Lecture 8: Word Segmentation

Generated by LectureMate

February 21, 2025

1 Introduction

In this lecture, we discussed the concept of word segmentation, which involves identifying word boundaries in a continuous speech stream. This task is crucial for language acquisition, as children must learn to recognize where words begin and end in spoken language.

2 Announcements

Next week, office hours will begin, hosted by TA Jon Watts. Students can drop in to ask questions regarding course materials, assignments, and tutorials.

3 Understanding Word Segmentation

3.1 Definition

Word segmentation is the process of determining where one word ends and another begins in a stream of speech. This task is not as straightforward as it may seem, as speech is continuous and lacks clear boundaries.

3.2 Importance in Language Acquisition

Children, starting from birth, must master word segmentation to acquire language. They initially produce limited sounds and gradually progress to single words and then to more complex utterances.

4 Theoretical Frameworks

4.1 Transitional Probability

Transitional probability is a key concept in understanding word segmentation. It refers to the likelihood of one sound following another in a speech stream. This concept is a form of conditional probability that helps in identifying word boundaries.

4.2 Minimum Description Length (MDL)

MDL is a principle used to derive a lexicon from segmented speech. It aims to find the most efficient way to describe the input data, balancing the size of the lexicon and the encoding of the data.

4.3 Probabilistic Modeling

Probabilistic models combine symbolic knowledge with statistical information, allowing for a more nuanced understanding of language acquisition. This approach assumes some innate knowledge while also incorporating learned probabilities.

5 Stages of Language Development

Children progress through several stages of language development:

- **Vegetative Sounds:** Limited sounds produced by infants.
- **Cooing and Babbling:** Occurs between 6 to 10 months.
- **Single Word Utterances:** Around 10 to 18 months.
- **Two-Word Utterances:** Around 18 months to 2 years.
- **Telegraphic Speech:** Simple sentences without function words, around 2 years.
- **Full Sentences:** By age 3, children can produce complex sentences.

6 Challenges in Word Segmentation

6.1 Continuous Speech Stream

In spoken language, there are no clear spaces between words, making it difficult for infants to identify word boundaries. Variability in pronunciation further complicates this task.

6.2 Statistical Learning

Children utilize statistical regularities in speech to segment words. They can identify patterns and make educated guesses about where word boundaries lie based on the frequency of sound sequences.

7 Experimental Evidence

7.1 Safran et al. Study

An experiment by Safran et al. demonstrated that 8-month-old infants could segment words from a stream of nonsense syllables. The infants were able to distinguish between familiar and unfamiliar sequences, indicating that they used statistical cues to identify word boundaries.

7.2 Head Turn Preference Procedure

This method involved measuring infants' attention to new versus familiar word sequences. Infants showed a preference for novel sequences, suggesting they had learned to recognize word boundaries during the training phase.

8 Conclusion

In summary, word segmentation is a fundamental aspect of language acquisition that involves identifying word boundaries in continuous speech. Children use various strategies, including transitional probabilities and statistical learning, to segment words effectively. The concept of minimum description length provides a framework for understanding how children might derive a lexicon from segmented speech.