# A Multi-tier Framework for Understanding Spoken Language

Steven Greenberg
The Speech Institute
steveng@cogsci.berkeley.edu

**To appear in: Listening to Speech: An Auditory Perspective**
Steven Greenberg and William Ainsworth, editors
Lawrence Erlbaum Associates Publisher

## 1. INTRODUCTION

Language can be approached from many different vantage points – neuroanatomy, psychology, phonetics, hearing, vision, physics, information theory, formal logic, and so on. Most scientific and linguistic accounts have focused on defining units of analysis, such as the phoneme, the word, and the phrase. Some of the more ingenious, such as Articulation Theory (Fletcher and Galt, 1940; French and Steinberg, 1947; Fletcher, 1953; Allen, 1994) and the Speech Transmission Index (Houtgast and Steeneken, 1985) characterize speech communication mostly in terms of equations. To date, all such efforts have failed, largely because language is extraordinarily multi-dimensional and not particularly amenable to simplifying assumptions. The great American linguist, Edward Sapir, characterized the problem very succintly: "All grammars leak..." (Sapir, 1921, p. 39). No language can be captured entirely in terms of a closed system of equations; there are always exceptions to the rules. Humans are particularly adept at learning exceptions and accepting them as "normal."

In this sense, no single perspective is likely to explain key linguistic phenomena. However, a theoretical framework can delineate certain principles of potential utility for developing future-generation speech and audio technology. In addition, such a theory can be used to synthesize seemingly unrelated material into a coherent intellectual framework.

"Listening to Speech" focuses on hearing as an explanatory framework for understanding speech communication. This proposition is not universally accepted; many still argue that much of what makes speech unique is the human vocal tract (Lieberman, 1984, 1990). This chapter summarizes a broad range of data, including some of my own, consistent with the primacy of hearing for shaping the principal properties of spoken language. But, a theoretical orientation focused entirely on hearing is insufficient – more than the auditory system is required for a comprehensive framework. The acoustic signal is often supplemented by visual cues, and it is the combination of these sensory streams that often accounts for the robustness of speech communication, particularly in noisy environments (Massaro, 1987; Grant, Walden & Seitz, 1998; Faulkner & Rosen in this volume). However, a theory based entirely on sensory coding is also inadequate for decribing the richness of spoken language. Another dimension is required to complete the picture. This dimension is information, which is the raison d'etre of speech communication: "We speak to be heard in order to be understood" (Jakobson, Fant & Halle, 1963). Throughout this chapter I shall emphasize how information acts as a controlling factor in defining many properties of spoken language. The auditory system is particularly adept at encoding information in a variety of ways, some of them quite subtle. What also distinguishes this chapter from others in this book is its emphasis on how speech is actually spoken, based on manual phonetic and prosodic annotation of casual conversations (Greenberg, 1997, 1999). The annotated material provides an extremely useful tool with which to analyze the micro-structure of spoken language.

## 2. UNITS OF ANALYSIS

### *2.1 The Tyranny of Orthography*

In the Western world, most writing systems are based on the Roman alphabet (Sampson, 1985). In such orthographies, there is a strong tendency for a sound to be represented by a discrete symbol. The spaces between words serve as lexical delimiters, while the orthographic symbols represent the sound shape. Certain languages, such as Spanish, lend themselves easily to Roman orthography; these tongues have a relatively transparent grapheme-to-phoneme relationship – words are pronounced pretty much as they are spelled and with some measure of consistency. Other languages, English among them, are not so easily portrayed in terms of such symbols.

Orthography has had an enormous influence on linguistic models of spoken language (Greenberg, 2003). From the orthographic perspective, words are sequences of sounds represented as discrete symbols. The concept of the "phoneme" is derived from this orthographic framework, particularly in terms of its abstract representation. A word may not be pronounced exactly as written (a common occurrence in English), but its abstract, underlying form is supposedly immutable. It is mainly a matter of mapping a phoneme's surface manifestation to its abstract representation – precisely how this mapping is performed is usually unspecified.

Other writing systems, among them Chinese, Japanese, Sanskrit (Hindi), Hangul (Korean) and Arabic, do not use the phoneme as their basis of orthography (Sampson, 1985). In the Semitic orthographies, such as Hebrew and Arabic, vowels are essentially invisible (their identity is highly predictable from context). Each symbol in the Japanese katakana and hiragana represents a simple form of syllable (referred to as a "mora"). Orthographies generally concentrate their symbolic focus on a single level of analysis; but this does not mean other levels are irrelevant.

### *2.2 The Unreality of the Phoneme*

Within the theoretical framework described in this chapter, the phoneme is not a privileged unit; in fact, it does not even exist, except as a means of translating the output of other levels into a conventional linguistic form.

What are the relevant units of analysis, if not the phoneme (and phone)? The answers lie in the brain and the sensori-motor systems responsible for its interaction with the external world (Figure 25.1).

### *2.3 The Articulatory Basis of Speech*

We first consider the articulatory system, for it establishes a convenient frame of reference for what follows (see Boersma, 1998; Stevens, 1998; Diehl & Lindblom, 2004 for in-depth treatments of this topic). When a person speaks, it is not individual phones that are uttered, but syllables. Articulatory gestures associated with the opening and closing of the jaw and lips are synchronized to the syllable. Articulating a consonant separately from the adjacent vowel is not considered speech in most contexts.

Associated with the articulatory gesture (and the syllable) is a fluctuation in energy. When the lips are closed the energy is very low. As the lips open, the amount of energy coming from the mouth increases. During maximum oral aperture the energy reaches a peak, generally close to the center of the syllabic nucleus. As the lips begin to close, the energy is reduced; to what extent depends on other (prosodic) factors discussed in Section 5. This articulatory cycle is important, for it provides the production basis for the syllable, as well as the phonetic organization within this structural unit. The beginning of the articulatory cycle is known as the "onset," while the concluding phase is called the "coda." Generally, onsets and codas are composed of consonants, while vowels form the "nucleus." Onsets behave very differently from codas, even if composed of the same nominal segment (see Section 3; Greenberg, 2003).

### *2.4 Manner of Articulation*

Within a syllable, articulators can function in one of two basic ways. Their place of maximum articulatory constriction is discussed in Section 2.6. The other is their manner of production, which refers to how the sound is produced. For example, vowels are produced with a relatively open vocal tract, and the lips are usually open (except for rounded vowels). There is no firm occlusion and sound has a relatively free path through the vocal cavity and out of the mouth. The energy level of vowels is high, as much as 40 dB more intense than most consonantal segments.

At the opposite end of the energy spectrum are plosives and fricatives. Plosives (also known as "stops") are formed by a complete occlusion of the vocal tract, such as occurs in the sound [b], where the lips come together briefly. At the onset of a syllable, this short interval (usually 5-10 ms) is followed by a sudden release of energy, which dies down quickly, but is followed by a gradual increase in sound pressure that is substantially lower in amplitude than a vowel. Fricatives (e.g., [s], [z] and [f]) exhibit a substantial, though incomplete, occlusion that is often sustained for many tens of milliseconds. Their energy level is also substantially lower than vowels. Affricates combine elements of plosives and fricatives (e.g., [tʃ] "church", [dʒ] "judge"). Nasal segments, such as [n] and [m] are similar to vowels in certain ways, but air comes out through the nose rather than the mouth. As with plosives, there is an articulatory occlusion (the lips in [m], the tongue tip contacting the alveolar ridge in [n]), but because the airflow is not occluded, the sound pressure is almost as high as for vowels.

Liquids, such as [l] and [r] also have certain properties in common with vowels. The tongue tends to move somewhat differently (bunching or rolling laterally), creating a sound pressure slightly lower than vowels (particularly between 2 and 3 kHz). This slight reduction in energy is an important property of liquids, particularly when they separate syllables (see Section 6.2). In some languages (though not in English), [r] is trilled (Ladefoged, 2000), usually at the beginning of a syllable. When liquids occur at the end of a syllable, they often fuse with the preceding vowel becoming, for all intents and purposes, a vocalic constituent whose primary function is to serve as a syllable divider (see Section 6.2).

In articulatory terms, the concept of the phone (and phoneme) comes closest to manner of articulation. Temporally, manner is packaged in relatively discrete units much of the time and it is possible to segment the speech signal into phones largely by using manner of articulation as the underlying basis of the automatic classifiers (Chang, Wester & Greenberg, 2003). The basis for this discrete articulatory representation probably lies in the energy arc, which is discussed in Section 8.

### *2. 5 Voicing*

From perception's perspective, voicing is closely related to manner of articulation. Both affect the amplitude of the speech signal, though in different ways. Traditionally, voicing has been viewed primarily as a means to distinguish among closely related segments, such as [p] and [b], or as the medium by which pitch-relevant information is conveyed. From the perspective of multi-tier theory, voicing boosts the amplitude of the speech signal. The voiced parts of the syllable are considerably more intense than their unvoiced counterparts. The syllable nucleus is usually voiced, while the onset and coda may or may not be. In this manner, voicing may provide a cue for syllabic organization, as it always accompanies the most intense (i.e., sonorant) constituents. Moreover, unvoiced portions, when present, are always situated on the syllabic flanks. Within the syllable, voicing is restricted to a single, contiguous interval. Its focus is the nucleus, and it spreads both forward into the coda and back into the onset. Voicing is also highly sensitive to prosody, as discussed in Section 6.

### *2.6 Place of Articulation*

Place of articulation is the other key parameter of production; it refers to the locus of maximum articulatory constriction and is often associated with a specific phonetic constituent. The plosives [b], [d] and [g] differ from each other largely in terms of where the maximum constriction lies – at the lips in the case of [b], towards the back of the hard palate (the velar ridge) for [g], and somewhere in between for [d].

There are three special attributes of articulatory place that make this production parameter so interesting and important.

First, there is supposedly a systematic relationship between articulatory place and the portion of the spectrum with greatest energy ("locus" frequency; see Stevens, 1998). For plosives, the relation between locus frequency and place of articulation is as follows – the further forward the constriction, the lower the locus. For bilabial stops, such as [p] and [b], the locus is usually between 600 and 1,000 Hz. The velar stop ([k], [g]) locus is ca. 2,500 - 3,000 Hz, while the alveolar ([t], [d]) locus is somewhere between 1,800-2,400 Hz. For fricatives, the relationship is just the opposite (but discussion of this pattern lies outside the scope of this chapter).

Second, vision contributes significantly to place-of-articulation cues. This is particularly important in noisy backgrounds and for the hearing impaired Grant et al., 1998; Massaro &

Cohen, 1999). Speechreading cues allow the listener to interpret what might otherwise be highly ambiguous acoustic data (Grant et al., 1998; Grant, 2002). The gain in intelligibility derived from such visual information is enormous – as much as a 14-dB enhancement of signal-to-noise ratio (Grant, 2002). Such an improvement can mean the difference between 10% of the words decoded and 90% correctly understood. This is an instance where information derived from an auxiliary sensory stream has the ability to make or break the process of speech communication. We return to this issue in Section 7.2.

Third, place-of-articulation is the most stable articulatory parameter in terms of pronunciation and across historically related languages. A statistical analysis of conversational American English shows that in terms of pronunciation, consonantal place of articulation (particularly at the syllable's onset) is significantly less likely to differ from the canonical pronunciation (that contained in a standard dictionary) relative to other articulatory parameters.

Of potential relevance is the observation that historically related words ("cognates") often share the same (or highly similar) consonantal place of articulation (English [ð] and German [d] in words such as "the" and "der"). Historical patterns of sound change are closely related to synchronic variation at a single point in time, for the seeds of phonetic evolution are contained in the pronunciation variation observed across speakers. It is not surprising that both synchronic and diachronic analyses imply that place of articulation is a highly stable (and lexically discriminative) feature.

## 3. INFORMATION IS DISTRIBUTED UNEQUALLY THROUGHOUT THE SYLLABLE

### *3.1 The Contrast Between Onset and Coda Consonants*

Pronunciation varies across the syllable. Onset consonants are far more likely to be articulated canonically than their coda counterparts, which tend to reduce or disappear entirely (Greenberg, 1999). For example, the most common pronunciation for the word "and" is [æ] [n], not [æ] [n] [d]. A statistical analysis of pronunciation patterns in American English indicates that nearly three-fourths of coda consonants are either [t] [d] or [n], all of which are associated with the alveolar (coronal) place of articulation (Greenberg, Carvey, Hitchcock & Chang 2002). Although anterior (e.g., [b], "tab") and posterior (e.g., [g], "tag") consonants may occur in coda position, they are relatively rare (and when they do, tend to be pronounced canonically). In contrast, the statistical distribution of onset consonants (with respect to articulatory place) is far more uniform (Greenberg et al., 2002).

In contexts where the identity of a constituent is highly predictable, that element carries far less information than in contexts where its identity can vary among many options (this is a fundamental principle of information theory – see Raisbeck, 1963). In this sense, the onset carries far more information than the coda. Reduction and deletion of coda consonants (particularly coronals) is consistent with their intrinsically lower entropy. Sections 6.3 and 7 discuss why coda consonants delete so often without significant impact on intelligibility.

### *3.2 Auditory Basis of Onset Primacy*

Onset consonants are far more likely to be articulated in canonical fashion that codas (or vowels – see Section 3.3). The primacy of onsets reflects how the auditory system (and brain) encodes information. Neurons have evolved to detect the initial portion of novel events far more than what follows (Nicholls, Wallace, Fuchs & Martin, 2001). This sensitivity to onsets is reflected in the much higher response rates observed in the auditory nerve at the beginning of a stimulus relative to what follows (Greenberg, 1996). From an ecological perspective, this sensitivity to onsets makes eminent sense. Rapid detection of a snapping twig could mean the difference between eating dinner that evening and being served as dinner to a predator.

The importance of onset coding is even more apparent at higher levels of the auditory pathway. In the auditory cortex, most neurons respond only close to stimulus onset, and are otherwise quiescent (see the chapter by Schreiner and colleagues in this volume). For a communication system emphasizing reliability, placing most of the information at the beginning of a packet (in this instance, a syllable) would focus the greatest amount of neural "attention" to the most informative constituents of the speech signal.

Most auditory cortical neurons are unresponsive to the final portion of an acoustic event. An "intelligent" communication system rarely places important information in that position. And when it does, such information is likely to be presented with a "bang" (e.g., a fully released plosive burst).

Consistent with this de-emphasis of consonantal codas is a statistical analysis of a German corpus (Jaeger, 2003), which reveals that it is not only English that places relatively little information in the terminal consonant of the syllable. A similar distribution of consonantal place of articulation is observed. This trend is taken a step further in Mandarin Chinese; there are only two consonants allowed in coda position – [n] and [ŋ]. It is therefore likely that de-emphasis of coda consonants is a linguistic universal (Greenberg, 1978). But the motivation for this pattern originates in the auditory system and the brain, not in the dictionary or the vocal tract.

## 4. THE SYLLABLE NUCLEUS

In English, the nucleus is usually a vowel (in rare exceptions, it can be a liquid or a nasal). It holds the constituents of the syllable together, and it is not an exaggeration to state that speech would hardly be possible without the nucleus. Moreover, the nucleus defines the energy arc's contour, serving to shape the signal's modulation spectrum for "consumption" by the auditory system and brain (see Section 8).

Within the traditional phonetic framework, vowels are often treated in the same way as consonants – just another set of phonemes. But, this egalitarian approach does not accurately portray the intricate patterns of pronunciation variation observed in everyday speech. Vowels function quite differently than consonants, and certain orthographies, such as Hebrew and Arabic, recognize this explicitly (the vowels are marked as diacritics, if at all).

The pronunciation of vowels is highly mutable, and depends on the talker's dialect, speaking style and emotional context. Moreover, the identity of a vowel often depends on the syllable's accent; most non-canonically pronounced vowels, accounting for 40% of all vocalic instances, occur in unaccented syllables (Greenberg et al., 2002). The pronunciation patterns of vowels are discussed more fully in Section 5.

## 5. THE IMPORTANCE OF PROSODY

Besides serving as the “glue” that binds constituents of the syllable together, the nucleus also functions as the primary medium through which prosody is realized. More than any other constituent, the nucleus provides information pertinent to the syllable’s accent, which is critical for decoding speech. Accent denotes how much information is carried in a syllable and also allows the listener to deduce consonantal identity more accurately than would otherwise be the case. Accent’s crucial role is the result of inherent ambiguity in the acoustic cues of speech. Their interpretation depends on context, and accent provides a lot of contextual information.

Just how important are prosodic cues for understanding spoken language? In an ingenious experiment (Bansal, 1969; Huggins, 1972), native speakers of Hindi were asked to read English words aloud. For example, the word “character” was spoken by Hindi speakers with the primary accent on the second syllable; in contrast, native speakers of English place the primary accent on the first syllable. When native speakers of American English wrote down the words spoken by the Hindi speakers, “character” was transcribed as “director,” and so on. Listeners appeared to interpret the sounds largely on the basis of their prosody, not on the actual phonetic composition of the word.

The phonetic properties of heavily accented syllables are different from those without accent (Figure 25.2). The coda consonants are far more likely to be canonically pronounced and the number of consonantal deletions is much smaller (Greenberg et al., 2002). Moreover, vowels in accented syllables are far more likely to come from the lower part of the articulatory space (e.g., [a], [ae], [aw], [ay], [ɔ]) (Figure 25.3), and are usually longer and more intense than their counterparts in unaccented syllables (Figure 25.4).

In unaccented syllables about three quarters of the vowels are either [I] [i] or [ax], all of which are high vowels where the tongue tip is arched towards the front or center of the vocal cavity. As with coda consonants, vocalic identity is largely predictable in unaccented syllables, and hence carries little intrinsic information. In heavily accented syllables there is a much broader range of vowels articulated, a situation analogous to that observed among consonants in the syllable onset. But the specific identity of the vowel, even in heavily accented syllables, is highly mutable. For example, in American English the vowels in such words as “orange” and “caught” reflect regional variation. In the Northeast U.S., the first vowel in “orange” is pronounced as an [ɑ] (a low, central vowel), while the vowel in “caught” is pronounced as an [ɔ] (a low, back vowel). In the Mid-western and Western regions, the vowels are reversed in pattern. There are many such examples in American English, as well as from languages around the world, where such vocalic interchanges occur. The point of interest is that the interchanged vowels (which are in “free variation”) are closely related. In the example given, they are both low vowels that are in articulatory proximity with respect to the front-back dimension (i.e., back and central articulations). It would be unusual for vocalic free variation to involve a fully front and fully back vowel, or a high and low vowel. The interchanges usually involve just one "notch" distance in the articulatory vowel space. This is no accident.

Correlated with vowel height and vocalic identity is segmental duration and amplitude. Low vowels are, on average, nearly twice as long as their high vocalic counterparts. Low vowels are also louder than high vowels. The three acoustic parameters mostly closely correlated with prosodic accent in (American) English are (1) vowel duration, (2) vocalic energy, and (3) vowel identity (Greenberg, Carvey & Hitchcock, 2002; Greenberg, 2005). In this sense, accent and vowel identity are closely related. The proportion of high vowels in fully accented syllables is very small, while most low vowels occur in accented syllables (Hitchcock & Greenberg, 2001; Figure 25.5).

In summary, the vocalic nucleus conveys information about prosodic accent that helps the listener interpret other phonetic information contained in the syllable. Moreover, other prosodic information, such as intonation, is conveyed mostly through the vocalic portion of the syllable.

## 6. THE RELATION BETWEEN PROSODY AND ARTICULATORY FEATURES

The traditional phonetic framework envisions little, if any, relationship between prosody and articulation. Articulatory features are considered attributes of phonetic segments; whether these features are produced "as advertised" is of little concern (except to students of pronunciation variation).

Phonetic variability (and its relation to prosody) is largely ignored by contemporary linguistics because there is no "official" place for pronunciation variation in the conventional theoretical framework. Such variation is essentially considered "noise."

Multi-tier theory turns the conventional phonetic framework on its head. From its perspective, phonetic variation is both extremely informative and meaningful; it conveys nuance and emotional shading that are the very essence of communication. In order to understand how such variability can be reliably interpreted, we examine the syllabic microstructure in some detail.

### *6.1 Voicing*

Voicing is considered the least phonetically informative of articulatory features. This is partly because nearly 80% of the speech signal is voiced, and this feature does not appear to be important for distinguish lexical identity. Within the traditional phonetic framework, a segment is either voiced or not – [b] vs. [p], [g] vs. [k], [z] vs. [s] – and this is the extent of its linguistic significance. However, even a cursory examination of the acoustic signal reveals the fallacy of this assumption. Many theoretically voiced segments are partially or entirely unvoiced. A common example in American English is [z] in coda position (e.g., "says"). Often the coda [z] is unvoiced throughout its entire length. However, this segment is not really an [s], for it neither sounds nor looks like [z]'s unvoiced counterpart.

Voiced plosives in the syllable onset are another example. It is unusual for voicing to occur throughout the segment's entire length. In English, the onset of voicing generally occurs between 20 and 40 ms after articulatory release. For a "voiced" velar constituent, such as [g], much of its length is actually unvoiced.

The segment [p] in a word such as “spin” provides yet another example. Phonemically, this constituent is a voiceless plosive. However, the segment sounds more like [b] than [p] when listening to it in isolation. The voicing associated with the nucleus has intruded into the onset. If voicing were purely a segmental feature, the word “spin” would be represented as [s] [b] [ih] [n]. But [s] [b] clusters are “illegal” in English.  What’s actually going on?

Such paradoxes and theoretical inconsistencies stem from the assumption that voicing is a segmental feature distinguishing voiced from unvoiced counterparts. The examples above can be readily explained if voicing is thought of not as a segmental feature, but rather as one reflecting prosodic accent at the syllabic level.

For example, most instances of unvoiced [z] in the Switchboard corpus occur in unaccented (or lightly accented) syllables. As described in Section 2.5, voicing spans a contiguous interval within a syllable, emerging from the core of the nucleus. Voicing may spread forward into the coda and back towards the onset. How much it spreads is largely determined by prosody. In heavily accented syllables voicing tends to spread deep into the syllabic flanks. For canonically voiced segments, such as [b] and [z], the entire syllable in the word “boys” may be voiced, including some pre-voicing prior to the [b]’s articulatory release. In unaccented syllables the [b]’s voice onset time is likely to be long, and the [z] entirely unvoiced. Although voicing may distinguish a voiced segment from an unvoiced counterpart, there are many instances where it does not perform such a contrastive function.

Voicing is probably the articulatory feature most sensitive to prosody. In German, it has played an important role in the language’s sound pattern. Historically, there was a distinction between voiced and unvoiced plosives in coda position – but no longer. In principle, all coda stop consonants are phonetically unvoiced, regardless of their phonemic status. This segmental shift probably began as a prosodic change, which carried the voicing with it. Consistent with this interpretation is the incomplete neutralization of the voiced/unvoiced distinction observed in certain contexts (Port & O’Dell, 1985; Port & Crawford, 1989) where prosody is likely to play a role.

In contemporary German, certain dialects (mostly in the north) voice the initial [s] of a word, so that the initial consonant is pronounced as a [z] rather than as an [s], while speakers in Bavaria and Austria tend to retain the original pronunciation ([s]). This dialectal variation likely reflects prosodic patterns that differ geographically.

### *6.2 Manner of Articulation*

Prosody affects manner of articulation, but in ways more subtle than voicing. Articulatory manner is closely associated with the energy arc (see Section 8). Because prosody also affects the energy arc, there is an intrinsic relation between manner and prosody. In English the two often interact in a variety of contexts. For example, it is common for a nasal segment in coda position to “delete” and leave its residue on the preceding vowel in the form of nasalization. This occurs rather frequently, particularly in unaccented or lightly accented syllables. Something akin to this process occurred in French several hundred years ago. Nasal segments in the coda are usually no longer pronounced (except in elision with a following vowel), but the preceding vowel is heavily nasalized. The exception to this French rule occurs when the initial constituent of the following syllable is vocalic (or a glide). Then, the nasal coda is pronounced as a full-fledged segment. But this exception lends further credence to the notion that manner and prosody are intimately related.

There are many instances where the primary function of a manner class is to separate syllables. Elision in French is one example of this function, in which certain latent phonetic properties only become manifest when dividing syllables of unequal accent. In English, liquids ([l], [r]), glides ([y] [w]) and flaps ([dx], [nx]) often serve in this capacity. In each instance, such syllable dividers are characterized by a dip in energy over a portion of the spectrum. In flaps the energy dip is very brief (20-40 ms) and spans most of the spectrum, while in liquids the modulation in amplitude is often restricted to a narrow portion of the spectrum (ca. 2.0-3.5 kHz). Such constituents are not really segments, but serve instead as "junctures" separating syllables (usually of unequal accent). The prosodic pattern determines the specific way in which such constituents are phonetically realized. Because most of these junctures are phonetically labeled as coda segments, which intrinsically carry little information, their phonetic realization rarely matters within the conventional phonetic framework – such variation is merely "noise."

From an historical perspective, manner of articulation is more stable than voicing, but less so than articulatory place. In cognate words, such as "those" and "das" or "valley" and "tal" the manner often differs (as does the voicing), while place of articulation (when normalized for manner) rarely does – a point addressed in the following section.

### *6.3 Place of Articulation*

Place of articulation is probably the key phonetic feature for distinguishing among words. It is the most stable in terms of pronunciation and historically over time. Place is also the articulatory feature least influenced by prosodic accent.

The prosodic pattern of an utterance reflects many factors, including dialect, speaking style and emotional mood. Stable properties of speech must be relatively insensitive to prosodic factors if they are to reliably convey information across a broad range of environmental and speaking conditions. Therefore, it is not surprising that articulatory place is relatively resistant to prosodic accent, both in onset and coda position. When prosody does affect the phonetic realization of a segment it is with respect to reduction or deletion, not in terms of a change in articulatory place; and this impact is usually confined to the coda (the apparent exception is deletion of [ð] in such words as "the," "them" and "those" in highly predictable contexts, where the intrinsic entropy associated with the onset segment is low).

There is far more to place of articulation's stability than its relative immunity to prosodic patterns (see Section 7).

## 7. PLACE OF ARTICULATION – THE KEY ARTICULATORY DIMENSION FOR LEXICAL IDENTITY

There is a paradox concerning place of articulation's stability. Its articulation rarely differs from the canonical (particularly in the syllable onset), and this dimension is important for distinguishing among words. And, as observed, place cues are relatively insensitive to variation in prosody. Yet, such information is also the most vulnerable to background noise (Miller and Nicely, 1955). How can this be?

There are several key properties of articulatory place that resolve this paradox.

### *7. 1 Manner-dependent place of articulation analysis*

The first concerns the entropy associated with place of articulation. In principle, there are between eight and ten distinct loci of articulatory constriction in English. The constriction can be achieved by both lips coming together ("bilabial"), or by the teeth abutting the tongue tip (labio-dental) or the tip of the tongue contacting the palatal ridge ("velar"), and so on. In practice, there are usually just two or three (or at most four) distinct loci of constriction for any given manner class. For English plosives, the constriction can be achieved bilabially, at the alveolar ridge or further back towards the velum. The pattern of constriction for nasals is comparable (except that the velar [ŋ], "si<u>ng</u>" occurs only in coda position). For fricatives, the locus of constriction varies from labio-dental ([f], [v]), to interdental ([θ], [ð]), alveolar ([s], [z]) and palatal ([ʃ], [ʒ]). The entropy associated with place of articulation is relatively low – once the associated manner of articulation has been determined. Except for fricatives, the classification of articulatory place reduces to a ternary set – anterior, central and posterior. Thus, for the listener, the task's complexity is typically reduced to a choice of three alternatives.

The interdependence of place and manner of articulation has some interesting consequences, as it predicts that historical sound changes affecting manner will necessarily impact place of articulation if the loci of constriction are not concordant (e.g., plosives and fricatives). This is precisely what has happened in the historical divergence of German and English. The German alveolar ([d], [t]) often becomes inter-dental in English ([θ], [ð]), The relational context of their manner class remains fixed (central), even though the <u>specific</u> locus of constriction has changed slightly.

### *7.2 Importance of visual cues for place of articulation*

The traditional phonetic framework assumes that formant trajectories are the primary acoustic cue for place of articulation (Kewley-Port, 1983; Kewley-Port & Neel in this volume). Some believe that the locus of energy in the stop burst is more important than the trajectory (Blumstein & Stevens, 1979). However, both cues are vulnerable to background noise.

A third cue, the visual motion of the lips, teeth, tongue and jaw ("speechreading") provides crucial information under adverse acoustic conditions. When the acoustic and visual cues are in conflict, the visual cues often sway the listener to revise his/her interpretation of the speech signal. This phenomenon is known as the "McGurk" effect (after the psychologist who first described the illusion – McGurk & McDonald, 1976), and could not occur if the acoustic cues were truly robust. In the classic experiment, a listener is acoustically presented a nonsense sequence, such as [aba], and watches the same talker articulate [aga]. Under such circumstances, the listener reports "hearing" [ada], an alveolar fusion between the bilabial [b] and the velar [g]. Although the specific mechanism responsible for this fusion is unknown, some experimental evidence suggests that the sensory processes underlying the McGurk effect occurs at a relatively early ("pre-categorical") stage of analysis and involves temporal processing of some kind (see Massaro, 1987; Braida, 1991 and Grant et al., 1998 for discussion of this issue).

Normally, the visual and acoustic cues are in register; this audio-visual coordination probably accounts for the special nature of articulatory place information. Almost all of the phonetic cues conveyed through speechreading concern place-of-articulation (Grant et al., 1998). There is virtually no information pertaining to manner of articulation and voicing transmitted through visual motion.

In most communication settings – even in the age of the telephone – speaker and listener converse face-to-face; the visual cues reinforce the acoustic stream (and vice versa). The robustness of place-of-articulation cues is a consequence of the bi-modal nature of the information contained in the speech signal.

When the acoustic and visual cues are in conflict (as in the McGurk effect), the brain needs to mediate between the two. However, the situation is not quite as simple as it seems. The visual stream is unable to over-ride the acoustic cues under all circumstances. When the acoustic stimulus is [ada], it is difficult for any visual signal to sway the listener's percept away from [d] (K. Grant, personal communication). Multi-tier theory predicts this, as it interprets the McGurk effect as the consequence of inherently ambiguous acoustic cues. If the acoustic cues were not ambiguous (as is the case with [d], due to the relatively flat formant transitions), then there would be little opportunity for visual interaction to alter the phonetic percept.

### *7. 3 Place of Articulation is a Demi-syllabic Feature*

Place of articulation is often treated as a segmental (usually consonantal) feature. This is an oversimplification. Articulatory place transcends the segment, particularly in the context of consonantal clusters. It is rare for two distinct places of articulation to occur in contiguous consonants within a syllable; even across syllables there is a tendency for homo-organic assimilation of place, particularly when the assimilated constituent carries little information by itself.

For syllable onsets, in particular, the constituent is treated as a single unit (with respect to articulatory place) rather than as an ensemble of independent segments. In this sense, place is not really a segmental feature but rather operates at the level of the demi-syllable (i.e., the onset or coda). In instances where exceptions occur (such as in the words "split" and "skit"), the distribution of place features within that context effectively neutralizes the place feature associated with [s] (i.e., [s] does not carry contrastive place information within a consonantal cluster, it merely serves as an acoustic conditioner associated with the onset component of the energy arc).

### *7. 3 The Relation Between Place of Articulation and Linguistic Information*

Place of articulation cues often convey important information. This is why there is usually only a single articulatory place associated with the onset and coda of a syllable. In effect, the brain learns to associate each consonantal component of the syllable with a specific place of articulation. In English it is exceedingly rare for two distinct places of articulation to occur in the onset. In contrast, the coda may contain two distinct loci of constriction, but when it does, the grammatical consequences are significant. In words, such as “kept” or “slept,” the second constriction is associated with the bound, past-tense morpheme [t], which modifies the meaning of the word (along with a change in vowel from [iy] in “keep” and “sleep” to [ɛ] in “kept” and “slept”).

Such examples suggest that place-of-articulation cues function primarily to convey linguistic information. Three-quarters of coda consonants are alveolars (in contrast to a relatively equitable distribution of place in the onset), consistent with the relatively predictable nature of that syllabic constituent (Greenberg et al., 2002). When the coda constriction differs from alveolar, it is usually associated with a relatively uncommon word. And when two distinct constrictions occur, this is a sure sign that both places of articulation convey important information (hence “kept”). And conversely, when the coda consonant is not highly informative, this is often signaled by a subtle shift in place of articulation (e.g., going [g] [ow] [ih] [ŋ] > goin’ [g] [ow] [ih] [n]).

## 8. THE ENERGY ARC AND ITS ASSOCIATION WITH MANNER OF ARTICULATION

Within a syllable it is rare for contiguous constituents to share the same manner of articulation (Chang, 2002; Greenberg et al., 2002), even though there are words in English whose canonical pronunciations contain such sequences (e.g., “laughs” – [l] [ae] [f] [s]). Why should this be so?

As mentioned in Section 2.4, each manner class is associated with a distinct energy level – vowels are the most intense, fricatives and plosives are the softest, with the energy level of other manner types lying in between. In English and other languages, the order in which phonetic segments may occur within a syllable obeys a principle closely associated with the energy arc. Sequences of the form [s] [t] [r] [ɔ] [ŋ] “strong” are valid, while others, such as [r] [z] [b] are not. A concept – the “sonority hierarchy” – was proposed over a century ago to account for such phonotactic patterns (Jespersen, 1897-1899, 1904). However, the sonority hierarchy is mainly a descriptive tool, lacking a firm theoretical foundation and fails to account for a variety of phonotactic patterns (Archangeli & Pulleyblank, 1994; Ohala, 1992; Ohala & Kawasaki-Fukumori, 1997; Zec, 1996).

Within the multi-tier framework, manner of articulation is the primary means through which the production system shapes the energy contour of the speech signal into a form that is easily “digested” by the auditory system. In abstract terms, the shape of this contour is an arc that rises to a peak in the nucleus and then descends. The specific way in which the arc ascends to the peak depends on the phonetic composition of the syllable’s onset. The rise can be gradual, in which case the onset is likely to contain several constituents (such as the [s] [t] [r] sequence in “strong”) or it can be more abrupt, as occurs when a syllable contains just a vowel or begins with a stop consonant. The same principle holds for the coda; however, the phonetic composition of this constituent often varies from that of the onset, for reasons described in Section 3. The contour associated with the onset and coda ultimately reflects the phonetic composition of the syllable as well as the linguistic information contained within. In this sense, the energy arc is highly sensitive to prosody, which affects both the height and length of the contour. Heavily accented syllables are more intense and longer than unaccented syllables (Figure 25.6). Through careful analysis of the energy arc a listener is able to deduce many important properties of the syllable, including (1) its linguistic importance, which is related to prosodic accent, (2) its voicing profile, (3) the likely number of phonetic constituents, and (4) the manner classes associated with each of these constituents. This coarse analysis sets the stage for more fine-grained phonetic analysis based on place of articulation. In many contexts articulatory place cues are not required to accurately decode the words spoken. But in situations where such information is required, the visual cues can be extremely useful, not only in decoding place-of-articulation information, but also for synchronizing manner cues across the syllable.

The length of the energy arc in English averages 200 ms. It is rarely shorter than 50 ms or longer than 400 ms. About two thirds of the arcs range in duration between 100 and 300 ms. These statistics coincide with those of the syllable (Greenberg, 1999), which is the energy arc's linguistic manifestation.

The modulation spectrum is closely related to both the energy arc and the syllable (Figure 25.7). It provides a convenient means with which to quantify the frequency, magnitude and phase of the energy arc and has been shown to correlate highly with intelligibility in a broad range of listening conditions (Houtgast and Steeneken, 1985; Drullman, Festen and Plomp, 1994a, 1994b; Drullman in this volume; Greenberg & Arai, 2004).

The modulation spectrum reflects the prosodic properties of speech. The peak of the modulation spectrum is approximately 5 Hz, coinciding with the average duration of the syllable. But the modulation spectrum contains appreciable energy between 2 and 12 Hz. Energy in the lower region of the spectrum (2-4 Hz) reflects heavily accented syllables, while higher modulation frequencies (6-20 Hz) reflect mostly unaccented syllables. The spectrum's peak represents the overlap between accented and unaccented syllables (Greenberg, Carvey, Hitchcock and Chang, 2003). Low-pass filtering the modulation spectrum below 4-6 Hz significantly decreases intelligibility, as does high-pass filtering above 8 Hz (Drullman et al., 1994a, 1994b; Drullman in this volume). Such manipulation of the speech signal is likely to degrade intelligibility through its interaction with the prosodic and syllable pattern of spoken material.

## 9. A MULTI-TIER LEXICAL REPRESENTATION

Within the traditional phonetic framework spoken words are represented as sequences of phonemes, analogous to the manner in which they are represented in written form (Section 2.1). Phonemic sequencing is the standard method for representing words in many language-related fields, including lexicography, descriptive and applied linguistics, foreign language instruction, verbal interaction studies, and most importantly for the present discussion, automatic speech recognition and synthesis (see Section 10).

It is difficult to describe spoken language entirely (or even largely) in terms of phonetic segments. There is far too much variability in utterances than can be accommodated with the standard phonetic approach (Sections 2.6, 3, 5 and 6). Moreover, there is psychological evidence in support of representational units other than the phoneme.

When attempting to recall a word that is on "the tip of one's tongue" (Brown & McNeill, 1966) a speaker is usually able to recall certain features with precision. The most reliable properties associated with such "missing" words are: (1) the number of syllables, (2) the prosodic accent pattern, and (3) the initial consonant. Vowels, as well as consonants in unaccented syllables, do not figure importantly in such tip-of-the-tongue phenomena.

Slips of the tongue and the ear are also consistent with the syllable's primacy in lexical representation. Transpositions of sounds in so-called "spoonerisms" usually involve syllabic onsets (e.g., "queer old dean" in place of "dear old queen") (Fromkin, 1973).

Phoneme monitoring tasks also appear to support the syllable as a basic unit of lexical representation. Listeners respond more quickly to a specific syllable than to a designated phone within the syllable, even when the phone occurs at the onset (Segui, Dupoux & Mehler, 1990).

The multi-tier framework suggests that representing words is neither simple nor straightforward. Just as syllables, phonetic constituents and articulatory features vary with respect to the amount of entropy conveyed, so do words. The mental lexicon is likely to be sensitive to entropy, and to adapt its representation of words depending on this parameter. In this sense, there is unlikely to ever be a single linguistic representation for any given word. Words are highly mutable, combining with other words to form a very broad range of meanings. Some of these are well represented by phoneme sequences, but most are not.

Because prosody so dramatically influences the pronunciation of words, any "robust" lexical representation must take this tier into account. The psychological evidence described above supports this conclusion, and is also consistent Todd's chapter in this volume. But precisely how does prosody factor into the mental lexicon? Heavily accented syllables appear to be more reliable "sign posts" than their unaccented counterparts. In the Switchboard corpus (Greenberg, 1997) the overwhelming majority ( > 80%) of the words spoken contain only a single syllable. Of the remainder, three-quarters contain two syllables. Most of the other polysyllabic words contain three syllables (Greenberg, 1999). Among words containing two or more syllables, those that are infrequently encountered in the corpus are far more likely to have two accented syllables than those that are relatively common. In other words, more sign posts are required when the word is rare or unfamiliar than when it is common currency.

Multi-tier theory also suggests that certain articulatory dimensions are more important than others. Place of articulation, particularly at the onset of a heavily accented syllable, is likely to provide a lot of discriminative information, particularly when combined with articulatory manner cues for the same syllable. Voicing is likely to carry the least amount of information in the mental lexicon.

In a highly literate society, such as ours, orthographic influences may be considerable. However, orthography's impact is likely to be greatest when it converges with the prosodic and syllabic phenomena described in this chapter.

Perhaps the most visible impact of lexical representation is in reading and foreign language instruction. In languages, such as English, where the correspondence between pronunciation and orthography is oblique, the time and energy expended to learn the language will necessarily be greater than those where the phoneme-to-grapheme correspondence is more transparent. Children experience far greater difficulty learning to read English or Danish compared to languages such as Spanish and Italian. Moreover, the proportion of dyslexic children is far higher in English-speaking countries than in Italy (Paulesu et al., 2001). Given the growing importance of English throughout the world, the relation between orthography and pronunciation is likely to be of paramount concern in language instruction for years to come.

## 10. APPLICATION OF MULTI-TIER THEORY TO SPEECH AND HEARING TECHNOLOGY

Automatic speech recognition (ASR) and other areas of speech technology have improved dramatically over the past decade (e.g., Waibel & Lee, 1990; Huang, Acero & Hon, 2001; Morgan, Bourlard & Hermansky, 2004). Automatic dictation programs are now available for many of the world's languages, and ASR is routinely used in commercial interactions over the telephone. There have been concomitant advances in speech synthesis over the same period of time (Keller, Bailly, Monaghan, Terken & Huckvale, 2001; Narayanan & Alwan, 2004). Despite this substantial progress, there is a growing belief that both technologies have reached their practical limits and that other approaches will be required to substantially enhance the quality of automatic recognition and synthesis of speech.

Speech recognition technology currently uses phoneme-based models for lexical units. State-of-the-art systems try to transcend the limitations imposed by the phonemic approach by using context-dependent phone models (usually consisting of either three or five contiguous phones), but with only partial success. Part of the problem is that syllables contain a variable number of phones, thereby guaranteeing that no strictly multi-phone model will ever be perfectly synchronized with all syllables in an utterance. This lack of synchrony introduces an unknown amount of "noise" in the acoustic models used. Moreover, multi-phone models require enormous amounts of training data in order to garner a sufficient number of exemplars for each phonetic context modeled. This is one of the reasons why it is so expensive and time consuming to develop speech recognition applications.

Another problem with current-generation ASR systems concerns the way they approach pronunciation variation. Most systems treat such variation as a source of noise, something to be minimized or eliminated if at all possible. The way in which such variability is handled in the recognition lexicon is revealing. Typically, the two or three most common pronunciations are incorporated into the dictionary, usually in terms of phoneme sequences. Unfortunately, much of the variability encountered cannot be representation in terms of phonemes (or phones) but require a fine-grained approach. Because the acoustic models are based on phone sequences, there is little that can be done within the conventional systems. The most sophisticated pronunciation models use highly sophisticated statistical methods to deduce the most appropriate phone sequences that capture the essence of the variability. Without realizing it, these methods often incorporate knowledge about syllable structure, prosodic accent and articulatory features. Much of their success can probably be attributed to such implicit knowledge derived from non-phonemic tiers.

The situation in speech synthesis differs from that of recognition technology. Long ago, it was recognized by practitioners in the field that the phoneme is not an ideal unit for synthesizing natural-sounding speech. As early as the 1970s Fujimura proposed the syllable as a more practical unit for synthesis (Fujimura, 1979; Fujimura & Lovins, 1978). Most of the current synthesis systems use either demi-syllables (Fujimura & Lovins, 1978) or a sophisticated unit-selection method that employs constituents of variable length (Black & Campbell, 1995; Campbell, 1994). Both approaches produce high-quality speech. However, the ability to synthesize a variety of speaking styles and emotional content is quite limited. The unit-selection method requires recording several hours of spoken material and cannot be extended to speaking styles other than those recorded. Moreover, the quality of the synthesis depends on the specific properties of the speaker's voice. Although demi-syllable synthesis is not limited to a specific speaker, it is quite time consuming to generate a wide range of speech. Usually, the synthesis is based on rules that have been laboriously developed over many years (e.g., Koutny, Olaszy & Olaszi, 2000). Such rule-based systems are not easily extensible to different speaking styles or emotional contexts.

For different reasons the technologies used in speech recognition and synthesis are unsuited for commercial exploitation in the long term. Ideally, both technologies should be capable of being deployed quickly and inexpensively, and capable of handling all sorts of speech material ranging from the formal to the casual. Currently, neither technology is up to the task. This situation is unlikely to change until the models underlying the technologies form a more accurate description of spoken language. For ASR technology this will require the use of visual cues that are so important for decoding the speech signal in noisy environments and melding them with acoustic information to form multi-tier linguistic representations. Future-generation synthesis systems will have to employ sophisticated speaker-independent algorithms that incorporate a lot of prosodic knowledge that can be combined with detailed articulatory models.

Perhaps the greatest potential impact of speech technology will be in future-generation hearing aids and other auditory prostheses. Currently, hearing aids rely on sophisticated signal processing methods to enhance the intelligibility of speech for the hearing impaired (Edwards, 2004). The efficacy of such prosthetic devices is limited in many contexts, particularly in the presence of background noise or reverberation. Unfortunately, it is unclear how the current signal processing methods actually enhance intelligibility or what could be done to improve the prostheses' efficacy. Nor is it clear whether signal processing, by itself, can compensate for a sensori-neural hearing loss. Our knowledge of how speech is processed in the auditory system is still at a rudimentary stage. What should be done to restore intelligibility to that of a normal-hearing individual is unclear. Research from Robert Shannon's group (see his chapter in this volume) suggests that only a coarse spectral profile of the speech signal may be sufficient to restore intelligibility. Faulkner and Rosen's work, on the other hand, suggests that adding some form of visual information could also be extremely useful (see their chapter in this volume).

Whatever happens in the field of auditory prostheses, it is likely that both speech recognition and synthesis technology will ultimately be incorporated into prosthetic devices. This will be necessary to overcome the problems associated with noisy, reverberant environments that make pure signal processing approaches less than ideal. Some form of reconstitution of the original speech signal will be required to impart the sort of clarity required by those with a significant hearing impairment.

## 11. CONCLUSION

Language is what distinguishes Homo sapiens from all other species in the animal kingdom, and is likely to have played an important role in the rapid evolution of human culture and society. This pace of change is likely to accelerate as technology becomes an increasingly important component of daily life. Many of this coming century's technological breakthroughs are likely to involve language as part of a broad effort to make human-machine communication more transparent. Such technological progress will require concomitant advances in scientific knowledge pertaining to the biological bases of speech communication. This broad-based scientific effort is likely to concentrate on the interaction between the auditory and visual modalities, as well as on the relation between the perception and production of speech. Underlying the function of this complex neural machinery is the transmission and processing of information. Linking the information-processing component of speech communication with its biological foundations is likely to form the focus of spoken language research over the coming decades.

## REFERENCES

Allen, J. B. (1994). How do humans process and recognize speech? IEEE Transactions on Speech and Audio Processing, 2, 567-577.

Archangeli, D., & Pulleyblank, D. (1994). Grounded phonology. Cambridge, MA: MIT Press.

Bansal, R. K. (1969). The intelligibility of Indian English: Measurements of the intelligibility of connected speech and sentence and word material, presented to listeners of different nationalities. Doctoral thesis, University of London.

Black, A. W., & Campbell, N., (1995). Optimising selection of units from speech databases for concatenative synthesis. Proceedings of the 4th European Conference on Speech Communication and Technology (Eurospeech-95).

Blumstein, S. E., & Stevens, K. N. (1979). Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants, Journal of the Acoustical Society of America, 66, 1001-1017.

Boersma, P. (1998). Functional phonology. Formalizing the interactions between articulatory and perceptual drives. The Hague: Holland Academic Graphics.

Braida, L. D. (1991). Crossmodal integration in the identification of consonant segments. Quarterly Journal of Experimental Psychology, 43A, 647-677.

Brown, R. & McNeill, D. (1966). The “tip-of-the-tongue” phenomenon. Journal of Verbal Learning and Verbal Behavior, 5, 325-337.

Campbell, N., (1994). Prosody and the selection of units for concatenation synthesis. Proceedings of the Second ESCA/IEEE Workshop on Speech Synthesis, pp. 61- 64.

Chang, S. (2002). A syllable, articulatory-feature, and stress-accent model of speech. Ph.D. Thesis, University of California, Berkeley [available as ICSI Technical Report 02-007].

Chang, S., Wester, M., & Greenberg, S. (2003). An elitist approach to automatic articulatory-acoustic feature classification for phonetic characterization of spoken language. Submitted.

Diehl, R., & Lindblom, B. (2004). Explaining the structure of feature and phoneme inventories: The role of auditory distinctiveness. In S. Greenberg, W. A. Ainsworth, A. N. Popper & R. R. Fay (Eds.), Speech processing in the auditory system (pp. 101-162). New York: Springer-Verlag.

Drullman, R., Festen, J. M., & Plomp, R. (1994a). Effect of temporal envelope smearing on speech reception. Journal of the Acoustical Society of America, 95, 1053-1064.

Drullman, R., Festen, J. M., & Plomp, R. (1994b). Effect of reducing slow temporal modulations on speech reception. Journal of the Acoustical Society of America, 95, 2670-2680.

**Edwards, B. (2004). Hearing aids and hearing impairment. In S. Greenberg, W. A. Ainsworth, A. N. Popper & R. R. Fay (Eds.), Speech processing in the auditory system (pp. 339-421). New York: Springer-Verlag.**

**Fletcher, H. (1953). Speech and hearing in communication. New York: Van Nostrand.**

**Fletcher, H., & Gault, R. H. (1950). The perception of speech and its relation to telephony. Journal of the Acoustical Society of America, 22, 89-150.**

**French, N. R., & Steinberg, J. C. (1947). Factors governing the intelligibility of speech sounds. Journal of the Acoustical Society of America, 19, 90-119.**

**Fromkin, V. (Ed.) (1973). Speech errors as linguistic evidence. Los Angeles: University of California Press.**

**Fujimura, O. (1979). An analysis of English syllables as cores and affixes. Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforchung, 32, 471-476.**

**Fujimura, O., & Lovins, J. (1978). Syllables as concatenative phonetic units. In A. Bell & J. B. Hooper (Eds.), Syllables and segments (pp. 107-120), Amsterdam: North Holland.**

**Grant, K. W. (2002). Measures of auditory-visual integration for speech understanding: A theoretical perspective. Journal of the Acoustical Society of America, 112, 30-33.**

**Grant, K. W., Walden, B. E., & Seitz, P. F. (1998). Auditory-visual speech recognition by hearing-impaired subjects: Consonant recognition, sentence recognition, and auditory-visual integration. Journal of the Acoustical Society of America, 103, 2677-2690.**

**Greenberg, J. (1978). Some generalizations concerning initial and final consonant clusters. In J. Greenberg (Ed.), Universals of human language (Volume 2: Phonology, pp. 243-280). Stanford, CA: Stanford University Press.**

**Greenberg, S. (1996) Auditory processing of speech. In N. Lass (Ed.). Principles of experimental phonetics (pp. 362-407). St. Louis: Mosby.**

**Greenberg, S. (1997). The Switchboard transcription project. Research Report #24, 1996 Large Vocabulary Continuous Speech Recognition Summer Research Workshop Technical Report Series. Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD.**

**Greenberg, S. (1999). Speaking in shorthand: A syllable-centric perspective for understanding pronunciation variation. Speech Communication, 29, 159-176.**

**Greenberg, S. (2003). Pronunciation variation is key to understanding spoken language. Proceedings of the International Congress of Phonetic Sciences, pp. 219-222.**

**Greenberg, S. (2005). From here to utility – Melding phonetic insight with speech technology. In W. Barry & W. Domelen (Eds.), Integrating phonetic knowledge with speech technology. Dordrecht: Kluwer.**

**Greenberg, S., & Ainsworth, W. A. (2004). Speech processing in the auditory system: An overview. In S. Greenberg, W. A. Ainsworth, A. N. Popper & R. R. Fay (Eds.), Speech processing in the auditory system (pp. 1-62). New York: Springer-Verlag.**

**Greenberg, S., & Arai, T. (2004). What are the essential cues for understanding spoken language? IEICE Transactions on Information and Systems, 87, 1059-1070.**

**Greenberg, S., Carvey, H. M., & Hitchcock, L. (2002). The relation of stress accent to pronunciation variation in spontaneous American English discourse. Proceedings of the ISCA Workshop on Prosody and Speech Processing, pp. 53-56.**

**Greenberg, S., Carvey, H. M., Hitchcock, L., & Chang, S. (2002). Beyond the phoneme – A juncture-accent model for spoken language. Proceedings of the Second International Conference on Human Language Technology Research, pp. 36-43.**

**Greenberg, S., Carvey, H. M., Hitchcock, L., & Chang, S. (2003). Temporal properties of spontaneous speech – A syllable-centric perspective. Journal of Phonetics, 31, 465-485.**

**Hitchcock, L., & Greenberg, S. (2001). Vowel height is intimately associated with stress accent in spontaneous American English discourse. Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech-2001), pp. 79-82.**

**Houtgast, T., & Steeneken, H. (1985). A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria. Journal of the Acoustical Society of America, 77, 1069-1077.**

**Huang, X. D., Acero, A., Hon, H.-W. (2001). Spoken language processing: A guide to theory, algorithm and system development. Englewood Cliffs, NJ: Prentice Hall.**

**Huggins, A. W. F. (1972). On the perception of temporal phenomena in speech. Journal of the Acoustical Society of America, 51, 1279-1290.**

**Jaeger, M. (2003) Perception of lost contrast. Proceedings of the International Congress of Phonetic Sciences, pp. 1735-1738.**

**Jakobson, R., Fant, G., & Halle, M. (1963). Preliminaries to speech analysis. The distinctive features and their correlates. Cambridge, MA: MIT Press.**

**Jespersen, O. (1897-1899). Fonetik. En systematisk fremstilling af laeren om sproglyd. Copenhagen: DetSchubotheske Forlag.**

**Jespersen, O. (1904). Phonetische Grundfragen. Leipzig and Berlin: Teubner.**

**Keller, E., Bailly, G., Monaghan, A., Terken, J., & Huckvale, M. (2001). Improvements in speech synthesis. New York: Wiley.**

**Kewley-Port, D. (1983). Time-varying features as correlates of place of articulation in stop consonants. Journal of the Acoustical Society of America, 73, 322-335.**

**Koutny, I., Olaszy, G., & Olaskzi, P. (2000). Prosody prediction from text in Hungarian and its realization in TTS conversion. International Journal of Speech Technology, 3, 187-200.**

**Ladefoged, P. (2000). A course in phonetics (4th ed.). Boston: Heinle.**

**Lieberman, P. (1984). The biology and evolution of language. Cambridge, MA: Harvard University Press.**

**Lieberman, P. (1990). Uniquely human: The evolution of speech, thought and selfless behavior. Cambridge, MA: Harvard University Press.**

**Massaro, D. W. (1987). Speech perception by ear and eye: A paradigm for psychological inquiry. Hillsdale, NJ: Lawrence Erlbaum Associates.**

**Massaro, D. W., & Cohen, M. M. (1999). Speech perception in hearing-impaired perceivers: Synergy of multiple modalities. Journal of Speech, Language, and Hearing Research, 42, 21-41.**

**McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices: A new illusion. Nature, 264, 746-748.**

**Morgan, N., Bourlard, H., & Hermansky, H. (2004). Automatic speech recognition: An auditory perspective. In S. Greenberg, W. A. Ainsworth, A. N. Popper & R. R. Fay (Eds.), Speech processing in the auditory system (pp. 309-338). New York: Springer-Verlag.**

**Narayanan, S., & Alwan, A. (2004) Text to speech synthesis : New paradigms and advances. Harlow, UK: Pearson Educational.**

**Nicholls, J. G., Wallace, B. G., Fuchs, P. A., & Martin, A. R. (2001). From neuron to brain (4th ed.). Sunderland, MA: Sinauer.**

**Ohala, J. J. (1992). Alternatives to the sonority hierarchy for explaining segmental sequential constraints. Papers from the Parasession on the Syllable. Chicago: Chicago Linguistics Society, pp. 319-338.**

**Ohala, J. J., & Kawasaki-Fukumori, H. (1997). Alternatives to the sonority hierarchy for explaining the shape of morphemes. In S. Eliasson & E. H. Jahr (Eds.), Studies for Einar Haugen (pp. 343-365). Berlin: Mouton de Gruyter.**

**Paulesu, E., Démonet, J.-F., Fazio, F., McCrory, E., Chanoine, V., Brunswick, N., Cappa, S. F., Cossu, G., Habib, M., Frith, C. D., & Frith, U. (2001). Dyslexia: Cultural diversity and biological unity. Science, 291, 2165-2167.**

**Port, R. F., & Crawford, P. (1989). Incomplete neutralization and pragmatics in German. Journal of Phonetics, 17, 257-282.**

**Port, R. F., & O'Dell, M. L. (1985). Neutralization of syllable-final voicing in German. Journal of Phonetics, 13, 455-471.**

**Raisbeck, G. (1963), The mathematical theory of information. Cambridge, MA: MIT Press.**

**Sampson, G. (1985). Writing systems. Stanford, CA: Stanford University Press.**

**Sapir, E. (1921). Language. New York: Harcourt.**

**Segui, J., Dupoux, E., & Mehler, J. (1990) The role of the syllable in speech segmentation,phoneme identification, and lexical access. In G. T. M. Altmann (Ed.), Cognitive models of speech processing: Psycholinguistic and computational perspectives (pp. 263-280). Cambridge, MA: MIT Press.**

**Stevens, K. N. (1998). Acoustic phonetics. Cambridge, MA: MIT Press.**

**Waibel, A., & Lee, K.-F. (1990). Readings in speech recognition. San Mateo, CA: Morgan Kaufmann**

**Zec, D. (1995). Sonority constraints on syllable structure. Phonology, 12, 85-129.**

## FIGURE CAPTIONS

25.1 A temporal perspective of speech processing in the auditory system. The time scale associated with each component of auditory and linguistic analysis is shown, along with the presumed anatomical locus of processing. The auditory periphery and brainstem is thought to engage solely in pre-linguistic analysis relevant for spectral analysis, noise robustness and source segregation. The neural firing rates at this level of the auditory pathway are relatively high (100-800 spikes/s). Phonetic and prosodic analyses are probably the product of auditory cortical processing given the relatively long time intervals required for evaluation and interpretation at this linguistic level. Lexical processing probably occurs beyond the level of the auditory cortex, involves both memory and learning. The higher-level analyses germane to syntax and semantics (i.e., meaning) is probably a product of many different regions of the brain and requires hundreds to thousands of milliseconds to complete. From Greenberg and Ainsworth (2004).

25.2 The impact of prosodic accent on pronunciation variation in the Switchboard corpus, partitioned by syllable position and the type of pronunciation deviation from the canonical form. The height of the bars indicates the percent of segments associated with onset, nucleus and coda components that deviate from the canonical phonetic realization. The magnitude of the deviation is also shown in terms of percentage figures for each bar. Note that the magnitude scale differs for each panel. The sum of the “Deletions,” (upper right panel) “Substitutions” (lower left) and “Insertions” (lower right) equals the total “Deviation from Canonical” shown in the upper left panel. Canonical onsets = 10,241, nuclei = 12,185, codas = 7,965. Adapted from Greenberg et al. (2002).

25.3 Spatial representation of the mean proportion of nuclei associated with syllables that are heavily accented or completely unaccented as a function of vocalic identity. Vowels are segregated into diphthongs and monophthongs for illustrative clarity. Note that the polarization of the y-axis scale for the unstressed syllables is the reverse of that associated with the heavily accented syllables (in order to highlight the spatial organization of the data). The x-axis refers to the front-back dimension in the horizontal plane and is intended purely for illustrative purposes. Data were computed from the SWITCHBOARD corpus. Adapted from Greenberg et al. (2002).

25.4 Spatial representation of the mean duration and amplitude (as well as their product, integrated energy) of vocalic nuclei in the annotated SWITCHBOARD corpus organized by prosodic accent magnitude and dynamic status of the vowel. The x-axis refers to the front-back dimension in the horizontal plane and is intended purely for illustrative purposes. Note that the durational scale on the y-axis differs across the plots. The vocalic labels are derived from the Arpabet orthography (see Greenberg, 1997 for a listing of the symbols used). Adapted from Hitchcock and Greenberg (2001).

25.5 The impact of prosodic accent ("Heavy" and "None") on the number of instances of each vocalic segment type in the corpus. The vowels are partitioned into their articulatory configuration in terms of horizontal tongue position ("Front," "Central" and "Back") as well as tongue height ("High," "Mid" and "Low"). Note the concentration of vocalic instances among the "Front" and "Central" vowels associated with "Heavy" accent and the association of high-front and high-central vowels with unaccented syllables. The data shown pertain solely to canonical forms realized as such in the corpus. The skew in the distributions would be even greater if non-canonical forms were included. Adapted from Greenberg et al. (2002).

25.6 An illustration of the energy arc principle through an example of a spectro-temporal profile (STeP) for a single, di-syllabic word, "seven" taken from the OGI Numbers95 corpus. The STeP is derived from the energy contour across time and frequency associated with many hundreds of instances of "seven" spoken by as many different speakers. The spectrum was partitioned into fifteen one-quarter-octave bands distributed between 300 and 3400 Hz (i.e., telephone bandwidth). The duration of each time frame is 10 ms. The amplitude was computed over a 25-ms window in terms of logarithmic (base e) units relative to the utterance mean. Each instance of a word was aligned with the other words at its arithmetic center. The mean duration of all instances of "seven" is shown by the red rectangle. The STeP has been labeled with respect to its segmental and syllabic components in order to indicate the relationship between onset, nucleus, coda and realizations within the syllable and their durational properties. The accented ("stressed") and unaccented ("unstressed") syllables are also indicated. Adapted from Greenberg et al. (2003).

25.7 The relation between (b) the modulation spectrum in the frequency band betw1 and 2 kHz, and (a) the distribution of syllable durations for fifteen minutes of spontaneous material (plotted in terms of equivalent modulation frequency for the sake of comparison). The syllable duration data were computed from 30 minutes of material from SWITCHBOARD, while the modulation spectrum was computed from 2 minutes of material from the same corpus. From Greenberg et al. (2003).

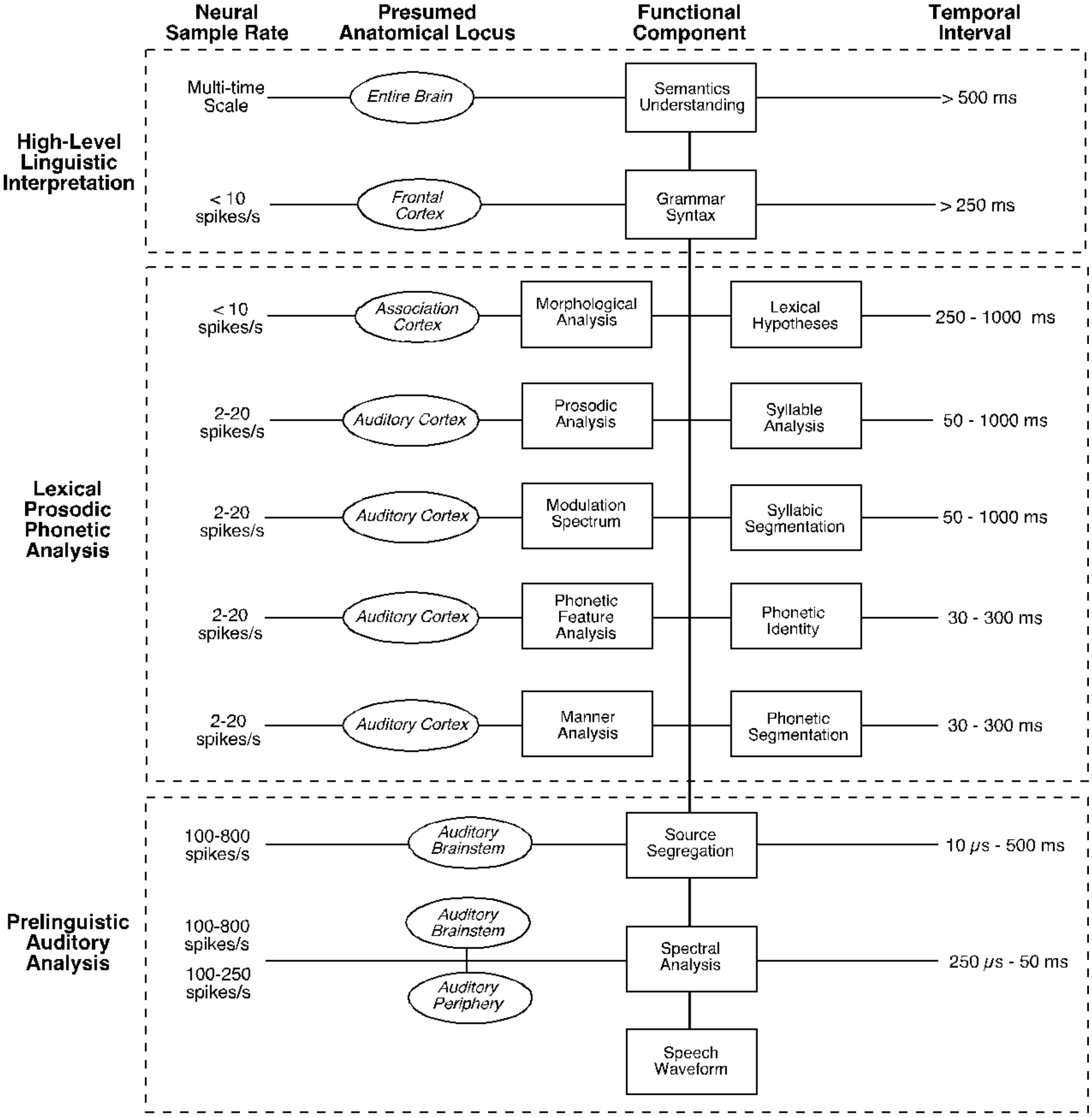
Neural Sample Rate
Presumed Anatomical Locus
Functional Component
Temporal Interval
High-Level Linguistic Interpretation
Multi-time Scale
Entire Brain
Semantics Understanding
> 500 ms
< 10 spikes/s
Frontal Cortex
Grammar Syntax
> 250 ms
Lexical Prosodic Phonetic Analysis
< 10 spikes/s
Association Cortex
Morphological Analysis
Lexical Hypotheses
250 - 1000 ms
2-20 spikes/s
Auditory Cortex
Prosodic Analysis
Syllable Analysis
50 - 1000 ms
2-20 spikes/s
Auditory Cortex
Modulation Spectrum
Syllabic Segmentation
50 - 1000 ms
2-20 spikes/s
Auditory Cortex
Phonetic Feature Analysis
Phonetic Identity
30 - 300 ms
2-20 spikes/s
Auditory Cortex
Manner Analysis
Phonetic Segmentation
30 - 300 ms
Prelinguistic Auditory Analysis
100-800 spikes/s
Auditory Brainstem
Source Segregation
10 μs - 500 ms
100-800 spikes/s
Auditory Brainstem
100-250 spikes/s
Auditory Periphery
Spectral Analysis
250 μs - 50 ms
Speech Waveform

**Figure 25.1**

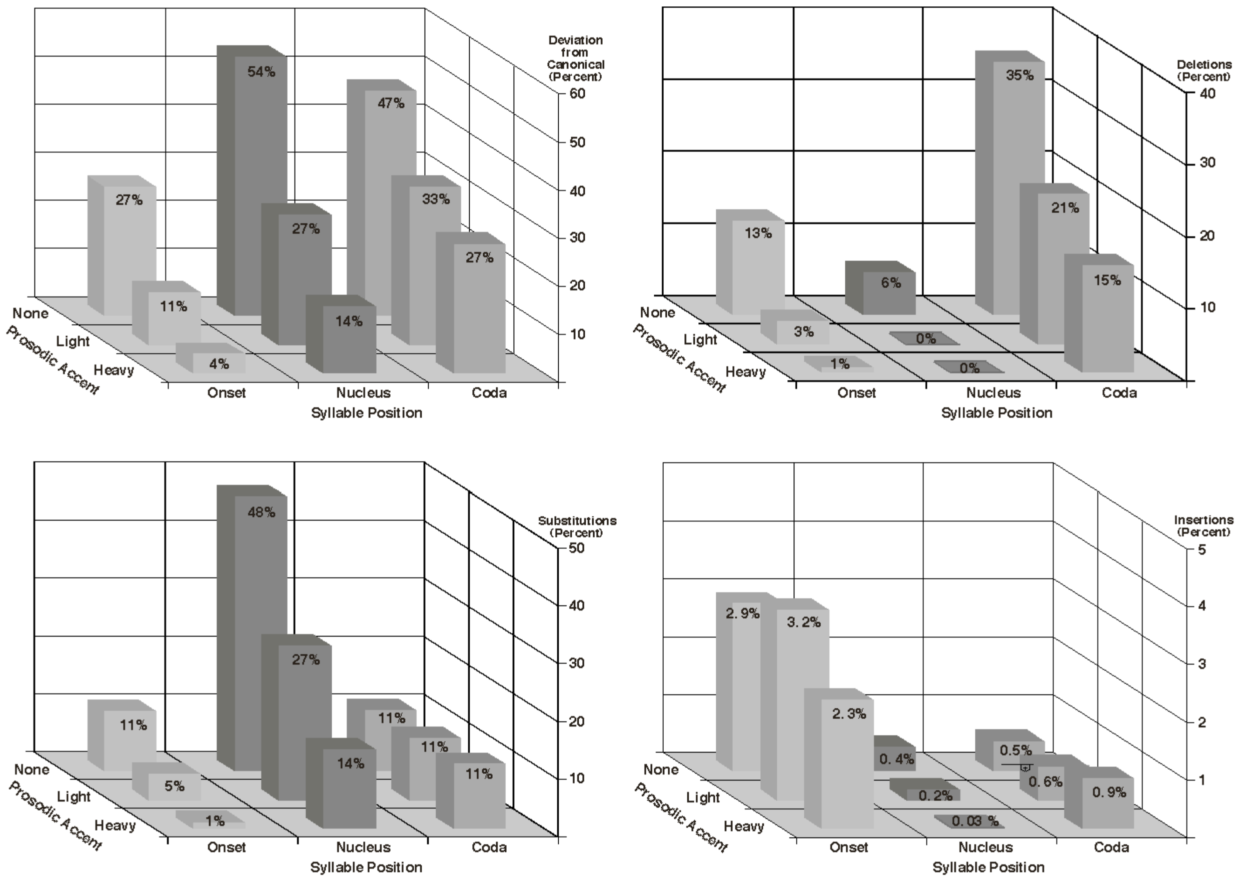
Deviation from Canonical (Percent)
60
50
40
30
20
10
27%
11%
4%
54%
27%
14%
47%
33%
27%
None
Light
Heavy
Prosodic Accent
Onset
Nucleus
Coda
Syllable Position
Deletions (Percent)
40
30
20
10
13%
3%
1%
6%
0%
0%
35%
21%
15%
None
Light
Heavy
Prosodic Accent
Onset
Nucleus
Coda
Syllable Position
Substitutions (Percent)
50
40
30
20
10
11%
5%
1%
48%
27%
14%
11%
11%
11%
None
Light
Heavy
Prosodic Accent
Onset
Nucleus
Coda
Syllable Position
Insertions (Percent)
5
4
3
2
1
2.9%
3.2%
2.3%
0.4%
0.2%
0.03 %
0.5%
0.6%
0.9%
None
Light
Heavy
Prosodic Accent
Onset
Nucleus
Coda
Syllable Position

**Figure 25.2**

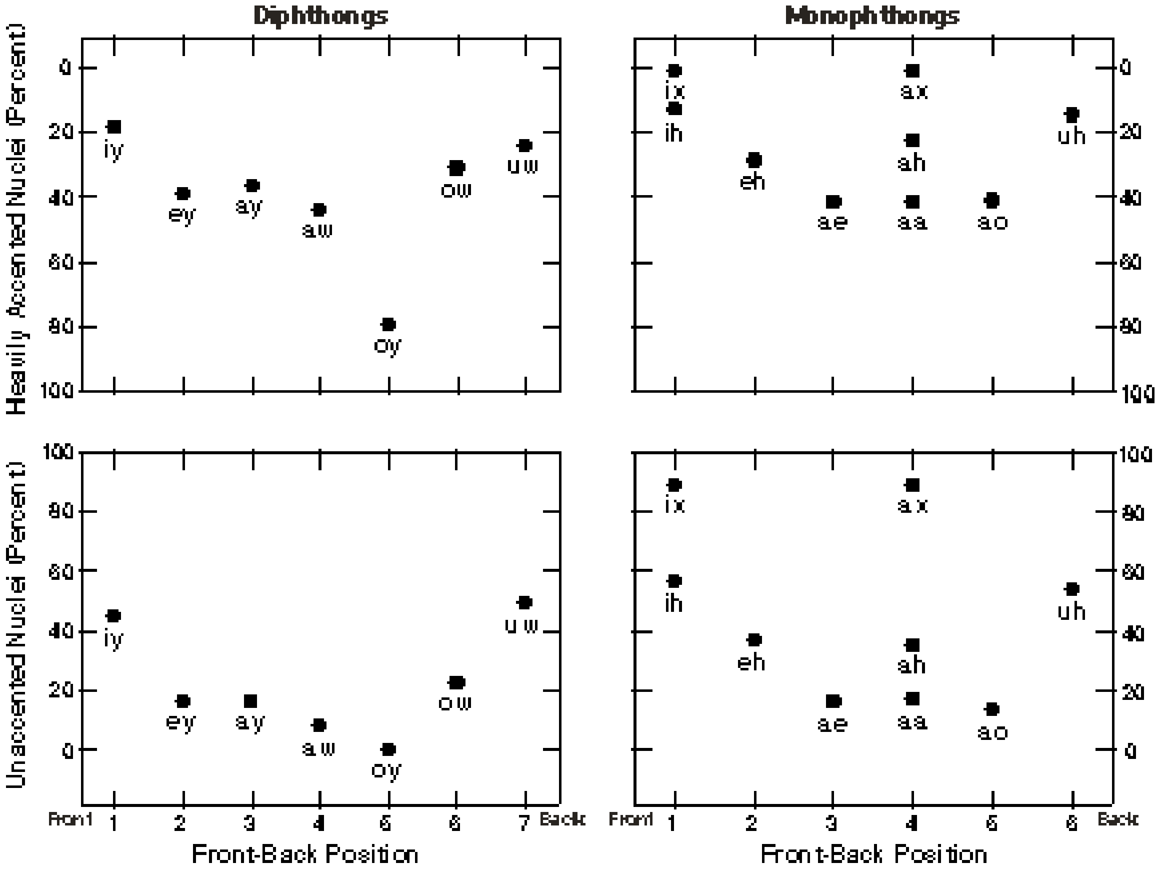
Diphthongs
Monophthongs
Heavily Accented Nuclei (Percent)
Unaccented Nuclei (Percent)
0
20
40
60
80
100
iy
ey
ay
aw
oy
ow
uw
ix
ih
eh
ae
ax
ah
aa
ao
uh
Front
1
2
3
4
5
6
7
Back
Front-Back Position

Figure 25.3

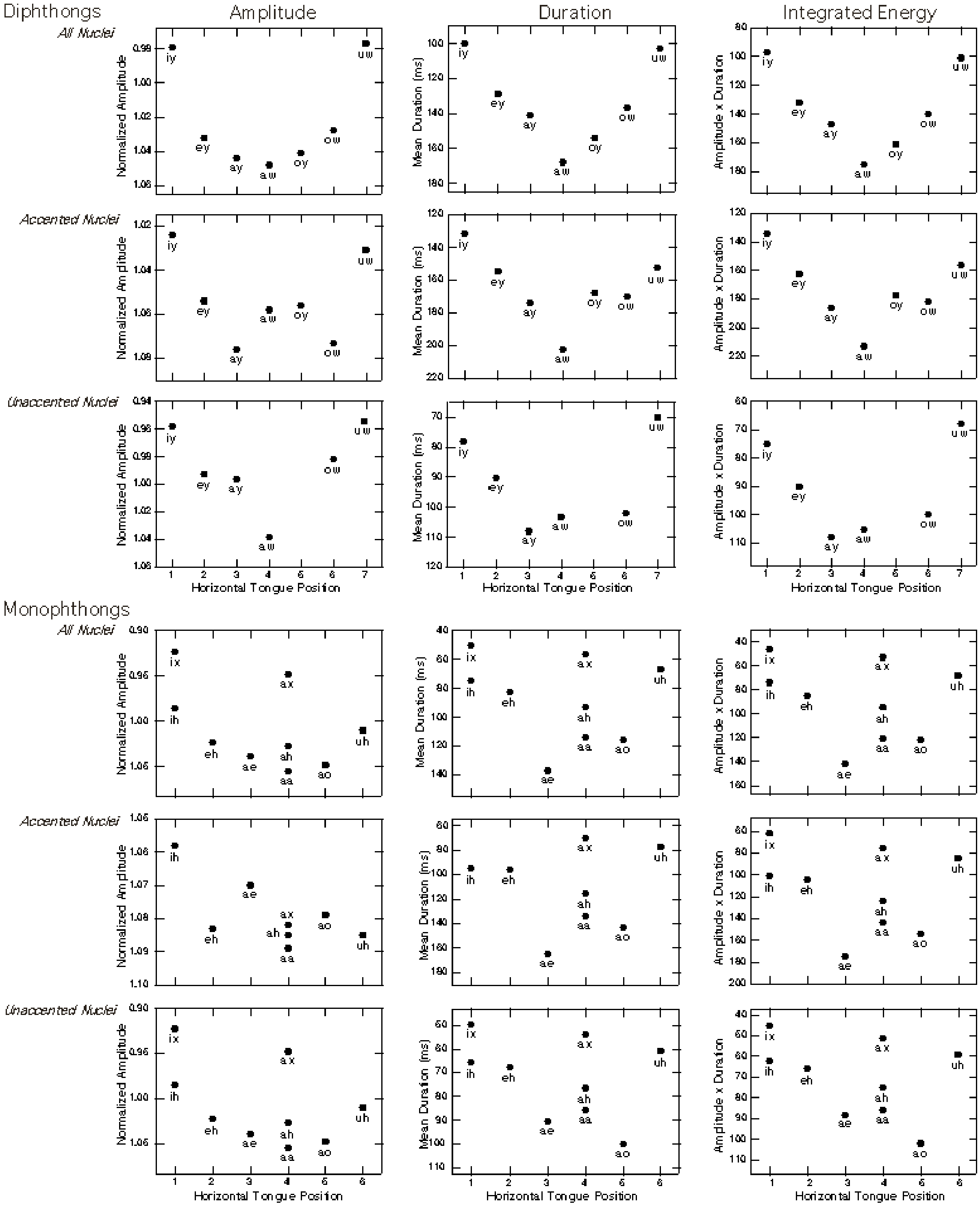
Diphthongs
Amplitude
Duration
Integrated Energy
All Nuclei
Accented Nuclei
Unaccented Nuclei
Normalized Amplitude
Mean Duration (ms)
Amplitude x Duration
iy
ey
ay
aw
oy
ow
uw
Horizontal Tongue Position
1
2
3
4
5
6
7
Monophthongs
ix
ih
eh
ae
ax
ah
aa
ao
uh

**Figure 25.4**

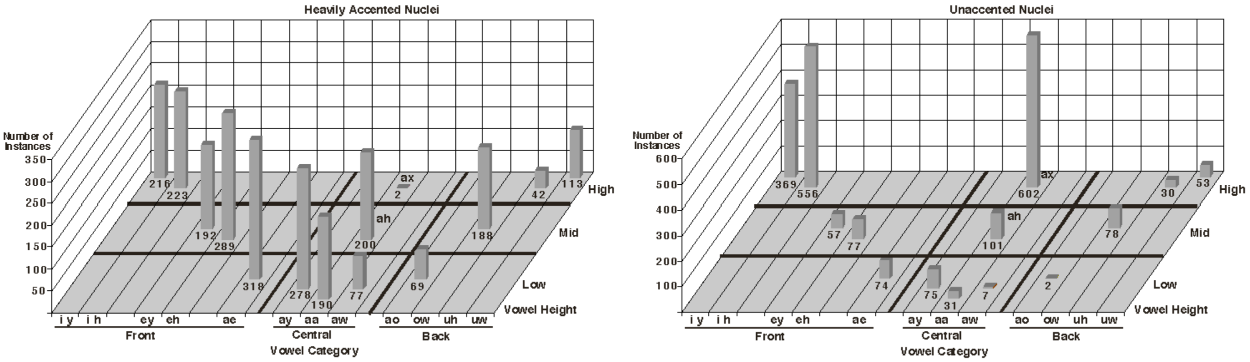
Heavily Accented Nuclei
Number of Instances
350
300
250
200
150
100
50
216
223
192
289
318
278
190
77
200
ah
ax
2
69
188
42
113
High
Mid
Low
Vowel Height
i y
i h
ey
eh
ae
ay
aa
aw
ao
ow
uh
uw
Front
Central
Back
Vowel Category
Unaccented Nuclei
Number of Instances
600
500
400
300
200
100
369
556
57
77
74
75
31
7
101
ah
602
ax
2
78
30
53
High
Mid
Low
Vowel Height
i y
i h
ey
eh
ae
ay
aa
aw
ao
ow
uh
uw
Front
Central
Back
Vowel Category

**Figure 25.5**

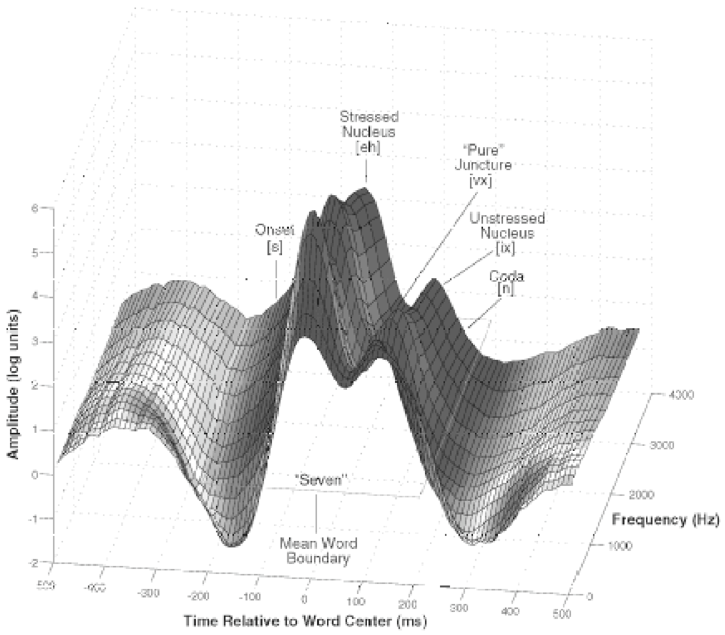
Stressed
Nucleus
[eh]
"Pure"
Juncture
[vx]
Unstressed
Nucleus
[ix]
Coda
[n]
Onset
[s]
"Seven"
Mean Word
Boundary
Amplitude (log units)
6
5
4
3
2
1
0
-1
-2
-500
-400
-300
-200
-100
0
100
200
300
400
500
Time Relative to Word Center (ms)
Frequency (Hz)
0
1000
2000
3000
4000

**Figure 25.6**

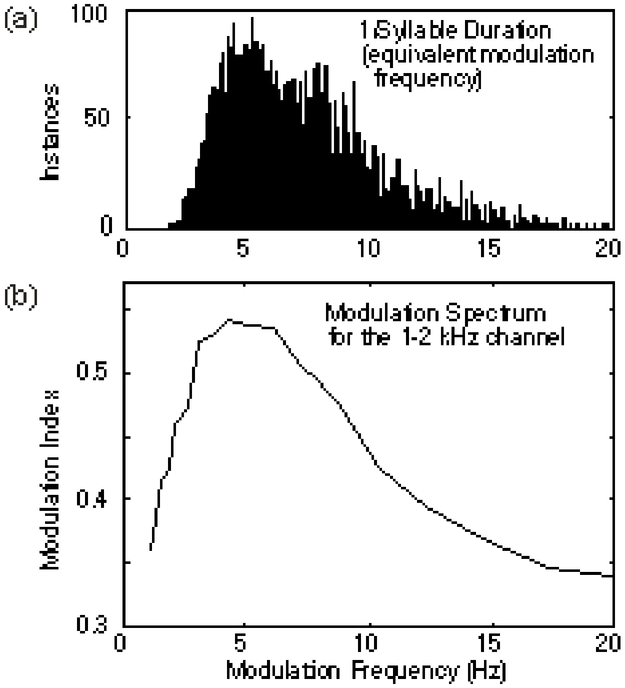
(a)
100
50
0
Instances
1/Syllable Duration
(equivalent modulation
frequency)
0
5
10
15
20
(b)
0.5
0.4
0.3
Modulation Index
Modulation Spectrum
for the 1-2 kHz channel
0
5
10
15
20
Modulation Frequency (Hz)

**Figure 25.7**