SilentSpeak: A Proposed Multilingual Deep Learning Framework for Lip Reading in English, Hindi, and Marathi to Assist Individuals with Speech and Hearing Impairments

Priyanka Pramod Kurde, Mahek Shaukat Patel, Sakshi Ganesh Muley and Mrs. Aarti Patil

D. Y. Patil Institute of Master of Computer Applications and Management, Akurdi, Pune.

Abstract— Lip reading, or visual speech recognition, can recover spoken content when audio is missing, unusable or inaudible, which is valuable for people with hearing or speech impairments. Most published research targets English, and the authors' literature search found no publicly documented word-level lip-reading dataset or system for Hindi or Marathi. This paper proposes SilentSpeak, a conceptual framework for word-level lip reading in English, Hindi and Marathi that is designed to run on a standard webcam without an internet connection. The proposed pipeline locates the lips with MediaPipe Face Mesh, describes each frame with a small convolutional network, and models the mouth movement over time with a bidirectional long short-term memory network. An English base model trained on the GRID corpus is to be adapted to Hindi and Marathi through transfer learning, by freezing the convolutional layers and retraining the recurrent and output layers on small pilot vocabularies. The framework also proposes confidence scoring with a top-three fallback, an emergency-keyword alert and offline speech output. The paper reviews the relevant literature, identifies the research gap, specifies the methodology and system architecture, and outlines a plan for evaluation. This is a proposed research direction, and no experimental results are claimed in this paper.

Keywords: Assistive technology, Hindi, Lip reading, Marathi, Transfer learning, Visual speech recognition

1. INTRODUCTION

Speech is usually treated as sound, but it can also be seen. When audio is missing, unusable or inaudible, because of hearing loss, a speech impairment, a noisy place or the need to stay silent, the movements of the lips can still carry part of the message. The World Health Organization reports that hearing loss affects more than 1.5 billion people today and is projected to affect nearly 2.5 billion by 2050 [1]. Clinical writing from India describes the burden of deafness and the difficulty of rehabilitation in the country [2], and the WHO report counts speech reading among the options that matter for many deaf people [1]. Vision is a useful ally: Sumby and Pollack showed that seeing a speaker's face improves speech understanding in noise [3].

Automated lip reading, also called visual speech recognition, maps a sequence of mouth images to words. Deep learning has moved the field quickly. LipNet decoded GRID sentences end to end and reported 95.2% sentence-level accuracy, a figure that the same paper sets against 52.3% for hearing-impaired human lipreaders [4]. The LRW and LRS datasets took the task to television broadcasts [5], [6], and later work raised word accuracy with residual networks and temporal convolutional networks [7], [8]. A recent survey compares the databases, feature extractors and classifiers used in automated lip reading up to early 2021 [9].

Most of the work reviewed here was built and tested on English or Mandarin material. India is multilingual, and tools that assume English leave many potential users behind. In our search of IEEE Xplore, Google Scholar, Springer and Kaggle we found no publicly documented word-level lip-reading dataset or system for Hindi or Marathi that is comparable with the English benchmarks. This is a finding of our search and not a claim that none exists.

This paper proposes SilentSpeak, a conceptual framework for word-level lip reading in English, Hindi and Marathi that is designed for a standard webcam and for use without an internet connection. Its main elements are (i) a compact pipeline that combines MediaPipe-based lip extraction with a convolutional front end and a bidirectional LSTM; (ii) transfer learning from an English base model to small Hindi and Marathi vocabularies; (iii) confidence scoring, an emergency-keyword alert and offline speech output; and (iv) a plan for evaluating the framework. SilentSpeak has not yet been implemented or tested. **This is a proposed research direction, and no experimental results are claimed in this paper.**

The rest of the paper is organized as follows. Section 2 reviews the literature, Section 3 identifies the research gap, Section 4 describes the proposed methodology, Section 5 presents the proposed system, Section 6 discusses what the literature allows us to say and how the framework will be evaluated, and Sections 7 and 8 give the conclusion and future scope.

2. LITERATURE REVIEW

2.1 Visual speech and its ambiguity

Seeing the speaker helps listeners understand speech in noise [3], so visual speech carries real information. It is also ambiguous, because several sounds share the same visible mouth shape. The authors of LipNet note that almost all of its few erroneous predictions occur within such groups of look-alike sounds (visemes),

since context is sometimes insufficient for disambiguation [4]. A closed vocabulary, as proposed here, reduces this ambiguity but does not remove it.

2.2 Deep lip reading

LipNet combines spatiotemporal convolutions, recurrent layers and a connectionist temporal classification (CTC) loss and was evaluated on the constrained-grammar GRID corpus [4]. Chung and Zisserman assembled word-level data from television (LRW) [5], and Chung et al. proposed the WLAS model, curriculum learning and the LRS sentence dataset [6]. Stafylakis and Tzimiropoulos combined a residual network with an LSTM and reported 83.0% word accuracy on LRW [7]. Martinez et al. replaced the recurrent back end with a temporal convolutional network and reported absolute gains of 1.2% on LRW and 3.2% on the Mandarin LRW-1000 dataset [8]. Afouras et al. described transformer-based models for lip reading and audio-visual recognition [10], and Ma et al. evaluated one approach on English, Mandarin, Spanish, Italian, French and Portuguese [11]. A broader survey is given in [9]. Table 1 summarizes these studies.

Table 1. Representative published lip-reading studies and the findings they report.

Ref.	Task and method	Reported finding
[4]	GRID sentences; spatio-temporal CNN, recurrent layers, CTC	95.2% sentence-level accuracy; 52.3% for hearing-impaired lipreaders
[5]	LRW word-level data from television	Introduced a large word-level benchmark
[6]	LRS sentences; WLAS with curriculum learning	Outperformed earlier benchmarks and a professional lip reader on BBC clips
[7]	LRW; residual network and LSTM	83.0% word accuracy on LRW
[8]	LRW and LRW-1000; temporal convolutional network	+1.2% (LRW) and +3.2% (LRW-1000) over the previous best
[10]	Sentence-level lip reading; transformer models with CTC and sequence-to-sequence losses	Also extended to audio-visual recognition
[11]	Six languages, one approach	No Indian language included
[9]	Survey of databases, features and classifiers	Covers work up to early 2021

The accuracies in Table 1 refer to different tasks, vocabularies and recording conditions. They should not be compared with one another, and they should not be read as the expected performance of SilentSpeak.

2.3 Transfer learning and regularization

Transfer learning improves learning in a target domain by reusing knowledge from a source domain [12]. Yosinski et al. showed that the first layers of deep networks tend to be general while the last layers are specific to the task, and that features can be kept frozen when the target dataset is small, because fine-tuning many parameters on little data risks over-fitting [13]. Dropout [14] and data augmentation [15] are common ways to limit over-fitting. The LSTM [16] models sequences, and the bidirectional recurrent network reads a sequence in both time directions [17].

2.4 Tools and data

MediaPipe is a framework for building perception pipelines [18], and its Face Mesh model estimates 468 three-dimensional face landmarks from a single camera image in real time [19]. The GRID corpus contains recordings of 34 talkers, each saying 1,000 sentences of a fixed form, and is available for research [20]. TensorFlow is a widely used platform for building and training neural networks [21].

3. RESEARCH GAP

Four gaps follow from the review and are summarized in Table 2. The first two are findings of our literature and dataset search and are worded accordingly.

Table 2. Research gaps and the response proposed in SilentSpeak.

Research gap	Proposed response
G1. No word-level lip-reading system or dataset for Hindi or Marathi was found in our search	A trilingual English–Hindi–Marathi pipeline
G2. Benchmark data are English or Mandarin and were recorded in controlled or broadcast conditions; evaluation on Indian-language speakers with webcams is unexplored	Small pilot vocabularies recorded with a standard webcam
G3. Urgent-word handling, visible confidence and offline operation are not a focus of the reviewed studies	Emergency-keyword alert, confidence score with top-three fallback, offline speech
G4. Compact, locally running designs are rarely discussed	A small CNN–BiLSTM intended for a laptop

The third gap deserves a note on confidence. Guo et al. show that modern neural networks are often poorly calibrated [22], so a softmax score cannot be shown to a user as an exact probability without care. The framework therefore treats confidence as a practical indicator, to be checked during evaluation.

4. METHODOLOGY

4.1 Proposed research design

We propose an applied, experimental design. A working prototype would be built, trained, evaluated and compared with a simpler baseline. The work is organized in seven phases: literature survey, identification of the research gap, problem formulation, methodology, data collection, implementation with analysis, and reporting. This paper covers the first four. The remaining phases have not been carried out and are planned as future work.

4.2 Proposed data sources

For English, the GRID corpus [20] would serve as secondary data. Its word-level alignments allow each sentence video to be cut into single-word clips, and the English base model would be trained on a subset of its vocabulary.

For Hindi and Marathi, we propose two pilot vocabularies of about 20 words each, chosen for everyday assistance: basic needs, responses and courtesies, a few medical words, simple actions and directions, and emergency words (Table 3). Native-speaker volunteers of both genders would repeat each word several times in front of a standard 1080p webcam at 30 frames per second, under more than one lighting condition. Participation would be voluntary, with written informed consent. Volunteers would be identified by code only, and the recordings would be used for academic research. Each clip would be labeled with the word, language, volunteer code, lighting condition and repetition number. A speaker-independent split into training, validation and test sets is planned, so that no volunteer appears in more than one set, and an ethics review of the plan should precede recording.

Table 3. Proposed pilot vocabularies (Devanagari with transliteration).

Category	Hindi	Marathi
Basic needs	पानी (paani), खाना (khana)	पाणी (paani), जेवण (jevan)
Responses and courtesy	हाँ (haan), नहीं (nahi), शुक्रिया (shukriya), नमस्कार (namaskar)	हो (ho), नाही (nahi), धन्यवाद (dhanyavaad), नमस्कार (namaskar)
Medical	दर्द (dard), डॉक्टर (doctor), थकान (thakaan)	दुखणे (dukhne), डॉक्टर (doctor)
Emergency	मदद (madad), मुझे बचाओ (mujhe bachao)	मदत करा (madad kara), वाचवा (vaachva), मला वाचवा (mala vachva)
Actions and directions	आओ (aao), जाओ (jao), रुको (ruko), पीछे (peechhe)	या (ya), थांबा (thamba), बघ (bagh), पुढे (pudhe)
Comfort and home	ठंडा (thanda), गरम (garam), आराम (aaraam), घर (ghar), भाई (bhai)	थंड (thaand), उष्ण (ushna), विश्रांती (vishraanti), घर (ghar), भाऊ (bhau)

4.3 Proposed preprocessing

Each frame would be passed through MediaPipe Face Mesh [19]. The landmarks on the outer and inner lip contours define a bounding box, to which a small margin is added. The crop would be resized to 64×64 pixels, converted to grayscale and scaled to [0, 1]. Frames would be sampled to a fixed sequence of 15 frames, about 0.5 s at 30 fps, so that clips from 25 fps GRID video and 30 fps webcam video have the same shape. Cleaning steps would remove corrupt or duplicate files, flag clips in which the face is lost, and verify every label. Horizontal flipping, brightness jitter and mild Gaussian noise would be applied to the training set only [15].

5. PROPOSED SYSTEM

5.1 Architecture overview

The proposed system works in seven stages (Fig. 1): webcam capture, face-landmark detection, lip extraction, preprocessing, word classification, confidence and language handling, and text output with offline speech. An emergency-keyword check sits between the last two stages.

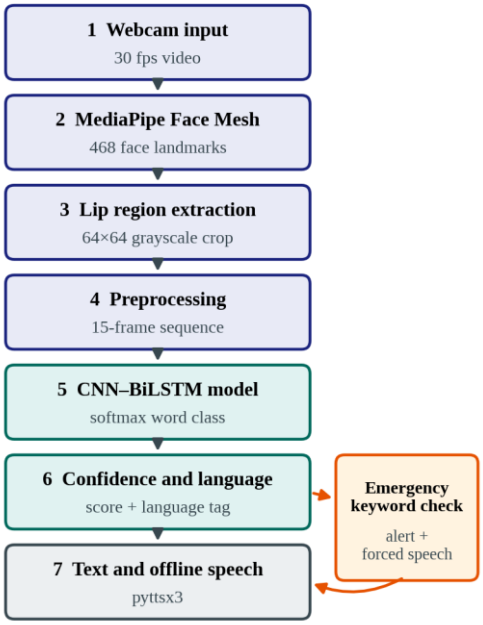


Fig. 1. Proposed seven-stage SilentSpeak pipeline with the emergency-keyword check.

5.2 CNN–BiLSTM model

Three convolutional blocks (32, 64 and 128 filters, 3×3, ReLU) would turn each frame into an 8×8×128 feature map, which is flattened to 8,192 values. The 15 frame vectors would then enter a bidirectional LSTM [16], [17] with 256 units per direction, followed by dropout with a rate of 0.5 [14] and a softmax layer with one output per word. Fig. 2 and Table 4 give the proposed configuration.

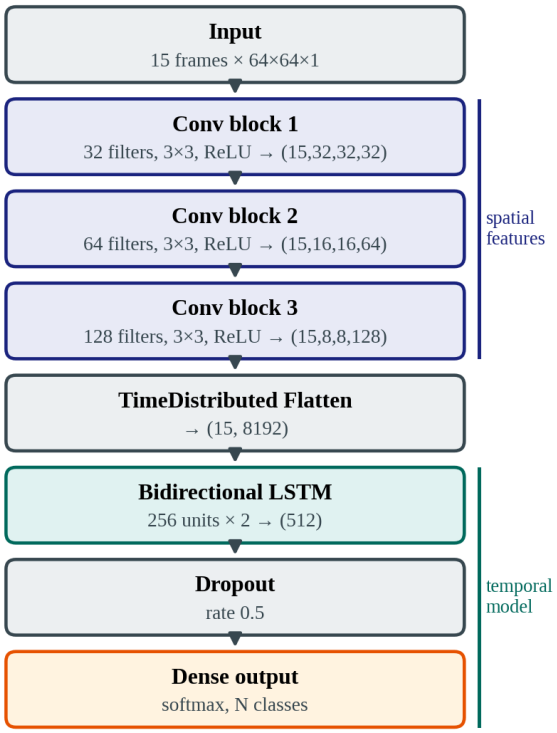


Fig. 2. Proposed CNN–BiLSTM architecture with tensor shapes.

Table 4. Proposed layer configuration of the CNN–BiLSTM model.

Layer	Configuration	Output shape
Input	15 frames × 64×64×1	(15, 64, 64, 1)
Conv block 1	32 filters, 3×3, ReLU	(15, 32, 32, 32)
Conv block 2	64 filters, 3×3, ReLU	(15, 16, 16, 64)
Conv block 3	128 filters, 3×3, ReLU	(15, 8, 8, 128)
TimeDistributed Flatten	—	(15, 8192)
Bidirectional LSTM	256 units × 2	(512)
Dropout	Rate 0.5	(512)
Dense output	Softmax, N classes	(N)

5.3 Transfer learning

The English base model would be trained first on GRID word clips. For Hindi and Marathi, a new model would be initialized from the English weights, the convolutional blocks would be frozen, and the output layer would be replaced by a layer with one output per regional word (Fig. 3). Only the recurrent and output layers would be fine-tuned on the regional data. Freezing is motivated by the observation that early layers are general [13] and by the small size of the planned regional datasets. The assumption that basic mouth shapes are shared across the three languages is plausible but untested, and the evaluation plan is designed to test it.

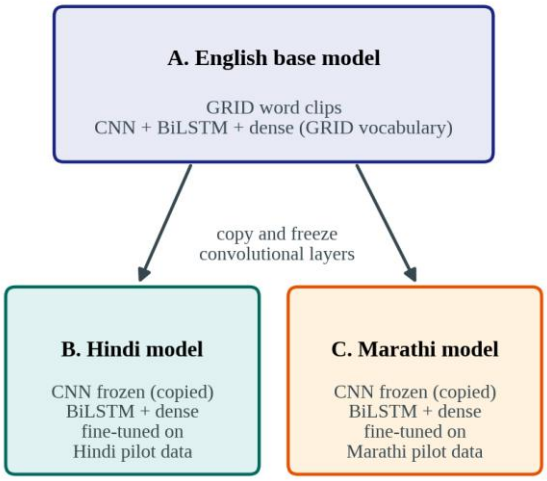


Fig. 3. Proposed transfer learning scheme from the English base model to Hindi and Marathi.

5.4 Confidence, emergency detection and speech

The top softmax probability would be shown as the confidence of each prediction. Below a threshold, proposed as 60% and to be tuned on validation data, the three most likely words would be shown instead of a single answer. Every predicted word would be compared with an emergency dictionary (Table 5). A match would raise a flashing red border and force speech output. Speech would be produced offline with the pyttsx3 library, which relies on the speech engine installed on the computer, so Hindi and Marathi speech depends on suitable voices being available.

Table 5. Proposed emergency keyword dictionary.

English	Hindi	Marathi
Help	मदद (madad)	मदत करा (madad kara)
Fire	आग (aag)	आग (aag)
Pain	दर्द (dard)	दुखणे (dukhne)
Doctor	डॉक्टर (doctor)	डॉक्टर (doctor)
Save me	मुझे बचाओ (mujhe bachao)	मला वाचवा (mala vachva)

A keyword can be flagged only if it belongs to the vocabulary of the active model. In the current design, four of the five Hindi entries and four of the five Marathi entries do, whereas fire is not in either 20-word vocabulary, and the English entries are not in the GRID vocabulary. These words are held in the dictionary for a later extension of the vocabularies.

5.5 Proposed implementation environment

The system is proposed to be written in Python with TensorFlow/Keras [21] and MediaPipe [18], trained on a cloud GPU runtime, and run on an ordinary laptop CPU at inference time.

6. RESULTS AND DISCUSSION

6.1 Status of results

SilentSpeak has not been implemented or experimentally evaluated. **This is a proposed research direction, and no experimental results are claimed in this paper.** No accuracy, speed, training curve, confusion matrix or dataset statistic is reported for the framework.

6.2 What the literature allows us to say

Published studies show that convolutional front ends with recurrent or temporal back ends can recognize words and sentences from video of the mouth [4], [7], [8], that most of the benchmarks are in English or Mandarin [5], [8], and that a single approach can be extended to several languages when enough data are available [11]. These findings support the design choices of SilentSpeak. They do not tell us how well the framework will perform on Hindi or Marathi. The figures in Table 1 describe other systems on other tasks and are not predictions for this one.

6.3 Proposed evaluation plan

Table 6 lists the measures planned for a future implementation. Two hypotheses would be tested: that the bidirectional LSTM improves word-level accuracy over a CNN-only model, and that a transfer-learned model recognizes Hindi and Marathi words clearly above chance. For a balanced 20-word vocabulary the chance level is 5%.

Table 6. Proposed evaluation plan (planned measures, no results).

Measure	Planned protocol
Word-level accuracy	Speaker-independent test sets for English, Hindi and Marathi
Ablation	CNN-only model against CNN–BiLSTM on the same data and splits
Transfer benefit	Transfer-learned models against Hindi and Marathi models trained from scratch
Confusion matrix	Row-normalized, to inspect viseme-related confusions
Confidence analysis	Reliability of softmax scores [22]; threshold tuned on validation data
Emergency keywords	Recall and false-alarm rate on held-out clips
Prediction time	Per-word latency on a laptop CPU
Repeatability	Several random seeds; means with confidence intervals

6.4 Anticipated challenges and limitations

Several difficulties are expected, although none has been measured. Visually similar words will be confused, because of viseme ambiguity [4]. The regional vocabularies are small, so any future results will be a proof of concept and not a benchmark. Speakers, lighting, head pose and webcams will vary, and the English model is trained on controlled GRID video that differs from webcam recordings. The system targets isolated words, not sentences. Hindi and Marathi speech output depends on installed voices. Finally, collecting data from volunteers requires ethics review and consent procedures before any dataset is shared.

7. CONCLUSION

This paper proposed SilentSpeak, a multilingual framework for word-level lip reading in English, Hindi and Marathi. It combines MediaPipe-based lip extraction, a compact CNN–BiLSTM model, transfer learning from English to two regional languages, confidence scoring, an emergency-keyword alert and offline speech output. The review of the literature showed a gap in resources for Hindi and Marathi and little attention to practical features such as urgent-word handling. The methodology, architecture and evaluation plan are specified so that the framework can be implemented and tested. **This is a proposed research direction, and no experimental results are claimed in this paper.**

8. FUTURE SCOPE

- Implement the framework and evaluate it according to the plan in Table 6, including a from-scratch baseline and repeated runs.
- Collect the pilot datasets after ethics review, with documented consent and privacy procedures, and consider releasing them for research.
- Extend from isolated words to sentences, for example with a CTC loss as used in LipNet [4], and add a language model.
- Calibrate the confidence score [22] and add a “no word” class for a closed mouth.
- Deploy on mobile devices, extend to other Indian languages, and test the prototype with people who have hearing or speech impairments and with their caregivers.

REFERENCES

[1] World Health Organization, World Report on Hearing. Geneva, Switzerland: World Health Organization, 2021, ISBN 978-92-4-002048-1.

[2] S. Varshney, "Deafness in India," Indian J. Otol., vol. 22, no. 2, pp. 73–76, 2016, doi: 10.4103/0971-7749.182281.

[3] W. H. Sumby and I. Pollack, "Visual contribution to speech intelligibility in noise," J. Acoust. Soc. Amer., vol. 26, no. 2, pp. 212–215, 1954, doi: 10.1121/1.1907309.

[4] Y. M. Assael, B. Shillingford, S. Whiteson, and N. de Freitas, "LipNet: End-to-end sentence-level lipreading," arXiv:1611.01599, 2016. [Online]. Available: https://arxiv.org/abs/1611.01599

[5] J. S. Chung and A. Zisserman, "Lip reading in the wild," in Proc. Asian Conf. Comput. Vis. (ACCV), 2016, ser. Lecture Notes in Computer Science, vol. 10112, pp. 87–103, doi: 10.1007/978-3-319-54184-6_6.

[6] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Lip reading sentences in the wild," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, doi: 10.1109/CVPR.2017.367.

[7] T. Stafylakis and G. Tzimiropoulos, "Combining residual networks with LSTMs for lipreading," in Proc. Interspeech, 2017, pp. 3652–3656, doi: 10.21437/Interspeech.2017-85.

[8] B. Martinez, P. Ma, S. Petridis, and M. Pantic, "Lipreading using temporal convolutional networks," in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), 2020, pp. 6319–6323, doi: 10.1109/ICASSP40776.2020.9053841.

[9] S. Fenghour, D. Chen, K. Guo, B. Li, and P. Xiao, "Deep learning-based automated lip-reading: A survey," IEEE Access, vol. 9, pp. 121184–121205, 2021, doi: 10.1109/ACCESS.2021.3107946.

[10] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Deep audio-visual speech recognition," IEEE Trans. Pattern Anal. Mach. Intell., vol. 44, no. 12, pp. 8717–8727, Dec. 2022, doi: 10.1109/TPAMI.2018.2889052.

[11] P. Ma, S. Petridis, and M. Pantic, "Visual speech recognition for multiple languages in the wild," Nature Mach. Intell., vol. 4, no. 11, pp. 930–939, Nov. 2022, doi: 10.1038/s42256-022-00550-z.

[12] S. J. Pan and Q. Yang, "A survey on transfer learning," IEEE Trans. Knowl. Data Eng., vol. 22, no. 10, pp. 1345–1359, Oct. 2010, doi: 10.1109/TKDE.2009.191.

[13] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?," in Advances in Neural Information Processing Systems, vol. 27, 2014, pp. 3320–3328.

[14] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," J. Mach. Learn. Res., vol. 15, no. 56, pp. 1929–1958, 2014.

[15] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," J. Big Data, vol. 6, Art. no. 60, 2019, doi: 10.1186/s40537-019-0197-0.

[16] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Comput., vol. 9, no. 8, pp. 1735–1780, 1997, doi: 10.1162/neco.1997.9.8.1735.

[17] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," IEEE Trans. Signal Process., vol. 45, no. 11, pp. 2673–2681, Nov. 1997, doi: 10.1109/78.650093.

[18] C. Lugaresi et al., "MediaPipe: A framework for building perception pipelines," arXiv:1906.08172, 2019. [Online]. Available: https://arxiv.org/abs/1906.08172

[19] Y. Kartynnik, A. Ablavatski, I. Grishchenko, and M. Grundmann, "Real-time facial surface geometry from monocular video on mobile GPUs," arXiv:1907.06724, 2019. [Online]. Available: https://arxiv.org/abs/1907.06724

[20] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," J. Acoust. Soc. Amer., vol. 120, no. 5, pp. 2421–2424, Nov. 2006, doi: 10.1121/1.2229005.

[21] M. Abadi et al., "TensorFlow: A system for large-scale machine learning," in Proc. 12th USENIX Symp. Oper. Syst. Design Implement. (OSDI), Savannah, GA, USA, Nov. 2016, pp. 265–283.

[22] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in Proc. 34th Int. Conf. Mach. Learn. (ICML), vol. 70, 2017, pp. 1321–1330.