Human Labeling Bias Audit Analysis

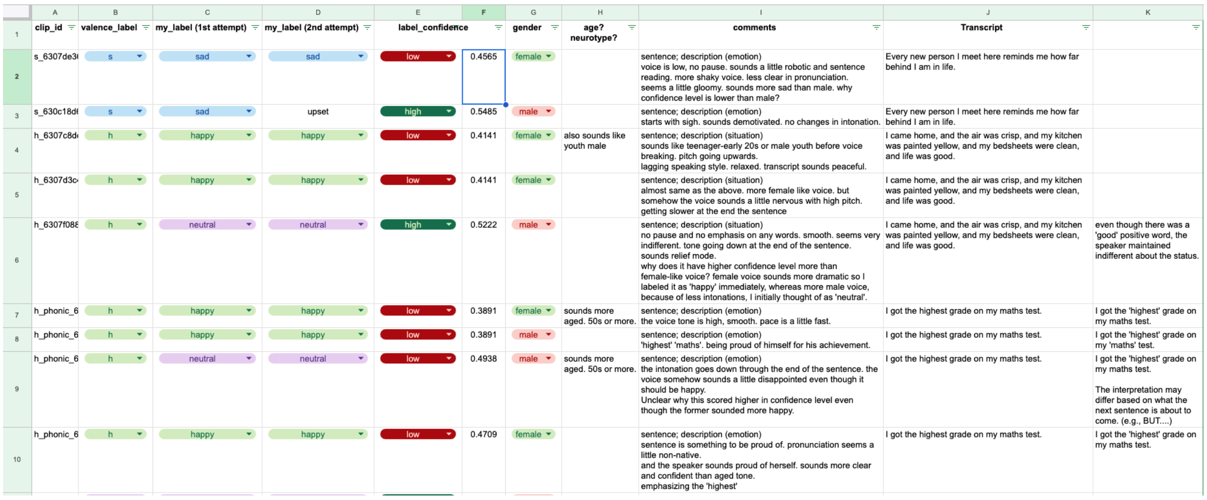
Data Track Beginner Track
Nov. 2025
Jiyae Choi

Methodology

Data Labeling

To minimize confirmation bias, I analyzed all 55 clips twice every 3 days (1st and 2nd attempts), initially without seeing the filenames that contained emotion indicators (h, a, s, n). I labeled and analyzed audio data based on:

- **Emotion labels:** basic 4 (happy, neutral, angry, sad), added (anxious, annoyed, upset, frustrated, scared)
- **Voice characteristics**: pitch (high, low), pacing (slow, fast), pauses, volume, pronunciation clarity
- **Linguistic content**: word choice, functional types of sentence (declarative, interrogative, imperative, exclamatory), explicit emotional language
- **Estimated demographics**: gender, age, neurotype
- **Label confidence**: confidence level evaluated by the provided API (over 0.5 confidence threshold labeled as "high")



	A	B	C	D	E	F	G	H	I	J	K
1	clip_id	valence_label	my_label (1st attempt)	my_label (2nd attempt)	label_confidence		gender	age? neurotype?	comments	Transcript	
2	s_6307de3	s	sad	sad	low	0.4565	female		sentence; description (emotion) voice is low, no pause. sounds a little robotic and sentence reading. more shaky voice. less clear in pronunciation. seems a little gloomy. sounds more sad than male. why confidence level is lower than male?	Every new person I meet here reminds me how far behind I am in life.	
3	s_630c18d	s	sad	upset	high	0.5485	male		sentence; description (emotion) starts with sigh. sounds demotivated. no changes in intonation.	Every new person I meet here reminds me how far behind I am in life.	
4	h_6307c8d	h	happy	happy	low	0.4141	female	also sounds like youth male	sentence; description (situation) sounds like teenager-early 20s or male youth before voice breaking. pitch going upwards. lagging speaking style. relaxed. transcript sounds peaceful.	I came home, and the air was crisp, and my kitchen was painted yellow, and my bedsheets were clean, and life was good.	
5	h_6307d3c	h	happy	happy	low	0.4141	female		sentence; description (situation) almost same as the above. more female like voice. but somehow the voice sounds a little nervous with high pitch. getting slower at the end the sentence	I came home, and the air was crisp, and my kitchen was painted yellow, and my bedsheets were clean, and life was good.	
6	h_6307f08	h	neutral	neutral	high	0.5222	male		sentence; description (situation) no pause and no emphasis on any words. smooth. seems very indifferent. tone going down at the end of the sentence. sounds relief mode. why does it have higher confidence level more than female-like voice? female voice sounds more dramatic so I labeled it as 'happy' immediately, whereas more male voice, because of less intonations, I initially thought of as 'neutral'.	I came home, and the air was crisp, and my kitchen was painted yellow, and my bedsheets were clean, and life was good.	even though there was a 'good' positive word, the speaker maintained indifferent about the status.
7	h_phonic_6	h	happy	happy	low	0.3891	female	sounds more aged. 50s or more.	sentence; description (emotion) the voice tone is high, smooth. pace is a little fast.	I got the highest grade on my maths test.	I got the 'highest' grade on my maths test.
8	h_phonic_6	h	happy	happy	low	0.3891	male		sentence; description (emotion) 'highest' 'maths'. being proud of himself for his achievement.	I got the highest grade on my maths test.	I got the 'highest' grade on my 'maths' test.
9	h_phonic_6	h	neutral	neutral	low	0.4938	male	sounds more aged. 50s or more.	sentence; description (emotion) the intonation goes down through the end of the sentence. the voice somehow sounds a little disappointed even though it should be happy. Unclear why this scored higher in confidence level even though the former sounded more happy.	I got the highest grade on my maths test.	I got the 'highest' grade on my maths test. The interpretation may differ based on what the next sentence is about to come. (e.g., BUT....)
10	h_phonic_6	h	happy	happy	low	0.4709	female		sentence; description (emotion) sentence is something to be proud of. pronunciation seems a little non-native. and the speaker sounds proud of herself. sounds more clear and confident than aged tone. emphasizing the 'highest'	I got the highest grade on my maths test.	I got the 'highest' grade on my maths test.

Attached the data analysis file to the last page of the report.

Data Labeler (Me)

As the sole data labeler, my interpretations were shared by:

- Cultural background: Korean woman (30s) from a homogenous society valuing indirect communication
- Language: Non-native English speaker, Korean as a mother language
- Limited exposure: Minimal prior contact with neurodivergent communication patterns, initially coding atypical voice clips as “unusual” rather than “different”

Data Label Analysis

Label Distribution

Out of 55 clips, Valence and I agreed on 41 labels (74.6% match rate) after the 2nd attempt. Valence's mean confidence scores are 0.43. For convenience, I have set the typical threshold at 0.5, with values above this threshold labeled as “high” and those below it as “low.”

Label	Happy	Sad	Angry	Neutral	Else	Total
valence_label	14	7	2	32	0	55
1st attempt	10	4	3	35	3	55
2nd attempt	10	2	3	33	7	55
Match Rate **(valence_label vs. 2nd attempt)**	71.43%	28.57%	50.00%	90.62%	-	74.64%

valence_label	my_label (2nd attempt) (Bracket is the count)	Total Count
angry	Upset (1)	1
happy	Neutral (4)	4
neutral	Angry (1), frustrated (1), sad (1)	3
sad	Annoyed&upset (1), anxious (2), anxious&angry (1), scared&anxious (1), upset (1)	6
Total Mismatch		**14**

My second attempt revealed nuanced distinctions. Valence’s analysis of “sad” diverged into a broader range of related negative emotions, such as “upset,” “scared,” “anxious,” or “annoyed.” This indicates a need for more granular labeling rather than distinct emotions.

Valence's 0.43 mean confidence also demonstrates consistent uncertainty, yet forces single labels. In video conferencing engagement metrics, these uncertain predictions systematically misread or stigmatize neurodivergent communication as disengagement or instability.

Bias Audit Analysis

Four patterns emerged from 14 mismatched clips. Each reveals how bias in training data creates systematic accessibility barriers when emotion AI is deployed in workplace communication tools.

Pattern 1: Positive valence with Flat tone

Example 1**:** A woman says, *"You are the most wonderful person I've ever met."*

- **Valence**: Happy (low confidence: 0.73)
- **My label**: Neutral
- **Reasoning**: “The voice tone is low and down. Pronunciation doesn't seem native. Why label confidence high? Doesn't shound so happy or emotional.”

Looking back, this could have been a neurodivergent speaker expressing genuine appreciation. The words were clearly positive (“most wonderful”). However, I have assumed that real happiness requires vocal enthusiasm with a high pitch and dramatic tone.

Example 2: A male voice says, *"I came home, and the air was crisp, and my kitchen was painted yellow, and my bedsheets were clean, and life was good."*

- **Valence:** Happy (high confidence: 0.52)
- **My label:** Neutral
- **Reasoning:** "No pause, no emphasis. Smooth and indifferent. Tone going down at the end."

Example 3: *Male voice, seemingly at the age of more than 50s, "I got the highest grade on my maths test"*

- **Valence:** All happy (average confidence: 0.4357)
- **My label:** 3 happy, <u>one neutral</u>
- **Reasoning:** "The intonation goes down through the end of the sentence. The voice somehow sounds a little disappointed, even though it should be happy."

From these three examples, positive content with non-enthusiastic delivery created cognitive dissonance for my data labeling. I have realized that I struggled with monotone speech, defaulting to "neutral" regardless of content.

Individuals with autism who exhibit a flat tone or certain speech disorders may often convey emotion without vocal inflection. When video conferencing utilizes emotion AI for engagement scoring, flat affect can become a professional barrier. Systems may flag them as ‘disengaged,’ which can affect meeting participation and performance evaluations.

Pattern 2: The Gender Intonation Bias

Example: 10 clips of *"Please tell him he will receive the letter in about five days,"* in both female and male-identified voices.

- **Valence**: All Neutral (confidence ranging from 0.3295 to 0.5342)
- **My label**: Neutral 8, Angry 1, Else 1 (seems frustrated)
- **Reasoning**: "With emphasis on ‘please' and 'five'. Slow-paced. sounds bothering and frustrating." “Pronunciation is collapsing and almost sounds like murmuring. Based on the context, the speaker seems bothered to speak.”

There were two interesting findings from this pattern 2. The male average confidence level was slightly higher (0.45) than that of the females (0.36). In addition, female voices with a clear voice seemed neutral, whereas male voices with specific emphasis seemed “angry” or “frustrated.” Although the metric is not as significant, the gender-confidence disparity can suggest that female voices require more evidence of emotion.

Pattern 3: The Age and Voice Quality Confusion

Example: Voice clips of *“It is a good idea to study before the big test.”*, including assumed older adults over the age of 50.

- **Valence**: All Neutral (high confidence: 0.52)
- **My label**: All Neutral
- **Reasoning**: "Voice sounds a little nervous with high pitch.", "Voice seems shaky, a little tensed."

I associated age-related vocal tremor with anxiety or nervousness despite neutral content, voice quality overriding semantics. In video conferencing, this could affect their perceived competence or emotional stability, as they may be systematically misread as anxious or unstable in professional settings.

Pattern 4: The Context Collapse

Example: *"Please don't leave me all alone"*

- **Valence**: Sad (presumably based on the words)
- **My label**: Stressed/annoyed (not sad)
- **Reasoning**: "A little pause after 'don't.' The voice sounds direct, annoyed."

Emotion recognition from voice alone can misinterpret the proper intention. Video conferencing systems often make judgments without understanding relationship dynamics or individual communication styles, which can cause professional harm, especially for neurodivergent users.

The Framework

Based on these patterns, I have identified some potential possibilities that emotion recognition systems can address to serve diverse populations truly.

1. Multimodal Weighting with Transparency

Current approach: Equally treating all signals (voice, content, etc), then output one single emotion label and confidence score.

Suggested approach:

- Show the highest label confidence scores for each signal's emotion separately.
  - Example: Voice (happy: 0.6 confident), Content (neutral: 0.4 confident)
- Flexible weighting across multi-modalities depending on the users' preferences or needs. With autistic users, vocal traits might be less weighted than the explicit content.

2. Expanded Emotion Taxonomy

Current approach: Only four categories: happy/sad/angry/neutral.

Suggested approach: Allow for complexity when multiple interpretations or uncertainties exist.

- Include 'intensity' level for subtlety ("Somewhat happy", "Very happy", etc.)
- Expand the emotion label for detail ("Satisfactory", "Enthusiastic", etc.)
- Multiple emotion labels ("Positive content with flat tone")
- Explicit uncertainty ("Multiple interpretations possible")

3. Data Labeling Transparency

Current approach: Four emotion labels and confidence level as results.

Suggested approach: Include brief explanations of why emotions and confidence level were labeled as such.

- Publish annotator demographics (race, gender, age, neurotype, native language) and for labeling context.
- Explanations of why annotators labeled data as such for labeling transparency
- Review the three sentences preceding and following the label data to understand the context.

4. Mandatory Inclusive Data Standards

Suggested approach: Introduce a more inclusive approach regarding neurodivergent or atypical communication in training, labeling, and testing datasets.

- **Training**
  - Require dataset showing representation across gender, age ranges, neurotypes, accents and dialects, and cultural backgrounds
  - Minimum 30% neurodivergent speakers (autistic, ADHD, alexithymia, etc.)
  - Never use monotone or flat affect as a negative training signal
- **Labeling**
  - Prioritize explicit content over vocal delivery
  - Include neurodivergent annotators on every labeling team
  - Mandatory training on neurodivergent communication patterns
- **Testing**
  - Test separately on neurotypical voice clips and neurodivergent clips
  - Never deploy until validated on both user segments

Reflections

The model and I agreed on 75% of labels, but this match reflects shared bias, not the model's accuracy. We were both confused with flat tone, found more confidence in male and young voices, and struggled with atypical prosody, common communication patterns among neurodivergent populations. What I found wasn't just algorithmic bias but a mirror reflecting how we all make flawed assumptions about how emotions "should" sound. Emotion is not a performance for others to judge but a genuine expression of oneself. It's one part of the expressions along with voice, face, and language.

The biggest takeaway from this challenge is that there is no absolute objective label for emotion. If we are to build emotion recognition systems for accessibility, we owe it to neurodivergent users to make them “with” them, not for them. The path forward requires humility and inclusion. The question isn't just about improving the accuracy of emotion models, but respecting diverse ways of communication. We should be aware of the curb-cut effect when building accessible and inclusive technology. Designing for people with mental or physical challenges can eventually improve experiences for everyone in the society, and for this video conferencing context, in professional settings.

Limitations

This analysis is limited by single-annotator bias and decontextualized clips that lack the conversational dynamics essential for real video conferencing interpretation.

Attached Files

Copy of Labeling Spreadsheet Template_Jiyae Choi