Vocoder-Based Speech Synthesis from Silent Videos

Michelsanti, Daniel; Slizovskaia, Olga; Haro, Gloria; Gómez, Emilia; Tan, Zheng-Hua; Jensen, Jesper

Published in:
Interspeech 2020

DOI (link to publication from Publisher):
10.21437/Interspeech.2020-1026

Publication date:
2020

Document Version
Publisher's PDF, also known as Version of record

INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Vocoder-Based Speech Synthesis from Silent Videos

Daniel Michelsanti[1], *Olga Slizovskaia*[2], *Gloria Haro*[2], *Emilia Gómez*[2], *Zheng-Hua Tan*[1], *Jesper Jensen*[1,3]

[1] Aalborg University, Aalborg, Denmark
[2] Universitat Pompeu Fabra, Barcelona, Spain
[3] Oticon A/S, Smørum, Denmark

{danmi, zt, jje}@es.aau.dk {olga.slizovskaia, gloria.haro, emilia.gomez}@upf.edu

Abstract

Both acoustic and visual information influence human perception of speech. For this reason, the lack of audio in a video sequence determines an extremely low speech intelligibility for untrained lip readers. In this paper, we present a way to synthesise speech from the silent video of a talker using deep learning. The system learns a mapping function from raw video frames to acoustic features and reconstructs the speech with a vocoder synthesis algorithm. To improve speech reconstruction performance, our model is also trained to predict text information in a multi-task learning fashion and it is able to simultaneously reconstruct and recognise speech in real time. The results in terms of estimated speech quality and intelligibility show the effectiveness of our method, which exhibits an improvement over existing video-to-speech approaches.

Index Terms: Speech synthesis, lip reading, deep learning, vocoder

1. Introduction

Most of the events that we experience in our life consist of visual and acoustic stimuli. Recordings of such events may lack the acoustic component, for example due to limitations of the recording equipment or technical issues in the transmission of the information. Since acoustic and visual modalities are often correlated, methods to reconstruct audio signals using videos have been proposed [1, 2, 3].

In this paper, we focus on one particular case of the aforementioned problem: *speech reconstruction (or synthesis) from a silent video*. Solving this task might be useful to automatically generate speech for surveillance videos and for extremely challenging speech enhancement applications, e.g. hearing assistive devices, where noise completely dominates the target speech, making the acoustic signal worth less than its video counterpart.

A possible way to tackle the problem is to decompose it into two steps: first, a *visual speech recognition* (VSR) system [4, 5, 6] predicts the spoken sentences from the video; then, a *text-to-speech* (TTS) model [7, 8, 9] synthesises speech based on the output of the VSR system. However, at least two drawbacks can be identified when such an approach is used. In order to generate speech from text, each word should be spoken in its entirety to be processed by the VSR and the TTS systems, imposing great limitations for real-time applications. Furthermore, when the TTS method is applied, useful information that should be captured by the system, such as emotion and prosody, gets lost, making the synthesised speech unnatural. For these reasons, approaches that estimate speech from a video, without using text as an intermediate step, have been proposed.

Le Cornu and Miller [10, 11] developed a video-to-speech method with a focus on speech intelligibility rather than quality. This is achieved by estimating spectral envelope (SP) audio features from visual features and then reconstructing the time-domain signal with the STRAIGHT vocoder [12]. Since the vocoder also requires other audio features, i.e. the fundamental frequency (F0) and the aperiodic parameter (AP), these are artificially created independently of the visual features.

Ephrat and Peleg [13] treated speech reconstruction as a regression problem using a neural network which takes as input raw visual data and predicts a line spectrum pairs (LSP) representation of linear predictive coding (LPC) coefficients computed from the audio signal. The waveform is reconstructed from the estimated audio features using Gaussian white noise as excitation, producing unnatural speech. This issue is tackled in a subsequent work [14], where a neural network estimates the mel-scale spectrogram of the audio from video frames and optical flow information derived from the visual input. The time-domain speech signal is reconstructed using either example-based synthesis, in which estimated audio features are replaced with their closest match in the training set, or speech synthesis from predicted linear-scale spectrograms.

Akbari et. al. [15] tried to reconstruct natural sounding speech using a neural network that takes as input the face region of the talker and estimates bottleneck features extracted from the auditory spectrogram by a pre-trained autoencoder. The time-domain signal is obtained with the algorithm in [16]. This approach shows its effectiveness when compared to [13].

All the methods reported until now have a major limitation: they estimate either a magnitude spectrogram, SPs or LSPs, which do not contain all the information of a speech signal. Vougioukas et al. [17] addressed this issue and proposed an end-to-end model that can directly synthesise audio waveforms from videos using a generative adversarial network (GAN). However, their direct estimation of a time-domain signal causes artefacts in the reconstructed speech.

In this work, we propose an approach, *vid2voc*, to estimate WORLD vocoder [18] features from the silent video of a speaker[1]. We trained the systems using either the whole face or the mouth region only, since previous work [13] shows a benefit in using the entire face. Our method differs from the work in [10, 11], because we predict all the vocoder features (not only SP) directly from raw video frames. The estimation of F0 and AP, alongside with SP, allows to have a framework with a focus on speech intelligibility (as in [10, 11]) and speech quality, able to outperform even the recently proposed GAN-based approach in [17] in several conditions. In addition, we train a system that can simultaneously perform speech reconstruction

[1] Although this paper aims at synthesising speech from frontal-view silent videos, it is worth mentioning that some methods using multi-view video feeds have also been developed [19, 20, 21, 22].

http://dx.doi.org/10.21437/Interspeech.2020-1026

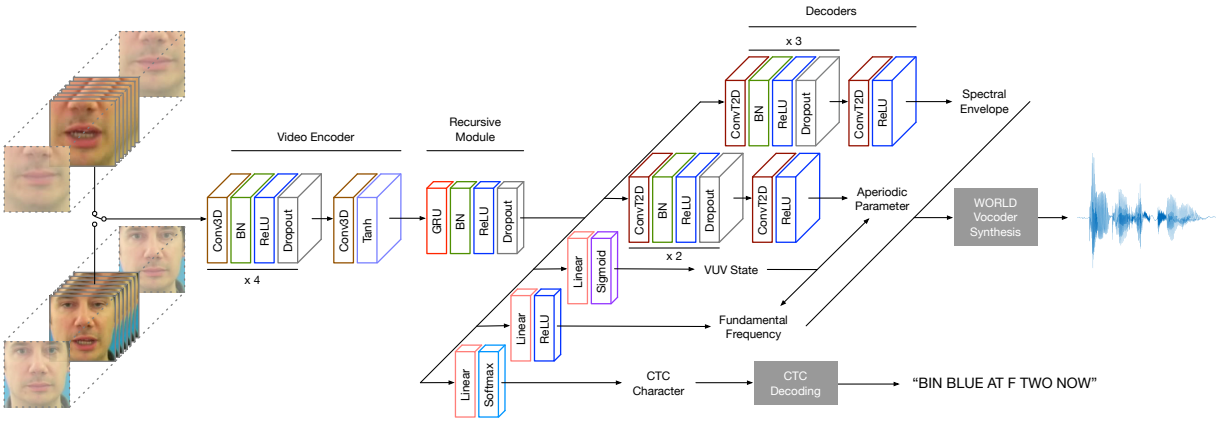

Figure 1: *Pipeline of our system. Conv3D: 3-D convolution. BN: Batch normalisation. GRU: Gated recurrent unit. ConvT2D: 2-D transposed convolution. VUV: Voiced-unvoiced. CTC: Connectionist temporal classification.*

(our main goal) and VSR, in a multi-task learning fashion. This can be useful in all the applications that require video captioning without adding considerable extra complexity to the system. Although Kumar et al. [21] incorporate a text-prediction model in their multi-view speech reconstruction pipeline, this model is trained separately from the main system and it is quite simple: it classifies encoded audio features estimated with a pre-trained network into 10 text classes. This makes the method dependent on the number of different sentences of the specific database used for training and not suitable for real-time applications. Instead, we make use of the more flexible connectionist temporal classification (CTC) [23] sequence modelling which has already shown its success in VSR [4].

Additional material, including samples of reconstructed speech that the reader is encouraged to listen to for a better understanding of the effectiveness of our approach, can be found in `https://danmic.github.io/vid2voc/`.

2. Methodology and Experimental Setup

2.1. Audio-Visual Speech Corpus

Experiments are conducted on the GRID corpus [24], which consists of audio and video recordings from 34 speakers (s1−34), 18 males and 16 females, each of them uttering 1000 six-word sentences with the following structure: <command> <color> <preposition> <letter> <digit> <adverb>. Each video has a resolution of 720×576 pixels, a duration of 3 s and a frame rate of 25 frames per second. The audio tracks have the same duration as the videos and a sample frequency of 50 kHz. In addition, text transcription for every utterance is provided.

As in [17], we evaluate our systems in speaker dependent and speaker independent settings. Regarding the speaker dependent scenario, the data from 4 speakers (s1, s2, s4, s29) is pooled together, then 90% of the data is used for training, 5% for validation and 5% for testing. Regarding the speaker independent scenario, the data from 15 speakers (s1, s3, s5−8, s10, s12, s14, s16, s17, s22, s26, s28, s32) is used for training, the data from 7 speakers (s9, s20, s23, s27, s29, s30, s34) for validation and the data from 10 speakers (s2, s4, s11, s13, s15, s18, s19, s25, s31, s33) for testing.

2.2. Audio and Video Preprocessing

The acoustic model used in this work is based on the WORLD vocoder [18], with a sample frequency of 50 kHz and a hop size of 250 samples[2]. WORLD consists of three analysis algorithms to determine SP, F0 and AP features, and a synthesis algorithm which incorporates these three features. Here, we use SWIPE [25] and D4C [26] to estimate F0 and AP, respectively. As done in [27], a dimensionality reduction of the features is applied: SP is reduced to 60 log mel-frequency spectral coefficients (MFSCs) and AP is reduced to 5 coefficients according to the D4C band-aperiodicity estimation. In addition, a voiced-unvoiced (VUV) state is obtained by thresholding the F0 obtained with SWIPE. All the acoustic features are min-max normalised using the statistics of the training set as in [28].

As in [17], videos are preprocessed as follows: first, the faces are aligned to the canonical face[3]; then, the video frames are normalised in the range $[-1, 1]$, resized to 128×96 pixels and, for the models that use only the mouth region as input, cropped preserving the bottom half; finally, the videos are mirrored with a probability of 0.5 during training.

2.3. Architecture and Training Procedure

As shown in Figure 1, our network maps video frames of a speaker to vocoder features and consists of a *video encoder*, a *recursive module* and five decoders: *SP decoder*, *AP decoder*, *VUV decoder*, *F0 decoder* and *VSR decoder*. We also tried not to use the VSR decoder, to see whether it has any impact on the performance.

The video encoder is inspired by [17]: it takes as input one video frame concatenated with the three previous and the three next frames and applies five 3-D convolutions (conv3D). Each of the first four convolutional layers is followed by batch normalisation (BN) [30], ReLU activation and dropout [31], while the last one is followed by Tanh activation.

To model the sequential nature of video data, a recursive module is used: it consists of a single-layer gated recurrent unit (GRU) [32], BN, ReLU activation and dropout.

Each decoder takes the GRU features as input. For ev-

[2]The window length is automatically determined by the WORLD algorithm.

[3]We use the face processor library in `https://github.com/DinoMan/face-processor`, which makes use of [29].

ery video frame the SP decoder produces an eight-frame-long estimate $\widehat{W}_{se} \in \mathbb{R}^{60\times 8}$ of the normalised dimensionality-reduced SP, W_{se}, through three 2-D transposed convolutions (convT2D), each followed by BN, ReLU activation and dropout, and another convT2D followed by ReLU activation.

The VUV decoder consists of a linear layer followed by ReLU activation. A threshold of 0.2 is applied to the output obtaining $\widehat{W}_{vuv} \in \mathbb{R}^{8}$, an estimate of the VUV state, W_{vuv}.

The AP decoder has a structure similar to the SP decoder, with a total of three convT2D in this case. Its output, $O_{nap} \in \mathbb{R}^{5\times 8}$, together with $\widehat{W}_{vuv}$ is used to get $\widehat{W}_{nap}$, an estimate of $W_{nap} = I_{5,8} - W_{ap}$, where $I_{5,8}$ indicates an all-ones matrix with 5 rows and 8 columns, and W_{ap} is the normalised dimensionality-reduced AP:

$$(\widehat{W}_{nap})_i = (O_{nap})_i \odot \widehat{W}_{vuv} \quad \text{for } i \in \{1,\dots,5\} \tag{1}$$

where $(A)_i$ indicates the i-th row of A and $\odot$ denotes the element-wise product.

The F0 decoder has a linear layer followed by a sigmoid activation function. Its output, $O_{f0} \in \mathbb{R}^{8}$, is point-wise multiplied with $\widehat{W}_{vuv}$ to obtain $\widehat{W}_{f0}$, an estimate of the normalised F0, W_{f0}:

$$\widehat{W}_{f0} = O_{f0} \odot \widehat{W}_{vuv}. \tag{2}$$

Finally, the VSR decoder, consisting of a linear and a softmax layers, outputs a CTC character that will be used to predict the text transcription of the utterance.

The system is trained to minimise the following loss:

$$J = \frac{\lambda_1}{\lambda} J_{se} + \frac{\lambda_2}{\lambda} J_{nap} + \frac{\lambda_3}{\lambda} J_{f0} + \frac{\lambda_4}{\lambda} J_{vuv} + \frac{\lambda_5}{\lambda} J_{vsr} \tag{3}$$

where $\lambda_1 = 600$, $\lambda_2 = 50$, $\lambda_3 = 10$, $\lambda_4 = 10$, $\lambda_5 = 1$, $\lambda = \sum_{i=1}^{5} \lambda_i$ and:

- J_{se}: mean squared error (MSE) between W_{se} and $\widehat{W}_{se}$.
- J_{nap}: MSE between W_{nap} and $\widehat{W}_{nap}$.
- J_{f0}: MSE between W_{f0} and $\widehat{W}_{f0}$.
- J_{vuv}: MSE between W_{vuv} and $\widehat{W}_{vuv}$.
- J_{vsr}: CTC loss [23] between the target text transcription and the estimated one.

Details regarding architecture and training hyperparameters can be found in Table 1.

2.4. Waveform Reconstruction and Lipreading

The network outputs are used to reconstruct the speech waveform with the WORLD synthesis algorithm [18] and to get a text transcription adopting the best path CTC decoding scheme [23].

2.5. Evaluation Metrics

The system is evaluated in terms of perceptual evaluation of speech quality (PESQ) [35] and extended short-time objective intelligibility (ESTOI) [36], two of the most used measures that provide estimates of speech quality and speech intelligibility, respectively. PESQ scores are in the range from -0.5 to 4.5 and ESTOI scores practically lie between 0 and 1. In both cases, higher values correspond to better performance.

For the systems having the VSR decoder, we also provide the word error rate (WER), a standard metric for automatic speech recognition systems. In this case, lower values correspond to better performance.

Table 1: *Architecture and training hyperparameters. Activation, batch normalisation and dropout omitted for brevity.*

Input Size					
$B\times S\times C\times F\times H\times W$					
Video Encoder					
Layer	Input Channels	Output Channels	Kernel Size	Stride	Padding
Conv3D	3	64	(7,4,4)	(1,2,2)	(0,1,1)
Conv3D	64	128	(1,4,4)	(1,2,2)	(0,1,1)
Conv3D	128	256	(1,4,4)	(1,d_1,2)	(0,1,1)
Conv3D	256	512	(1,4,4)	(1,2,2)	(0,1,1)
Conv3D	512	128	(1,d_2,6)	(1,1,1)	(0,0,0)
Recursive Module					
Layer	Input Size		Hidden Size		
GRU	128		128		
Spectral Envelope (SP) Decoder					
Layer	Input Channels	Output Channels	Kernel Size	Stride	Padding
ConvT2D	128	256	(1,6)	(1,1)	(0,0)
ConvT2D	256	128	(2,4)	(1,2)	(0,0)
ConvT2D	128	64	(4,4)	(1,2)	(0,0)
ConvT2D	64	1	(4,2)	(1,2)	(0,0)
Aperiodic Parameter (AP) Decoder					
Layer	Input Channels	Output Channels	Kernel Size	Stride	Padding
ConvT2D	128	128	(4,1)	(1,1)	(0,0)
ConvT2D	128	64	(3,3)	(1,1)	(0,0)
ConvT2D	64	1	(3,3)	(1,1)	(0,0)
Voiced-Unvoiced (VUV) Decoder					
Layer	Input Size		Output Size		
Linear	128		8[a]		
Fundamental Frequency (F0) Decoder					
Layer	Input Size		Output Size		
Linear	128		8[a]		
Visual Speech Recognition (VSR) Decoder					
Layer	Input Size		Output Size		
Linear	128		28[b]		
Extra Information					

The system is implemented in Pytorch [33] and trained for N iterations using the Adam optimizer [34] with a learning rate of 0.0001, β_1=0.5 and β_2=0.9. The model that performs the best in terms of PESQ on the validation set is used for testing.
S=75 (sequence length). C=3 (image channels).
F=7 (consecutive video frames). W=96 (video frame width).
If the full face is used as input:
B=16 (batch size). H=128 (video frame height). d_1=3. d_2=5.
If only the mouth is used as input:
B=24 (batch size). H=64 (video frame height). d_1=2. d_2=4.
In the speaker dependent case, the dropout probability of each dropout layer is p_d=0.2. N=300000.
In the speaker independent case, p_d=0.5 for the video encoder and the GRU, and p_d=0.2 for the rest. N=185000.

[a]Eight is the number of the output audio frames corresponding to the video frame used as input (together with its context).
[b]The 28 CTC characters consist of the 26 letters of the English alphabet, one space character and one blank token.

Table 2: *Systems used in this study.*

	Input	
	Mouth	Face
w/o VSR Decoder	vid2voc-M	vid2voc-F
w/ VSR Decoder	vid2voc-M-VSR	vid2voc-F-VSR

Table 3: *Results for the speaker dependent and the speaker independent cases. Best performance (except WORLD) in bold.*

	Speaker Dependent			Speaker Independent		
Mean Scores	PESQ ↑	ESTOI ↑	WER ↓	PESQ ↑	ESTOI ↑	WER ↓
Approach in [15][a]	1.82	-	-	-	-	-
Approach in [17]	1.71	0.329	-	1.24	0.198	-
vid2voc-M	1.89	0.448	-	1.20	0.214	-
vid2voc-M-VSR	**1.90**	**0.455**	15.1%	1.23	**0.227**	**51.6%**
vid2voc-F	1.85	0.439	-	1.19	0.202	-
vid2voc-F-VSR	1.88	0.447	**14.4%**	**1.25**	0.210	69.3%
WORLD[b]	3.06	0.759	-	3.03	0.759	-

[a]Value taken from the experiments in [17].
[b]WORLD indicates the reconstruction retrieved from the vocoder features of the clean speech signals and it is a performance upper bound of our systems.

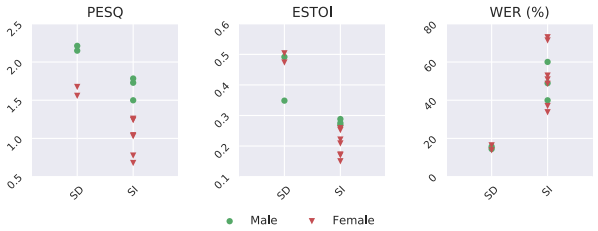

Figure 2: *Results of the vid2voc-M-VSR models for the speaker dependent (SD) and the speaker independent (SI) cases. Each marker indicates the mean score of a speaker.*

3. Results and Discussion

As shown in Table 2, four systems are trained based on the input (mouth or full face) and the presence of the VSR decoder (only speech synthesis or speech synthesis and VSR).

The systems are compared with the recently proposed GAN-based approach in [17]. As an additional baseline, we also report the PESQ score for [15], since this method, which makes use of bottleneck features extracted from auditory spectrograms, outperforms [17] in terms of estimated speech quality for the speaker dependent case.

3.1. Speaker Dependent Case

Table 3 (left part) shows the speaker dependent results. We observe that our models outperform the approach in [17] in terms of both PESQ and ESTOI by a considerable margin. Vougioukas et al. [17] mention that their system produces low-power hum artefacts that affect the performance. They tried to solve the issue by applying average filtering to the output of their network, experiencing a rise of the PESQ score from 1.71 to 1.80 (not shown in Table 3), comparable to [15], but still appreciably lower than the results we achieve. However, this filtering negatively affected the intelligibility of the produced speech signals, and was not used in the final system.

Among the systems we developed (cf. Table 2), we observe that including the VSR decoder in the pipeline is beneficial for the speech reconstruction task (see Table 3). Moreover, the use of the mouth as input not only is sufficient to synthesise speech, but it also allows to achieve higher estimated speech quality and intelligibility if compared to the models that use the whole face of the speaker as input. This might be explained by the fact that handling an input with a larger dimensionality is harder if we want to keep roughly the same deep architecture with a similar number of parameters. However, when the whole face is used as input, the WER is slightly lower, indicating that there might be a performance trade-off between VSR and speech reconstruction that should be further investigated in future work in relation with other multi-task learning techniques.

3.2. Speaker Independent Case

Regarding the speaker independent scenario (cf. right part of Table 3), we observe that the performance gap between the approach in [17] and our systems is not as large as for the speaker dependent case. Although our models appear to perform slightly better than [17] in terms of ESTOI, the PESQ scores are similar. This can be explained by the fact that some speech characteristics, e.g. F0, cannot be easily estimated for unseen speakers. Since it is reasonable to think that people having similar facial characteristics (e.g. due to gender, age etc.) have similar speech characteristics (cf. [37], where the face of a person was predicted from a speech signal), we expect that training a network with a dataset that includes more speakers might be beneficial: such a network can produce an average voice of speakers from the training set that share similar facial traits with an unseen talking face.

Among the systems we developed, the presence of the VSR decoder still gives an advantage for speech reconstruction. Unlike the speaker dependent case, the WER for the model that uses the whole face as input is higher than the system using only the mouth. This is due to the early stopping technique that we adopt, which tends to favour speech reconstruction over VSR, indicating again the trade-off between these two tasks.

Finally, Figure 2 shows the results for the vid2voc-M-VSR models by speaker. We can see that the spread of the scores is much higher for the speaker independent case in particular for WER. This is in line with the observations reported in [17], suggesting the different performance between the estimated speech of subjects whose facial traits substantially differ from the speakers in the training set and the others.

4. Conclusion

In this study, we reconstructed speech from silent videos using a deep model that estimates WORLD vocoder features. We tested our approach in both speaker dependent and speaker independent scenarios. In both cases, we were able to obtain speech signals with estimated speech quality and intelligibility generally higher if compared to a recently proposed GAN-based approach. In addition, we designed our system to simultaneously perform visual speech recognition by using a decoder that estimates CTC characters from a given video sequence.

Future work includes: (a) the adoption of self-paced multi-task learning techniques; (b) the improvement of the visual speech recognition performance, e.g. with a beam search decoding scheme; (c) the design of a system that can generalise well to unseen speakers in noncontrolled environments.

5. Acknowledgment

The authors would like to thank Konstantinos Vougioukas, Stavros Petridis, Pritish Chandna and Merlijn Blaauw.

This research is partially funded by: the William Demant Foundation; the TROMPA H2020 project (770376); the Spanish Ministry of Economy and Competitiveness under the María de Maeztu Units of Excellence Program (MDM-2015-0502) and the Social European Funds; the MICINN/FEDER UE project (PGC2018-098625-B-I00); the H2020-MSCA-RISE-2017 project (777826 NoMADS).

6. References

[1] A. Davis, M. Rubinstein, N. Wadhwa, G. J. Mysore, F. Durand, and W. T. Freeman, "The visual microphone: Passive recovery of sound from video," *ACM Transactions on Graphics*, vol. 33, no. 4, pp. 79–, 2014.

[2] A. Owens, P. Isola, J. McDermott, A. Torralba, E. H. Adelson, and W. T. Freeman, "Visually indicated sounds," in *Proc. of CVPR*, 2016.

[3] Y. Zhou, Z. Wang, C. Fang, T. Bui, and T. L. Berg, "Visual to sound: Generating natural sound for videos in the wild," in *Proc. of CVPR*, 2018.

[4] Y. M. Assael, B. Shillingford, S. Whiteson, and N. De Freitas, "Lipnet: End-to-end sentence-level lipreading," in *Proc. of GTC*, 2017.

[5] T. Stafylakis and G. Tzimiropoulos, "Combining residual networks with LSTMs for lipreading," in *Proc. of Interspeech*, 2017.

[6] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Lip reading sentences in the wild," in *Proc. of CVPR*, 2017.

[7] J. Sotelo, S. Mehri, K. Kumar, J. F. Santos, K. Kastner, A. Courville, and Y. Bengio, "Char2wav: End-to-end speech synthesis," in *Proc. of ICLR Workshop*, 2017.

[8] W. Ping, K. Peng, A. Gibiansky, S. O. Arik, A. Kannan, S. Narang, J. Raiman, and J. Miller, "Deep voice 3: 2000-speaker neural text-to-speech," in *Proc. of ICLR*, 2018.

[9] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan *et al.*, "Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions," in *Proc. of ICASSP*, 2018.

[10] T. Le Cornu and B. Milner, "Reconstructing intelligible audio speech from visual speech features," in *Proc. of Interspeech*, 2015.

[11] ——, "Generating intelligible audio speech from visual speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 9, pp. 1751–1761, 2017.

[12] H. Kawahara, I. Masuda-Katsuse, and A. De Cheveigne, "Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, no. 3-4, pp. 187–207, 1999.

[13] A. Ephrat and S. Peleg, "Vid2speech: Speech reconstruction from silent video," in *Proc. of ICASSP*, 2017.

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

[15] H. Akbari, H. Arora, L. Cao, and N. Mesgarani, "Lip2audspec: Speech reconstruction from silent lip movements video," in *Proc. of ICASSP*, 2018.

[16] T. Chi, P. Ru, and S. A. Shamma, "Multiresolution spectrotemporal analysis of complex sounds," *The Journal of the Acoustical Society of America*, vol. 118, no. 2, pp. 887–906, 2005.

[17] K. Vougioukas, P. Ma, S. Petridis, and M. Pantic, "Video-driven speech reconstruction using generative adversarial networks," in *Proc. of Interspeech*, 2019.

[18] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: A vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Transactions on Information and Systems*, vol. 99, no. 7, pp. 1877–1884, 2016.

[19] Y. Kumar, M. Aggarwal, P. Nawal, S. Satoh, R. R. Shah, and R. Zimmermann, "Harnessing AI for speech reconstruction using multi-view silent video feed," in *Proc. of ACM-MM*, 2018.

[20] 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 *Proc. of ISM*, 2018.

[21] Y. Kumar, R. Jain, K. M. Salik, R. R. Shah, Y. Yin, and R. Zimmermann, "Lipper: Synthesizing thy speech using multi-view lipreading," in *Proc. of AAAI*, 2019.

[22] S. Uttam, Y. Kumar, D. Sahrawat, M. Aggarwal, R. R. Shah, D. Mahata, and A. Stent, "Hush-hush speak: Speech reconstruction using silent videos," in *Proc. of Interspeech*, 2019.

[23] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proc. of ICML*. ACM, 2006.

[24] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, 2006.

[25] A. Camacho, "SWIPE: A sawtooth waveform inspired pitch estimator for speech and music," Ph.D. dissertation, University of Florida Gainesville, 2007.

[26] M. Morise, "D4C, a band-aperiodicity estimator for high-quality speech synthesis," *Speech Communication*, vol. 84, pp. 57–65, 2016.

[27] M. Blaauw and J. Bonada, "A neural parametric singing synthesizer modeling timbre and expression from natural songs," *Applied Sciences*, vol. 7, no. 12, p. 1313, 2017.

[28] P. Chandna, M. Blaauw, J. Bonada, and E. Gomez, "A vocoder based method for singing voice extraction," in *Proc. of ICASSP*, 2019.

[29] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2D & 3D face alignment problem? (and a dataset of 230,000 3D facial landmarks)," in *Proc. of ICCV*, 2017.

[30] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. of ICML*, 2015, pp. 448–456.

[31] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[32] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *Proc. of EMNLP*, 2014.

[33] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in NeurIPS*, 2019.

[34] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. of ICLR*, 2015.

[35] 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," in *Proc. of ICASSP*, 2001.

[36] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2009–2022, 2016.

[37] T.-H. Oh, T. Dekel, C. Kim, I. Mosseri, W. T. Freeman, M. Rubinstein, and W. Matusik, "Speech2face: Learning the face behind a voice," in *Proc. of CVPR*, 2019.