



# Impact of Background Noise and Contribution of Visual Information in Emotion Identification by Native Mandarin Speakers

Minyue Zhang, Hongwei Ding\*

Speech-Language-Hearing Center, School of Foreign Languages, Shanghai Jiao Tong University, Shanghai, China

Zhang.my@sjtu.edu.cn, hwding@sjtu.edu.cn

## Abstract

Many studies on emotion processing considered little about the issue of ecological validity and insufficient attention has been drawn to uni-sensory and multisensory emotion perception in challenging environments. The current research explored how adding multi-talker babble noise impacts emotion perception and how visual information affects the results in comparison with the audio-alone conditions. Forty native Mandarin participants (21 females and 19 males) were asked to identify the emotion according to the auditory or audiovisual information they received. Results showed that the emotion identification accuracy was significantly lower in noisy conditions than in noiseless ones, whether additional visual information was presented simultaneously or not. In noisy environments, providing multisensory emotional information greatly facilitated recognition performances even when the visual information was less reliable. To conclude, multi-talker babble noise had a corrupting effect on emotion identification, which worked in both unisensory and multisensory settings, and emotion perception is a robust multisensory situation that follows the inverse effectiveness principle.

**Index Terms:** multisensory integration, emotion perception, babble noise, audiovisual

## 1. Introduction

Successful understanding of emotions plays an important role in social interaction, allowing people to be aware of the feelings or intentions of others and communicate effectively [1]. Emotional cues can be expressed in several forms, e.g., affective prosody, facial expressions, semantic meaning, and body gestures. In face-to-face daily communication, these auditory and visual cues are usually conveyed simultaneously, rendering the integration of such cues essential and inevitable [2].

The multisensory nature of emotion perception in social contexts has elicited a multitude of studies over the past decades that explored how emotional information from different modalities is processed and merged [3]. Many researchers found that presenting congruent affective prosody and emotional facial expressions at the same time facilitates emotion recognition [4, 5] and a series of behavioral and neurophysiological studies have robustly replicated this finding [6-8]. Participants generally responded more accurately and faster to audiovisual emotional stimuli as compared to auditory stimuli alone, demonstrating a remarkable visual contribution [8]. Nevertheless, these studies

were mostly conducted in quiet laboratory settings while in the real world, background noise is nearly ubiquitous, which makes it questionable to what extent the multisensory benefits observed in these experiments are ecologically valid and can be generalized to realistic scenarios.

Ecological validity, referring to whether one can generalize from observed behavior in laboratories to natural behavior in the real world [9], has become a concern in multiple dimensions of experimental work. Some studies investigating multisensory emotion perception took this issue into consideration and adopted ecologically relevant materials. Collignon, et al. [6] used dynamic visual and vocal clips of emotional expressions and obtained bimodal stimuli by simultaneously presenting the visual and the auditory clips. Compared with previous experiments using static facial expressions as visual stimuli, their stimuli approximated real-life conditions of social communication much more, but certain imperfections still existed. First, the auditory and the visual clips of emotional expressions that were used to combine into bimodal stimuli came from two separate stimulus sets and it was not explicitly stated whether the dynamic facial expressions and the emotional voices were synchronized with each other during the “bimodal” presentation. It was thus quite possible that in the bimodal condition, the sounds that participants heard were more or less not in sync with the pictures they saw, which would substantially decrease the degree of ecological validity. Another issue for the research is that its noisy conditions were created by adding white Gaussian noise, which was less encountered in real-life social environments [10, 11]. In the study conducted by de Boer, et al. [12], they did not use white noise to mask the auditory channel but chose to mimic a degradation similar to age-related sensorineural hearing loss by manipulating the phonetic properties of the auditory clips. However, in order to avoid the potential confounding from linguistic content, they used pseudo-language as the content of auditory stimuli, which is less ecologically valid since pseudo-language generally would not appear in real-life conversations.

To provide insights into multisensory emotion perception with a higher ecological value, we conducted the present study, attempting to explore the impact of background noise on unisensory and multisensory emotion recognition and the contribution of adding synchronous visual information on recognition performances in challenging listening conditions. We used highly ecological stimuli and examined how emotion recognition accuracy fluctuated with the disturbance of multi-talker babble noise in unisensory (auditory) and multisensory (audiovisual) emotion identification tasks by 40 native

\* Corresponding author

Mandarin participants. For the multisensory tasks, we also tested whether the visual information was clean or degraded affected its contribution. In accordance with multisensory benefits in emotion processing, we predicted a significant contribution of visual information to emotion recognition in noisy conditions (Hypothesis 1). It was also expected that despite the visual contribution, the corruption brought by background babble noise would still undermine performances in multisensory emotion recognition tasks (Hypothesis 2). We hope to provide more ecologically valid understandings of how emotional signals integrate and interact in multisensory emotion processing.

## 2. Methods

### 2.1. Participants

Forty healthy young adults studying at Shanghai Jiao Tong University (21 females and 19 males, mean age  $\pm$  SD: 22.19  $\pm$  2.76) participated in the experiment. All participants spoke Mandarin Chinese as their primary language and the dominant language in daily use and had normal or corrected-to-normal vision and normal hearing as verified by standard audiological screening [13]. None reported a history of speech, language, or hearing impairment or any psychological or neurological conditions. Each of them completed written informed consent before the experiment started, and was financially compensated after the experiment for their participation.

### 2.2. Stimuli

The stimuli were eight audiovisual monosyllabic interjections (嘿, 啊, 哎, 呀, 哈, 诶, 咳, and 哦 (International Phonetic Alphabet [xei], [a], [ai], [ya], [xa], [ei], [xai], and [o])) with four emotions, i.e., happy, sad, angry, and calm. We used interjections as stimuli because they are important devices in real-life conversations to express a mental or emotional attitude or state [14] and are meanwhile semantically empty [15].

Table 1: Duration (milliseconds) of the stimuli.

Emotion	Mean	SD
Happy	423.50	72.76
Sad	765.94	207.99
Angry	443.19	75.98
Calm	473	78.13
<b>Overall</b>	<b>526.41</b>	<b>184.75</b>

Each interjection was produced in a soundproof booth by two amateur actors (one male and one female) who were native Mandarin speakers, resulting in  $8 \times 4 \times 2 = 64$  videos. The images were captured using a Sony HDR-PJ580 camera and the sounds were recorded using a Neumann U87 Ai condenser studio microphone (Georg Neumann, Berlin, Germany) and a Fireface UFX soundboard (RME Fireface; RME Inc.). The recorded images and sounds were combined into videos with the soundtrack in sync with the picture. The resulting videos were then edited with Adobe Premiere Pro CC (Adobe Systems, California) into video clips with a resolution of  $1440 \times 1080$  and a digitization rate of 25 frames per second (1 frame = 40 ms). The soundtracks were digitized at a sampling rate of 44,100 Hz with a 16-bit amplitude resolution, and normalized peak value (90%) using Adobe Audition CC (Adobe Systems, California). The duration

measures of the stimuli are shown in Table 1. The 64 videos and the 64 soundtracks were respectively validated by 30 native Mandarin Chinese who did not participate in the current experiment. The mean identification rates for the videos and the soundtracks were both above 88%.

The stimuli were presented in six conditions (Figure 1): auditory-only (AO), unimodal auditory stimuli plus acoustic noise (AnO); clean audiovisual (AV), audiovisual stimuli plus visual noise (AVn), audiovisual stimuli plus acoustic noise (AnV), and audiovisual stimuli plus both acoustic and visual noise (AnVn). The acoustic noise was an 8-talker babble noise created by Chen, Hu, and Yuan [16]. Multi-talker babble noise, one of the real-life noise types [11], is a common speech interference that many listeners encounter regularly in everyday cocktail-party listening [17]. The babble noise was normalized peak value (90%) using Adobe Audition CC (Adobe Systems, California) and added to the sound channel with the signal-to-noise ratio (SNR) being  $-13\text{dB}$ , which was confirmed through the pilot testing to avoid ceiling performance yet partially mask the speech token. The presentation of the auditory babble noise began about 500 ms prior to the beginning of the stimulus token and ended about 500 ms following token offset. Visual information was degraded by overlaying an image of white noise onto the video track using Adobe Premiere Pro CC (Adobe Systems, California) in the proportion of 70% to significantly lower the accurate identification of the stimuli presented only visually ( $t(39) = 7.20, p < 0.001$ ) and meanwhile allow analysis of emotional meaning [6, 18]. Pilot testing confirmed that this proportion of white noise could avoid ceiling performance yet partially mask the visual channel.

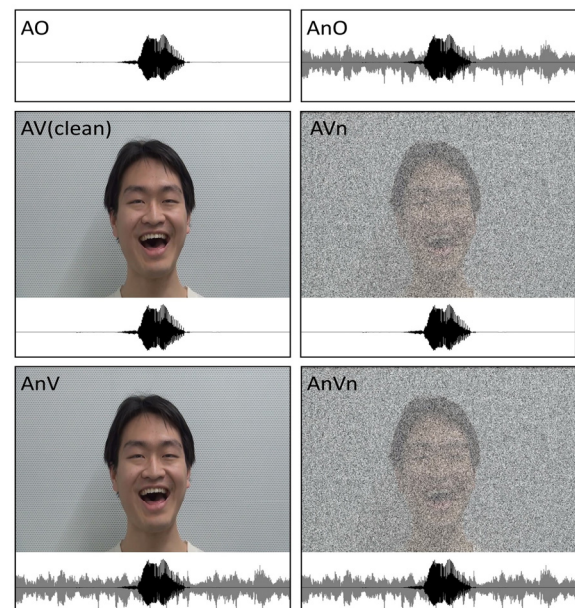


Figure 1: Examples of the stimuli for each condition.

### 2.3. Procedure

The visual channel was displayed in the center of an LCD screen on a white background and the auditory channel was presented binaurally using high-fidelity circumaural headphones (Sennheiser HD 280 Pro) at a comfortable level (70 dB SPL). The participants were presented a total of 384 stimuli ( $4$  [emotions]  $\times 8$  [interjections]  $\times 2$  [actors]  $\times 6$

[conditions]). These stimuli were presented in six blocks (one for each condition) and the blocks were counterbalanced across participants. Each block included 64 trials, which were presented in a pseudorandom order. On each trial, participants watched and/or heard an interjection produced by the male or female speaker. Between trials, a fixation cross was placed on the screen to center participants' gaze before hearing/viewing the next audio/video. Participants were asked to respond as quickly and accurately as possible by pressing one of the four response keys that were mapped to the four emotional categories. Correspondences between the category and the key were counterbalanced across participants but were held constant throughout the experiment for each participant. The experimenter carefully checked the participants' understanding of the general procedures as well as the correspondence between the keys and the emotional categories, and then started the experiment. Each block started with a training test consisting of four trials. Participants needed to reach 100% accuracy in the practice test before entering the test phase. Breaks were encouraged between blocks to maintain concentration and prevent fatigue.

#### 2.4. Statistical analyses

A repeated-measures ANOVA was performed in R (version 4.0.2), using function *ezANOVA* from the *ez* package (version 4.4-0), with the percentage correct data as the dependent variable and auditory condition (noisy/noiseless), visual condition (no visual information/degraded visual information/clean visual information), and their interaction as fixed-effects variables. The Greenhouse-Geisser correction was performed in cases of a violation of the sphericity assumption with effect sizes reported as generalized eta-squared (*ges*). Significant main effects were followed up by post-hoc tests with Bonferroni correction to test which specific conditions were significantly different from each other.

### 3. Results

The recognition performance (mean accuracy) is illustrated for the various conditions in Figure 2. Participants achieved the highest mean accuracy in the AV condition (Mean  $\pm$  SD = 96.95%  $\pm$  3.73%), followed successively by the AVn condition (Mean  $\pm$  SD = 95.86%  $\pm$  3.01%), the AnV condition (Mean  $\pm$  SD = 95.12%  $\pm$  3.47%), the AnVn condition (Mean  $\pm$  SD = 93.40%  $\pm$  5.27%), the AO condition (Mean  $\pm$  SD = 88.87%  $\pm$  5.86%), and the AnO condition (Mean  $\pm$  SD = 54.06%  $\pm$  9.82%).

As displayed in Figure 2, overall, participants performed better in the noiseless settings than in the noisy ones and their performances were also better in the multisensory conditions compared with those in the unisensory ones. This was corroborated by a two-way repeated-measures analysis of variance (ANOVA) (2 Auditory Conditions  $\times$  3 Visual Conditions) that revealed a significant main effect of the auditory condition ( $F(1, 39) = 756.77, p < 0.001, ges = 0.57$ ) and a significant main effect of the visual condition on the recognition accuracy ( $F(1.23, 47.97) = 563.08, p < 0.001, ges = 0.80$ ). The ANOVA also detected a significant interaction effect between the auditory condition and the visual condition ( $F(1.45, 56.60) = 279.90, p < 0.001, ges = 0.65$ ). Mauchly's test indicated that the assumption of sphericity had been violated for the main effects of visual condition ( $W = 0.37, p < 0.001, \epsilon = 0.80$ ), and the auditory condition  $\times$  visual condition interaction ( $W = 0.62, p < 0.001, \epsilon = 0.65$ ). Therefore, degrees

of freedom were corrected using Greenhouse-Geisser estimates of sphericity, and the significant effects remained. The interaction effect indicates that the simultaneous presentation of visual information had different effects on the emotion recognition accuracy, depending on whether the speech token was embedded in babble noise. As demonstrated in Figure 2, the enhancement brought by visual aids was more evident in the noisy conditions than the noiseless conditions.

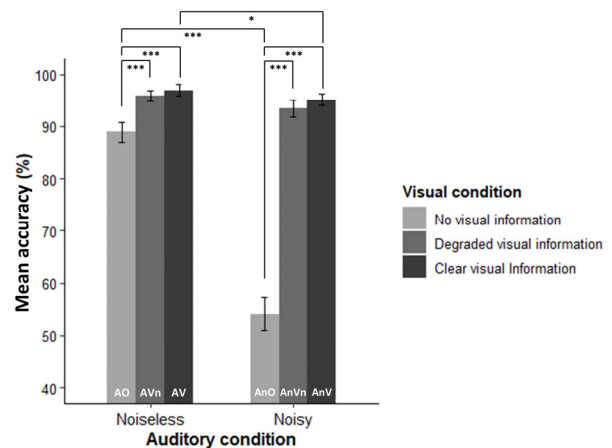


Figure 2: Emotion recognition accuracy for the six conditions. (\*:  $p < 0.05$ ; \*\*\*:  $p < 0.001$ )

Bonferroni post hoc tests revealed significant differences between AO and AnO ( $p < 0.001$ ), AO and AV ( $p < 0.001$ ), AO and AVn ( $p < 0.001$ ), AnO and AnV ( $p < 0.001$ ), AnO and AnVn ( $p < 0.001$ ), and AV and AnV ( $p = 0.02$ ). Between the unisensory tasks, the recognition accuracy of the AnO condition was significantly lower than the AO condition. Similarly, for the multisensory AV condition, adding babble noise also reduced the identification performance, demonstrated by the significantly lower accuracy in the AnV condition. When visual information was added into the two unisensory conditions separately, the performances all enhanced and resulted in higher accuracy for the AnV condition and the AnVn condition compared with the AnO condition as well as higher accuracy for the AV condition and the AVn condition compared with the AO condition.

### 4. Discussion

This study examined the effects of multi-talker babble noise on emotion recognition performances by native Mandarin Chinese speakers and how adding visual information affects the outcomes. In line with our first hypothesis, the results showed that in both noiseless and noisy environments, adding visual information was able to improve the recognition accuracy, which was still true when the visual aid was less reliable, and the facilitating effect was more pronounced in the noisy condition compared to the quiet one. As predicted in our second hypothesis, the study also found that multi-talker babble noise significantly undermined the emotion recognition performances, even when visual information was presented simultaneously.

Multi-talker babble noise has a corrupting effect on emotion recognition, which works in both unisensory and multisensory settings. When only auditory information was

presented, the recognition accuracy in the noisy condition decreased substantially compared with the noiseless condition. This indicates that babble noise can severely degrade the sound channel and limit the perception of emotions in affective prosody. Similar detrimental effects brought by babble noise have been observed in speech intelligibility [19] and lexical tone recognition [20]. Taken together, the results of our study and the previous studies demonstrate that speech-like signals are prone to the impact of multi-talker babble noise, whether the target information to be decoded from the signals is semantic meaning, phonological features, or paralinguistic information. This aligns with former investigations in the realm of human-computer interactions, which recognize the particular challenges that babble noise can pose to speech-related systems [21] and suppose that the difficulties may be explained by the similar spectro-temporal modulation between speech signals and babble noise [22].

In the multisensory audiovisual setting, the deleterious effect induced by babble noise persisted, but the decrease was much less severe than in the unisensory setting. This indicates that when the channel that conveys emotional signals (in this study, the prosodic channel) is corrupted, adding a stream of emotional information from another channel (e.g., visual) can partially compensate for the deterioration. Moreover, we found that even when the visual cues were degraded and became less reliable, the compensatory effect still existed. This is important because in real-world applications (e.g., clinical intervention or human-computer interaction), visual cues can be limited due to sensory impairments or loss of image quality, and our results, consistent with recent clinical studies [23] and engineering practices [24], highlight, with due consideration of ecological validity, the robustness of multisensory integration of emotion to visual unreliability.

The multisensory benefits are evident in both ideal (noiseless) and challenging (noisy) auditory environments, as simultaneously presenting multimodal emotional signals significantly improved recognition performances, whether there was background noise or not. However, the enhancement is greater in the challenging environment than in the ideal condition. That is, the multisensory gain *increased* as responses to the *decrease* in the unisensory stimuli. Our confirmation of this characteristic of multisensory advantages provides a coherent picture with the conclusions of previous studies on emotion processing that multisensory emotion perception follows the “inverse effectiveness” principle, i.e., the degree of multisensory benefits is in inverse proportion to the effectiveness of the stimuli [6, 25]. This implies that in the real world where auditory noise is inevitable, a multisensory approach would be especially helpful for training or rehabilitation purposes in emotion recognition.

The current study has a limitation that might restrict the interpretation of the findings, which highlights a promising future research direction. In our study, we did not equate the SNR levels when adding noise to the auditory and the visual channels because, in degraded conditions, the SNRs of the auditory and the visual channels were unable to be the same, if both of the two channels needed to be partially masked without either ceiling or floor effect. In a pilot test, we found that when the facial expression was partially masked, the corresponding emotional prosody embedded in noise with the same SNR level could seldom be distinguished, and when the emotional prosody was partially masked, recognition of the equivalently masked facial expression showed an unavoidable ceiling effect. We considered two possible reasons for this

phenomenon. One possibility is that in multimodal emotion presentation, the facial channel generally conveys more salient emotional information than the prosodic channel [26] (see the validation results of the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) that confirmed a face-bias in emotion recognition [27]). The other possibility is that emotional prosody might generally be more susceptible to noise degradation than emotional facial expressions, as discovered by research in affective computing [28]. Whichever possibility turns out to be correct, it is beyond the essential questions that this preliminary study seeks to answer. Therefore, in the present study, we did not heedlessly equate the SNRs of the degraded auditory stimuli and the degraded visual stimuli, though it indeed limited our interpretation of the results within the scope of visual compensation for challenging listening conditions. As for the question of which modality is more dominant and more relied on in noise-disturbed multisensory emotion processing, the investigation into which might require the equivalence of auditory and visual SNRs, it is beyond the scope of this study and remains to be examined. Future efforts are thus highly recommended to explore ways of equalizing the SNRs of degraded auditory and visual channels without sacrificing the naturalness of the stimuli so that further comparisons and discussion can be facilitated on the complementation and competition of the two modalities in emotion recognition in adverse perceptual environments.

## 5. Conclusions

The current study investigated how emotion identification performances by native Mandarin participants are influenced by background multi-talker babble noise and the presence of synchronous visual information. Results revealed that multi-talker babble noise significantly contaminated auditory emotional information and that the damaging effect could be partially counteracted by adding dynamic facial expressions, which still held when the visual aid was less reliable. The facilitation brought by synchronous visual information worked in both ideal and challenging listening environments and was more pronounced in the latter condition. To conclude, multi-talker babble noise can significantly decrease emotion identification performances, which becomes less severe with the presence of visual aid. The multisensory contribution is robust to noise disturbance and is more prominent in adverse listening conditions. These findings provide ecologically valid evidence for multisensory benefits in emotion perception and enrich current knowledge in this research field by involving data obtained from tonal language speakers. The use of acoustic cues for expressing emotional information in tonal languages such as Mandarin Chinese are different from those in non-tonal languages [29] and different language and cultural backgrounds have been found to influence the ways people perceive emotions [30-32]. The results of our study may inspire future comparisons that explore linguistic and cultural factors in multisensory emotion processing in challenging conditions.

## 6. Acknowledgements

This research was funded by the major Programs of National Social Science Foundation of China (No. 18ZDA293, No. 13&ZD189).

## 7. References

- [1] A. H. Fischer and A. S. R. Manstead, "Social functions of emotion," in *Handbook of Emotions (3rd ed)*, M. Lewis, J. Haviland-Jones, and L. F. Barrett Eds. New York: Guilford Press, 2008.
- [2] B. De Gelder and J. Vroomen, "The perception of emotions by ear and by eye," *Cognition and Emotion*, vol. 14, no. 3, pp. 289-311, 2000.
- [3] M. Klasen, Y.-H. Chen, and K. Mathiak, "Multisensory emotions: Perception, combination and underlying neural processes," *Reviews in the Neurosciences*, vol. 23, no. 4, pp. 381-392, 2012.
- [4] D. W. Massaro and P. B. Egan, "Perceiving affect from the voice and the face," *Psychonomic Bulletin & Review*, vol. 3, no. 2, pp. 215-221, 1996.
- [5] J. Vroomen, J. Driver, and B. D. Gelder, "Is cross-modal integration of emotional expressions independent of attentional resources?," *Cognitive, Affective, & Behavioral Neuroscience*, vol. 1, no. 4, pp. 382-387, 2001.
- [6] O. Collignon *et al.*, "Audio-visual integration of emotion expression," *Brain Research*, vol. 1242, pp. 126-135, 2008.
- [7] B. Kreifelts, T. Ethofer, W. Grodd, M. Erb, and D. Wildgruber, "Audiovisual integration of emotional signals in voice and face: An event-related fMRI study," *Neuroimage*, vol. 37, no. 4, pp. 1445-1456, 2007.
- [8] L. Lambrecht, B. Kreifelts, and D. Wildgruber, "Gender differences in emotion recognition: Impact of sensory modality and emotional category," *Cognition and Emotion*, vol. 28, no. 3, pp. 452-469, 2014.
- [9] M. A. Schmuckler, "What is ecological validity? A dimensional analysis," *Infancy*, vol. 2, no. 4, pp. 419-436, 2001.
- [10] Y. Ghanbari and M. R. Karami-Mollaei, "A new approach for speech enhancement based on the adaptive thresholding of the wavelet packets," *Speech Communication*, vol. 48, no. 8, pp. 927-940, 2006.
- [11] H. Sheikhzadeh and H. R. Abutalebi, "An improved wavelet-based speech enhancement system," presented at the 7th European Conference on Speech Communication and Technology, Aalborg, Denmark, Sept. 3-7, 2001.
- [12] M. J. de Boer, T. Jürgens, F. W. Cornelissen, and D. Başkent, "Degraded visual and auditory input individually impair audiovisual emotion recognition from speech-like stimuli, but no evidence for an exacerbated effect from combined degradation," *Vision Research*, vol. 180, pp. 51-62, 2021.
- [13] T. K. Koerner and Y. Zhang, "Differential effects of hearing impairment and age on electrophysiological and behavioral measures of speech in noise," *Hearing Research*, vol. 370, pp. 130-142, 2018.
- [14] F. Ameka, "Interjections: The universal yet neglected part of speech," *Journal of Pragmatics*, vol. 18, no. 2, pp. 101-118, 1992.
- [15] M. E. M. Meinard, "Distinguishing onomatopoeias from interjections," *Journal of Pragmatics*, vol. 76, pp. 150-168, 2015.
- [16] F. Chen, Y. Hu, and M. Yuan, "Evaluation of noise reduction methods for sentence recognition by Mandarin-speaking cochlear implant listeners," *Ear and Hearing*, vol. 36, no. 1, 2015.
- [17] V. Viswanathan, B. G. Shinn-Cunningham, and M. G. Heinz, "Temporal fine structure influences voicing confusions for consonant identification in multi-talker babble," *The Journal of the Acoustical Society of America*, vol. 150, no. 4, pp. 2664-2676, 2021.
- [18] J. Pilarczyk and M. Kuniecki, "Emotional content of an image attracts attention more than visually salient features in various signal-to-noise ratio conditions," *Journal of Vision*, vol. 14, no. 12, pp. 1-19, 2014.
- [19] J. L. Hall and J. L. Flanagan, "Intelligibility and listener preference of telephone speech in the presence of babble noise," *The Journal of the Acoustical Society of America*, vol. 127, no. 1, pp. 280-285, 2010.
- [20] J. Shao, C. Zhang, G. Peng, Y. Yang, and W. Wang, "Effect of noise on lexical tone perception in Cantonese-speaking amusics," in *Interspeech 2016*, 2016, pp. 272-276.
- [21] S. Karimi and M. H. Sedaaghi, "Robust emotional speech classification in the presence of babble noise," *International Journal of Speech Technology*, vol. 16, no. 2, pp. 215-227, 2013.
- [22] T.-S. Chi, L.-Y. Yeh, and C.-C. Hsu, "Robust emotion recognition by spectro-temporal modulation statistic features," *Journal of Ambient Intelligence and Humanized Computing*, vol. 3, no. 1, pp. 47-60, 2012.
- [23] R. J. Hirst *et al.*, "The effect of eye disease, cataract surgery and hearing aid use on multisensory integration in ageing," *Cortex*, vol. 133, pp. 161-176, 2020.
- [24] E. M. Benssassi and J. Ye, "Investigating multisensory integration in emotion recognition through bio-inspired computational models," *IEEE Transactions on Affective Computing*, pp. 1-13, 2021.
- [25] B. E. Stein and M. A. Meredith, *The merging of the senses*. Cambridge (MA): The MIT Press, 1993.
- [26] U. Hess, A. Kappas, and K. R. Scherer, "Multichannel communication of emotion: Synthetic signal production," in *Facets of Emotion: Recent Research*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc., 1988, pp. 161-182.
- [27] S. R. Livingstone and F. A. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVD ESS): A dynamic, multimodal set of facial and vocal expressions in North American English," *PLoS ONE*, vol. 13, no. 5, pp. 1-35, 2018.
- [28] A. Shukla, S. Petridis, and M. Pantic, "Does visual self-supervision improve learning of speech representations for emotion recognition," *IEEE Transactions on Affective Computing*, pp. 1-15, 2021.
- [29] E. D. Ross, J. A. Edmondson, and G. B. Seibert, "The effect of affect on various acoustic measures of prosody in tone and non-tone languages: A comparison based on computer analysis of voice," *Journal of Phonetics*, vol. 14, no. 2, pp. 283-302, 1986.
- [30] A. Tanaka, A. Koizumi, H. Imai, S. Hiramatsu, E. Hiramoto, and B. de Gelder, "I feel your voice: Cultural differences in the multisensory perception of emotion," *Psychological Science*, vol. 21, no. 9, pp. 1259-1262, 2010.
- [31] M. Kawahara, D. A. Sauter, and A. Tanaka, "Culture shapes emotion perception from faces and voices: changes over development," *Cognition and Emotion*, vol. 35, no. 6, pp. 1175-1186, 2021.
- [32] P. Chen, A. Chung-Fat-Yim, and V. Marian, "Cultural experience influences multisensory emotion perception in bilinguals," *Languages*, vol. 7, no. 1, p. 12, 2022.