix $Cov(\mathbf{X})$ must be a symmetric matrix.

There is an important theorem regarding the mean vector and covariance matrix for a linear transformation of random vector $\mathbf{X}$. Suppose $\mathbf{X}$ is an n-dimensional vector as specified by Eq. (3.53), with mean vector $E(\mathbf{X})$ and covariance matrix $Cov(\mathbf{X})$. Now, assume $\mathbf{Y}$ is a m-dimensional random vector which is a linear transform of random vector $\mathbf{X}$ by the

relation: $\mathbf{Y}=\mathbf{AX}+\mathbf{B}$, where $\mathbf{A}$ is a $m\times n$ transformation matrix whose elements are constants, and $\mathbf{B}$ is a m-dimensional constant vector. Then we have the following two equations:

$$E(\mathbf{Y})=\mathbf{A}E(\mathbf{X})+\mathbf{B} \tag{3.56}$$

$$Cov(\mathbf{Y})=\mathbf{A}Cov(\mathbf{X})\mathbf{A}^t \tag{3.57}$$

3.1.6. Some Useful Distributions

In the following two sections, we will introduce several useful distributions that are widely used in applications of probability and statistics, particularly in spoken language systems.

3.1.6.1. Uniform Distributions

The simplest distribution is uniform distribution where the pf or pdf is a constant function. For uniform discrete random variable X, which only takes possible values from $\{x_i \mid 1\le i\le n\}$, the pf for X is

$$P(X=x_i)=\frac{1}{n} \quad 1\le i\le n \tag{3.58}$$

For uniform continuous random variable X, which only takes possible values in the real interval $[a,b]$, as shown in Figure 3.5, the pdf for X is

$$f(x)=\frac{1}{b-a} \quad a\le x\le b \tag{3.59}$$

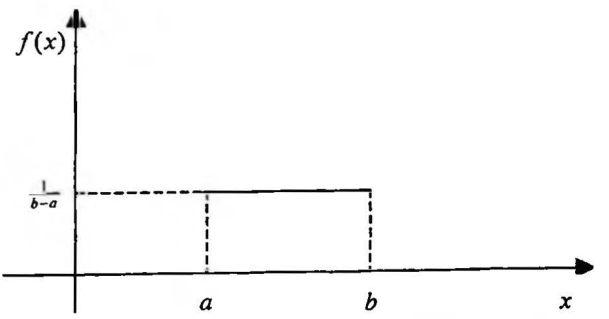

Figure 3.5 A uniform distribution for pdf in Eq. (3.59).

3.1.6.2. Binomial Distributions

The *binomial distribution* is used to describe binary-decision events. For example, suppose that a single coin toss will produce heads with probability p and produce tails with probability $1-p$. Now, if we toss the same coin n times and let X denote the number of heads observed, then the random variable X has the following *binomial* pf:

$$P(X = x) = f(x \mid n, p) = \binom{n}{x} p^x (1-p)^{n-x} \tag{3.60}$$

It can be shown that the mean and variance of a binomial distribution are:

$$E(X) = np \tag{3.61}$$

$$Var(X) = np(1-p) \tag{3.62}$$

Figure 3.6 illustrates three binomial distributions with p = 0.2, 0.3, and 0.4, and n = 10.

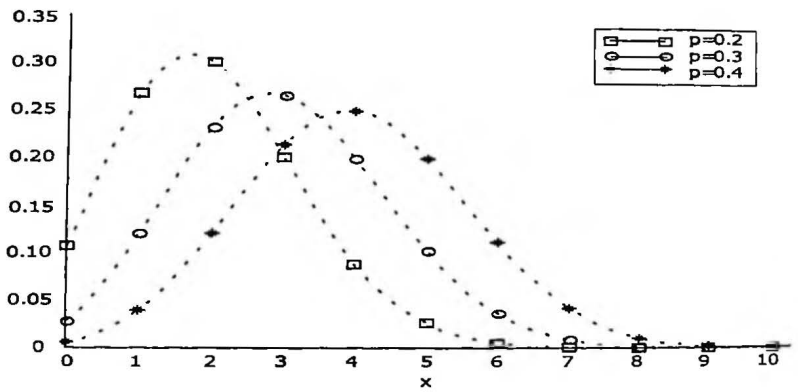

Figure 3.6 Three binomial distributions with p = 0.2, 0.3, and 0.4, and n = 10.

3.1.6.3. Geometric Distributions

The geometric distribution is related to the binomial distribution. As in the independent coin toss example, heads has a probability p and tails has a probability $1-p$. The geometric distribution is to model the time until tails appears. Let the random variable X

be the time (the number of tosses) until the first tail-up is shown. The pdf of X is in the following form:

$$P(X = x) = f(x \mid p) = p^{x-1}(1-p) \quad x = 1, 2, \ldots \text{ and } 0 < p < 1 \tag{3.63}$$

The mean and variance of a geometric distribution are given by:

$$E(X) = \frac{1}{1-p} \tag{3.64}$$

$$Var(X) = \frac{1}{(1-p)^2} \tag{3.65}$$

One example for the geometric distribution is the distribution of the state duration for a hidden Markov model, as described in Chapter 8. Figure 3.7 illustrates three geometric distributions with $p = 0.1$, 0.4, and 0.7.

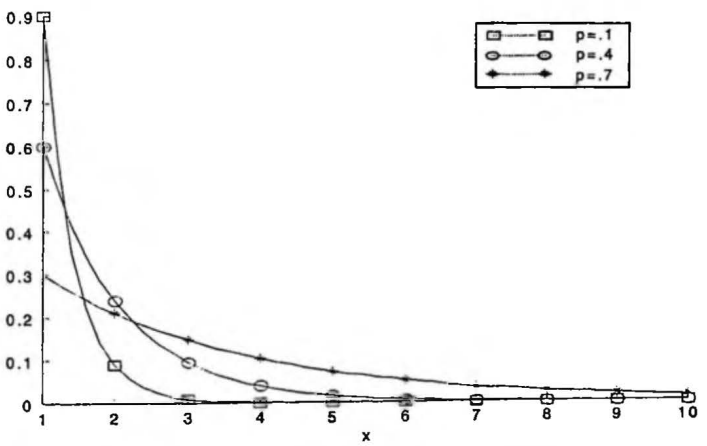

Figure 3.7 Three geometric distributions with different parameter p.

3.1.6.4. Multinomial Distributions

Suppose that a bag contains balls of k different colors, where the proportion of the balls of color i is p_i. Thus, $p_i > 0$ for $i = 1, \ldots, k$ and $\sum_{i=1}^{k} p_i = 1$. Now suppose that n balls are randomly selected from the bag and there are enough balls ($> n$) of each color. Let X_i denote

the number of selected balls that are of color i. The random vector $\mathbf{X} = (X_1, \ldots, X_k)$ is said to have a *multinomial distribution* with parameters n and $\mathbf{p} = (p_1, \ldots, p_k)$. For a vector $\mathbf{x} = (x_1, \ldots, x_k)$, the pf of $\mathbf{X}$ has the following form:

$$P(\mathbf{X} = \mathbf{x}) = f(\mathbf{x} \mid n, \mathbf{p}) = \begin{cases} \dfrac{n!}{x_1! \ldots x_k!} p_1^{x_1} \ldots p_k^{x_k} & \text{where } x_i \geq 0 \ \forall i = 1 \ldots k \\ & \text{and } x_1 + \cdots + x_k = n \\ 0 & \text{otherwise} \end{cases} \tag{3.66}$$

It can be shown that the mean, variance and covariance of the multinomial distribution are:

$$E(X_i) = np_i \quad \text{and} \quad Var(X_i) = np_i(1 - p_i) \quad \forall i = 1, \ldots, k \tag{3.67}$$

$$Cov(X_i, X_j) = -np_i p_j \tag{3.68}$$

Figure 3.8 shows a multinomial distribution with $n = 10$, $p_1 = 0.2$, and $p_2 = 0.3$. Since there are only two free parameters x_1 and x_2, the graph is illustrated only using x_1

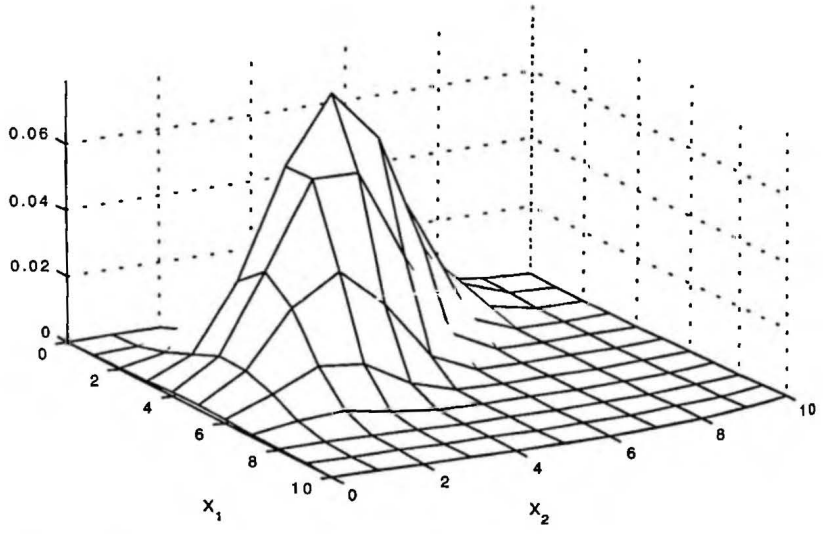

Figure 3.8 A multinomial distribution with n=10, $p_1 = 0.2$, and $p_2 = 0.3$.

and x_2 as axis. Multinomial distributions are typically used with the χ^2 test that is one of the most widely used goodness-of-fit hypotheses testing procedures described in Section 3.3.3.

3.1.6.5. Poisson Distributions

Another popular discrete distribution is the *Poisson distribution*. The random variable X has a Poisson distribution with mean λ ($\lambda > 0$) if the pf of X has the following form:

$$P(X = x) = f(x \mid \lambda) = \begin{cases} \dfrac{e^{-\lambda}\lambda^x}{x!} & \text{for } x = 0,1,2,\ldots \\ 0 & \text{otherwise} \end{cases} \tag{3.69}$$

The mean and variance of a Poisson distribution are the same and equal λ:

$$E(X) = Var(X) = \lambda \tag{3.70}$$

Figure 3.9 illustrates three Poisson distributions with $\lambda = 1$, 2, and 4. The Poisson distribution is typically used in queuing theory, where x is the total number of occurrences of some phenomenon during a fixed period of time or within a fixed region of space. Examples include the number of telephone calls received at a switchboard during a fixed period of time. In speech recognition, the Poisson distribution is used to model the duration for a phoneme.

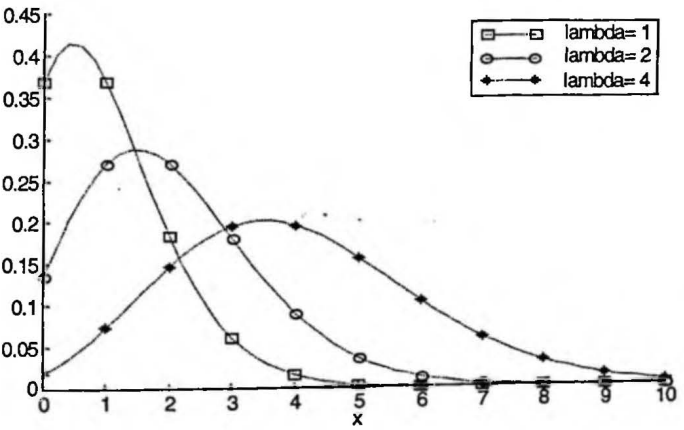

Figure 3.9 Three Poisson distributions with $\lambda = 1$, 2, and 4.

3.1.6.6. Gamma Distributions

A continuous random variable X is said to have a *gamma distribution* with parameters α and β ($\alpha > 0$ and $\beta > 0$) if X has a continuous pdf of the following form:

$$f(x \mid \alpha, \beta) = \begin{cases} \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} & x > 0 \\ 0 & x \le 0 \end{cases} \tag{3.71}$$

where

$$\Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1} e^{-x} dx \tag{3.72}$$

It can be shown that the function Γ is a factorial function when α is a positive integer.

$$\Gamma(n) = \begin{cases} (n-1)! & n = 2,3,... \\ 1 & n = 1 \end{cases} \tag{3.73}$$

The mean and variance of a gamma distribution are:

$$E(X) = \frac{\alpha}{\beta} \text{ and } Var(X) = \frac{\alpha}{\beta^2} \tag{3.74}$$

Figure 3.10 illustrates three gamma distributions with $\beta = 1.0$ and $\alpha = 2.0$, 3.0, and 4.0. There is an interesting theorem associated with gamma distributions. If the random variables $X_1, \ldots, X_k$ are independent and each random variable X_i has a gamma distribution with parameters α_i and β, then the sum $X_1 + \cdots + X_k$ also has a gamma distribution with parameters $\alpha_1 + \cdots + \alpha_k$ and β.

A special case of gamma distribution is called *exponential distribution*. A continuous random variable X is said to have an *exponential distribution* with parameters β ($\beta > 0$) if X has a continuous pdf of the following form:

$$f(x \mid \beta) = \begin{cases} \beta e^{-\beta x} & x > 0 \\ 0 & x \le 0 \end{cases} \tag{3.75}$$

It is clear that the exponential distribution is a gamma distribution with $\alpha = 1$. The mean and variance of the exponential distribution are:

$$E(X) = \frac{1}{\beta} \text{ and } Var(X) = \frac{1}{\beta^2} \tag{3.76}$$

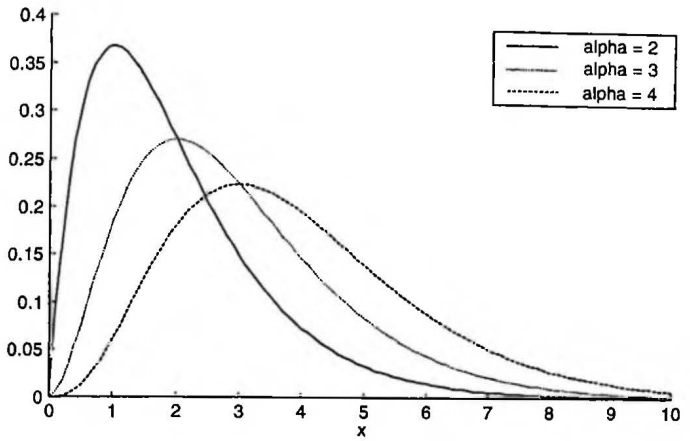

Figure 3.10 Three Gamma distributions with $\beta = 1.0$ and $\alpha = 2.0, 3.0$, and 4.0.

Figure 3.11 shows three exponential distributions with $\beta = 1.0$, 0.6, and 0.3. The exponential distribution is often used in queuing theory for the distributions of the duration of a service or the inter-arrival time of customers. It is also used to approximate the distribution of the life of a mechanical component.

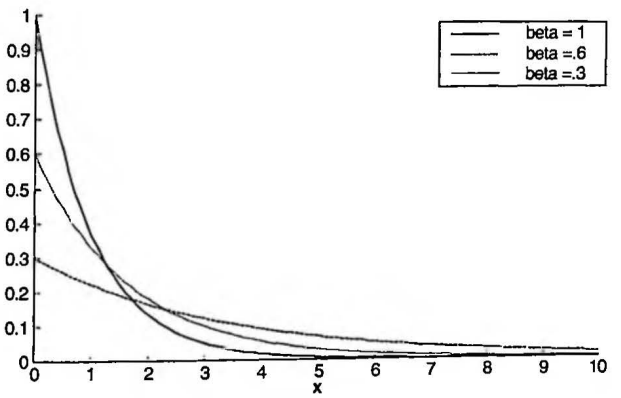

Figure 3.11 Three exponential distributions with $\beta = 1.0$, 0.6, and 0.3.

3.1.7. Gaussian Distributions

Gaussian distribution is by far the most important probability distribution mainly because many scientists have observed that the random variables studied in various physical experiments (including speech signals), often have distributions that are approximately Gaussian. The Gaussian distribution is also referred to as normal distribution. A continuous random variable X is said to have a *Gaussian distribution* with mean μ and variance σ^2 $(\sigma > 0)$ if X has a continuous pdf in the following form:

$$f(x \mid \mu, \sigma^2) = N(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2} \right] \tag{3.77}$$

It can be shown that μ and σ^2 are indeed the mean and the variance for the Gaussian distribution. Some examples of Gaussians can be found in Figure 3.4.

The use of Gaussian distributions is justified by the *Central Limit Theorem*, which states that observable events considered to be a consequence of many unrelated causes with no single cause predominating over the others, tend to follow the Gaussian distribution [6].

It can be shown from Eq. (3.77) that the Gaussian $f(x \mid \mu, \sigma^2)$ is symmetric with respect to $x = \mu$. Therefore, μ is both the mean and the median of the distribution. Moreover, μ is also the mode of the distribution, i.e., the pdf $f(x \mid \mu, \sigma^2)$ attains its maximum at the mean point $x = \mu$.

Several Gaussian pdfs with the same mean μ, but different variances are illustrated in Figure 3.4. Readers can see that the curve has a *bell* shape. The Gaussian pdf with a small variance has a high peak and is very concentrated around the mean μ, whereas the Gaussian pdf with a large variance is relatively flat and is spread out more widely over the x-axis.

If the random variable X is a *Gaussian distribution* with mean μ and variance σ^2, then any linear function of X also has a Gaussian distribution. That is, if $Y = aX + b$, where a and b are constants and $a \neq 0$, Y has a Gaussian distribution with mean $a\mu + b$ and variance $a^2\sigma^2$. Similarly, the sum $X_1 + \cdots + X_k$ of random variables $X_1, \ldots, X_k$, where each random variable X_i has a Gaussian distribution, is also a Gaussian distribution.

3.1.7.1. Standard Gaussian Distributions

The Gaussian distribution with mean 0 and variance 1, denoted as $N(0,1)$, is called the *standard Gaussian distribution* or *unit Gaussian distribution*. Since the linear transformation of a Gaussian distribution is still a Gaussian distribution, the behavior of a Gaussian distribution can be solely described using a standard Gaussian distribution. If the random variable X is a Gaussian distribution with mean μ and variance σ^2, that is, $X \sim N(\mu, \sigma^2)$, it can be shown that

$$Z = \frac{X - \mu}{\sigma} \sim N(0,1) \tag{3.78}$$

Based on Eq. (3.78), the following property can be shown:

$$P(|X-\mu| \le k\sigma) = P(|Z| \le k) \tag{3.79}$$

Equation (3.79) demonstrates that every Gaussian distribution contains the same total amount of probability within any fixed number of standard deviations of its mean.

3.1.7.2. The Central Limit Theorem

If random variables $X_1,\ldots,X_n$ are i.i.d. according to a common distribution function with mean μ and variance σ^2, then as the random sample size n approaches ∞, the following random variable has a distribution converging to the standard Gaussian distribution:

$$Y_n = \frac{n(\bar{X}_n - \mu)}{\sqrt{n\sigma^2}} \sim N(0,1) \tag{3.80}$$

where $\bar{X}_n$ is the sample mean of random variables $X_1,\ldots,X_n$ as defined in Eq. (3.41).

Based on Eq. (3.80), the sample mean random variable $\bar{X}_n$ can be approximated by a Gaussian distribution with mean μ and variance σ^2/n.

The central limit theorem above is applied to i.i.d. random variables $X_1,\ldots,X_n$. A. Liapounov in 1901 derived another central limit theorem for independent but not necessarily identically distributed random variables $X_1,\ldots,X_n$. Suppose $X_1,\ldots,X_n$ are independent random variables and $E(|X_i-\mu_i|^3) < \infty$ for $1 \le i \le n$; the following random variable will converge to standard Gaussian distribution when $n \to \infty$.

$$Y_n = \left(\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \mu_i\right) \Big/ \left(\sum_{i=1}^{n} \sigma_i^2\right)^{1/2} \tag{3.81}$$

In other words, the sum of random variables $X_1,\ldots,X_n$ can be approximated by a Gaussian distribution with mean $\sum_{i=1}^{n} \mu_i$ and variance $\left(\sum_{i=1}^{n} \sigma_i^2\right)^{1/2}$.

Both central limit theorems essentially state that regardless of their original individual distributions, the sum of many independent random variables (effects) tends to be distributed like a Gaussian distribution as the number of random variables (effects) becomes large.

3.1.7.3. Multivariate Mixture Gaussian Distributions

When $\mathbf{X} = (X_1,\ldots,X_n)$ is an n-dimensional continuous random vector, the multivariate Gaussian pdf has the following form:

$$f(\mathbf{X}=\mathbf{x}\,|\,\boldsymbol{\mu},\boldsymbol{\Sigma}) = N(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right] \tag{3.82}$$

where $\boldsymbol{\mu}$ is the n-dimensional mean vector, $\boldsymbol{\Sigma}$ is the $n \times n$ covariance matrix, and $|\boldsymbol{\Sigma}|$ is the determinant of the covariance matrix $\boldsymbol{\Sigma}$.

$$\boldsymbol{\mu} = E(\mathbf{x}) \tag{3.83}$$

$$\boldsymbol{\Sigma} = E\left[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^t\right] \tag{3.84}$$

More specifically, the i-j^{th} element σ_{ij}^2 of covariance matrix $\boldsymbol{\Sigma}$ can be specified as follows:

$$\sigma_{ij}^2 = E\left[(x_i - \mu_i)(x_j - \mu_j)\right] \tag{3.85}$$

If $X_1, \ldots, X_n$ are independent random variables, the covariance matrix $\boldsymbol{\Sigma}$ is reduced to diagonal covariance where all the off-diagonal entries are zero. The distribution can be regarded as n independent scalar Gaussian distributions. The joint pdf is the product of all the individual scalar Gaussian pdfs. Figure 3.12 shows a two-dimensional multivariate Gaussian distribution with independent random variables x_1 and x_2 with the same variance. Figure 3.13 shows another two-dimensional multivariate Gaussian distribution with independent random variables x_1 and x_2 that have different variances.

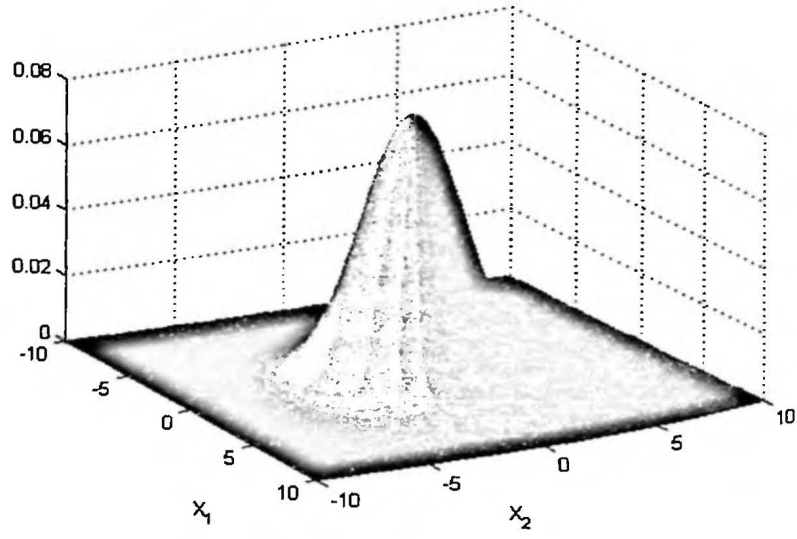

Figure 3.12 A two-dimensional multivariate Gaussian distribution with independent random variables x_1 and x_2 that have the same variance.

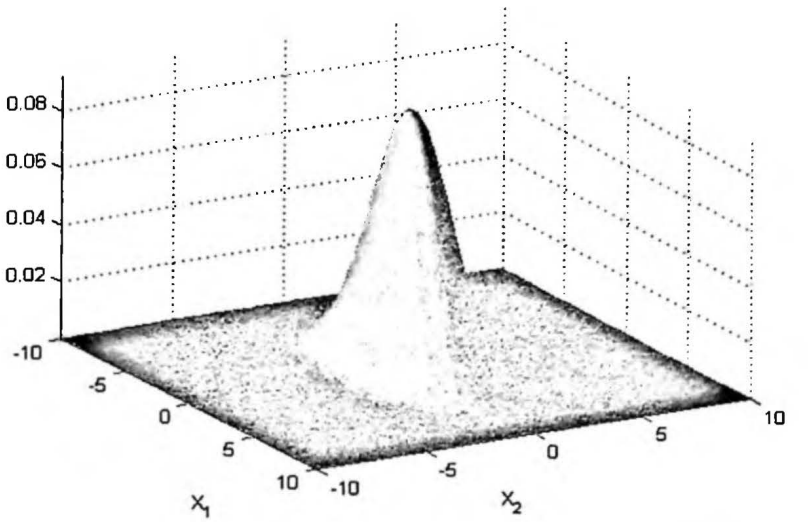

Figure 3.13 Another two-dimensional multivariate Gaussian distribution with independent random variable x_1 and x_2 which have different variances.

Although Gaussian distributions are unimodal,[2] more complex distributions with multiple local maxima can be approximated by Gaussian mixtures:

$$f(\mathbf{x}) = \sum_{k=1}^{K} c_k N_k(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{3.86}$$

where c_k, the mixture weight associated with kth Gaussian component, is subject to the following constraint:

$$c_k \geq 0 \text{ and } \sum_{k=1}^{K} c_k = 1$$

Gaussian mixtures with enough mixture components can approximate any distribution. Throughout this book, most continuous probability density functions are modeled with Gaussian mixtures.

3.1.7.4. χ^2 Distributions

The gamma distribution with parameters α and β is defined 1). For any given positive integer n, the gamma distribution for which $\alpha = n/2$ a is called the χ^2

[2] A unimodal distribution has a single maximum (bump) for the distribution. F ibution, the maximum occurs at the mean.

distribution with n degrees of freedom. It follows from Eq. (3.71) that the pdf for the χ^2 distribution is

$$f(x|n)=\begin{cases}\dfrac{1}{2^{n/2}\Gamma(n/2)}x^{(n/2)-1}e^{-x/2} & x>0\\ 0 & x\le 0\end{cases} \tag{3.87}$$

χ^2 distributions are important in statistics because they are closely related to random samples of Gaussian distribution. They are widely applied in many important problems of statistical inference and hypothesis testing. Specifically, if the random variables $X_1,\ldots,X_n$ are independent and identically distributed, and if each of these variables has a standard Gaussian distribution, then the sum of square $X_1^2+\ldots+X_n^2$ can be proved to have a χ^2 distribution with n degrees of freedom. Figure 3.14 illustrates three χ^2 distributions with $n=2$, 3, and 4.

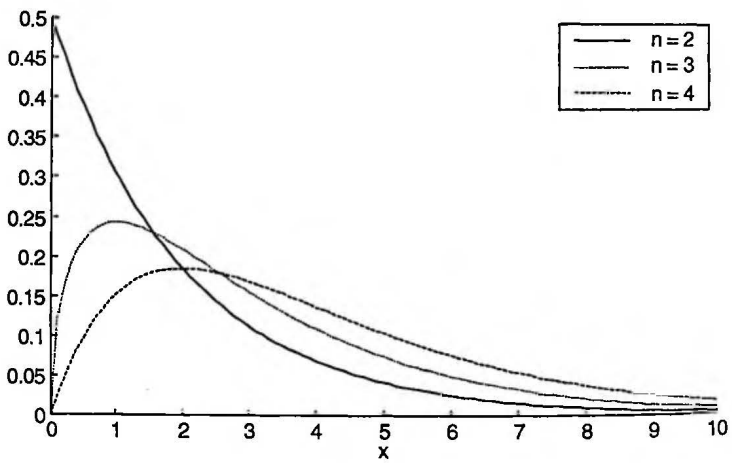

Figure 3.14 Three χ^2 distributions with $n=2$, 3, and 4.

The mean and variance for the χ^2 distribution are

$$E(X)=n \text{ and } Var(X)=2n \tag{3.88}$$

Following the additivity property of the gamma distribution, the χ^2 distribution also has the additivity property. That is, if the random variables $X_1,\ldots,X_n$ are independent and if X_i has a χ^2 distribution with k_i degrees of freedom, the sum $X_1+\ldots+X_n$ has a χ^2 distribution with $k_1+\ldots+k_n$ degrees of freedom.

3.1.7.5. Log-Normal Distribution

Let x be a Gaussian random variable with mean μ_x and standard deviation σ_x, then

$$y = e^x \tag{3.89}$$

follows the *lognormal* distribution

$$f(y \mid \mu_x, \sigma_x) = \frac{1}{y\sigma_x\sqrt{2\pi}} \exp\left\{-\frac{(\ln y - \mu_x)^2}{2\sigma_x^2}\right\} \tag{3.90}$$

shown in Figure 3.15, and whose mean is given by

$$\begin{aligned} \mu_y = E\{y\} = E\{e^x\} &= \int_{-\infty}^{\infty} \exp\{x\} \frac{1}{\sqrt{2\pi}\sigma_x} \exp\left\{-\frac{(x-\mu_x)^2}{2\sigma_x^2}\right\} dx \\ &= \int_{-\infty}^{\infty} \exp\{\mu_x + \sigma_x^2/2\} \frac{1}{\sqrt{2\pi}\sigma_x} \exp\left\{-\frac{(x-(\mu_x+\sigma_x^2)^2}{2\sigma_x^2}\right\} dx = \exp\{\mu_x + \sigma_x^2/2\} \end{aligned} \tag{3.91}$$

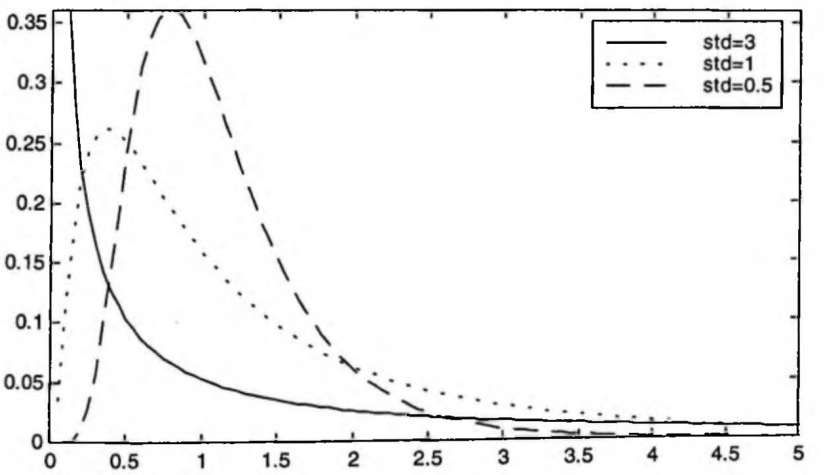

Figure 3.15 Lognormal distribution for $\mu_x = 0$ and $\sigma_x = 3$, 1, and 0.5, according to Eq. (3.90).

where we have rearranged the quadratic form of x and made use of the fact that the total probability mass of a Gaussian is 1. Similarly, the second order moment of y is given by

$$E\{y^2\} = \int_{-\infty}^{\infty} \exp\{2x\} \frac{1}{\sqrt{2\pi}\sigma_x} \exp\left\{-\frac{(x-\mu_x)^2}{2\sigma_x^2}\right\} dx$$
$$= \int_{-\infty}^{\infty} \exp\{2\mu_x + 2\sigma_x^2\} \frac{1}{\sqrt{2\pi}\sigma_x} \exp\left\{-\frac{(x-(\mu_x+2\sigma_x^2))^2}{2\sigma_x^2}\right\} dx = \exp\{2\mu_x + 2\sigma_x^2\} \tag{3.92}$$

and thus the variance of y is given by

$$\sigma_y^2 = E\{y^2\} - (E\{y\})^2 = \mu_y^2 (\exp\{\sigma_x^2\} - 1) \tag{3.93}$$

Similarly, if $\mathbf{x}$ is a Gaussian random vector with mean $\boldsymbol{\mu}_x$ and covariance matrix $\boldsymbol{\Sigma}_x$, then random vector $\mathbf{y} = e^{\mathbf{x}}$ is log-normal with mean and covariance matrix [8] given by

$$\boldsymbol{\mu}_y[i] = \exp\{\boldsymbol{\mu}_x[i] + \boldsymbol{\Sigma}_x[i,i]/2\}$$
$$\boldsymbol{\Sigma}_y[i,j] = \boldsymbol{\mu}_y[i]\boldsymbol{\mu}_y[j](\exp\{\boldsymbol{\Sigma}_x[i,j]\} - 1) \tag{3.94}$$

using a similar derivation as in Eqs. (3.91) to (3.93).

3.2. ESTIMATION THEORY

Estimation theory and *significance testing* are two of the most important theories and methods of *statistical inference*. In this section, we describe estimation theory while significance testing is covered in the next section. A problem of statistical inference is one in which data generated in accordance with some unknown probability distribution must be analyzed, and some type of inference about the unknown distribution must be made. In a problem of statistical inference, any characteristic of the distribution generating the experimental data, such as the mean μ and variance σ^2 of a Gaussian distribution, is called a parameter of the distribution. The set Ω of all possible values of a *parameter* Φ or a group of parameters $\Phi_1, \Phi_2, \ldots, \Phi_n$ is called the *parameter space*. In this section we focus on how to estimate the parameter Φ from sample data.

Before we describe various estimation methods, we introduce the concept and nature of the estimation problems. Suppose that a set of random variables $\mathbf{X} = \{X_1, X_2, \ldots, X_n\}$ is iid according to a pdf $p(x|\Phi)$ where the value of the parameter Φ is unknown. Now, suppose also that the value of Φ must be estimated from the observed values in the sample. An *estimator* of the parameter Φ, based on the random variables $X_1, X_2, \ldots, X_n$, is a real-valued function $\theta(X_1, X_2, \ldots, X_n)$ that specifies the estimated value of Φ for each possible set of values of $X_1, X_2, \ldots, X_n$. That is, if the sample values of $X_1, X_2, \ldots, X_n$ turn out to be $x_1, x_2, \ldots, x_n$, then the estimated value of Φ will be $\theta(x_1, x_2, \ldots, x_n)$.

We need to distinguish between *estimator*, *estimate*, and *estimation*. An estimator $\theta(X_1, X_2, \ldots, X_n)$ is a function of the random variables, whose probability distribution can be derived from the joint distribution of $X_1, X_2, \ldots, X_n$. On the other hand, an estimate is a specific value $\theta(x_1, x_2, \ldots, x_n)$ of the estimator that is determined by using some specific

sample values $x_1, x_2, \ldots, x_n$. Estimation is usually used to indicate the process of obtaining such an estimator for the set of random variables or an estimate for the set of specific sample values. If we use the notation $\mathbf{X} = \{X_1, X_2, \ldots, X_n\}$ to represent the vector of random variables and $\mathbf{x} = \{x_1, x_2, \ldots, x_n\}$ to represent the vector of sample values, an estimator can be denoted as $\theta(\mathbf{X})$ and an estimate $\theta(\mathbf{x})$. Sometimes we abbreviate an estimator $\theta(\mathbf{X})$ by just the symbol θ.

In the following four sections we describe and compare three different estimators (estimation methods). They are *minimum mean square estimator*, *maximum likelihood estimator*, and *Bayes' estimator*. The first one is often used to estimate the random variable itself, while the latter two are used to estimate the parameters of the distribution of the random variables.

3.2.1. Minimum/Least Mean Squared Error Estimation

Minimum mean squared error (MMSE) estimation and *least squared error* (LSE) estimation are important methods for random variables since the goal (minimize the squared error) is an intuitive one. In general, two random variables X and Y are i.i.d. according to some pdf $f_{X,Y}(x,y)$. Suppose that we perform a series of experiments and observe the value of X. We want to find a transformation $\hat{Y} = g(X)$ such that we can predict the value of the random variable Y. The following quantity can measure the goodness of such a transformation:

$$E(Y-\hat{Y})^2 = E(Y-g(X))^2 \tag{3.95}$$

This quantity is called *mean squared error* (MSE) because it is the mean of the squared error of the predictor $g(X)$. The criterion of minimizing the mean squared error is a good one for picking the predictor $g(X)$. Of course, we usually specify the class of function G, from which $g(X)$ may be selected. In general, there is a parameter vector $\mathbf{\Phi}$ associated with the function $g(X)$, so the function can be expressed as $g(X,\mathbf{\Phi})$. The process to find the parameter vector $\hat{\mathbf{\Phi}}_{MMSE}$ that minimizes the mean of the squared error is called *minimum mean squared error estimation* and $\hat{\mathbf{\Phi}}_{MMSE}$ is called the *minimum mean squared error estimator*. That is,

$$\hat{\mathbf{\Phi}}_{MMSE} = \arg\min_{\mathbf{\Phi}} \left[E\left[(Y - g(X,\mathbf{\Phi}))^2 \right] \right] \tag{3.96}$$

Sometimes, the joint distribution of random variables X and Y is not known. Instead, samples of (x,y) pairs may be observable. In this case, the following criterion can be used instead,

$$\hat{\mathbf{\Phi}}_{LSE} = \arg\min_{\mathbf{\Phi}} \sum_{i=1}^{n} [y_i - g(x_i,\mathbf{\Phi})]^2 \tag{3.97}$$

The argument of the minimization in Eq. (3.97) is called *sum-of-squared-error* (SSE) and the process of finding the parameter vector $\hat{\mathbf{\Phi}}_{LSE}$, which satisfies the criterion is called *least*

squared error estimation or *minimum squared error estimation*. LSE is a powerful mechanism for curve fitting, where the function $g(x,\Phi)$ describes the observation pairs (x_i, y_i). In general, there are more points (n) than the number of free parameters in function $g(x,\Phi)$, so the fitting is over-determined. Therefore, no exact solution exists, and LSE fitting becomes necessary.

It should be emphasized that MMSE and LSE are actually very similar and share similar properties. The quantity in Eq. (3.97) is actually n times the sample mean of the squared error. Based on the law of large numbers, when the joint probability $f_{X,Y}(x,y)$ is uniform or the number of samples approaches to infinity, MMSE and LSE are equivalent.

For the class of functions, we consider the following three cases:

- Constant functions, i.e.,

$$G_c = \{g(x) = c, c \in R\} \tag{3.98}$$

- Linear functions, i.e.,

$$G_l = \{g(x) = ax + b \quad a,b \in R\} \tag{3.99}$$

- Other non-linear functions G_{nl}

3.2.1.1. MMSE/LSE for Constant Functions

When $\hat{Y} = g(x) = c$, Eq. (3.95) becomes

$$E(Y - \hat{Y})^2 = E(Y - c)^2 \tag{3.100}$$

To find the MMSE estimate for c, we take the derivatives of both sides in Eq. (3.100) with respect to c and equate it to 0. The MMSE estimate c_{MMSE} is given as

$$c_{MMSE} = E(Y) \tag{3.101}$$

and the minimum mean squared error is exactly the variance of Y, $Var(Y)$.

For the LSE estimate of c, the quantity in Eq. (3.97) becomes

$$\min \sum_{i=1}^{n} [y_i - c]^2 \tag{3.102}$$

Similarly, the LSE estimate c_{LSE} can be obtained as follows:

$$c_{LSE} = \frac{1}{n} \sum_{i=1}^{n} y_i \tag{3.103}$$

The quantity in Eq. (3.103) is the sample mean.

3.2.1.2. MMSE and LSE for Linear Functions

When $\hat{Y} = g(x) = ax + b$, Eq. (3.95) becomes

$$e(a,b) = E(Y - \hat{Y})^2 = E(Y - ax - b)^2 \tag{3.104}$$

To find the MMSE estimate of a and b, we can first set

$$\frac{\partial e}{\partial a} = 0, \text{ and } \frac{\partial e}{\partial b} = 0 \tag{3.105}$$

and solve the two linear equations. Thus, we can obtain

$$a = \frac{\text{cov}(X,Y)}{Var(X)} = \rho_{XY}\frac{\sigma_Y}{\sigma_X} \tag{3.106}$$

$$b = E(Y) - \rho_{XY}\frac{\sigma_Y}{\sigma_X}E(X) \tag{3.107}$$

For LSE estimation, we assume that the sample $\mathbf{x}$ is a d-dimensional vector for generality. Assuming we have n sample-vectors $(\mathbf{x}_i, y_i) = (x_i^1, x_i^2, \cdots, x_i^d, y_i)$, $i = 1 \ldots n$, a linear function can be represented as

$$\hat{\mathbf{Y}} = \mathbf{XA} \text{ or } \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1^1 & \cdots & x_1^d \\ 1 & x_2^1 & \cdots & x_2^d \\ \vdots & \vdots & & \vdots \\ 1 & x_n^1 & \cdots & x_n^d \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_d \end{pmatrix} \tag{3.108}$$

The sum of squared error can then be represented as

$$e(\mathbf{A}) = \| \hat{\mathbf{Y}} - \mathbf{Y} \|^2 = \sum_{i=1}^{n} \left(\mathbf{A}^t \mathbf{x}_i - y_i \right)^2 \tag{3.109}$$

A closed-form solution of the LSE estimate of $\mathbf{A}$ can be obtained by taking the gradient of $e(\mathbf{A})$,

$$\nabla e(\mathbf{A}) = \sum_{i=1}^{n} 2(\mathbf{A}^t \mathbf{x}_i - y_i)\mathbf{x}_i = 2\mathbf{X}^t(\mathbf{XA} - \mathbf{Y}) \tag{3.110}$$

and equating it to zero. This yields the following equation:

$$\mathbf{X}^t\mathbf{XA} = \mathbf{X}^t\mathbf{Y} \tag{3.111}$$

Thus the LSE estimate $\mathbf{A}_{LSE}$ will be of the following form: