Acoustic Features of Cantonese Speech Acts: Prosodic Evidence from Words and Sentences

Meixuan Li, Bingxin Liu and Si Chen

The Hong Kong Polytechnic University

Objectives

Social interaction depends on the speaker's and listener's mutual understanding of not only *what* is said but *why* it is said. Action-theoretic accounts and recent studies demonstrate that speech acts or intentions can be conveyed by prosody alone in single words and nonwords. Whether this applies to Cantonese—a highly tonal language—remains an open question. Addressing this gap, the present study explores the mapping between basic speech acts and core prosodic parameters and the interface between phonetic form and pragmatic function.

Methodology

Eight native Hong Kong Cantonese speakers (4 males, 4 females; M age = 20.4 yr) with no self-reported hearing or speech disorders participated in a recording task. After a brief practice block, they were instructed to produce two sets of utterances presented in a randomized order: (i) isolated words and (ii) short carrier sentences. On every trial the screen displayed the description of a brief daily social scenario (e.g., "You are travelling with your classmates; it's night-time, everyone is sitting around a campfire, and you all feel bored") and an AI-generated voice provided the partner's preceding turn. Participants then were required to speak the target word or sentence as if they were part of the scenario, conveying one of the six speech acts—statement, doubt, command, suggestion, celebration and complaint. Recordings were made in a sound-proof booth with a head-mounted condenser microphone at 44.1 kHz/16-bit. All tokens were manually segmented and pitch-normalized by converting F0 to semitones relative to each talker's gender-specific mean. Praat was used to extract four prosodic acoustic values—mean F0, F0 range, mean intensity, and duration. These values served as input to the descriptive statistics and the subsequent linear discriminant analysis.

Results

Descriptive statistics showed intention-specific prosodic patterns in both word and sentence conditions. Doubt and Suggestion were realized with the highest mean F0 and the widest pitch spans (~2.3 st), whereas Statement and Complaint occupied the lowest pitch region and Command exhibited the narrowest span. Command and Celebration were the loudest (≈ 65 dB), Complaint the longest (≈ 0.8 s in words, 1.3 s in sentences), and Command the shortest. A jack-knife linear discriminant analysis using only these four cues—mean F0, F0 range, intensity and duration—classified tokens at 35.3 % accuracy for words and 31.8 % for sentences, roughly double the 16.7 % chance level ($\chi^2 > 184$, $p < 10^{-25}$). Commands and Statements were recognized best (≈ 46–54 % correct), indicating that global prosody alone carries sufficient information for above-chance decoding of Cantonese speech acts, yet leaves room for confusion among acoustically similar intentions.

Discussion

Overall, the results show that Cantonese speakers reliably produce pitch height, pitch span, loudness, and timing patterns to convey different speech acts. These global prosodic patterns are strong enough to yield machine classification at twice-chance accuracy, yet not distinctive enough to prevent specific confusions. Expanding the measurements to dynamic and spectral features and recruiting a larger speaker group should improve performance and explain how acoustic features jointly encode intentions.