Segmentation and Units

- The units that words are built from could be: triphones, diphones, phonemes, syllables, morphemes,…
- The process of segmenting units (phonemes, syllables and/or words) could be explicit & cue driven or implicit in the recognition process.
- Languages may differ in both units and segmentation.
- Any representation and segmentation process must maintain the serial order of the units.

Early Attempts

Savin & Bever had listeners monitor for initial phonemes (e.g. /p/) or syllables (e.g. /pa/) in lists of syllables. They were faster to respond to syllables (the larger unit) than they were to phonemes (the smaller unit).

Doe this mean that the syllable is the basic unit of recognition?

Early Attempts - 2

McNeil & Lindig asked listeners to listen for phonemes, syllables or words. They were fastest when the unit they listened for matched the units that were presented. So, the task of listening for a particular unit does not tell you what the basic unit is.

Listeners are very accommodating, they will attempt to do the task that is given to them. The fact that they can do a task does not imply that the units they are using constitute *the* basic units of perception.

Serial Order

Regardless of the representational unit chosen, "something" is needed to maintain the serial order of the units.

That is, in recognizing the word /kæt/, we need to somehow keep the order of the units as /k/ then /æ/, then /t/. Otherwise, the talker could say /kæt/ and we hear /tæk/ or /ækt/.

Lashley described this problem in its general form in 1951.

Solutions to Serial Order

One approach is to organize the units into higher level structures. In production, the word controls the internal organization of the phonemes.

In perception, the activation of a phoneme could lead to partial activation of words. These, in turn, accumulate phoneme activation and maintain the order.

This does not solve the problem. As /t/ in /tæk/ is perceived, what is to keep it from activating the /t/ in /kæt/? You must *assume* that the /t/ is the beginning of the word and only activate the words beginning with /t/.

A Units Based Solution: Triphones

Wickelgren (1969) proposed a solution based on context sensitive allophones (triphones or Wickelphones).

Each unit represents both that phoneme and the adjacent phonemes (before and after).

So, cat is /$_{\#}k_{æ\ k}æ_{t\ æ}t_{\#}$/.

By coding the units on either side, the units are unique with respect to their position in the word. They maintain their serial order by virtue of being context sensitive.

Triphone Limits

The cost of this approach is that the number of basic units is larger (instead of 40 - 100, it is in the 1000s). However, there is little evidence that human memory and perception is built from a small number of primitives.

The other "cost" to this approach is that:

a) You still need to know where a word begins to start the coding with the correct initial segment.

and/or

b) There are acoustic cues to phoneme position that listeners use in perception.

Segmentation - Implicit

One approach is "implicit" segmentation. As embodied in Marslen-Wilson's early Cohort approach, and to some extent in TRACE, identifying one word tells you where the beginning of the next one is.

Lexical knowledge is certainly a powerful influence - Try listening to a language you do not speak.

The problem for lexically driven segmention is that a sequence like /aɪskrim/ is ambiguous.

Is this "I scream" or "ice cream". Does this imply that we do explicit segmentation based on acoustic information?

Segmentation - Explicit

English does have acoustic correlates to the position of the phoneme in a syllable and word and to the boundaries between syllables and words.

1) Allophonic cues. The precise realization of a phoneme depends upon syllable position. For example, the /k/s in “I scream” and “ice cream” are different.

2) Stress and metrical structure. Most English words start with a stressed syllable. In French, the last syllable in the word is stressed.

In English, stress and allophonic cues are correlated, providing information about word boundaries.

Acoustic Phonetic Information

As an example, consider /simigɚ/, /gretaɪ/, and /silivs/.

In all three cases, the first two syllables are stressed. All three are ambiguous in this abstract phonemic representation.

If the first is “seem eager”, then there is laryngal or glottal pulsing before the second vowel. If it is “see meager”, then there is nasalization in the closure and the first vowel is longer.

These are acoustic correlates to phoneme position and word juncture in English and listeners are sensitive to them. (see Umeda & Coker, 1974)

Metrical Structure (Stress)

Cutler & Norris proposed that listeners use the metrical structure of English to segment words. Most English words start with a strong syllable. So, segment in front of strong syllables.

Strong syllables in English are stressed and have full vowels.

The task they used was to embed words in nonsense strings. Listeners had to respond when they detected a word.

Word Spotting Task

Cutler and colleagues (McQueen, Norris, etc.) have used variations on a word spotting task to explore segmentation.

A word is embedded in a multi-syllabic context (e.g. "mint" - /mɪnt/ in /mintaɪf/ or "apple" in /vifæpəl/, /fæpəl/). Then, manipulate aspects of the signal that might influence segmentation to see what influences the listener's performance at detecting an embedded word.

For comparison with "mint", "sin" might be embedded in context to make /sintaɪf/.

Word Spotting Data

Mint is harder to pull out of /mintaɪf/ than /mintəf/. Sin is equally easy in both /sintaɪf/ and /sintəf/.

Since /taɪf/ is a strong syllable, it triggers segmentation between the /n/ and the /t/. This makes it hard to recognize “mint” in /mintaɪf/ because the /t/ has to be re-grouped with the first syllable.

Metrical Structure?

Is this really evidence for metrical structure?

The second syllable was always pronounced as /taɪf/ or /təf/, so there are junctural cues when the /t/ is syllable initial. The second syllable is stressed in /taɪf/, unstressed in /təf/. The second syllable has a full vowel in /taif/, schwa in /təf/. Which of these sets of information is influencing listeners?

When each of these is manipulated separately, juncture cues and stress (syllable duration and amplitude) are used by listeners, vowel quality is not.

Possible Word Constraint

Cutler, McQueen and Norris revised their proposal. The Possible Word Constraint (PWC) proposes that listeners segment the speech stream so that the segments could be words in the language.

Constraints include:

1. Allophonic details
2. Phonotactics & Neighborhoods
3. Syllable structure (/f/ can't be a word in English)
4. Accent, Stress (Metrical structure)

Word Spotting Data - 2

Listeners try to spot embedded words:

“fapple”, “vuhfapple”, “veefapple”

Spotting apple in fapple is harder than in vuhfapple. The explanation is that /f/ can’t be a syllable or word, but /vʌf/ can.

But, what other “cues” are present?

Word Spotting Errata

RTs are long, so it is not clear that this task reflects on-line word recognition.

Neighborhoods for the non-word syllable that remains behind did not influence performance.

Juncture cues and the identity of the juncture consonant do influence performance.

The identity of the vowel in the non-word syllable has no influence.

Segmentation Hierarchy

Mattys (2004; Mattys, White & Melhorn, 2005) proposed that there is a hierarch of cues to word segmentation.

Tier 1 - Lexical Influences of semantics, pragmatics, knowledge of words of language combine. This information dominates listening.

Tier 2 - Segmental Phonotactics, acoustic phonetics and related segmental information has a lesser influence but dominate if lexical information is poor.

Tier 3 - Metrical Word stress is least influential but can dominate if other sources are poor.

Hierarchy Data

Mattys (2004; Mattys, White & Melhorn, 2005) used a range of tasks including cross-modal priming.

When the signal was clear, stress played little or no role in segmentation when pitted against segmental or lexical cues. As segmental or lexical cues were degraded (e.g. using noise to mask segmental information), stress was influential.

Similarly, when lexical and allophonic cues are pitted against one another, lexical information dominates. When the lexical information is degraded, allophonics are most influential.

Hierarchy Data - 2

Fernandes, Ventura & Kolinsky (2007) used artificial language learning to investigate phonotactics and coraticulation.

When the signal was clear, coarticulation was the major cue to word boundaries. When noise was added, transitional probabilities (phonotactic information) dominated in influencing word boundaries.

This suggests that segmental information is not a single, homogeneous set of cues but should be subdivided.

Summary

Across a variety of tasks, segmentation of the speech signal into words seems to exploit a wide range of information. Is word recognition a process of "constraint satisfaction"?

In addition, not all cues are equally potent in all situations. Lexical information appears to be the strongest source of constraint with allophonic details also making a major contribution.

Most of these studies have relatively poor control of the acoustic phonetic details, so interpretation must be done cautiously in terms of the nature of segmental information.

Additional Reading

Davis, Marslen-Wilson & Gaskell (2002) *Journal Experimental Psychology: Human Perception and Performance, 28*(1), 218-244.

McQueen, Otake & Cutler (2001) Rhythmic cues and possible-word constraints in Japanese speech segmentation. *Journal of Memory and Language, 45*(1), 103-132.

Newman, Sawusch & Wunnenberg, (2011) Cues and cue interactions in segmenting words in fluent speech. *Journal of Memory and Language, 64*(4), 460-476.

Norris, McQueen, Cutler, Butterfield & Kearns (2001) Language-universal constraints on speech segmentation. *Language and Cognitive Processes, 16*(5/6), 637-660.