Behavior Research Methods
https://doi.org/10.3758/s13428-023-02270-7

ORIGINAL MANUSCRIPT

A Cantonese Audio-Visual Emotional Speech (CAVES) dataset

Chee Seng Chong[1] · Chris Davis[1] · Jeesun Kim[1]

Accepted: 11 October 2023

Abstract

We present a Cantonese emotional speech dataset that is suitable for use in research investigating the auditory and visual expression of emotion in tonal languages. This unique dataset consists of auditory and visual recordings of ten native speakers of Cantonese uttering 50 sentences each in the six basic emotions plus neutral (angry, happy, sad, surprise, fear, and disgust). The visual recordings have a full HD resolution of 1920 × 1080 pixels and were recorded at 50 fps. The important features of the dataset are outlined along with the factors considered when compiling the dataset. A validation study of the recorded emotion expressions was conducted in which 15 native Cantonese perceivers completed a forced-choice emotion identification task. The variability of the speakers and the sentences was examined by testing the degree of concordance between the intended and the perceived emotion. We compared these results with those of other emotion perception and evaluation studies that have tested spoken emotions in languages other than Cantonese. The dataset is freely available for research purposes.

Keywords Cantonese dataset · Auditory and visual expressions · Emotional speech · Dataset evaluation

The study of emotional expression, both production and perception, is important for research areas interested in communication, knowledge representation, and culture. For example, understanding emotion expression and perception can provide a solid basis of research on human–human and human–machine interaction, how linguistic and paralinguistic information are simultaneously expressed, and whether and how social norms impact the production and perception of emotion. As such, progress in these research areas is underwritten by the development and availability of appropriate materials (corpora or datasets). The current work presents our work on a dataset of Cantonese audio-visual emotional speech (CAVES); below we present the background and aims of this endeavor.

A single dataset cannot serve the diverse interests of researchers who aim to study the expression of emotion. That is, the choices made in constructing a particular dataset predispose it to certain types of investigation. Consider how past studies have differed in what has been investigated and how these investigations were carried out. For example, many early studies of human emotion recognition were designed to maximize the control and standardization of experimental stimuli, and so often employed static facial expressions of emotions as conveyed by photographs (e.g., Ebner et al., 2010; Langner et al., 2010). This focus created a need for a standard set of expressions that would allow results to be compared across studies. Thus, rather than a dataset per se, standard sets of emotional face expressions were developed (e.g., the Pictures of Facial Affect, Ekman & Friesen, 1976).

In contrast, recent work has sought to compile rich emotion datasets that can consist of a very large number of multimodal (visual and auditory) instances depicting a broad range of emotions (e.g., Vidal et al., 2020). Typically, such datasets consist of extracts of talk shows, movies, interviews, real-life drama, etc., that have been selected based on various criteria which are then rated in terms of which emotion

The study was supported by a MARCS Institute PhD scholarship to the first author and by ARC for the other authors (DP130104447). The authors would like to thank the House Ear Institute for the use of the Cantonese Hearing in Noise Test sentences list and Dr. Vincent Aubanel for the MATLAB scripts used for the segmentation of the recordings. We have no conflicts of interest to disclose.

✉ Jeesun Kim
j.kim@westernsydney.edu.au

Chee Seng Chong
chongcheeseng138@gmail.com

Chris Davis
chris.davis@westernsydney.edu.au

[1] The MARCS Institute for Brain, Behaviour and Development, Western Sydney University, Locked Bag 1797, Penrith, NSW 2751, Australia

Published online: 28 November 2023

was expressed, or by a rating-by-comparison method (where two clips are presented and the one judged more positive selected, see Baveye et al., 2013). These stimuli better reflect real-life social experience; however, they can present numerous challenges with respect to measurement and comparison due to the large number of different talkers, differences in how the stimuli were collected (in recording quality, noise levels, etc.), differences in content, the variability of emotional displays, poses, framing and so on. As such, these datasets are better suited to machine learning applications rather than investigations that require controlled contrasts of specific variables.

In designing the current dataset, we adopted an approach intermediate between the two outlined above. That is, we made high-quality, consistent recordings of stimuli designed to allow a high degree of experimental control; but importantly, these stimuli captured the multimodal, i.e., auditory-visual dynamics of spoken emotional expressions. To get an idea of where the current dataset fits with others, Table 1 provides an overview of some selected auditory-visual speech emotion datasets. The table lists these datasets, the language of production, the number of speakers that contributed, the number of utterances recorded, the emotions expressed, how emotions were elicited and what access is available. As can be seen most of these datasets consist of spoken English, the number of speakers and utterances differ markedly, a core set of emotions were expressed, mostly acted, and typically the datasets are not publicly available.

The construction of the current CAVES dataset had several motivations, which can be summed up in terms of the dataset representativeness and research application. With respect to representativeness, the main aim of this work was to establish an emotional speech dataset for Cantonese. Cantonese is a major world language, and as far as we can tell no AV speech emotion datasets have been compiled or are available. Moreover, Cantonese is a tonal language (with two more phonetic tones than Mandarin), and somewhere between 40 and 60% of the world's languages are classified as tonal languages (Maddieson, 2013; Yip, 2002), yet AV spoken emotion datasets almost exclusively use non-tonal languages (see Table 1).

In terms of research application, an important aim for CAVES was to provide a resource to examine a range of issues concerning how linguistic and emotional information are simultaneously expressed. More specifically, since spoken emotions and lexical tone is expressed via similar acoustic properties, e.g., fundamental frequency (F0) plays a key role in the production and perception of lexical tone and emotional prosody, the issue of how spoken tones affect the expression of emotion is a particularly interesting one. Studies (e.g., Anolli et al., 2008; Chong et al., 2015; Wang & Lee, 2015) have examined how variation in F0 is utilized to concurrently express linguistic and emotion information, but these studies were based on limited samples and emotions. A more extensive dataset would allow for a more comprehensive examination of the trade-offs between linguistic and emotional prosody for both auditory and visual expressions, e.g., the influence of dynamic versus static tones, the role of segmental and suprasegmental timing, whether the difference in F0 between female and male speech plays a role, and so on.

CAVES versus other tonal language AV emotional speech datasets

As can be seen in Table 1, several other tonal language spoken emotion AV datasets have been developed. These, however, were established for a range of purposes different from the linguistic and psycholinguistic research application outlined above. In what follows, we identify some issues with these datasets; not intrinsic limitations per se, but rather limitations with respect to the current interest. To make these issues concrete, we focus on dataset scope and context (see Douglas-Cowie et al., 2003) and, from a practical viewpoint, on dataset availability.

The first dataset is the Multimedia Human–Machine Communication (MHMC) dataset (Lin et al., 2012). The language is most likely Taiwanese Mandarin although the language was not explicitly indicated. As far as we could determine, this dataset is not publicly available. The dataset used posed emotions produced by seven actors (both females and males, but no details of how many of each). Each actor said 30 sentences in four emotions (happiness, sadness, anger, neutral) and each utterance was repeated, thus there were 1680 utterances in total (2 × 30 × 4 × 30). The design context of the MHMC was that it was developed for automatic recognition of human emotions from audio-visual signals. The recordings were made in an office environment that had a reasonable level of foreground and background noise and video images were captured using a Logitech QuickCam at a low resolution of 320 × 240 pixels with a frame rate of 30 frames per second. The main issues with this dataset for current purposes are its lack of availability; the limited number of emotions captured, and the low quality of the AV recordings.

The second dataset is the National Tsing Hua University-National Taiwan University of Arts Chinese Interactive Emotion (NNIME) dataset (Chou et al., 2017). This dataset is not available online; and researchers must contact the first author to gain access. The design context was to capture dyadic human–human communication. This involved simultaneous recording of audio, video, and ECG signals collected during spoken interactions between 22 pairs of actors performing a spontaneous 3-min dialogue. The language is Taiwanese Mandarin. The emotions targeted for

Table 1 A comparison of key features of auditory-visual speech emotion datasets (2006–2021)

Dataset	Language	Speakers	Utterance	Emotion	Type	Access
CAVES	Cantonese	10	10425	A,D,F,H,N,S,Su	Acted	Yes
MHMC[1]	Mandarin	7	1680	A,H,N,S	Acted	No
NNIME[2]	Mandarin	44 (24 F)	3 min speech	A,D*,F*,H,N,S,Su	Acted	Contact
CHEAVD 1.0[3]	Mandarin	238 (113 F)	2322	A,D,H,N,S	Acted*	Not found
CHEAVD 2.0[4]	Mandarin	527 (219 F)	7030	A,A*D,H,N,O,S,W	Acted*	Not found
TUM AVIC[5]	English	21 (10 F)	3901	Nonverbal	Interaction	No
AFEW[6]	English	330	1426	A,D,F,H,N,S,Su	Natural	Contact
Fiction[7]	English	28 (12 F)	152 turns	F, negative, N,O	Acted*	No
RAVDESS[8]	English	24 (12 F)	7356 clips	A,C,D,F,H,N,S,Su	Acted	Yes
SEMAINE[9]	English	150	959	A,C*,D,F,H,S	Interaction	Request
SAVEE[10]	English	4 (M)	480	A,D,F,H,N,S,Su	Acted	No
eNTERFACE[11]	English	42 (8 F)	1186	A,D,F,H,N,S,Su	Acted	No
IEMOCAP[12]	English	10	3060	A,D,F,H,N,S,Su,O	Actors	Form
MELD[13]	English	6 (84%)	13000	A,D,F,J,N,S,Su	Acted*	Yes
SAFE[14]	English (80%)	400	4073	F, negative, N, positive	Acted	No
Kim & Davis[15]	English	5	10	A,D,F,H,N,S,Su	Acted	No
RECOLA[16]	French	46 (27 F)	4 min speech	5 Social Affects	Natural	Link down
Vera[17]	German	47	947	Valance, active, dom	Acted*	Link down
JAVED[18]	Japanese	14 (4 M)	100 min.	A, Content, H,N,S	Acted	No
HEU (part2)[19]	Multilingual	967	2435	A,B,D,D*,F,H,N,S,Su	Acted*	No
MDESVG[20]	Polish	16 (8 F)	560	A,D,F,H,N,S,Su	Acted	No
RAMAS[21]	Russian	10	581	A,D,H,S,S*,Su	Acted	Link down
BAUM-1[22]	Turkish	31 (13 F)	1222	A,D,F,H,S,Su,O	Interaction	Link down

Acted* = TV, movie, talk-show; A = anger; A* = anxious; B = bored; C = calm; C* = contempt; D = disgust; D* = disappointed; F = fear; F* = frustration; H = happy; J = joy; N = neural; S = sad; S* = scared; Su = surprise; O = other; W = worry

[1]Lin et al. (2012). [2]Chou et al. (2017). [3]Li et al. (2017). [4]Li et al. (2018). [5]Schuller et al. (2009). [6]Dhall et al. (2012). [7]Clavel et al. (2004). [8]Livingstone et al. (2018). [9]McKeown et al. (2011). [10]Jackson & Haq (2015). [11]Martin et al. (2006). [12]Busso et al. (2008). [13]Poria et al. (2018). [14]Clavel et al. (2006). [15]Kim & Davis (2012). [16]Ringeval et al. (2013). [17]Grimm et al. (2008). [18]Lubis et al. (2016). [19]Chen et al. (2021). [20]Sapiński et al. (2018). [21]Perepelkina et al. (2018). [22]Zhalehpour et al. (2016).

each performance were anger, sadness, happiness, frustration, surprise, and neutral. Emotion annotation was carried out by 49 annotators (the majority of which were peer reports from other actors). In all, 102 recording sessions were rated using a five-point scale on dimensions activation and valence. Video recordings were made via a fixed position camera that was positioned at a considerable distance from the actors who were free to move as within the field view of the camera. The main issues with this dataset are its lack of immediate availability; that the emotions captured do not cover the basic six emotions and were expressed using different utterances; and the low AV recording quality of facial expression due to the fixed distant camera, freely moving actors and profile views.

There are two other datasets, the CASIA Natural Emotional Audio–Visual Datasets (CHEAVD, Li et al., 2017) and CHEAVD 2.0 (Li et al., 2018) that appear to be available online for a fee (from the Chinese Linguistic Data Consortium), although how to arrange this was not apparent. The design context of these datasets was that they gather many speakers to enable speaker-independent emotion analysis. As such, the datasets were collected from media excerpts, e.g., for CHEAVD from 34 films, two television series, two television shows, one impromptu speech and one talk show (in Mandarin Chinese). Segments were selected based on clips that did not have high levels of background music or vocal overlap; the segment had only a single talker's speech and face and contained a complete utterance. CHEAVD had 238 speakers (47.5% female) and CHEAVD 2.0 527 speakers (41.6% female). Segments were rated by four annotators that resulted in 26 emotional labels being used (neutral, angry, happy, sad, worried, anxious, disgust, surprise, blamed, sarcastic, aggrieved, curious, fearful, embarrassing, nervous, confused, proud, helpless, hesitant, contemptuous, frustrated, serious, anticipated, shy and guilty). The distribution of emotion labels was highly skewed (with some emotions having many and others with relatively few instances). It is unclear how the movie and television video clips were captured; however, videos were saved in 640 × 480 pixels with a frame rate of 25 fps, audio was saved at 44.1 kHz, stereo and 16 bit. The main issues with these two datasets are that they are not free for researchers; the emotions captured are unbalanced (i.e., few examples for some emotions); the emotions are expressed using different utterances and talkers; and the extent of background noise (both auditory and visual) was not specified.

CAVES contrasts with the above datasets. First, while there are some emotional speech datasets available for tonal languages, none exist for Cantonese. This is despite Cantonese being a language that is widely spoken and has a rich set of six phonetic tones. Also, compared to these datasets as well as most other ones in Table 1, the CAVES dataset is freely available for research purposes at https://forms.office.com/r/3VfeWQnAVa. For this style of dataset, the CAVES has a reasonably large number of talkers ($N = 10$), that had equal representation of female and male speakers. In addition, it has a large number of items ($N = 50$) that has good coverage of different tones at sentence initial and final positions (see below), and each sentence is expressed in each of the basic six emotions plus a neutral expression. Note that the six basic emotions were selected as they are considered to be universally expressed and recognized across cultures and are well studied and represented within the literature on cross-cultural research (Ekman, 1992). In all, then, there are 3500 AV stimuli and the same number of visual only and auditory only stimuli.

The context of the CAVES dataset design was for it to facilitate investigations of how linguistic and affective expressions interact, that is, how emotion and speech modulate facial expressions. To do this meant having controlled recording of talkers expressing the same lexical content in different emotions as this allows for differences in the expressive behaviors associated with each emotion to be readily contrasted. Having the same lexical content also controls for possible tone context effects and the influence of syntactic form on emotional expression (Cole, 2015). Moreover, having controlled talker recordings of predetermined content allowed a neutral semantic content to be used for all the emotions; this alleviates problems of perceivers being influenced by lexical content.

Below we present in more detail, a description of the design and development of the CAVES dataset (Part 1); and then report on a perceptual validation study (Part II). Note that this study was conducted in line with the principles of the Declaration of Helsinki. Approval was granted by the Ethics Committee of Western Sydney University (H10442).

Part I: Dataset design and development

As mentioned above, the CAVES dataset contains six basic emotions (anger, disgust, fear, happy, sad, surprise) plus a neutral expression to serve as a baseline. We selected a set of semantically neutral carrier sentences so that all the six basic emotions could be expressed for each sentence without any semantic interference. Having a set of common carrier sentences allows the different emotions (and neutral baseline) to be straightforwardly compared. We also selected sentences that had a good coverage of the different lexical tones both in initial and final sentence positions. This selection allows for an examination of how the onset and offsets of emotion intonation change as a function of the different tones. It is not clear, for example, whether Cantonese speakers utilize different acoustic cues in expressing anger (associated with a rising tone) for sentences that have an onset/offset falling versus rising tones.

Methods

Participants

Ten native speakers of Cantonese (five females) who were born and raised in Hong Kong were invited to participate for monetary reimbursement. The average age of the participants was 29.1 years (SD = 4.9). All the participants also spoke English.

Materials

Fifty sentences each containing ten syllables/characters were used. These sentences were selected from the 240 Cantonese Hearing In Noise Test (CHINT) sentence list[1] (Wong & Soli, 2005). The selection was made on the basis that the sentences had a good distribution of different tones at initial and final positions. In selecting sentences in terms of the spread of the tones, a six-tone system was used (see Table 2).

Table 3 shows the number of sentences with each tone in the initial and final positions for the 240 sentences in the CHINT and the 50 sentences that were selected to have a good balance of tones in the initial and final positions.

The CHINT sentences were developed to be used for hearing in noise tests, as such they included a list of

[1] The CHINT was used under licence from the University of Hong Kong; and House, Ear Institute, which restricts the full listing of the sentences used. In the CHINT, sentence-final particles (SFP) were avoided. The list of 50 sentences (no SFP) is available from the first author on reasonable request.

Table 2 Example of the six tone classification system using the homophone /fan/

Tone Number	Tone	Description	/fan/
1	˥	high level	分
2	˧˥	mid rising	粉
3	˧	mid level	訓
4	˨˩	low falling	墳
5	˩˧	low rising	忿
6	˨	low level	份

Here /fan/ is expressed in the six different tones and the English translation of these Chinese words from tone 1 to 6 are: “point”, “noodles”, “discipline”, “grave”, “angry” and “portion”

Table 3 Numbers of each tone in initial and final sentence position for the 50 items selected from the CHINT (the data for the full CHINT list are given for comparison)

Tones	240 original		50 selected	
	Initial	Final	Initial	Final
1	36	61	8	12
2	47	45	8	9
3	38	36	8	7
4	19	46	8	8
5	86	7	14	5
6	14	45	4	9

The numbers in each cell represent the number of sentences with the indicated tone in the initial and final position.

parenthesized words that could be substituted and considered as a correct answer in a speech identification in noise paradigm. For example, in this sentence: 教授(就快 or 就嚟)去美國做研究[2] both 就快 and 就嚟 have the same meaning, “soon”, so for the purposes of this dataset, we decided to use the second pair of characters 就嚟 to maximize the number of different tones within that sentence. To illustrate this, the original sentence represented in terms of tones would be: 3, 6, (6, 3 or 6, 4), 3, 5, 3, 6, 4, 3. So, to balance out the ratio of tone 3 to tone 4 characters we picked the second pair of characters so the sentence represented with these tones would be: 3, 6, 6, 4, 3,5, 3, 6, 4, 3. The same strategy was used for all the selected sentences to obtain emotional utterances that had a balance of all six tones. Including a balance of lexical tones will allow for the examination of how speech tones affect emotion expressions, e.g., how pitch changes in Cantonese especially on the different tones such as a high-level pitch contour in tone 2 or low/falling in tone 4 influence or are influenced by the expression of emotion.

Recording setup

The recording was conducted in a sound attenuated booth at Western Sydney University. In the booth, participants were seated in front of a 20.1” LCD video monitor (Diamond Digital DV201B) that was used to present the stimulus sentences to the participant. Directly above the monitor was a video camera (Sony NXCAM HXR-NX30p) where participants were requested to fixate prior to uttering the sentences. The videos were recorded at 1920 × 1080 full HD resolution at 50 fps. To capture participants’ utterances a microphone (AT 4033a Transformerless Capacitor Studio Microphone) was placed about 20 cm away from the participants’ lips and out of the field of view of the camera (see Fig. 1 for a depiction of the setup). Audio captured using the microphone was fed into the Motu Ultralite mk3 audio interface with FireWire connection to a PC running CueMix FX digital mixer and then to Audacity which captured the sound at a sampling rate of 48 kHz. The camera, screen, and microphone heights were adjusted to suit each participant. The audio and video outputs were monitored by the experimenter who was seated outside of the booth.

All participants did a short trial session of three utterances to determine the best gain values for the microphones. Once an acceptable level had been achieved, the recording commenced.

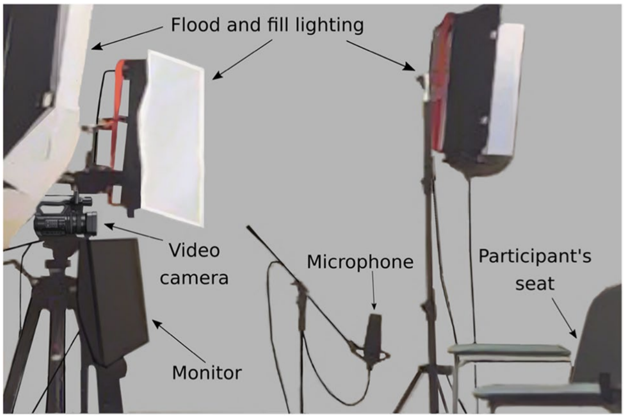

Fig. 1 The setup in the recording booth showing the camera, screen microphone, lighting and participants’ seat

[2] In English, this sentence reads “The professor will soon be travelling to America to conduct research”.

Procedure: Elicitation of emotion

Participants were briefed regarding the recording procedure before signing a written consent form indicating that recordings could be used for research purposes (the instructions and emotion labels were conveyed in English). Participants were instructed to be as natural as possible in how they expressed themselves and were asked to produce the emotions with the intent of communicating their emotional feelings to an observer. That is, rather than focusing on the inner experience of an emotion, the current interest was in the expression of emotional expressions for the purpose of communication, i.e., the emphasis was on the non-verbal signals people use to communicate emotion. As such, no emotional induction procedure was used and although the emotions were acted, the participants did strive to express each emotion as if she/he was conveying emotional information to another person. Table 1 shows that this type of emotion elicitation method is the most commonly used (see also Scherer, 2003).

The recording session was blocked by emotion type and the order of presentation was randomized across participants. In each block, participants were first informed of the emotion to express, and were then instructed to imagine themselves expressing this emotion to an interlocutor. They were then given three practice trials prior to the start of each block. Stimuli sentences were then displayed one at a time in a random order and the participants produced the utterances when ready. Participants were given feedback via the screen if they were required to repeat a sentence (e.g., if they misread the sentence or did not fixate on the camera while producing the expressions). They were given a break after the successful production of every 25 sentences. By the end of the recording session, all participants produced a total of 350 utterances (50 sentences × 7 emotional expressions including neutral).

Data segmentation

The audio recordings were high-pass filtered (100 Hz) to remove noise and a noise-shaped dither was then applied. To reduce processing time, a down-sampled (16 kHz) copy of the audio recordings was used for segmentation purposes only. The Audio Segmentation Toolkit (Gravier et al., 2010) was used to automatically segment speech events from silence and the segmentations were exported as PRAAT textgrids (Boersma & Weenink, 2014). The segmentations were manually checked and corrected using MTRANS (Cooke et al., 2011). A buffer of 500 ms before and after the utterance was applied to capture articulatory and expressive gestures that may unfold before and after the utterance.

The corrected segmentations were used to extract individual sentence recordings from the uncompressed video (MTS, Advanced Video Coding High Definition format) and the audio recordings (48-kHz recordings). Audio-visual stimuli were then created by combining the video and audio recordings. The video files were cropped to a size of 1000 × 1000 pixels, framing the participants head within the center of the recording (see Fig. 2 for an example). We used a static cropping method which crops a predefined area (rather than a dynamic method which tracks and centers head movement). This method was used to capture the expressive head movements made by the participants to preserve rigid head motion which may carry emotion information (Davis & Kim, 2006; Kim et al., 2014).

All segmented video clips were kept in their original format (.MTS) to preserve the quality of the stimuli. These clips (10 speakers × 7 (6 basic emotions + neutral) × 50 sentences = 3500) were labeled by speaker ID, emotion type and then by sentence ID, except 25 sentences produced by one female speaker in the Sad condition that were lost due to technical issues.

Part II: Validation of the CAVES dataset

The recorded expressions from the CAVES dataset were evaluated using an emotion identification experiment. The primary aim of this experiment was to assess the validity of the produced spoken emotion in terms of the degree of concordance between the perceived emotion expression and the intended emotion expressed by the speaker. For this, we examined how emotion recognition accuracy rates and the distribution of response frequencies from each spoken

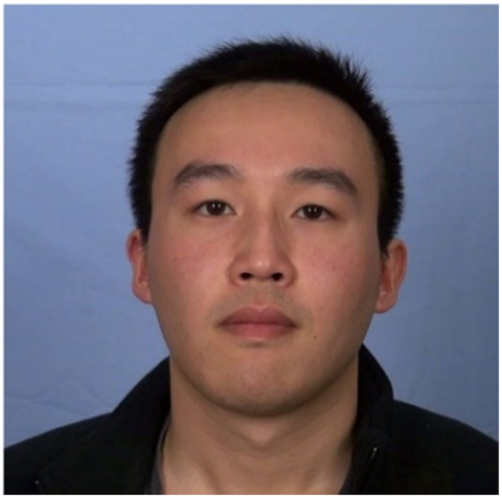

Fig. 2 A single frame extracted from video clip to illustrate the extent to which the video was cropped

expression varied as a function of emotion type and presentation condition. The results were also compared to the findings of other emotion perception and evaluation studies particularly using spoken emotions in languages other than Cantonese.

The secondary objective of this experiment was to determine the reliability of the dataset by examining how variable speakers and sentences were in terms of accuracy rates (for speakers and sentences). Results from this examination can be used as a reference for the selection of stimuli in future studies.

Methods

Participants

Fifteen participants (including ten females, mean age = 23.2 years, SD = 3.9) took part in this study for a small payment. All participants were native listeners of Cantonese who were born and raised in Hong Kong. The participants spoke both English and Mandarin. The majority of the participants were recruited through word of mouth and snowball sampling procedures. This sample size was selected based on it being a feasible sample size to run given how long each participant took to complete the experiment and the availability of Cantonese speakers. This sample size yielded effects sizes for each of the main contrasts (i.e., AV vs. VO; AV vs. AO and VO vs. AO) for each emotion that ranged from small to medium (Cohen, 1988) (see Table 3). Effect sizes were calculated based on the methodology published in Westfall et al. (2014).

Stimuli

All speech recordings (50 sentences × 7 (6 emotions and neutral) expressions × 10 speakers including five females) were used as stimuli, except 25 sentences by one female speaker in the Sad condition. These recordings were presented in three presentation conditions, auditory-visual (AV), visual-only (VO) and auditory-only (AO), resulting in a total of 10,425 stimulus items.

Design and procedure

Due to the large number of stimuli, each participant was tested over multiple sessions. Each session was conducted on a separate day and consisted of a total of 900 trials (50 sentences × 3 presentation conditions × 6 emotions, note: the neutral expression was used as a speaker-specific baseline, see below) from a random selection of either male or female only speakers.

Although we aimed to conduct a fully within-subjects design by having each participant rate all the speech recordings, not all participants were able to fully complete the validation study due to limits on availability (judging the large number of recordings in this dataset required a considerable time commitment). Nevertheless, participants were encouraged to participate in as many sessions as possible, up to the maximum of ten (where all recordings would be tested). As an eligibility criterion to participant in the experiment, all participants agreed to participate in at least five sessions of the study. In all, 51,408 judgments were obtained from participants. The number of these judgments were split evenly across emotion type (17% each), presentation type (AV = 31%; VO = 37%; AO = 32%) and female/male (38 and 62 %) and individual talker (8–13%).

The stimuli were presented using DMDX (Forster & Forster, 2003) on a 15.6-inch laptop (Lenovo T520) that is connected to an EDIROL UA-25ex soundcard with Sennheiser HD550 headsets. Participants were tested individually in sound-attenuated IAC booths at Western Sydney University.

Participants were given written instructions and a short practice session prior to the start of the experiment. In the practice session, participants were first presented with two video clips of the speaker uttering a sentence in a neutral expression. These neutral expressions were included to help familiarize the participants with the speaker and acted as a speaker specific baseline against which to judge the emotional expressions. The neutral utterances were followed by 12 practice trials. Each trial consisted of one emotion expression and participants were required to identify the emotion by responding to a six alternative forced choice task using the mouse. Note, these practice trials were presented later as experimental trials to be again rated by the participants. The researcher remained with the participant during the practice session to ensure that the participants understood the task.

The experimental trials were presented in the same format as the practice trials. The trials were blocked by presentation condition for each speaker. Participants were always given two sentences in a neutral expression at the beginning of each block. The presentation order of the blocks was counterbalanced across the speakers. With no time limit imposed, participants could proceed at a pace that they were comfortable with. Participants were given a 5-min break every 150 trials and reimbursed for their time at the end of each session.

Analysis

The first set of analyses was conducted on the participants' accuracy data in recognizing the different types of emotions across the three presentation conditions (the analysis was also conducted using unbiased Hu scores, Wagner, 1993).

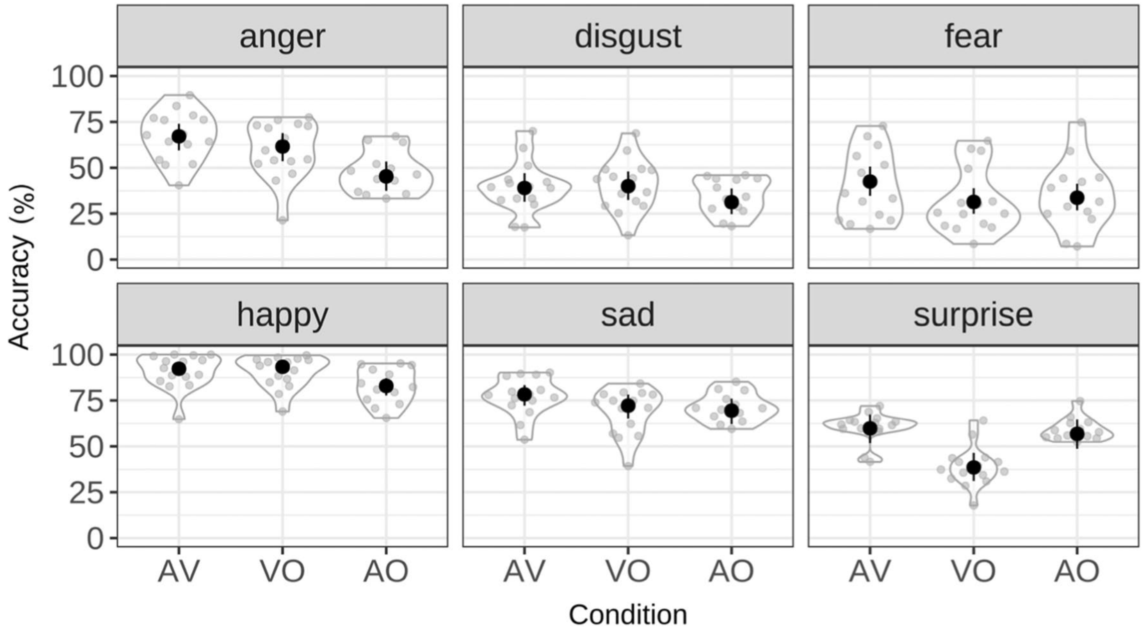

Fig. 3 Percent accuracy scores for all emotion types by presentation conditions (model-based standard error)

Using the findings of other studies as a benchmark, it was expected that (1) bimodal emotion expressions (AV) be recognized with higher accuracy than the unimodal expressions of VO and AO (see Kim & Davis, 2012); and (2) recognition accuracy would vary as a function of emotion type, e.g., expressions of Happy were expected to be recognized with the highest accuracy while expressions of Fear at the lowest accuracy (Ebner et al., 2010; Langner et al., 2010; Tanaka et al., 2015; also see Scherer, Banse & Wallbott, 2001 for auditory only expressions). The rest of the analyses consist of providing descriptive statistics, examining confusion matrices, and exploring speaker and item level differences in accuracy scores. The perception data is available at https://forms.office.com/r/3VfeWQnAVa.

Results and discussion

Accuracy

Figure 3 presents the recognition accuracy scores across the six emotion types for each presentation condition. A generalized linear mixed model fit by maximum likelihood (Laplace Approximation) [glmerMod], Family: binomial (logit) was fitted to the data to examine if recognition accuracy varied as a function of presentation condition, emotion type and the interaction between the two. The R afex package (Singmann et al., 2021) was used to build the model and the ggplot package (Wickham, 2016) and afex_plot package (Singmann et al., 2021) was used to generate all the graphs presented in this paper. Speaker, participants, and sentence were entered as random factors, emotion type and presentation condition as fixed factors, and recognition accuracy as the dependent variable (formula: mixed(Accuracy ~ Emotion type * Presentation condition + (1|Participant) + (1|Item) +(1|Speaker), data = CAVES_data, family = binomial, method = "LRT", all_fit = TRUE)). Note that attempting to generate maximal or near maximal models (e.g., add in random slopes to any of the random variables) led to failures to converge, thus we accepted a simpler model, rather than risk the problems associated with fitting overparameterized models (see Matuschek et al., 2017).

Both the main effects and the interaction between them were significant; emotion type, $\chi2(5) = 7714.38$, $p < .001$; Presentation condition, $\chi2(2) = 327.77$, $p < .001$; and the interaction between these variables, $\chi2(10) = 600.16$, $p < .001$. One-sample t tests with Bonferroni correction indicate that expressions of all emotion types were recognized at above chance accuracy (16.7% = 100% divided by six possible response options) across all presentation conditions.

As mentioned above, the accuracy scores were also converted to unbiased Hu scores (Wagner, 1993); since these are proportions these scores were arcsine transformed. The transformed Hu scores were then analyzed with a within measures repeated ANOVA (using the AFEX package, formula: Hu score (arcsine) ~ Emotion type*Presentation condition + Error(participant/Emotion type*Presentation condition). The outcome of the analysis agreed with that of the unadjusted recognitions scores, Emotion type $F(2.24, 26.92) = 55.28$, $p < .001$, Partial Eta squared = .822; Presentation condition, F(1.91, 22.96) = 16.97, p < .001, Partial Eta squared = .586; and the interaction between Emotion type and Presentation condition, F(4.39, 52.7) = 16.25, p < .001, Partial Eta squared = .575.

Further statistical significance testing was conducted using the emmeans package (Lenth, 2021) and the results are shown in Table 4. P values were adjusted using the Tukey method for comparing a family of three estimates. These tests were conduction on the recognition scores and the (arcsine transformed) Hu scores. As can be seen in the table, the pattern of the outcome of the analysis was the same for the simple recognition scores and the unbiased (transformed) Hu scores.

In general, the patterns in accuracy rates observed in this study were similar to those of other AV speech studies, e.g., Kim and Davis (2012) that examined spoken expressions of English presented in the three different presentation conditions. Accuracy in the AV condition was significantly higher than both VO and AO conditions for all emotion types except for Disgust and Happy; VO was as accurate as AV for these two emotion types. Comparing the VO to AO condition, accuracy was significantly higher in the VO condition for Anger, Disgust and Happy. This result is also closely aligned with the findings of Kim and Davis (2012) that Anger, Disgust and Surprise were recognized more accurately in the VO compared to AO condition. In the current results, Surprise was the only emotion type where accuracy in the AO condition was significantly higher than the VO condition.

Collapsing across presentation modalities, Happy was recognized at significantly higher accuracy rates than other emotion expression (p < .001). Disgust and Fear were recognized at significantly lower accuracy than all other expressions (p < .001). A similar pattern of emotion recognition was reported in the evaluation of the Faces, Radboud and Karolinska Directed Emotional Faces dataset (Ebner et al., 2010; Langner et al., 2010; Goeleven et al., 2008). The finding that spoken expressions of Fear were recognized with the lowest accuracy was similar to that observed in the Tanaka et al. (2015) study which examined spoken expressions of emotions produced by Japanese and Dutch speakers.

Confusion matrices

Tables 5–7 show the confusion matrices for the three presentation conditions. Expressions of Anger were either misidentified as Disgust (AV and AO) or Sad (VO). Expressions of Disgust were either misidentified as Anger (AO) or Sad (AV and VO). Confusion between Anger and Disgust is a common finding observed in evaluations of facial and spoken expressions (Kim & Davis, 2012; Tanaka et al., 2015). It was further observed that negative emotions such as Anger, Disgust and Fear were typically misidentified as Sad; a finding that aligns with previous evaluations of static facial expressions which found Sad to be the most frequently selected response (Goeleven et al., 2008).

Fear was misidentified as Sad across all presentation modalities. This was similar to the results reported by Tanaka et al. (2015) and by Banse and Scherer (2001). This is, however, in contrast to some studies that have reported that Surprise is the most likely alternative response (see Goeleven

Table 4 Effect size (d) and p values of the pairwise contrasts between presentation condition and emotion type for the simple recognition (*middle*) and bias-corrected Hu scores (*bottom row*)

Emotion	AV vs. VO	AV vs. AO	VO vs. AO

Anger	d = 0.05
	Z = 4.4, p < .001
	t = 6.54, p < .0001
		d = 0.20
		Z = 16.5, p < .001
		t = 11.28, p < .0001
			d = 0.15
			Z = 13.1, p < .001
			t = 4.60, p < .001

Disgust	d < 0.01
	Z = – 0.70, *n.s.*
	t = 0.30, p = .77, n.s
		d = 0.08
		Z = 6.2, p < .001
		t = 3.42, p < .006
			d = 0.08
			Z = 7.4, p < .001
			t = 2.89, p < .05

Fear	d = 0.11
	Z = 9.2, p < .001
	t = 3.47, p < .0046
		d = 0.08
		Z = 7.0, p < .001
		t = 2.73, p < .05
			d = 0.02
			Z = – 2.0, *n.s*
			t = – 0.71, p = .49, n.s

Happy	d = 0.01
	Z = – 1.1, *n.s.*
	t = – 0.42, p = .68, n.s
		d = 0.09
		Z = 7.6, p < .001
		t = 3.75, p < .01
			d = 0.11
			Z = 9.3, p < .001
			t = 9.29, p < .0001

Sad	d = 0.06
	Z = 4.9, p < .001
	t = 3.75, p < .003
		d = 0.08
		Z = 6.7, p < .001
		t = 2.63, p < .05
			d = 0.02
			Z = 2.1, *n.s.*
			t = – 1.88, p = .08, n.s

Surprise	d = 0.20
	Z = 17.0, p < .001
	t = 5.50, p < .0001
		d = 0.03
		Z = 2.1, *n.s.*
		t = 0.01, p = .99, n.s
			d = 0.18
			Z = – 15.3, p < .001
			t = – 3.77, p < .01

Table 5 Confusion matrix for the AV condition

		Response					
		Anger	Disgust	Fear	Happy	Sad	Surprise
Presented emotion	Anger	**69.1**	10.2	3.4	8.5	7.3	1.5
	Disgust	6.4	**41.6**	17.4	5.6	18.4	10.7
	Fear	4.2	11	**43.9**	8	26	7
	Happy	0.4	0.4	0.8	**91.5**	1.7	5.1
	Sad	2.4	6.5	4.8	7.8	**78**	0.4
	Surprise	1.4	1.7	5.7	27.7	2.5	**61**

For Tables 5, 6 and 7, percent correct emotion identification is in bold.

Table 6 Confusion matrix for the VO condition

		Response					
		Anger	Disgust	Fear	Happy	Sad	Surprise
Presented emotion	Anger	**62.3**	8.2	5.2	7.2	14.9	2.2
	Disgust	7.8	**42**	19.3	5.2	21.5	4.1
	Fear	9.1	13.9	**33.7**	8.1	27.5	7.6
	Happy	0.6	1.1	0.9	**92.8**	3.5	1.1
	Sad	7.6	7.7	6	5.6	**72.1**	0.9
	Surprise	7.5	4.7	5.8	34.6	6.7	**40.6**

Table 7 Confusion matrix for the AO condition

		Response					
		Anger	Disgust	Fear	Happy	Sad	Surprise
Presented emotion	Anger	**48.7**	29.7	2.4	10.6	5.8	2.9
	Disgust	22.6	**35**	10.9	7.1	13.4	11
	Fear	4.9	6.6	**37.4**	14.7	30.3	6.1
	Happy	2	4.8	1.9	**83.6**	6.3	1.4
	Sad	0.9	6.9	7.2	12.9	**71.1**	1.1
	Surprise	6.4	2.8	4.2	23.9	3.3	**59**

et al., 2008; Biehl et al., 1997). Expressions of Happy and Fear were rarely confused with other emotion types.

Interestingly, expressions of Sad in the AO condition were at times misidentified as Happy, further investigation of the data suggests that this was mainly driven by the stimuli produced by one of the male speakers (M5 whose expressions were recognized with the lowest accuracy scores, see Fig. 4). Across all presentation conditions, expressions of Surprise were most likely to be misidentified as Happy which is also a commonly observed finding (see Kim & Davis, 2012; Tanaka et al., 2015).

Variability of speakers

Figure 4 shows participants' mean percent accuracy score for identifying emotion expressions that were produced by each of the ten speakers in the CAVES dataset. The variability is reasonably low, i.e., there were high agreement across the perceivers in identifying the speakers' emotions, indicating the judgments and the expressions were reliable (Biehl et al., 1997). We see also calculated Gwet's AC1 inter-rater reliability scores (Gwet, 2014) using the irrCAC R package, version 1.3 (Gwet, 2023). This was done using the recognition scores as a measure of inter-rater reliability, i.e., by assessing how often a presented emotion was recognized as that emotion (the correct count data). Given that the interpretation of AC1 scores is like the kappa statistic, AC1 ranges from almost perfect disagreement (– 1.00) to almost perfect agreement (+ 1.00) and a score of zero indicates chance reliability, we can use the nomenclature of Landis and Koch (1977). Based on this, the CAVES emotion reliability ranged from slight (0.00–0.20) to almost perfect

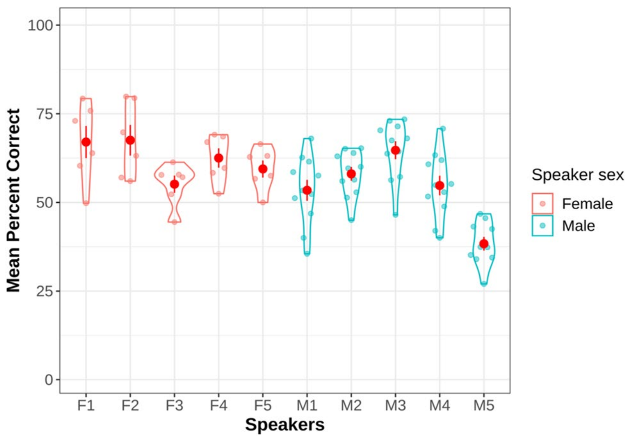

Fig. 4 Mean percent correct recognition score for each speaker in the CAVES dataset. *Note.* Female speakers were given identifiers that started with 'F' with a number from 1 to 5 to denote each individual speaker. Similarly, males were given identifiers that started with 'M'

(0.81–1.00), with most falling in the fair (0.21–0.40) to moderate range (0.41–0.60), i.e., AC1 disgust = 0.15 (SE = .030); AC1 fear = 0.2 (SE = .037); AC1 = 0.393 SE = .023); AC1 anger = 0.45 (SE = .048); AC1 sad = 0.6 (SE = .041); AC1 happy = 0.84 (SE = .039).

A Kruskal–Wallis test indicated that emotion expressions produced by female speakers were recognized at a higher accuracy than male speakers $\chi2(1) = 318.5$, $p < .001$. This is a common finding in the literature of emotion perception studies (for example, see Wells, Gillespie & Rotshtein, 2016).

Tukey HSD tests indicated that expressions produced by speaker F2 and F1 were recognized at significantly higher accuracy rates than all other participants ($p < .001$) except M3. The difference between F1 and F2 was not significant. Accuracy at recognizing the expressions produced with speaker M5 was significantly lower than all the other speakers ($p < .001$).

Variability of sentences

There were a total of 50 different sentences that were recorded in the CAVES dataset. Collapsing across emotion type and speakers, participants recognized all sentence stimuli within the range of 54–63%. For 39 out of 50 sentences, the emotion recognition accuracy rates were within a range of 56 to 59%. Figure 5 shows the distribution of scores for all sentences for each emotion type and the outliers (i.e., sentences with identifiers of 0410 and 0510).

Overall, the variability in accuracy scores across individual sentences is relatively low, indicating that these sentences are similar in their characteristics relevant to emotion expressions by Cantonese speakers, thus confirming the validity of our emotion selection and manipulation procedure (see Table 2). Further, the low variability suggests that the selected sentences (with the exception of a few outliers) which were semantically neutral did not bias responses to any particular emotion type.

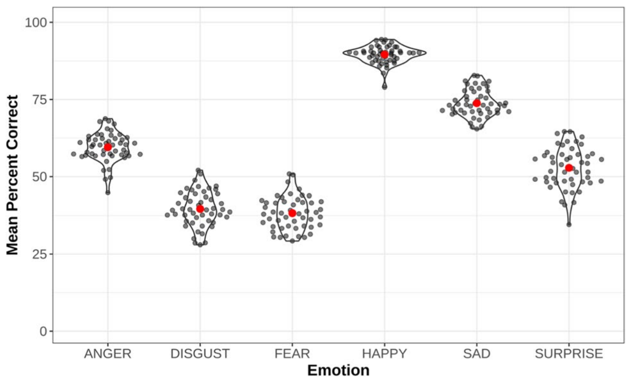

Fig. 5 Mean percent correct recognition scores for all 50 sentences across the six emotion types

In sum, the emotion expressions of the CAVES dataset can be recognized at above chance accuracy rates; show a significant AV benefit effect, and the response distribution frequencies are more or less similar to those of other emotion perception and evaluation studies (e.g., Kim & Davis, 2012; Li et al., 2017) but do differ from the RAVDESS (Livingstone & Russo, 2018). While differences with other studies were found in the patterning of the confusion matrices, these differences were likely driven in part by idiosyncratic individual differences (speaker M5 for example). Cultural or language differences between Cantonese speakers and the other language speakers in the other studies may also have contributed to different outcomes (e.g., English speakers in Kim & Davis 2012; Japanese and Dutch in Tanaka et al., 2015; and German expressers of the Radboud dataset, Ebner et al., 2010).

It is also worth pointing out a caveat related to the evaluation study. As mentioned above, due to its size, it was impractical to have all participants assess all items in the dataset. Thus, although there were approximately equal number of stimulus judgments made at the level of emotion and presentation type there were unequal numbers of judgments per item. This approach to evaluation could introduce variability in participant judgments for different items that may have affected the overall findings.

Discussion and potential applications

We have developed a dataset of auditory and visual expressive speech in Cantonese (CAVES). This dataset provides a new research tool for studying spoken emotional expression in a tonal language (Cantonese). The dataset consists of 3000 high-quality Audio-visual recordings by ten (five female) native Cantonese speakers. Items were validated by 15 native Cantonese perceivers. Overall, each emotion was recognized at accuracy levels greater than chance (Anger = 60.3%, Disgust = 39.5%, Fear = 38.3%, Happy = 89.3, Sad= 73.7, Surprise = 53.3, chance = 16.7%). The expression 'happy' was the most accurate for all presentation modes (AV, VO, and AO).

The dataset provides a useful resource for research on such topics as the auditory and visual expressive speech (see Kim, Bailey & Davis, 2018), on how emotions are expressed in a tonal language, and on how emotions are expressed in speech. For example, in the auditory domain, the background-noise free, high quality auditory stimuli allow the precise measurement of the acoustic parameters associated with tones (e.g., F0-based measures, time-based measures) that can be contrasted across emotions and compared to the neutral condition (here, the ProsodyPro suite of Praat scripts is an excellent measurement resource, Xu, 2013).

The CAVES dataset provides an excellent resource to systematically explore how emotional speech is expressed in a tonal language for both auditory, visual, and auditory-visual presentation. In general, it has been shown that when vocal emotions are expressed in a tonal language, a restriction in pitch variation occurs compared to non-expressive speech (Ross et al., 1986; Anolli et al., 2008; Chong et al., 2015). More recent work in Mandarin (that has four phonetic tones, tone 1 - high level; tone 2 - rising; tone 3 - low/dipping, and tone 4 - falling), has explored this interaction between the expression of lexical tone and emotion in more detail. For example, in a series of studies that examined the auditory expression of emotional speech in Mandarin, Wang and colleagues found that variation in F0 in emotion expression was reduced for tone 1 (Wang & Lee, 2015; Wang & Quian, 2018; Wang et al., 2018). That is, Wang and Lee (2015) examined the expression of happy, neutral, and sad and found that a restriction in F0 variation occurred in all high, level tone sequences (tone 1 group) for the expression of happiness but did not happen for the dynamic tones. Wang and Quian (2018) replicated Wang and Lee (2015) but added in two more emotions (anger and fear). The results confirmed that the expression of different emotions had a restricted F0 range for tone 1.

These studies also showed that even though tone 1 had restricted pitch variation, the expressed emotions were well recognized. To explain this, it was proposed that speakers and listeners used other cues (e.g., duration, pitch and intensity) to express and identify emotions, i.e., that when a certain cue (e.g., pitch) is restricted in one language, other cues will be exaggerated to allow the vocal emotions to be identified (Wang et al., 2018). To determine whether emotion expression or the expression of tones is affected more by their interaction, Chang et al. (2023) conducted a perception experiment in Mandarin for angry, fear, happy, sad, and neutral expressions. Listeners were asked to identify the tones or the emotion expressed and the results showed that emotions affect Mandarin tone identification to a greater extent than Mandarin tones affect emotion recognition.

In addition to examining how lexical tone affects the perception of emotion and vice-versa, the Chang et al. (2023) study also conducted acoustic analyses of F0 (mean and range),mean amplitude and duration. Although they found that emotional expression influences Mandarin tone production and did so to different degrees depending on which Mandarin tones were spoken and which emotions expressed, the pattern of their results was at odds with earlier work showing a restriction F0 range for tone 1 (e.g., for tone 1, anger had a larger F0 range than the emotion neutral baseline). Chang et al. suggested that differences in methodology and materials may have led to the discrepancy between their results and others.

In our view, the CAVES dataset offers a unique resource for further probing and understanding how the expression of lexical tone and emotion type interact. For example, the CAVES dataset, with its large number of instances, allows for a machine learning approach to the issue of how lexical tone and emotional expressive speech may interact in the richer tone space of Cantonese. Recently Kanwal et al. (2022) have developed an auditory emotion classification method using robust features that achieved state of the art correct emotion classification from the English language RAVDESS emotion dataset (Burkhardt et al., 2005) and the German language EMO-DB emotion dataset (Livingstone & Russo, 2018). Applying this method to the CAVES dataset, and comparing the overall correct classification rates with those of the RAVDESS and EMO-DB, as well as comparing the confusion matrices, will reveal whether and how the expression of lexical tone affects emotion classification.

Likewise, in the visual domain, the high-quality videos allow for markerless tracking of head and face motion (e.g., using the openface software package, Baltrusaitis et al., 2018). A comparison of emotion classification based on the visual properties of emotion across tone and non- tonal language datasets is interesting given claims that the expression of lexical tone affects head and face motion (Burnham et al., 2022). In addition to using classification models, the relative performance of human perceivers (for both auditory only and visual only presentation) on these datasets would also be an option.

Speech in noise (emotion in noise) recognition is also a research area that will benefit from the CAVES dataset. Recent studies have suggested that emotional speech is more intelligible in noise than neutral speech (e.g., Gordon & Ancheta, 2017; cf., Davis et al., 2017). The CAVES dataset provides an opportunity to investigate the basis of this effect for a larger set of emotions and talkers. The CAVES dataset also provides the materials to investigate the perception of lexical tones in noise. This type of investigation is important for assessing design choice for the language model of cochlear implants for tone speakers (see Wong et al., 2018). Not only does the CAVES dataset allow for comparison of different tones and tone position to be tested, it also enables any effect of emotional expression and visual speech to be determined.

Another research area for which the CAVES dataset will prove useful concerns the factors that modulate emotion recognition performance, i.e., why some depictions of spoken emotion are better recognized than others. That is, it has been proposed that stimuli from talkers who produced more consistent emotion portrayals will be better recognized (see Davis & Kim, 2019). Here, the high quality auditory and visual stimuli of the CAVES dataset allow within and across talker consistency of auditory and visual properties to be easily assessed.

Open practices statement The CAVES Dataset and the data from the validation experiment are available (only for research purposes) at https://forms.office.com/r/3VfeWQnAVa .

Funding Open Access funding enabled and organized by CAUL and its Member Institutions

References

Anolli, L., Wang, L., Mantovani, F., & De Toni, A. (2008). The voice of emotion in Chinese and Italian young adults. *Journal of Cross-Cultural Psychology, 39*(5), 565–598. https://doi.org/10.1177/0022022108321178

Baltrusaitis, T., Zadeh, A., Lim, Y. C., & Morency, L. P. (2018). Openface 2.0: Facial behavior analysis toolkit. In *Proceedings of 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)* (pp. 59–66). IEEE. https://doi.org/10.1109/FG.2018.00019

Baveye, Y., Bettinelli, J. N., Dellandréa, E., Chen, L., & Chamaret, C. (2013). A large video dataset for computational models of induced emotion. In *Proceedings of 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction* (pp. 13–18). IEEE. https://doi.org/10.1109/ACII.2013.9

Biehl, M., Matsumoto, D., Ekman, P., Hearn, V., Heider, K., Kudoh, T., & Ton, V. (1997). Matsumoto and Ekman's Japanese and Caucasian facial expressions of emotion (JACFEE): Reliability data and cross-national differences. *Journal of Nonverbal Behavior, 21*(1), 3–21. https://doi.org/10.1023/A:1024902500935

Boersma, P., & Weenink, D. (2014). *Praat: Doing Phonetics by Computer*. http://www.praat.org/.

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F., & Weiss, B. (2005). A dataset of German emotional speech. *In Ninth European Conference on Speech Communication and Technology.* https://www.researchgate.net/profile/Felix-Burkhardt-2/publication/221491017_A_dataset_of_German_emotional_speech/links/00b7d5226f45d66e38000000/A-dataset-of-German-emotional-speech.pdf

Burnham, D., Vatikiotis-Bateson, E., Barbosa, A. V., Menezes, J. V., Yehia, H. C., Morris, R. H., et al. (2022). Seeing lexical tone: Head and face motion in production and perception of Cantonese lexical tones. *Speech Communication, 141*, 40–55. https://doi.org/10.1016/j.specom.2022.03.011

Busso, C., Bulut, M., Lee, C. C., Kazemzadeh, A., Mower, E., Kim, S., & Narayanan, S. S. (2008). Iemocap: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation, 42*(4), 335–339. https://doi.org/10.1007/s10579-008-9076-6

Chang, H. S., Lee, C. Y., Wang, X., Young, S. T., Li, C. H., & Chu, W. C. (2023). Emotional tones of voice affect the acoustics and perception of Mandarin tones. *Plos one, 18*(4), e0283635. https://doi.org/10.1371/journal.pone.0283635

Chen, J., Wang, C., Wang, K., Yin, C., Zhao, C., Xu, T., et al. (2021). HEU Emotion: A large-scale database for multimodal emotion recognition in the wild. *Neural Computing and Applications, 33*, 8669–8685. https://doi.org/10.1007/s00521-020-05616-w

Chong, C., Kim, J., & Davis, C. (2015). Exploring acoustic differences between Cantonese (tonal) and English (non-tonal) spoken expressions of emotions. In *Sixteenth Annual Conference of the International Speech Communication Association* (pp. 1522–1525). https://www.isca-speech.org/archive_v0/interspeech_2015/papers/i15_1522.pdf

Chou, H. C., Lin, W. C., Chang, L. C., Li, C. C., Ma, H. P., & Lee, C. C. (2017). NNIME: The NTHU-NTUA Chinese interactive multimodal emotion corpus. In *Proceedings of the Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)* (pp. 292–298). https://doi.org/10.1109/ACII.2017.8273615

Clavel, C., Vasilescu, I., Devillers, L., & Ehrette, T. (2004). *Fiction database for emotion detection in abnormal situations* (pp. 2277–2280). *Proceedings of ICSLP*.

Clavel, C., Vasilescu, I., Devillers, L., Richard, G., Ehrette, T., & Sedogbo, C. (2006). The SAFE Corpus: illustrating extreme emotions in dynamic situations. In *First International Workshop on Emotion: Corpora for Research on Emotion and Affect (International conference on Language Resources and Evaluation (LREC 2006))* (pp. 76–79).

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Routledge Academic.

Cole, J. (2015). Prosody in context: A review. *Language, Cognition and Neuroscience, 30*(1–2), 1–31. https://doi.org/10.1080/23273798.2014.963130

Davis, C., Chong, C. S., & Kim, J. (2017). The effect of spectral profile on the intelligibility of emotional speech in noise. In *Proceedings of INTERSPEECH 2017* (pp. 581–585). https://doi.org/10.21437/Interspeech.2017-948

Davis, C., & Kim, J. (2019). Auditory and visual emotion recognition: Investigating why some portrayals are better recognized than others. In *Proceedings of the 15th International Conference on Auditory-Visual Speech Processing* (pp. 33–37). https://www.isca-speech.org/archive_v0/AVSP_2019/pdfs/AVSP_2019_paper_18.pdf

Davis, C., & Kim, J. (2006). Audio-visual speech perception off the top of the head. *Cognition, 100*(3), B21–B31. https://doi.org/10.1016/j.cognition.2005.09.002

Dhall, A., Goecke, R., Lucey, S., & Gedeon, T. (2012). Collecting large, richly annotated facial-expression databases from movies. *IEEE multimedia, 19*(3), 34. https://doi.org/10.1109/MMUL.2012.26

Douglas-Cowie, E., Campbell, N., Cowie, R., & Roach, P. (2003). Emotional speech: Towards a new generation of datasets. *Speech Communication, 40*(1–2), 33–60. https://doi.org/10.1016/S0167-6393(02)00070-5

Ekman, P., & Friesen, W. V. (1976). *Pictures of facial affect*. Consulting Psychologists Press.

Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion, 6*(3–4), 169–200. https://doi.org/10.1080/02699939208411068

Ebner, N. C., Riediger, M., & Lindenberger, U. (2010). FACES—A dataset of facial expressions in young, middle-aged, and older women and men: Development and validation. *Behavior Research Methods*, *42*(1), 351–362. https://doi.org/10.3758/BRM.42.1.351

Forster, K. I., & Forster, J. C. (2003). DMDX: A Windows display program with millisecond accuracy. *Behavior research methods, instruments, & computers, 35*, 116–124. https://doi.org/10.3758/BF03195503

Goeleven, E., De Raedt, R., Leyman, L., & Verschuere, B. (2008). The Karolinska directed emotional faces: A validation study. *Cognition and Emotion, 22*(6), 1094–1118. https://doi.org/10.1080/02699930701626582

Gordon, M. S., & Ancheta, J. (2017). Visual and acoustic information supporting a happily expressed speech-in-noise advantage. *The Quarterly Journal of Experimental Psychology, 70*(1), 163–178. https://doi.org/10.1080/17470218.2015.1130069

Gravier, G., Betser, M., & Ben, M. (2010). *AudioSeg: Audio Segmentation Toolkit*, release 1.2. IRISA. https://gforge.inria.fr/projects/audioseg. Groningen corpus. http://www.elsnet.org/groningen.html

Grimm, M., Kroschel, K., & Narayanan, S. (2008). The Vera am Mittag German audio-visual emotional speech database. In *IEEE International Conference on Multimedia and Expo* (pp. 865–868). IEEE.

Gwet, K. L. (2014). *Handbook of Inter-Rater Reliability* (4th Edition). Advanced Analytics, . ISBN:978-0970806284.

Gwet, K. L. (2023). irrCAC: Computing the Extent of Agreement among Raters with Chance-Corrected Agreement Coefficient (CAC). *R package version, 1*, 3.

Jackson, P., & Haq, S. (2015). *Surrey Audio-Visual Expressed Emotion (SAVEE) Database*. http://kahlan.eps.surrey.ac.uk/savee/

Kanwal, S., Asghar, S., & Ali, H. (2022). Feature selection enhancement and feature space visualization for speech-based emotion recognition. *PeerJ Computer Science, 8*, e1091. https://doi.org/10.7717/peerj-cs.1091

Kim, J., Bailey, G., & Davis, C. (2018). Introduction to the Special Issue on Auditory-visual expressive speech and gesture in humans and machines. *Speech Communication, 98*, 63–67. https://doi.org/10.1016/j.specom.2018.02.001

Kim, J., Cvejic, E., & Davis, C. (2014). Tracking eyebrows and head gestures associated with spoken prosody. *Speech Communication*, *57*, 317–330. http://dx.doi.org/https://doi.org/10.1016/j.specom.2013.06.003

Kim, J., & Davis, C. (2012). Perceiving emotion from a talker: How face and voice work together. *Visual Cognition, 20*(8), 902–921. https://doi.org/10.1080/13506285.2012.713874

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 3*(1), 159–174. https://doi.org/10.2307/2529310

Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D. H., Hawk, S. T., & Van Knippenberg, A. D. (2010). Presentation and validation of the Radboud Faces Database. *Cognition and Emotion, 24*(8), 1377–1388. https://doi.org/10.1080/02699930903485076

Lenth, R. V. (2021). *emmeans: Estimated Marginal Means, aka Least-Squares Means*. R package version 1.6.3. https://CRAN.R-project.org/package=emmeans. https://doi.org/10.1080/2F00031305.1980.10483031

Li, Y., Tao, J., Chao, L., Bao, W., & Liu, Y. (2017). CHEAVD: a Chinese natural emotional audio–visual dataset. *Journal of Ambient Intelligence and Humanized Computing, 8*(6), 913–924. https://doi.org/10.1007/s12652-016-0406-z

Li, Y., Tao, J., Schuller, B., Shan, S., Jiang, D., & Jia, J. (2018). Mec 2017: Multimodal emotion recognition challenge. In *2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia)* (pp. 1–5). IEEE.

Lin, J. C., Wu, C. H., & Wei, W. L. (2012). Error weighted semi-coupled hidden Markov model for audio-visual emotion recognition. *IEEE Trans. Multimedia, 14*, 142–156. https://doi.org/10.1109/TMM.2011.2171334

Livingstone, S. R., & Russo, F. A. (2018). The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PLOS One, 13*(5), e0196391. https://doi.org/10.1371/journal.pone.0196391

Lubis, N., Gomez, R., Sakti, S., Nakamura, K., Yoshino, K., Nakamura, S., & Nakadai, K. (2016). Construction of Japanese audio-visual emotion database and its application in emotion recognition. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (LREC'16) (pp. 2180–2184).

Maddieson, I. (2013). Consonant Inventories. In M. S. Dryer & M. Haspelmath (Eds.), *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology http://wals.info/chapter/1

Martin, O., Kotsia, I., Macq, B., & Pitas, I. (2006). The eNTERFACE'05 audio-visual emotion database. In *In 22nd International Conference on Data Engineering Workshops (ICDEW'06)* (pp. 8–8). IEEE.

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing type I error and power in linear mixed models. *Journal of Memory and Language, 94*, 305–315. https://doi.org/10.1016/j.jml.2017.01.001

McKeown, G., Valstar, M., Cowie, R., Pantic, M., & Schroder, M. (2011). The SEMAINE Database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE transactions on affective computing, 3*(1), 5–17.

Perepelkina, O., Kazimirova, E., & Konstantinova, M. (2018). Ramas: Russian multimodal corpus of dyadic interaction for affective computing. In *International Conference on Speech and Computer* (pp. 501–510). Springer.

Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., & Mihalcea, R. (2018). Meld: A multimodal multi-party dataset for emotion recognition in conversations. arXiv preprint arXiv:1810.02508

Ringeval, F., Sonderegger, A., Sauer, J., & Lalanne, D. (2013). Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. In *In 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)* (pp. 1–8). IEEE.

Ross, E. D., Edmondson, J. A., & Seibert, G. B. (1986). 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*, *14*(2), 283–302. https://doi.org/10.1016/S0095-4470(19)30669-2

Sapiński, T., Kamińska, D., Pelikant, A., Ozcinar, C., Avots, E., & Anbarjafari, G. (2018). Multimodal database of emotional speech, video and gestures. In *International Conference on Pattern Recognition* (pp. 153–163). Springer.

Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. *Speech communication, 40*(1–2), 227–256. https://doi.org/10.1016/S0167-6393(02)00084-5

Scherer, K. R., Banse, R., & Wallbott, H. G. (2001). Emotion inferences from vocal expression correlate across languages and cultures. *Journal of Cross-Cultural Psychology*, *32*(1), 76–92. http://dx.doi.org/https://doi.org/10.1177/0022022101032001009

Schuller, B., Müller, R., Eyben, F., Gast, J., Hörnler, B., Wöllmer, M., Rigoll, G., Hothker, A., & Konosu, H. (2009). Being bored? Recognising natural interest by extensive audiovisual integration for real-life application. *Image and Vision Computing, 27*(12), 1760–1774. https://doi.org/10.1016/j.imavis.2009.02.013

Singmann, H., Bolker, B., Westfall, J., Aust, F., Ben-Shachar, M. S., Højsgaard, S., Fox, J., Lawrence, M. A., Mertens, U., Love, J., Lenth, R., Christensen, R. H. B. (2021). *afex: Analysis of Factorial Experiments.* R package version 1.0–1. https://CRAN.R-project.org/package=afex. https://cran.r-project.org/web/packages/afex/index.html

Tanaka, A., Takagi, S., Hiramatsu, S., In't Veld, E. H., & de Gelder, B. (2015). Towards the development of facial and vocal expression dataset in east Asian and Western cultures. In *of the 13th International Conference on Auditory-Visual Speech Processing* (pp. 63–66). https://www.isca-speech.org/archive_v0/avsp15/papers/av15_063.pdf

Vidal, A., Salman, A., Lin, W. C., & Busso, C. (2020). Msp-face corpus: A natural audiovisual emotional dataset. In *Proceedings of the 2020 international conference on multimodal interaction* (pp. 397–405). https://doi.org/10.1145/3382507.3418872

Villegas, J., Cooke, M., Aubanel, V. & Piccolino-Boniforti, M. A. (2011). *M TRANS: A multi-channel, multi-tier speech annotation tool.* [Online]. http://www.laslab.org/tools/mtrans/

Wagner, H. L. (1993). On measuring performance in category judgment studies of nonverbal behavior. *Journal of Nonverbal Behavior, 17*, 3–28.

Wang, T., & Lee, Y. C. (2015). Does restriction of pitch variation affect the perception of vocal emotions in Mandarin Chinese? *The Journal of the Acoustical Society of America, 137*(1), EL117-EL123. https://doi.org/10.1121/1.4904916

Wang, T., Lee, Y. C., & Ma, Q. (2018). Within and across-language comparison of vocal emotions in Mandarin and English. *Applied Sciences, 8*(12), 2629. https://doi.org/10.3390/app8122629

Wang, T., & Qian, Y. (2018). Are pitch variation cues indispensable to distinguish vocal emotions. In Proceedings of the 9th International Conference on Speech Prosody (pp. 324–328). 10.21437/SpeechProsody.2018-66.

Wells, L. J., Gillespie, S. M., & Rotshtein, P. (2016). Identification of emotional facial expressions: Effects of expression, intensity, and sex on eye gaze. *PLOS One, 11*(12), e0168307. https://doi.org/10.1371/journal.pone.0168307

Westfall, J., Kenny, D. A., & Judd, C. M. (2014). Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. *Journal of Experimental Psychology: General, 143*(5), 2020. https://doi.org/10.1037/xge0000014

Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. *Springer-Verlag New York. ISBN 978-3-319-24277-4,* https://ggplot2.tidyverse.org

Wong, L. L. N., & Soli, S. D. (2005). Development of the Cantonese Hearing In Noise Test (CHINT). *Ear and Hearing, 26*(3), 276–289. https://doi.org/10.1097/00003446-200506000-00004

Wong, P., Cheng, S. T., & Chen, F. (2018). Cantonese tone identification in three temporal cues in quiet, speech-shaped noise and two-talker babble. *Frontiers in Psychology, 9*, 1604. https://doi.org/10.3389/fpsyg.2018.01604

Xu, Y. (2013). *ProsodyPro—A tool for large-scale systematic prosody analysis*. Laboratoire Parole et Langage, France.

Yip, M. (2002). *Tone*. Cambridge University Press.

Zhalehpour, S., Onder, O., Akhtar, Z., & Erdem, C. E. (2016). BAUM-1: A spontaneous audio-visual face database of affective and mental states. *IEEE Transactions on Affective Computing, 8*(3), 300–313. https://doi.org/10.1109/TAFFC.2016.2553038

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.