# Indirect Acquisition of Pitch Comprehension in Speech-to-Speech Large Language Models

Tace McNamara[1], Jon McCormack[2], and Jesper Kjeldskov[3]

[1,2]SensiLab
[1,2,3]Monash University

The ability to draw meaning from changes in pitch is common to both music and spoken communication [11, 3, 15, 8]. However, humans do not require musical education to recognise and appreciate features of pitch in musical contexts [2]. In a similar way, we hypothesised that some degree of pitch comprehension may have emerged in speech-to-speech (S2S) Large Language Models as an indirect result of real-time conversation training.

To test this, we developed a zero-shot method to evaluate monophonic pitch comprehension and assessed four S2S models: Gemini 2.5 Pro [5], GPT-4o, Qwen2.5-Omni-7B [4] and Phi-4 [13]. The factorial experiment design utilises matched audio-prompt pairs to compare performance across three auditory conditions: speech, song, and instrumental sound sources. Organised to reflect a continuum of increasing complexity and music-specific knowledge requirements, the test evaluates five domains of pitch processing: direction, distance, contour, tonality, and melody.

The speech condition achieved the highest mean performance of all sound sources with an above-chance accuracy of 45%. Interestingly, both song and instrumental conditions shared a slightly above-chance mean accuracy of 31%, suggesting that models favour speech rather than the human voice *per se*. Pitch direction classification tasks achieved the highest mean accuracy (47.7%) while models performed worst on contour classification (18%) where bi-directional pitch changes proved challenging. High inter-model agreement suggests a commonality in system architecture that currently lacks the capacity for dynamic tracking of pitch changes over time.

These results indicate that dedicated musical training may be necessary to perform complex pitch processing however, improvement in this area has the potential to improve prosodic language comprehension Toh et al. [16] and linguistic encoding Wong et al. [17], Lee and Hung [11], and Glenn Schellenberg and Winner [6]. This research contributes to the growing literature on emergent abilities in LLMs [7, 14, 9] with the potential to further our understanding of the relationship between music and language [2, 1], and their uncertain origins in human evolution [10, 12].

## References


[1] Mireille Besson and Daniele Schön. "Comparison between Language and Music". In: *Annals of the New York Academy of Sciences* 930.1 (2001), pp. 232–258. ISSN: 1749-6632. DOI: 10.1111/j.1749-6632.2001.tb05736.x. (Visited on 06/27/2025).

[2] Dana Boebinger et al. "Music-Selective Neural Populations Arise without Musical Training". In: *Journal of Neurophysiology* 125.6 (June 2021), pp. 2237–2263. ISSN: 0022-3077. DOI: 10.1152/jn.00588.2020. (Visited on 06/20/2025).

[3] Steven Brown. "Are Music and Language Homologues?" In: *Annals of the New York Academy of Sciences* 930.1 (June 2001), pp. 372–374. ISSN: 0077-8923, 1749-6632. DOI: 10.1111/j.1749-6632.2001.tb05745.x. (Visited on 06/27/2025).

[4] Yunfei Chu et al. *Qwen-Audio: Advancing Universal Audio Understanding via Unified Large-Scale Audio-Language Models*. Dec. 2023. DOI: 10.48550/arXiv.2311.07919. arXiv: 2311.07919 [eess]. (Visited on 06/17/2025).

[5] Gemini Team, Google. *Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities*. Tech. rep. Google DeepMind, June 2025. (Visited on 07/01/2025).

[6] E. Glenn Schellenberg and Ellen Winner. "Music Training and Nonmusical Abilities: Introduction". In: *Music Perception* 29.2 (Dec. 2011), pp. 129–132. ISSN: 0730-7829, 1533-8312. DOI: 10.1525/mp.2011.29.2.129. (Visited on 06/27/2025).

[7] Thilo Hagendorff. *Machine Psychology: Investigating Emergent Capabilities and Behavior in Large Language Models Using Psychological Methods*. Oct. 2023. arXiv: 2303.13988 [cs]. (Visited on 12/14/2023).

[8] Ilana Harris et al. "Is Song Processing Distinct and Special in the Auditory Cortex?" In: *Nature Reviews Neuroscience* 24.11 (Nov. 2023), pp. 711–722. ISSN: 1471-0048. DOI: 10.1038/s41583-023-00743-4. (Visited on 06/20/2025).

[9] Michal Kosinski. "Theory of Mind Might Have Spontaneously Emerged in Large Language Models". In: (Nov. 2023). DOI: 10.48550/arXiv.2302.02083. arXiv: 2302.02083 [cs]. (Visited on 01/03/2024).

[10] Lily N. C. Law and Marcel Zentner. "Assessing Musical Abilities Objectively: Construction and Validation of the Profile of Music Perception Skills". In: *PLoS ONE* 7.12 (Dec. 2012), e52508. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0052508. (Visited on 06/19/2025).

[11] Chao-Yang Lee and Tsun-Hui Hung. "Identification of Mandarin Tones by English-speaking Musicians and Nonmusicians". In: *The Journal of the Acoustical Society of America* 124.5 (Nov. 2008), pp. 3235–3248. ISSN: 0001-4966, 1520-8524. DOI: 10.1121/1.2990713. (Visited on 06/20/2025).

[12] Erin Mcmullen and Jenny R. Saffran. "Music and Language: A Developmental Comparison". In: *Music Perception* 21.3 (Mar. 2004), pp. 289–311. ISSN: 0730-7829, 1533-8312. DOI: 10.1525/mp.2004.21.3.289. (Visited on 06/27/2025).

[13] Microsoft et al. *Phi-4-Mini Technical Report: Compact yet Powerful Multimodal Language Models via Mixture-of-LoRAs*. Mar. 2025. DOI: 10.48550/arXiv.2503.01743. arXiv: 2503.01743 [cs]. (Visited on 06/18/2025).

[14] Stefano Nolfi. "On the Unexpected Abilities of Large Language Models". In: *Adaptive Behavior* 32.6 (Dec. 2024), pp. 493–502. ISSN: 1059-7123. DOI: 10.1177/10597123241256754. (Visited on 02/12/2025).

[15] Adam Tierney et al. "Speech versus Song: Multiple Pitch-Sensitive Areas Revealed by a Naturally Occurring Musical Illusion". In: *Cerebral Cortex* 23.2 (Feb. 2013), pp. 249–254. ISSN: 1460-2199, 1047-3211. DOI: 10.1093/cercor/bhs003. (Visited on 06/24/2025).

[16] Xin Ru Toh et al. "Enduring Musician Advantage among Former Musicians in Prosodic Pitch Perception". In: *Scientific Reports* 13.1 (Feb. 2023), p. 2657. ISSN: 2045-2322. DOI: 10.1038/s41598-023-29733-3. (Visited on 06/26/2025).

[17] Patrick C M Wong et al. "Musical Experience Shapes Human Brainstem Encoding of Linguistic Pitch Patterns". In: *Nature Neuroscience* 10.4 (Apr. 2007), pp. 420–422. ISSN: 1097-6256, 1546-1726. DOI: 10.1038/nn1872. (Visited on 06/20/2025).