Virtual Immortality: Reanimating characters from TV shows

James Charles, Derek Magee, David Hogg

School of Computing,
University of Leeds
{j.charles,d.r.magee,d.c.hogg}@leeds.ac.uk

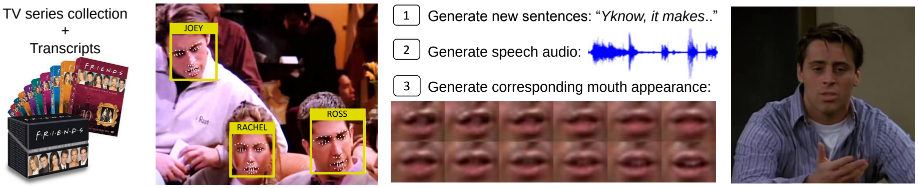
Abstract. The objective of this work is to build virtual talking avatars of characters fully automatically from TV shows. From this unconstrained data, we show how to capture a character's style of speech, visual appearance and language in an effort to construct an interactive avatar of the person and effectively immortalize them in a computational model. We make three contributions (i) a complete framework for producing a generative model of the audiovisual and language of characters from TV shows; (ii) a novel method for aligning transcripts to video using the audio; and (iii) a fast audio segmentation system for silencing non-spoken audio from TV shows. Our framework is demonstrated using all 236 episodes from the TV series Friends [34] ($\approx$ 97hrs of video) and shown to generate novel sentences as well as character specific speech and video.

Keywords: Visual speech, video synthesis, video alignment.

1 Introduction

For many years humans have been enthralled with recording people and their activities using e.g. sculpture, paintings, photographs, video and sound. We strive to modernize the existing set of recording methods by building a generative computational model of a person's motion, appearance, speech, language and their style of interaction and behavior. This model is trained from unconstrained pre-recorded material of a person but grants one with the ability to generate *brand-new* and *interactive* content, effectively rendering the person virtually immortal. Uses of such a system include a natural interface between human and computer, possibly putting a face and personality to existing voice-only assistants such as Apple's Siri, Microsoft's Cortana or Amazon's Alexa. Such a model could also be used as an effortless way to generate ground truth audiovisual data for training AI interactive systems.

A system capable of learning to generate virtual talking avatars of characters appearing in TV shows is proposed. Such a task is very challenging due to different camera angles, shot changes, camera motion, scale variations, lighting, appearance changes and background audio, e.g music and laughter. Transcripts (a written record of character dialog) supplement the videos and are used to help form training labels. However, as they contain no timing information, using them to infer where, who and when someone is talking on screen is non-trivial.


TV series collection
+
Transcripts
JOEY
RACHEL
ROSS
1 Generate new sentences: "Yknow, it makes.."
2 Generate speech audio:
3 Generate corresponding mouth appearance:

(a) Training corpus (b) Character data collection (c) Generating new content (example for Joey) (d) Final avatar synthesis

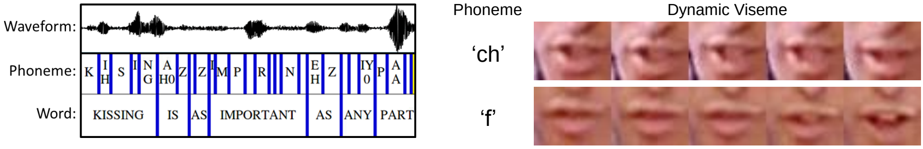
Fig. 1. System overview for learning to generate a virtual talking avatar of a target TV show character. Example shown for the *Friends* TV series and the character *Joey*. See text for details.

Related work. Our work is most closely related to visual text to speech systems, which take a sentence, in the form of text, and synthesize speech audio and a visual head with corresponding mouth motion. Our virtual model is a 2D rendering of a character [31] and is trained to generate visual speech using a concatenative unit selection system [15, 32] where short clips of mouth motion, called dynamic visiemes [30], are stitched together to form the visual signal. A similar approach is also taken for the audio [4]. Traditionally, visual speech systems are built from data captured in controlled environments [29, 1]. However, in our case the audiovisual data from TV shows is unconstrained and not designed for the task of training a visual speech system. Our work also differs from previous methods as a model of character language is trained for producing new sentences in character specific style. Furthermore, we also produce background video content including upper body motion and gestures. For illustrative purposes throughout this paper, we train our generative audiovisual and language model on *Joey* from the popular TV show *Friends*, allowing him to say new sentences in his own style and voice. Next we give an overview of the system.

2 System overview

A high level overview of our system is illustrated in Fig 1. Our goal is to build a system capable of learning to generate new audiovisual content of a chosen TV character (the *target*) from a TV show, in the form of a moving, gesturing and speaking 2D avatar. A collection of episodes and transcripts from a popular TV show, Fig 1(a), can provide a large training corpus (e.g. over 200 episodes of video for *Friends*) for learning a generative model of the avatar. Our system automatically labels audio and video with phonetic boundaries, character face bounding boxes with character names, and facial landmarks, as shown in Fig 1(b) and Fig 2(a). From this, our generative model is trained.

Novel avatar content is produced for a target character, as shown in Fig 1(c)-(d), by: (1) first generating new short sentences sampled from a character specific language model. Then, corresponding audiovisual data is generated in a two phase approach [17] whereby (2) the text is converted to a phonetic sequence with phoneme durations and an audio signal is generated. (3) the visual speech engine uses this phonetic information for producing the visual element of the avatar, synthesizing a video of mouth appearance. Finally, mouth synthesis is



(a) Example audio to transcript alignment

(b) Example dynamic visemes

Fig. 2. Audio to transcript alignment and dynamic visemes.

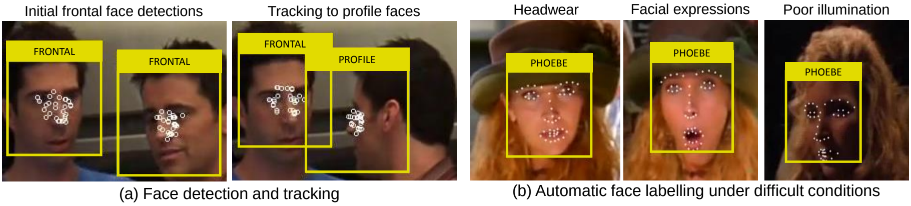
blended on to a moving background of the target, showing the full face together with upper body and background scene, as if the target were performing and gesturing in the original TV show, example frame in Fig 1(d).

3 Character data collection

Data collection is non-trivial and involves multiple stages of processing: (i) Muting of non-spoken audio, (ii) phonetically labeling the speech, (iii) face detection and tracking, (iv) automatic character labeling, (v) facial landmark detection, and finally (vi) producing phonetic databases (units of speech audio) and visemic databases (units of mouth appearance) for each character. From these databases new audiovisual content can be generated. Each of these stages is now explained in detail.

Muting non-spoken audio. A critical task for training our speech synthesizer is first detecting spoken audio. In TV shows, speech is mixed with background noise, e.g., music, traffic noise and in particular canned/audience laughter. Speech audio is automatically detected and background noise muted prior to further audio analysis. Although prior work on this exists for radio broadcasts [26], news broadcasts [16] and feature films [3], speech detection in TV shows is a very different domain. Comedy shows exhibit canned/audience laughter more so than in films. Also, one should leverage the consistent nature of episodes, i.e., same characters, same environments and similar music scores, to help improve speech detection. To this end, we build a speech detection system capable of generalizing well across episodes, yet trained from only one manually labeled episode. A sliding window classifier (random forest) labels the speech into spoken and non-spoken audio. The audio is represented with Mel Frequency Cepstral Coefficients (MFCCs), common in automatic speech recognition [18, 20].

Aligning transcripts to video (using audio). Time-aligned transcripts act as supervisory labels for training our system. In particular, previous works use subtitle-aligned transcripts for learning to recognize and label characters in movies and TV shows [9, 14, 6, 21]. Subtitles have words missing and the timing is rather ad-hoc and does not provide accurate timing level information necessary for learning visual speech systems where phoneme-level precision is required. Instead, we align the transcripts to speech audio, producing much greater timing



(a) Face detection and tracking

(b) Automatic face labelling under difficult conditions

Fig. 3. Face detection, tracking and automatic labeling with character names.

precision and accuracy. Transcripts are force-aligned to speech-only audio for each episode using dynamic programming and an American-English language model, as in [24]. Alignment results in word and phoneme boundaries, as shown in Fig 2(a).

Face detection and tracking. An online multi-face detection and tracking system produces face tracks for all episodes. Highly confident face detections [33] initialize the face tracking. Face bounding boxes are tracked [25] to the next frame by tracking keypoints in the box center (Fig 3(a) white circles). Boxes are tracked from frame to frame provided enough keypoints (> 3) are matched, and the previous box covers at least 90% of the tracked box. By tracking points only in the center of the face we limit the possibility of tracking background points and circumvent drifting, particularly between shots. While only frontal faces are detected, tracking leads to profile faces being captured (Fig 3(a)). Note, this stage can produce tracks for main characters as well as supporting actors.

Automatic character labeling. Automatic labeling of characters in TV shows has received much attention over the years [10, 8, 14, 7, 6, 23, 21]. Our approach here is similar in principle to the founding work by Everingham *et al.* [9] where subtitle-aligned transcripts were used. However, we demonstrate that improved precision from audio aligned transcripts leads to a relatively simple and accurate approach while also removing the need for visual speech detection. As in [21], we automatically label whole face tracks from ConvNet based face features [22]. To initialize, we transfer character labels from the aligned transcripts to solitary face tracks (no other faces on screen). A multi-class linear SVM classifier is trained in an iterative manner to solve the labeling problem for the remaining face-tracks. At each iteration the classifier is trained from current labels and then applied to *all* tracks. Only easy tracks (high classifier confidence) are labeled. At each iteration, progressively more tracks become labeled. We found 8 iterations sufficient. One classifier per episode is initially trained and later a single classifier per series is trained from current labels and applied across every episode of that series. In this way, knowledge about appearance variation (e.g. changes in lighting) can be shared across episodes (example variations in appearance shown for the character *Phoebe* from *Friends* in Fig 3(b)). Empirically it was found that training in this manner led to better results than simply training only one classifier from all data in the series.

Dynamic Visemes. As in [30] we generate dynamic, concatenative units of visual speech called *dynamic visemes*. In our case, a dynamic viseme is a small video clip of mouth appearance capturing coarticulation. A one-to-many mapping from phoneme to viseme is formed by building a database per character, example visemes for the character *Joey* from *Friends* are shown in Fig 2(b). Mouths are detected using a facial landmark detector [2], example landmark detections are shown as white dots in Fig 1(b) and Fig 3(b). RGB pixel values for each frame of the dynamic viseme is represented as a set of PCA coefficients. The PCA model is trained from all frames over all visemes, one model per target character. The previous and next phoneme in the phonetic sequence is also assigned to each viseme, providing contextual features. We next describe how to train our model and generate the virtual avatar.

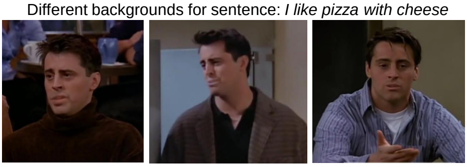
4 Text to visual speech

Language model. New sentences in the style of the target character are generated with a deep Long Short Term Memory (LSTM) [12] Recurrent Neural Network (RNN). Each letter of the sentence is generated given the previous letter [28, 11, 13]. RNNs capture long-range dependencies between letters to form words and sentence structure. In our case, a letter-level modeling approach is more appropriate than a word-level language model [19, 27] as it has the ability to learn person specific spelling and sentence structure, such as “yknow” meaning “you know”. A two hidden layer network with 128 nodes at each hidden layer is trained with backpropagation through time (unrolled to 50 time steps).

Text to speech (TTS). Speech audio is generated using a cluster unit selection based method [5] and trained from the phonetic labeling of the audio. At run time, input text is converted to a phonetic sequence and for each phoneme a corresponding unit of audio is selected based on the phonetic context and word position. A speech waveform is generated by stringing the selected audio units together. We use the Festival auditory speech synthesizer [4] software for building and running the TTS model.

Visual speech. Generating visual speech follows a similar approach to generating speech audio, except phonetic duration (from the TTS) guides the synchronization of mouth motion with the speech audio. Concatenating dynamic visemes together (in time) forms visual mouth motion. Visemes are selected based on their phonetic label and context and visual smoothness is enforced by matching the PCA coefficients of the last frame of one viseme to the first frame of the next, optimized using the Viterbi algorithm. A post processing method of temporal smoothing is applied. The number of frames for each viseme is either linearly upsampled or downsampled to match the phonetic duration.

Avatar synthesis. A *moving* background section of video (containing only the target character in a frontal facing pose, perhaps gesturing) acts as a canvas


Different backgrounds for sentence: *I like pizza with cheese*

Character	**#Sent**	**#Wrds**	**Speaking-time**
Joey	26	286	1.05 min
Monica	28	278	1.76 min
Chandler	27	284	1.70 min
Ross	30	327	2.11 min
Rachel	30	328	2.09 min
Phoebe	24	265	1.49 min

Fig. 4. Left: example frames of avatar *Joey* showing same sentence with different backgrounds. **Right**: Table of average word/sentence statistics per episode.

for "pasting" on a generated mouth video over the target's mouth. The mouth video is scaled and rotated according to facial landmarks on the background video and blended using alpha mapping. The coloring of the mouth video is altered to match the color of the background mouth using histogram specification. Generated speech-audio is combined with the video to form the final synthesis.

5 Avatar from Friends

The system is applied to the TV show *Friends* [34] where we "virtually immortalize" the character *Joey*. A demo video can be viewed at http://tinyurl.com/ztvgeat.

Dataset. Audiovisual material is obtained from The Friends complete collection DVD boxset of 236 episodes in total, each episode approx. 22mins in length. The first 3 seasons (73 episodes) are processed for data collection. Various statistics of the data extracted from automatic data collection is shown in Fig 4.

Training. Episode 2 from season 1 was manually labeled with speech and non-speech for training the speech detector. All other training processes are fully automatic given the transcripts. The language model is trained from sentences with 5-10 words across all 236 episodes (1857 sentences in total).

Generating new video and speech. Example new generated sentences sampled from the language model include: "Hey Ross do you want me to talk to some lady" and "I want to do something wrong" more examples are in the supplementary material. Example generated output frames of *Joey* saying the new sentence "I like pizza with cheese" is shown in Fig 4, produced using various moving backgrounds.

6 Summary and extensions

We have presented a semi-supervised method for producing virtual talking avatars of celebrities from TV shows. Given only the transcripts and one manually segmented episode (3hrs of manual work) one can process data from all episodes for any chosen character fully automatically. The character *Joey* from *Friends* was "virtually immortalized" in a generative model, enabling him to say new sentences in his style and appearance. We plan to improve the rendering of the avatar and extend our model to include interaction with real people and also between avatars.

Acknowledgments. We thank Andrew Zisserman for useful discussions.

References

1. Anderson, R., Stenger, B., Wan, V., Cipolla, R.: Expressive visual text-to-speech using active appearance models. In: Proc. CVPR (2013)
2. Asthana, A., Zafeiriou, S., Cheng, S., Pantic, M.: Incremental face alignment in the wild. In: Proc. CVPR (2014)
3. Benatan, M., Ng, K.: Cross-covariance-based features for speech classification in film audio. Journal of Visual Languages & Computing 31, 215–221 (2015)
4. Black, A., Taylor, P., Caley, R., Clark, R., Richmond, K., King, S., Strom, V., Zen, H.: The festival speech synthesis system. http://www.cstr.ed.ac.uk/projects/festival/ (2001)
5. Black, A.W., Lenzo, K.A.: Building synthetic voices. Language Technologies Institute, Carnegie Mellon University and Cepstral LLC (2003)
6. Bojanowski, P., Bach, F., Laptev, I., Ponce, J., Schmid, C., Sivic, J.: Finding actors and actions in movies. In: Proc. ICCV (2013)
7. Cinbis, R.G., Verbeek, J., Schmid, C.: Unsupervised metric learning for face identification in tv video. In: Proc. ICCV (2011)
8. Cour, T., Sapp, B., Nagle, A., Taskar, B.: Talking pictures: Temporal grouping and dialog-supervised person recognition. In: Proc. CVPR (2010)
9. Everingham, M., Sivic, J., Zisserman, A.: "Hello! My name is... Buffy" – automatic naming of characters in TV video. In: Proc. BMVC (2006)
10. Everingham, M., Zisserman, A.: Automated detection and identification of persons in video using a coarse 3-D head model and multiple texture maps. IEE Proceedings on Vision, Image and Signal Processing 152(6), 902–910 (2005)
11. Graves, A.: Generating sequences with recurrent neural networks. arXiv preprint arXiv:1308.0850 (2013)
12. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation 9(8), 1735–1780 (1997)
13. Karpathy, A.: The unreasonable effectiveness of recurrent neural networks. http://karpathy.github.io/2015/05/21/rnn-effectiveness/, accessed: 07-25-2016
14. Köstinger, M., Wohlhart, P., Roth, P.M., Bischof, H.: Learning to recognize faces from videos and weakly related information cues. In: Proc. Advanced Video and Signal-Based Surveillance (2011)
15. Liu, C., Yuen, J., Torralba, A.: Sift flow: Dense correspondence across scenes and its applications. IEEE PAMI (2011)
16. Lu, L., Zhang, H.J., Jiang, H.: Content analysis for audio classification and segmentation. IEEE Transactions on speech and audio processing 10(7), 504–516 (2002)
17. Mattheyses, W., Verhelst, W.: Audiovisual speech synthesis: An overview of the state-of-the-art. Speech Communication 66, 182–217 (2015)
18. Mermelstein, P.: Distance measures for speech recognition, psychological and instrumental. Pattern recognition and artificial intelligence 116, 374–388 (1976)
19. Mikolov, T., Karafiát, M., Burget, L., Cernockỳ, J., Khudanpur, S.: Recurrent neural network based language model. In: Interspeech (2010)
20. Mogran, N., Bourlard, H., Hermansky, H.: Automatic speech recognition: An auditory perspective. In: Speech processing in the auditory system (2004)
21. Parkhi, O.M., Rahtu, E., Zisserman, A.: It's in the bag: Stronger supervision for automated face labelling. In: ICCV Workshop: Describing and Understanding Video & The Large Scale Movie Description Challenge (2015)
22. Parkhi, O.M., Vedaldi, A., Zisserman, A.: Deep face recognition. In: Proc. BMVC (2015)

23. Ramanathan, V., Joulin, A., Liang, P., Fei-Fei, L.: Linking people in videos with their names using coreference resolution. In: Proc. ECCV (2014)
24. Rubin, S., Berthouzoz, F., Mysore, G.J., Li, W., Agrawala, M.: Content-based tools for editing audio stories. In: ACM symposium on User interface software and technology (2013)
25. Shi, J., Tomasi, C.: Good features to track. In: Proc. CVPR (1994)
26. Sonnleitner, R., Niedermayer, B., Widmer, G., Schlüter, J.: A simple and effective spectral feature for speech detection in mixed audio signals. In: Proc. International Conference on Digital Audio Effects (2012)
27. Sundermeyer, M., Schlüter, R., Ney, H.: Lstm neural networks for language modeling. In: Interspeech (2012)
28. Sutskever, I., Martens, J., Hinton, G.E.: Generating text with recurrent neural networks. In: Proc. ICML (2011)
29. Taylor, S., Theobald, B.J., Matthews, I.: The effect of speaking rate on audio and visual speech. In: Proc. ICASSP (2014)
30. Taylor, S.L., Mahler, M., Theobald, B.J., Matthews, I.: Dynamic units of visual speech. In: Proc. ACM SIGGRAPH (2012)
31. Tiddeman, B., Perrett, D.: Prototyping and transforming visemes for animated speech. In: Proc. Computer Animation (2002)
32. Verma, A., Rajput, N., Subramaniam, L.V.: Using viseme based acoustic models for speech driven lip synthesis. In: Proc. Multimedia and Expo (2003)
33. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: Proc. CVPR (2001)
34. Warner Bros. Television and Bright/Kauffman/Crane Productions: Friends, seasons 1-10 (1994-2004)