# Dynamic Temporal Alignment of Speech to Lips

Tavi Halperin
Hebrew University of Jerusalem
**16** PUBLICATIONS **148** CITATIONS
SEE PROFILE

Shmuel Peleg
Hebrew University of Jerusalem
**216** PUBLICATIONS **11,443** CITATIONS
SEE PROFILE

# Dynamic Temporal Alignment of Speech to Lips

TAVI HALPERIN*, ARIEL EPHRAT*, and SHMUEL PELEG, The Hebrew University of Jerusalem

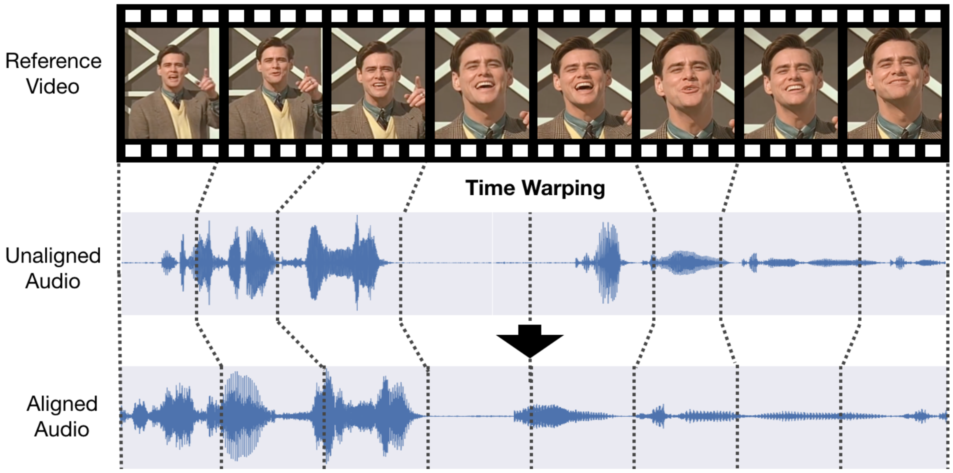

Fig. 1. Given a speech video and a segment of corresponding, but unaligned, audio, we align the audio to match the lip movements in the video.

Many speech segments in movies are re-recorded in a studio during post-production, to compensate for poor sound quality as recorded on location. Manual alignment of the newly-recorded speech with the original lip movements is a tedious task. We present an audio-to-video alignment method for automating speech to lips alignment, stretching and compressing the audio signal to match the lip movements. This alignment is based on deep audio-visual features, mapping the lips video and the speech signal to a shared representation. Using this shared representation we compute the lip-sync error between every short speech period and every video frame, followed by the determination of the optimal corresponding frame for each short sound period over the entire video clip. We demonstrate successful alignment both quantitatively, using a human perception-inspired metric, as well as qualitatively. The strongest advantage of our audio-to-video approach is in cases where the original voice in unclear, and where a constant shift of the sound can not give a perfect alignment. In these cases state-of-the-art methods will fail.

## 1 INTRODUCTION

In movie filming, poor sound quality is very common for speech recorded on location. Maybe a plane flew overhead, or the scene itself was too challenging to record high-quality audio. In these cases, the speech is re-recorded in a studio during post-production using a process called “Automated Dialogue Replacement (ADR)” or “looping”. In “looping” the actor watches his or her original performance in a loop, and re-performs each line to match the wording and lip movements.

ADR is a tedious process, and requires much time and effort by the actor, director, recording engineer, and the sound editor. One of the most challenging parts of ADR is aligning the newly-recorded audio to the actor’s original lip movements, as viewers are very sensitive to audio-lip discrepancies. This alignment is especially difficult when the original on-set speech is unclear.

In this work we temporally align audio and video of a speaking person by using innovative deep audio-visual (AV) features that were suggested by [Chung and Zisserman 2016]. These features map the lips video and the speech signal to a shared representation. Unlike the original synchronization method of Chung and Zisserman [2016], which shifts the audio or the video clip by a global offset, we use dynamic temporal alignment, stretching and compressing the signal dynamically within a clip. This is usually a three-step process [Hosom 2000]: (*i*) features are calculated for both the reference and the unaligned signals; (*ii*) optimal alignment which maps between the two signals is found using dynamic time warping (DTW) [Rabiner and Juang 1993]; (*iii*) a warped version of the unaligned signal is synthesized so that it temporally matches the reference signal [Ninness and Henriksen 2008]. In this paper we leverage the pre-trained AV features of Chung and Zisserman [2016] to find an optimal audio-visual alignment, and then use dynamic time warping to obtain a new, temporally aligned speech video.

We demonstrate the benefits of our approach over a state-of-the-art audio-to-audio alignment method, and over Chung and Zisserman [2016], using a human perception-inspired quantitative metric. Research has shown that the detectability thresholds of lack of synchronization between audio and video is +45 ms when the audio leads the video and -125 ms when the audio is delayed relative to the video. The broadcasting industry uses these thresholds in their official broadcasting recommendations [BT.1359 1998]. In order to evaluate the perceptive quality of our method’s output, our quantitative error measure is therefore the percentage of aligned frames which are mapped outside the undetectable region, relative to ground truth alignment. It should be noted that comparison to an audio-to-audio alignment method can only be performed when a clear reference audio signal exists, which may not always be the case.

*Indicates equal contribution

In that scenario, dynamic audio-to-visual or visual-to-visual alignment is the only option, a task which, to the best of our knowledge, has not yet been addressed.

To summarize, our paper's main contribution is a method for fully automated dialogue replacement in videos (ADR). We leverage the strength of deep audio-visual speech synchronization features of Chung and Zisserman [2016] and suggest a dynamic temporal alignment method. To the best of our knowledge, our paper is the first to propose a method for dynamic audio-to-visual time alignment.

## 2 RELATED WORK

We briefly review related work in the areas of audio and video synchronization and alignment, as well as speech-related video processing.

*Audio-to-video synchronization.* Audio-to-video synchronization (AV-sync), or *lip-sync*, refers to the relative timing of auditory and visual parts of a video. Automatically determining the level of AV-sync in a video has been the subject of extensive study within the computer vision community over the years, as lack of synchronization is a common problem. In older work, such as Lewis [1991], *phonemes* (short units of speech) are recognized and subsequently associated with mouth positions to synchronize the two modalities. Morishima et al. [2002] classifies parameters on the face into *visemes* (short units of visual speech), and uses a viseme-to-phoneme mapping to calculate synchronization. Zoric and Pandzic [2005] train a neural network to solve this problem.

In more recent work, methods have been proposed which attempt to find audio-visual correspondences without explicitly recognizing phonemes or visemes, such as Bredin and Chollet [2007] and Sargin et al. [2007] who use canonical correlation analysis (CCA). Marcheret et al. [2015] train a neural network-based classifier to determine the synchronization based on pre-defined visual features. In a recent pioneering work Chung and Zisserman [2016] have proposed a model called *SyncNet*, which learns a joint embedding of visual face sequences and corresponding speech signal in a video by predicting whether a given pair of face sequence and speech track are synchronized or not. They show that the learned embeddings can be used to detect and correct lip-sync error in video to within human-detectable range with greater than 99% accuracy.

The common denominator of the above works is that they attempt to detect and correct a global lip-sync error, i.e. the global shift of the audio signal relative to the video. In this work, we leverage the audio-visual features of SyncNet to perform dynamic time alignment, which can stretch and compress very small units of the unaligned (video or audio) signal to match the reference signal.

*Automatic time alignment of sequences.* Dynamic time warping (DTW) [Sakoe and Chiba 1978] uses dynamic programming to find the optimal alignment mapping between two temporal signals by minimizing some pairwise distance (e.g. Euclidean, cosine, etc.) between sequence elements. This algorithm has been used extensively in the areas of speech processing [Hosom 2000; King et al. 2012; Sakoe and Chiba 1978] and computer vision [Gong and Medioni 2011; Halperin et al. 2018; Zhou et al. 2008] for e.g. temporal segmentation and frame sampling, among many scientific disciplines.

King et al. [2012] propose a new noise-robust audio feature for performing automatic audio-to-audio speech alignment using DTW. Their feature models speech and noise separately, leading to improved ADR performance when the reference signal is degraded by noise. This method of alignment essentially uses audio as a proxy for aligning the re-recorded audio with existing lip movements. When the reference audio is very similar to the original, this results in accurate synchronization. However, when the reference signal is significantly degraded (as a result of difficult original recording conditions), our method overcomes this problem per performing audio-to-video alignment directly, resulting in higher-quality synchronization. In addition, when a reference audio signal is unavailable, direct audio-to-video alignment is the only option.

*Speech-driven video processing.* There has been increased interest recently within the computer vision community in leveraging natural synchrony of simultaneously recorded video and speech for various tasks. These include audio-visual speech recognition [Feng et al. 2017; Mroueh et al. 2015; Ngiam et al. 2011], predicting a speech signal or text from silent video (*lipreading*) [Chung et al. 2016; Ephrat et al. 2017; Ephrat and Peleg 2017], and audio-visual speech enhancement [Afouras et al. 2018; Ephrat et al. 2018; Owens and Efros 2018].

A large and relevant body of audio-visual work is speech-driven facial animation, in which, given a speech signal as input, the task is to generate a face sequence which matches the input audio [Bregler et al. 1997; Cao et al. 2004; Chang and Ezzat 2005; Furukawa et al. 2016; Taylor et al. 2017]. We do not attempt to provide a comprehensive survey of this area, but mention a few recent works. Garrido et al. [2015] propose a system for altering mouth motion of an actor in a video, so that it matches a new audio track containing a translation of the original audio (*dubbing*. Suwajanakorn et al. [2017] use an RNN to map audio features to a 3D mesh of a specific person, and Chung et al. [2017] train a CNN which takes audio features and a still face frame as input, and generates subject-independent videos matching the audio. Thies et al. [2016] don't use audio explicitly, but propose a real-time system for reenacting the face movement of a source sequence on a target subject. While the above works succeed in producing impressive results, they require the subject to be in a relatively constrained setting. This is oftentimes not the case in difficult-to-record movie scenes which require ADR, where the goal is to align the video and audio without modifying the pixels in the video frames.

## 3 DYNAMIC ALIGNMENT OF SPEECH AND VIDEO

Our speech to lips alignment is comprised of three main stages: audio-visual feature extraction, finding an optimal alignment which maps between audio and video, and synthesizing a warped version of the unaligned signal to temporally match the reference signal. An overview of our method is illustrated in Figure 2.

### 3.1 Audio-Visual Feature Extraction

We use SyncNet [Chung and Zisserman 2016] to extract language-independent and speaker-independent audio-visual embeddings. The network was trained to synchronize audio and video streams which were recorded simultaneously. This type of synchronization

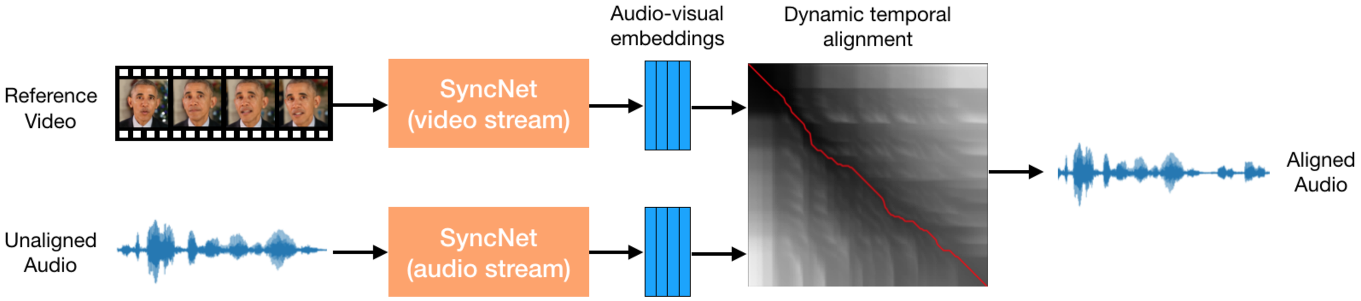

Fig. 2. **High-level diagram of our speech to lips alignment:** Given unaligned video and speech: (i) SyncNet features are computed for both; (ii) dynamic time warping is performed for optimal alignment between the features; (iii) A new speech is synthesized that is now aligned with the video.

is termed 'linear' as the audio is shifted by a constant time delta throughout the entire video. SyncNet encodes short sequences of 5 consecutive frames with total duration of 200 ms. or the equivalent amount of audio into a shared embedding space. We use the network weights provided by the authors, which were trained to minimize $l_2$ distance between embeddings of synchronized pairs of audio and video segments while maximizing distance between non matching pairs. We define the data term for our Dynamic Programing cost function to be pairwise distances of these embeddings.

### 3.2 Dynamic Time Warping for Audio-Visual Alignment

Naturally, as the number of possible mouth motions is limited, there are multiple possible low cost matches for a given short sequence. For example, segments of silence in different parts of the video are close in embedding space. SyncNet solves this by averaging time shift prediction over the entire video. We, however, are interested in assigning per frame shifts, therefor we use dynamic time warping.

Our goal here is to find a mapping ('path') with highest similarity between two sequences of embeddings $A = (a_1, ..., a_N), B = (b_1, ..., b_M)$, subject to non decreasing time constraint: if the path contains $(a_i, b_j)$ then later frames $a_{i+k}$ may only match later audio segments $b_{j+l}$. Additional preferences are (i) audio delay is preferred over audio advance with respect to reference video (a consequence of the different perception of the two); (ii) smooth path, to generate high quality audio; (iii) computationally efficient. We will now describe how we meet these preferences.

We solve for optimal path using Dijkstra's shortest path algorithm [Dijkstra 1959]. We construct a data cost matrix $C$ as pairwise dot products between embeddings from the reference video and the embeddings of an unaligned audio. Each matrix element is associated with a graph node, and edges connect node $(i, j)$ to $\{(i+1, j), (i, j+1), (i+1, j+1)\}$ so that the non decreasing time constraint holds.

Classically, the weight on an edge pointed at $(i, j)$ is the matrix value of the target element $C_{i,j}$. To better fit the perceptual attributes of consuming video and audio we modify the cost to prefer a slight delay by assigning the weight $0.5 * C_{i,j} + 0.25 * C_{i-1,j} + 0.25 * C_{i-2,j}$. Relative improvement which stems from this modification is studied in Section 4.

We assume the two modalities are cut roughly to the same start and end points, so we find a minimal path from $(0, 0)$ to $(N, M)$. We experimented with looser constraint by adding quiet periods on start and end points, and did not find any significant difference in results.

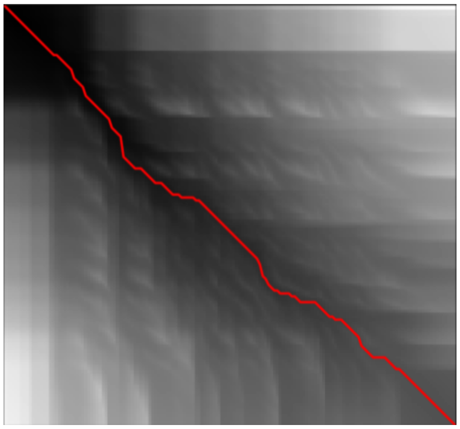

Fig. 3. **Cumulative cost matrix for dynamic programming:** The figure shows a sample matrix containing cumulative frame matching costs for a reference video and unaligned audio pair. Each matrix element contains the cumulative cost of matching an audio frame (row) and a video frame (column). Darker entries correspond to lower cost, and the optimal alignment is the path which minimizes the overall matching cost (shown here in red). Note that there are many similar structures because, for example, two different silent segments would have similar rows.

If other modalities exist, i.e reference audio and unaligned video, we compute 4 cross distances between embeddings of reference and unaligned, and assign the matrix element with the *minimal* of all four. This helps mitigate effects of embedding noise from e.g face occlusion or sudden disrupting sounds. We found out that even in the absence of such noise, combining different modalities improves the alignment.

In terms of our cost matrix, Syncnet's global shift corresponds to selecting the path as a diagonal on the matrix.

To avoid unnecessary computations, we only compute costs of nodes and edges in a strip around the 'diagonal' $(0, 0) \rightarrow (N, M)$. A sample full matrix is shown in Figure 3 for visualization.

### 3.3 Smoothing the Path

While the optimal path between sequences of embeddings is found, the quality of the generated audio based on that path may be degraded due to strong accelerations in the alignment. We first smooth the path with a Laplacian filter, then with a Gaussian. The amount of smoothing is chosen adaptively so that the smoothed path will not deviate from the original by more than a predefined value $\lambda$. Usually we set $\lambda < 0.1$ seconds, well within the boundaries of undetectable misalignment. This value may be changed for signals with specific characteristics or for artistic needs. After smoothing, the path is no longer integer valued, and interpolation is needed for voice synthesis.

### 3.4 Synthesis of New Signal

In standard ADR, the task is to warp the audio without modifying the original video frames. Therefore, we use the alignment to guide a standard audio warp method. We use a fairly simple phase vocoder [Laroche and Dolson 1999] to stretch and compress the audio stream according to the alignment, without affecting the pitch. This method uses the short-time Fourier transform (STFT) computed over the audio signal. We used audio sampled at 16KHz, each STFT bin (including complex values for each time-frequency cell) was computed on a window of size 512 audio samples, with 1/2 window overlap between consecutive STFT bins. The STFT magnitude is time warped, and phases are fixed to maintain phase differences between consecutive STFT windows. Since our alignment is based on video frames, its accuracy is only at time steps of 40ms, while the time step between STFT bins 16 ms. We create the alignment between STFT bin by re-sampling the frame-level alignment.

## 4 EXPERIMENTS AND RESULTS

Our main motivating application is fully automating the process of dialogue replacement in movies, such that the re-recorded speech can be combined with the original video footage in a seamless manner. We tested our method, both quantitatively and qualitatively, in a variety of scenarios.

*Evaluation.* Quantitative evaluation was performed using a human perception-inspired metric, based on the maximum acceptable audio-visual asynchrony used in the broadcasting industry. According to the International Telecommunications Union (ITU), the auditory signal should not lag by more than 125 ms or lead by more than 45 ms. Therefore, the error metric we use is the percentage of frames in the aligned signal which fall outside of the above acceptable range, compared to the ground truth alignment.

### 4.1 Alignment of Dually-Recorded Sentences

In this task, given a sentence recorded twice by the same person—one *reference* signal, and the other *unaligned*—the goal is to find the optimal mapping between the two, and warp the unaligned audio such that it becomes aligned with the reference video.

To our knowledge, there are no publicly available audio-visual datasets containing this kind of dually-recorded sentences, which are necessary for evaluating our method. To this end, we collected recordings of the same two sentences (*sa1* and *sa2* from the TIMIT dataset [S Garofolo et al. 1992]) made by four male speakers and one female speaker. The only instruction given to the speakers was to speak naturally. Therefore, the differences in pitch and timing between the recordings were noticeable, but not extremely distinct.

The dataset for this experiment was generated by mixing the original *unaligned* recordings with two types of noise, at varying signal-to-noise (SNR) levels. The types of noise we used, *crowd* and *wind*, are characteristic of interferences in indoor and outdoor recording environments, respectively. In order to demonstrate the effectiveness of our approach in noisy scenarios, we generated noise at three different levels: 0, $-5$, and $-10$ dB.

Alignment of each segment is performed using the following dynamic programming setups: (*a*) Alignment of unaligned audio to reference video (*Audio-to-video*); (*b*) Audio-to-video alignment with the additional delay constraint detailed in Section 3.2 (*Audio-to-video + delay*); (*c*) All combinations of modality-to-modality alignment, namely, audio-to-audio, audio-to-video, video-to-audio, and video-to-video, taking the step with minimum cost at each timestep (*All combinations*); (*d*) All modality combinations, with the additional delay constraint (*All combinations + delay*).

We compare our method to the state-of-the-art audio-to-audio alignment method of King et al. [2012], which has been implemented as the *Automatic Speech Alignment* feature in the Adobe Audition digital audio editing software [Wixted 2012]. This method uses noise-robust features as input to a dynamic time warping algorithm, and obtains good results when the reference signal is not badly degraded. As a baseline, we also compare to the method of Chung and Zisserman [2016] for finding a global offset between signals, whose audio-visual features we use as input to our method.

Since we have no ground truth mapping between each pair of recorded sentences, we adopt the method used by King et al. [2012] for calculating a "ground truth" alignment. They use conventional Mel-Frequency Cepstral Coefficients (MFCCs) to calculate alignment between reference and unaligned audio clips, with no noise added to the reference. Time-aligned synthesized "ground truth" signals were manually verified to be satisfactory, by checking audio-visual synchronization and comparing spectrograms.

Table 1 shows the superiority of our approach, with the error expressed as percentage of aligned frames outside the undetectable asynchrony range (-125 to +45 ms). The results demonstrate that at even at lower noise levels, our AV and combined modality approaches give improved performance over existing methods. At extreme noise levels, e.g crowd noise at -10 dB, our combined method has an average of only around 2% of frames outside the undetectable region, whereas the method of King et al. [2012] has over 10.5%. Alignment using SyncNet results in 88% of the frames outside the undetectable region, has it attempts to find an optimal global offset.

Figure 4 shows examples of reference, unaligned and aligned video and audio waveforms for one of the dually-recorded sentences in our dataset. The videos of this example can be viewed in our supplementary video.

### 4.2 Alignment of Synthetically Warped Sentences

In this task, we set out to investigate the limits of our method, in terms of degradation of both the audio and video parts of the reference signal as well as optimal segment duration for alignment.

Table 1. **Quantitative analysis and comparison with prior art:** This table shows the superiority of our approach over (i) a state-of-the-art audio-to-audio alignment method, implemented as feature in Adobe Audition [King et al. 2012], and (ii) SyncNet [Chung and Zisserman 2016]. The error is expressed as percentage of aligned frames outside the undetectable asynchrony range (-125 to +45 ms). The results demonstrate that even at lower noise levels, our Audio-to-Video and our combined modality (Audio to Video+Audio) approaches have improved performance over existing methods. At extremely high noise levels, our method has a clear and significant advantage. The delay is described in Sec. 3.2.

| | **"Crowd" noise** | | | **"Wind" noise** | | |
|---|---|---|---|---|---|---|
| | 0 dB | -5 dB | -10 dB | 0 dB | -5 dB | -10 dB |
| SyncNet [Chung and Zisserman 2016] | 88.49 | 88.49 | 88.49 | 88.49 | 88.49 | 88.49 |
| Adobe Audition [King et al. 2012] | 4.07 | 10.23 | 10.61 | 4.85 | 4.93 | 10.09 |
| Audio-to-Video | 7.26 | 7.26 | 7.26 | 7.26 | 7.26 | 7.26 |
| Audio-to-Video (with delay) | 4.12 | 4.12 | 4.12 | 4.12 | 4.12 | 4.12 |
| Audio to Video+Audio | 2.03 | 1.98 | **2.03** | 3.77 | 5.04 | 5.85 |
| Audio to Video+Audio (with delay) | **0.61** | **0.88** | 4.25 | **1.21** | **1.22** | **4.03** |

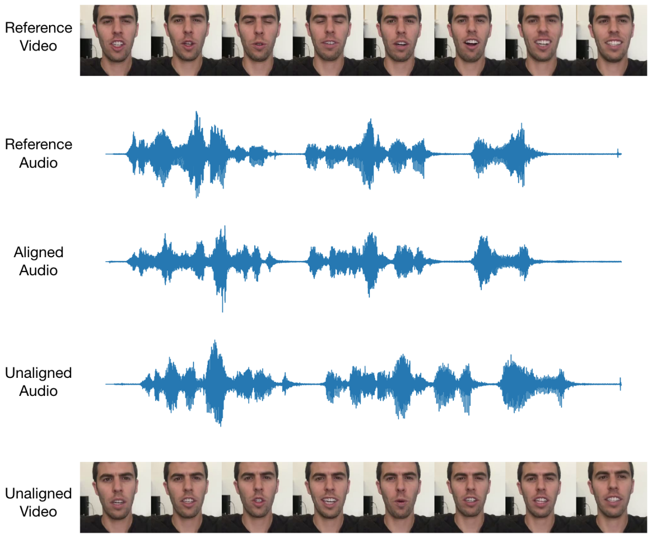

Fig. 4. **Example of reference and unaligned waveforms:** This figure shows examples of reference, unaligned and aligned video and audio waveforms for one of the dually-recorded sentences in our dataset.

To this end, we use segments from a dataset containing weekly addresses given by former president Barack Obama, which are synthetically warped using mappings obtained from the dataset we created for the previous experiment. These mappings are representative of the natural variation in pronunciation when people record the same sentence twice. The goal in this experiment is to find the optimal alignment between the original reference video and the synthetically warped video.

*Robustness to signal degradation.* In order to test the robustness of our method to various forms of degraded reference signals, we start with 100 same-length segments from the Obama dataset, and degrade the reference signals in following ways: (*i*) by adding crowd noise at −10 dB to each reference audio signal; (*ii*) by silencing a random one-second segment of each reference audio signal; (*iii*) by occluding a random one-second segment of each reference video sequence with a black frame; (*iv*) by combining random silencing and random occlusions (*ii* + *iii*).

Each reference degradation is tested using the dynamic programming setups used in the previous experiment, namely: *Audio-to-video*, *Audio-to-video + delay*, *All combinations*, and *All combinations + delay*. Here too, we compare to the global offset method of Chung and Zisserman [2016], and add the error percentage of frames in the *Unaligned* signal as a baseline.

Table 2 shows the results of this experiment. When the audio is severely degraded with either loud noise or random silence, performing direct audio-to-video alignment performs best. When the reference video signal is degraded with occlusions, our method relies more on the audio signal, and combining both the audio and video of the reference video works best. Example videos of degraded reference video and the resulting alignment can be viewed in our supplementary video.

*Effect of segment duration.* In order to investigate the effect segment duration has on alignment performance, we performed alignment on 100 segments from the Obama dataset of various durations between 3 to 15 seconds. There was no clear trend in the results of this study, leading us conclude that segment duration (within the aforementioned range) has a negligible effect on the performance of our method.

### 4.3 Alignment of Two Different Speakers

While not the main focus of our work, various additional alignment scenarios can be addressed using our audio-visual alignment method. One of these is alignment of two different speakers.

Since audio and visual signals are mapped to a joint synchronization embedding space which, presumably, places little emphasis on the identity of the speaker, we can use our method to align of two different speakers saying the same text. For this task, we use videos from the TCD-TIMIT dataset [Harte and Gillen 2015], which consists of 60 volunteer speakers reciting various sentences from the TIMIT dataset [S Garofolo et al. 1992]. We evaluated our results qualitatively, and included an example in our supplementary video, involving alignment between male and female subjects. Figure 5

Table 2. **Analysis of robustness to degraded reference signals:** Alignment performance when the reference signal has undergone several types of degradation: (i) High noise, (ii) random 2-second silence in audio and (iii) 2-second blackout of video frames. The error is expressed as percentage of aligned frames outside the undetectable asynchrony range. Note that using SyncNet [Chung and Zisserman 2016], performing just a time shift, resukted in more errors than even the unaligned sound.

| | **Crowd noise (-10dB)** | **Random silence** | **Random occlusion** | **Silence + occlusions** |
|---|---|---|---|---|
| Unaligned Voice | 33.77 | 34.03 | 37.87 | 32.73 |
| SyncNet [Chung and Zisserman 2016] | 71.37 | 73.34 | 78.74 | 84.0 |
| Audio-to-Video | **2.63** | 2.92 | 16.16 | 15.17 |
| Audio-to-Video (with delay) | 3.35 | **2.62** | 14.17 | 9.76 |
| Audio to Video+Audio | 2.83 | 2.71 | **3.08** | 5.78 |
| Audio to Video+Audio (with delay) | 5.45 | 3.14 | 4.12 | **5.04** |

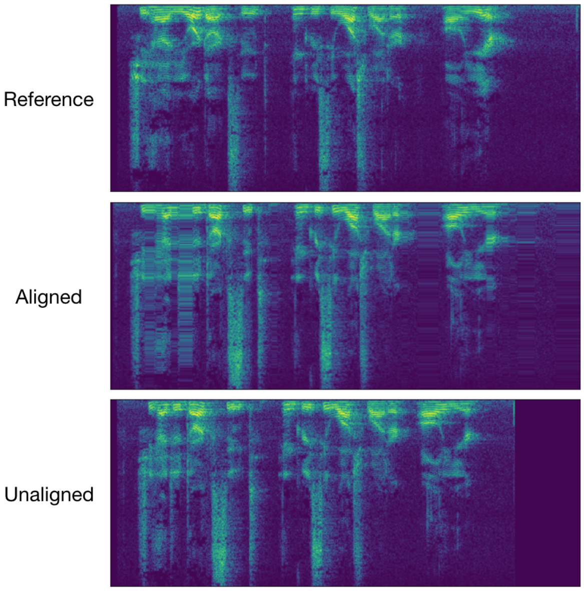

Fig. 5. **Spectrograms of alignment of different speakers:** This figure shows spectrograms of three signals: speech of one speaker used as reference (top); speech of a different speaker who we would like to align to the reference (bottom); aligned speech using our audio-to-video method (middle).

shows example spectrograms of reference, unaligned and aligned signals of two different speakers.

## 5 LIMITATIONS AND CONCLUSION

Our method is currently limited by the quality of its synthesized speech, which is sometimes of poorer quality than original due to challenging warps. Also, in cases of clean reference speech, our method is comparable to existing audio-to-audio alignment.

In conclusion, a method was presented to align speech to lip movements in video using dynamic time warping. The alignment is based on deep features that map both the face in the video and the speech into a common embedding space. Our method makes it easy to create accurate Automated Dialogue Replacement (ADR), and have shown it to be superior to existing methods, both quantitatively and qualitatively. ADR is possible using speech of the original speaker, or even the speech of another person.

## REFERENCES

Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. 2018. The Conversation: Deep Audio-Visual Speech Enhancement. *arXiv preprint arXiv:1804.04121* (2018).

Hervé Bredin and Gérard Chollet. 2007. Audiovisual speech synchrony measure: application to biometrics. *EURASIP Journal on Applied Signal Processing* 2007, 1 (2007), 179–179.

Christoph Bregler, Michele Covell, and Malcolm Slaney. 1997. Video rewrite: Driving visual speech with audio. In *Proceedings of the 24th annual conference on Computer graphics and interactive techniques*. ACM Press/Addison-Wesley Publishing Co., 353–360.

BT.1359. 1998. Relative Timing of Sound and Vision for Broadcasting. *ITU* (1998).

Yong Cao, Petros Faloutsos, Eddie Kohler, and Frédéric Pighin. 2004. Real-time speech motion synthesis from recorded motions. In *Proceedings of the 2004 ACM SIGGRAPH/Eurographics symposium on Computer animation*. Eurographics Association, 345–353.

Yao-Jen Chang and Tony Ezzat. 2005. Transferable videorealistic speech animation. In *Proceedings of the 2005 ACM SIGGRAPH/Eurographics symposium on Computer animation*. ACM, 143–151.

Joon Son Chung, Amir Jamaludin, and Andrew Zisserman. 2017. You said that? *arXiv preprint arXiv:1705.02966* (2017).

Joon Son Chung, Andrew W. Senior, Oriol Vinyals, and Andrew Zisserman. 2016. Lip Reading Sentences in the Wild. *CoRR* abs/1611.05358 (2016).

Joon Son Chung and Andrew Zisserman. 2016. Out of time: automated lip sync in the wild. In *Asian Conference on Computer Vision*. Springer, 251–263.

Edsger W Dijkstra. 1959. A note on two problems in connexion with graphs. *Numerische mathematik* 1, 1 (1959), 269–271.

Ariel Ephrat, Tavi Halperin, and Shmuel Peleg. 2017. Improved speech reconstruction from silent video. In *ICCV 2017 Workshop on Computer Vision for Audio-Visual Media*.

Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T Freeman, and Michael Rubinstein. 2018. Looking to Listen at the Cocktail Party: A Speaker-Independent Audio-Visual Model for Speech Separation. *ACM Transactions on Graphics (TOG)* 37, 4 (2018).

Ariel Ephrat and Shmuel Peleg. 2017. Vid2speech: speech reconstruction from silent video. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 5095–5099.

Weijiang Feng, Naiyang Guan, Yuan Li, Xiang Zhang, and Zhigang Luo. 2017. Audio visual speech recognition with multimodal recurrent neural networks. In *Neural Networks (IJCNN), 2017 International Joint Conference on*. IEEE, 681–688.

Shoichi Furukawa, Takuya Kato, Pavel Savkin, and Shigeo Morishima. 2016. Video reshuffling: automatic video dubbing without prior knowledge. In *ACM SIGGRAPH 2016 Posters*. ACM, 19.

Pablo Garrido, Levi Valgaerts, Hamid Sarmadi, Ingmar Steiner, Kiran Varanasi, Patrick Perez, and Christian Theobalt. 2015. Vdub: Modifying face video of actors for plausible visual alignment to a dubbed audio track. In *Computer Graphics Forum*, Vol. 34. Wiley Online Library, 193–204.

Dian Gong and Gerard Medioni. 2011. Dynamic manifold warping for view invariant action recognition. In *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 571–578.

Tavi Halperin, Yair Poleg, Chetan Arora, and Shmuel Peleg. 2018. Egosampling: Wide view hyperlapse from egocentric videos. *IEEE Transactions on Circuits and Systems for Video Technology* 28, 5 (2018), 1248–1259.

Naomi Harte and Eoin Gillen. 2015. TCD-TIMIT: An audio-visual corpus of continuous speech. *IEEE Transactions on Multimedia* 17, 5 (2015), 603–615.

John-Paul Hosom. 2000. Automatic time alignment of phonemes using acoustic-phonetic information. (2000).

Brian King, Paris Smaragdis, and Gautham J Mysore. 2012. Noise-robust dynamic time warping using PLCA features. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on.* IEEE, 1973–1976.

Jean Laroche and Mark Dolson. 1999. New phase-vocoder techniques for pitch-shifting, harmonizing and other exotic effects. In *Applications of Signal Processing to Audio and Acoustics, 1999 IEEE Workshop on.* IEEE, 91–94.

John Lewis. 1991. Automated lip-sync: Background and techniques. *Computer Animation and Virtual Worlds* 2, 4 (1991), 118–122.

Etienne Marcheret, Gerasimos Potamianos, Josef Vopicka, and Vaibhava Goel. 2015. Detecting audio-visual synchrony using deep neural networks. In *Sixteenth Annual Conference of the International Speech Communication Association.*

Shigeo Morishima, Shin Ogata, Kazumasa Murai, and Satoshi Nakamura. 2002. Audio-visual speech translation with automatic lip syncqronization and face tracking based on 3-d head model. In *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, Vol. 2. IEEE, II–2117.

Youssef Mroueh, Etienne Marcheret, and Vaibhava Goel. 2015. Deep multimodal learning for audio-visual speech recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on.* IEEE, 2130–2134.

Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y. Ng. 2011. Multimodal Deep Learning. In *ICML*.

Brett Ninness and Soren John Henriksen. 2008. Time-scale modification of speech signals. *IEEE Transactions on Signal Processing* 56, 4 (2008), 1479–1488.

Andrew Owens and Alexei A Efros. 2018. Audio-Visual Scene Analysis with Self-Supervised Multisensory Features. *arXiv preprint arXiv:1804.03641* (2018).

Lawrence R Rabiner and Biing-Hwang Juang. 1993. *Fundamentals of speech recognition.* Vol. 14. PTR Prentice Hall Englewood Cliffs.

J S Garofolo, Lori Lamel, W M Fisher, Jonathan Fiscus, D S. Pallett, N L. Dahlgren, and V Zue. 1992. TIMIT Acoustic-phonetic Continuous Speech Corpus. (11 1992).

Hiroaki Sakoe and Seibi Chiba. 1978. Dynamic programming algorithm optimization for spoken word recognition. *IEEE transactions on acoustics, speech, and signal processing* 26, 1 (1978), 43–49.

Mehmet Emre Sargin, Yücel Yemez, Engin Erzin, and A Murat Tekalp. 2007. Audiovisual synchronization and fusion using canonical correlation analysis. *IEEE Transactions on Multimedia* 9, 7 (2007), 1396–1403.

Supasorn Suwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman. 2017. Synthesizing obama: learning lip sync from audio. *ACM Transactions on Graphics (TOG)* 36, 4 (2017), 95.

Sarah Taylor, Taehwan Kim, Yisong Yue, Moshe Mahler, James Krahe, Anastasio Garcia Rodriguez, Jessica Hodgins, and Iain Matthews. 2017. A deep learning approach for generalized speech animation. *ACM Transactions on Graphics (TOG)* 36, 4 (2017), 93.

Justus Thies, Michael Zollhöfer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. 2016. Face2face: Real-time face capture and reenactment of rgb videos. In *Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on.* IEEE, 2387–2395.

Ellen Wixted. 2012. Interview With the Creator of the Automatic Speech Alignment Feature in Audition CS6. https://blogs.adobe.com/creativecloud/interview-with-the-creator-of-the-automatic-speech-alignment-feature-in-audition-cs6/. (2012). Accessed: 2018-06-04.

Feng Zhou, Fernando De la Torre, and Jessica K Hodgins. 2008. Aligned cluster analysis for temporal segmentation of human motion. In *Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference on.* IEEE, 1–7.

Goranka Zoric and Igor S Pandzic. 2005. A real-time lip sync system using a genetic algorithm for automatic neural network configuration. In *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on.* IEEE, 1366–1369.