CREMA-D

Crowd-sourced Emotional Multimodal Actors Dataset

Cao, Cooper, Keutmann, Gur, Nenkova & Verma

IEEE Transactions on Affective Computing, 5(4), 377–390 (2014)

Section of Biomedical Image Analysis & Brain Behavior Laboratory, University of Pennsylvania

Seminar presentation · Part 2 of 2 · 40 minutes

Roadmap

35 minutes of material, then 5 minutes for discussion

1	**Bridge and motivation** The three gaps IEMOCAP left open	4 min
2	**Corpus specification** 91 actors, 12 sentences, 6 emotions, 4 intensity levels	7 min
3	**Recording protocol** How the clips were produced and screened	5 min
4	**The crowd-sourced perception study** 2,443 raters and three presentation modalities	7 min
5	**Results** Modality, emotion category, and perceived intensity	7 min
6	**Comparison and appraisal** CREMA-D against IEMOCAP; strengths and limits	5 min

Where we left off

Three things IEMOCAP could not tell us

Who is speaking

Ten actors, all English-speaking, all recruited from one university drama department. No purchase on how emotional expression varies with age, sex or background.

What is being said

Emotion and lexical content are entangled: people say different things when they are angry than when they are sad, so a model may be reading the words.

Which channel carries it

Every rater saw and heard each turn together. The corpus therefore cannot say how much emotion is in the voice and how much is in the face.

CREMA-D's answer to all three

Recruit ninety-one actors spanning ages twenty to seventy-four and five racial and ethnic groups. Fix the linguistic content to twelve sentences with deliberately neutral meaning, so that every emotion is carried by delivery alone. Then present each clip to raters three separate ways — voice only, face only, and both together — and measure how well each channel communicates.

The price: no dialogue, no partner, no context, and clips lasting a couple of seconds.

A corpus built for two audiences at once

Which explains several design choices that look odd from a machine-learning perspective

Affective computing

A large, balanced, freely redistributable training set for audio, visual and audio-visual emotion recognition — with far more speakers than anything then available, and with human perceptual judgements attached to every clip rather than just the director's intent.

Clinical neuroscience

A validated stimulus set for studies of social cognition. Emotion-recognition deficits are characteristic of schizophrenia and of several other conditions, and testing for them requires stimuli whose difficulty and perceptual properties are already quantified in a healthy population.

Why this matters for how we read the paper

A clinical stimulus set has requirements a training corpus does not: every item must be short, isolated and comparable to every other item; the linguistic content must not vary; and you need to know in advance how a healthy population perceives each item, so that a patient's response can be scored against it.

Those requirements produce exactly the design we are about to see. The absence of dialogue is not an oversight — it would make the stimuli unusable for the clinical purpose.

1 Corpus specification

What is actually in the 7,442 clips

The corpus in eight numbers

Everything is released openly under an Open Database Licence

7,442

audio-visual clips

91

actors — 48 male, 43 female

20–74

age range, in years

5

racial and ethnic groups represented

12

fixed sentences, identical across every actor

6

emotions: anger, disgust, fear, happy, neutral, sad

4

intensity levels: low, medium, high, unspecified

2,443

crowd raters, each judging 90 unique clips

Roughly five hours of speech in total, with an average clip length of about two and a half seconds — against IEMOCAP's twelve hours and four-and-a-half-second turns. CREMA-D is the smaller corpus by duration and the far larger one by speaker count. That inversion is the whole design.

Twelve deliberately boring sentences

Semantically neutral, so that emotion can only be carried by delivery

It's eleven o'clock	That is exactly what happened	I'm on my way to the meeting	I wonder what this is about
The airplane is almost full	Maybe tomorrow it will be cold	I would like a new alarm clock	I think I have a doctor's appointment
Don't forget a jacket	I think I've seen this before	The surface is slick	We'll stop in a couple of minutes

Why this matters methodologically

In any corpus of natural speech, what people say correlates with how they feel. A classifier can then reach high accuracy by reading the words rather than the voice — and it will fail the moment the vocabulary changes.

Holding the sentence set fixed removes that shortcut entirely. Any signal a model finds in CREMA-D audio is prosodic, spectral or articulatory. Nothing is lexical.

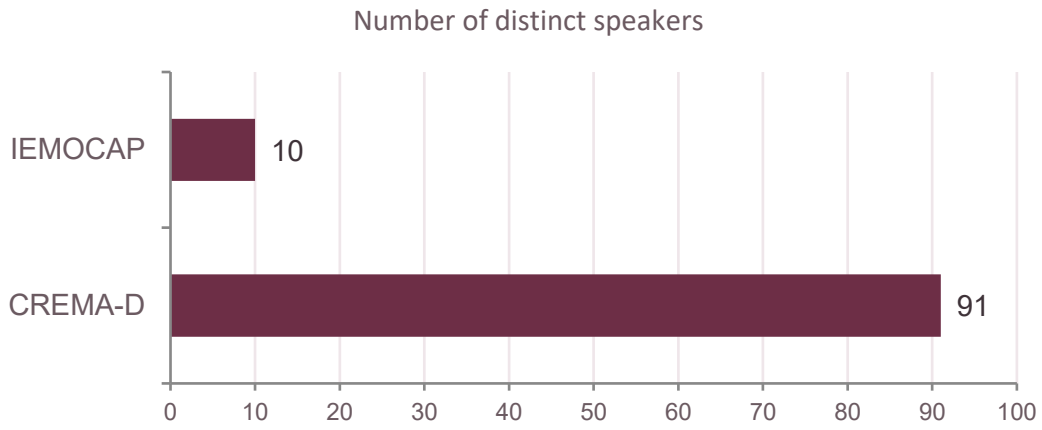
And what it costs

Twelve sentences is a vanishingly small slice of English. Models trained here see almost no phonetic or syntactic variety, so they may not transfer to open-vocabulary speech.

There is also a subtler risk: with the same twelve sentences repeated by every actor, a model can learn sentence identity and use it as a nuisance feature unless the splits are built carefully.

Speaker diversity is the headline contribution

Ninety-one actors, against ten in IEMOCAP



Composition

48 male and 43 female actors, aged from twenty to seventy-four, drawn from African American, Asian, Caucasian and Hispanic backgrounds plus an unspecified group. A handful of actors are over sixty, which makes the corpus one of very few emotional speech resources containing older voices at all.

- Why it matters: emotional expression is not demographically uniform. Vocal ageing changes pitch, jitter and speaking rate; conventions for displaying emotion vary across communities.
- A model trained on ten young actors from one drama school has no way to represent that variation, and no way to reveal that it cannot.
- Ninety-one speakers also make genuinely speaker-independent evaluation practical — you can hold out a substantial test group without gutting the training set.
- For the clinical use case this is not optional. A stimulus set for assessing patients must not be confounded with the demographics of a handful of performers.

Nine times the speakers on less than half the audio — breadth across people bought with duration per person.

Where 7,442 comes from

Intensity is a designed factor, not an afterthought

The recording plan per actor

Twelve sentences × six emotions = 72 clips at unspecified intensity.

One sentence — 'It's eleven o'clock' — was additionally recorded at low, medium and high intensity for each of the five non-neutral emotions, adding 10 further clips.

That gives 82 clips per actor, and 91 × 82 = 7,462 planned. The released set contains 7,442; a handful of recordings were lost or rejected.

Why one sentence carries the intensity design

Directing every actor through three graded intensities of six emotions across twelve sentences would have been prohibitively long and would have exhausted the performers.

Concentrating the intensity manipulation in a single, repeated sentence keeps everything else constant, so intensity can be studied cleanly — at the cost of studying it in only one phonetic context.

Filenames encode the design directly: actor ID, three-letter sentence code, three-letter emotion code, and intensity — so '1001_IEO_ANG_HI' is actor 1001 saying the eleven-o'clock sentence in high-intensity anger. The corpus is self-documenting.

Note the consequence for class balance: the corpus is close to uniform across the six emotions, with a modest surplus of the five non-neutral ones from the intensity clips. Compare IEMOCAP, where fear and disgust are under one percent each.

2 Recording protocol

How the clips were produced, directed and screened

Producing the clips

Studio recording with professional direction

Casting	Ninety-one professional actors recruited to span sex, age and ethnic background — the demographic spread was a recruitment target, not an accident.
Direction	Actors were given the target emotion together with a short contextual scenario for the sentence, then coached by a director, rather than simply told to sound angry.
Capture	Recorded in a controlled studio setting with consistent framing, lighting and microphone placement across all actors, so that clips are comparable as stimuli.
Takes and selection	Multiple takes were recorded; the team retained one clip per sentence-emotion-intensity cell, screening for technical faults and unusable performances.
Segmentation	Each clip is trimmed to a single sentence, roughly two and a half seconds long, with no surrounding dialogue or context.

Compare IEMOCAP: no partner, no unfolding dialogue, no rehearsal of a play — a single actor performing a fixed line to camera.

Two kinds of label, and the corpus ships both

A quiet but important methodological choice

Intended emotion

What the actor was directed to convey, together with the intensity level they were asked for. Available for every clip by construction, at zero annotation cost.

This is the label most acted corpora ship, and the one IEMOCAP's authors explicitly refused to trust.

Perceived emotion

What raters actually judged, separately for voice only, face only, and both together — plus a continuous intensity rating.

Expensive to collect, but it is the only thing that tells you whether a clip communicates what it was meant to communicate.

Why having both is more useful than having either

The gap between intended and perceived is itself the measurement. A clip the actor performed as fear and that nobody hears as fear is not noise to be discarded — it is evidence about how weakly fear is carried by the voice.

It also lets each user pick the label that suits their question. If you are training a recogniser of communicative signals, use the perceived label. If you are testing whether a patient can decode a signal that healthy raters decode reliably, you need both: the intended label as the key, and the perceived distribution as the difficulty calibration.

3

The crowd-sourced perception study

Two thousand raters, three presentation modalities

The rating design

Each clip is judged three separate times, once per modality

Voice only

Audio played, no video. The rater hears the sentence and nothing else.

Face only

Video played silently. The rater sees the facial expression with no vocal cue.

Audio-visual

Both channels together — the condition closest to ordinary perception.

- Every rater judged 90 unique clips: 30 in voice only, 30 in face only, 30 in audio-visual. No rater saw the same clip twice.
- For each clip they chose one of the six emotion categories and then rated how intense the emotion was.
- 2,443 raters × 90 clips ≈ 220,000 individual judgements, spread over 7,442 clips × 3 modalities ≈ 22,300 clip-modality items.
- That works out at roughly ten ratings per clip per modality; 95% of clips received more than seven.

Ten independent judgements per item is an order of magnitude more annotation effort per clip than IEMOCAP's three raters per turn — and it is what makes the perception results trustworthy.

Making crowd ratings trustworthy

The obvious objection to crowdsourcing, and how the paper answers it

Screening the setup

Raters in the audio conditions completed a sound check before starting, confirming they could hear target words correctly. Without it, a silent or broken playback would look exactly like a failure to recognise emotion.

Redundancy

Around ten independent raters per clip per modality means individual carelessness averages out, and it also yields a distribution over categories rather than a single vote — which is far more informative than a majority label.

Reported agreement

Inter-rater reliability is reported with Krippendorff's alpha, which handles many raters and incomplete rating matrices — the right choice here, where no two raters see the same set of clips.

Roughly where the agreement lands

Alpha of about 0.42 across the corpus — 'moderate' by conventional interpretation, and in the same broad territory as IEMOCAP's Fleiss kappa of 0.27.

Emotion perception is simply not a task on which people agree strongly. Any corpus claiming near-perfect agreement has probably restricted itself to unambiguously extreme displays.

A caution about comparing the two figures

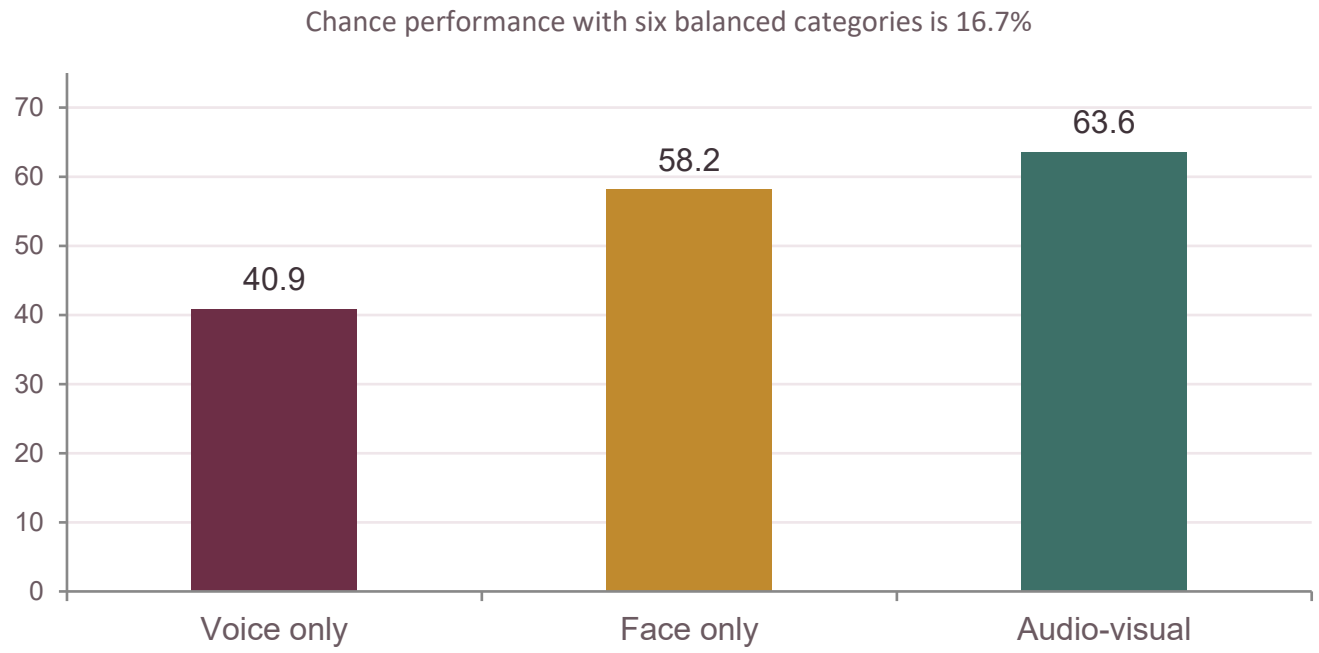
Kappa and alpha are different statistics, computed over different label sets — nine categories with multi-labelling allowed in IEMOCAP, six forced-choice categories here — and over different stimuli. Treat them as broadly comparable in magnitude, not as a ranking.

4 Results

What the perception study found

The headline result

How often raters recovered the emotion the actor intended



The face dominates

Seeing the expression alone recovers the intended emotion far more often than hearing the voice alone — a gap of some seventeen points.

But the voice still adds

Audio-visual beats face-only. If the face were doing all the work, adding sound would change nothing. Raters are integrating the two channels.

Read the absolute numbers carefully

Even with both channels, raters and actors agree less than two-thirds of the time. Emotional communication through short, context-free utterances is intrinsically lossy — which sets a realistic ceiling for what any recogniser on this corpus should be expected to achieve.

Not all emotions travel equally well

Reported ordering of recognition rates, from most to least reliably recognised

1	**Neutral**	The easiest category, and partly by construction: it is what raters fall back on when nothing distinctive is present.
2	**Happy**	Carried strongly in the face by the smile, and in the voice by raised pitch and energy.
3	**Anger**	High arousal and high energy make it salient in both channels.
4	**Disgust**	A face emotion above all — the characteristic nose and upper-lip movement has no clear vocal counterpart.
5	**Fear**	Frequently confused with surprise and with sadness; low-arousal fear in particular is hard to read.
6	**Sad**	The hardest category, and a standing problem in the field — sadness is low-arousal and quiet, so short clips carry little evidence.

The paper reports this ordering rather than a single headline figure per emotion, because the rates differ across the three modalities.

The finding that matters most

Different emotions live in different channels

Face-carried

Disgust is the clearest case. The characteristic configuration of nose, upper lip and brow is highly distinctive visually, while there is no comparably reliable vocal signature.

Give a recogniser audio only, and disgust becomes very hard.

Voice-supported

High-arousal states such as anger benefit substantially from the acoustic channel: loudness, pitch range and voice quality all shift markedly.

These are the emotions where adding audio to video yields the largest gain.

Weak in both

Sadness and fear are poorly recovered whichever channel is available, and adding the second channel does not rescue them.

Their signal is either too subtle for a two-second clip, or genuinely requires context.

Why this is the paper's most valuable contribution

No single-modality corpus can produce this result, and IEMOCAP — where every rating was made with both channels present — cannot produce it either. It takes a design in which the same clips are judged three separate ways by independent raters.

The practical consequence for modelling: a fusion architecture should not weight the two streams equally for every class, because the channels do not carry the classes equally. The optimal weighting is emotion-dependent, and this corpus is what lets you measure it.

Perceived intensity

The second thing every rater judged

What was collected

Alongside the categorical choice, raters gave a continuous intensity judgement — so every clip carries not only 'which emotion' but 'how strongly', in each of the three modalities separately.

Because one sentence was recorded at directed low, medium and high levels, perceived intensity can also be checked against intended intensity.

The main finding

Average perceived intensity is highest in the visual-only condition.

Seeing an expression without hearing it makes the emotion feel stronger than hearing it without seeing it — and, notably, stronger than the combined presentation.

- Intensity and category are separable. A clip can be confidently identified as sad and still be judged mild, or be ambiguous in category yet clearly intense.
- This gives CREMA-D something IEMOCAP provides through its dimensional annotation, but organised differently: intensity per modality, rather than valence and arousal from a general-purpose scale.
- For the clinical use case it is essential — assessment items need graded difficulty, and a low-intensity fear clip is a harder item than a high-intensity one.
- For modelling it enables ordinal and regression targets, and it supports curriculum-style training where easy, high-intensity items come first.

5

Comparison and appraisal

The two corpora side by side, and what each is good for

IEMOCAP and CREMA-D side by side

Two coherent answers to the same design problem

	IEMOCAP (2008)	CREMA-D (2014)
Speakers	10 actors, one drama department	91 actors, ages 20–74, five ethnic groups
Volume	~12 hours, 10,039 turns	~5 hours, 7,442 clips
Unit	Dialogue turn, ~4.5 s, in context	Isolated sentence, ~2.5 s, no context
Elicitation	Dyadic plays and improvised scenarios	Directed portrayal of a fixed sentence
Linguistic content	Free — varies with emotion	Fixed — 12 neutral sentences
Emotion classes	9 rated; 4 usable in practice	6, near-balanced, all usable
Intensity	Dimensional scales on a subset	Designed factor plus perceptual ratings
Raters per item	3 categorical, 2 dimensional	≈10 per clip, per modality
Modality separation	None — always audio-visual	Voice, face and audio-visual rated apart
Label shipped	Perceived only	Intended and perceived
Access	Licence agreement with USC	Open Database Licence, public repository
Official splits	None	None

The last row is the one thing neither corpus got right — and the reason results on both are so hard to compare across papers.

What CREMA-D gets right

Assessed on the same four axes we used for IEMOCAP

	Design claim	Verdict
Scope	Ninety-one speakers across a wide age and ethnic range, six balanced emotion classes, three modalities.	*Strongest of any comparable corpus of its period, and still unusual today.*
Naturalness	Directed portrayal with a scenario prompt, by professional actors, in a studio.	*The weakest axis. This is portrayal, and closer to the acted corpora IEMOCAP criticised.*
Context	Single sentences, no interlocutor, no discourse history.	*Deliberately sacrificed — and defensible only because the clinical use requires it.*
Descriptors	Intended and perceived category, plus perceived intensity, in three modalities, from about ten raters each.	*Exceptionally rich. The per-modality rating design is genuinely novel.*

Limitations

Some acknowledged, some structural, some only visible with hindsight

Portrayal, not experience

Actors performing on cue in a studio. Whatever IEMOCAP's dyadic elicitation bought, CREMA-D gives back.

No context whatsoever

Two seconds, one sentence, no interlocutor. Emotions that depend on discourse — frustration above all — are simply absent.

Twelve sentences

Almost no phonetic or syntactic variety. Transfer to open-vocabulary speech is untested and cannot be tested here.

No dimensional annotation

Category plus intensity, but no valence, arousal or dominance scales, so the corpus does not serve dimensional prediction directly.

Six forced-choice categories

No 'other', no blends, no multi-labelling. Raters must place ambiguous displays into one of six boxes, which inflates the apparent cleanliness of the labels.

Still no official splits

The same problem as IEMOCAP. Worse here, because the twelve repeated sentences make careless splits leak in an extra way.

How CREMA-D is used today

Twelve years on, it is one of the two default benchmarks

Audio-visual fusion

The canonical testbed for multimodal emotion models, precisely because the per-modality human baselines are published alongside the data.

Speech model evaluation

A standard downstream task for self-supervised speech representations, appearing in emotion-recognition benchmark suites.

Cross-corpus transfer

Often paired with IEMOCAP to test generalisation — train on one, evaluate on the other, and watch the accuracy fall.

Generative work

Used to train and evaluate emotional talking-face synthesis, where the intensity labels provide a control signal.

The cross-corpus point is worth dwelling on

Models trained on one of these corpora and evaluated on the other lose a great deal of accuracy. That is not a flaw in either resource — it is a measurement of how much of what we call 'emotion recognition' is really corpus recognition: the elicitation method, the recording chain, the actors, the label set.

Reading the two papers together is what makes that number interpretable rather than merely disappointing.

What the pair of papers teaches

The synthesis for the seminar as a whole

There is no neutral corpus design

Every choice about elicitation, unit length, label set and rater pool writes an assumption into the data. Those assumptions become the ceiling on what any model trained there can learn.

Control and naturalness are a dial

IEMOCAP turns it toward naturalness and pays in speaker count and confounded lexical content. CREMA-D turns it toward control and pays in context. Both positions are defensible; neither is free.

Annotation is measurement

Agreement in the fair-to-moderate band is normal for emotion. Report it, design around it, and treat human performance as the realistic ceiling rather than a hundred percent.

Three practical rules to take away

Match the corpus to the claim. If you want to argue that a model recognises emotion in speech, a corpus with twelve fixed sentences cannot support the claim; if you want to argue it works across speakers, ten actors cannot support it either.

State your split, your class merges and your discards, every time. On both of these corpora those choices are worth more than most modelling contributions.

Report cross-corpus results. Within-corpus accuracy measures how well you have fitted one elicitation protocol; the gap between corpora is the honest estimate of what you have actually built.

Questions for discussion

1. CREMA-D ships both the intended and the perceived label. If the two disagree for a given clip, which one belongs in your training set — and does the answer change for a recogniser versus a clinical assessment?

2. Fixing the sentence set removes the lexical confound, but a model trained on twelve sentences has seen almost no language. Is that a fair trade, or does it just relocate the problem?

3. The face beats the voice by a wide margin. What does that imply for the many papers reporting audio-only emotion recognition — are they solving a harder problem than they claim, or a different one?

4. Both corpora sit in the fair-to-moderate agreement band. Should emotion recognition be evaluated against a majority label at all, or against the full rater distribution?

5. If you were designing a corpus today with the benefit of both papers, which of the two designs would you start from, and what is the first thing you would change?

Cao et al. (2014), IEEE Transactions on Affective Computing 5(4), 377–390 · Busso et al. (2008), Language Resources and Evaluation 42(4), 335–359