Deliverable D3.5: A Software Package for Audio Processing Assisted by Visual Information

Due Date: 01/06/2023

Main Author: Sharon Gannot (BIU)

Contributors:

Dissemination: Public Deliverable

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No. 871245.

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No. 871245.

DOCUMENT FACTSHEET

Deliverable	D3.5: A Software Package for Audio Processing Assisted by Visual Information
Responsible Partner	BIU
Work Package	WP3: Robust Audio-visual Perception of Humans
Task	T3.2 & T3.3
Version & Date	27/05/23
Dissemination	Public Deliverable

CONTRIBUTORS AND HISTORY

Version	**Editor**	**Date**	**Change Log**
1	BIU	27/05/23	First Draft
2	INRIA	xx/05/2023	Second Draft

APPROVALS

Authors/editors	Sharon Gannot (BIU)
Task Leader	BIU
WP Leader	BIU

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No. 871245.

Contents

This project has received funding from the European Union's Horizon 2020
Research and Innovation Programme under Grant Agreement No. 871245.

Executive Summary

Deliverable 3.5 reports on software packages incorporating the visual modality into the audio pipeline. We report on two achievements: 1) an instantaneous direction-of-arrival (DOA) estimation for a single active speaker, incorporated into the video tracker, and 2) voice generation from silent video.

Audio-only DOA estimation: We have already reported in D3.2 on both the visual and audio trackers (Task T3.1). We will shortly deploy under ROS and evaluate the convolutional neural network (CNN)-based multiple-speaker audio tracker [9]. In the meantime, relying on the accuracy of the visual tracker, we have implemented a simple audio-based localizer. By doing so, we can relate the audio and visual identities of the speakers, thus facilitating speaker diarisation.[1]

Voice generation from silent video: In the lip-to-speech task, we are given a soundless video of a person talking and are required to accurately and precisely generate the missing speech. Such a task may occur, e.g., when the speech signal is completely obfuscated due to background noises. This algorithm can serve for speaker extraction of a desired speaker in adverse conditions. In the near future, we will leverage this visual information to, hopefully, improve the separation and diarisation results that were reported in D3.3 and D3.4.[2]

[1] Code will be available at `https://gitlab.inria.fr/spring/wp3_av_perception/audio_gcc_doa/-/tree/main?ref_type=heads`
[2] Code will be available at `https://gitlab.inria.fr/spring/wp3_av_perception/lipvoicer`

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No. 871245.

1 Introduction

This deliverable is part of WP3 of the H2020 SPRING project. The objective of WP3 is "the robust extraction, from the raw auditory and visual data, of users' low-level characteristics, namely: position, speaking status and speech signal." Following this objective, WP3 has two main outcomes:

1. The Multi-Person Tracking module, jointly exploiting auditory and visual raw data to detect, localise and track multiple speakers (corresponds to T3.1).
2. The Diarisation and Separation and the Speech Recognition modules, extracting the desired speaker(s) from a speech dynamic mixture and recognising the speech utterances from the separated sources, for a static T3.2 and a moving T3.3 robot.

In this context, the current deliverable D3.5 is complementary to D3.1, D3.2, D3.3, and D3.4. Here, we present two software tools:

1. a simple audio localisation algorithm to match the visual localisation readings. Note that the full integration of this tool is not yet accomplished, but all components are available.
2. an audio generation tool from a silent video that is capable of generating high-quality audio with a reasonably low word error rate (WER). The tool can be used in acoustically adverse conditions and will be later extended to a full-fledged audio-visual separation algorithm. This tool is not yet integrated under ROS.

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No. 871245.

2 Audio-Visual Localisation and Tracking

In D3.1, a state-of-the-art multi-person visual tracker known as fair multi-objective tracking (FairMOT) [22] was introduced. In D3.2, some of the original FairMOT models based on the residual neural network (ResNet34) [11] architecture have been compared with newly trained models that are better adapted to the non-rectilinear perspective characteristics of the fisheye camera. In D3.2 we reported on the integration of the visual tracker with a simple audio tracker known as Open embeddeD Audition System (ODAS) [7], specifically implemented for robot audition tasks. In this report, we will discuss a simple localization algorithm based on the GCC-PHAT. The audio tracking module will be substituted in the near future with a CNN-based system developed by BIU [9].

2.1 Audio-Visual Fusion based on ODAS

ODAS implements a sound source localisation algorithm, which combines the classical steered response power with phase transform (SRP-PHAT) method, enhanced by hierarchical search with directivity model and automatic calibration (HSDA), followed by a tracking algorithm supported by a Kalman filter. The package can be used out-of-the-box for ARI's microphone array. Nevertheless, it required some software development to share the hardware (specifically, to be able to use the microphones simultaneously by other sound processing modules), and ROS integration. Tracked sound sources are given by ODAS as unit vectors pointing to them (i.e. direction-only), in the microphone frame. Since we do not actually know the distance of the sound source but only its direction, we have to set an arbitrary distance (2 or 3 m, for example) in order to obtain an approximated 3D position of the sound source in the microphone frame.

2.2 Audio-Visual Fusion based on GCC-PHAT and VAD

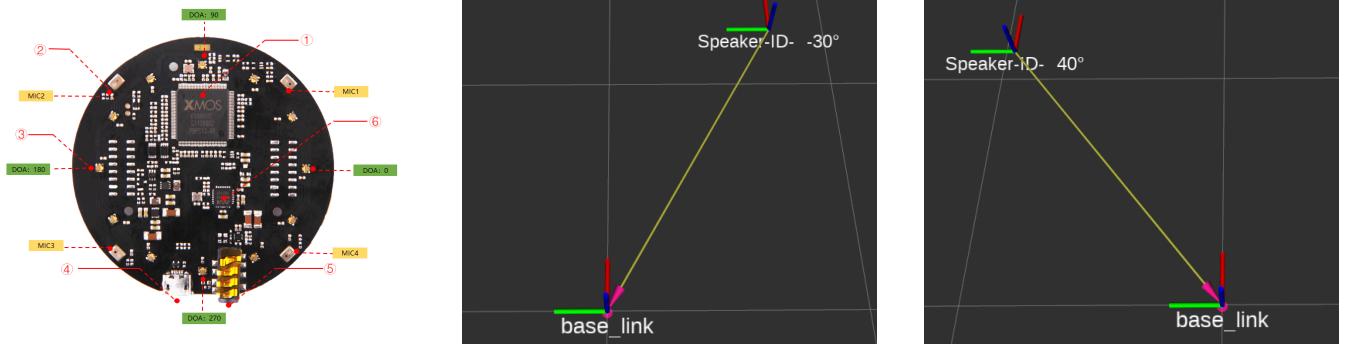
The GCC-PHAT is a classical time difference of arrival (TDOA) estimator, based on the maximization of the cross-correlation between the microphone signals. To alleviate the influence of the reverberation, the cross-spectrum between the microphone signals is first calculated. Then, the cross-spectrum is normalized by its absolute value to obtain the corresponding phase. Finally, the normalized cross-spectrum is back-transformed to the time domain. The peak of the resulting generalized cross-correlation (GCC) corresponds to the TDOA between the signals. We applied parabolic interpolation to the GCC series to obtain a higher resolution peak-finding.

The ReSpeaker microphone array installation in ARI is perpendicular to the floor. We will therefore only use the two upper microphones rather than the entire 4-microphone array (see Fig. 2.1a) and estimate the DOA in the azimuth plane. The distance between microphones 1 and 2 is 45.7mm. In the near-field, the relation between the TDOA and the DOA depends on the distance between the source and the microphone. We have therefore prepared a lookup table with the proper correspondence for several distances and will select the relevant table based on the depth information. A significant drawback of the GCC-PHAT is its inability to produce meaningful results in the multiple-speaker case. Although several cures to this problem are available in the literature, we decided to take a different path. In any case, in the near future, we plan to substitute this module with the CNN-based algorithm [9], which can track multiple concurrent speakers.

The audio-based DOA readings will be fused with the visual-based DOA readings to form a comprehensive audio-visual id of the speakers that are engaged with ARI. Moreover, the single-microphone separation algorithm (see D3.4) provides two separated soundtracks together with their activity patterns. These VAD signals and the DOA readings will be used to properly separate, diarise, and identify the speakers. Note that when the activities of the sources overlap, they will be separated by the algorithm, but the DOA readings will be unreliable and therefore discarded.

Sample DOA estimates for a single-speakers scenario are depicted in Figs. 2.1b, 2.1c.[1]

[1] Animation of a source moving on a semi-circle of radius 1 m can be found in `https://www.dropbox.com/s/ry2tn7ea85t5rof/gcc_phat_lookuptable_3_pos.mp4?dl=0`



(a) Respeaker array. The distance between microphones 1 and 2 is 45.7 mm.

(b) DOA estimation for a source at $-30°$, 1 m from ARI.

(c) DOA estimation for a source at $40°$, 1 m from ARI.

Figure 2.1: The ReSpeaker microphone array and sample DOA estimates.

This project has received funding from the European Union's Horizon 2020
Research and Innovation Programme under Grant Agreement No. 871245.

3 LipVoicer

Lip-to-speech involves generating a natural-sounding speech synchronized with a soundless video of a person talking. Despite recent advances, current methods still cannot produce high-quality speech with high levels of intelligibility for challenging and realistic datasets. In this work, we present *LipVoicer*, a novel method that generates high-quality speech, even for in-the-wild and rich datasets, by incorporating the text modality. Given a silent video, we first predict the spoken text using a pre-trained lip-reading network. We then condition a diffusion model on the video and use the extracted text through a classifier-guidance mechanism where a pre-trained automatic speech recognition (ASR) serves as the classifier. LipVoicer outperforms multiple lip-to-speech baselines on LRS2 and LRS3, which are in-the-wild datasets with hundreds of unique speakers in their test set and an unrestricted vocabulary. Moreover, our experiments show that the inclusion of the text modality plays a major role in the intelligibility of the produced speech, readily perceptible while listening, and is empirically reflected in the substantial reduction of the WER metric.

3.1 Introduction

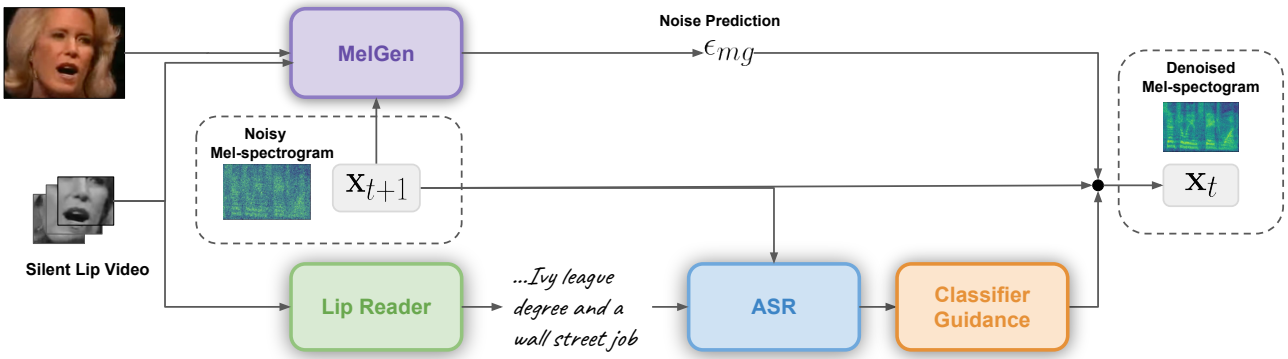
In the lip-to-speech task, we are given a soundless video of a person talking and are required to accurately and precisely generate the missing speech. Such a task may occur, e.g., when the speech signal is completely obfuscated due to background noises. This task poses a significant challenge as it requires the generated speech to satisfy multiple criteria. This includes intelligibility, synchronization with lip motion, naturalness, and alignment with the speaker's characteristics such as age, gender, accent, and more. Another major hurdle for lip-to-speech techniques is the ambiguities inherent in lip motion, as several phonemes can be attributed to the same lip movement sequence. Resolving these ambiguities requires the analysis of lip motion in a broader context within the video.

Generating speech from a silent video has seen significant progress in recent years, partly due to advancements made in deep generative models. Specifically in applications such as text-to-speech and mel-spectogram-to-audio (neural vocoder) [18, 14]. Despite these advancements, many lip-to-speech methods produce satisfying results only when applied to datasets with a limited number of speakers, and constrained vocabularies, like GRID [4] and TCD-TIMIT [10]. Therefore, speech generation for silent videos in-the-wild still lags behind. We found that these methods struggle to reliably generate natural speech with a high degree of intelligibility on more challenging datasets like LRS2 [1] and LRS3 [2].

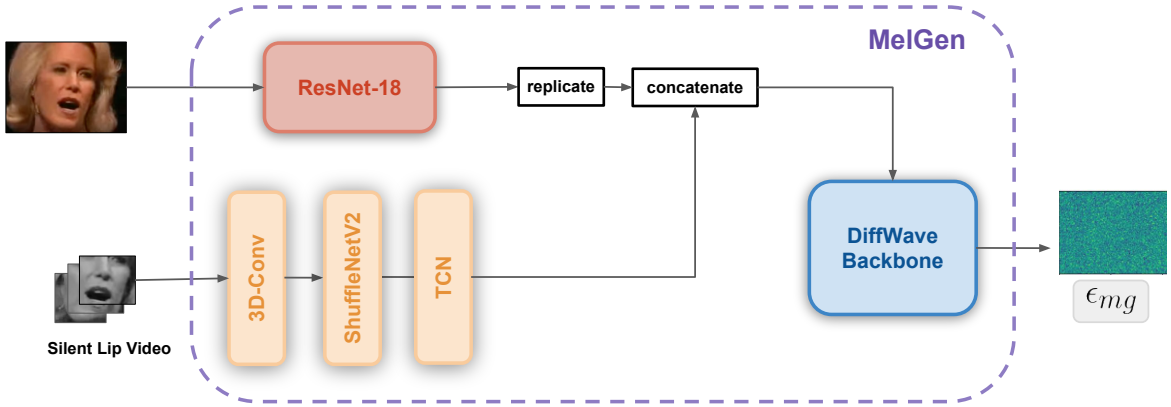
3.2 LipVoicer: Methodology

We introduce *LipVoicer*, a novel approach for producing high-quality speech for silent videos. The first and crucial part of LipVoicer is leveraging a lip-reading model at inference time, for extracting the transcription of the speech we wish to generate. Next, we train a diffusion model, conditioned on the video, to generate mel-spectrograms. This generation process is guided by both the video and the predicted transcription, as illustrated in Fig. 3.1a. Consequently, our model successfully intertwines the information conveyed by textual modality with the dynamics and characteristics of the speaker, captured by the diffusion model. Incorporating the inferred text has an additional benefit, as it allows LipVoicer to alleviate the lip motion ambiguity to a great extent. Finally, we use the DiffWave [18] neural vocoder to generate the raw audio. A diagram of our approach is depicted in Fig. 3.1.

Previous methods often use text to guide the generation process at train time. We, however, utilize it at inference time. The text, transcribed using a lip-reader, allows us to utilize guidance [6, 12] which ensures that the text of the generated audio corresponds to the target text. Guidance, with or without a classifier, is an important part of diffusion models and a key feature in recent advancements in text-to-image [20, 21] and text-to-speech [14, 13].



(a)



(b)

Figure 3.1: An illustration of LipVoicer, a dual-stage framework for lip-to-speech. (a) To generate the speech from a given silent video, a pre-trained lip-reader provides additional guidance by predicting the text from the video. An ASR steers MelGen, which generates the mel-spectrogram, in the direction of the extracted text using classifier guidance, such that the generated mel-spectrogram reflects the spoken text. (b) MelGen, our diffusion denoising model that generates mel-spectrograms conditioned on a face image and a mouth region video extracted from the full-face video using classifier-free guidance.

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No. 871245.

3.3 LipVoicer: details

This section details the LipVoicer scheme for lip-to-speech generation. Given a silent talking-face video, LipVoicer generates a mel-spectrogram that corresponds to a high likelihood underlying speech signal. The proposed method comprises three main components:

1. A mel-spectrogram generator (MelGen) that learns to generate a mel-spectrogram image from the video.
2. A pre-trained lip-reading network that infers the most likely text from the silent video.
3. An ASR system that anchors the mel-spectrogram recovered by MelGen to the text predicted by the lip-reader.

At first, we train MelGen, a conditional denoising diffusion probabilistic models (DDPM) model trained to generate a mel-spectrogram waveform conditioned on the video. Similar to diffusion-based frameworks in text-to-speech, e.g. [13], we use a DiffWave [18] residual backbone for MelGen. When considering the representation for the video, we wish it to encapsulate all the required information to generate the mel-spectrogram, i.e., the content (spoken words) and dynamics (accent, intonation) of the underlying speech, the timing of each part of speech, as well as the identity of the speaker, e.g. gender, age, etc. However, we wish to remove all irrelevant information to help train and remove unnecessary computational costs. To this end, the video is pre-processed by creating a cropped mouth region video and randomly choosing a single full-face image, corresponding to the content and dynamics and to the voice characteristics, respectively. The mouth cropping was implemented according to the procedure in [19].

Next, a DDPM is trained to generate the mel-spectrogram with and without the conditioning on the video embedding following the classifier-free mechanism [12]. In order to make MelGen robust to scenarios characterized by an unconstrained vocabulary, we use the text modality as an additional source of guidance. In general, syllables uttered in a silent talking-face video can be ambiguous, and may consequently lead to an incoherent reconstructed speech. It can therefore be beneficial to harness recent advances in lip-reading and ground the generated mel-spectrogram to the text predicted by a pretrained lip-reading network.

To circumvent the challenge of aligning text with video content, we employ text guidance by harnessing the classifier guidance approach [6], similarly to [14], by using a powerful ASR model. Note that we use an audio ASR rather than audio-video ASR, to encourage the model to focus on audio generation. Classifier guidance allows us to train MelGen that is solely conditioned on the video and use a pre-trained ASR to guide that the generated speech is matched. As a result, the ASR is responsible for the precise words in the estimated speech, and MelGen provides the voice characteristics, synchronization between the video and the generated audio, and the continuity of the speech, see Fig. 3.1 for an illustration of our system.

One additional advantage of this approach is the modularity and ease of substituting both the lip-to-text and the ASR modules. If one wishes to substitute these models with improved versions in the future, the process can be accomplished effortlessly without requiring any re-training. Finally, DiffWave [18] is used as the vocoder that transforms the reconstructed mel-spectrogram to a time-domain speech signal.

3.4 Evaluation

We evaluate our LipVoicer model on the challenging LRS2 and LRS3 datasets. These datasets are "in-the-wild" videos, with hundreds of unique speakers and with an open vocabulary. We show that our proposed design leads to the best results on these datasets in both human evaluations as well as WER of an ASR system.

To the best of our knowledge, LipVoicer is the first method to use text inferred by lip-reading to enhance lip-to-speech synthesis. The inclusion of the text modality in inference removes the uncertainty of deciphering which of the possible candidate phonemes correspond to the lip motion. Additionally, it helps the diffusion model to focus on creating naturally synced speech. The speech generated by LipVoicer is intelligible, well synchronized to the video, and sounds natural. Finally, LipVoicer achieves state-of-the-art results for highly challenging in-the-wild datasets.

We evaluated our method with WER and synchronization metrics. For a fair comparison, we evaluate the WER using the ASR model from [8] that is distinct from the one we use for guidance. For synchronization, we use the pre-trained SyncNet [3] model to evaluate the LSE-C and LSE-D metrics. As a result of the disparity in image shapes between LRS2 and the expected input shape SyncNet was trained on, we refrain from providing synchronization metrics for LRS2. Despite our efforts to mitigate this challenge through image padding to align with SyncNet's expected input shape, this approach resulted in significant artifacts that adversely impacted SyncNet's performance across all methods. For SVTS, we report WER and synchronization metrics only for LRS3, since the authors did not open-source their code and only released the generated test files for LRS3. From the WER scores, it is clear that our method significantly improves over competing baselines. It is also clear that this improvement is solely due to the ASR guidance, as without it the WER plunges from 24.1% to 84.9% on LRS3. In addition to generating high-quality content, LipVoicer demonstrates

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No. 871245.

commendable synchronization scores, ensuring that the generated speech aligns seamlessly with the accompanying video. Interestingly, the incorporation of text classifier guidance enhances the intelligibility performance while leading to a slight degradation in the LSE-C synchronization metric. We postulate that this observation may occur in cases where the predicted text is significantly different from the ground-truth text.

	LRS2	LRS3		
	WER	WER	LSE-C ↑	LSE-D ↓
GROUND TRUTH	6.1%	2.5%	6.880	7.638
LIP2SPEECH [16]	58.2%	61.7%	5.231	8.832
SVTS [5]	-	75.6%	6.018	8.290
VCA-GAN [15]	95.1%	87.5%	5.255	8.913
LIPVOICER W/O ASR (OURS)	81.2%	84.9%	**6.318**	8.310
LIPVOICER (OURS)	**33.9%**	**24.1%**	6.239	**8.266**

Table 3.1: Comparison of LRS2 & LRS3 word error rate (WER) and Lip-Speech Synchronization.

We also created a demo showcasing LipVoicer's superiority in producing natural, synchronized, and intelligible speech, providing additional evidence of its effectiveness.[1]

3.5 Conclusions

We presented LipVoicer, a novel method that shows promising results in generating high-quality speech from silent videos. LipVoicer achieves this by utilizing text inferred from a lip-reading model to guide the generation of natural audio. We train and test LipVoicer on multiple challenging datasets comprised of in-the-wild videos. We empirically show that text guidance is crucial to create intelligible speech, as measured by the WER. Furthermore, we show through human evaluation that LipVoicer faithfully recovers the ground truth speech and surpasses recent baselines in intelligibility, naturalness, quality, and synchronization. The impressive achievements of LipVoicer in lip-to-speech synthesis not only advance the current state-of-the-art but also paves the way for intriguing future research directions in this domain.

[1] https://lipvoicer.github.io

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No. 871245.

4 Discussion

This document reports the progress in joint audio-visual processing. We first discussed the audio tracker based on GCC-PHAT [17] that adds the audio directional information to the already established visual directional information. Together with the activity patterns of the sources provided by the speaker separation module, a full diarisation-separation module can be implemented. In the near future, the simple GCC-PHAT algorithm will be substituted with a CNN-based algorithm [9]. A second module, reported in this document, is a lip-to-voice algorithm that generates an audio signal from a silent video. This algorithm can serve as a speech enhancement module in extreme acoustic conditions, but more importantly, it will serve as a platform for a joint audio-visual speaker separation.

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No. 871245.

Bibliography

[1] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman. Deep audio-visual speech recognition. In *arXiv:1809.02108*, 2018.

[2] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. LRS3-TED: a large-scale dataset for visual speech recognition. *arXiv preprint arXiv:1809.00496*, 2018.

[3] J. S. Chung and A. Zisserman. Out of time: automated lip sync in the wild. In *Workshop on Multi-view Lip-reading, Asian Conference on Computer Vision (ACCV)*, 2017.

[4] Martin Cooke, Jon Barker, Stuart P. Cunningham, and Xu Shao. An audio-visual corpus for speech perception and automatic speech recognition. *The Journal of the Acoustical Society of America*, 120:2421–2424, 2006.

[5] Rodrigo Schoburg Carrillo de Mira, Alexandros Haliassos, Stavros Petridis, Björn W. Schuller, and Maja Pantic. SVTS: scalable video-to-speech synthesis. In *Interspeech*, 2022.

[6] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[7] François Grondin, Dominic Létourneau, Cédric Godin, Jean-Samuel Lauzon, Jonathan Vincent, Simon Michaud, Samuel Faucher, and François Michaud. Odas: Open embedded audition system. *arXiv preprint arXiv:2103.03954*, 2021.

[8] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*, 2020.

[9] Hodaya Hammer, Shlomo E Chazan, Jacob Goldberger, and Sharon Gannot. Dynamically localizing multiple speakers based on the time-frequency domain. *EURASIP Journal on Audio, Speech, and Music Processing*, 2021(1):1–10, 2021.

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

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE conference on computer vision and pattern recognition (CVPR)*, pages 770–778, 2016.

[12] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS Workshop on Deep Generative Models and Downstream Applications*, 2021.

[13] Myeonghun Jeong, Hyeongju Kim, Sung Jun Cheon, Byoung Jin Choi, and Nam Soo Kim. Diff-TTS: A denoising diffusion model for text-to-speech. *INTERSPEECH*, 2021.

[14] Heeseung Kim, Sungwon Kim, and Sungroh Yoon. Guided-TTS: A diffusion model for text-to-speech via classifier guidance. In *The 39th International Conference on Machine Learning (ICML)*, volume 162, pages 11119–11133, July 2022.

[15] Minsu Kim, Joanna Hong, and Yong Man Ro. Lip to speech synthesis with visual context attentional GAN. *Advances in Neural Information Processing Systems*, 34:2758–2770, 2021.

[16] Minsu Kim, Joanna Hong, and Yong Man Ro. Lip-to-speech synthesis in the wild with multi-task learning. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

[17] Charles Knapp and Glifford Carter. The generalized correlation method for estimation of time delay. *IEEE transactions on acoustics, speech, and signal processing*, 24(4):320–327, 1976.

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No. 871245.

[18] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. In *International Conference on Learning Representations (ICLR)*, 2021.

[19] Pingchuan Ma, Stavros Petridis, and Maja Pantic. Visual speech recognition for multiple languages in the wild. *Nature Machine Intelligence*, 4:930–939, 2022.

[20] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv:2204.06125*, 2022.

[21] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022.

[22] Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenjun Zeng, and Wenyu Liu. Fairmot: On the fairness of detection and re-identification in multiple object tracking. *International Journal of Computer Vision*, 129:3069–3087, 2021.