The SAGE Encyclopedia of Human Communication Sciences and Disorders

Segmentation of Speech

Contributors: Isabel Falé
Edited by: Jack S. Damico & Martin J. Ball
Book Title: The SAGE Encyclopedia of Human Communication Sciences and Disorders
Chapter Title: "Segmentation of Speech"
Pub. Date: 2019
Access Date: May 18, 2019
Publishing Company: SAGE Publications, Inc.
City: Thousand Oaks,
Print ISBN: 9781483380834
Online ISBN: 9781483380810
DOI: http://dx.doi.org/10.4135/9781483380810.n540
Print pages: 1650-1651

Segmentation of speech refers to the human ability to identify the boundaries of discrete language units (e.g., phonemes, syllables, words) in the continuous speech signal. Although the speech signal has many silent parts, these are not necessarily related to the boundaries of language units or pauses between boundaries. For instance, some of these silent portions can be due to speech articulation (e.g., stops closure). This identification process, which is automatic and fast for adults when listening to a familiar language, relies on several cues that work in a probabilistic way, merging information from different linguistic sources: lexico-semantic, syntactic, and prosodic. However, word segmentation cues are not completely reliable and they are language-specific. Native language experience plays an important role in defining relevant cues for speech segmentation in each language. Adults perform speech segmentation very easily in their own language provided that conversation or the speech signal is not affected by contextual conditions, such as a noisy environment. This entry discusses speech segmentation as a relevant issue for speech perception, language acquisition, language learning, and automatic speech recognition (ASR).

Identifying word forms in the oral stream and mapping them onto meaningful units is a complex task because words in speech are not produced by how they are in isolation, and they have different characteristics according to the phonetic and phonotactic context in which they occur. For example, in the speech stream, words are produced sequentially, and sometimes, concatenation and coarticulation occur between segments or even syllables. Both concatenation and coarticulation create an impact on the acoustic properties of words, changing the signal significantly. This means that the way a word sounds can be highly variable.

Speech segmentation is also a challenging task in language acquisition because infants do not have access to a mental lexicon, at least in their first year of life, so they cannot use it to identify language units (i.e., words) in the oral stream. The way infants deal with speech segmentation has been a research issue with contradictory results. Early studies proposed that infants perform speech segmentation first by learning words in isolation and then by identifying them within the speech continua. This claim has been attacked because the speech that infants hear is not produced word by word in isolation: Words are embedded in the speech stream. So infants must have another way to identify meaningful units. Another strong claim supports the hypothesis that infants rely on their ability to segment speech by using *sublexical segmentation*, that is, by using patterns provided by metrical cues, acoustic cues, and phonotactic cues. *Metrical cues* are related to strong and weak syllables and how they are ordered in a language. Researchers tested the idea that strong syllables in continuous speech could be considered word onsets by the listener, and results of studies on speech perception supported this view. *Acoustic cues*, such as longer segment duration or initial word segment aspiration, are cues that indicate the location of word beginnings in some languages. *Phonotactic cues* indicate possible word boundaries. Disallowed phonemic sequences in the language are key factors in aiding word segmentation.

Language learning by adults is quite different from language acquisition in children because adults already speak a language, having access not only to a lexicon but also to the experience of processing speech cues. In the process of learning a new language, nonnative speakers behave like infants, using their growing second language (L2) lexicon to add new words. Without the existence of a consolidated lexicon, successful segmentation in L2 depends on the hierarchies of nonlexical cues (metrical, acoustic, and phonotactic) in mother tongue or first language (L1) and L2. If the target language (L2) has juncture regularities similar to those of L1, segmentation process will be easier than if L2 has different juncture features, which can cause interference.

There are multiple cues available in the speech signal, which provide information to the listener, directing the segmentation process. Listeners must combine these cues, solving eventual conflicts, in order to be able to identify the input. Segmentation of speech seems to be an a priori bottom-up process because acoustic and statistical cues are the first to be available for listener processing. Moreover, research has provided evidence that the use of bottom-up segmentation works whenever other information (e.g., lexical) is missing or when there are problems in the speech signal (e.g., noise or distortion). Otherwise, higher level information surpasses bottom-up cues in the segmentation process. This assumption challenges the classical view that bottom-up is the only process utilized in speech segmentation and word identification. In fact, some researchers support the claim that segmentation of continuous speech is a product of the combination of bottom-up and top-down

processing. Listeners start to build their mental representation of the speech streaming through physical cues present in the signal (bottom-up), which allows them to activate potential lexical candidates. Whenever more signal and cues became available, lexical candidates are eliminated (top-down) and the word is recognized.

Research has also focused attention on the effects of aging in language production and perception. Speech segmentation is affected, in the first place, by the auditory decline usual in aging. However, this decline is not necessarily followed by a vocabulary or language decay. For instance, some studies report that vocabulary decline is small when compared to other abilities. Listeners may be forced to change strategies for speech segmentation in this period of life because auditory issues (e.g., loss of auditory abilities, higher sensitivity to environmental conditions) result in a growing cognitive load on the segmentation process, namely to process and resolve conflicting cues.

Speech segmentation is also a fundamental issue for natural language processing, namely for ASR. ASR systems usually analyze the speech signal as a concatenation of acoustic patterns that are classified to correspond to language units. Most of them use acoustic-based segmentation, which allows for an efficient use of a language model. The most common approach to speech segmentation in ASR is the probabilistic, which means that the acoustic signal corresponds to a word in the vocabulary with a certain degree of probability. After a word is spoken, a score for the match between the speech signal and a vocabulary item is computed. The word or word sequence that achieves the highest score is chosen to be the recognition result.

See also Acoustic Phonetics; Lexicon; Perception; Speech Perception, Theories of; Speech Recognition; Syllable

Isabel Falé
http://dx.doi.org/10.4135/9781483380810.n540
10.4135/9781483380810.n540

Further Readings

Cutler, A., & Norris, D. (1988). The role of strong syllables in segmentation for lexical access. Journal of Experimental Psychology: Human Perception and Performance, 14, 113–121.
Klatt, D. (1980). Speech perception: A model of acoustic-phonetic analysis and lexical access. In R. Cole (Ed.), Perception and production of fluent speech (pp. 243–288). Hillsdale, NJ: Erlbaum.
Kuhl, P. (2004). Early language acquisition: Cracking the speech code. Nature Reviews Neuroscience, 5, 831–843.
Mattys, S., & Bortfeld, H. (2017). Speech segmentation. In G. Gaskell & J. Mirković (Eds.), Speech perception and spoken word recognition (pp. 55–75). London, UK: Routledge.
McQueen, J. M. (1988). Segmentation of continuous speech using phonotactics. Journal of Memory and Language, 39, 21–46.