Hearing smiles in the crowd: how babble noise shapes smiled speech perception

Rong Li[1], Esther Janse[2], Dirk Heylen[1], Khiet P. Truong[2,1]
[1]Human Media Interaction, University of Twente, the Netherlands,
[2]Centre for Language Studies, Radboud University, the Netherlands

Smiles are ubiquitous in everyday interaction and are conveyed not only visually but also through the voice, where they are reflected in systematic variations in prosody and voice quality. For instance, smiling can increase the second formant (F2) frequency, particularly in rounded vowels, and has been associated with increased intensity and changes in higher-formant energy (Barthel & Quené, 2015; Ponsot et al., 2018). Listeners can also detect smiles in speech from voice alone. In particular, spread-lip speech primarily reflects the acoustic consequences of the facial gesture, whereas amused speech additionally involves prosodic modulation beyond these mechanical effects (Aubergé & Cathiard, 2003; Drahota et al., 2008). Yet auditory smile perception has largely been studied under clean laboratory listening conditions. We therefore investigated whether smile-related vocal information remains robust under multi-talker babble noise.

We examined two aspects of auditory smile perception: (1) detecting a smile-like quality in speech and (2) categorizing smile-like speech as amused or spread-lip. Stimuli were selected from the AMuS corpus (El Haddad et al., 2017) and comprised neutral, amused, and spread-lip productions from one French-speaking male and one English-speaking female (with the French materials presented to French-fluent and English materials to English-fluent listeners). Recordings were presented in three listening conditions: Quiet, −3 dB, and −6 dB SNR, using multi-talker cafeteria babble from the NOISEX-92 database (Varga & Steeneken, 1993).

Seventy-five adults (38 English-fluent and 37 French-fluent listeners) completed two online binary forced-choice tasks. Task 1 contrasted smile-like speech (amused and spread-lip) with neutral speech, whereas Task 2 contrasted amused with spread-lip speech. Sentence materials were rotated across noise conditions so that each sentence occurred only once per participant. Trial-level classification accuracy was analyzed using generalized linear mixed-effects models (GLMMs), complemented by Signal Detection Theory (SDT) analyses (Hautus et al., 2021) of perceptual sensitivity (d'), decision criterion (c), and Area Under the ROC Curve (AUC).

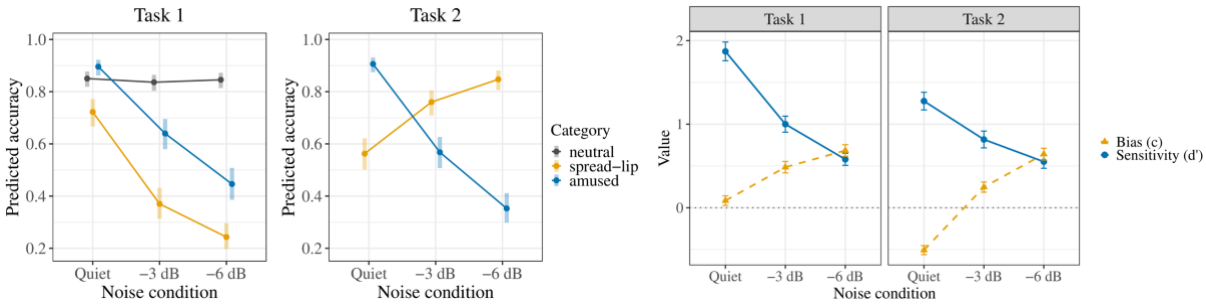
In Task 1, classification accuracy for both spread-lip and amused speech deteriorated substantially with increasing noise, whereas neutral-speech detection remained relatively stable, yielding a significant Noise x Category interaction. Amused speech was recognized more accurately than spread-lip speech across all noise levels. Perceptual sensitivity decreased systematically with increasing noise, accompanied by a progressive shift toward a more conservative decision criterion (Figure 1). Thus, participants not only had less sensory evidence available for smile judgments but also became less inclined to report hearing a smile.

Task 2 showed a pronounced interaction between noise condition and smile category. Under quiet conditions, amused speech was classified substantially more accurately than spread-lip speech. Under increasing noise, amused accuracy declined while spread-lip accuracy increased, producing a crossover pattern (Figure 1). Sensitivity for distinguishing the two smile types nevertheless decreased steadily, while a liberal bias toward "amused" responses in quiet reversed to a conservative bias under noise.

Across both tasks, noise masking therefore reduced the sensory evidence available for smile judgments while also shifting listeners toward a more conservative decision strategy as uncertainty increased. The findings highlight the importance of separating perceptual sensitivity from response bias when studying smiled speech in realistic noisy environments and when evaluating speech technologies operating without visual context.

Figure 1

Effects of noise on auditory smile perception



Note. Left: GLMM-predicted accuracy by task, category, and noise condition (95% CIs). Right: mean sensitivity (d') and decision criterion (c) across noise conditions for both tasks (SE). Reproduced from Li et al. (2026).

Author Note

This work has been published in the *Proceedings of Interspeech 2026* (Li et al., 2026). The present abstract summarizes the same study and results.

References

Aubergé, V., & Cathiard, M. (2003). Can we hear the prosody of smile? *Speech Communication, 40*(1), 87–97.

Barthel, H., & Quené, H. (2015). Acoustic-phonetic properties of smiling revised: Measurements on a natural video corpus. In *Proceedings of the 18th International Congress of Phonetic Sciences*.

Drahota, A., Costall, A., & Reddy, V. (2008). The vocal communication of different kinds of smile. *Speech Communication, 50*, 278–287.

El Haddad, K., Torre, I., Gilmartin, E., Çakmak, H., Dupont, S., Dutoit, T., & Campbell, N. (2017). Introducing AmuS: The Amused Speech Database. In *Statistical Language and Speech Processing* (pp. 229–240). Springer International Publishing.

Hautus, M. J., Macmillan, N. A., & Creelman, C. D. (2021). *Detection theory: A user's guide*. Routledge.

Li, R., Janse, E., Heylen, D., & Truong, K. P. (2026). Hearing smiles in the crowd: How babble noise shapes smiled speech perception. *Proceedings of Interspeech 2026*, 4690–4694.

Ponsot, E., Arias, P., & Aucouturier, J.-J. (2018). Uncovering mental representations of smiled speech using reverse correlation. *The Journal of the Acoustical Society of America, 143(*1), EL19–EL24.

Varga, A., & Steeneken, H. J. M. (1993). Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems. *Speech Communication, 12*(3), 247–251.