Video-Driven Speech Reconstruction using Generative Adversarial Networks

Konstantinos Vougioukas[1,2], Pingchuan Ma[1], Stavros Petridis[1,2], and Maja Pantic[1,2]

[1]iBUG Group, Imperial College London
[2]Samsung AI Centre, Cambridge, UK

June 17, 2019

Abstract

Speech is a means of communication which relies on both audio and visual information. The absence of one modality can often lead to confusion or misinterpretation of information. In this paper we present an end-to-end temporal model capable of directly synthesising audio from silent video, without needing to transform to-and-from intermediate features. Our proposed approach, based on GANs is capable of producing natural sounding, intelligible speech which is synchronised with the video. The performance of our model is evaluated on the GRID dataset for both speaker dependent and speaker independent scenarios. To the best of our knowledge this is the first method that maps video directly to raw audio and the first to produce intelligible speech when tested on previously unseen speakers. We evaluate the synthesised audio not only based on the sound quality but also on the accuracy of the spoken words.

Index Terms: speech synthesis, generative modelling, visual speech recognition

1 Introduction

Lipreading is a technique that involves understanding speech in the absence of sound, primarily used by people who are deaf or hard-of-hearing. Even people with normal hearing depend on lip movement interpretation, especially in noisy environments.

The ineffectiveness of audio speech recognition (ASR) methods in the presence of noise has lead to the research of automatic visual speech recognition (VSR) methods. Thanks to recent advances in machine learning VSR systems are capable of performing this task with high accuracy [1, 2, 3, 4, 5]. Most of these systems output text but there are many applications such as videoconferencing in noisy or silent environments that would benefit from the use of video-to-speech systems.

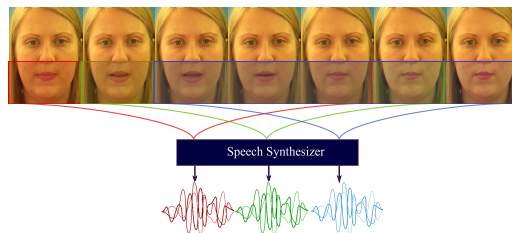* The first author performed this work during his internship at Samsung

Figure 1: The speech synthesizer accepts a sequence of frames and produces the corresponding audio sequence.

One possible approach for developing video to speech systems is to combine VSR with text-to-speech systems (TTS), with text serving as an intermediate representation. However, there are several limitations to using such systems. Firstly, text-based systems require transcribed datasets for training, which are hard to obtain because they require laborious manual annotation. Secondly, generation of the audio can only occur at the end of each word, which imposes a delay on the pipeline making it unsuitable for real-time applications such as videoconferencing. Finally, the use of text as an intermediate representation prevents such systems from capturing emotion and intonation, which results in unnatural speech.

Direct video-to-speech methods do not have these drawbacks since audio samples can be generated with every video frame that is captured. Furthermore, training of such systems can be done in a self-supervised manner since most video comes paired with the corresponding audio. For these reasons video-to-speech systems have been recently considered.

Such video-to-speech systems are proposed by Le Cornu and Miller in [6, 7], which use Gaussian Mixture Models and Deep Neural Networks (DNN) respectively to estimate audio features, which are fed into a vocoder to produce audio, from visual features. However, the hand-crafted visual features used in this approach are not capable of capturing the pitch and intonation of the speaker and in order to produce intelligble results they have be artificially generated.

Convolutional neural networks (CNNs) have been shown to be powerful feature extractors for images and videos and have replaced handcrafted features in more recent works. One such system is proposed in [8] to predict line spectrum pairs (LSPs) from video. The LSPs are converted into waveforms but since excitation is not predicted the resulting speech sounds unnatural. This method is extended in [9] by adding optical flow information as input to the network and by adding a post-processing step, where generated sound features are replaced by their closest match from the training set. A similar method that uses multi-view visual feeds has been proposed in [10]. Finally, Akbari et. al. [11] propose a model that uses CNNs and recurrent neural networks (RNNs) to transform a video sequence into audio spectrograms, which are later transformed into waveforms using the algorithm proposed in [12].

In this work, we propose an end-to-end model that is capable of directly converting silent video to raw waveform, without the need for any intermediate handcrafted features as shown in Fig. 1. Our method is based on generative adversarial networks (GANs), which allows us to produce high fidelity (50KHz) audio sequences of realistic intelligible speech. Contrary to the aforementioned works our model is capable of generating intelligible speech even in the case of unseen speakers[1]. The generated speech is evaluated using standard audio reconstruction and speech synthesis metrics. Additionally, we propose using a speech recognition model to verify the accuracy of the spoken words and a synchronisation method to quantify the audio-visual synchrony.

2 Video-driven Speech Reconstruction

The proposed model for speech reconstruction is made up of 3 sub-networks. The generator network, shown in Fig. 2 is responsible for transforming the sequence of video frames into a waveform. During the training phase the critic network drives the generator to produce waveforms that sound similar to natural speech. Finally, a pretrained *speech encoder* is used to conserve the speech content of the waveform.

2.1 Generator

The generator is made up of a *content encoder* and an *audio frame decoder*. The *content encoder* consists of a *visual feature encoder* and an RNN. The *visual feature encoder* is a 5 layer 3D CNN responsible for encoding information about the visual speech which is present in a window of N consecutive frames. These encodings z_s, which are produced at each time step are fed to a single layer gated recurrent unit (GRU) network which produces a series features z_c describing the content of the

video. The *audio frame decoder* receives these features as input and generates the corresponding window of audio samples. Batch normalization [13] and ReLU activation functions are used throughout the entire generator network except for the last layer in the *visual feature encoder* and *decoder*, where the hyperbolic tangent non-linearity is used without batch normalization.

2.2 Critic

The critic is a 3D CNN which is given audio clips of fixed length t_d from the real and generated samples. During training the critic learns a 1-Lipschitz function D which is used to calculate the Wasserstein distance between the distribution of real and generated waveforms. In order to enforce the Lipschitz continuity on D the gradients of critic's output with respect to the inputs are penalized if they have a norm larger than 1 [14]. The penalty is applied independently to all inputs and therefore batch normalization should not be used in any layers of the critic since it introduces correlation between samples of the batch. The audio clips which are provided as input to the critic are chosen at random from both the real and generated audio sequences using a uniform sampling function S.

2.3 Speech Encoder

The *speech encoder* network is used to extract speech features from the real and generated audio. This network is taken from the pretrained model proposed in [15], which performs speech-driven facial animation. Similarly to our model this network is trained in a self-supervised manner and learns to produce encodings that capture speech content that can be used to animate a face. Using this module we are able to enforce the correct content onto the generated audio clips through a perceptual loss.

2.4 Training

Our model is based on the Wasserstein GAN proposed in [14], which minimises the Wasserstein distance between the real and fake distribution. Since optimising the Wasserstein distance directly is intractable the Kantorovich-Rubinstein duality is used to obtain a new objective [16]. The adversarial loss of our model is shown in Equation 1, where D is the critic function, G is the generator function, x is a sample from the distribution of real clips P_r and $\widetilde{x}$ is a sample from the distribution of generated clips P_g. The gradient penalty shown in Equation 2 is calculated with respect to the input $\hat{x}$ sampled from distribution $P_{\hat{x}}$, which contains all linear interpolates between P_r and P_g.

$$\mathcal{L}_{adv} = \mathbb{E}_{x \sim P_r}[D(S(x))] - \mathbb{E}_{\widetilde{x} \sim P_g}[D(S(\widetilde{x}))] \quad (1)$$

$$\mathcal{L}_{gp} = \mathbb{E}_{\hat{x} \sim P_{\hat{x}}}[(||\nabla_{\hat{x}} D(\hat{x})||_2 - 1)^2] \quad (2)$$

---

[1]Generated samples online at the following website:
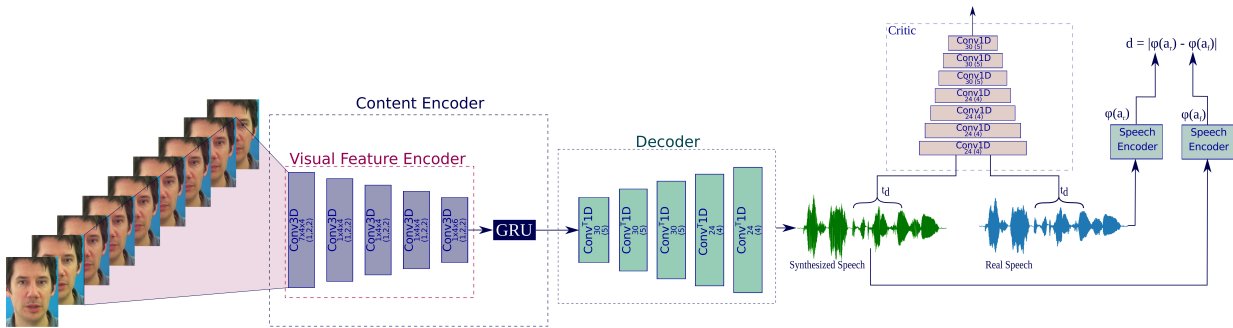https://sites.google.com/view/speech-synthesis/home

Figure 2: Architecture for the generator network consisting of a content encoder and an audio frame decoder. Convolutional layers are represented with Conv blocks with the kernel size depicted as $k_{time} \times k_{height} \times k_{width}$ and with strides for the respective axes appearing in parentheses. The critic accepts as input samples of $t_d = 1s$ and determines whether they come from real or generated audio.

The *speech encoder* maps a waveform x to a feature space through a function ϕ. During training a perceptual loss, corresponding to the L_1 distance between the features obtained from mapping real and generated audios is minimized. This forces the network to learn high-level features correlated with speech. The perceptual loss is shown in Equation 3.

$$\mathcal{L}_p = |\phi(x) - \phi(\widetilde{x})| \tag{3}$$

An L_1 reconstruction loss is also used to ensure that the generated waveform closely matches the original. Finally, we use a total variation (TV) regularisation factor in our loss described in Equation 4, where T is the number of samples in the audio. This type of regularization penalizes non-smooth waveforms and thus reduces the amount of high frequency noise in the synthesized samples.

$$\mathcal{L}_{TV} = \frac{1}{T} \sum_t |\widetilde{x}_{t+1} - \widetilde{x}_t| \tag{4}$$

The final objective used to obtain the optimal generator is a combination of the above losses as shown in Equation 5. The loss factors are weighted so that each loss has approximately equal contribution. The weights we used were $\lambda_{L_1} = 150$, $\lambda_{tv} = 120$, $\lambda_{gp} = 10$ and $\lambda_p = 70$.

$$\mathcal{L} = \min_G \max_D \mathcal{L}_{adv} + \lambda_{L_1} \mathcal{L}_{L_1} + \lambda_{TV} \mathcal{L}_{TV} + \lambda_p \mathcal{L}_p + \lambda_{gp} \mathcal{L}_{gp} \tag{5}$$

When training a Wasserstein GAN the critic should be trained to optimality [17]. We therefore perform 6 updates on the critic for every update of the generator. We use the Adam [18] optimizer with a learning rate of 0.0001 for both the generator and critic and train until no improvement is seen on the mel-cepstral distance between the real and generated audio from the validation set for 10 epochs. We use a window of $N = 7$ frames as input for the generator and input a clip of $t_d = 1s$ for the critic.

3 Experimental Setup

Our model is implemented in Pytorch and takes approximately 4 days to train on a Nvidia GeForce GTX 1080 Ti GPU. During inference our model is capable of generating audio for a 3s video recorded at 25 frames per second (fps) in 60ms when running on a GPU and 6s when running on the CPU.

3.1 Dataset

The GRID dataset has 33 speakers each uttering 1000 short phrases, containing 6 words taken from a limited dictionary. The structure of a GRID sentence is described in Table 1. We evaluate our method in both speaker dependent and speaker independent settings. Subjects 1, 2, 4 and 29 were used for the subject dependent task and videos were split into training, validation and test sets using a random 90%-5%-5% split respectively. This setup is similar to that used in [11].

In the subject independent setting the data is divided into sets based on the subjects. We use 15 subjects for training, 8 for validation and 10 for testing. We use the split proposed in [15].

Table 1: Structure of a typical sentence from the GRID corpus.

Command	Color	Preposition	Letter	Digit	Adverb
bin	blue	at	A-Z	0-9	again
lay	green	by	\{W}		now
place	red	in			please
set	white	with			soon

As part of our preprocessing all faces are aligned to the canonical face using 5 anchor points taken from the edge of the eyes and the tip of the nose. Video frames are normalised, resized to 128x96 and cropped keeping only the bottom half, which contains the mouth. Finally, data augmentation is performed by mirroring the training videos.

3.2 Metrics

Evaluation of the generated audio is not a trivial task and there is no single metric capable of assessing all aspects of speech such as quality, intelligibility and spoken word accuracy. For this reason we employ multiple methods to evaluate the different aspects of our generated samples. In order to measure the quality of the produced samples we use the mean mel-cepstral distortion (MCD) [19], which measures the distance between two signals in the mel-frequency cepstrum and is commonly used to assess the performance of speech synthesizers. We also use Short Term Objective Intelligibility (STOI) metric which measures the intelligibility of the generated audio clip.

Speech quality is measured with the perceptual evaluation of speech quality (PESQ) metric [20]. PESQ was originally designed to quantify degradation due to codecs and transmission channel errors. However, this metric is sensitive to changes in the speaker's voice, loudness and listening level [20]. Therefore, although it may not be an ideal measure for this task we still use it in order to be consistent with previous works.

In order to determine the synchronisation between the video and audio we use the pretrained SyncNet model proposed in [21]. This model calculates the euclidean distance between the audio and video encodings for multiple locations in the sequence and produces the audio-visual offset (measured in frames) based on the location of the smallest distance and audio-visual correlation (confidence) based on the fluctuation of the distance.

Finally, the accuracy of the spoken message is also measured using the word error rate (WER) as measured by an ASR system, which is trained on the GRID training set. On ground truth audio the ASR system achieves 4% WER.

4 Results

Our proposed model is capable ,of producing intelligible, high quality speech at high sampling rates such as 50KHz. We evaluate our model on the subject dependent scenario and compare it to the *Lip2AudSpec* model proposed in [11]. Furthermore we present results when the system is tested on unseen speakers.

4.1 Speaker Dependent Scenario

Examples of real and generated audio samples are compared in Fig. 3. In order to present a fair comparison our audio is sub-sampled to match the rate of the sample produced by *Lip2AudSpec*. Through inspection it is noticeable that our model captures more information than *Lip2AudSpec* especially for very low and very high frequency content. This results in words being more clearly articulated when using our model. Our model introduces artifacts in the form of a low-power persistent hum in the waveform, which is also visible in the spectral domain.

Table 2: Metrics for the evaluation of the generated audio waveforms when testing on seen speakers. Best performance is shown in bold.

Measure	*Lip2AudSpec*	Proposed Model
PESQ	**1.82**	1.71
WER	32.5%	**26.6%**
AV Confidence	3.5	**4.4**
AV Offset	**1**	**1**
STOI	0.446	**0.518**
MCD	38.14	**22.29**

Table 3: Ablation study performed in the subject dependent setting. Best performance is shown in bold.

Model	PESQ	WER	AV Conf/Offset	STOI	MCD
Full	**1.71**	**26.6%**	**4.4(1)**	**0.518**	**22.29**
w/o $\mathcal{L}_{L_1}$	1.45	33.2%	3.9(1)	0.450	26.87
w/o $\mathcal{L}_{TV}$	1.44	31.3 %	3.9(1)	0.483	25,82
w/o $\mathcal{L}_p$	1.14	83.3%	2.0(5)	0.378	30.12

We compare the samples produced by our model to those produced by *Lip2AudSpec* using the metrics described in section 3.2. The results are shown in Table 2. Our method out-performs *Lip2AudSpec* in all the intelligibility tests (STOI, WER). Furthermore, our generated samples achieve better spectral accuracy as indicated by the smaller MCD. The *Lip2AudSpec* achieves a better PESQ score which is likely due to the artifacts that are created using our model. Indeed, if we apply average filtering to the signal, which reduces these artifacts the PESQ increases to 1.80. However, this is done at the expense of sharpness and intelligibility since STOI drops to 4.7 and the WER increases to 36%. Finally, both methods have similar scores in audio-visual synchrony, which is expected since they both use similar architectures to extract the visual-features.

In order to quantify the effect of each loss term we perform an ablation study by removing one loss term at a time. The results of the ablation study are shown in Table 3. It is clear that removing each term makes the performance worse and the full models is the best over all performance measures. We note that the adversarial loss is necessary for the production of speech and when the system was evaluated without it generation resulted in noise. The other important contribution is made from the perceptual loss, without which the speech produced does not accurately capture the content of the speech. This is evident from the large increase in the WER, although this is reflected in the other metrics as well.

4.2 Speaker Independent Scenario

We present the results of our model on unseen speakers in Table 4. Comparison with other methods is not possible for this setup since other methods are speaker dependent (training the *Lip2AudSpec* model on the same data

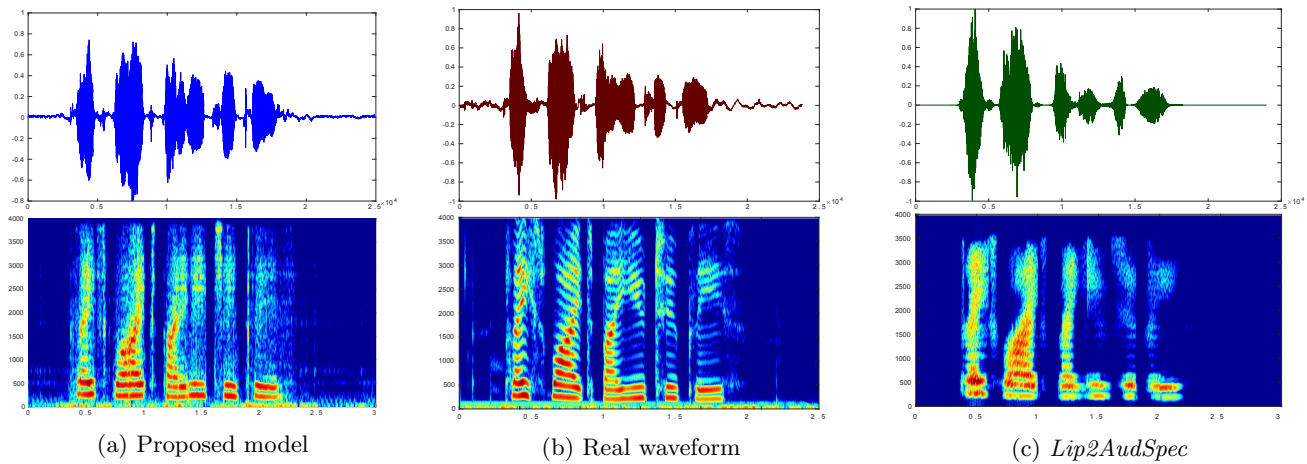(a) Proposed model (b) Real waveform (c) *Lip2AudSpec*

Figure 3: Examples of real and generated waveforms and their corresponding spectrograms

Table 4: Metrics for the evaluation of the generated audio waveforms when testing on unseen speakers

PESQ	WER	AV Confidence	AV Offset	STOI	MCD
1.24	40.5 %	4.1	1	0.445	24.29

did not produce intelligible results). From the results is evident that the speaker independent setting is a more complex problem. One of the reasons for this is the fact that the model can not learn the voice of unseen speakers. This results in generating a voice that does not correspond to the real speaker (i.e. female voice for a male speaker). Furthermore, in certain cases the voice will morph during the phrase. These factors likely account in large part for the drop in the PESQ metric, which is sensitive to such alterations. Morphing voices likely also affects the intelligibility of the audio clip, which is reflected in the WER, STOI and MCD. Finally we notice a slight improvement in audio visual synchrony which may be due to the increased number of samples seen during training.

It is important to note that the model has a different performance for each unseen subject. The WER fluctuates from 40% to 60% depending on the speaker. This is to be expected especially for subjects whose appearance differs greatly from the subjects in the training set. Additionally, since GRID is made up mostly of native speakers of English we notice that unseen subjects who are non-native speakers have worse WER.

5 Conclusions

In this work we have presented an end-to-end model that reconstructs speech from silent video and evaluated in two different scenarios. Our model is capable of generating intelligible audio for both seen and unseen speakers. Future research should focus on producing more natural and coherent voices for unseen speakers as well as im-

proving intelligibility. Furthermore, we believe that such systems are capable of reflecting the speakers emotion and should be tested on expressive datasets. Finally, a major limitation of this method is the fact that it operates solely on frontal faces. Therefore the natural progression of this work will be to reconstruct speech from videos taken in the wild.

6 Acknowledgements

References

[1] Y. M. Assael, B. Shillingford, S. Whiteson, and N. De Freitas, "Lipnet: End-to-end sentence-level lipreading," *arXiv preprint arXiv:1611.01599*, 2016.

[2] S. Petridis, Y. Wang, Z. Li, and M. Pantic, "End-to-end multi-view lipreading," in *British Machine Vision Conference*, London, September 2017.

[3] S. Petridis, T. Stafylakis, P. Ma, F. Cai, G. Tzimiropoulos, and M. Pantic, "End-to-end audiovisual speech recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 6548–6552.

[4] T. Stafylakis and G. Tzimiropoulos, "Combining residual networks with lstms for lipreading," in *Interspeech*, 2017.

[5] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Lip reading sentences in the wild," in *IEEE CVPR*, July 2017.

[6] T. L. Cornu and B. Milner, "Reconstructing intelligible audio speech from visual speech features,"

in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[7] T. Le Cornu and B. Milner, "Generating intelligible audio speech from visual speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 9, pp. 1751–1761, Sep. 2017.

[8] A. Ephrat and S. Peleg, "Vid2speech: Speech reconstruction from silent video," in *Int'l Conf. on Acoustics, Speech and Signal Processing*, 2017, pp. 5095–5099.

[9] A. Ephrat, T. Halperin, and S. Peleg, "Improved speech reconstruction from silent video," *ICCV 2017 Workshop on Computer Vision for Audio-Visual Media*, 2017.

[10] Y. Kumar, R. Jain, M. Salik, R. r. Shah, R. Zimmermann, and Y. Yin, "Mylipper: A personalized system for speech reconstruction using multi-view visual feeds," in *2018 IEEE International Symposium on Multimedia (ISM)*, Dec 2018, pp. 159–166.

[11] H. Akbari, H. Arora, L. Cao, and N. Mesgarani, "Lip2audspec: Speech reconstruction from silent lip movements video," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 2516–2520.

[12] T. Chi, P. Ru, and S. Shamma, "Multiresolution spectrotemporal analysis of complex sounds," *The Journal of the Acoustical Society of America*, vol. 118, pp. 887–906, 09 2005.

[13] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ser. ICML'15. JMLR.org, 2015, pp. 448–456.

[14] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of wasserstein gans," in *Advances in Neural Information Processing Systems*, 2017, pp. 5767–5777.

[15] K. Vougioukas, S. Petridis, and M. Pantic, "End-to-End Speech-Driven Facial Animation with Temporal GANs," in *British Machine Vision Conference*, 2018.

[16] C. Villani, *Optimal transport: old and new.* Springer Science & Business Media, 2008, vol. 338.

[17] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *International Conference on Machine Learning*, 2017, pp. 214–223.

[18] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[19] R. Kubichek, "Mel-cepstral distance measure for objective speech quality assessment," in *Proceedings of IEEE Pacific Rim Conference on Communications Computers and Signal Processing*, vol. 1, May 1993, pp. 125–128 vol.1.

[20] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," ser. ICASSP. IEEE Computer Society, 2001, pp. 749–752.

[21] J. S. Chung and A. Zisserman, "Out of time: automated lip sync in the wild," in *Workshop on Multiview Lip-reading, ACCV*, 2016.