Guanghua School of Management
Peking University

因思想 而光华

Large scale voice and video data for financial forecasting

Feng Li

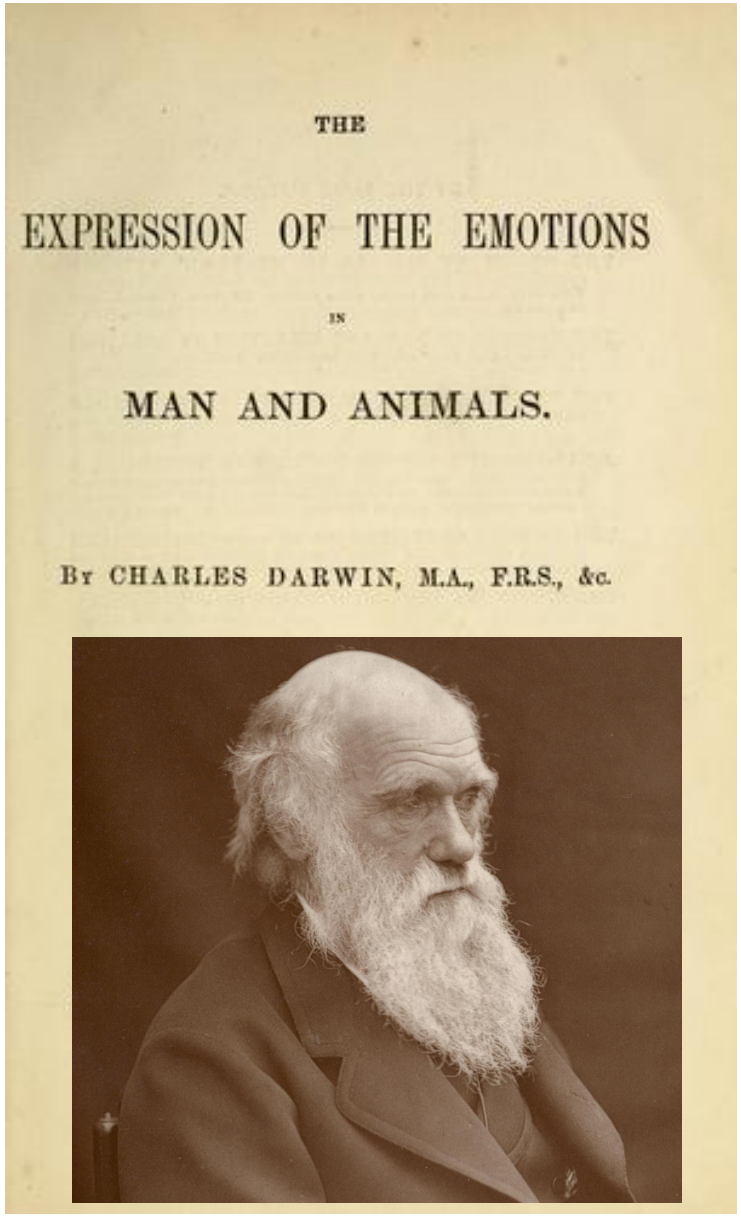
Guanghua School of Management

Peking University

feng.li@gsm.pku.edu.cn

Nonverbal communication

- Communication through a nonverbal platform such as **eye contact**, **body language**, **voice**, etc.
- **Darwin**, for the first time, studied nonverbal communication in *The Expression of the Emotions in Man and Animals* in **1872**
- He noticed the **interactions** between animals such as lions, tigers, dogs etc. and realized they also communicated by gestures and expressions.



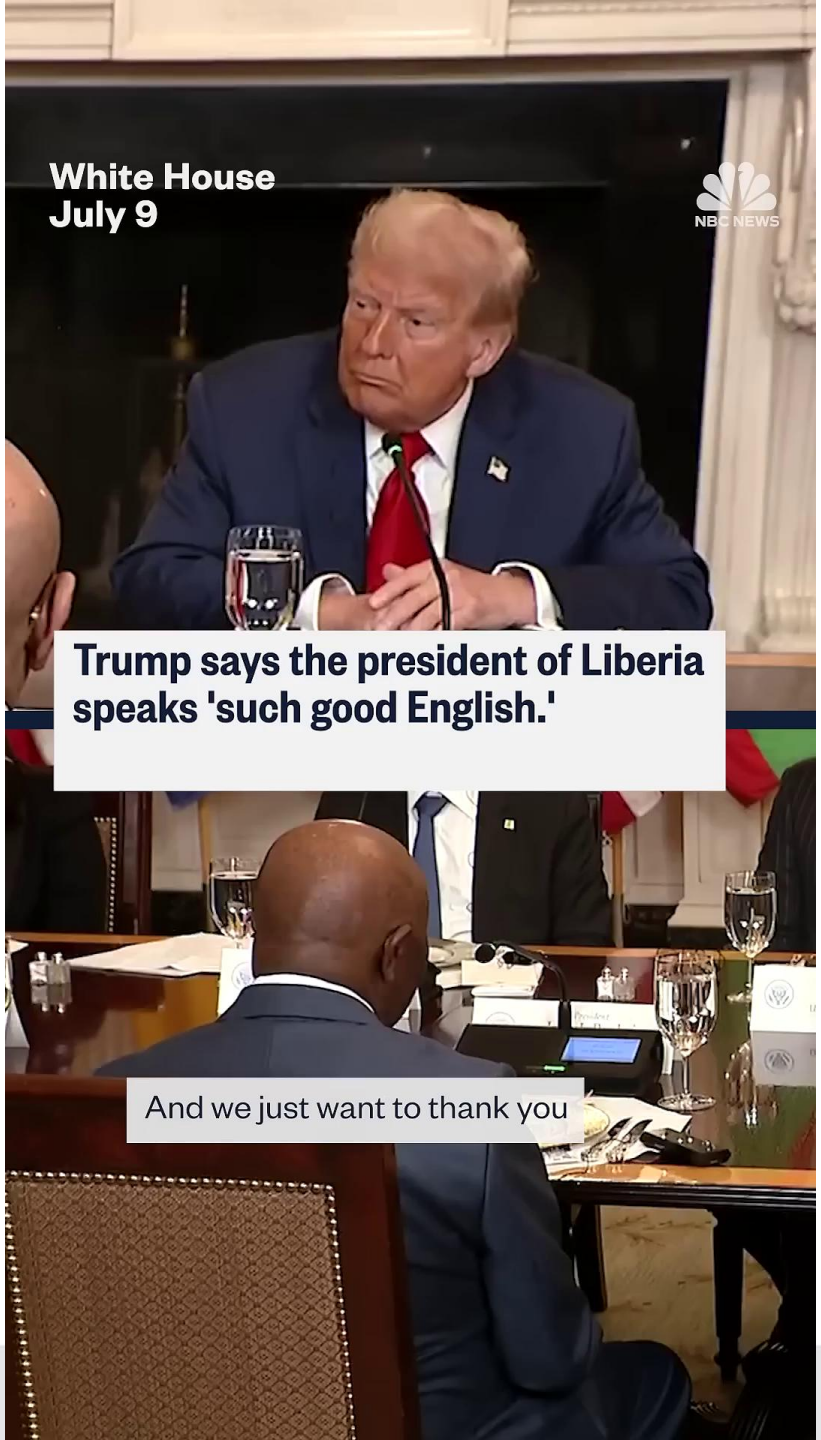
Impact of Accent on Communication Efficiency

- Psychology and linguistics show that **accented speech reduces processing fluency**. (Lev-Ari & Keysar, 2010; Munro & Derwing, 1999;Clarke & Garrett, 2004)
- **Listeners perceive accented speakers as less precise or credible**—even with identical content. (Fuertes et al., 2010; Lippi-Green, 1997)
- Some cases in the video
  - President of Liberia Speaking English
  - Member of Parliament Scottish Accent Baffles British Parliamentarian
  - Trump Skipping Question On Anti-India Activities, Can't Understand Tough Accent (Blaming on the accent?)
  - SNL Clip – Scottish Air Traffic Controller

https://www.youtube.com/watch?v=1jHfY0dDZxA

https://www.youtube.com/watch?v=NBsGCdFgjH8

President of Liberia Speaking English



https://www.youtube.com/shorts/eEkPxQxFZvM

Communication Shapes Market Efficiency

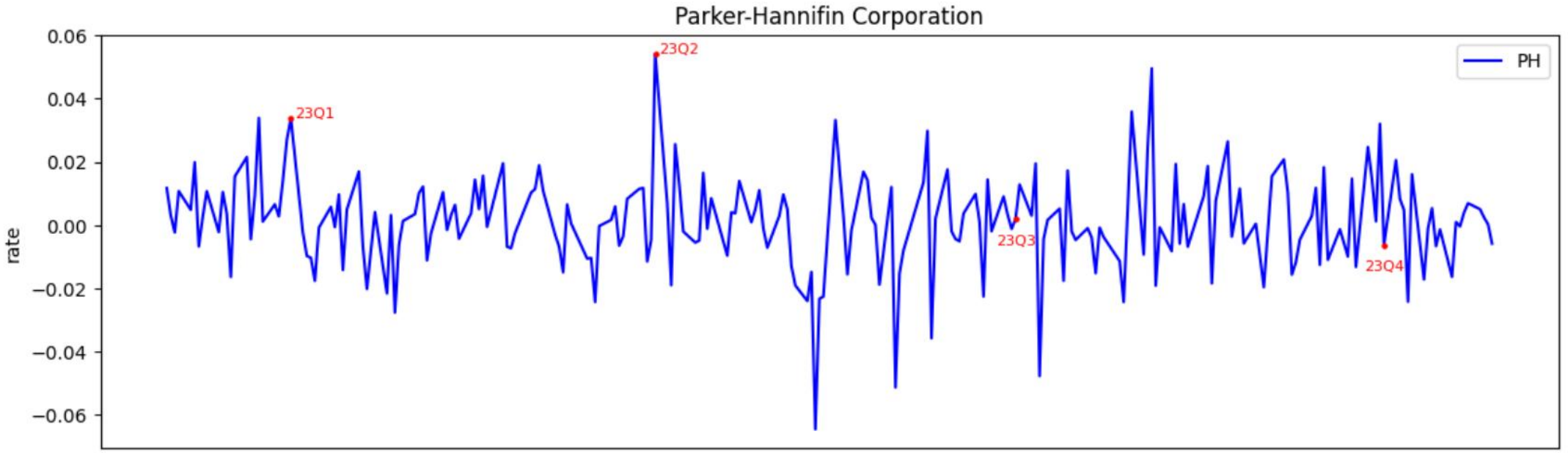
- Financial analysts translate complex disclosures into actionable information.
- Earnings calls—especially Q&A—are where private signals become public.
- Prior research: informativeness depends on **WHAT is said** (Textual analysis).
- We study whether the informativeness depends on **HOW they say** (Audio analysis)
- apart from what they say.

Accents meaningfully shape information efficiency in finance

- **Communication efficiency** depends on who speaks and how information is delivered—not just on textual content.
- **Linguistic diversity** can create frictions but can also enhance informational value when aligned with audience background.
- Firms and analysts may benefit from awareness of **how accents and communication styles interact with listener composition**.
- Future directions: interaction with other vocal cues, repeated interactions, and global settings.

Read Our Paper:
Accent Matters: Communication Costs and Information Content in Analyst–Manager Dialogue
(Feng Li, Tengjia Shu and Mengxi Yu)

Your tone matters



$$y_{i,t,n} = \frac{stockprice_{i,t+n} - stockprice_{i,t+n-1}}{stockprice_{i,t+n-1}} - \frac{NSDQindex_{t+n} - NSDQindex_{t+n-1}}{NSDQindex_{t+n-1}}$$

2023Q1_earningscall.mp3

VS

2023Q4_earningscall.mp3

Q1: disappointing guidance

Q4: optimistic outlook

Earnings Call: Data

- **Data Source**: Public company quarterly earnings calls, retrieved from earningscall.biz
- **Audio recordings**: Spoken statements by CEOs, CFOs, analysts
- **High-frequency time series** (44kHz): 44,000 samples per second.
- **Scale**
  - **61,000 of calls** from 2017- 2025
  - **5000+ firms** across sectors (e.g., tech, finance, healthcare)
  - Audio duration ranges from **60 - 90 minutes per call**
  - **Total size: 7.6 TB**

Finfluencers: Data

- A large-scale, multimodal dataset from financial content creators
- Sourced from Qifutong (Zhongtai Securities) platform focused on financial education and commentary
- **Video Scale:**
  - **Size:** 7600 videos from 100 finance influencers, 2.5TB
  - **Video content**: finance-related clips (30-60 min)
  - **Time resolution:** 24-60 frames per second
- Video ID, post time, views, likes.

IPO Roadshow: Data

- **A pre-IPO marketing campaign** where company executives present to institutional investors
- Aimed at explaining:
  - Business model
  - Financials and forecasts
  - Competitive landscape
  - Investment rationale
- Delivered via **video presentations**, **voice narration**, and **Q&A sessions**
- **2.0 TB**, 30–60 minutes long for each video
- **Time span**: From 2013 onward, with 2979 IPOs across sectors

IPO Roadshow: Forecasting Applications

- IPO Day Price Movement
- Underpricing Risk or likelihood of first-day “pop”
- Probability of Post-IPO Volatility
- Investor sentiment modeling from Q&A dynamics
- Automated “confidence scoring” of management tone over time

Why These Data Matter

- **Rich Behavioral Signals:** Tone, hesitation, facial expressions, and gestures reflect confidence, stress, or uncertainty
- **Go beyond what is said** — focus on how it's said
- **High-Frequency and Time-Aligned**
  - Audio and video provide continuous time series, frame-by-frame or second-by-second
  - Enable modeling fine-grained dynamics over time (e.g., tone shifts, gesture bursts)
- **Real-World Impact**
  - These communications drive investor sentiment, market reactions, and policy expectations
  - Small changes in delivery can trigger volatility, repricing, or attention shifts
- **Financial Stability Relies on Public Understanding**
  - A well-informed public helps reduce irrational herding, panic, and speculative bubbles
  - Clear communication from policymakers and firms supports market confidence
  - Better education → more stable behavior → more stable markets
  - Platforms like Qifutong and IPO roadshows serve as living laboratories for this cycle

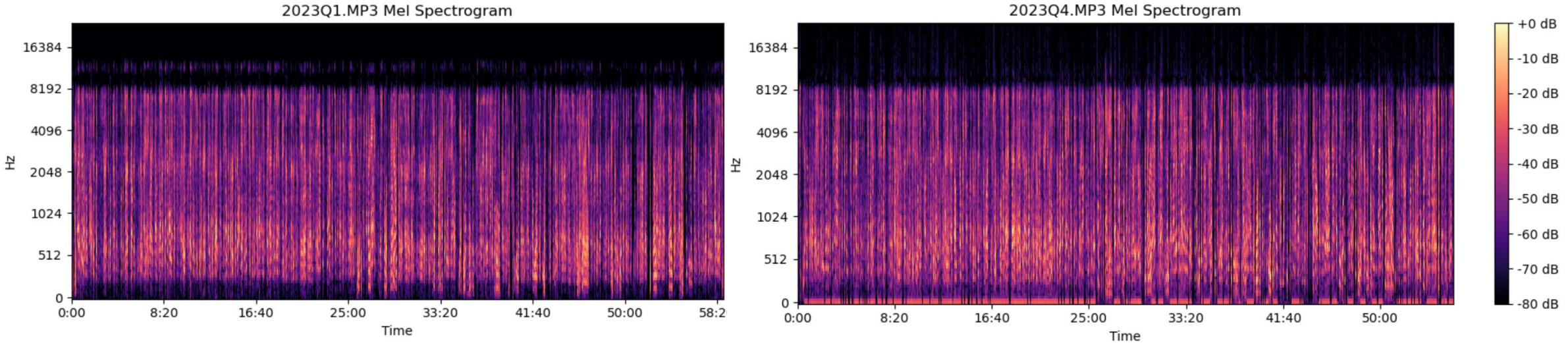
Voice Features

- **Short-time Zero Crossing Rate:** The number of times the signal waveform crosses zero per second within a window, normalized by the window length.
- **Short-time Energy:** The energy level of the speech signal per second within a window.
- **Spectral Centroid:** The center of mass of the audio spectrum within a window, calculated as the weighted mean of frequencies.
- **Loudness (LUFS):** Integrated loudness measured in Loudness Units Full Scale (LUFS) per second.
- **Sharpness:** The proportion of high-frequency energy (typically > 2 kHz) to total energy within a window.
- **Mel-Frequency Cepstral Coefficients (MFCC):** Cepstral coefficients derived from a Mel-scaled filter bank, mimicking human auditory perception.
- **Speech Rate:** The number of vowels spoken per second by the speaker.
- **Number of Questions:** The count of questions raised by financial analysts during an Earnings Call

Voice Features

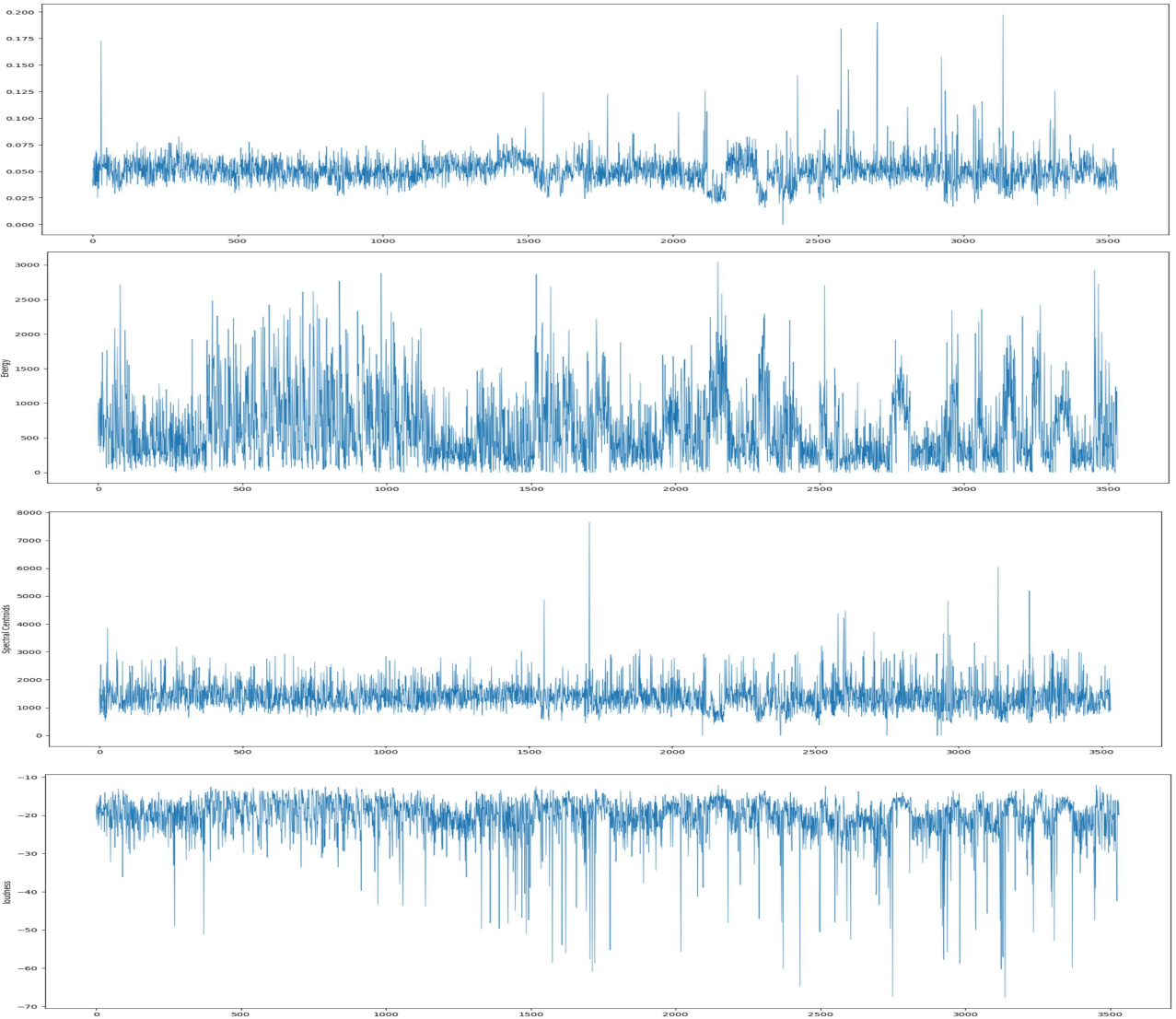
Mel-Spectrogram: heatmap of sound

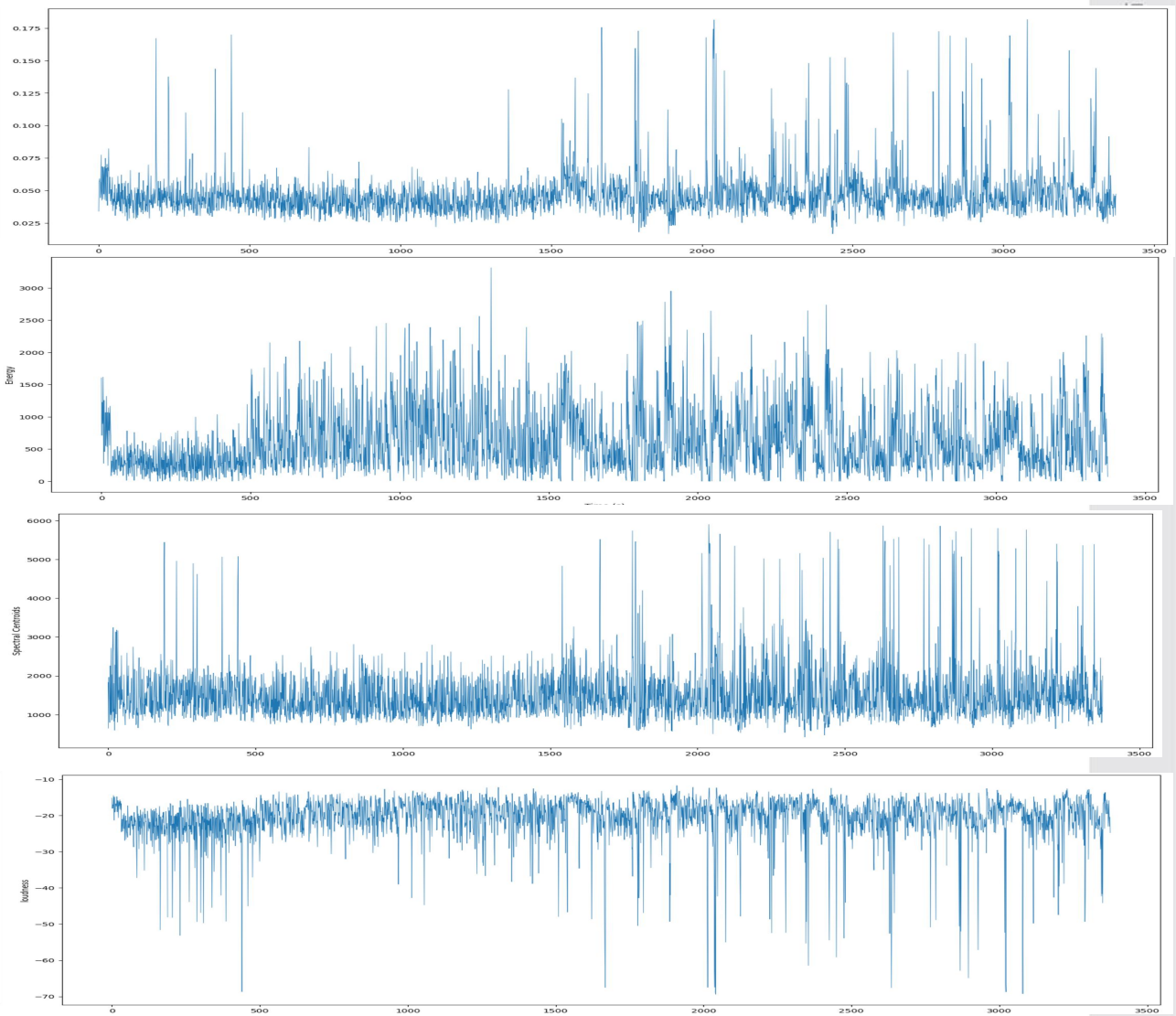
- A **visual representation** of sound over time
- Shows how **energy** is distributed across frequencies
- Uses the Mel scale, which reflects **how humans perceive pitch**
- Detect emotional arcs, stress points, speaking patterns by "see" **patterns in sound**
- Detect emotional arcs, stress points, speaking patterns



Mel-spectrogram. A brighter color indicates more energy at that frequency at that point in time.

Voice Features: Time Series

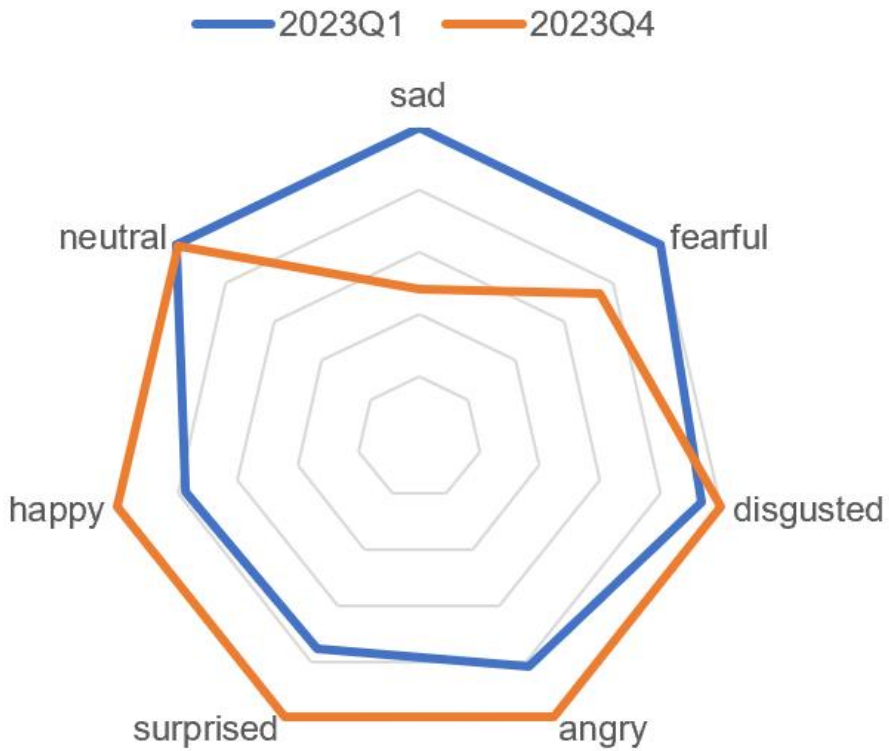




Short-time Zero Crossing Rate, Short-time Energy, Spectral Centroid, Loudness (LUFS) for 2023Q1 and 2023Q3

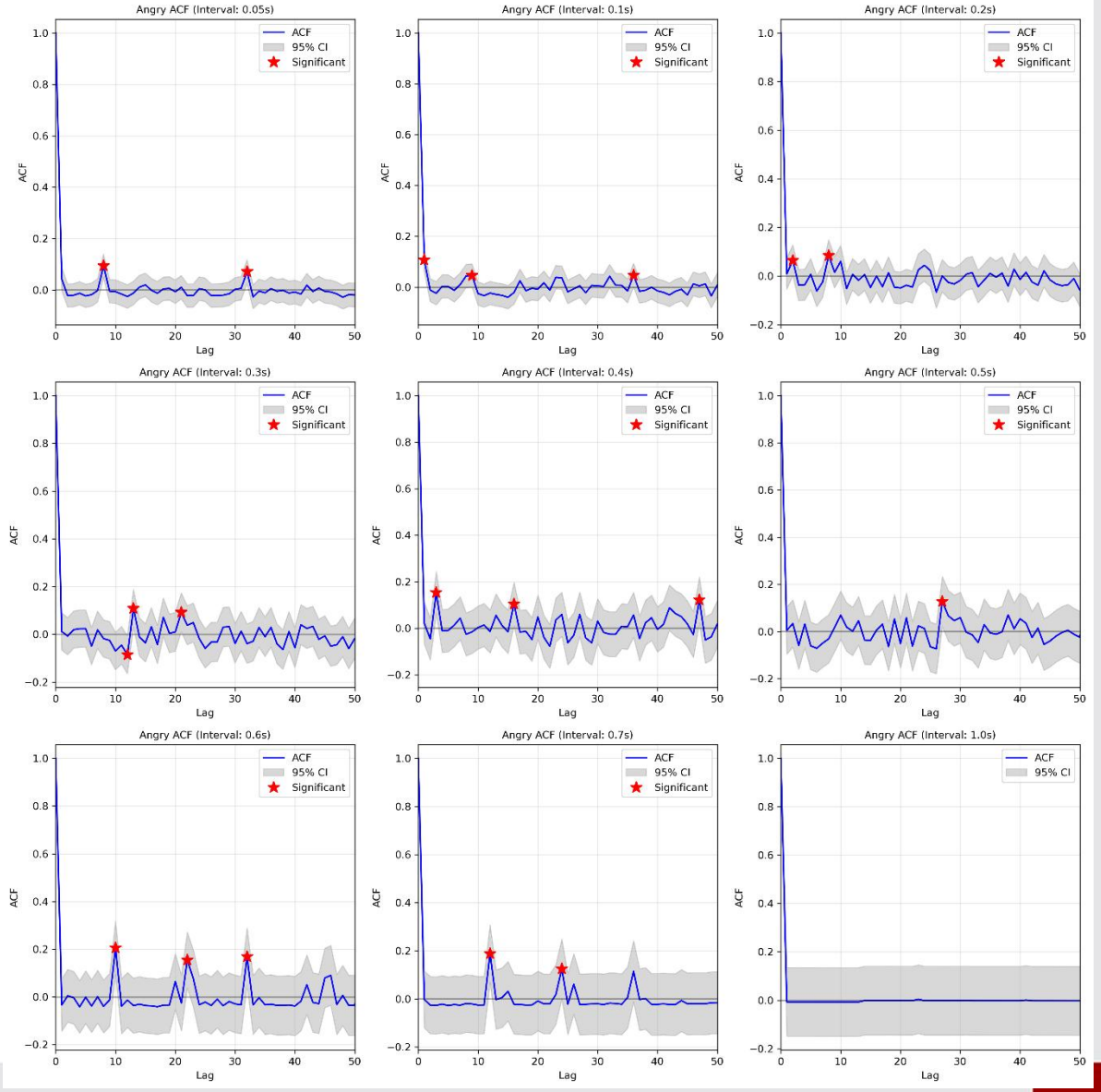
Emotional Features

- **Definition**
  - Emotional features represent affective states extracted from voice, video
  - Capture how people feel — not just what they say
  - Essential for understanding human intent, confidence, and uncertainty
- **Time Series Representation**
  - Emotional signals are dynamic, not static
  - Can be measured per frame, per second, or per sentence
- **Models**
  - Transformers + Emotion Embeddings: for multimodal fusion
  - Pretrained emotion models: e.g., **emotion2vec**, FER+, MSP-IMPROV

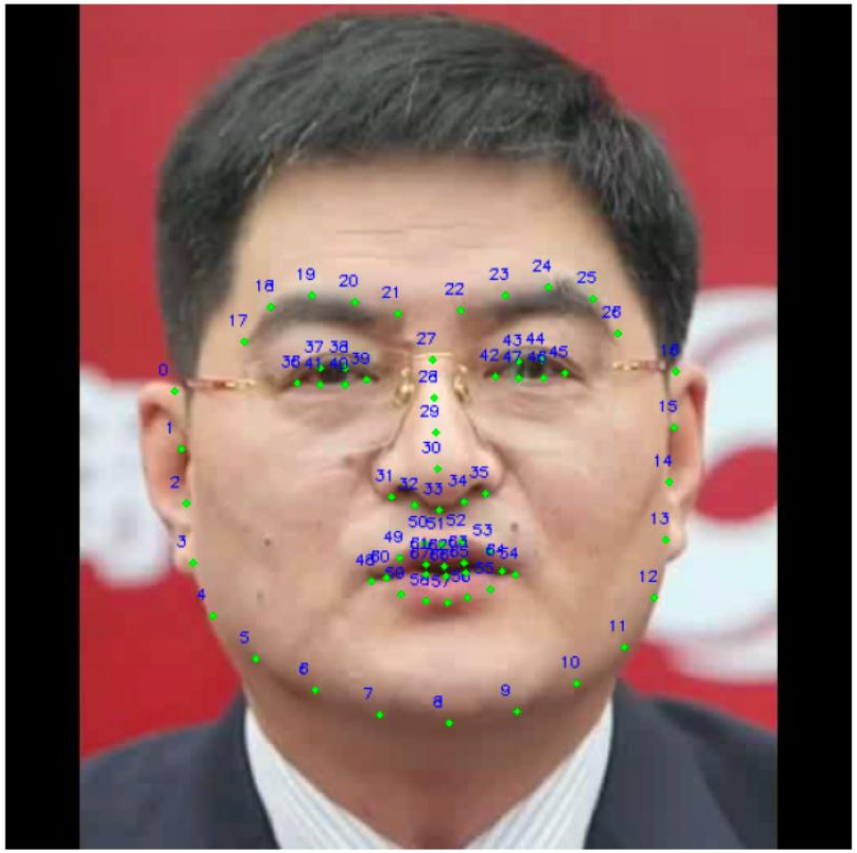


Sentiment analysis captures 7 emotional dimensions from **vocal cues**

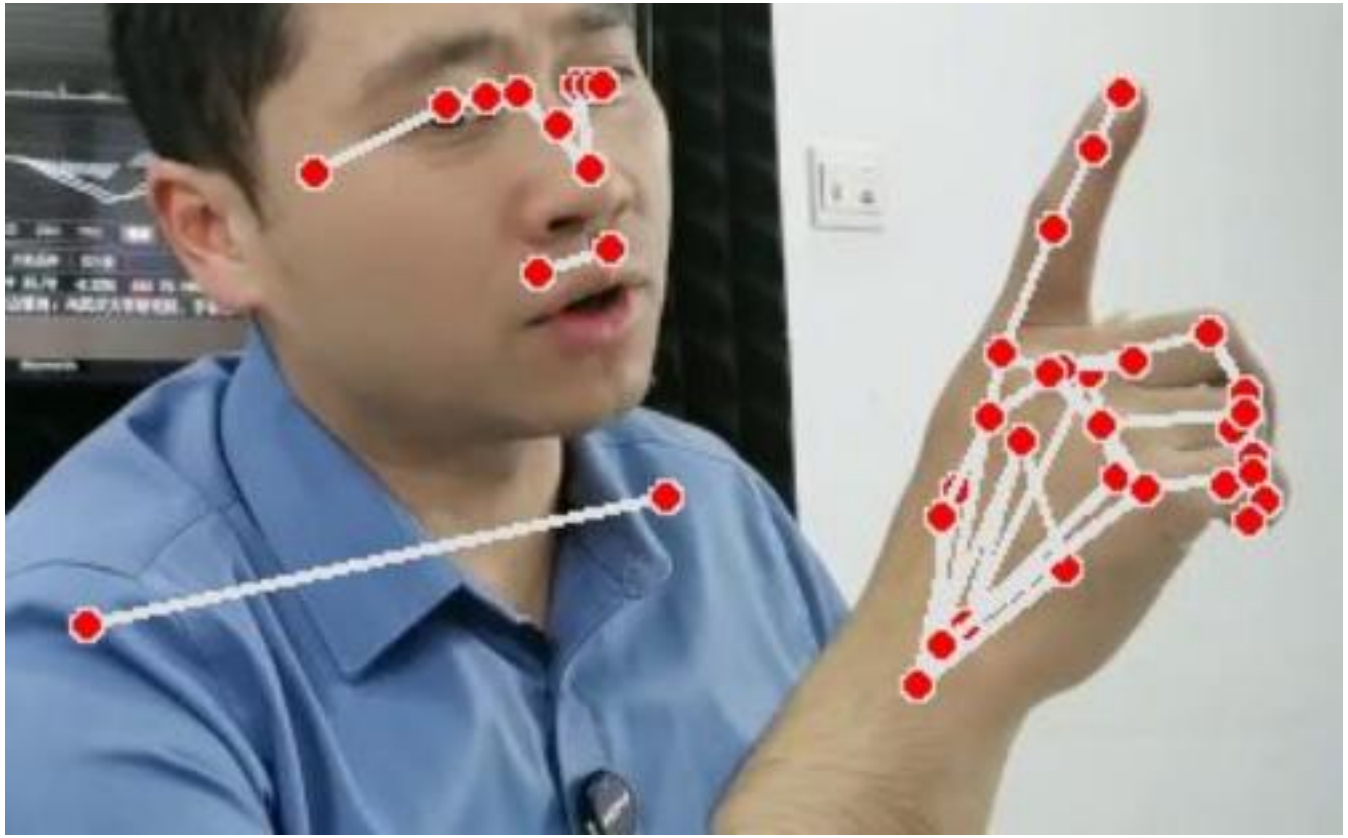
Emotional Features: Second Order



Pose and hand landmarks



Frame level facial landmarks (**IPO Roadshow**)



Frame level pose and hands landmarks (**Finfluencers**)

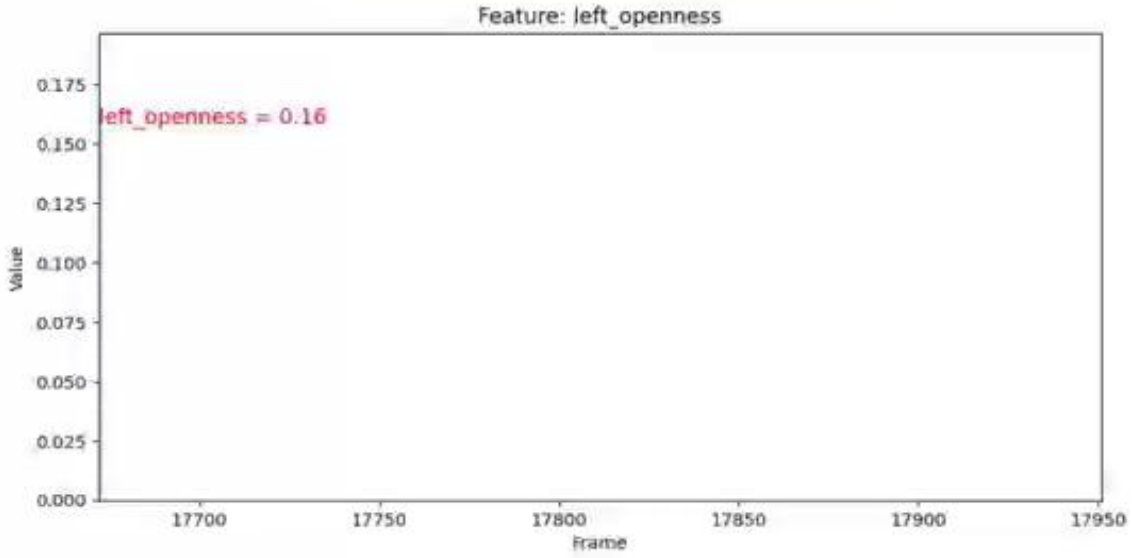
Understanding gestures using landmarks

Gestures matter

- **Reflect nonverbal communication**, emotion, and speaker intent
- **Emphasize or contradict spoken content** (e.g., nodding while denying)
- **Reveal cognitive states**: confidence, anxiety, certainty, rehearsal

Common gesture signals for forecasting

- **Hand movement intensity** (e.g., large vs. minimal motion)
- **Frequency** of gestures per minute
- **Asymmetry** or irregularity in movement
- **Face-hand-vocal coordination** (e.g., face touching, nose scratching)



因思想 而光华

Gesture clustering

- We found raw coordinate data do not work well in gesture clustering.
- We calculate **per frame statisitcal features** based on landmarks including
  - **Hand distance**
  - **Adjacent fingertips distance**
  - **Fingertip-wrist distance**
  - **Finger bending degree**
  - **Palm closure**
  - **Palm angle**
  - **Hand movement speed**

- **Hands openness**

large

small

- **Hands distance**

large

small

Gesture forecasting

- **Definition**
  - Predicting future human gestures from past motion sequences using statistical, machine learning, or deep learning models.
  - Applied in human–computer interaction, virtual reality, robotics, healthcare, and behavioral analysis.

- **High variability** across individuals and contexts.
- **Noisy sensor** in video data.
- Balancing accuracy, latency, and interpretability in real-time applications.

- **Applications**
  - Gesture-based control in AR/VR and gaming.
  - Assistive technologies (sign language translation, rehabilitation monitoring).
  - Predictive human–robot collaboration.

Pipeline tasks for video and voice data

- **Detecting and grouping streamers** from large collection of videos
- **Speaker diarization** (speech activity detection, speaker change detection, overlapped speech detection, speaker embedding)
- **Tracking pose and hands landmarks**
- **Extracting second aligned transcripts**

Thanks!

- Supported by National Natural Science Foundation of China
- My two RA students: **Ding, Pengcheng and Yu, Mengxi**

- **Our Python Library** https://github.com/feng-li/videofeatures

- **Computing environment**
  - **Local processing**: 104 CPU cores, 1TB RAM, 80TB HD, one 4090 GPU card
  - **Cloud**: eight 4090 GPU cards, 8 muxi GPU cards, 8 A100 GPU cards