AI in Animation: Expressions, Lip Sync, Articulated Parts, and Audio

Deepali Aneja

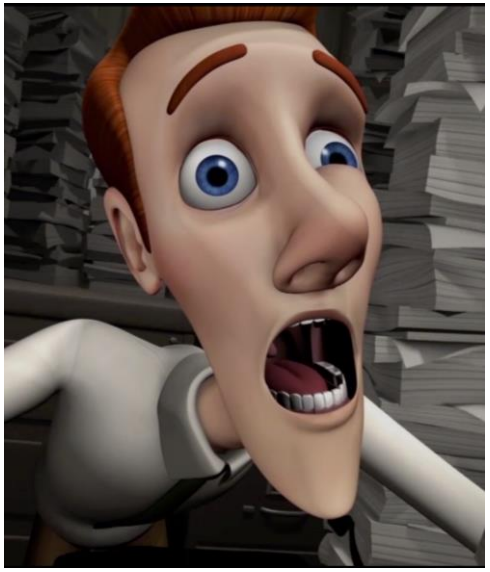
PAUL G. ALLEN SCHOOL
OF COMPUTER SCIENCE & ENGINEERING

Adobe RESEARCH

Expressive characters

Animated Shorts from Animation Research Labs, University of Washington.

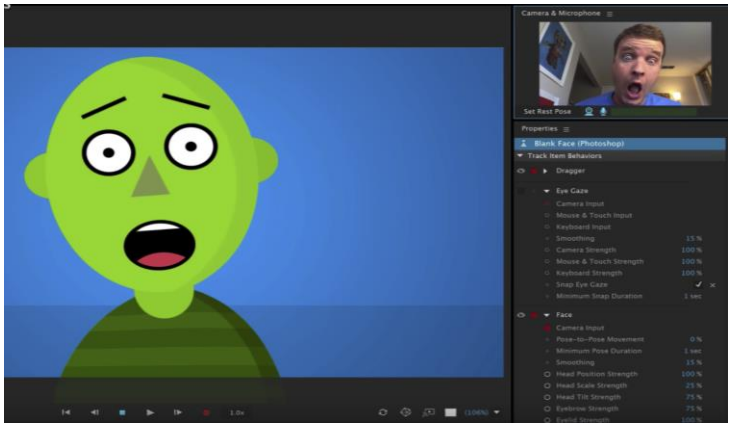
Understand and model how people perceive expressive characters



Animated Shorts from Animation Research Labs, University of Washington.

Performance-based animation

Motion Capture



Character Animator

FaceWare

Animoji

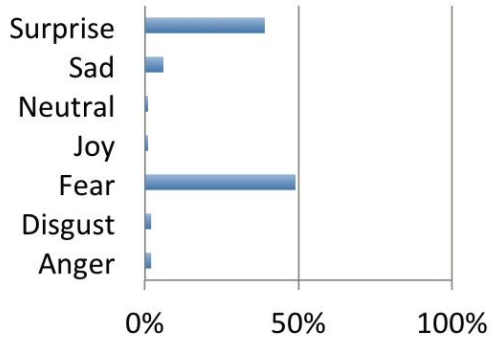
Outline

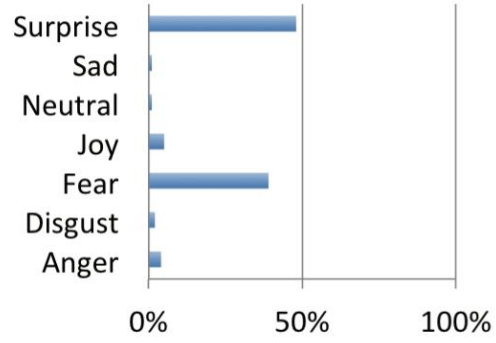
- Expressions
- Lip Sync
- Articulated Parts
- Interactive Audio

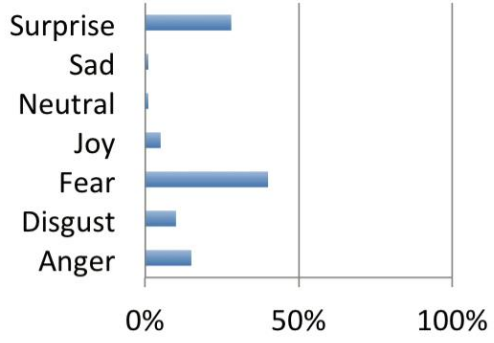
Outline

- Expressions
- Lip Sync
- Articulated Parts
- Interactive Audio

Artist-based animation







Expression transfer: Faceshift

Geometric features transfer, expressions often do not!

Actual : Disgust

Perceived : Angry

Actual : Sad

Perceived : Confused

Beyond human geometry

Retargeting convincing 3D character expressions

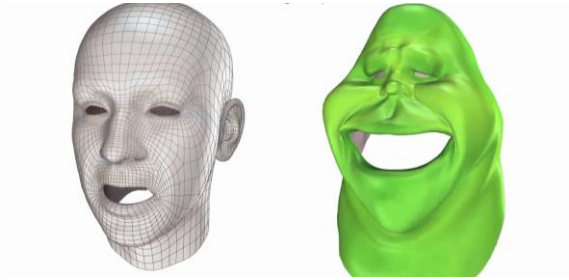
Problem: Geometry/FACS-based expression retargeting is not always perceptually accurate and often lacks expressiveness

Approach: Develop data-driven perceptually-valid expression retargeting model

- Modeling Stylized Character Expressions via Deep Learning
  Aneja et al., ACCV 2016
- *Learning to Generate 3D Stylized Character Expressions from Humans*
  Aneja et al., WACV 2018

Related Work

Cao et al. (SIGGRAPH 2013)



Weise et al. (SIGGRAPH 2011)

Li et al. (SIGGRAPH 2013)

Cao et al. (SIGGRAPH 2016)

Bouaziz et al. (EPFL 2014)

Bouaziz et al. (SIGGRAPH 2013)

Expression Retrieval

Retrieve characters using

perceptual model mapping and human geometry

- Use deep learning to learn mappings between
  - Human expressions and characters expressions
  - Humans and humans
  - Characters and characters
- This is not only geometric mapping
  - It is perceptual modelling of expressions!

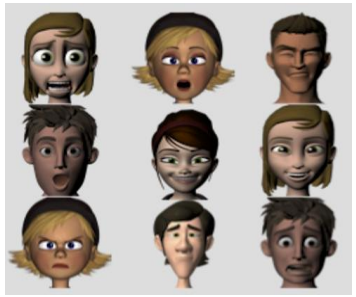
Expression Retrieval

Step 1

Use deep learning to create a perceptual model of human expressions

Human feature space

$f($  $)$



Step 2

Learn analogous character model

Character feature space

$f'($  $)$

$f($  $)$

Step 3

Learn Mapping

$f'($  $)$

Step 4

Retrieve characters using perceptual model mapping and human geometry

Training Data

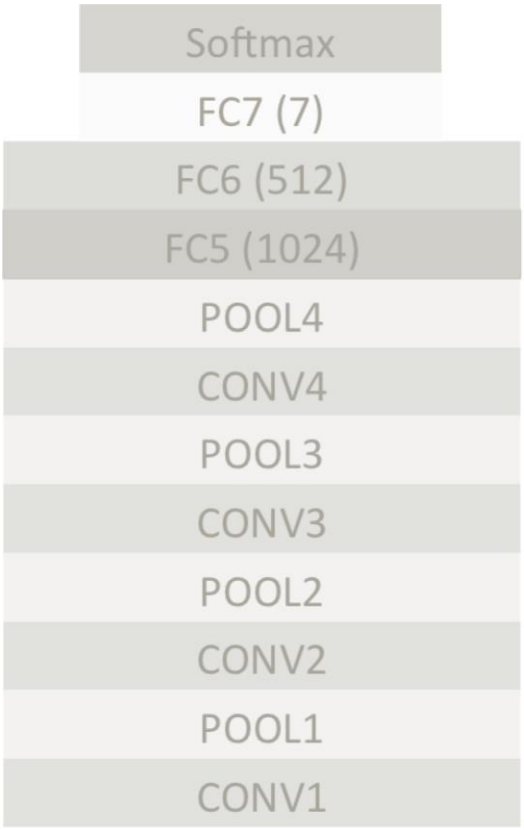
- Seven classes : Anger, Disgust, Fear, Joy, Neutral, Sad and Surprise

- Stylized Characters expression database
  - Total of 70K images
  - **Facial Expression Research Group (FERG-DB) is publicly available.**

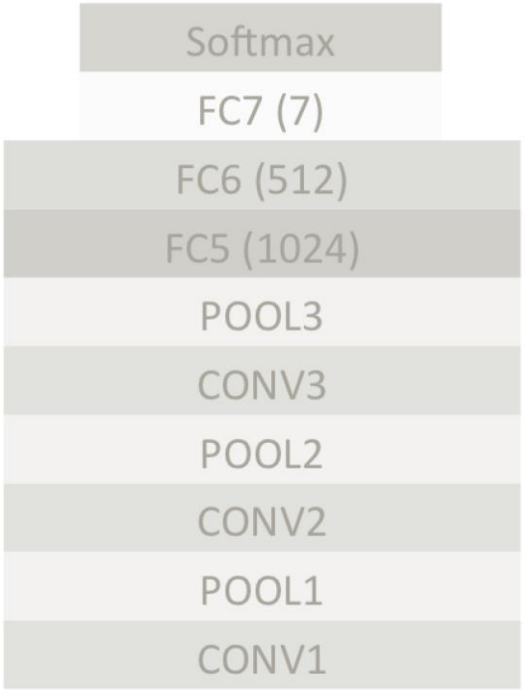
- Human expression database
  - Total of 75K images
    - CK+: The Extended Cohn-Kanade
    - DISFA: Denver Intensity of Spontaneous Facial Actions
    - KDEF: The Karolinska Directed Emotional Faces
    - MMI

FERG-DB Stylized character expression database: http://grail.cs.washington.edu/projects/deepexpr/ferg-db.html

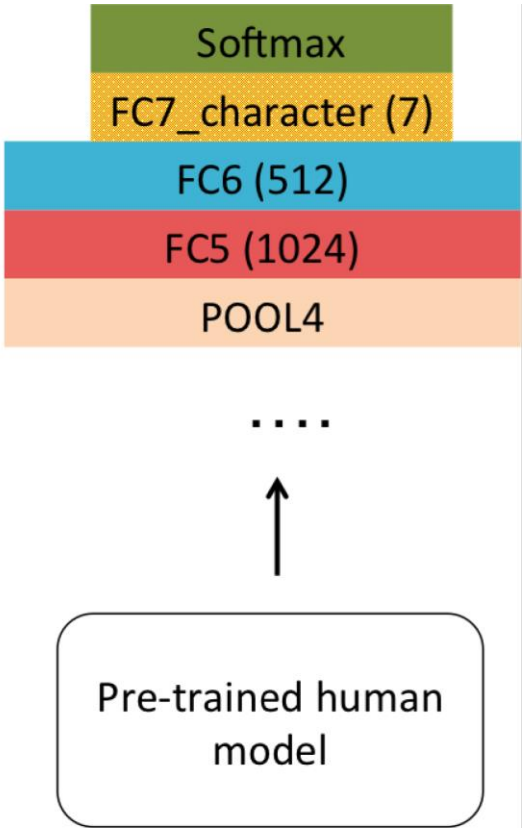
Network Architecture



Human CNN (HCNN)

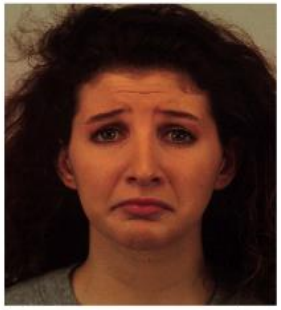


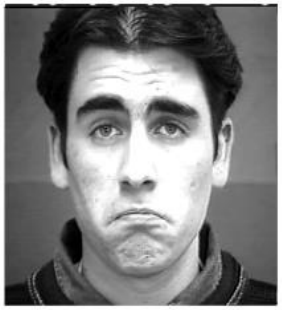
Character CNN (CCNN)

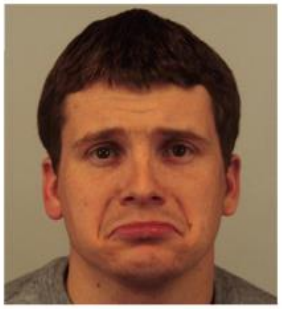


Transfer Learning Shared CNN (SCNN)

Retrieval Results







Query (Sad)

Query (Surprise)

Distance Metrics

- Extracted features from the **last fully connected layer (FC6)** of both the models: HCNN and SCNN and normalized the feature vectors

$$\phi_d = \alpha\,|\text{JS Distance}| + \beta\,|\text{Geometric Distance}|$$

Expression feature vectors
(N-1) Layer features

Geometry feature vectors

Lin, J.: Divergence measures based on the shannon entropy. Information Theory, IEEE Transactions on 37 (1991) 145–151 44.

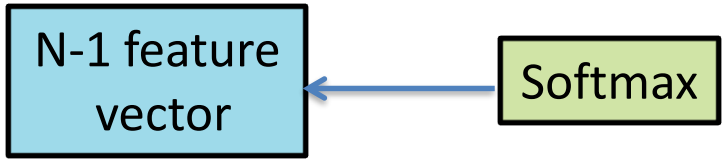
Distance Metrics

- Extracted features from the **last fully connected layer (FC6)** of both the models: HCNN and SCNN and normalized the feature vectors

$$\phi_d = \alpha\,|\text{JS Distance}| + \beta\,|\text{Geometric Distance}|$$

Expression feature vectors
(N-1) Layer features

Geometry feature vectors



Back propagation

Distance Metrics

- Extracted features from the **last fully connected layer (FC6)** of both the models: HCNN and SCNN and normalized the feature vectors

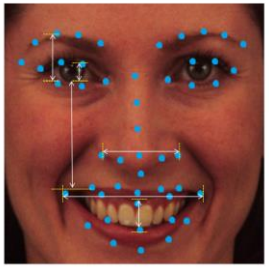
$$\phi_d = \alpha\,|\text{JS Distance}| + \beta\,|\text{Geometric Distance}|$$

Expression feature vectors
(N-1) Layer features

Geometry feature vectors

N-1 feature vector ← Softmax

Back propagation



Character Retrieval

Query

Multiple retrieval results for the joy query image

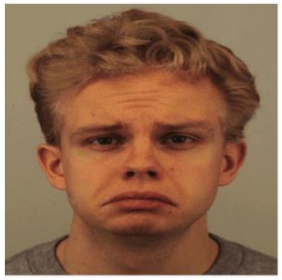
Query

$$\phi_d = \alpha \left| \text{JS Distance} \right|$$

Character Retrieval

Query

Character retrievals sorted by geometry



$$\phi_d = \alpha\,|\text{JS Distance}| + \beta\,|\text{Geometric Distance}|$$

Results

Query

Top matches of Character retrievals

Average Retrieval Score
(for each expression across all characters)

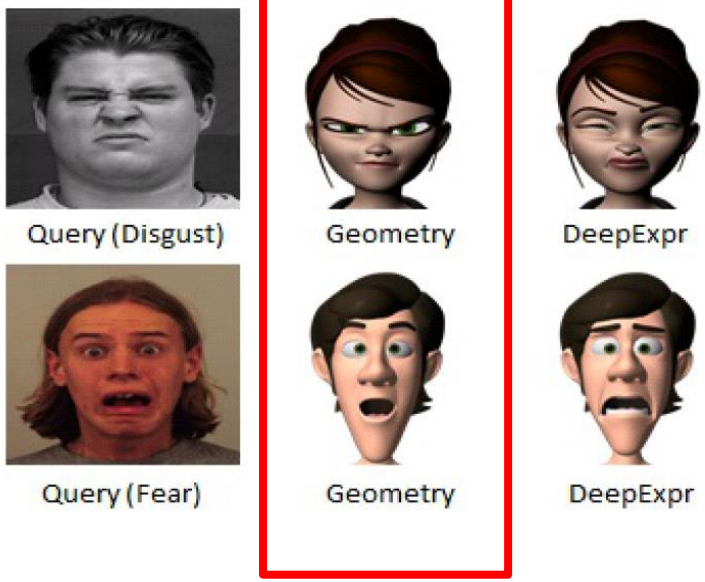
$$score(q) = \frac{1}{1 - N \cdot N_{rel}} \left(\sum_{k=1}^{N_{rel}} R_k - \frac{N_{rel}(N_{rel}+1)}{2} \right)$$

Expression	Geometry	DeepExpr
Anger	0.384	0.213
Disgust	0.386	0.171
Fear	0.419	0.228
Joy	0.276	0.106
Neutral	0.429	0.314
Sad	0.271	0.149
Surprise	0.322	0.125

Average Retrieval Score
(for each expression across all characters)

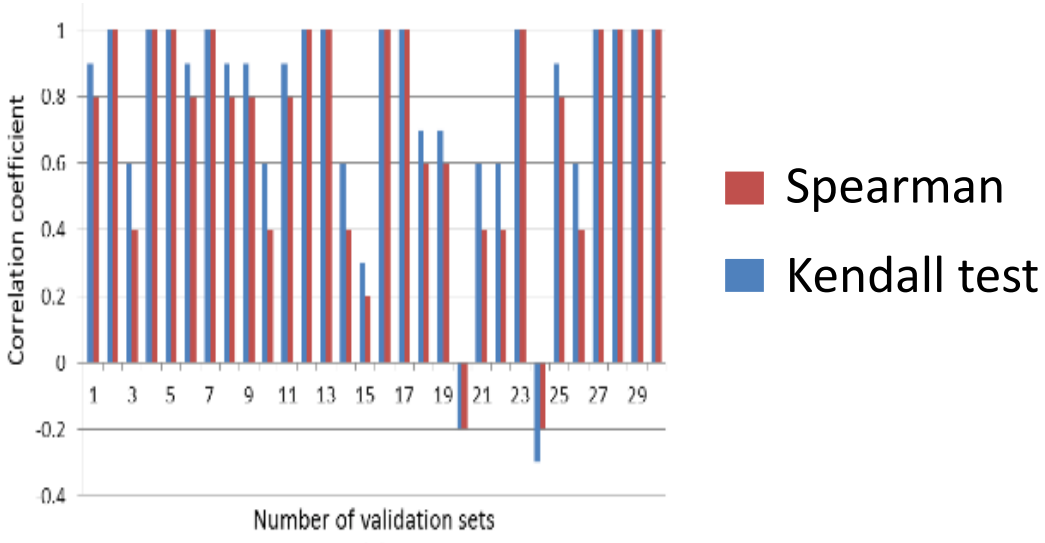
$$score(q) = \frac{1}{1 - N \cdot N_{rel}} \left(\sum_{k=1}^{N_{rel}} R_k - \frac{N_{rel}(N_{rel}+1)}{2} \right)$$

Expression	Geometry	DeepExpr
Anger	0.384	0.213
Disgust	0.386	0.171
Fear	0.419	0.228
Joy	0.276	0.106
Neutral	0.429	0.314
Sad	0.271	0.149
Surprise	0.322	0.125



Correlation

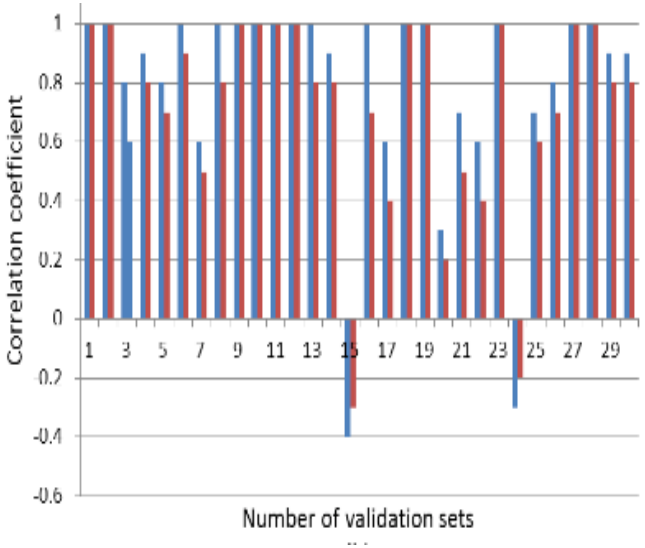
Correlation with Expert



Spearman

Kendall test

Correlation with MT subjects

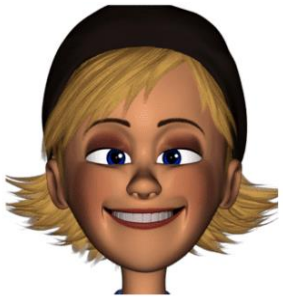


- Spearman correlation with expert best rank is **0.934** and with MT best rank is **0.942**
- Kendall correlation with expert best rank is **0.910** and with MT best rank is **0.927**

Failure cases

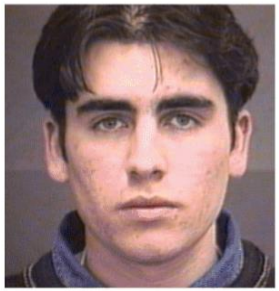
Query

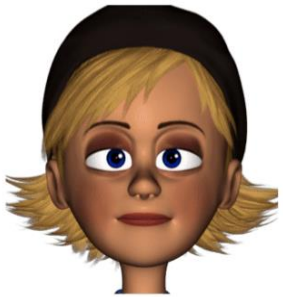
Top matches of Character retrievals



Joy

Surprise

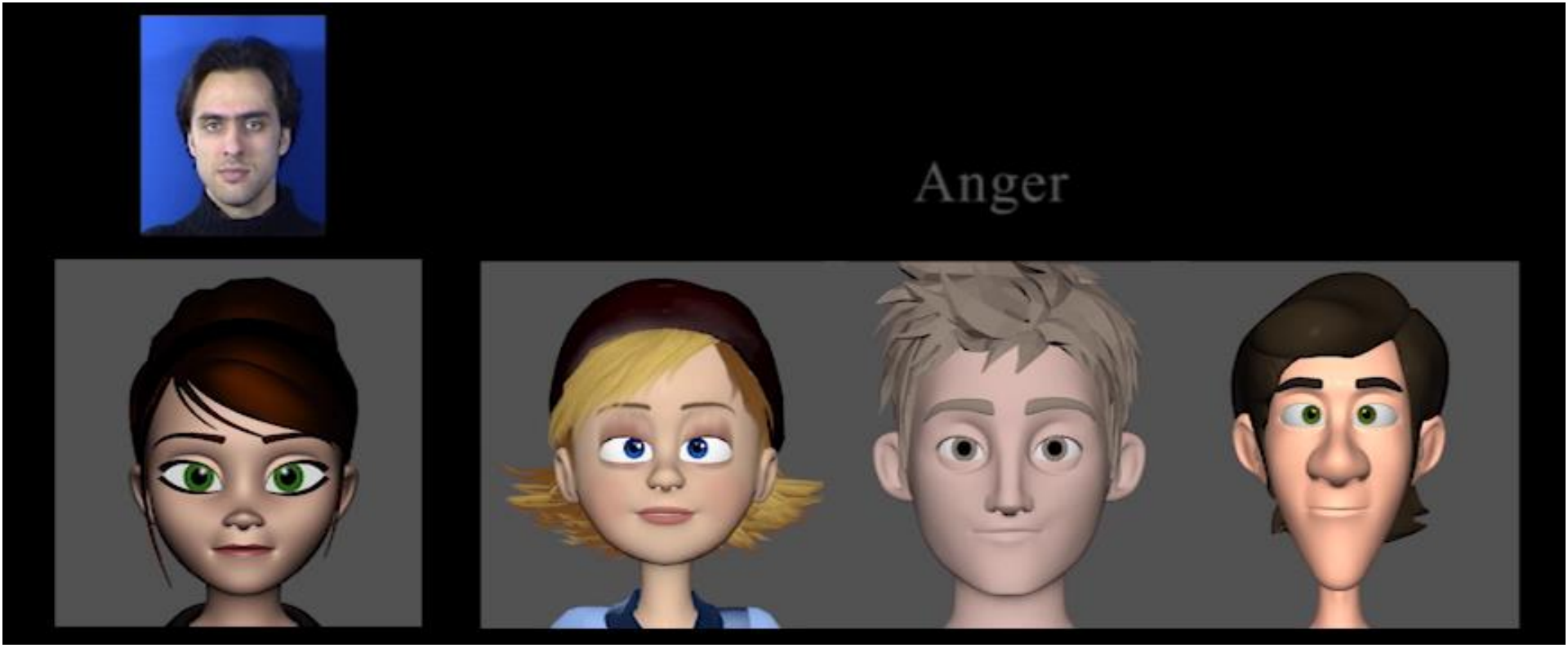




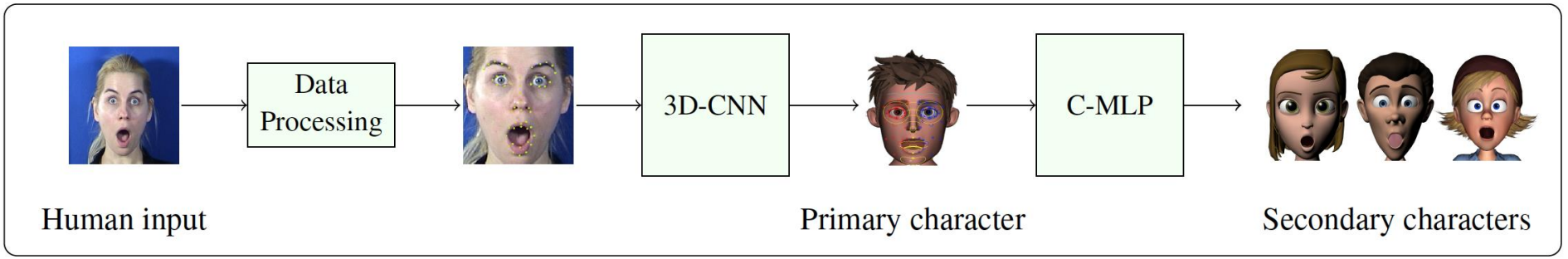
Neutral

Sad

Expression Retargeting in 3D



Approach

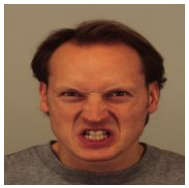


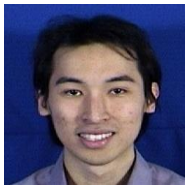
3D-CNN: Deep convolutional neural network for human to character transfer

C-MLP: Lightweight multi-layer perceptron for character to character transfer

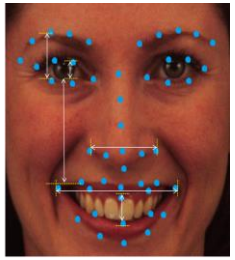
Data and Preprocessing

Expression classes : Anger, Disgust, Fear, Neutral, Joy, Sadness, Surprise





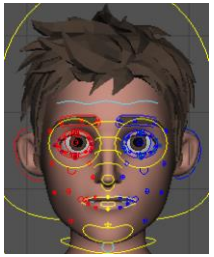
SFEW, CK+, MMI, KDEF and DISFA
~100k images

Bounding box and landmark detection; frontalization and geometry extraction

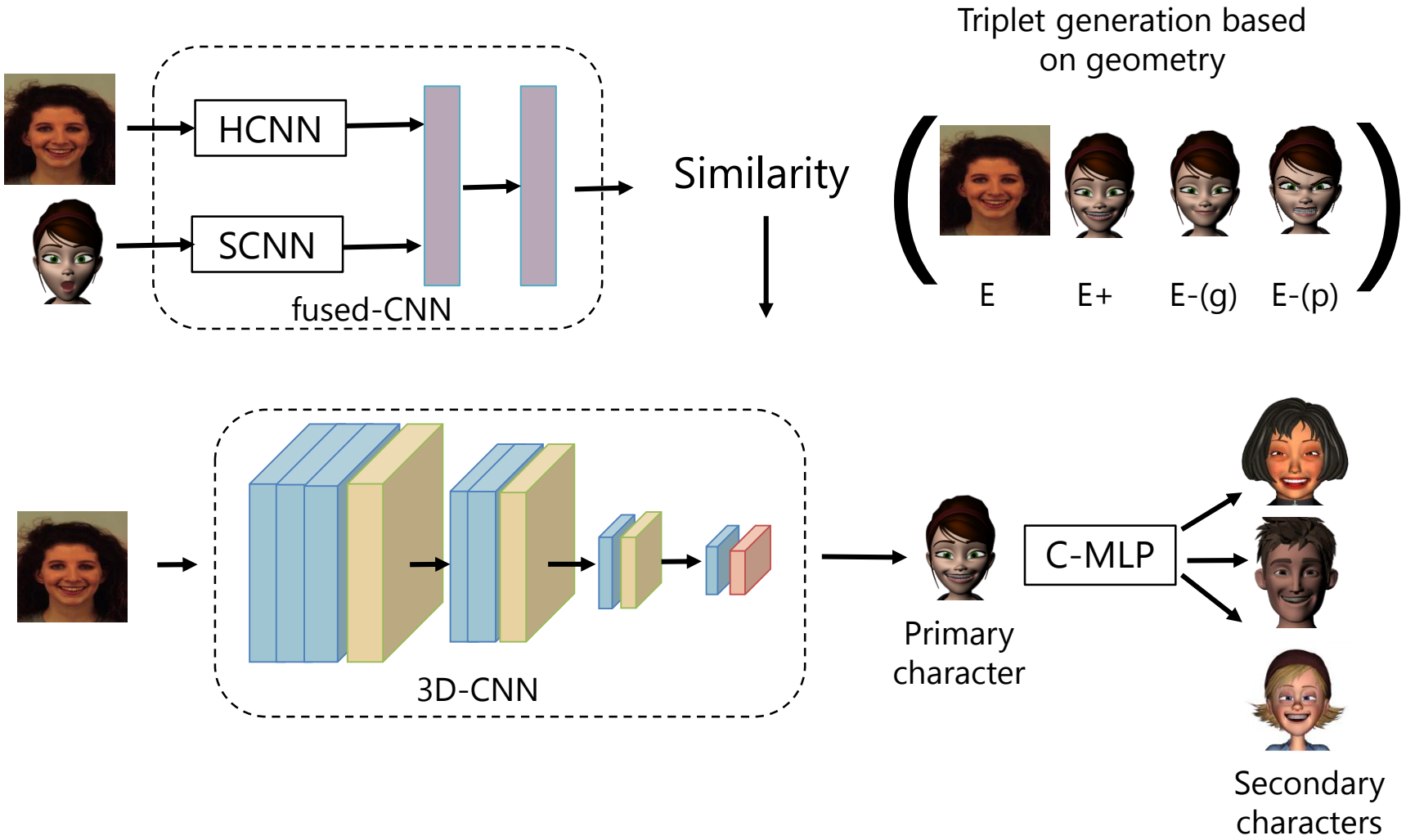


FERG-DB + 3 new characters
~55K images




3D rig parameters; geometry extraction

Network Architecture

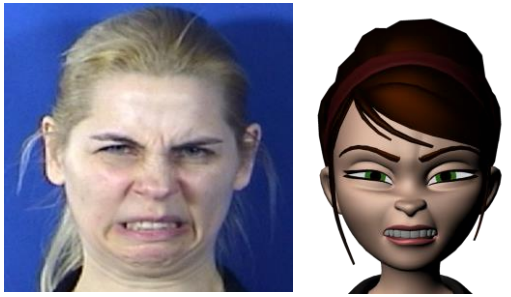


Results with Primary Character

Expression: Sadness

Expression: Surprise

Expression: Disgust



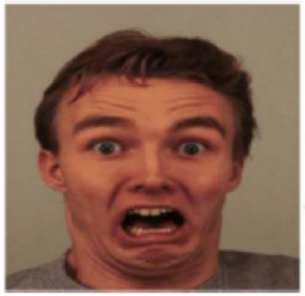
Expression: Anger

Expression: Fear

Expression: Joy

Results with multiple characters

Input Primary Secondary New characters

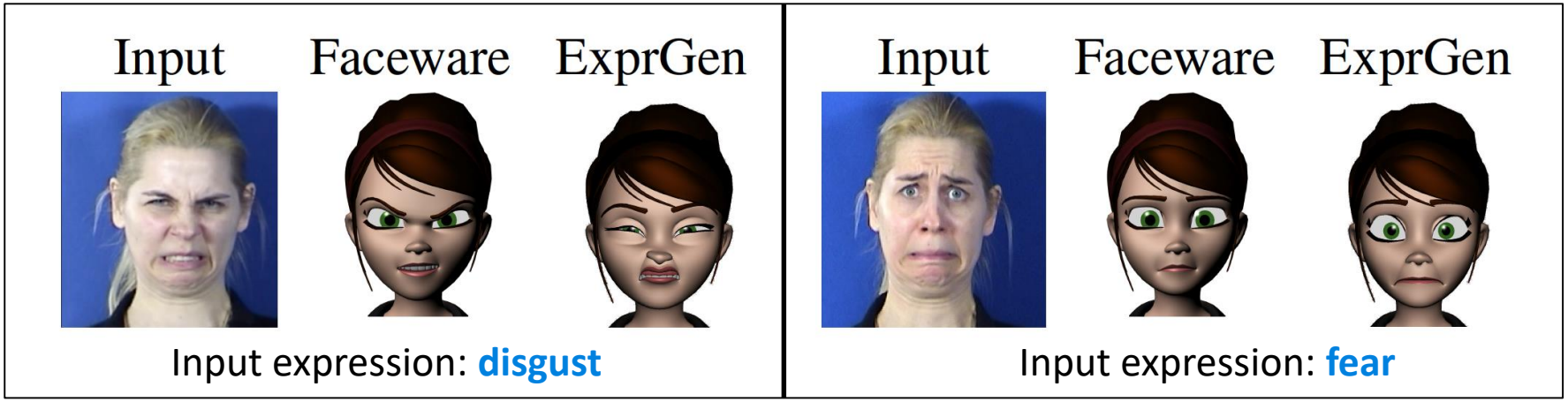




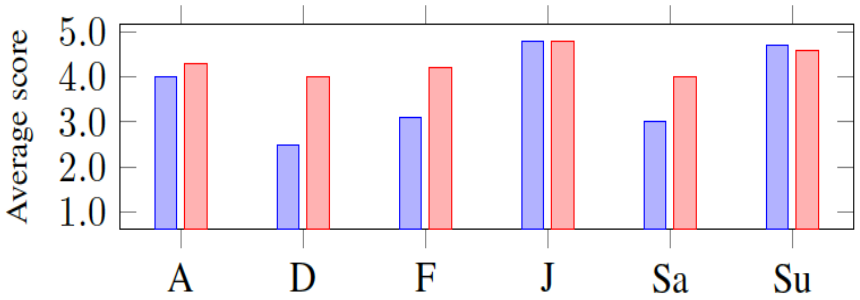




Evaluation



Blue: Faceware, Red: Ours



Quantitative comparison
Faceware (blue bars) and ExprGen (red bars)

Contributions

- Developed a perceptually valid method to map human facial expressions to 3D stylized character rig controls
- Novel stylized character data set (FERG-DB)
- Semi-supervised method to enable expression transfer between multiple characters

Outline

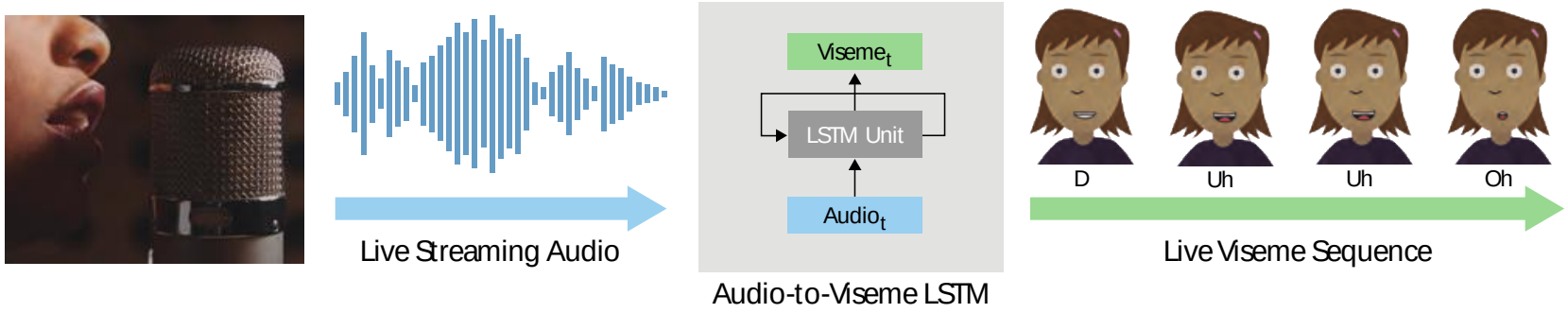
- Expressions
- Lip Sync
- Articulated Parts
- Interactive Audio

Real-Time Lip Sync for Live 2D Animation

Problem: Automatic lip sync has errors in timing (viseme transition) and classification accuracy (which viseme)

Approach: Learn data-driven mapping from audio to visemes

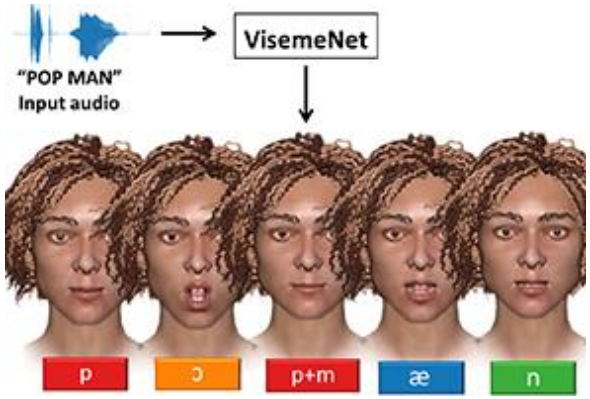
- Improved Lip sync (Adobe Character Animator CC 2018)
  Aneja and Li, shipped in version 1.0



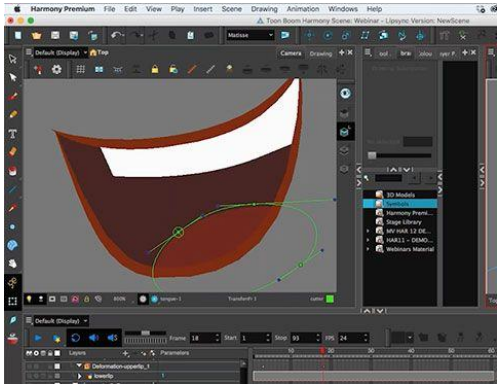
Related Work

Suwajanakorn et al (SIGGRAPH 2017)

Taylor et al (SIGGRAPH 2017)



Zhou et al (SIGGRAPH 2018)

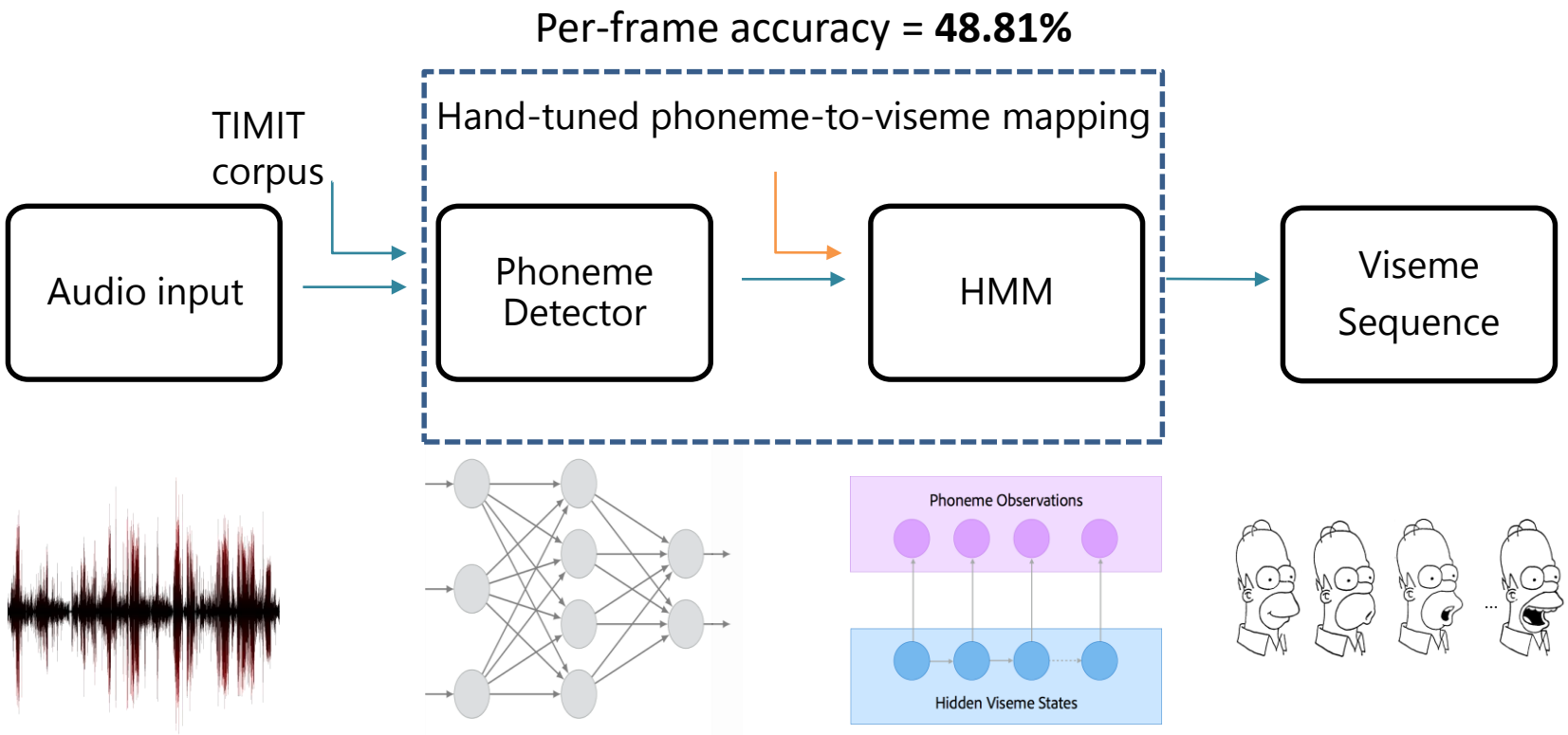


ToonBoom Harmony

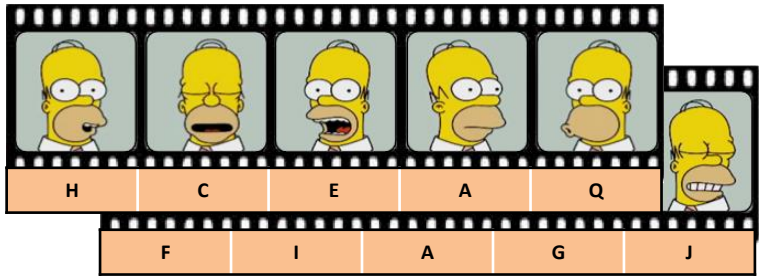
2D Live Lip Sync challenges

- Constrained palette
- Transition timing
- Hand-authoring is time-intensive

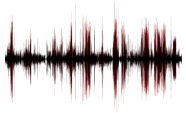
Character Animator Pipeline (2016)



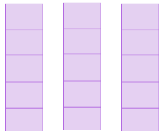
Improved Lip Sync



Viseme labels



Audio



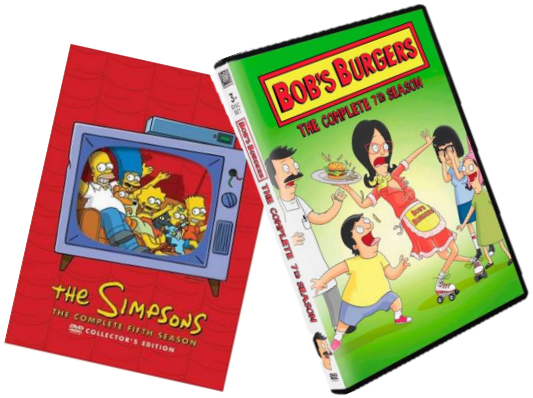
Phoneme Probabilities

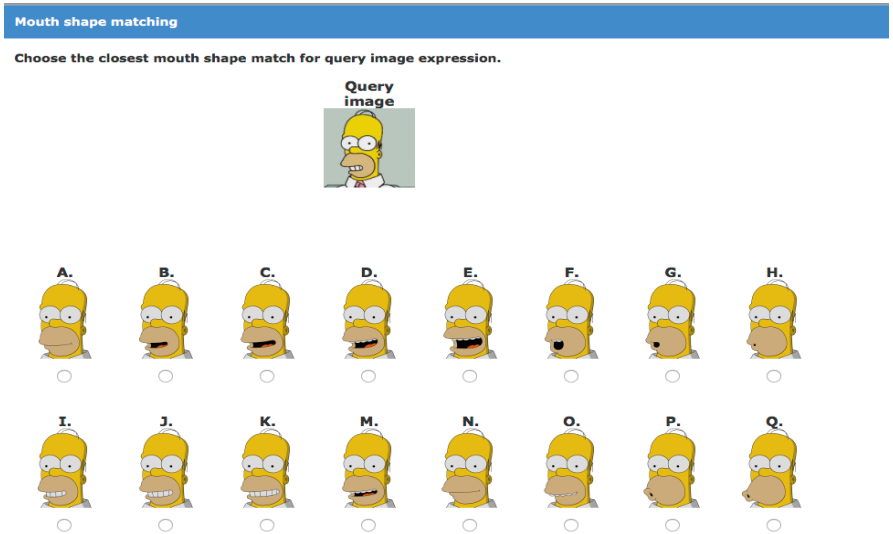
Learn per frame phoneme-to-viseme mapping

- ~200 secs of Simpsons labelled data
- Implemented a data-driven HMM
  - Learnt transition probabilities
  - Included emission probabilities

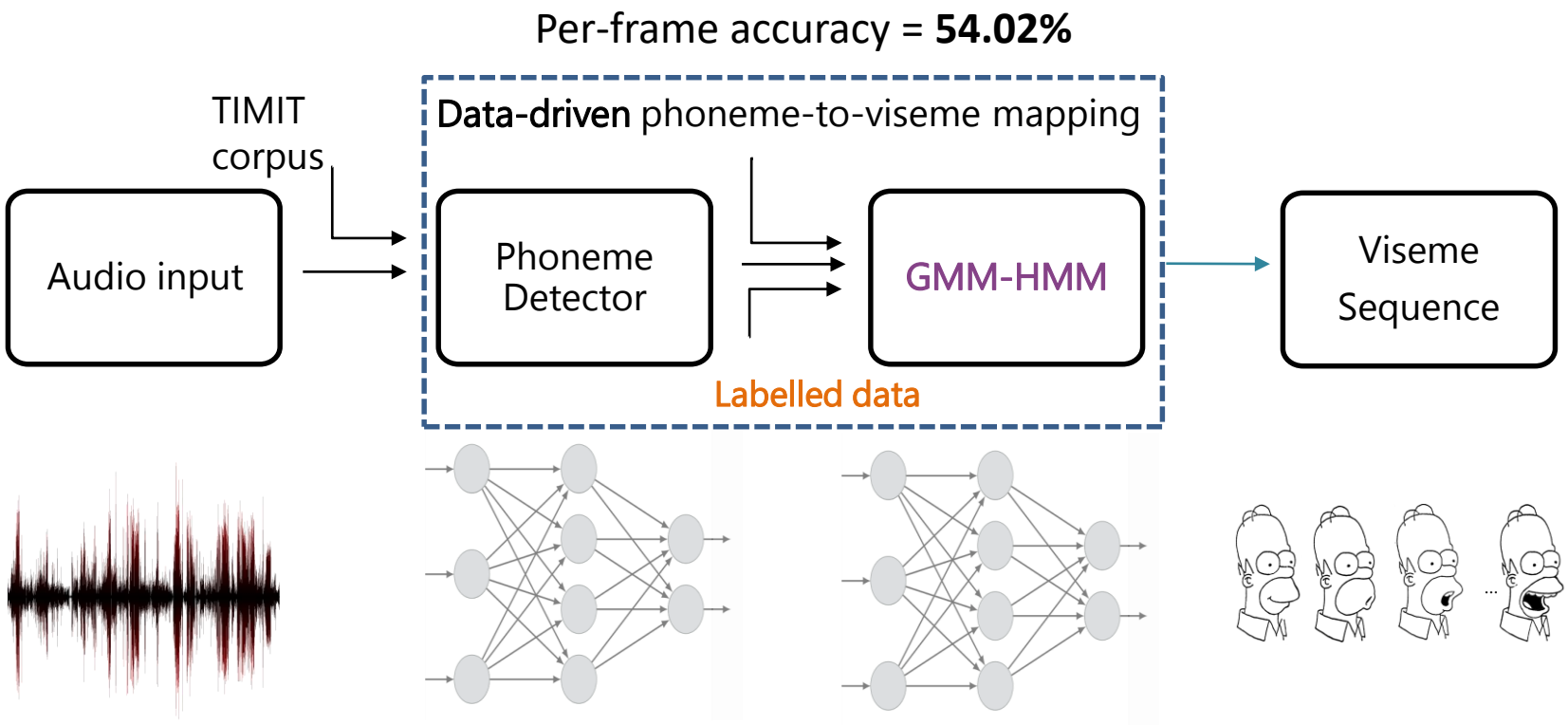
Data Collection (2017)

- Crowdsourcing (Mechanical Turk)
- Viseme labels for ~18 mins of Simpsons and Bob's Burger videos

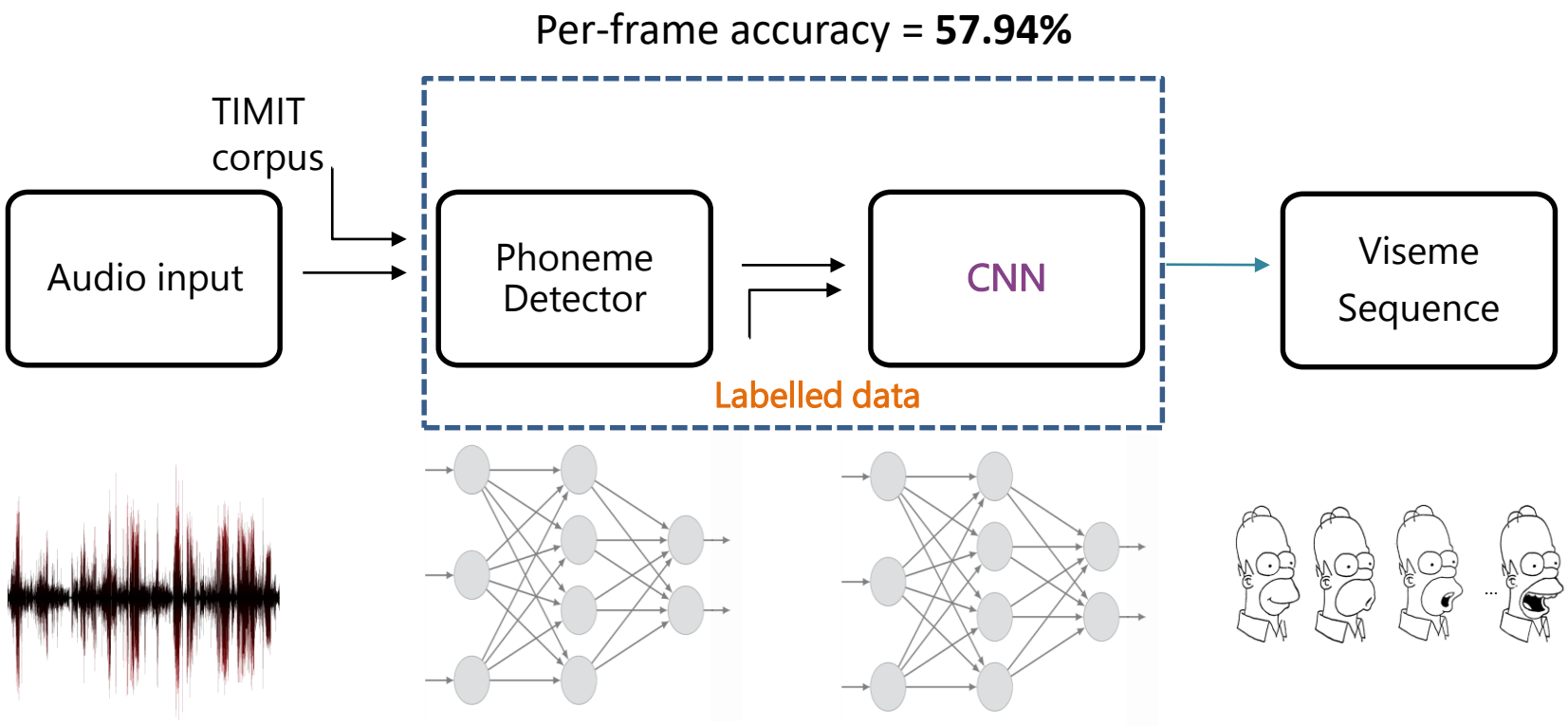




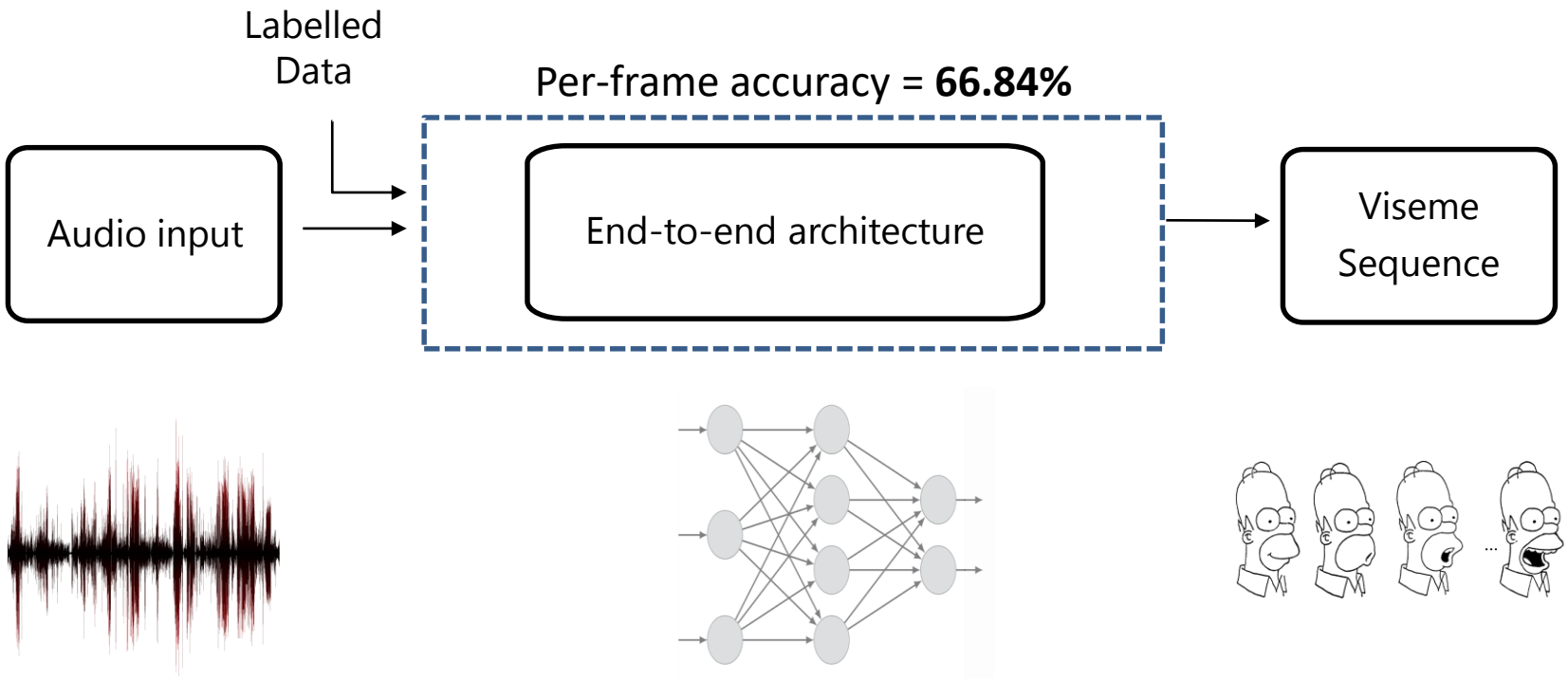
Character Animator Pipeline (2017)



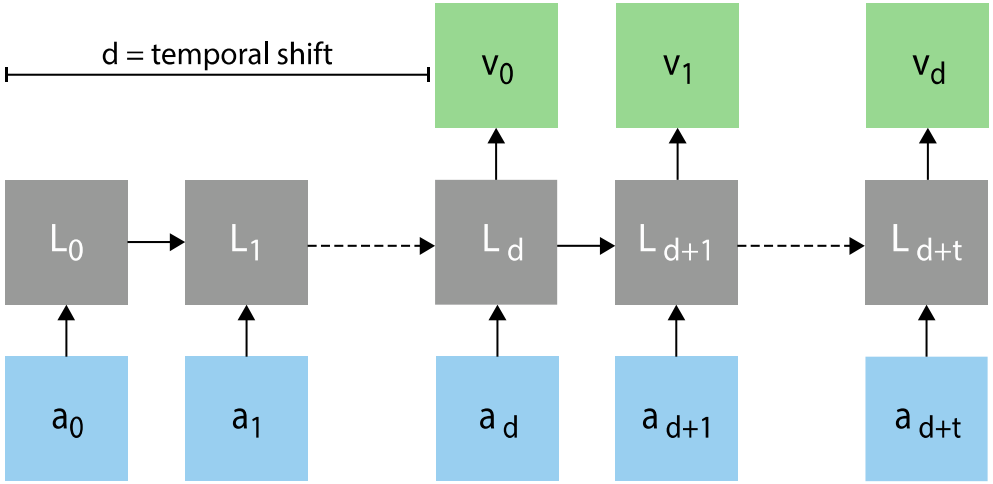
Character Animator Pipeline (2017)



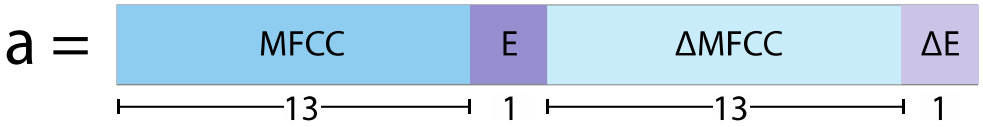
Character Animator Pipeline (2017)



Model



(a) LSTM Architecture

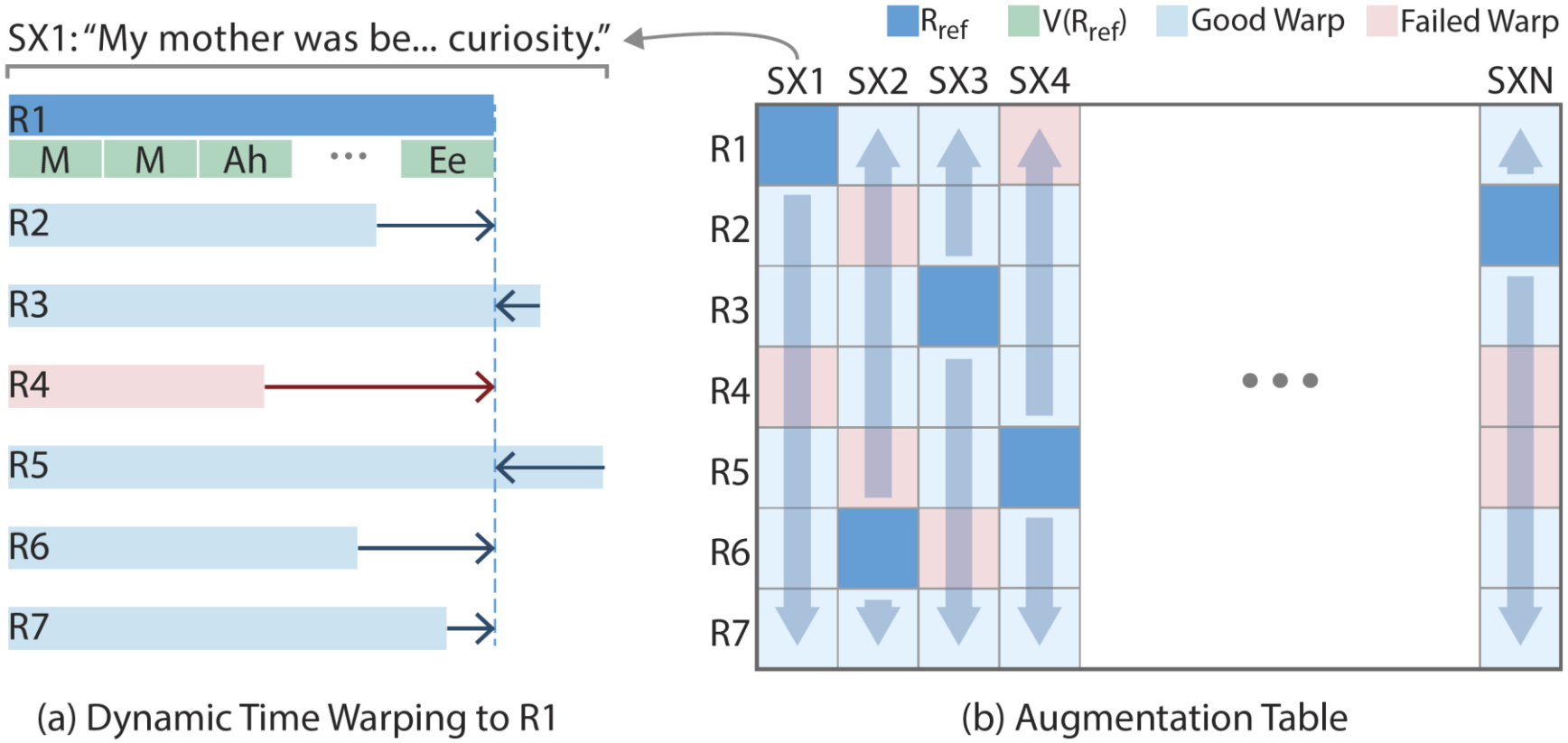


(b) Audio Feature Vector

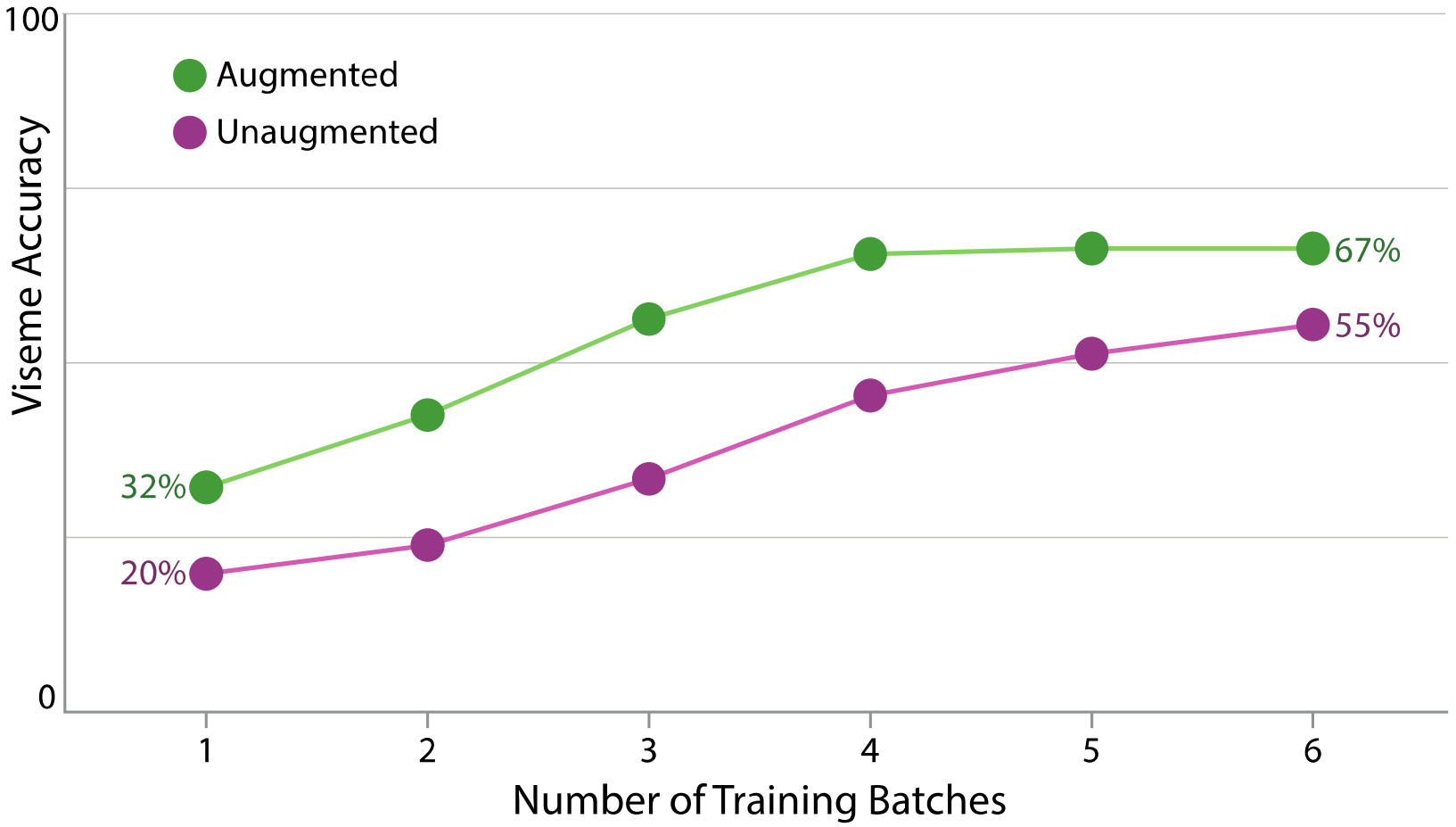
Data

- Hand – animated audio sequences (20 mins)
- Training data:
  - TIMIT dataset (3-4 secs each)
  - Data Augmentation [80 mins of data]
- Test data: 50 test recordings (3-4 secs each)
  - TIMIT test set : 25 recordings (random)
  - Non-TIMIT test set : 25 recordings (random)

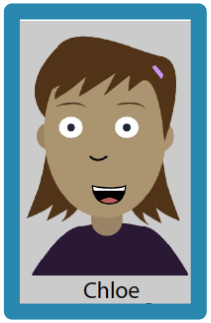
Dataset Augmentation



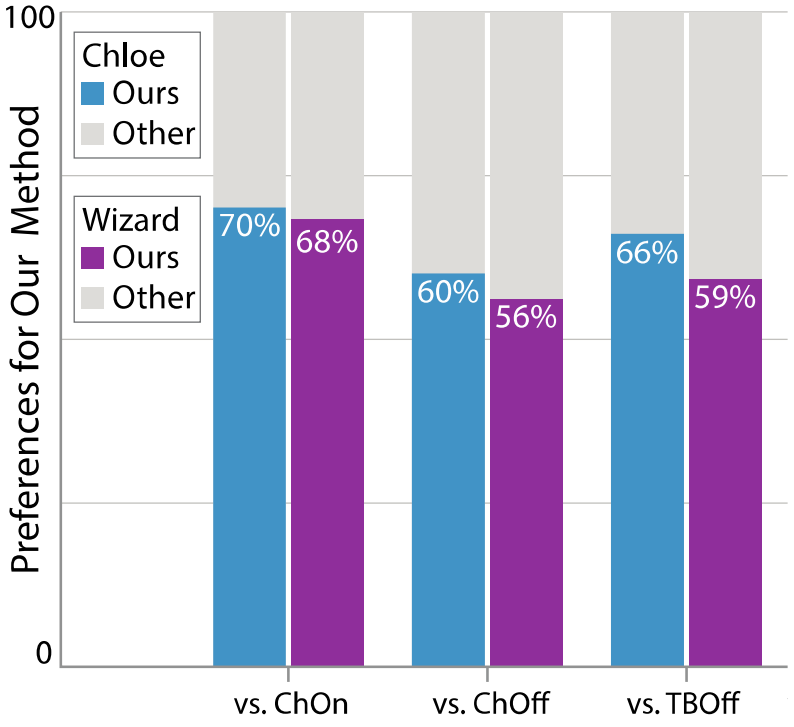
Impact of Data Augmentation



Comparisons



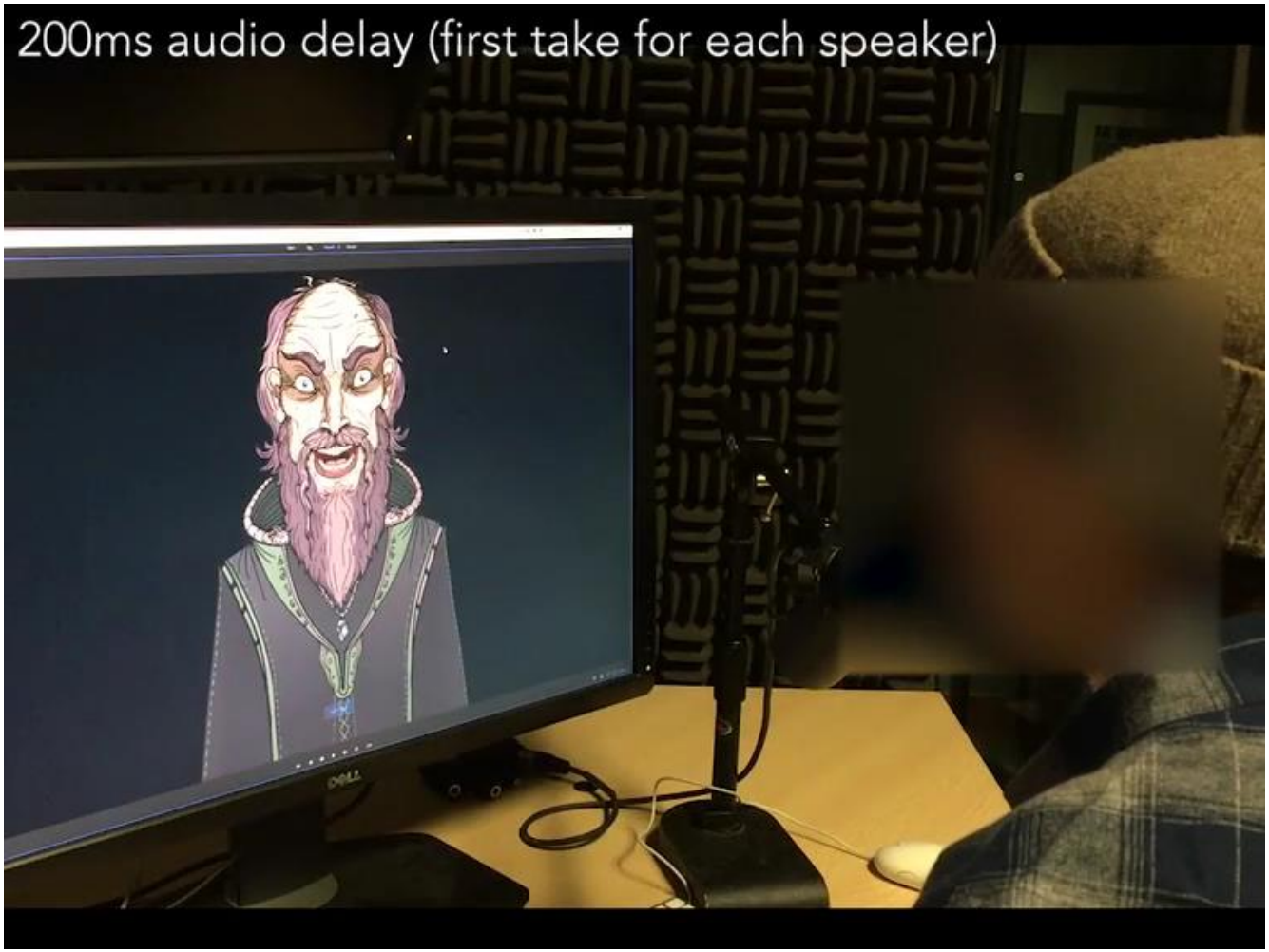
We collected **20** judgements for every recording (10 for each puppet), which resulted in **1000 judgements** for each competing method.



ChOn: Character Animator Online
ChOff: Character Animator Offline
TBOff : ToonBoom Offline

Comparisons with Competing Methods

Live Lip Sync



Contributions

- Developed an accurate, speaker generalizable and **low-latency** lip sync model for 2D live animation

- Performed **extensive human judgement experiments** demonstrating that our technique improves upon existing state-of-the-art 2D lip sync engines, most of which require offline processing.

Outline

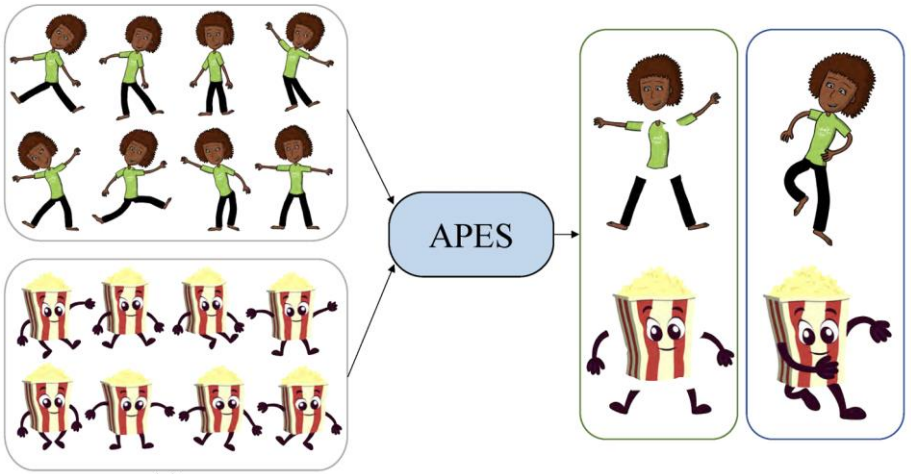
- Expressions
- Lip Sync
- **Articulated Parts**
- Interactive Audio

Articulated Parts

Problem: Partitioning characters into independently moving parts is tedious and time-consuming

Approach: Automatically identify articulated parts from a small set of character poses shown in sprite sheet using ML techniques

- Apes: Articulated part extraction from sprite sheets
  Xu et al., CVPR 2022



APES: Articulated Part Extraction from Sprite Sheets

Zhan Xu[1,2] Matthew Fisher[2] Yang Zhou[2] Deepali Aneja[2]
Rushikesh Dudhat[1] Li Yi[3] Evangelos Kalogerakis[1]

[1]University of Massachusetts Amherst [2]Adobe Research [3]Tsinghua University

Contributions

- A method for analyzing a sprite sheet and creating a corresponding articulated character that can be used as a puppet for character animation
- A neural architecture to predict pixel motions and cluster pixels into articulated moving parts without relying on a known character template
- An optimization algorithm for selecting the character parts that can best reconstruct the given sprite poses

Outline

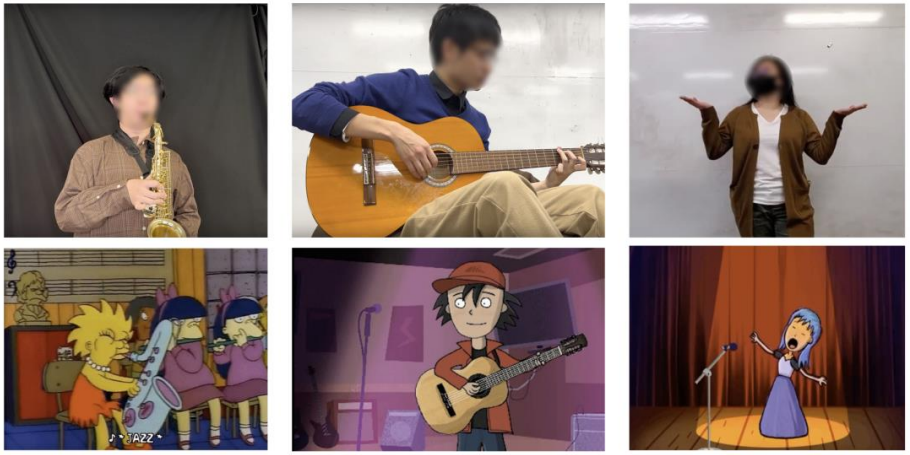
- Expressions
- Lip Sync
- Articulated Parts
- Interactive Audio

Interactive Audio

Problem: Existing audio-driven live animation tools focus on speech and have little or no support for non-speech sounds

Approach: Exemplar-based authoring tool for interactive, audio-driven animation focusing on non-speech sounds

- SoundToons: Exemplar-Based Authoring of Interactive Audio-Driven Animation Sprites
  Chong et al., IUI 2023



SoundToons: Exemplar-based Authoring of Interactive Audio Driven Animation Sprites

Toby Chong[1,2] Deepali Aneja[2] Takeo Igarashi[1] Valentina Shin[2]

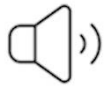
東京大学 THE UNIVERSITY OF TOKYO 1

Adobe 2

SoundToons

Exemplar-Based Authoring of Interactive Audio-driven Animation Sprites.

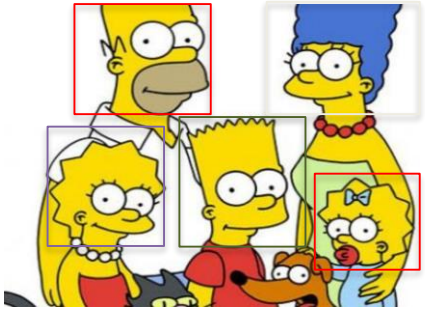
Supplemental Video

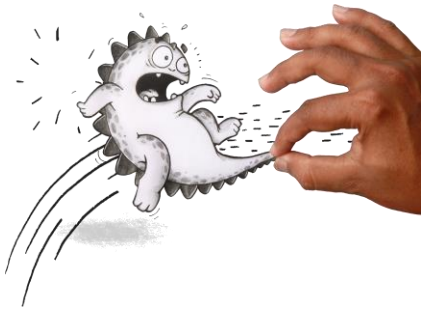
Contains Audio 

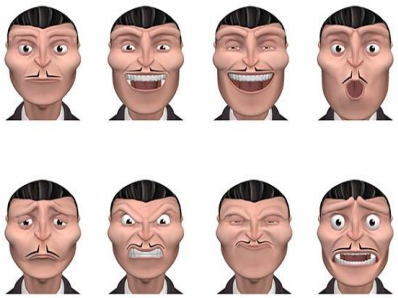
Contributions

- Developed an interactive tool for creating custom, interactive audio-driven animation sprites through exemplar-based authoring
- Adaptation and extension of a few-shot sound detection model to detect multiple different sound events and trigger corresponding animation events
- Developed an optimization method for inferring a smooth and continuous mapping between audio features and animation parameters

Applications







Thanks!