CMU MOSI Dataset

- YouTube Videos express opinions about various subjects

BUT I CAN SAFELY ASSURE YOU THAT EVEN IF THEY DIDNT IT WOULD STILL BE A HIT

Sentiment Score: + 2.4

FOR THIS ONE I JUST DIDNT FEEL LIKE IT

Sentiment Score: - 2.0

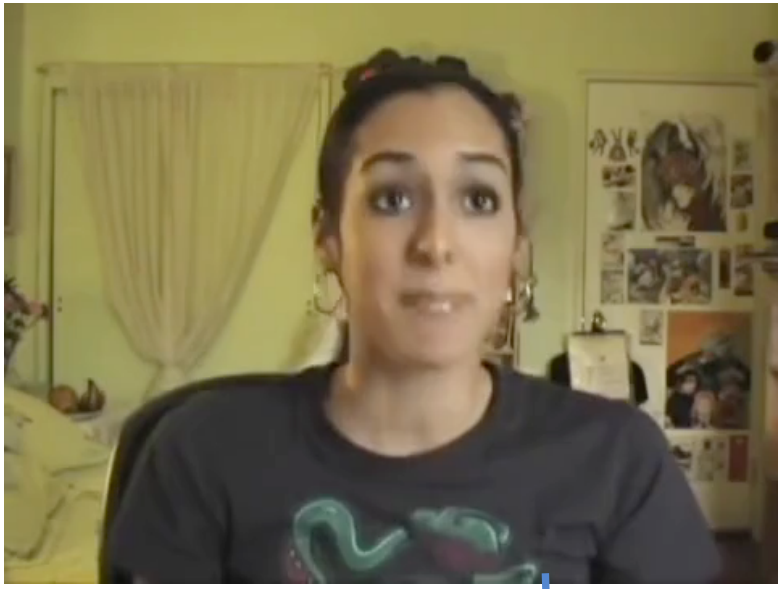
CMU MOSI Dataset

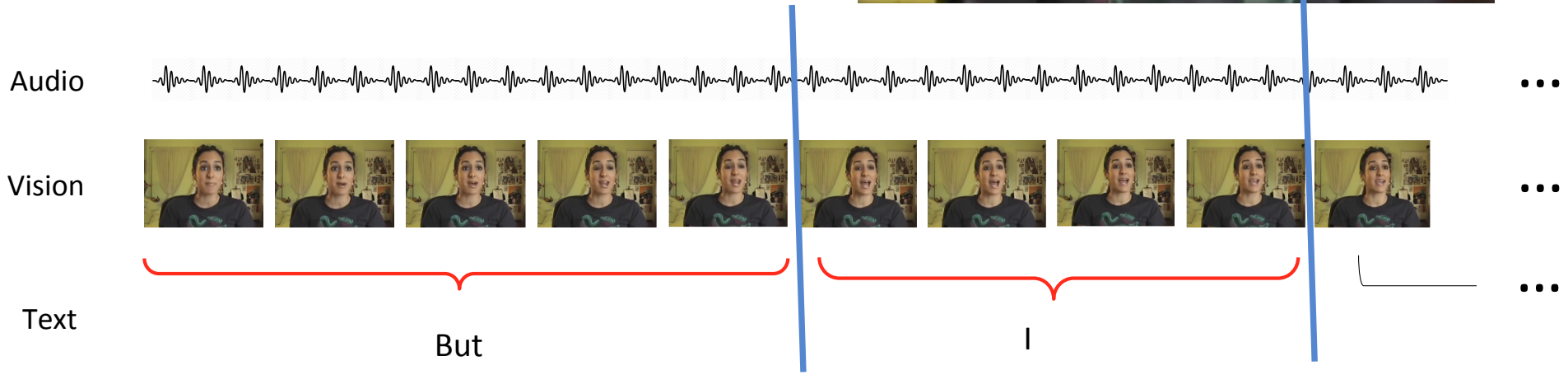
- MOSI: Multimodal Opinion Sentiment Intensity (MOSI) Dataset
  - Multimodal Observations: transcribed speech (language), visual face gestures (visual), and audio features (acoustic)
  - Opinion-level subjectivity segmentation:
    "[I love Shawshank Redemption]"
  - Sentiment Intensity Annotations: {-3, -2, -1, 0, 1, 2, 3}
    Highly Negative -… - Neutral - … - Highly Positive
  - Alignment: between words, visual, and acoustic features.

Example

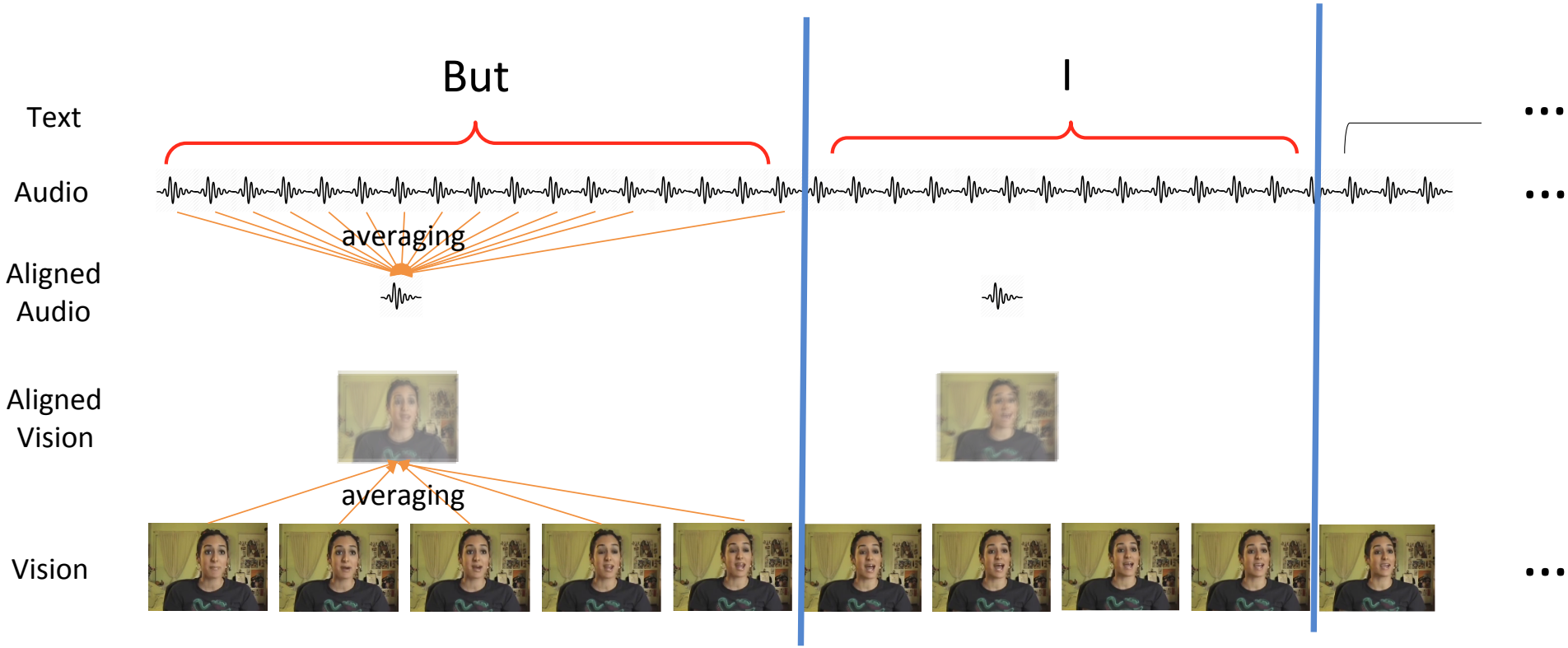
BUT I CAN SAFELY ASSURE YOU THAT EVEN IF THEY DIDNT IT WOULD STILL BE A HIT

Sentiment Score: + 2.4





Data Preprocessing: Averaging



Statistics

- # Train: 1283 videos
- # Valid: 229 videos
- # Test: 686 videos
- Average length of videos: 4.2 sec
- Average word count per videos: 12

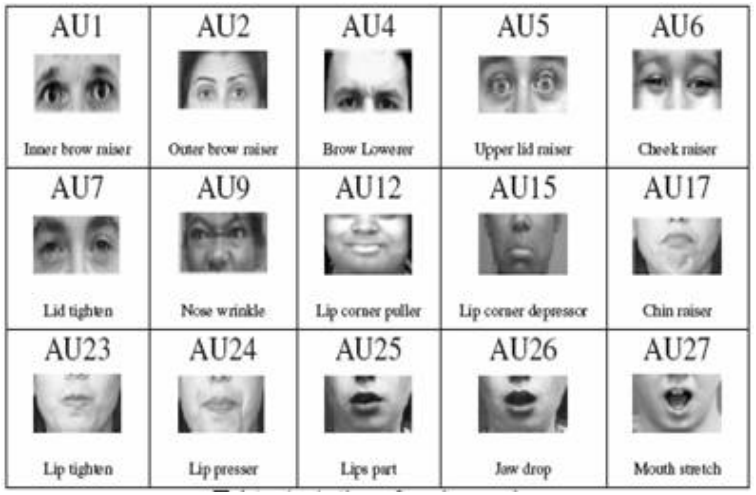
Language Features

- Transcribed manually.
- 300-dim. Glove word embeddings [1] from 840 billion tokens from the Common Crawl Dataset [2]

1. [Pennington,Socher,andManning2014] Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation.
2. http://commoncrawl.org

Visual Features

- Generated automatically at 30 Hz.
- 46-dim. extracted from Facet [1]
  - Facial Action Coding Systems (FACS) [2] with 46 observable action units (AUs)
  - AU intensity on a 5 point scale



1. https://imotions.com
2. Ekman, Paul, and Erika L. Rosenberg, eds. *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, USA, 1997.

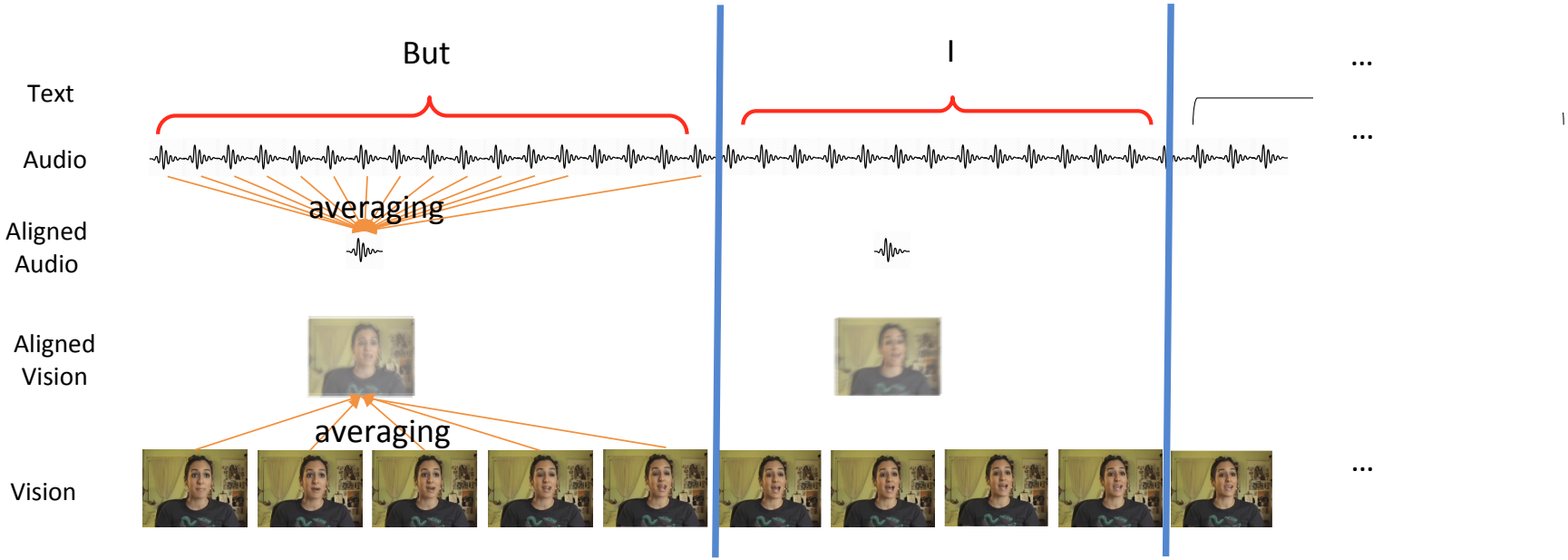
Audio Features

- Generated automatically at 100 Hz.
- 74-dim. extracted from COVAREP [1]
  - 12 Mel-frequency cepstral coefficients, pitch tracking and voiced/ unvoiced segmenting features [2], glottal source parameters [3], peak slope parameters and maxima dispersion quotients [4].

1. Degottex,G.;Kane,J.;Drugman,T.;Raitio, T.; and Scherer, S. 2014. Covarepa collaborative voice analysis repository for speech technologies. In *Acoustics, Speech and Signal Processing (ICASSP), 2014.*
2. Drugman, T., and Alwan, A. 2011. Joint robust voicing detection and pitch estimation based on residual harmonics. In *Interspeech*, 1973–1976.
3. Childers, D. G., and Lee, C. 1991. Vocal quality factors: Analysis, synthesis, and perception. the Journal of the Acoustical Society of America 90(5):2394–2410.
4. Kane,J.,andGobl,C.2013.Waveletmaxima dispersion for breathy to tense voice discrimination. IEEE Transactions on Audio, Speech, and Language Processing 21(6):1170–1179.

Features Alignment & Pre-processing

- Perform forced alignment of three modalities using P2FA [1].
- Interval duration of each word utterance as a time-step.



1. Yuan, J., and Liberman, M. 2008. Speaker identification on the scotus corpus. *Journal of the Acoustical Society of America* 123(5):3878.

Features Alignment & Pre-processing

- Perform forced alignment of three modalities using P2FA [1].
- interval duration of each word utterance as a time-step.
- Calculate the expected video and audio features over the word utterance time interval.
- Pre-processing:
  - If word count >= 20: take the last 20 words.
  - If word count < 20: fill insufficient features with 0s.

1. Yuan, J., and Liberman, M. 2008. Speaker identification on the scotus corpus. *Journal of the Acoustical Society of America* 123(5):3878.

Provided Dataset

- Training Data
  - train_labels = (1283,1)
  - train_visual = (1283, 20, 46)
  - train_audio = (1283, 20, 46)
  - train_text = (1283, 20, 300)

Validation Data

valid_labels = (229,1)
valid_visual = (229, 20, 46)
valid_audio = (229, 20, 46)
valid_text = (229, 20, 300)

- Test Data
  - test_labels = (686,1)
  - test_visual = (686, 20, 46)
  - test_audio = (686, 20, 46)
  - test_text = (686, 20, 300)

- Other features: an Ipython Notebook for downloading aligned features over visual features or audio features

Baselines

	7-way Multiclass	**Regression**
Method	Error rate	MAE
Early Fusion SVM	26.5	1.1
Early Fusion LSTM	33.5	1.02

- MAE: Mean Average Precision