DATAHACK
SEP 2016

SimPhony

A Data-Driven Approach for Video Creation

Doron Kukliansky

Zach Moshe, Neta Livneh, Yahel Guberman, Ofer Fridman

Outline

- Hackathon project
- Speaker recognition
- Semantic sentence similarity

The Problem

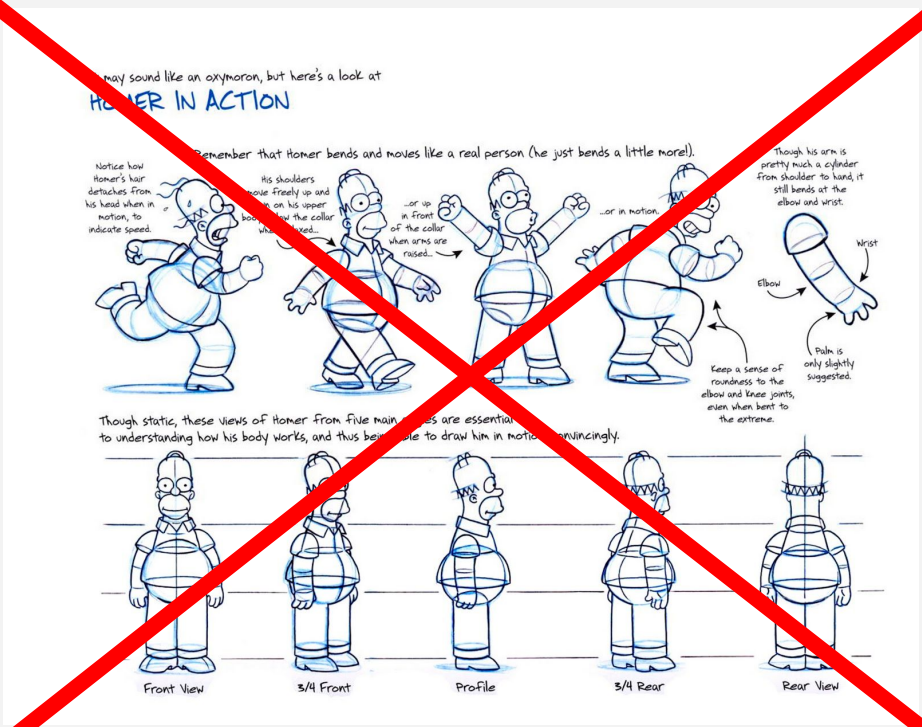
- Not enough Simpsons episodes!
- There are only
  - 27 seasons
  - 596 episodes
  - 14000 minutes
  - 90GB of data

The Problem

There aren’t episodes on every subject.

Goal: Given a script, automatically create a new Simpsons episode.

Solution


may sound like an oxymoron, but here's a look at
HOMER IN ACTION
Remember that Homer bends and moves like a real person (he just bends a little more!).
Notice how Homer's hair detaches from his head when in motion, to indicate speed.
His shoulders
...or up in front of the collar when arms are raised...
...or in motion.
Though his arm is pretty much a cylinder from shoulder to hand, it still bends at the elbow and wrist.
Wrist
Elbow
Palm is only slightly suggested.
Keep a sense of roundness to the elbow and knee joints, even when bent to the extreme.
Though static, these views of Homer from five main
are essential
to understanding how his body works, and thus bei
le to draw him in motio
nvincingly.
Front View
3/4 Front
Profile
3/4 Rear
Rear View

Data-Driven Approach

- Use existing sub-scenes as building blocks for a dialog-based script
- Example:
  - Homer: I want dinner.
  - Marge: All we have is beer.
  - Homer: I love beer!

Why Simpsons?

- Lots of data
- Characters don’t age
- Characters don’t change clothes
- Speech isn’t perfect

Scene Indexing

- We need to cut video precisely
  - Speech to text
  - Watch all simpsons episodes!
  - If only someone watched all Simpsons episodes...

Subtitles

```
82
00:04:25,760 --> 00:04:27,591
Come on, fat boy, run!

83
00:04:27,680 --> 00:04:28,829
I'm not fat!

84
00:04:28,920 --> 00:04:30,512
I'm just... I'm unfit.
```

Subtitles don’t include the speaker...

Speaker Recognition

- Find data set
- Using text
- Using images
- Using speech

Speaker Recognition Using Speech

- Multiclass classification problem
  - 4 labels + 1 other
- Inference on sound extracted from episodes

Speech Training Data

- Download labeled waves
  - Not the exactly same setting
  - The prior is lost
  - Sound encoding issues?
- Episode scripts
  - Created a script-subtitle alignment script

Speaker Recognition

- Deep learning
- Multiclass logistic regression
  - How to define features?

Mel Cepstrum Features

MFCCs are commonly derived as follows:[1][2]

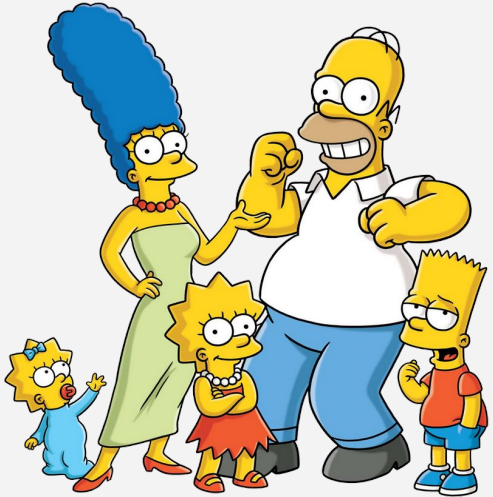
1. Take the Fourier transform of (a windowed excerpt of) a signal.
2. Map the powers of the spectrum obtained above onto the mel scale, using triangular overlapping windows.
3. Take the logs of the powers at each of the mel frequencies.
4. Take the discrete cosine transform of the list of mel log powers, as if it were a signal.
5. The MFCCs are the amplitudes of the resulting spectrum.

```
~$ pip install python_speech_features
```

https://en.wikipedia.org/wiki/Mel-frequency_cepstrum

Classification

- Features are calculated on time windows
  - Combine predictions
- Multiclass logistic regression
  - 80% accuracy on 4 labels



Not Enough Data

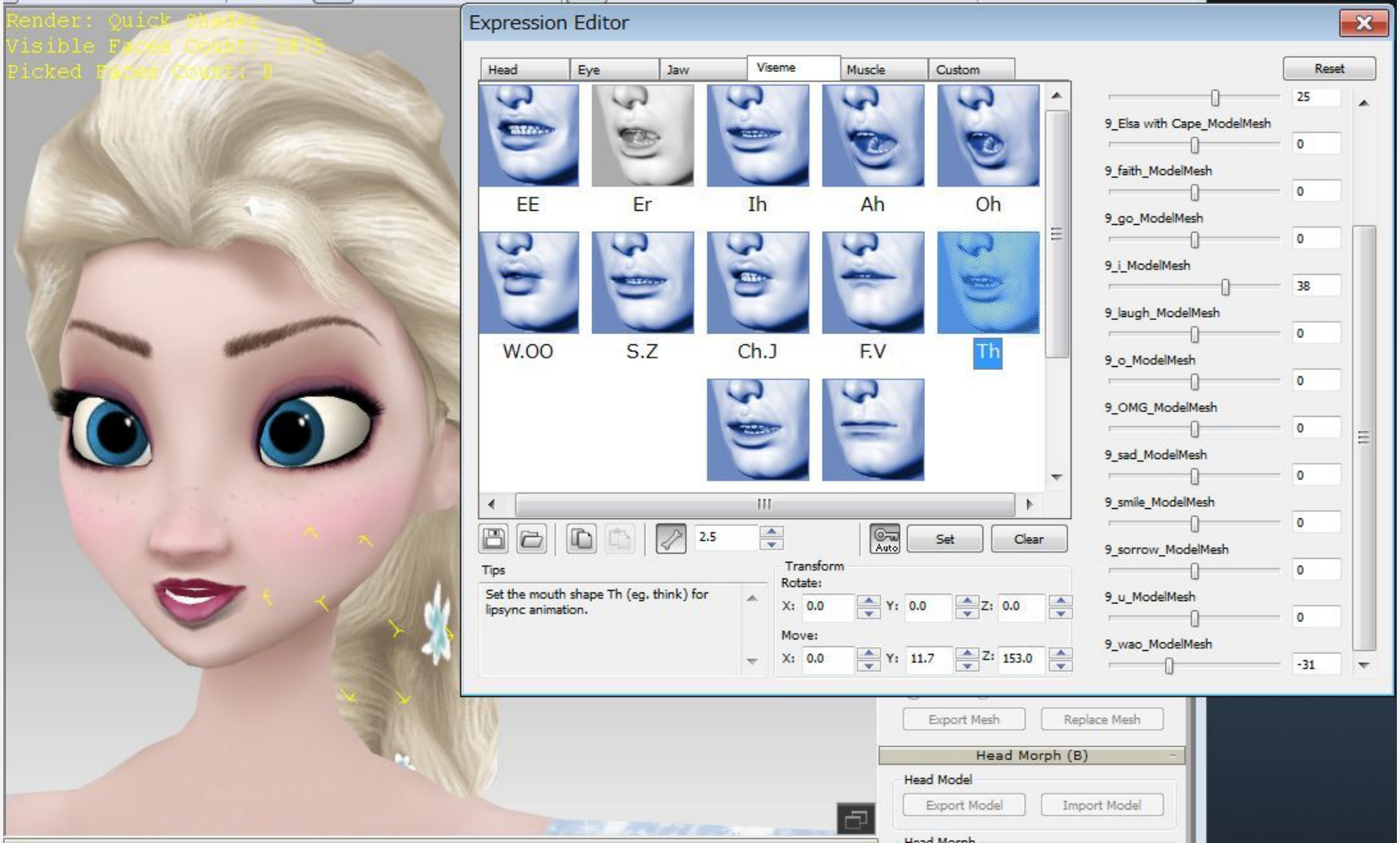
- Not every sentence appears in the episodes we have
- Approaches
  - Text to speech
  - Find similar sentence

Text to speech

- Generate speech from text
- Most lip movement is in the vowels
  - Find sentences with “lip similarity”
    - Food -> “CooC” <- Good

Phonetic Symbols	*Sounds*	*Photos*	*Drawings*
æ, eɪ	at, and, ate		
ʊ, ɜ, ə, r	look, bird, supply, red		
ɑ, ʌ, aɪ	dog, cut, ice		
ɛ, ɪ	end, it		
i, j, s, ʃ z, ʒ	eat, yes, so, show, zoo, vision		
u, oʊ, w	you, no, were		
b, m, p	but, man, pet		
tʃ, t	chat, tea		
d, g dʒ, k, n, ŋ	dim, go, jog, king, new, sing		
ð, l, θ	the, lie, think		
f, v	fat, view		

https://theendisnearus.deviantart.com/art/Phoneme-Chart-505829828


Render: Quick Shader
Visible Faces Count: 3875
Picked Faces Count: 0
Expression Editor
Head
Eye
Jaw
Viseme
Muscle
Custom
Reset
EE
Er
Ih
Ah
Oh
W.OO
S.Z
Ch.J
F.V
Th
2.5
Auto
Set
Clear
Tips
Set the mouth shape Th (eg. think) for lipsync animation.
Transform
Rotate:
X: 0.0
Y: 0.0
Z: 0.0
Move:
X: 0.0
Y: 11.7
Z: 153.0
25
9_Elsa with Cape_ModelMesh
0
9_faith_ModelMesh
0
9_go_ModelMesh
0
9_i_ModelMesh
38
9_laugh_ModelMesh
0
9_o_ModelMesh
0
9_OMG_ModelMesh
0
9_sad_ModelMesh
0
9_smile_ModelMesh
0
9_sorrow_ModelMesh
0
9_u_ModelMesh
0
9_wao_ModelMesh
-31
Export Mesh
Replace Mesh
Head Morph (B)
Head Model
Export Model
Import Model

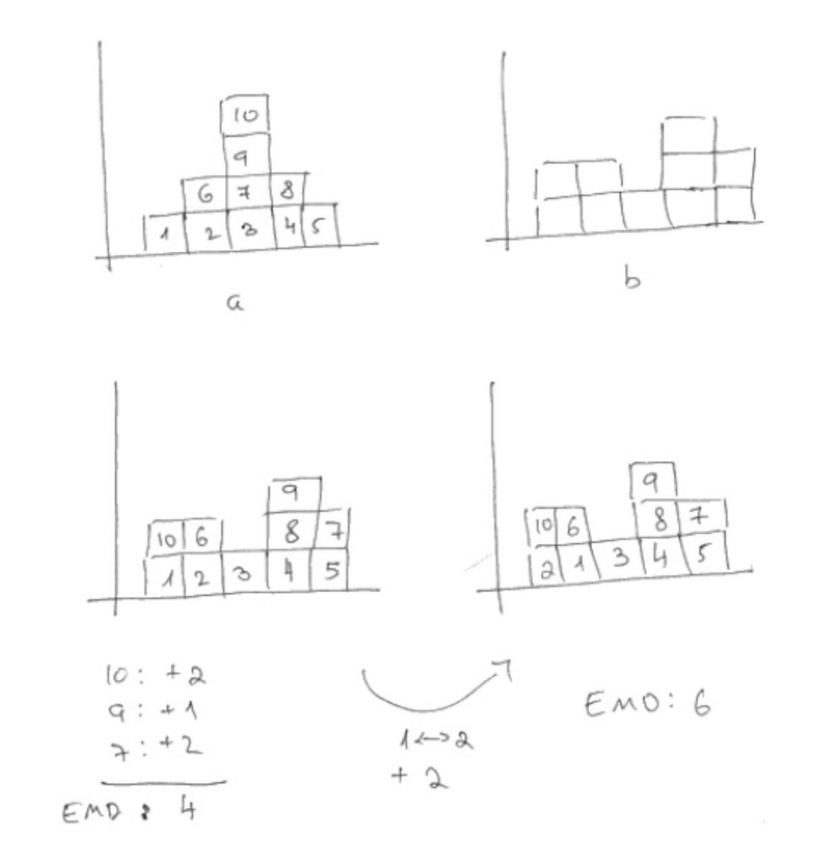
Text to Speech - Fail

- Speech feels unnatural
- Actually, it’s a Text to Dubbing problem
  - Much harder!

Sentence Similarity

- Edit distance
  - Similar words, different meaning
- Semantic distance
  - Word Mover’s Distance

Earth Mover's Distance



b

a	1	2	3	4	5	Σ
1	1					1
2		2				2
3	1		1	1	1	4
4				2		2
5					1	1
Σ	2	2	1	3	2	10

https://www.slideshare.net/BarVinograd/wasserstein-gan

EMD Formulation

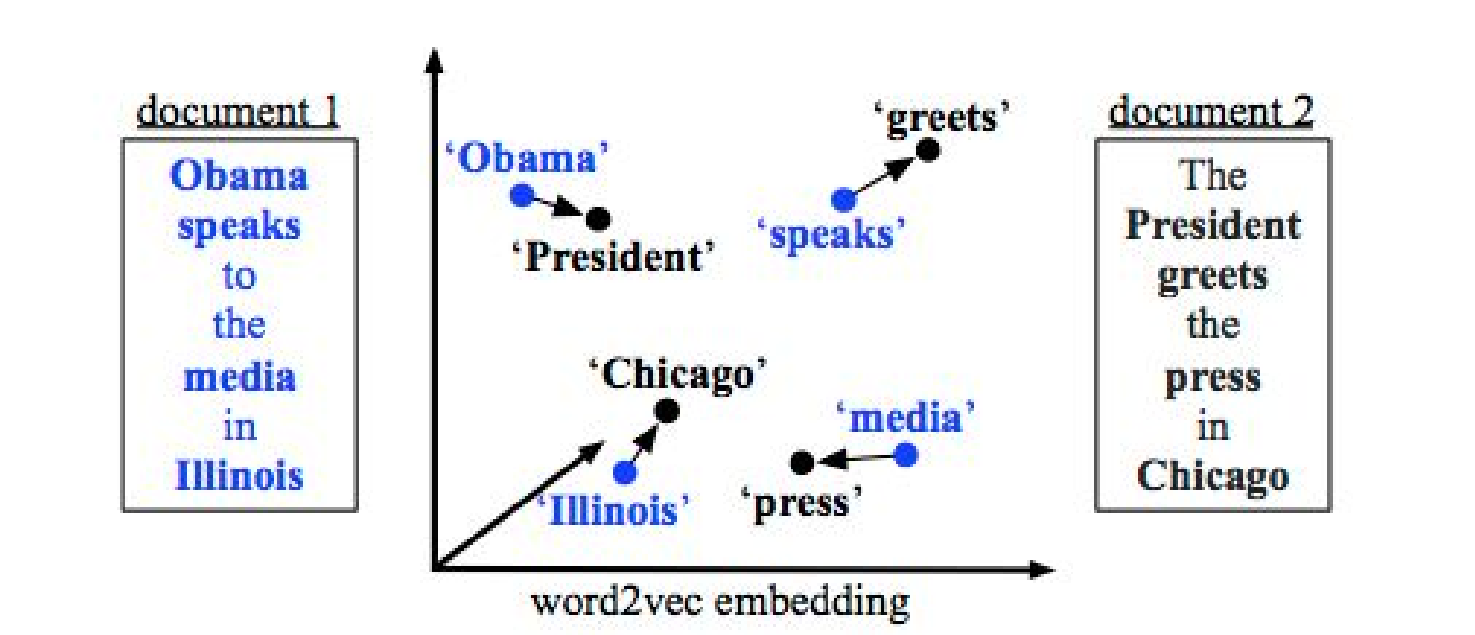
d_i – amount of source dirt in i
d'_j – amount of destination dirt in j
$c(i,j)$ – distance from i to j
T_{ij} – dirt flow from i to j. Unknown.

$$\min_{\mathbf{T}\geq 0} \sum_{i,j=1}^{n} \mathbf{T}_{ij} c(i,j)$$

$$\text{subject to: } \sum_{j=1}^{n} \mathbf{T}_{ij} = d_i \quad \forall i \in \{1,\ldots,n\}$$

$$\sum_{i=1}^{n} \mathbf{T}_{ij} = d'_j \quad \forall j \in \{1,\ldots,n\}.$$

Word Mover Distance



http://proceedings.mlr.press/v37/kusnerb15.pdf

WMD Formulation

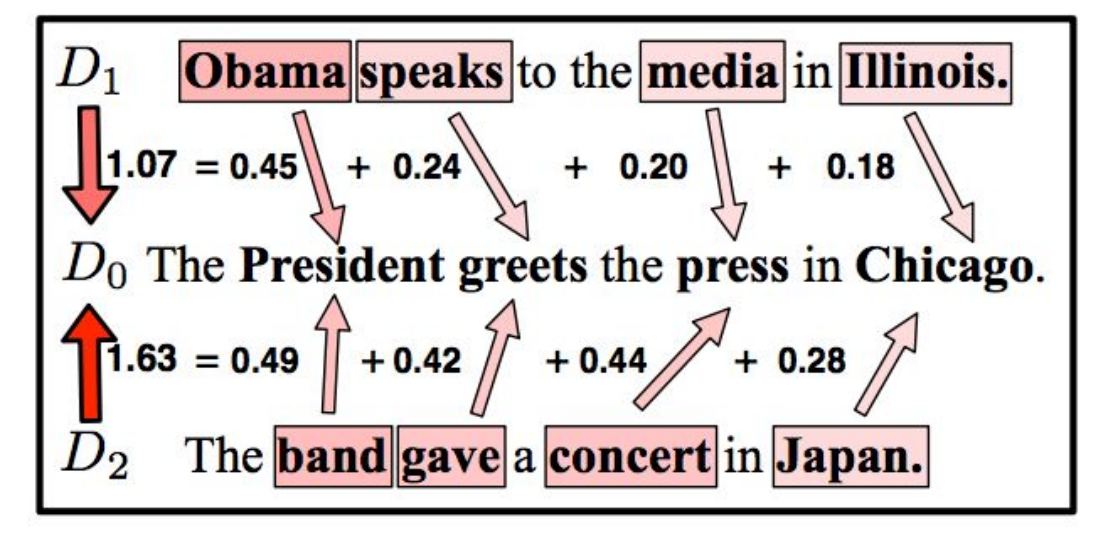
d_i – count of word i in sentence 1 / sentence length
d'_j – count of word j in sentence 2 / sentence length
$c(i, j)$ – distance from word i to word j
T_{ij} – dirt flow from i to j. Unknown.

$$\min_{\mathbf{T} \geq 0} \sum_{i,j=1}^{n} \mathbf{T}_{ij} c(i, j)$$

$$\text{subject to: } \sum_{j=1}^{n} \mathbf{T}_{ij} = d_i \quad \forall i \in \{1, \ldots, n\}$$

$$\sum_{i=1}^{n} \mathbf{T}_{ij} = d'_j \quad \forall j \in \{1, \ldots, n\}.$$

Word Mover Distance



In [1]:
```
from gensim.models import Word2Vec
model = Word2Vec.load_word2vec_format('GoogleNews-vectors-negative300.bin', binary=True)
model.wmdistance(sentence1, sentence2)
```

Sentence Similarity

	orig	script
0	go to bed	Go to sleep.
1	your father said eat your carrot	Your mother said\neat your broccoli.
2	we had lunch	And we had dinner.
3	you missed lunch	you missed breakfast.

Pivot

- Speaker isn’t important enough
- New task: Simpsons’ style text

Demo

Extras

- CNN based Simpsons detector
  - http://zachmoshe.com/2017/05/03/simpsons-detector.html
- Our GitHub repository
  - https://github.com/yahelg/simphony

Thank You!


I will not create fake Simpsons episodes.
I will not create fake Simpsons episodes.
I will not create fake Simpsons episodes.
I will not create fake Simpsons episodes.
I will not create fake Simpsons episodes.
I will not create fake Simpsons episodes.
I will not create fake Simpsons episodes.
I will not create fake Simpsons episodes.
I will not create fake Simpsons episodes.
I will not create fake Simpsons episodes.
I will not create fake Simpsons episodes.