MARSLEN-WILSON, WILLIAM, ed. 1989. *Lexical representation and process*. Cambridge, Mass.: MIT Press.
PERKELL, JOSEPH S., & DENNIS H. KLATT, eds. 1985. *Invariance and variability in speech processes*. Hillsdale, N.J.: Erlbaum.
PICKETT, JAMES M. 1980. *The sounds of speech communication*. Baltimore: University Park Press.

Psycholinguistic Aspects

A listener who hears a spoken utterance has to recognize what was said. Any act of recognition involves matching an input to a pre-stored representation. In the case of speech recognition, the input is a sound pattern, and the pre-stored representation of meaning is conceptual; thus speech recognition consists in the translation of sound to meaning. The peripheral auditory system acts first on the incoming sound signal, and passes on to the brain a representation in terms of auditory features. The speech perception process, as studied by psycholinguists, begins at this point. This process takes the auditory features as input, and turns them into a representation which the hearer can use to make a selection from a stored set of sound/meaning associations.

The characteristics of these stored conceptual representations partly determine the nature of the 'translation' process. The set of potential messages is infinite; however, recognizers do not have infinite storage capacity. Therefore the stored set of meaning representations—the LEXICON—cannot include every message with which a recognizer might potentially be presented. The set of representations in the lexicon must be finite and must consist of discrete units. Yet the size of the store must be very large: the educated adult language user's vocabulary has been estimated to be around 150,000 words. Furthermore, its contents are highly heterogeneous; entries may range from grammatical morphemes, through words of widely varying length and structure, to quite complex idiomatic phrases. [*See* Processing; Production of Language.]

Part of the process of translating sound into meaning, therefore, consists in determining which portions of a signal correspond to which discrete stored units. This is essentially a problem of SEGMENTATION. The only segmentation that is logically required is that of a speech signal into lexical units. But auditory linguistic input extends over time; a portion of input corresponding to a particular lexical form is not instantaneously available in its entirety. Moreover, only rarely are recognizers presented with isolated lexical items. Most speech signals are made up of an effectively continuous stream of words. Momentary discontinuities within it do not correspond systematically to its linguistic structure.

Segmentation would be unproblematic if explicit boundary markers indicated which parts of the signal belonged together in a single lexical unit; however, reliable cues to lexical boundaries have not so far been discovered. One way around this problem is simply to match arbitrary portions of the auditory input against LEXICAL TEMPLATES. This crude process, in a number of different guises, is the basis of all systems for automatic SPEECH RECOGNITION [*q.v.*] currently in commercial use. However, such template-matching procedures are extremely inefficient. First, they involve a large number of futile access attempts, since the heterogeneity of lexical units means that the duration of the string to be tested cannot be predicted. Second, since they invoke a simple search procedure, the large size of the lexical stock means that each attempt at access requires a long search. This explains why all current commercial automatic speech recognizers are limited to very small vocabularies.

The largest problem, both for the realization of speech recognition by machines and for the explanation of speech perception by humans, is that the speech signal corresponding to a particular lexical representation is not a fixed acoustic form. It is no exaggeration to say that even two productions of the same utterance by the same person, speaking on the same occasion at the same rate, will not be completely identical. But within-speaker variability is tiny compared to the enormous variability across speakers and across occasions. Speakers differ in the length and shape of their vocal tracts, as a function of age, sex, and other physical characteristics; productions of a given sound by a large adult male and by a small child have little in common. Situation-specific variations include the speaker's current physiological state; thus the voice can change when the speaker is tired—or as a result of temporary changes in vocal tract shape such as a swollen or anaesthetized mouth, a pipe clenched between the teeth, or a mouthful of food. Other situational variables include distance between speaker and hearer, intervening barriers, and background noise. For all these reasons, acoustic signals vary greatly; if they are to be perceived as the 'same' speech entity, there must be some way of factoring out speaker- and situation-specific contributions. This is called the problem of NORMALIZATION across speakers.

A further source of variability results from the different varieties of a given language. Sounds can be articulated very differently in different dialects: compare English /r/ as spoken in Kansas, Boston, Bombay, Aberdeen, Sydney, Somerset, and Surrey. Dialects also differ in how they distinguish between sounds; thus Southern British English uses three different vowels in *foot, strut,* and *goose*; but Scottish has the same vowel in *foot* and *goose,* while Northern British has the same vowel in *foot* and *strut*. Listeners must therefore normalize for dialect variability as well. [*See* Dialectology.]

At the word level, variability also arises from speech style or register, and from the often related factor of speech rate. Consider the two words *did you*. In formal speech they would be pronounced [dɪdju]; a phonetic transcription shows five separate segments. A more casual style allows the [d] and [j] to merge into an affricate, giving [dɪdʒu]. If the words occur at the beginning of a phrase, the entire first syllable will often be dropped, leaving only the affrication as a trace of the word *did*: thus, [dʒu] *get paid yet?* Finally, in appropriate contexts the vowel of *you* can be reduced or lost entirely: [dʒə] *get it?* or [dʒæv] *any luck?* In the latter phrase, the affricate [dʒ] is performing the function of [dɪdju] in a formal, precise utterance of *Did you have any luck?*; yet there is virtually no overlap between the two transcriptions.

This extreme variability means simply that, if the lexicon were to store an exact acoustic representation for every possible form in which a given lexical unit might be presented as a speech signal, it would need infinite storage capacity. Therefore the lexical representation of the input signal, i.e. the sound component of the sound/meaning pairing, must be in a relatively abstract or normalized form. In consequence, the progression from auditory features to the input representation for lexical access necessarily involves a process of transformation.

These considerations together lead to the conclusion that the mapping from auditory features to lexical input representation should not be direct. On one hand, the problem of segmentation under conditions of continuity suggests that prelexical classification of speech signals into some representation below the word level would permit a more efficient system of lexical access. A sublexical representation overcomes the necessity for simple search procedures in lexical access, and hence removes the problem of the impracticable amount of time required to search a vocabulary of the size used by human recognizers. But the greatest advantage of a sublexical representation is that the set of potential units can be very much smaller than the set of units in the lexicon. However large and heterogeneous the lexical stock, sublexical representations allow any lexical item to be decomposed into a selection from a small and finite set of units.

On the other hand, the problem of the necessity of transformation also argues for an intermediate level of representation between auditory features and lexical input. If transformation is necessary in any case, then transforming the input into a small set of possibilities will be far easier than into a large set of possibilities.

The most obvious candidates for the role of intermediate representation have been the units of analysis used by linguistics. The PHONEME has been the most popular choice because, by definition, it is the smallest unit into which speech can be sequentially decomposed. Unsurprisingly, the central issue here is again speech variability—and the degree to which acoustic cues to phonemes possess constant, invariant properties which are necessarily present whenever the phoneme is uttered. At the phoneme level, variability is compounded by the phenomenon of coarticulation. Phonetic segments are context-sensitive, which means that a given segment may be spoken quite differently as a function of the other segments which surround it. Stop consonants are particularly sensitive to the identity of the following vowel; thus spectrograms of the words *day* and *do* look quite different in the consonant as well as the vowel portions. In some cases, these differences can be noticed even by the speaker; thus /k/ is articulated further forward in the vocal tract in *key* than in *caw*. Moreover, coarticulation effects are not limited to immediately adjacent segments; they can extend both forwards and backwards over several segments. Consider the utterance *She has to spruce herself up*: in most cases, the lip-rounding for the [u] in *spruce* is fully in place by the utterance of the word-initial [s], or even during the preceding syllable, and it does not disappear until well into the word *herself*. [*See* Coarticulation and Timing.]

The number of possible phonetic contexts in a language is not infinite, so the problem of variability which results from coarticulation might be considered tractable in principle. But the number of potential speakers and the number of potential speech situations are each truly infinite. Phoneme perception has been the primary research topic of phonetic work in this area, and much is now known about how listeners can use acoustic cues

to identify and distinguish between phonemes; nevertheless, no comprehensive solution to the normalization problem has yet been found. This means that no machine recognition system has been developed that can accurately identify phonemes. Further, it is not yet known whether human listeners use phonemes as an intermediate representation between auditory features and lexical representations.

The phoneme is not the only intermediate perceptual unit to have been proposed by speech scientists. Other such units include those above the phonemic level, such as SYLLABLES, and those below it, such as featural representations or spectral templates. In general, models of auditory word recognition assuming a level of representation using linguistic units—such as phonological features, phonemes, or syllables—have been developed within cognitive psychology, and have not directly addressed questions of machine implementation of speech recognition. Non-linguistic units such as 'diphones' or 'demisyllables' have been proposed by researchers who are concerned more with machine implementation than with psychological modeling.

In the above discussion, a simplifying assumption has been adopted, namely that there is only one type of auditory feature, one type of input representation to the lexicon, and one type of intermediate representation (if any). While this may be true for auditory features (which are constrained by the physiology of the auditory system), it is not necessarily true for the other two levels of representation. For example, there is variation in what may potentially constitute a lexical unit; relatively uninflected languages like Chinese contrast with highly inflected languages like Turkish. Similarly, there is variation in the potential characteristics of lexical input representations. Here there is a major distinction between languages which use prosody to distinguish between lexical units, and those which do not. The former group includes tone languages, like Chinese and Thai, and lexical-stress languages like English and Russian. The latter group includes fixed-stress languages like Polish or Hungarian, as well as non-tone non-stress languages like French. Finally, there is considerable variation across languages as to what linguistic units are viable candidates for prelexical representation. In particular, syllable structure can vary, from languages which allow only consonant-vowel syllables, to those like English in which syllables may be as different in structure as *a* and *scrounge*, and in which stress patterns result in a wide discrepancy in acoustic-phonetic clarity between the realization of stressed and unstressed syllables. [*See* Syllables.] Syllable boundaries, likewise, may be phonologically distinct or indistinct. These types of variation could perhaps encourage cross-linguistic differences in how speech is segmented, which could in turn imply that cross-linguistic differences also exist in the nature of prelexical or lexical input representations. That is, the very structure of a language may affect the way it is processed.

For a summary of research on acoustic cues to phonemes, see Borden & Harris 1984. Both phonetic and psycholinguistic work are reviewed by Jusczyk 1986 and by Pisoni & Luce 1986. A collection of the best current research on variability is Perkell & Klatt 1986.

ANNE CUTLER

BIBLIOGRAPHY

BORDEN, GLORIA J., & KATHERINE S. HARRIS. 1984. *Speech science primer: Physiology, acoustics, and perception of speech*. 2d ed. Baltimore: Williams & Wilkins.

JUSCZYK, PETER W. 1986. A review of speech perception research. In *Handbook of perception and human performance*, vol. 2, *Cognitive processes and performance*, edited by Kenneth R. Boff et al., pp. 1–57. New York: Wiley.

PERKELL, JOSEPH S., & DENNIS H. KLATT, eds. 1986. *Invariance and variability in speech processes*. Hillsdale, N.J.: Erlbaum.

PISONI, DAVID B., & PAUL A. LUCE. 1986. Speech perception: Research, theory, and the principal issues. In *Pattern recognition by humans and machines*, vol. 1, *Speech perception*, edited by Eileen C. Schwab & Howard C. Nusbaum, pp. 1–50. New York: Academic Press.

PERSIAN is a southwestern member of the Iranian language family [*q.v.*], in the Indo-Iranian branch of Indo-European. Its three major varieties, all official languages, are Persian of Iran (thirty million speakers), Dari in Afghanistan (five million, alongside East Iranian Pashto), and Tajiki in Soviet Tajikistan (2.2 million). Persian is the native tongue of about half the population of Iran; about 25 percent speak non-Persian Iranian languages such as Kurdish, Baluchi, and Pashto.

Among reference works, Lumsden 1810 is still the only extensive grammar that makes thorough use of indigenous Muslim grammar. Phillott 1919 is the most extensively documented description to date. Jensen 1931 is a descriptive and comparative grammar of Classical Persian, with notes on the modern language. Lazard