Notes\_on\_datasets

I use a text editor (Sublime Text) to help me count occurrences within datasets. Excel was helpful for quickly counting unique speakers.

Unless otherwise noted, most of the quoted sections below are from the paper associated with that dataset.

# aesdd

From the readme

For the creation of v.1 of the database, 5 (3 female and 2 male) professional actors were recorded. 19 utterances of ambiguous out of context emotional content were chosen. The actors acted these 19 utterances in every one of the 5 chosen emotions. One extra improvised utterance was added for every actor and emotion. The guidance of the actors and the choice of the final recordings were supervised by a scientific expert in dramatology. For some of the utterances, more that one takes were qualified. Consequently, around 500 utterances occured in the final database.

No gender labels, but I manually determined that speakers 1, 2, and 5 are female-sounding and speakers 3, 4, and 6 are male-sounding. Contrary to the documentation, the dataset actually contains 6 different speakers.

“The recordings took place in the sound studio of the Laboratory of Electronic Media (Aristotle University of Thessaloniki, Greece), offering an appropriate acoustic environment, thus ensuring high quality recordings. Since the current work is focused on theatrical productions, the spoken /recorded phrases had to derive from theatrical scripts. Specifically, 19 utterances were chosen from different theatrical plays for the database formulation, based on the criterion of the ambiguity of their emotional context. The actors expressed these 19 sentences in Greek language in 5 different emotional contexts, namely happiness, sadness, anger, fear and disgust, because of their undisputed form, as stated above. Moreover, for every emotion, one extra /improvised utterance was recorded, while more than one recording were used for some utterances, resulting in around 500 utterances of emotional speech (5 actors x 5 emotions x 20 utterances). As all actors recorded the same utterances for all 5 emotions, it is ensured that the training process is not user-dependent or the verbal content of the utterances. A scientific expert in dramatology was present in order to supervise the recordings, to guide the actors and to make the proper adjustments/ corrections when needed, ensuring the quality and suitability of the acted speech.”

“Concerning the audio technical characteristics, all the recordings were coded in PCM wav format, with a sampling frequency of 44100Hz and 16-bit dynamic range.”

sadness/s05 (3).wav is corrupted.

# BAUM1

There are misspellings in the emotion labels.

“Subject Label”, “Clip”, and “Clip Name” don’t always match in the acted annotations: S019\_006, S019\_008, S019\_009, S019\_010, S019\_011, S019\_012, S019\_013, S019\_014, S019\_015, S019\_016, S019\_017, S019\_018, S019\_019, S019\_020, S019\_021, S019\_022, S019\_023

I have manually corrected these in preclean.tsv

Not all files appear in the annotations excel files. I looked and found 72 such files.

To recode these, I looked for other samples with similar elicitation prompts. This worked fine for the acted samples (8), but the spontaneous samples (64) are more heterogenous. I matched the subtitles of 3 of these exactly with others and carried over the labels. The rest were too subjective for me to feel confident rating them on my own. 61 records dropped

They made the actors watch some disturbing stuff to elicit emotions…

From the paper

The data was collected from 31 subjects, 17 of which are female, which are shown in Fig. 2. All subjects are native speakers of Turkish, and have an age range of 19-65.

…



The paper is inconsistent on the actual number of participants coded female.

S015\_007 spontaneous emo label does not match emo code in annotations!

Some mp4s no audio? Check downstream

Boredom is mapped to negative valence (van Tilburg & Igou, 2017; Kort, Reilly, & Picard, 2001); Regan, Mandryk, & Atkins (2007)

Interest is mapped to positive valence (Kort, Reilly, & Picard, 2001); remapped to curiosity; see notes in emoreact section on curiosity

Baum1 paper: “*Interest (Curiosity)*: You want to learn about your friend’s secret.”

Contempt mapped to negative (Trnka, Mana, & Kuška, 2021; Ferran 2017; Melwani, Mueller, & Overbeck, 2012)

Surprise is negative (Noordewier & Breugelmans 2013; Koch, Alves, & Krüger 2016; Noordewier, Topolinski, & Van Dijk 2016)

Unsure->uncertain

Bothered can be negative too

# BAUM2

Extensive manual Pre-cleaning done in data\_dir.xlsx; stages of processing go from leftmost sheet to rightmost sheet

Recoded to valence and recounted majority vote

Turkish and English

# cafe

Surprise is default negative (see…)

“This dataset includes six different sentences, pronounced by twelve actors, in six basic emotions plus one neutral emotion. The basic emotions are acted in two different intensities. This represents a total of 936 different audio samples.”

“The sentences had to be emotionally neutral from a semantic point of view, yet well suited to be acted in various emotions. They also had to be reasonably easy to pronounce. Finally, they had to be composed of the same number of syllables so that neither one has more importance than the others.”

All samples accounted for

Québec French

“This dataset contains six different sentences, pronounced by six male and six female actors, in six basic

emotions plus one neutral emotion. The six basic emotions are acted in two different intensities. The audio is digitally recorded at high-resolution (192 kHz sampling rate, 24 bits per sample).”

# CREMA-D

From livingstone & russo 2018:

“The CREMA-D consists of 91 English-speaking actors, expressing six spoken emotions. One sentence was produced at three levels of intensity, the other 11 sentences with unspecified intensity. This extensive set of 7442 recordings was validated by 2443 raters using crowdsourced participants (Survey Sampling International) in an Internet-presented format, providing approximately 10 ratings per clip.”

“There are 91 actors, 48 male and 43 female (51 actors worked with one director, 40 with the another). The actors were between the ages of 20 and 74 with a mean age of 36. Table 2 provides detailed age information. Several racial and ethnic backgrounds were represented in the actor group: Caucasian, African American, Hispanic, and Asian. Table 3 provides a detailed breakdown of the racial and ethnic groups.”

FROM README:   
CREMA-D is a data set of 7,442 original clips from 91 actors. These clips were from 48 male and 43 female actors between the ages of 20 and 74 coming from a variety of races and ethnicities (African America, Asian, Caucasian, Hispanic, and Unspecified).

Actors spoke from a selection of 12 sentences. The sentences were presented using one of six different emotions (Anger, Disgust, Fear, Happy, Neutral, and Sad) and four different emotion levels (Low, Medium, High, and Unspecified).

Participants rated the emotion and emotion levels based on the combined audiovisual presentation, the video alone, and the audio alone.

Crowd-sourced ratings

“Binomial majority is used to define majority recognition. Unlike traditional majority, which is defined as more than 50% of raters having selected the specific emotion, binomial majority is achieved when a binomial test would reject at the 95% confidence level the null hypothesis that the most commonly chosen label is selected randomly from the six possible labels.”

* Recode intended emotions to valence
* Recode votes to valence
  + A, D, F, S -> -1, N -> 0, H -> 1
  + H:N -> 1
  + Remaining -> -1 if the vote string doesn’t contain “N”
* Retain the record if any of VoiceVote valence, FaceVoteValence, or MultiModalVote valence matches the intended valence

Discarded 569, kept 6873

Intended emotion used for final valence label

# filelist is copy-pasted from data\_selection.xlsx

The file 1040\_ITH\_SAD\_XX.wav does not exist, but 1040\_ITH\_SAD\_X.wav does.

# dzafic

Just 6 samples, so I created the tsv manually

# ekorpus

The corpus contains 1,234 Estonian sentences that express anger, joy and sadness, or are neutral. [867 retained]

Female voice, 44.1 KHz, 16Bit, Mono;

wav, textgrid: phonemes, words, sentences.

The audio-recordings and text of sentences can be downloaded and saved.

^ <https://metashare.ut.ee/repository/browse/estonian-emotional-speech-corpus/4d42d7a8463411e2a6e4005056b40024a19021a316b54b7fb707757d43d1a889/>

Elicited spontaneously, but they read from preselected materials

Used <http://peeter.eki.ee:5000/reports/valence>

To filter only samples with >51% rater recognition for positive, negative, and neutral valence

[Text influence: all

Emotion: all

Min. recognition %: 51]

Filter op returned list of sample IDs

Used the corresponding textgrid field for emotion category label

69 Sample IDs missing from dataset: 466, 468, 475, 481, 485, 493, 495, 499, 509, 511, 521, 550, 552, 554, 556, 558, 564, 570, 572, 578, 582, 586, 588, 592, 594, 598, 608, 612, 626, 638, 642, 469, 471, 473, 507, 515, 519, 525, 527, 529, 531, 533, 535, 537, 541, 545, 547, 548, 560, 562, 566, 568, 574, 584, 590, 596, 600, 602, 604, 606, 610, 614, 620, 622, 624, 632, 634, 640, 644, 646

253 sample IDs discarded since perceived valence did not match intended valence:  
120325, 13875, 13701, 13935, 13737, 120706, 120732, 13673, 14039, 121322, 121324, 173, 217, 229, 279, 285, 13435, 13473, 13525, 13661, 13677, 13683, 13687, 13743, 13799, 13877, 13879, 13903, 13917, 13923, 13929, 13967, 13979, 120151, 120161, 120171, 120173, 120203, 120221, 120225, 120227, 120233, 120317, 120365, 120526, 120532, 120534, 120536, 120572, 120590, 120692, 120712, 120718, 120744, 120901, 120909, 120961, 121141, 121188, 121208, 121286, 121312, 171, 121330, 13639, 14131, 14163, 14167, 14179, 14353, 120263, 120311, 120403, 120568, 120596, 120598, 120600, 120616, 120676, 120738, 120819, 120857, 120943, 120951, 120955, 120963, 120971, 121003, 121123, 121145, 121156, 121194, 121206, 57, 91, 175, 269, 395, 13249, 13311, 13335, 13341, 13627, 43, 120279, 120281, 120845, 65, 69, 120323, 121113, 120337, 127, 13583, 121214, 120367, 120369, 121222, 167, 185, 195, 121121, 121266, 120401, 121232, 13643, 13649, 13773, 13949, 13983, 120923, 120445, 120451, 120457, 120540, 14125, 14129, 121137, 14137, 121139, 120586, 120588, 121234, 14165, 121150, 121268, 120967, 120975, 14201, 120979, 120634, 120642, 120995, 14249, 120662, 14259, 14261, 121174, 13001, 121007, 120670, 14281, 14283, 13033, 13071, 120680, 121182, 121051, 121184, 13091, 121186, 120742, 14375, 120750, 13179, 120754, 121276, 121067, 121071, 121073, 14451, 120768, 14463, 120770, 14477, 13255, 121288, 120776, 121198, 120780, 14543, 14547, 13331, 121202, 120814, 120163, 120175, 120177, 120189, 120829, 120831, 13451, 120833, 13459, 120835, 120839, 120696, 13009, 120782, 14295, 13251, 121075, 120219, 13679, 120877, 120881, 120895, 120931, 121041, 121152, 121244, 121310, 305, 13077, 13085, 13181, 13199, 13257, 13267, 13433, 13675, 13721, 13839, 13843, 13859, 13965, 14035, 14109, 14205, 14209, 120139, 120147, 120215, 120297, 120315, 120347, 120355, 120359, 120381, 120538, 120644, 120690, 120788

# EmoDB

From abstract…

Ten actors (5 female and 5 male) simulated the emotions, producing 10 German utterances (5 short and 5 longer sentences) which could be used in everyday communication and are interpretable in all applied emotions.

The recordings were taken in an anechoic chamber with high-quality recording equipment. In addition to the sound electro-glottograms were recorded. The speech material comprises about 800 sentences (seven emotions \* ten actors \* ten sentences + some second versions).

The complete database was evaluated in a perception test regarding the recognisability of emotions and their naturalness. Utterances recognised better than 80% and judged as natural by more than 60% of the listeners were phonetically labelled in a narrow transcription with special markers for voice-quality, phonatory and articulatory settings and articulatory features.

The database can be accessed by the public via the internet (<http://www.expressive-speech.net/emodb/>). 🡨 link is no good

From <https://www.kaggle.com/piyushagni5/berlin-database-of-emotional-speech-emodb> :

The EMODB database is the freely available German emotional database. The database is created by the Institute of Communication Science, Technical University, Berlin, Germany. Ten professional speakers (five males and five females) participated in data recording. The database contains a total of 535 utterances. The EMODB database comprises of seven emotions: 1) anger; 2) boredom; 3) anxiety; 4) happiness; 5) sadness; 6) disgust; and 7) neutral. The data was recorded at a 48-kHz sampling rate and then down-sampled to 16-kHz.

Every utterance is named according to the same scheme:

* Positions 1-2: number of speaker
* Positions 3-5: code for text
* Position 6: emotion (sorry, letter stands for german emotion word)
* Position 7: if there are more than two versions these are numbered a, b, c ....

Example: 03a01Fa.wav is the audio file from Speaker 03 speaking text a01 with the emotion "Freude" (Happiness).

Table

Description automatically generated

# EmoReact\_V\_1.0

Train-val-test split provided, keeping speakers separated between splits

Discard the split in favor of re-splitting, stratifying by language and dataset (and gender?); speaker gender was unbalanced by split too

multimodal emotion dataset of children between the ages of four and fourteen years old. The dataset contains 1102 audio-visual clips annotated for 17 different emotional states: six basic emotions, neutral, valence and nine complex emotions including curiosity, uncertainty and frustration.

Regan, Mandryk, & Atkins (2007) associate excitement with positive valence and frustration with negative valence and boredom is negative

Lang (1995); Russell, Weiss, & Mendelsohn (1989) – excitement is positive

Carleton (2016); Anderson, Carleton, Diefenbach, & Han (2019): uncertainty is negative

Curiosity gets mixed reviews: van Lieshout, Traast, de Lange, Cools (2019); Shin & Kim (2019); Noordewier & van Dijk (2017)

[Nojavanasghari, Baltrusaitis, Hughes, & Morency (2016)](https://www.researchgate.net/profile/Behnaz-Nojavanasghari/publication/305945989_The_Future_Belongs_to_the_Curious_Towards_Automatic_Understanding_and_Recognition_of_Curiosity_in_Children/links/57a6fd0f08ae455e8542cbe7/The-Future-Belongs-to-the-Curious-Towards-Automatic-Understanding-and-Recognition-of-Curiosity-in-Children.pdf) looked at EmoReact and found that curiosity was associated with positive valence. Hill, Fombelle, & Sirianni (2016) also treat it as positive. So for EmoReact, curiosity could be treated as positively valenced.

I was going to treat them all as positively valenced samples, but MELD also allowed for multiple surprise valences. Therefore, surprise will not have a set valence unless no valence score is provided by the dataset.

Not all emotions in the paper are in the dataset labels… because of v1?

Present: curiosity, excitement, happiness, uncertainty, surprise, disgust, fear, frustration

missing: exploration, confusion, anxiety, attentiveness, anger, sadness, embarrassment

regarding annotating speaker identity, the paper cites Florian et al. as reference #47, but the 47th reference is Schroff, Kalenichenko, & Philbin 2015: “A unified embedding for face recognition and

clustering.”, which doesn’t sound like it’s about annotating speaker identity/gender. There was no other Florian in the references, so I can’t trace this method.

Anyways, I’m manually annotating gender of the child subjects myself, mostly using the thumbnails and listening to the audio when I’m not as sure

Sometimes it’s the adult interviewer’s voice, not the child subject; sometimes both. It is unclear whether the label votes are for the interviewer or child subject in these cases, so where I identified it, I manually omit the observation: VCR107\_2.mp4, BULLYING27\_2.mp4, GAMEBOY19\_2.mp4, GAMEBOY29\_2.mp4, KIMCHI65\_2.mp4, OLDCOMPUTERS128\_2.mp4, TYPEWRITERS19\_2.mp4, TYPEWRITERS28\_2.mp4. I didn’t check every single one, however. The interviewer is usually (always?) male sounding. REBECCA13\_2.mp4 is just ambient music.

A lot of them have interviewer speech

Some are non-verbal vocalizations

Every sample has a valence rating

Multiple ethnicities, NA English

I went back and manually labeled speaker IDs for this dataset too. No speaker appears in more than one split folder. There are supposed to be 21 unique speakers per split folder

This took forever but was necessary in order not to leak speakers between splits

I tried my best but tbh I wasn’t always 100% sure

Fuzzy votes…

Perceived emotion neutral + perceived valence non-zero ::> unk, +/- 1 (perceived valence)

Perceived valence 0 + perceived emotion non-neutral ::> perceived emotion, +/- 1 (valence of perceived emotion)

Basically, records were discarded only if the valence of a perceived emotion w/ unambiguous valence is cross-aligned with the perceived valence. Unaligned (but not cross aligned) records were aligned.

# Emotional\_EMA

From readme…

‘’’

This Electromagnetic Articulography (EMA) database includes articulatory motions recorded by an EMA system.

Talkers produced simulated (acted) emotional speech.

A set of 10 sentences was commonly used for speech recording of a male (AB) and two females (JN, LS), who are native speakers of American English.

On top of the 10 sentences, there are 4 additional sentences used for recording by only AB.

Each sentence was produced five times for four different emotions, such as neutrality, anger, sadness and happiness.

…AB produced 280 utterances (14 sentences x 5 repetitions x 4 emotions), and JR and JN produced 200 utterances (10 sentences x 5 repetitions x 4 emotions).

Each utterance was digitalized in 12-bit amplitude resolution with 16kHz sampling rate.

Speech was recorded simultaneously by the EMA system so that speech and corresponding articulatory movements are aligned in time.

‘’’

Created valence\_scores\_per\_sample from DocumentationEma.txt

In DocumentationEma.txt, there were two filenames misspelled:

|  |  |  |
| --- | --- | --- |
| 4EMO\_~43.WAV | = | 4emo\_ls\_angry\_41\_041.wav |
| 4EMO\_~86.WAV | = | 4emo\_ls\_happy\_32\_032.wav |

For the emotion category votes, each sample obtained a majority vote for one category (3/4 or 4/4 votes). All of these matched the intended emotion. These were from the best\_xxx\_files.txt files.

Samples were also rated on valence separately (different evaluators). These ratings didn’t always match the valence of the intended emotion. I kept the samples where either the majority valence vote (if present) or the average valence rating matched the valence of the intended emotion. I allowed both criteria to increase the number of samples retained. 32 samples were discarded because of perceived-intended mismatch.

The best\_xxx\_files.txt files do not contain all the files listed in DocumentationEma.txt! The leftovers were assessed by valence only.

Discarded 58

“A set of 14 sentences (given below), which are mostly neutral in emotional content, was used for speech recording. A male native speaker of American English, who has no formal theatrical vocal training, produced each sentence five times in a random order. Four different emotions, i.e., neutral, angry, sad and happy, were simulated by the subject. The subject produced a set of 70 utterances in a row for each emotion resulting in a total of 280 utterances (14 sentences x 5 repetitions x 4 emotions). Each utterance was digitized in 12-bit amplitude resolution with 16kHz sampling rate. Speech was recorded simultaneously by the EMA system so that speech and corresponding articulatory movements are aligned in time.”

# EmoV-DB\_sorted

Belgian French and North American English

elicitation prompts based on CMU Arctic (en) and SIWIS (fr)

The download link I got only has 4 out of the 5 speakers mentioned in the paper.

The French speaker is missing, so all samples are English…

Table

Description automatically generated

“Amused speech can contain chuckling sounds which overlap and/or intermingle with speech called speech-laughs[22] or can be only amused smiled speech [10]. So, for the amused data in our database, in order to collect as much data as possible and considering the relatively limited time the actors provided us, we focused on amused speech with speech-laughs. This choice was motivated by our previous study showing that this type of amused speech was perceived is perceived as more amused than amused smiled speech (without speech-laugh). Also in another study, we show that including laughter in synthesized is always perceived as amused no matter the style of speech it is inserted in (neutral or smiled) [11]. Based on the previous studies made on amusement, the actors were encouraged, while simulating the other emotions, to use nonverbal expressions before and even while uttering the sentences if they felt the need to (e.g. yawning for sleepiness, affect bursts for anger and disgust).”

# enterface\_db

from livingstone & russo 2018:

“A fourth set, the eNTERFACE'05, also provides audiovisual expressions of vocal emotional communication [76]. The set consists of 42 English-speaking lay-expressers from different countries, expressing six emotions in scenario-elicited format. Five distinct sentences for each emotion were produced with unspecified intensity. Recordings were included based on the judgements of two trained investigators. However, no measures of accuracy or reliability were provided. As such, the set cannot be assessed and compared against the performance of the RAVDESS or other existing sets.”

“46 subjects were invited to react to six different situations, each of them eliciting one of the following emotions: happiness, sadness, surprise, anger, disgust and fear.”

“Two human experts decided whether or not the subject had expressed itself in such a way that an untrained human observer could without ambiguity recognize the emotion present in the reaction, for each of the emotions to be elicited. In a post-processing step, samples in which the emotion was not clearly recognized were discarded, so that the database would only contain video samples carrying relevant affective information. In this postprocessing step, decision was made to remove 4 subjects whose none of the video samples carried a believable affective message”

“The final version of the database thus contains 42 subjects, coming from 14 different nationalities.”

“Among the 42 subjects, a percentage of 81% were men, while the remaining 19% were women.”

“The recordings lasted for two weeks. All the experiments were driven in English. Each subject was told to listen to six successive short stories, each of them eliciting a particular emotion. They had then to react to each of the situations and two human experts judged whether the reaction expressed the emotion in an unambiguous way. If this was the case, the sample was added to the database. If not, it was discarded”

“Eventually, the database consists of a total of 1166 video sequences. Out of these 1166 video sequences, 264 concern women recordings (23%) and 902 men recordings (77%).”

Some of these .avi files are quite long, over 1.5 min. in some cases; there may need to be a downstream check

File Naming anomalies for some of subject 3 and subject 11 and subject 6

Subject 11 files are mislabeled as subject 12 files

Subject 11 and 12 and both male speakers

I labeled the speaker genders myself.

“The video sequences were processed using a 720x576 Microsoft AVI format. The frame rate was equal to 25 frames per second, while pixel aspect ration was D1/DV PAL (1.067). The video was compressed using a DivX 5.0.5 Codec, to ensure easy portability. The audio sample rate was 48000 Hz, in an uncompressed stereo 16-bit format.”

# Emotional Speech Dataset

English and Mandarin Chinese

“The dataset consists of 350 parallel utterances with an average duration of 2.9 seconds spoken by 10 native English and 10 native Mandarin speakers. For each language, the dataset consists of 5 male and 5 female speakers in five emotions summarized as follows: 1) happy, 2) sad, 3) neutral, 4) angry, and 5) surprise. Speech data are sampled at 16 kHz and saved in 16 bits.”

I manually labeled speaker gender

Speakers 1-10 are mandarin speakers, 11-20 are English speakers

As a native speaker of both languages, it really sounds like PRC Mandarin and US English to me (although I speak Chinese with a Taiwanese accent.) One of the male speakers has a US southern accent.

“Speech data are sampled at 16 kHz and saved in 16 bits.”

# EYASE

“In this work, an Egyptian Arabic speech emotion database is pre- sented that includes four different emotions: angry, happy, neutral and sad. The introduced database includes a total of 579 speech utterances for 3 male and 3 female subjects”

“In this work, an Egyptian Arabic semi-natural emotion speech database is created from the award winning Egyptian drama series Hatha Al-Masaa ( ) ( “Hatha Almasaa ”). Four basic emotions were considered in the introduced Egyptian Arabic speech emotion (EYASE) database: angry (A), happy (H), neutral (N) and sad (S). The EYASE database was recorded for three male and three female lead professional actors. At the time of filming, the actors were within the age range from 22 to 45 years old and had between 12 and 22 years of professional experience, with the exception of the youngest female actor who had about six years of acting experience. Initially, sound clips were recorded and labelled based on visual, audio and story narrative as well as on the depicted actor emotion.”

“In this work, a semi-natural Egyptian Arabic speech emotion (EYASE) database was introduced that includes 579 utterances from 3 male and 3 female professional actors for the angry, happy, neutral and sad emotions.”

“Next, context unaware annotation based solely on the recorded sound clips was performed by two different labelers. Finally, only utterances on which all labelers agreed on were included in the EYASE database. In total, the EYASE database includes 579 utterances. All speech samples were recorded using the open source Audacity software ( https://www.audacityteam.org/ ) at a sampling rate of 44.1 kHz where a single utterance duration ranged from 1 to 6 s.”

# jl-corpus

New Zealand English

“…the speech was recorded from 4 speakers. All the speakers (two male and two female) were trained voice actors (two current broadcasters, one broadcasting tutor and one broadcasting trainee) of New Zealand English”

“The speech signal was sampled at 44.1kHz and stored as 16-bit numbers.”

“In total, there are 4 (speakers) × 5 (primary emotions) × 2 (repetitions) × 15 (sentences) × 2 (sessions) = 1200 primary emotion sentences and 4 (speakers) × 5 (secondary emotions) × 2 (repetitions) × (13 (emotion neutral sentences) + 2 (emotion salient sentences)) × 2 (sessions) = 1200 secondary emotion sentences, making a total of 2400 sentences, with a footprint of 520 MB.”

I only have access to the “unchecked and unannotated” raws. So the emotion code is the intended emotion

Default sample rate: 44100Hz

Encoding: 16 bit PCM

Chanel: Mono

Format: WAV

File naming rule: (Gender)(speaker.ID)\_(Emotion)\_(Sentence.ID)(session.ID)

Chart

Description automatically generated

No valence provided for encouraging, concerned, or assertive 🡪 discard

There actually aren’t any samples labeled as enthusiastic or pensive, so that isn’t helpful

“Raw JL corpus (unchecked and unannotated)\Perception test material on Qualtrics\main\_training” and “Raw JL corpus (unchecked and unannotated)\Perception test material on Qualtrics\sub\_training” may contain additionally usable files, but these were meant for perception tests. The documentation isn’t so good so I didn’t use these.

# LimaCastroScott

Non-speech vocalizations

“Two female (27 and 33 years of age) and two male (28 and 34 years of age) speakers produced the vocalizations. They were European Portuguese native speakers and did not have formal acting training. Two of them (one male and one female) had training in music, including singing lessons.”

Raters : “A total of 40 undergraduate students took part in the study. Twenty were assigned to the forced choice task (mean age = 19.9 years, SD = 1.4; 19 females), and the other 20 to the rating task (mean age = 20.3 years, SD = 2; 19 females). They were recruited from University of Porto and received course credits for their participation. Five participants in the forced choice task had some degree of formal musical training, including instrumental practice (average years = 3.2, SD = 1.9); 3 participants in the rating experiment also had some degree of musical training (average years = 3.7, SD = 4.6). All participants were native speakers of European Portuguese and reported no hearing impairments or speech disorders, no psychiatric or neurological illnesses, and no head trauma or substance abuse.”

“The speakers were invited to participate in one recording session. They were provided with a list of the emotions they had to express, as well as with a list of illustrative real-life scenarios typically associated with the experience of each emotion (see the Appendix). After an initial briefing, the speakers read the emotion words and the corresponding scenarios and were asked to produce the vocal sounds they would make if they were experiencing that emotion. No guidance was provided as to the specific kind of sounds they should make, apart from general examples (e.g., some people laugh when they feel amused or sob when they feel sad). They were told that they should not produce sounds with verbal content (e.g., “yuck,” “yippee,” “phew”), only nonverbal vocalizations. After a short familiarization phase, several different exemplars of the same category were recorded from each speaker (approximately seven). Extra recordings were made whenever the vocalizations were deemed to be unrecognizable (as exemplars of the intended emotions) by the experimenter (first author). They were told to try to sound as natural and spontaneous as possible. It has been acknowledged that some emotion categories can be expressed in distinct manners—emotion families—and this variation might be linked with distinct acoustic profiles (e.g., Banse & Scherer, 1996; Ekman, 1992; Scherer, 2003; Scherer, Johnstone, & Klasmeyer, 2003); for instance, anger can be expressed in a hot explosive manner (rage) or in a cold controlled way. In our stimulus set, there is variability regarding this issue: Anger was produced mainly in a hot rather than in a cold manner, but there are also exemplars of more sustained anger; sadness vocalizations vary between a quiet and a mild form; and for fear, vocalizations vary between milder states and panic.

The vocalizations were recorded in the sound-insulated booth of the Speech Laboratory at University of Porto, using Pro Tools LE version 5.1.1 (Digidesign, Avid Technology) software and a high-quality microphone attached to an Apple Macintosh computer. Digitization was done at a 48- kHz sampling rate and 16-bit resolution” But the ones I actually found were at 44.1 kHz.

“Interparticipant reliability in ratings on emotion scales was also very high (Cronbach’s α = .966), further indicating that our set of vocalizations produces reliable responses.”

“We also extracted a derived measure of accuracy from the raw ratings. For each vocalization, when the highest of the eight ratings was provided on the “correct” scale, the response was considered as a correct categorization; otherwise, the response was considered as an incorrect categorization. These rates are depicted in the last column of Table 2. Such a derived accuracy index has been used in previous studies on emotion recognition”

“Interparticipant reliability was very high for valence (α = .982) and arousal ratings (α = .95); it was lower, although satisfactory, for authenticity ratings (α = .836).”

“We herein produced and validated a well-controlled corpus of nonverbal vocalizations, which we make available for future research on emotion processing. It includes 121 sounds, recorded by four different speakers, and represents similarly positive and negative emotion categories: achievement/ triumph, amusement, sensual pleasure, relief, anger, disgust, fear, and sadness. These vocalizations elicited high emotion recognition accuracy in a forced choice task (86 %, on average) and were consistently rated as communicating the intended emotions in a rating task. Perceptual data include also details concerning perceived valence, arousal, and authenticity for each vocalization. Furthermore, we have shown that acoustic cues alone contain sufficient information to automatically classify vocalizations’ emotion category and to predict listeners’ behavioral responses. Although the validation procedure was mostly based on young women, we have established that the vocalizations can be accurately recognized by both male and female listeners of different ages:”

I discarded four samples: relief\_MS\_13, relief\_MS\_14, relief\_MS\_15, fear\_T\_16. The perceived valence was cross-aligned for these samples. Unlike calm in MESS, I don’t have justification to recode these as neutral

# LEGOv2

All the audio files from 20061122 look funny; not sure if they are usable

Some of these audio clips are very short, not enough room for a full utterance or even a word

Abstract “Data from the Let’s Go Bus Information System from the Carnegie Mellon University in Pittsburgh has been formatted, parameterized and annotated with quality, emotion, and task success labels containing 347 dialogs with 9,083 system-user exchanges.”

Text

Description automatically generated

“We further introduce the negative emotional state of the user that is manually annotated by a human rater who chooses one of the labels garbage, non-angry, slightly angry, very angry for each single user turn. From all 4,832 user turns, 68.5% were non-angry, 14.3% slightly angry, 5.0% very angry and 12.2% contained garbage, i.e., non-speech events.”

Angry + slightly and very angry samples as angry samples: 797 angry samples

I count the elicitation context as discourse.

From readme: “Annotation Scheme: friendly, neutral, slightly angry, angry, very angry” + garbage

Although the paper didn’t mention it, “friendly” is a good hint for positive affect.

# MAV

Francophone

Abstract: “The Montreal Affective Voices consist of 90 nonverbal affect bursts corresponding to the emotions of anger, disgust, fear, pain, sadness, surprise, happiness, and pleasure (plus a neutral expression), recorded by 10 different actors (5 of them male and 5 female). Ratings of valence, arousal, and intensity for eight emotions were collected for each vocalization from 30 participants. Analyses revealed high recognition accuracies for most of the emotional categories (mean of 68%).”

“The goal of the present contribution is to make available to the research community a validated set of auditory stimuli designed as an auditory counterpart of the Ekman faces. The Montreal Affective Voices (MAV) consist of 90 short, nonverbal affect bursts corresponding to portrayed expressions of anger, disgust, fear, pain, sadness, surprise, happiness, and sensual pleasure (plus a neutral expression) recorded by 10 different actors. Emotional categories of pain and sensual pleasure were added to the six basic emotions of the Ekman faces because they are especially relevant on an evolutionary or welfare level. The variability in segmental structure was minimized by asking the actors to produce these vocalizations on the vowel /ɑ/ (as in apple). This set of vocalizations was validated on the basis of ratings of valence, arousal, and perceived intensity along the eight discrete emotions provided by a group of 30 decoders (participants).”

Twenty-two amateur or professional Francophone actors

“Vocalizations were recorded in the sound-proof room of the Vocal Neurocognition Laboratory (University of Montreal), using a UMT800 condenser microphone (Microtech Gefell) at a distance of approximately 30 cm. Recordings were preamplified using a Millennia Media HV-3B preamplifier and were digitized at a 96-kHz sampling rate and 16-bit resolution, using an Audiophile 2496 PCI soundcard (M-Audio). They were then edited in short, meaningful segments and normalized peak value (90% of maximum amplitude) and were downsampled at 44.1 kHz, using Adobe Audition (Adobe Systems, Inc.). For each actor and vocalization category, only the best occurrence, rated as successful display by the experimenter, was kept for the validation stage.”

“The 5 male and 5 female actors who produced the most successful displays, based on judgments from the 29 remaining participants, were selected out of the 22 actors. Vocalizations from those 10 actors were then included in further acoustical and statistical analyses.”

Table

Description automatically generated

“Physical characteristics are reported for the ten selected actors. The first column indicates the actor's label, which is found at the beginning of the filename for each of the seven sounds produced by an actor. M; male; f: female. Weight is indicated in pounds.”

Mean perceived valence for surprise samples was 40 for male actors and 38 for female actors. Surprise should be negative for this one.

I have 90 sound files, as expected.

# MELD

Data are clips from *Friends*

This dataset allows for negative surprise and positive surprise. What the authors describe here as sentiment closely matches the idea of valence.

“The MELD dataset has evolved from the EmotionLines dataset developed by Chen et al. (2018). EmotionLines contains dialogues from the popular sitcom Friends, where each dialogue contains utterances from multiple speakers. EmotionLines was created by crawling the dialogues from each episode and then grouping them based on the number of utterances in a dialogue into four groups of [5, 9], [10, 14], [15, 19], and [20, 24] utterances respectively. Finally, 250 dialogues were sampled randomly from each of these groups, resulting in the final dataset of 1,000 dialogues.”

“The utterances in each dialogue were annotated with the most appropriate emotion category. For this purpose, Ekman’s six universal emotions (Joy, Sadness, Fear, Anger, Surprise, and Disgust) were considered as annotation labels. This annotation list was extended with two additional emotion labels: Neutral and Non-Neutral.” I didn’t find any labeled “Non-Neutral”

5 mTurk raters

“We format the audio files as 16-bit PCM WAV files for further processing.”

“These scenarios justify both context and multimodality to be important aspects for emotion recognition in conversation.” <- context and multimodality will be unavailable to me

“...we use seven emotions for the annotation, i.e., anger, disgust, fear, joy, neutral, sadness, and surprise, across the training, development, and testing splits (see Table 4). It can be seen that the emotion distribution in the dataset is expectedly non-uniform with the majority emotion being neutral. We have also converted these fine-grained emotion labels into more coarse-grained sentiment classes by considering anger, disgust, fear, sadness as negative, joy as positive, and neutral as neutral sentiment-bearing class. Surprise is an example of a complex emotion which can be expressed with both positive and negative sentiment. The three annotators who performed the utterance annotation further annotated the surprise utterances into either positive or negative sentiment classes.”

“Multiple infrequent speakers (< 1% utterances) are grouped as Others.”

Table

Description automatically generatedChart, bar chart

Description automatically generated

Some actual dev files are missing from dev\_sent\_emo.csv

Train file count matches

“non-neutral” was not present in datasets.yaml. I think the authors recoded the EmotionLines data with multimodal context.

5 test files were present in the dataset but missing from datasets.yaml:

MELD.Raw/test/output\_repeated\_splits\_test/dia93\_utt5.mp4

MELD.Raw/test/output\_repeated\_splits\_test/dia93\_utt6.mp4

MELD.Raw/test/output\_repeated\_splits\_test/dia93\_utt7.mp4

MELD.Raw/test/output\_repeated\_splits\_test/dia108\_utt1.mp4

MELD.Raw/test/output\_repeated\_splits\_test/dia108\_utt2.mp4

All are in the test split

All are from dialogue #93 (x3) or #108 (x2)

I compared the records from ./test\_sent\_emo.csv and ./ MELD.Raw/test/test\_sent\_emo.csv for dialogues #93 and #108 and the were the same (but for some special punctuation characters). In fact, the five files above were missing from both as well.

Here’s what’s up with these

dia108\_utt1 and dia108\_utt1 are identical: I dunno Monica it feels funny just being here.

They also match dia108\_utt0, which was not missing. The text transcription of the utterance is incorrect in datasets.yaml, but this is irrelevant for me since the emotion label seems fine.

Context: “When Monica buys a new bed from Janice's ex-husband's store, she doesn't get the bed she expected. Meanwhile, Rachel and Ross go out for dinner with her father, and Joey teaches a soap opera acting class.” <https://tvquot.es/friends/the-one-with-the-race-car-bed/>

“Phoebe: **I don't know, Monica. It feels funny just being here.** If you buy a bed from Janice's ex-husband, it's like betraying Chandler.

Monica: Not at these prices.”

A picture containing person

Description automatically generated Phoebe looks disgusted to me?

The valence label is definitely alright anyways.

Anyways, there shouldn’t be a problem with this sample.

dia93\_utt5 and dia93\_utt6 were identical. dia93\_utt7 had no audio.

They were also identical to dia93\_utt4: “I don’t want to talk to you, Wayne. I hate you. You ruined my life.” Joey sounds pretty angry:

A person with his arms crossed

Description automatically generated with medium confidence

The label for test/dia93\_utt4 in datasets.yaml is incorrect! On inspection, the test/dia93\_utt8 and test/dia93\_utt9 videos are identical. The actual dialogue in this self-same footage is the concatenation of Joey’s utterances from dia93\_utt8 and dia93\_utt9 in datasets.yaml (plus a little Chandler speech not noted). dia93\_utt8 dia93\_utt9 of the test split are both neutral (emotion and valence) Joey utterances.

Some of the timestamps in this dataset aren’t great. For example, some samples contain more laugh track than any speech audio although the corresponding datasets.yaml entry says there should be words.

test/dia93\_utt10 is one of those that only has laugh track. The utterance transcription is not present in the audio.

Actions taken:

dia93\_utt4 manually recoded: emotion ang, sentiment negative, speaker Joey

dia93\_utt9 and dia93\_utt10 specifically discarded

some of the samples in this dataset may have multiple speakers saying the same thing in unison. I kept these since ostensibly the valence is expressed in unison too

I manually annotated speaker gender. Luckily, I’ve seen the show and can Google a lot of the characters. A lot of the names are unambiguously gendered anyways

e.g. <https://friends.fandom.com/wiki/Dr._Harad>

“Dr. Leedbetter” is a typo for “Dr. Ledbetter” (speakers)

The Stripper is from s04e22 (F)

These laugh tracks might be problematic.

“a waiter” and “the waiter” are the same speaker

I’m generally assuming that the same speaker label appearing in the same episode is actually the same speaker.

Test/dia71\_utt1 and Test/dia71\_utt2 videos are identical. dia71\_utt2 has the correct annotation in datasets.yaml, so Test/dia71\_utt1 is manually excluded.

I had to merge some identities like “Ross and Joey” with “Joey and Ross”.

Some of the video begin/end timings are so off they ruin the sample. Consider train/dia503\_utt10. Sentiment and emotion match the text and speaker. But the actual mp4 contains the next speaker’s utterance, which is clearly negatively valenced. I didn’t go searching for these, but this one I manually omitted.

Train/dia715\_utt0 is mislabeled. Phoebe: “Thank you! Thanks, Mon. Oh but Mon, if you touch my guitar… [one more time…]” is the actual audio. It’s mostly positive, but it looks like it’s transitioning to negative. This is too ambiguous, so I discard it.

5 dev files were present in the dataset but not in datasets.yaml:

MELD.Raw/dev\_splits\_complete/dia49\_utt4.mp4

MELD.Raw/dev\_splits\_complete/dia49\_utt5.mp4

MELD.Raw/dev\_splits\_complete/dia66\_utt9.mp4

MELD.Raw/dev\_splits\_complete/dia66\_utt10.mp4

Dev/dia66\_utt9 and dev/ dia66\_utt10 are darn near identical to dev/ dia66\_utt8, which is coded just fine. dia66\_utt9 and dia66\_utt10 are dropped

dia49\_utt4 and dia49\_utt5 are identical. It’s from an angry exchange between Ross and Susan:

ROSS: Please. This is so your fault.

SUSAN: How, how is this my fault?

ROSS: Look, Carol never threw me out of her room before you came along.

SUSAN: Yeah? Well, there's a lot of things Carol never did before I came along.

**ROSS: You tryin' to be clever?**

**SUSAN: You know what your problem is? You are so threatened by me.**

ROSS: Oh, I'm threatened by you?

SUSAN: Yes.

PHOEBE: Hey, hey, ok, all right, that's it!

That’s clearly angry talk that sounds angry.

action: dev/dia49\_utt4 manually coded: emotion ang, sentiment negative, speaker Ross and Susan (unique), gender u; dev/dia49\_utt5 ignored

there are two dia201\_utt1, one in train and another in test

dia4\_utt1 exists in all 3

train/dia4\_utt1 has usable no audio

the annotations for test/dia27\_utt0 and test/dia27\_utt1 are wrong. They are both of a clip of Joey mid-utterance saying “badges” excitedly. The valence is too ambiguous for me, so I’m discarding both of these.

Every speaker named “All” is treated as a new speaker per clip

There are files named like final\_videos\_testdia1\_utt0.mp4 in the test folder. I ignored these since I don’t have documentation/labels on them.

train/dia125\_utt3 is corrupted

# MESS

Angry, happy, calm, sad

Calm is pleasant here

“All recordings had a 36-kHz sampling rate, which maintained signal fidelity of spectral information through 18 kHz (sufficient for audible speech information).” I looked and the wav files I have are 24.4kHz.

“The 1,080 Theo–Victor–Michael (TVM) sentences used by Helfer and Freyman (2009) were used in the present work. These sentences were originally designed to investigate effects of speech-on-speech auditory processing, which requires a cue name to which the listener may be primed to identify a target talker amidst other distracting talkers using similar speech materials. For example, a listener would be instructed to listen to the sentence that starts with the cue name “Theo,” while three sentences are presented, each with a unique cue name. These sentences are structured as follows: “[Cue name] discussed the [word 1] and the [word 2] today,” where the cue name was replaced with the name “Theo,” “Victor,” or “Michael,” and the word placeholders were filled with one- or two-syllable common nouns. The specific structure of these sentences lends to experiments related to auditory perception, and these sentence features may be utilized in future research endeavors studying the effects of emotion on other aspects of auditory perception (e.g., competing speech paradigms with emotional targets and distractors).”

“All recruiting was carried out using procedures approved by the University of Utah Institutional Review Board (IRB). Six participants (three men, three women) for recording were recruited from the Department of Theater, University of Utah. Participants were all Caucasian young adult students (aged 19–21 years, M = 20 years, SD = 0.8 years) who were native speakers of American English and had completed all coursework in vocal production offered by their program. This population was targeted for recruitment to ensure that they had received some formal training on vocal emotion production and would be able to accurately and consistently produce the emotions desired for the study.”

“The talkers produced TVM sentences in four emotional styles. The emotion categories Angry, Happy, Sad, and Calm each represent a quadrant of the activation/pleasantness dimensional space, similar to the arousal/valence plane (Russell, 1980). Specifically, Angry is a high-activation, low-pleasantness emotion; Happy is a high-activation, high-pleasantness emotion; Sad is a low-activation, lowpleasantness emotion; and Calm is a low-activation, highpleasantness emotion. Previous studies typically employ Neutral as a fourth category instead of Calm; however, neutral speech has been demonstrated to contain a negative valence (Scherer, Banse, Wallbott, & Goldbeck, 1991), and so Calm was included in place of Neutral as an emotion that contrasts from the others in activation and pleasantness. Other recent work has also proposed the use of Calm as a reference emotion category similar to Neutral with which other emotions may be compared (Livingstone & Russo, 2018).”

“The 1,080 TVM sentences include 360 sentences for each talker of a given gender. The 360 sentences were divided into 90 sentences for each emotion, resulting in 30 sentences for each cue name, emotion, and talker (30 sentences × 3 cue names × 4 emotions × 3 talkers = 1,080 sentences).”

All talkers spoke American English:

“All recruiting was carried out using procedures approved by the University of Utah Institutional Review Board (IRB). Six participants (three men, three women) for recording were recruited from the Department of Theater, University of Utah. Participants were all Caucasian young adult students (aged 19–21 years, M = 20 years, SD = 0.8 years) who were native speakers of American English and had completed all coursework in vocal production offered by their program. This population was targeted for recruitment to ensure that they had received some formal training on vocal emotion production and would be able to accurately and consistently produce the emotions desired for the study. Preference was given in this instance to the experimental control of the emotional productions rather than to the naturalness of the stimuli (as could have been obtained with untrained talkers or spontaneous speech productions).”

91 Calm samples had valence ratings less than 50% (not by much, all within 6.25 points). (these were the only samples with valence mismatch) All of these Goodness of Fit >=60 as well as category\_accuracy and Percent\_C both >= 0.6; so these were intended as mild positive utterances yet were perceived as mild negative utterances and unambiguously as calm. Calm is usually treated as neutral or slightly positive in the literature. Rather than discarding these, I choose to keep them but re-label the valence as neutral

Chart, scatter chart

Description automatically generated

# oreau2

I’m using version 2 of this dataset, which contains only those that passed a perception check

I don’t think the speaker table at the end of Documentation.docx is fully accurate.

“32 speakers (8 female and 24 male), between the ages of 22 and 27, were participated in the design of Oréau database”

79 quotidian sentences

Actual file nomenclature:

<speaker\_gender>/sess<emo\_char>/<speaker\_##>a<utterance\_##><EMO\_CHAR>a.wav

Example: f/sessp/11a06Pa.wav

Speaker gender is f

Emotion is fear (peur)

Speaker #11

Utterance #6

“Utterances recognised better than 50% and judged as natural by listeners were retained, which constitutes the second version of database (OréauFR\_02).”

Surprise is negative for these

“Speech signals were first recorded at 44.1 kHz and then down sampled to the required 16kHz samples.” The ones I have are 44.1 Hz.

# ravdess

abstract: “The RAVDESS is a validated multimodal database of emotional speech and song. The database is gender balanced consisting of 24 professional actors, vocalizing lexically-matched statements in a neutral North American accent. Speech includes calm, happy, sad, angry, fearful, surprise, and disgust expressions, and song contains calm, happy, sad, angry, and fearful emotions. Each expression is produced at two levels of emotional intensity, with an additional neutral expression. All conditions are available in face-and-voice, face-only, and voice-only formats. The set of 7356 recordings were each rated 10 times on emotional validity, intensity, and genuineness. Ratings were provided by 247 individuals who were characteristic of untrained research participants from North America. A further set of 72 participants provided test-retest data. High levels of emotional validity and test-retest intrarater reliability were reported. Corrected accuracy and composite "goodness" measures are presented to assist researchers in the selection of stimuli. All recordings are made freely available under a Creative Commons license and can be downloaded at <https://doi.org/10.5281/> zenodo.1188976.”

“Filename convention. Each RAVDESS file has a unique filename. The filename consists of seven two-digit numerical identifiers, separated by hyphens (e.g., 02-01-06-01-02-01-12. mp4). Each two-digit numerical identifier defines the level of a different experimental factor. The identifiers are ordered: Modality±Channel±Emotion±Intensity±Statement±Repetition±Actor.mp4 or .wav. The numerical coding of levels is described in Table 1. For example, the filename ª02-01-06-01-02-01-12.mp4º refers to: Video-only (02)±Speech (01)±Fearful (06)± Intensity normal (01)±Statement ªdogsº (02)±First repetition (01)±Twelfth actor, female (12).”

Table

Description automatically generated

I have access to only 03 audio-only modality

“The gender of the actor is coded by the actor's number, where odd numbered actors are male, even numbered actors are female.”

“Experimental design. The RAVDESS contains 7356 recordings of 24 actors (12 male, 12 female). All actors produced 104 distinct vocalizations, consisting of 60 spoken utterances and 44 sung utterances. Each of the 104 vocalizations was exported to create three separate modality conditions: audio-video (face and voice), video-only (face, but no voice), and audio-only (voice, but no face). This produced 312 files per actor (104 × 3). The song recordings of one female participant were lost due to technical issues (132 files). Thus, 24 × 312±132 = 7356 files. This set is composed of 4320 speech recordings and 3036 song recordings.”

“Actors vocalized two distinct statements in the speech and song conditions. The two statements were each spoken with eight emotional intentions (neutral, calm, happy, sad, angry, fearful, surprise, and disgust), and sung with six emotional intentions (neutral, calm, happy, sad, angry, and fearful). All emotional conditions except neutral were vocalized at two levels of emotional intensity, normal and strong. Actors repeated each vocalization twice.”

“This set is composed of 4320 speech recordings and 3036 song recordings”

“Twenty-four professional actors, working in Toronto, Ontario, Canada were hired for stimulus creation (M = 26.0 years; SD = 3.75; age range = 21±33; 12 males and 12 females). Actors self-identified as Caucasian (N = 20), East-Asian (N = 2), and Mixed (N = 2, East-Asian Caucasian, and Black-Canadian First nations Caucasian). To be eligible, actors needed to have English as their first language, speak with a neutral North American accent, and to not possess any distinctive features (e.g., beards, facial tattoos, hair colorings, facial piercings). Participants were also required to identify text presented at 1.5 m distance without wearing glasses.”

“The microphone track for each actor was peak-normalized to -3 dBFS using Adobe Audition CS6. Peak normalization was chosen to retain the natural variation in loudness between emotional conditions [95, 126, 133]. The singing audio track was imported into Melodyne for pitch adjustment to ensure that the three melodies remained perceptually distinct. Intervals are perceived as ªin tuneº when mistuned by up to 35 cents [134, 135], and ªout of tuneº when …”

The emotional valence of song utterances… there is an underlying philosophical question here I am sidestepping; if the samples have passed some kind of perception test, I’ll chalk that up as empirical validation

Use the previous literature to justify surprise as negative

“Two baseline emotions. Third, the RAVDESS includes two baseline emotions, neutral and calm. Many studies incorporate a neutral or ªno emotionº control condition. However, neutral expressions have produced mixed perceptual results [70], at times conveying a negative emotional valence [71]. Researchers have suggested that this may be due to uncertainty on the part of the performer as to how neutral should be conveyed [66]. To compensate for this a calm baseline condition has been included, which is perceptually like neutral, but may be perceived as having a mild positive valence. To our knowledge, the calm expression is not contained in any other set of dynamic conversational expressions, and is present in one static facial image set [29].”

^ Accordingly, calm is mapped to 0 valence since it was intended as a baseline emotion

Speakers recruited from Toronto

I might consider dropping the song files to see what difference it makes. Also, calm might be recoded to positive.

“Stimuli. Two neutral statements were used (ªKids are talking by the doorº, ªDogs are sitting by the doorº). Statements were seven syllables in length and were matched in word frequency and familiarity using the MRC psycholinguistic database [102]. For the singing trials, statements were associated with melodies that were sounded using piano MIDI tones of fixed acoustic intensity, consisting of six eighth notes (300 ms) and ending with a quarter note (600 ms). The tonality of melodies associated with each emotion was tailored to be consistent with emotional association [103, 104]. The melody associated with the positively valenced emotions calm and happy was in the major mode (F3, F3, A3, A3, F3, E3, F3). The melody associated with the negatively valenced emotions sad, angry, and fearful was in the minor mode (F3, F3, Ab3, Ab3, F3, E3, F3). The melody associated with neutral emotion did not contain the third scale degree (F3, F3, G3, G3, F3, E3, F3) and was designed to be ambiguous in terms of major or minor mode. The perceived valence of song melodies was validated in a separate a perceptual task. Eight participants (5 female, 3 male, mean age = 27.4, SD = 9.2), from Ryerson University, Toronto volunteered to participate. Raters had varied amounts of private musical instruction (mean = 9.0 years, SD = 7.1). Participants were asked to rate the perceived valence of each of the three melodies (major-mode, neutral, minor-mode), using a 9-point valence scale from the self-assessmentmanikin (SAM) [105]. Results confirmed that the major-mode melody (M = 7.88, SD = 1.13) was rated as more positive than the neutral melody (M = 5.13, SD = 1.55), which in turn was rated as more positive than the negative melody (M = 3.0, SD = 1.77).”

“Microphone output was recorded using Pro Tools 8 and a Digidesign 003 mixing workstation, at a sampling rate of 48 kHz, 16 bit, with files saved in uncompressed wave format.”

# savee

from audiodata/info.txt:

“

--------------------------------------

Speakers

--------------------------------------

'DC', 'JE', 'JK' and 'KL' are four male speakers recorded for the SAVEE database

--------------------------------------

Audio data

--------------------------------------

Audio files consist of audio WAV files sampled at 44.1 kHz

There are 15 sentences for each of the 7 emotion categories.

The initial letter(s) of the file name represents the emotion class, and the following digits represent the sentence number.

The letters 'a', 'd', 'f', 'h', 'n', 'sa' and 'su' represent 'anger', 'disgust', 'fear', 'happiness', 'neutral', 'sadness' and 'surprise' emotion classes respectively.

E.g., 'd03.wav' is the 3rd disgust sentence.

“

Timeline

Description automatically generatedA close-up of a document

Description automatically generated with low confidence

Can I use the .wmv files in MetaData? These appear to be TV/movie clips used as prompts for elicitation or evaluation. So they should be fine as additional samples.

disgust01.wmv is a 35-second dialogue (ultimately discarded) between two male speakers (Harold & Kumar?), who are the same speakers in fear04.wmv. The entire scene is arguably negatively valenced. You might say the interlocutors are expressing concurrent fear and disgust and anger and surprise. It’s not on the utterance level. I almost chose to keep it since the valence is consistent (and gender is unambiguous). Although disgust is the focal emotion, I wouldn’t label it as just that. Ideally, I would manually crop the video/audio by speaker turns. But honestly that’s just too much work. If I did it here, I would feel compelled to do the same for all the messy speaker splits from the MELD *Friends* dataset.

Anyways, the assumption is that short dialogue (between same-gendered interlocuters) with constant (and contiguous?) valence can be roughly treated as equivalent to some other unambiguously valenced (possibly ambiguous emotion category label) at the utterance level for valence detection in training and evaluation

sadness03.wmv contains no vocals. Discard

I added 7 unique speakers from the MetaData folder.

maybe I should only keep surprise if the valence is coded?

Maybe I could save the emotional samples with ambiguous valence, train on the others, predict on the ambiguous for pseudolabels, and then proceed with a juiced up dataset?

Nah, surprise should be default negative.

See noordewier & breugelmans 2013:

“we explored the possibility that surprise may be a (mildly) negative emotion in a study of autobiographical recall of unexpected and surprising events (Experiment 1) and in two studies of facial expressions of surprise (Experiments 2 and 3a/b). The studies show that experience of surprise and the perception of surprise in others may indeed be initially negative.”

See Neta, Davis, & Whalen 2011:

“We offer evidence that the default interpretation of surprise is negative, as participants were faster to detect surprised faces when presented within a happy context (Exp. 2). Finally, we kept the valence of the contexts constant (i.e., surprised faces) and showed that participants were faster to detect happy than angry faces (Exp. 3). Together, these experiments demonstrate the utility of the oddball paradigm to serve as an implicit context to resolve the valence ambiguity of surprised facial expressions, but that this implicit context does not completely override the default negativity.”

# ShEMO

“This paper introduces a large-scale, validated database for Persian called Sharif Emotional Speech Database (ShEMO). The database includes3000 semi-natural utterances, equivalent to 3 hours and 25 minutes of speech data extracted from online radio plays. The ShEMO covers speech samples of 87 native-Persian speakers for five basic emotions including anger, fear, happiness, sadness and surprise, as well as neutral state. Twelve annotators label the underlying emotional state of utterances and majority voting is used to decide on the final labels. According to the kappa measure, the inter-annotator agreement is 64% which is interpreted as “substantial agreement".”

From radio plays broadcast online: 4 www.radionamayesh.ir

“We segmented each stream into smaller parts such that each segment would cover the speech sample of only one speaker without any background noise or effect. We recruited 12 annotators (6 males, 6 females) to label the affective state of the utterances on a 7-point scale (including anger, fear, neutrality, happiness, sadness, surprise, and none of the above). The annotators were all native speakers of Persian with no hearing impairment or psychological problems. The mean age of the annotators was 24.25 years (SD = 5.25 years), ranging from 17 to 33 years.”

"Sharif Emotional Speech Database (ShEMO) is a large-scale semi-natural database for Persian which contains 3 hours and 25 minutes of speech data from 87 native-Persian speakers (31 females, 56 males). There are 3000 utterances in .wav format, 16 bit, 44.1 kHz and mono which cover five basic emotions of anger, fear, happiness, sadness and surprise, as well as neutral state. The utterances are extracted from radio plays which are broadcast online 4."

"We selected 50 radio plays of various genres including comedy, romantic, crime, thrilled and drama as potential sources of emotional speech. We balanced out the differences of the audio streams using a free open-source audio editor software application, named Audacity. Since most streams (about 90% of them) had a sampling frequency of 44.1kHz, we upsampled the streams which had a lower sampling rate using cubic interpolation technique. We also converted the stereo-recorded streams to mono."

“

The characters used in the label of the utterances and their corresponding meaning:

A: anger emotion

F: female speaker (if used at the beginning of the label e.g.F14A09) or fear (if used in the middle of the label e.g. M02F01)

H : happiness

M : male speaker

N : neutral

S : sadness

W : surprise

e.g. F03S02 F means the speaker is female, 03 denotes the speaker code, S refers to the underlying emotion of the utterance which is sadness, 02 means this is the second utterance for this speaker in sad emotion.

“

Iranian Persian, as the radio show website is .ir (All rights reserved to the Voice of the Islamic Republic of Iran) and the researchers are affiliated with an Iranian university

I’ll call this a discourse context. I have a feeling a lot of the datasets with recordings from discourse contexts already have some silence trimmed, however. But silence should be trimmed on both sides for standalone utterances and only on the left for utterances from discourse.

# tess

Toronto, Canada

2 female speakers

anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral

“A set of 200 target words were spoken in the carrier phrase "Say the word \_\_\_\_\_' by two actresses (aged 26 and 64 years) and recordings were made of the set portraying each of seven emotions (anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral). There are 2800 stimuli in total. Two actresses were recruited from the Toronto area. Both actresses speak English as their first language, are university educated, and have musical training. Audiometric testing indicated that both actresses have thresholds within the normal range.”

Surprise is “pleasant surprise”, so definitely positive for tess

Single words

# urdu

from readme: “URDU dataset contains emotional utterances of Urdu speech gathered from Urdu talk shows. It contains 400 utterances of four basic emotions: Angry, Happy, Neutral, and Emotion. There are 38 speakers (27 male and 11 female).

This data is created from Youtube. Speakers are selected randomly. Anyone can use this data only for research purposes.

Nomenclature followed while naming the files in the dataset is to provide information about the speaker, gender, number of the file for that speaker and overall numbering of the file in particular emotion. Files are named as follows:”

“We have collected the first custom dataset of spontaneous emotional speech in the Urdu language. The data consists of audio recordings collected from the Urdu TV talk shows. There are overall 400 utterances for four basic emotions: angry, happy, sad, and neutral. There are 38 speakers (27 males and 11 females). This corpus contains spontaneous emotional excerpts from authentic and unscripted discussions between different guests of TV talk show.”

“…video clips are collected from YouTube based on the discussion and situations going on in the talk shows. Although emotional corpus formulation from TV shows is an easy and abundant task but obtaining samples in different emotions is very difficult. Sometimes the presence of music and other noise accompanies make the data collection worse. We cannot have all the emotional state compared to other databases that are recorded by experts in studio condition, therefore, we only collected videos on four basic emotions (i.e., happy, sad, angry, and neutral). After collecting video clips, we mixed and gave these files to four students from NUST3 and CIIT4 universities. They were asked to annotate the data with emotional labels to the speakers’ state using both audio and video content. Final labels were given to each utterance when at least 2 annotators assigned them same emotion. This dataset is publicly available for research purposes5.”

My unique speaker count is different. I have 11 female speakers and 18 male speakers for 29 total. I have female speakers SF1–SF11 and male speakers SM1–7 and SM17–27, so I’m missing male speakers 8 through 16 (nine male speakers), however, utterance count is the same (400).

# vivae

zenodo page

"The Variably Intense Vocalizations of Affect and Emotion Corpus (VIVAE) consists of a set of human non-speech emotion vocalizations. The full set, comprising 1085 audio files, features eleven speakers expressing three positive (achievement/ triumph, sexual pleasure, and surprise) and three negative (anger, fear, physical pain) affective states, each parametrically varied from low to peak emotion intensity. The smaller core set of 480 represents a fully crossed subsample of the full set..."

"Recordings took place at the Berklee College of Music in Boston MA, USA. The VIVAE full set contains 1085 (i.e., all recordings that passed an acoustic quality check and were free of linguistic content). The core set was selected based on authenticity judgements collected throughout the technical validation of the dataset. The technical validation comprises data from 30 individuals providing emotion, intensity, and authenticity judgements. Details on the construction and perceptual evaluation will be shared in an academic paper. Please contact the authors for any questions at natalie.holz@ae.mpg.de (<mailto:natalie.holz@ae.mpg.de)>."

Surprise is meant to be positive here.

Human non-speech vocalizations

From the paper:

“Stimuli

The stimuli are 480 nonverbal vocalizations, representing the Core Set of a validated corpus48. The database comprises six affective states (three positive and three negative) at four different intensity levels (low, moderate, strong, and peak emotion intensity; note that in this text, the term “intensity” exclusively refers to the emotional intensity, i.e., the variation from a very mildly sensed affective state to an extremely intense affective state and should not be confused with the auditory perception of signal intensity as loudness). The six affective states—achievement/triumph, anger, fear, pain, positive surprise, sexual pleasure-represent a suitable, well-studied sample of affective states for which variations in emotion intensity have previously been described3,4,10.

Vocalizations were recorded at the Berklee College of Music (Boston, MA). Ten female speakers, all non-professional actors, were instructed to produce emotion expressions as spontaneously and genuinely as possible. No restrictions were imposed on the specific sounds speakers should produce, only that vocalizations should have no verbal content as in words (e.g., “yes”) or interjections (e.g., “ouch”). Following a technical validation, the Core Set was developed as fully crossed stimulus sample based on authenticity ratings. Stimuli were recorded with a sampling rate of 44.1-kHz (16-bit resolution). Sound duration ranges from 400 to 2000 ms.”

Zenodo page:

“The recordings are digitized at a 44.1-kHz sampling rate and 16-bit resolution.”

“Each of the 1085 VIVAE files has a unique filename...”

e.g. S04\_surprise\_peak\_10.wav

Speaker (S01 to S11).

Emotion (achievement, anger, fear, pain, pleasure, and surprise).

Emotional intensity (low, moderate, strong, peak).

Item-ID (unique integer identi…

“eleven speakers expressing three positive(achievement/ triumph, sexual pleasure, and surprise) and three negative (anger, fear, physical pain) affective states,each parametrically varied from low to peak emotion intensity.”

Definitely 11 speakers according to the actual files. S11 is female sounding. I have the full set.

From supp.:

“To compare ratings between the forced choice task and the emotion ratings, we report response patterns visualized in confusion matrices for each task (Figure S1a and b). The matrices display the cumulative frequency scores and rating scores, respectively, for each combination of expressed and perceived emotion at each intensity (low to peak and 1–7) across all trials. By visual inspection, similar patterns of concordant and discordant classification can be identified across tasks. These descriptive results are supported by a high positive correlation (Pearson’s r = .89, p < .001) between forced choice responses and emotion ratings across all possible pairs of expressed and perceived emotion.”