IEMOCAP

Interactive Emotional Dyadic Motion Capture Database

Busso, Bulut, Lee, Kazemzadeh, Mower, Kim, Chang, Lee & Narayanan

Language Resources and Evaluation, 42(4), 335–359 (2008)

Speech Analysis and Interpretation Laboratory (SAIL), University of Southern California

Seminar presentation · Part 1 of 2 · 40 minutes

Roadmap

35 minutes of material, then 5 minutes for discussion

1. **Motivation and the 2008 landscape** — 6 min
   Why existing emotional corpora were not usable
2. **Corpus design** — 7 min
   Dyads, plays, and improvised scenarios
3. **Recording and post-processing** — 6 min
   Motion capture rig, segmentation, transcription
4. **Emotional annotation and reliability** — 8 min
   Categorical, dimensional, and how well raters agree
5. **Critical appraisal** — 4 min
   What the design buys and what it costs
6. **Legacy and modern use** — 4 min
   The de facto SER benchmark — and its pitfalls

Why we are reading a 2008 corpus paper

IEMOCAP is still the default benchmark for speech emotion recognition

~12 h
of synchronised audio, video and motion capture

10,039
dialogue turns, hand-segmented and professionally transcribed

61
reflective markers on face, head and hands

3,000+
citations — the most used multimodal emotion corpus

The claim the paper is making

Emotion is carried jointly by speech and gesture, so a corpus that captures only one channel cannot support models of expressive human communication. The contribution is not an algorithm — it is a resource plus a defensible methodology for eliciting emotion that is neither exaggerated acting nor uncontrollable real-world recording.

Why it still matters to us

Almost every speech-emotion result you will read this year — including results from large self-supervised speech models — is reported on IEMOCAP. If you do not understand how these labels were produced, you cannot interpret those numbers.

1 Motivation and the 2008 landscape

Why the authors concluded that no existing corpus would do

The problem, as the authors framed it

Following Douglas-Cowie et al. (2003), a corpus is judged on four axes

Scope

How many speakers, which emotions, which language, how much data?

Most corpora: a handful of speakers, a few hundred utterances.

Naturalness

Acted portrayal versus genuinely felt emotion.

Most corpora: actors told to 'be angry', producing caricature.

Context

Isolated sentences versus emotion unfolding in discourse.

Most corpora: read sentences with no interlocutor and no history.

Descriptors

How rich and how reliable is the emotional annotation?

Most corpora: the intended emotion used directly as the label.

The bind

Acted corpora give control but discard the blends and ambiguity of real affect — recognisers trained on them collapse in deployment. Broadcast and call-centre corpora give naturalness but bring copyright and privacy restrictions, uncontrolled microphone and camera placement, and wildly unbalanced emotion classes. Neither route yields synchronised, high-resolution gesture data.

The elicitation dilemma

Every option on the table in 2007 failed at least one requirement

Approach	What it gets right	What it gives up
Read acted speech	Full control of text and emotion	Portrayal, not experience; caricature
Broadcast / TV corpora	Genuinely felt, spontaneous affect	Copyright, privacy, uncontrolled A/V
Call-centre recordings	Real stakes, real frustration	Audio only, unbalanced classes
Wizard-of-Oz induction	Emotion induced, not portrayed	Weak, narrow emotional range
Recall of past experience	Authentic emotional memory	Monologue; no interaction

The authors' bet: skilled actors, engaged in interpersonal drama with a real partner, produce something closer to genuine emotion than any of the above — while keeping the recording controllable.

Six design requirements

Stated explicitly in Section 2 — the paper is unusually clear about this

Genuine realisations

Not simulated emotion on command

Dialogue, not monologue

Emotion elicited inside a conversation

Many experienced actors

Trained performers, plural, for generality

Controlled content

Emotional and linguistic content known in advance

Synchronised multimodal capture

Detailed gesture data alongside audio

Labels from human judgement

Perceived emotion, not the director's intent

Requirements 1 and 4 are in direct conflict

You cannot have fully spontaneous emotion and fully controlled content at the same time. The corpus resolves this by refusing to choose: half of it is scripted plays (control) and half is improvisation (naturalness). Almost every later analysis in the paper compares those two halves.

2 Corpus design

Ten actors, five dyads, three plays, eight improvised scenarios

The core idea: put the emotion inside an interaction

One participant is motion-captured; the other is a real conversational partner

- An emotion has a cause and an addressee. In a monologue it has neither.
- The partner supplies real-time uptake: interruption, sympathy, escalation.
- Dialogues run about five minutes, so emotion arrives with discourse history behind it.
- Raters later judge each turn in sequence, with that same context available.
- Actors reported the rig did not interfere with natural interaction.

Design consequence

Because the two speakers are in genuine dialogue, IEMOCAP supports research questions no single-speaker corpus can: entrainment, interruption, listener backchannels, and how one participant's affect drives the other's.

Two elicitation routes, deliberately kept separate

Roughly half the turns come from each

Scripted sessions · 5,255 turns

Three ten-minute plays, selected after reading more than a hundred candidates, each with one male and one female role so the data stays gender-balanced.

A theatre professional supervised selection to ensure the plays covered the five target emotions. Actors memorised and rehearsed the text.

Gives: identical linguistic content across sessions, so lexical effects can be controlled for. Emotion flows and changes as the narrative dictates.

Spontaneous sessions · 4,784 turns

Eight hypothetical scenarios, each assigning a target emotion to each partner. Actors improvised in their own words.

Scenario topics follow Scherer, Wallbott and Summerfield's survey of situations people actually report as emotionally significant — bereavement, separation, bureaucratic obstruction.

Gives: unrehearsed word choice, disfluency, genuine turn-taking. Costs: no control over what is said.

Target emotions designed into the material

Happiness · Anger · Sadness · Frustration · Neutral. Frustration was added to the usual four because it dominates real human–machine interaction. Four further categories were added later, at annotation time, to describe what the actors actually produced.

The improvised scenarios

Each scenario assigns a different target emotion to each partner

Scenario	Marked subject	Partner
Motor vehicle office rejects an application after an hour of queueing	Frustration	Anger
New parent deployed abroad; the couple must separate for a year	Sadness	Sadness
Telling a close friend about an engagement	Happiness	Happiness
Three years of unsuccessful job-hunting; hope running out	Frustration	Neutral
Airline loses a bag and offers a fraction of its value	Anger	Neutral
A close friend has died of cancer diagnosed a year earlier	Sadness	Neutral
Sharing news of acceptance to university with a best friend	Happiness	Happiness
Reaching a human operator after thirty minutes with a machine	Neutral	Anger

Note the asymmetry: in five of the eight scenarios the two partners are given different targets, so the corpus contains emotional mismatch, not just shared mood.

The actors and the sessions

Ten performers, five mixed-gender dyads

10

actors: 5 female, 5 male

7 + 3

professionals plus senior drama students (USC)

5

dyadic sessions, one actor and one actress each

≈6 h

per session, including rest periods

Selection and direction

Actors were chosen after reviewing audition tapes. An experienced professional acted as director throughout, checking that scripts were memorised and — critically — that the intended emotions were expressed genuinely rather than as caricature.

Why 'semi-natural', not 'natural'

The authors are candid: by Douglas-Cowie's taxonomy this is a semi-natural corpus, because actors may still exaggerate. Their argument is that the elicitation setting pushes it closer to natural than any previous acted corpus — an argument, not a demonstration.

Ten speakers is below the threshold Douglas-Cowie recommends. The authors acknowledge this directly and defend it as a first step — a limitation we will return to.

3

Recording and post-processing

Motion capture rig, synchronisation, segmentation, transcription

The capture rig

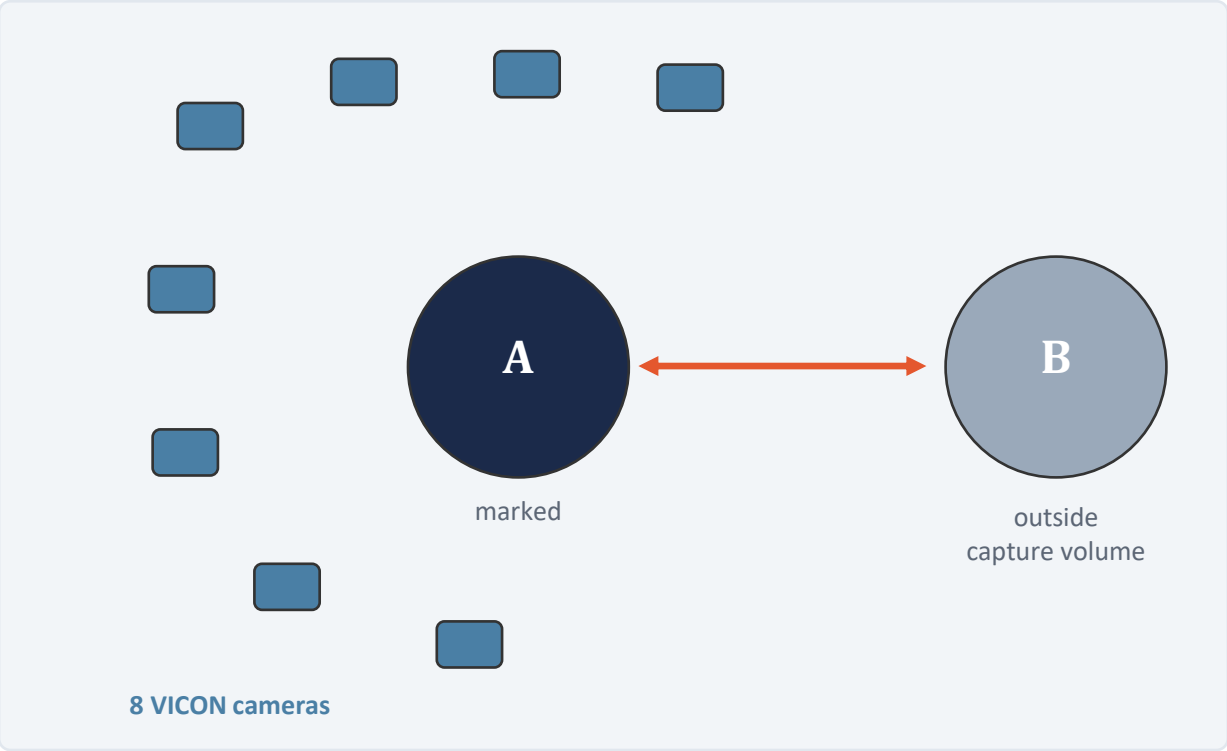
Everything is synchronised to a single reference

Motion capture	VICON system, 8 cameras placed about one metre from the marked subject, sampling at 120 frames per second.
Facial markers	53 markers roughly 4 mm across, positioned mostly on MPEG-4 facial feature points, spaced apart to aid reconstruction.
Head and hands	A headband with 2 markers compensates for head rotation; wristbands carry 2 markers each, plus one marker per hand — 61 markers in total.
Audio	Two Schoeps shotgun microphones, one aimed at each participant, recorded at 48 kHz.
Video	Two digital cameras capturing a semi-frontal view of each participant — these are what the annotators later watched.
Synchronisation	A clapboard fitted with reflective markers, visible to both the motion capture cameras and the video cameras, aligns every stream.

The subjects were seated so that gestures stayed inside the capture volume, but were told to gesture naturally and to keep their hands away from their faces.

Why only one subject wears markers

A constraint that shapes what the corpus can be used for

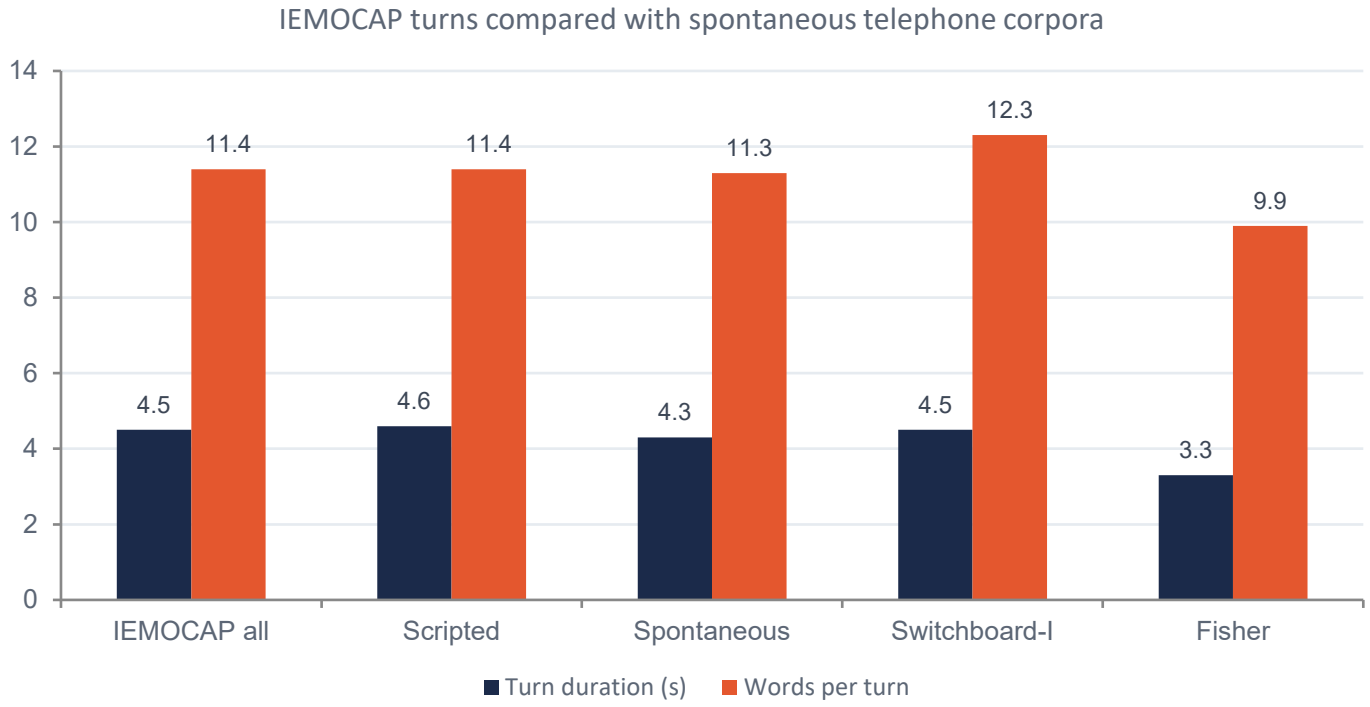


- The motion capture cameras respond to any reflective surface in view, so a second marker rig — plus its microphones, cameras and computer — would corrupt the recording.
- Aiming all eight cameras at one person also raises the spatial resolution of the capture.
- The consequence: the pair record the whole session, swap the markers, rest, and record it again.
- So the corpus has full motion capture for one side of every dialogue, and audio and video for both.

Practical implication: work using the motion capture channel has roughly half the data available to work using audio alone. Most published SER results on IEMOCAP quietly use only the audio.

Segmentation and transcription

Turns are the unit of analysis — and the unit of annotation



- 10,039 turns in total — 5,255 scripted and 4,784 spontaneous.
- A turn is a continuous stretch in which one actor is speaking. Short backchannels such as 'mm-hmm' were not segmented.
- Transcription was done professionally; word and phoneme boundaries came from forced alignment with Sphinx-III models trained on 360 hours of neutral speech.
- The comparison is the argument: IEMOCAP turns look statistically like real spontaneous telephone conversation, not like read speech.

Watch the spontaneous column: more than 20% of improvised turns are a single word — 'yeah', 'okay', 'what', 'right'. Backchannel-like turns are trivially easy to label neutral and inflate accuracy if you do not filter them.

4 Emotional annotation and reliability

Two annotation schemes, six raters, and an uncomfortable kappa

Categorical annotation

Labels come from perception, never from the actor's instruction

Who

Six evaluators, all fluent English speakers from USC. Sessions were arranged so that three different raters judged every turn. Evaluation ran in roughly 45-minute blocks with breaks.

How

The ANVIL annotation tool, with the video playing. Turns were judged in sequence so that raters had the audio, the visual channel and the preceding dialogue available — matching the context the actors themselves had.

What

Nine categories: anger, sadness, happiness, disgust, fear, surprise, frustration, excitement, neutral — plus a free-text 'other'. Multiple labels were allowed per turn to capture blends.

Label assignment: majority vote, and only when the top-voted category is unique.

74.6%

of all turns reached a unique majority label

66.9%

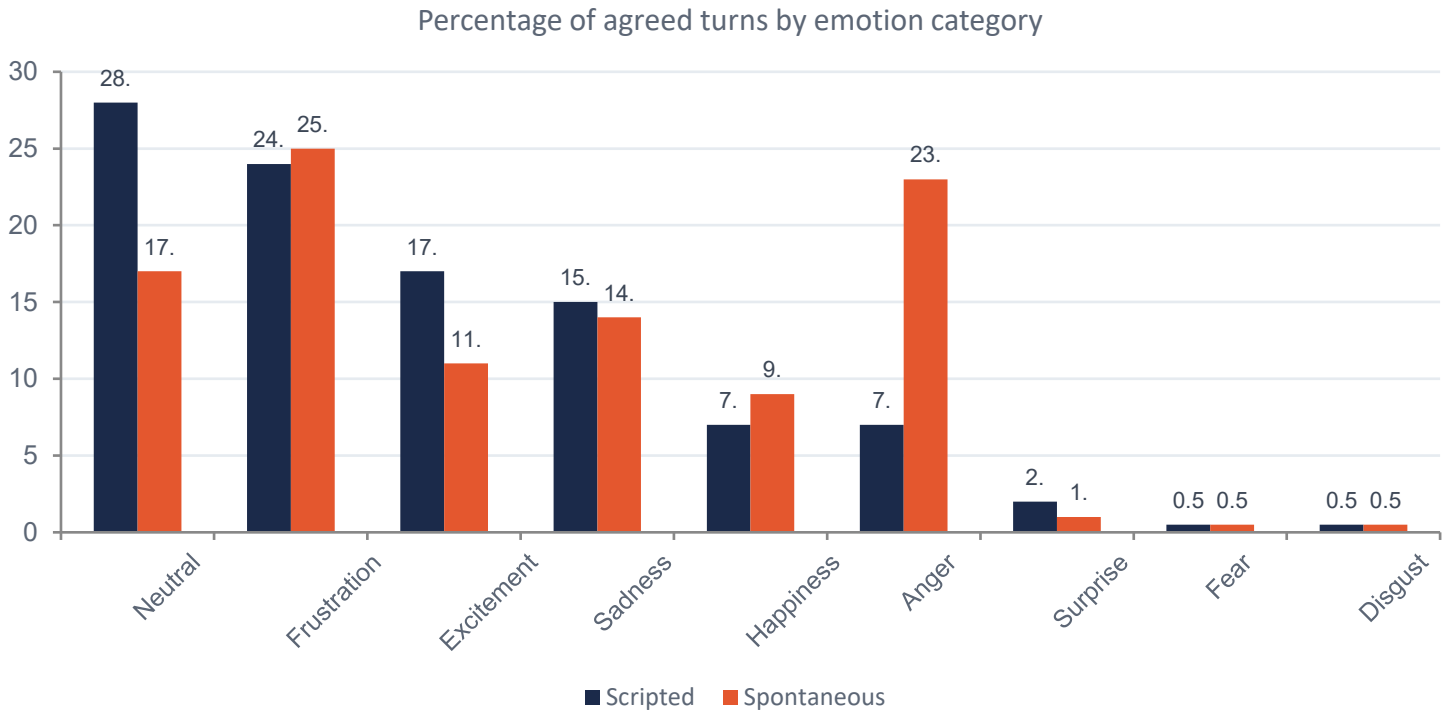
of scripted turns

83.1%

of spontaneous turns

What emotions the corpus actually contains

Distribution over turns that reached agreement; values below 1% shown as 0.5%

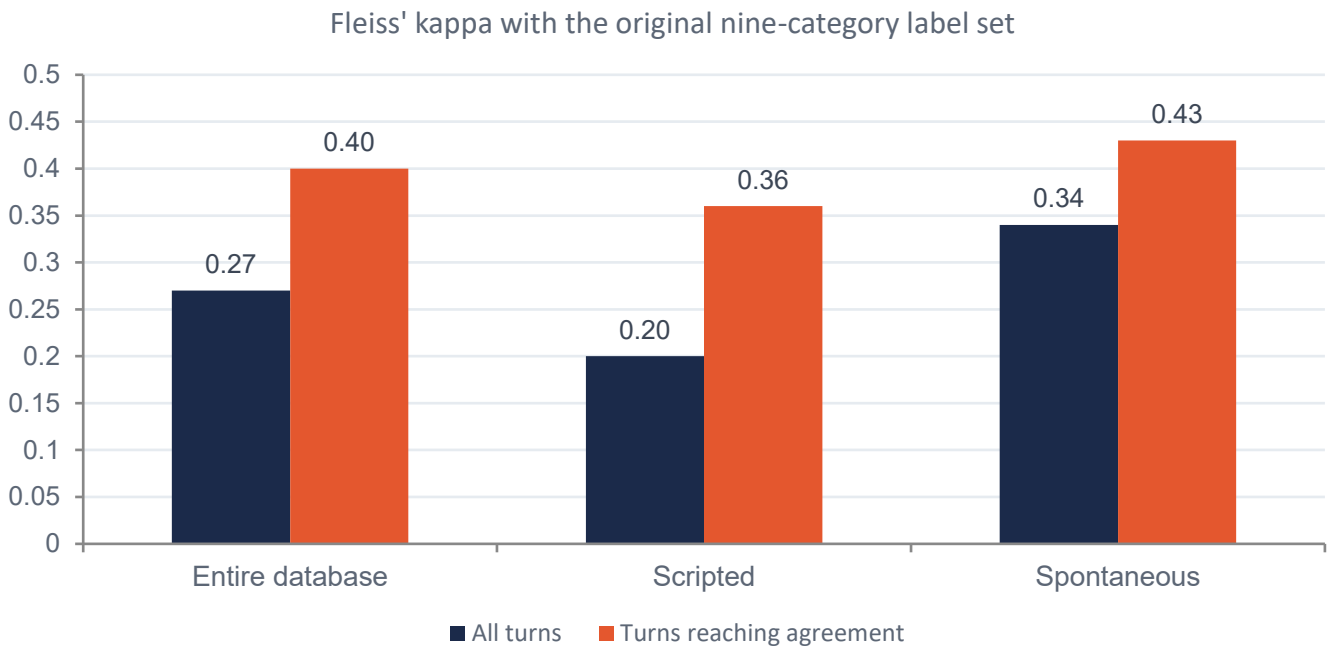


- The five target emotions dominate, as designed — the corpus is reasonably balanced across them.
- Fear and disgust are effectively absent. Nobody should train a fear detector on IEMOCAP.
- Anger is far more frequent in the improvised sessions (23% against 7%): the scenarios simply demanded it.
- Excitement is the reverse, and it is a category that was not planned for at all.
- Frustration is the most common non-neutral label in both halves.

Design determines contents: the emotion profile of this corpus is a profile of eight scenarios and three plays, not of human affect in general.

How much do the raters agree?

Fleiss' kappa, three raters per turn



Merging confusable classes helps

Collapsing happiness with excitement, and folding fear, disgust and surprise into 'other', raises kappa to 0.35 overall and 0.48 on agreed turns.

Landis and Koch scale

0.21–0.40 is 'fair'; 0.41–0.60 is 'moderate'. IEMOCAP sits at the boundary — which is typical for emotion annotation and is reported as such elsewhere in the literature.

The finding worth taking away

Improvised turns are consistently easier to agree on than scripted ones. The authors' explanation: a play moves gradually between emotional states as the narrative requires, producing ambiguous, blended displays, whereas each improvised scenario aims at one clear target.

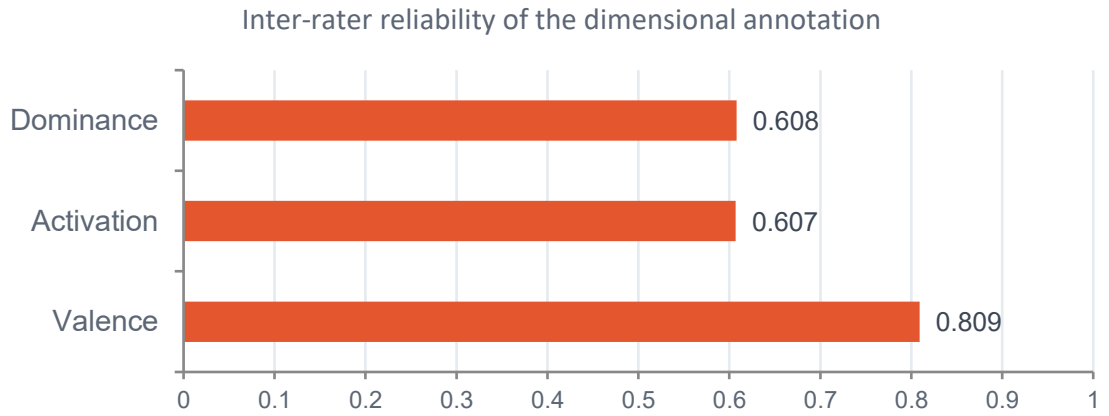
The second annotation scheme: dimensional attributes

Self-assessment manikins for valence, activation and dominance

Protocol

Two further raters scored each turn on three five-point scales using self-assessment manikins: valence from negative to positive, activation from calm to excited, dominance from weak to strong.

About 85.5% of the corpus had been scored at the time of publication. Scores were z-normalised per speaker to compensate for how differently individual raters use the scale.



- Why add a second scheme at all: categories say which emotion, but nothing about how intense it is. Two turns labelled 'anger' can be wildly different displays.
- Manikins are pictorial, so raters do not have to share a definition of a word like 'frustration' — a real advantage over categorical labels.
- Valence is markedly more reliable than activation or dominance. Judging pleasantness is easier than judging arousal from a short clip.
- This annotation is what makes IEMOCAP usable for regression-style emotion prediction, not just classification.

Two schemes, two purposes: categories tell you which emotion, dimensions tell you how strong. Most work uses the categories; the dimensional layer is why IEMOCAP also appears in valence and arousal regression papers.

Actors do not agree with their audience

Six actors re-evaluated their own improvised turns

0.79

average agreement when naive raters judge the turns

0.60

average agreement when the actors judge their own turns

- The ground truth here is the naive raters' majority vote, so their score is expected to be higher. Even so, a nineteen-point gap is striking.
- In follow-up work the authors found kappa fell when self-evaluations were added to the pool — the actors were the outliers.
- Actors were more selective in assigning labels: they knew what they meant to convey, and they knew the recording protocol.
- Individual variation is wide — one actress agreed with the raters on only 44% of her own turns.

Where human raters go wrong

Average agreement with the majority label: 72%

Frustration is the sink

Neutral, anger and disgust all leak into frustration — 13%, 17% and 17% of their instances respectively. It is the least distinct category in the corpus.

Happiness ↔ excitement

18% of happiness is read as excitement and 16% of excitement as happiness. The two differ mainly in arousal, not valence.

Sadness is comparatively safe

77% agreement, the highest of the emotional categories, and it leaks mostly into neutral.

Surprise is the weakest

65% agreement, spread across frustration and excitement — and with only about 2% of the data, it is not usable in practice.

This is the empirical basis for the conventions the field adopted

Because happiness and excitement are so heavily confused, most later work merges them into a single 'happy' class. Because fear, disgust and surprise are both rare and unreliable, they are dropped. What survives is the four-way task — angry, happy, sad, neutral — on which almost every IEMOCAP result is reported. That task is not defined anywhere in this paper; it emerged from these confusion patterns.

5 Critical appraisal

What the design buys, and what it costs

Scoring IEMOCAP on its own four axes

The authors' self-assessment, and whether it holds up

	Design claim	Verdict
Scope	Ten speakers, twelve hours, five well-populated emotion classes, three modalities in sync.	*Partly met. Ten is below the recommended minimum, and the authors say so.*
Naturalness	Emotion elicited through interaction with a partner, with a director suppressing caricature.	*Well argued but never measured. The planned naturalness study is not in this paper.*
Context	Five-minute dialogues; raters judged turns in sequence with the preceding discourse visible.	*Fully met, and it is the corpus's strongest claim over its predecessors.*
Descriptors	Categorical and dimensional annotation, both from human raters, plus full transcripts.	*Fully met — richer annotation than any comparable corpus of the period.*

Limitations

Some acknowledged in the paper, some visible only in hindsight

Ten speakers, one culture

All actors were English-speaking, recruited from one university drama department. Whatever a model learns about anger, it learns from ten Californians.

Still acted

Semi-natural by the authors' own classification. Portrayal under direction is not felt emotion, however good the elicitation.

The rig constrains behaviour

Subjects were seated, told to keep hands clear of the face, and kept inside the capture volume. Natural gesture is bounded by the equipment.

Two classes are unusable

Fear and disgust each account for under one percent. The corpus cannot support the six-way basic-emotion task at all.

A quarter has no label

Turns without a unique majority vote are usually discarded, silently removing the hardest and most ambiguous cases from every evaluation.

No official splits

The paper defines no train, validation or test partition. Every group invents its own, and results are not comparable across papers.

6 Legacy and modern use

How the field actually uses IEMOCAP — and what to watch for

What the field actually reports on

A community convention built on top of the released corpus

1

Keep four classes

Angry, happy, sad, neutral. Frustration is dropped despite being the most frequent emotional label in the corpus.

2

Merge excitement into happiness

Justified by the confusion matrix — but it roughly doubles the happy class and changes the task.

3

Discard unlabelled turns

Only turns with a unique majority vote survive. The result is about 5,500 utterances.

4

Invent a split

Usually leave-one-session-out over the five sessions, but random utterance splits are still published.

What this means when you read a results table

Two papers reporting 'weighted accuracy on IEMOCAP' may differ in whether excitement was merged, which turns were kept, and whether the test speaker was ever seen in training. Those choices are worth several accuracy points each — often more than the modelling contribution being claimed. Before comparing numbers, check the preprocessing paragraph. If a paper does not state its split, treat the comparison as uninformative.

What IEMOCAP enabled — and what it left open

Setting up the second half of the seminar

Enabled

Multimodal emotion recognition with genuinely synchronised channels. Context-sensitive modelling, where the preceding turns inform the current prediction. Dyadic phenomena: entrainment, interruption, listener behaviour. Speech-driven facial and head-motion synthesis. Dimensional emotion regression alongside classification.

Left open

Ten speakers cannot tell you how emotion expression varies across age, sex or cultural background. Emotion and lexical content are entangled, because what is said differs across emotions. And there is no separate measurement of how well each modality alone conveys emotion — every label was assigned with audio and video together.

Next: CREMA-D (Cao et al., 2014)

Six years later, a group at Pennsylvania attacks exactly those three gaps. Ninety-one actors instead of ten, spanning ages twenty to seventy-four and five ethnic groups. Twelve fixed, semantically neutral sentences, so lexical content is held constant. And every clip rated three times over — from audio alone, from video alone, and from both — by more than two thousand raters.

The cost of all that control is everything IEMOCAP was built to protect: no dialogue, no partner, no context. That trade is the through-line of the whole seminar.

Questions for discussion

1. The paper's labels come from perception, not from the actor's intent — and the two disagree substantially. Which one should count as ground truth, and does the answer depend on the application?

2. Improvised turns were easier to annotate than scripted ones. Does that mean scripts are the wrong tool for emotion elicitation, or that ambiguity is a feature we should be capturing rather than avoiding?

3. A kappa of 0.27 sets a ceiling on what any classifier can meaningfully achieve. How should papers report performance against a noisy human reference?

4. Fear and disgust are essentially absent. Is a corpus that covers only the emotions its scenarios happen to elicit a fair test of emotion recognition?

5. IEMOCAP ships no train/test split. How much of the last fifteen years of reported progress on this benchmark do you think is real?