DYNAMIC TEMPORAL ALIGNMENT OF SPEECH TO LIPS

Tavi Halperin[*1] *Ariel Ephrat*[*2] *Shmuel Peleg*[1]

[1] The Hebrew University of Jerusalem, Israel [2] Google Research

ABSTRACT

Many speech segments in movies are re-recorded in a studio during post-production, to compensate for poor sound quality as recorded on location. We present an audio-to-video 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 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. In these cases state-of-the-art audio only methods will fail.

Index Terms— Automatic Dialogue Replacement

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.

We propose an automatic method to temporally align audio and video of a speaking person by using innovative deep audio-visual (AV) features that were suggested by Chung and Zisserman [1]. These features map the lips video and the speech signal to a shared representation. We use these features for dynamic temporal alignment, stretching and compressing the signal dynamically within a clip. This is usually a three-step process [2]: (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) [3]; (iii) a warped version of the unaligned signal is synthesized so that it temporally matches the reference signal [4].

We demonstrate the benefits of our approach over a state-of-the-art audio-only alignment method [5], and over [1], 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 [6]. In order to evaluate the perceptive quality of our method's output, our quantitative error measure is therefore based on these thresholds. It should be noted that comparison to an audio-only alignment method can only be performed when a clear reference audio signal exists, which may not always be the case. In that scenario, 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 audio-visual fully automated dialogue replacement (AV-ADR). We leverage the strength of deep audio-visual speech synchronization features and suggest a dynamic temporal alignment method.

1.1. Automatic time alignment of sequences

Dynamic time warping (DTW) [7] 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 area of speech processing [7, 2, 5] as well as in computer vision for video [8, 9, 10] for e.g. temporal segmentation and frame sampling, among many scientific disciplines.

King et al. [5] proposed a 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 (or unavailable) audio only methods will fail. Our method overcomes this problem by performing audio-to-video alignment, resulting in higher-quality synchronization.

* Equal contribution. Ariel performed this work while at HUJI.

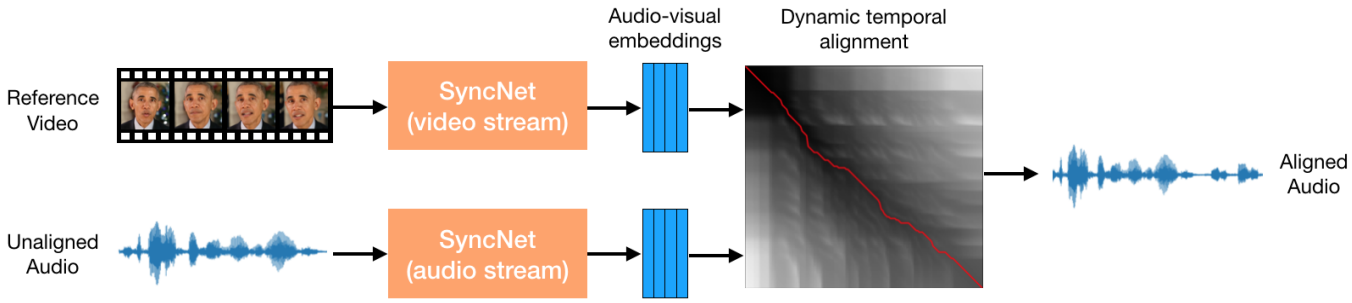

Fig. 1: **High-level diagram of our speech to lips alignment:** Given unaligned video and speech: (i) SyncNet features are computed for both; (ii) dynamic time warping is performed for optimal alignment between the features; (iii) A new speech is synthesized that is now aligned with the video.

1.2. 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. The common denominator of the AV-sync 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.

Several methods have been proposed which attempt to find audio-visual correspondences, such as [11] and [12] who use canonical correlation analysis (CCA).

Recently, there has been increased interest in leveraging natural synchrony of simultaneously recorded video and speech for various tasks. These include predicting a speech signal or text from silent video [13, 14, 15], and audio-visual speech enhancement [16, 17, 18]. In a recent pioneering work [1] 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. 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.

2. METHOD

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 1.

2.1. Audio-Visual Feature Extraction

We use SyncNet [1] 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 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.

2.2. 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, therefore 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 [19]. 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 ele-

ment 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 3.

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.

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' $(0, 0) \rightarrow (N, M)$.

2.3. 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 λ. 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.

2.4. Synthesis of New Signal

We use a fairly simple phase vocoder [20] to stretch and compress the audio stream according to the alignment, without affecting the pitch. We used audio sampled at 16KHz, each STFT bin was computed on a window of size 512, with $1/2$ window overlap.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 40 ms, while the time step between STFT bins 16 ms. We create the alignment between STFT bin by re-sampling the frame-level alignment.

3. EXPERIMENTS AND RESULTS

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.

3.1. 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 [21]) 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. An example is provided in our supplementary video[1].

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.

Alignment of each segment is performed using the following dynamic programming setups: (a) Alignment of unaligned audio to reference video (*A2V*); (b) A2V alignment with the additional delay constraint detailed in Section 2.2 (*A2V delay*); (c) All four combinations of modality-to-modality alignment, taking the step with minimum cost at each timestep (*AV2AV*); (d) All modality combinations, with the additional delay constraint (*AV2AV delay*).

We compare our method to [5], which has been implemented as the *Automatic Speech Alignment* feature in the Adobe Audition digital audio editing software [22]. 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 [1] for finding a global offset between signals.

[1] https://youtu.be/t7m0yEnBG7M

Table 1: Comparison to (i) a state-of-the-art audio-to-audio alignment method, implemented as a feature in Adobe Audition [5], and (ii) SyncNet [1]. The results demonstrate that even at lower noise levels, our Audio-to-Video (A2V) and our combined modality (AV2AV) approaches have improved performance over existing methods. The delay is described in Sec. 2.2.

	"Crowd" noise			"Wind" noise		
Noise dB	**0**	**-5**	**-10**	**0**	**-5**	**-10**
SyncNet [1]	88.4	88.4	88.4	88.4	88.4	88.4
Adobe Audition [5]	4.0	10.2	10.6	4.8	4.9	10.0
A2V	7.2	7.2	7.2	7.2	7.2	7.2
A2V delay	4.1	4.1	4.1	4.1	4.1	4.1
AV2AV	2.0	1.9	**2.0**	3.7	5.0	5.8
AV2AV delay	**0.6**	**0.8**	4.2	**1.2**	**1.2**	**4.0**

Since we have no ground truth mapping between each pair of recorded sentences, we adopt the method used by [5] 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.The results demonstrate that even at lower noise levels, our A2V and combined modality approaches give improved performance over existing methods. At extreme noise levels the improvement becomes even more significant.

3.2. Alignment of Synthetically Warped Sentences

In this task, we investigate the limits of our method, in terms of degradation of both the audio and video parts of the reference signal. 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.

3.3. 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 the following ways: (i) by adding crowd noise at -10 dB; (ii) by silencing a random one-second segment; (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).

Table 2: Alignment performance when the reference signal has undergone several types of degradation: (i) High noise (-10 dB), (ii) random 2-second silence in audio and (iii) 2-second blackout of video frames.

	Crowd noise	**Random silence**	**Random occlusion**	**Silence + occlusions**
Unaligned Voice	33.77	34.03	37.87	32.73
A2V	**2.63**	2.92	16.16	15.17
A2V delay	3.35	**2.62**	14.17	9.76
AV2AV	2.83	2.71	**3.08**	5.78
AV2AV delay	5.45	3.14	4.12	**5.04**

Each reference degradation is tested using the dynamic programming setups used in the previous experiment.We 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.

3.4. 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 two different speakers saying the same text. For this task, we used videos from the TCD-TIMIT dataset [23].We evaluated our results qualitatively, and included an example in our supplementary video, involving alignment between male and female subjects.

4. CONCLUSION

We presented a method 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 Audio-Visual Automated Dialogue Replacement (AV-ADR), and have shown state-of-the-art performance.

5. REFERENCES

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

[2] John-Paul Hosom, "Automatic time alignment of phonemes using acoustic-phonetic information," 2000.

[3] Lawrence R Rabiner and Biing-Hwang Juang, *Fundamentals of speech recognition*, vol. 14, PTR Prentice Hall Englewood Cliffs, 1993.

[4] Brett Ninness and Soren John Henriksen, "Time-scale modification of speech signals," *IEEE Transactions on Signal Processing*, vol. 56, no. 4, pp. 1479–1488, 2008.

[5] Brian King, Paris Smaragdis, and Gautham J Mysore, "Noise-robust dynamic time warping using plca features," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 1973–1976.

[6] BT.1359, "Relative timing of sound and vision for broadcasting," *ITU*, 1998.

[7] Hiroaki Sakoe and Seibi Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE transactions on acoustics, speech, and signal processing*, vol. 26, no. 1, pp. 43–49, 1978.

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

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

[10] Tavi Halperin, Yair Poleg, Chetan Arora, and Shmuel Peleg, "Egosampling: Wide view hyperlapse from egocentric videos," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 5, pp. 1248–1259, 2018.

[11] Hervé Bredin and Gérard Chollet, "Audiovisual speech synchrony measure: application to biometrics," *EURASIP Journal on Applied Signal Processing*, vol. 2007, no. 1, pp. 179–179, 2007.

[12] Mehmet Emre Sargin, Yücel Yemez, Engin Erzin, and A Murat Tekalp, "Audiovisual synchronization and fusion using canonical correlation analysis," *IEEE Transactions on Multimedia*, vol. 9, no. 7, pp. 1396–1403, 2007.

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

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

[15] Joon Son Chung, Andrew W. Senior, Oriol Vinyals, and Andrew Zisserman, "Lip reading sentences in the wild," *CoRR*, vol. abs/1611.05358, 2016.

[16] Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T Freeman, and Michael Rubinstein, "Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation," *ACM Transactions on Graphics (TOG)*, vol. 37, no. 4, 2018.

[17] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman, "The conversation: Deep audio-visual speech enhancement," *arXiv preprint arXiv:1804.04121*, 2018.

[18] Andrew Owens and Alexei A Efros, "Audio-visual scene analysis with self-supervised multisensory features," *arXiv preprint arXiv:1804.03641*, 2018.

[19] Edsger W Dijkstra, "A note on two problems in connexion with graphs," *Numerische mathematik*, vol. 1, no. 1, pp. 269–271, 1959.

[20] Jean Laroche and Mark Dolson, "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, 1999, pp. 91–94.

[21] J S Garofolo, Lori Lamel, W M Fisher, Jonathan Fiscus, D S. Pallett, N L. Dahlgren, and V Zue, "Timit acoustic-phonetic continuous speech corpus," 11 1992.

[22] Ellen Wixted, "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.

[23] Naomi Harte and Eoin Gillen, "Tcd-timit: An audio-visual corpus of continuous speech," *IEEE Transactions on Multimedia*, vol. 17, no. 5, pp. 603–615, 2015.