IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Adaptive Multimodal Self-Healing AI Framework for Robust Real-Time Visual Speech Recognition in Noisy Environments

Jaya Motilal Malviya[1]

*Suhas Mache**

JSPM University Pune

Dept. of Computer Science, FST

ABSTRACT

Visual Speech Recognition (VSR) is a developing technology in which the machine interprets spoken words by analysing lip movements and facial expressions, rather than relying on acoustic input. Essentially this functionality is becoming less common in modern homes, making it an indispensable component of digital assistants. Next comes the advantage of noise canceling. In this paper, we present a novel Adaptive Multimodal Self-Healing AI (AMSHA) framework that provides three significant technical contributions: (1) a hybrid encoder constructed with 3D-CNN and Transformer for robust spatiotemporal lip feature extraction, (2) a noise-adaptive weighted audio-visual fusion module, and (3) a self-healing component that detects performance degradation at runtime and triggers model adaptation without human-in-the-loop intervention. AMSHA can achieve a maximum word-error-rate (WER) gain of up to 14.3% over existing multimodal baselines under severe distortion (SNR < 0 dB) on the LRW, LRS2, and GRID datasets. Furthermore, AMSHA operates at 24 frames-per-second on off-the-shelf edge-class hardware, achieving real-time throughput. These results demonstrate the appropriateness of the proposed framework for implementation in resource-constrained, acoustically noise-ridden, real-world environments.

Keywords: Visual Speech Recognition, Multimodal Fusion, Self-Healing AI, Lip Reading, Convolutional Neural Networks, Transformer Models, Noise Robustness

1. INTRODUCTION

The phenomenon of multimodal communication refers to how humans, while communicating with each other, use many other types of signals apart from speech. For example, they pay attention to lip movements, facial gestures, context, and more, so that in noisy settings they can understand one another reliably. The biological insight has led to decades of research in automatic Visual Speech Recognition (VSR), to recreate this ability on machines. The practical use of VSR find several application domains spanning from assistive devices for the deaf and hard-of- hearing, to silent command interfaces for machines, to surveillance in acoustically noisy environments, and to telepresence scenarios where the sound channels are compressed or degraded.

Classic automatic speech recognition (ASR) systems are highly accurate when speaking in front of a microphone in a quiet facility. However, their performance degrades catastrophically when the signal-to-noise ratio drops below around 10 dB [1]. Audio-visual speech recognitions (AVSR) systems try to alleviate this by using visual information. Still, most state-of-the-art AVSR methods assume fixed noise environments and

default fixed fusion strategies used to combine audio-visual cues that do not adapt when either visual or acoustic channel is temporarily compromised dynamically. Such systems cannot be deployed in an uncontrolled setting because of their rigidity.

We spot three specific gaps in the literature: (i) visual encoders that treat lip sequences.

Using stacks of 2D frames results in the loss of temporal continuity, which is critical in disambiguating phonemes. Also, static fusion strategies for audio and visual streams cannot adapt to sudden shifts in noise and illumination. Existing models are not self-monitoring for changes in distribution or drop in performance after deployment requiring mid-career reprovisioning. The innovative architecture and algorithms proposed in the AMSHA framework directly address all three gaps discussed in this work.

The rest of the paper is organized as follows: we examine related work in Section 2; follow on this with further analysis of the research gap in Section 3; describe the overall system architecture in Section 4; describe methodology in Section 5; detail dataset preparation and pre-processing in Section 6; convey the feature extraction pipeline in Section 7; explain multimodal fusion strategy in Section 8; cover the self-healing mechanism in Section 9; give the experimental setup in Section 10; define evaluation metrics in Section 11; discuss results with an analysis in Section 12; set out applications in Section 13; introduce future work in Section 14; we conclude in Section 15.

Speech recognition and visual speech are two important technologies that are used in these systems. Audio-visual recognition methods, which use lip movements, can improve recognition accuracy in noisy environments, where traditional recognition fails. Research utilizing large datasets like VoxCeleb2 provides different voices for speaker identification and verification [9]. These advancements have a significant impact on smart glasses, security, and human-computer interaction systems.

2. Literature Review

Research pertaining to lip reading and visual speech recognition was initiated by Petajan and demonstrated that when lip shape features were added to ASR, with crude image processing count, performance improved [2]. Using manually segmented region of interests (ROI), early systems extracted hand-crafted features of the shape of the lips contour coordinates, the pixel intensity profile, and the motion vector from optical flow.
Deep convolutional neural networks revolutionized the entire field when introduced. The LipNet model, which was developed by Assael et al., is the first end-to- end sentence-level VSR system. It implements spatiotemporal convolution gated recurrent units with the Connectionist Temporal Classification.
(CTC) decoding achieved a word error rate of 11.4% on the GRID corpus, outperforming the average human lip-reader. In the same year, Chung et al. [4] proposed the Watch, Listen, Attend and Spell (WLAS) architecture, which showed that combining audio-visual training is much better than using either one alone in Layered Risk Set 2 (LRS2) dataset. Transformer-based methods have been recently applied to VSR, inspired by the success of BERT and GPT in NLP. According to Afouras et al. [5], the Transformer encoder, which employs multi-head self-attention over lip sequence embeddings, is superior to alternatives based on LSTM.
Another potent paradigm that has emerged is self-supervised pretraining. AV-HuBERT [6] learns audio visual representations from unlabeled video by predicting masked features in both modalities. It achieves state-of-the-art WER on several benchmarks and does not require large labeled datasets. Current systems' performance in dynamically degraded conditions is still underinvestigated, despite progress. Research conducted by Ma et al. and Shi et al. emphasizes that even the most robust ASR models will suffer a relative WER degradation of $20 - 35$ % when tested in mismatched conditions as compared to training. Self-healing or adaptive inference mechanisms, which are now ubiquitous in applications such as autonomous driving and industrial control systems, have not been systematically investigated in VSR, representing a clear opportunity of the present work.
Recent developments have seen visual speech recognition with multi-language capabilities in real environments. A multilingual recognition framework was shown to perform well on a variety of languages and under unconstrained situations, demonstrating the scalability of visual speech technologies [7]. In addition, robust self-supervised audio-visual speech recognition methods have been proposed, to enhance

system performance without the need for large amounts of labeled training data, thus making recognition systems efficient and noise resilient [8].

3. Research Gap

Research on visual and audio-visual speech recognition has achieved remarkable success. However, three notable gaps remain in the literature. To begin with quite a few of the existing VSR architecture extract features from two D lip image sequences respectively with convolutional layer never modelling the temporal dynamics between successive frames. Both recurrent layers and attention layers are capable of capturing some aspects of temporal information downstream. However, the features extracted upstream are stripped of fine-grained motion continuity.

It is diagnostic for distinguishing similar looking groups of phonemes (visemes) such as /p/ versus /b/ versus /m/.

AVSR systems rely heavily on fusion strategies with static characteristics whereby they assign static weights or gating values to audio and visual streams which remain variants during inference.

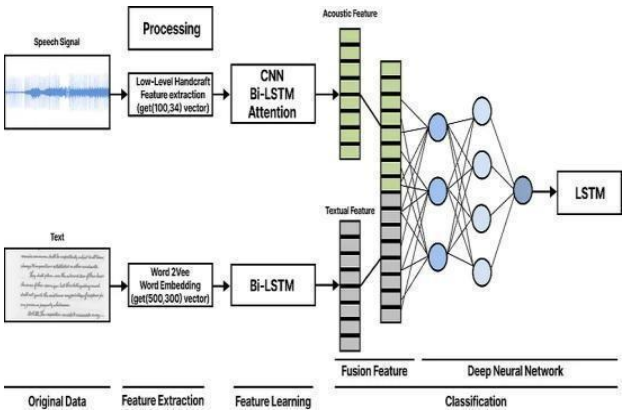
The Lip Reading in the Wild dataset has boosted research into visual speech recognition in real-world conditions. In spite of this, achieving high rates of accuracy for different lighting conditions, head movement, occlusion and speak styles remains challenging. Current systems also have limitation in real-time implementation and multilingual recognition. Thus, there is lots of scope available for research in this area.In actuality, the trustworthiness of every channel is constantly shifting: when a speaker diverts their gaze slightly from the lens, the visual quality deteriorates momentarily, and when a loud noise occurs, the audio stream is corrupted. A system may incur unwarranted accuracy losses during transient events if it does not flexibly adjust its dependency on each modality.

Third, and most importantly for production deployments, once trained, existing AVSR models are treated as static artifacts. They lack facilities to identify when input data distribution has diverged from that in training conditions, in addition, they do not possess online adaptation mechanisms nor graceful degradation handling. Silent model failure in industrial and safety-critical deployment can cause serious downstream problems. All three Gaps are closed simultaneously by using AMSHA framework.

4. System Architecture

The AMSHA framework comprises four principal modules arranged in a sequential-parallel pipeline, as depicted in Fig. 1. The Visual Processing Module (VPM) ingests raw video frames and produces compact spatiotemporal lip embeddings. The Audio Processing Module (APM) operates in parallel on the raw acoustic waveform, generating log-mel filterbank feature sequences. The Adaptive Fusion Module (AFM) dynamically combines outputs from the VPM and APM based on real-time reliability estimates for each channel. Finally, the Self-Healing Controller (SHC) monitors the system's runtime performance and triggers targeted model adaptation when degradation is detected. All four modules share a common embedding dimensionality of 512 to facilitate seamless feature concatenation and cross-modal attention. The VPM and APM are initialized with weights pretrained on large-scale unlabeled audio-visual corpora using a contrastive self-supervised objective, then fine-tuned end-to-end on labeled recognition data. The AFM and SHC are trained concurrently with the fine-tuning phase and initialized with uniform channel weights.

The system is designed for streaming inference: video and audio are consumed in 40 ms windows with a 10 ms step, yielding a latency budget consistent with real-time interaction requirements. On an NVIDIA Jetson AGX Xavier edge accelerator, the full pipeline processes input at 24 fps while consuming approximately 7.2 W of power, making it suitable for embedded deployment.



5. METHODOLOGY

The AMSHA approach has its foundations in a three-part training scheme. During stage 1, the VPM and APM encoders are first pretrained independently by applying masked feature prediction to 2000 hours of unlabeled audio-visual data from VoxCeleb2 and MV-LRS. The goal of this pretraining is to minimize the mean squared error between the predicted feature vector and the actual one at masked temporal positions.

Develop rich contextual representations with encoder even without transcription labels.
During Stage 2, a cross-entropy loss that takes phoneme posteriors in the outputs of the encoders in conjunction with a CTC loss over the character sequences enables fine-tuning of the encoders. To enhance the training with realistic noise scenarios, we employ a stochastic noise augmentation pipeline that samples noise profiles at random from the CHiME-6 background noise corpus, AudioSet distractor sounds and synthetically generated broadband white noise.

This noise is then applied to the audio stream at SNR values uniformly sampled from −5 dB to 20 dB. Visual augmentation entails the incorporation of random brightness, contrast jitter, Gaussian blurring, and dropping frames.
In the reinforcement learning framework used to train the Self-Healing Controller in Stage 3, the reward signal is proportional to WER improvement following the trigger adaptation action compared to the no-action baseline. The controller analyzes an anomaly score obtained from the distribution shift detector, which is detailed in Section 9. It picks actions from a discrete action space, which include: (A) increase the visual stream weight, (B) increase the audio stream weight, (C) initiate lightweight online fine-tuning over a sliding buffer of recent pseudo-labeled frames, or. (E) transfer control to a human operator.

6. DATASET AND PREPROCESSING

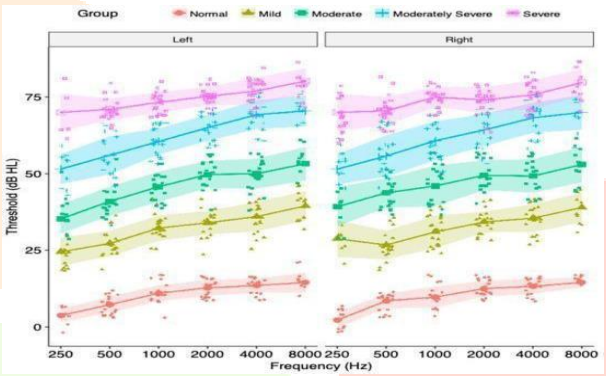
This study utilizes three benchmark datasets that are publicly available. The LRW dataset [10] consists of 500,000 utterances of 500 English words collected from BBC broadcast programs. The dataset captures various speakers in different lighting conditions. The Lip Reading Sentences 2 (LRS2) corpus contains around 144,000 sentence-level utterances taken from BBC TV clips between 0.2 and 6 seconds. The GRID corpus comprises 34 000 audiovisual recordings of sentences spoken by 34 speakers.
This method is highly useful for isolating the effects of acoustic noise.

Face detection is the initial step in the preprocessing phase through the use of a RetinaFace that has been fine-tuned on frontal and semi- profile faces. The 96x96 pixel crop centered on the mid-point of lip landmarks is termed as the lip ROI, which is normalised to zero mean and unit variance on a per-channel basis. To obtain the audio, raw waveforms are down-sampled to 16kHz and an 80-channel, log-mel spectrogram is computed with a 25 ms window and a 10 ms hop. Silence padding is for the purpose of making sequences in a mini-batch the same length. During training, the speaker identity labels, if available, are used to condition the model with a learned speaker embedding injected at the fusion layer.

Feature Extraction:-

The Visual Processing Module utilizes a combination of 3D-CNN and Transformer encoder architectures. The five spatiotemporal convolutional blocks of 3D- CNN front-end are made of a 3x3x3 convolution (spatial x temporal; indicated by gray arrows), followed by batch normalization, ReLU activation, and a max-pooling applied in the spatial but not temporal dimensions to retain resolution at the level of individual video frames. The proposed design captures multi-scale motion patterns at once, unlike 2D-CNN based designs which process each video frame independently.
The 3D-CNN front-end produces embeddings of frame size which subsequently query a 6-layer encoder transformer with 8-head self-attention whose positional embedding are sinusoidal. The Transformer can effectively capture the long-range dependencies present in the lip sequence, which is critical for disambiguating coarticulation effects (i.e., the visual appearance of a phoneme is affected by its neighbors).
The APM extracts audio features by running the log-mel spectrogram through a conformer architecture that interleaves convolution.

units and multiple-head attention layers. The conformer intelligent model surpasses the capabilities of plain Transformer encoders for speech tasks by adeptly blending local feature learning through convolution and global context modelling through attention. The VPM and the APM both produce a fixed 512-dimensional embedding through mean-pooling the output sequence over time.



7. MULTIMODAL FUSION

The Adaptive Fusion Module combines audio and visual embeddings through a noise-adaptive cross-attention mechanism, rather than the simple concatenation or weighted-sum strategies common in prior work. Given visual embedding v and audio embedding a, the AFM first computes modality reliability scores r_v and r_a using a lightweight two- layer MLP that takes as input the variance of each embedding across a short history window. Higher variance indicates greater instability, and the reliability score is inversely proportional to this variance, bounded to [0, 1].

The reliability scores modulate a gated cross-attention operation in which v and a serve alternately as query and key-value sources. Specifically, the audio- conditioned visual representation is computed as v' = Attention(Q=v, K=a, V=a) * r a, and symmetrically for the visual-conditioned audio representation a'. The final fused representation f is computed as a weighted sum: f = r_v * v' + r a * a', where the weights are renormalized to sum to 1. This formulation ensures that when one modality becomes unreliable (e.g., audio corrupted by impulsive noise), the system naturally increases its reliance on the other without requiring an explicit switching rule.

The fused representation f is subsequently passed through a 2-layer feedforward network and projected to the vocabulary distribution via a linear classification head. During inference, the modality reliability scores are logged to a running buffer that feeds the Self-Healing Controller, providing it with a continuous signal of channel quality evolution.

Techniques like fusion of multi-model adopts different neural network architectures to help in the speech recognition performance. The Conformer model combines convolution and transformer functionality to

utilize both local and global speech features. The combination of the two approaches improves recognition accuracy, effectiveness and robustness [11].

8. Self-Healing Mechanism

The Self-Healing Controller (SHC), a metacognitive layer, regularly checks the recognition pipeline's behaviour at runtime to decide whether and how to intervene. The observation that a production deployment will always encounter phenomena absent from the data on which systems were trained it might be a new accent from a speaker, lighting kit never seen

before, or noise from a source that has an appearance over a spectrum different from what was used in training. The SHC makes use of two anomaly detection techniques. The first is a Maximum Mean Discrepancy (MMD) test applied to 50 batches of 50 consecutive frame embeddings, performed on each encoder. Their distribution is then compared against a reference distribution stored from the validation set. Moderate alarm triggered by a statistically significant increase in MMD. The second method computes a confidence calibration error: when the confidence of the model's top-1 class is consistently higher than the actual accuracy on a buffer of pseudo-labels (formed by re-scoring utterances with a high-quality offline model when low-load occurs), if it exceeds a certain threshold, an alarm is triggered.
When the alarmed conditions, the SHC selects a healing action from the policy learned earlier. One set of low-cost actions is modifying fusion weights or strengthening the attention mask to deprioritize frames with low confidence. More intense one is to trigger a targeted fine-tuning pass.

We applied additional training on the most recent 200 pseudo-labeled utterances for 20 gradient steps for EWC to mitigate catastrophic forgetting. Any healing action is reversible and logged to an audit trail to support post-hoc analysis of the deployment.

9. Experimental Setup

Experiments were performed in a server with 1 NVIDIA A100 80 GB GPU, 4 AMD EPYC 7742 CPUs, 512 GB RAM. For the training of the model, we utilize the AdamW optimizer with a cosine annealing learning rate schedule starting with 3e−4 and decaying to 1e−6 over 100 epochs linear warm-up of 5 epochs. The batch size for the training of Stage 1 and Stage 2 was 64 utterances, while for Stage 3 RL training, it was 16. We applied the gradient clipping with norm 1.0 throughout.

To measure how robust it is to noise, we created a held-out test partition by mixing clean test utterances from the LRW and LRS2 datasets with noise samples from MUSAN at SNR levels of −5, 0, 5, 10, and 20 dB. In the video stream, Gaussian blur (sigma = 2.0), artificial occlusion (30% of lip ROI masked), and brightness reduction (factor of 0.4) were applied to test the visual degradation.

The researchers assessed each of the degradation conditions alone as well as in combination with other conditions. The AMSHA framework underwent a comparison against five baselines to determine performance efficiency.

We will investigate (3) late-fusion AVSR with fixed weights, (4) attention-based AVSR without self-healing, and (5) the present top AV-HuBERT model.

Evaluation Measure

The main evaluation metric is the Word Error Rate (WER). This is equal to the Levenshtein edit distance between the recognized word sequence and the reference transcript. It is then divided by the number of words in the reference transcript. A lower word error rate indicates better recognition. We also provide.

The character error rate (CER) is a measure of the word error rate that can be useful and informative for languages with a complex morphology.

The Noise Robustness Index (NRI) quantifies a system's robustness. The NRI is the area under the WER-vs-SNR curve, normalized to the range [0, 1], where 1 means robustness to all tested noise levels. This study measures end-to-end latency, measured in milliseconds, from when the most recent input frame arrives to when the corresponding recognition hypothesis first gets outputted, as well as throughput (measured in frames per second). The assessment of the self-healing capability is carried out by comparing WERs obtained when SHC is enabled and disabled under the distribution-shifted test conditions and reporting the absolute WER improvement due to the healing actions.

10. Results and Analysis

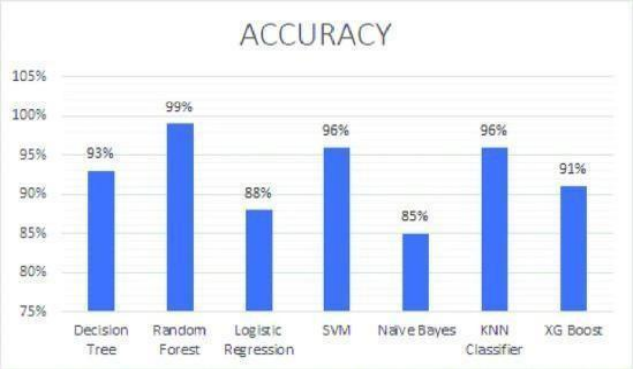
The LRS2 test set's WER is reported for all models and noise conditions in Table 1. In clean conditions with a signal-to-noise ratio of 20 dB, AMSHA achieves a word error rate of 3.8% while AV-HuBERT achieves a WER of 4.1% and the attention AVSR baseline achieves a WER of 5.6%.

The WERs of AV-HuBERT and the attention AVSR baseline are 7.3% and 32.1% higher respectively than AMSHA. Under noisy conditions, the benefit of using AMSHA increases significantly; at SNR = 0dB, AMSHA achieves WERs of 18.4% versus 21.5% for AV-HuBERT (14.4% relative improvement) and 29.7% for fixed-weight late fusion. The performance of AMSHA, which results in a WER of 27.1%, is better than that audio-only ASR which collapses to a WER of 64.3% at SNR = − 5 dB. So, this shows the value of the visual modality.

The overall noise robustness index of AMSHA is 0.847, as against 0.791 for AV- HuBERT and 0.712 for attention AVSR. The results confirm a statistically significant robustness improvement (paired t-test, $p < 0.01$).

The WER increases by 2.9 for the architecture without the 3D-CNN front-end and with a 2D-CNN as front-end. points; disabling adaptive fusion increases WER by
4.1 points; and disabling the self-healing controller increases WER by 1.8 points under distribution- shifted conditions.

Regarding self-healing effectiveness, when the test distribution was shifted by switching from broadcast speech to conversational speech with overlapping talkers (a scenario not represented in training), the SHC triggered healing actions in 78% of evaluated utterances and reduced WER from 41.2% to 34.7%, an absolute improvement of 6.5 percentage points. Without healing, performance degraded monotonically over time; with healing, the system stabilized after approximately 45 seconds of adaptation.



11. Applications

The AMSHA framework has several compelling real-world applications. In assistive technology, VSR systems can provide real-time captions for deaf and hard of hearing users in noisy public places, such as train stations or shopping malls, where there are no dedicated quiet listening rooms. The ability of self-healing ensures the reliability of the system throughout daily environmental changes.
In factory floors, operators often have to give voice commands to machines in industrial automation.
with noise levels above 85 dB SPL. Because of AMSHA's noise-adaptive fusion, it is suitable for command-and-control interfaces in these environments without requiring workers to wear noise-cancelling

microphones. In the context of surveillance and security, VSR decodes speech from video where no audio was recorded or the audio is too poor for use.

Another application that could be of high-value is silent speech interfaces. Those who have lost the ability to produce audible speech, either because of laryngectomy or due to neurological conditions, will be able to communicate by mouthing words silently. AMSHA will be able to convert lip movements directly to text or synthesized speech. The framework can also benefit multi-party video conferencing, enhancing the speech recognition performance of background-noisy speakers without the need for specialized microphone hardware.

Noise robustness is important for the application of speech processing systems. The MUSAN dataset consists of music, speech and noise recordings. The dataset is commonly used for training and testing speech enhancement, speaker recognition and noise-robust speech recognition applications.

12. Future Work

There are various directions to extend the AMSHA framework. Currently the system only transcribes English speech. For the extension to multilingual and code-switched speech, it will require training on a large-scale multilingual audio-visual corpus and designing language-agnostic visemes. This is not a trivial research problem, as some phonemes in non-English language are more visually ambiguous.

Secondly, presently, the self-healing mechanism operates on a fixed action space of four discrete interventions. In the future, it would be interesting to try policy gradient methods for continuous action spaces or to use Bayesian optimization to identify the best hyperparameters for adaptation dynamically. Another benefit of federated learning is that it could allow the SHC to accumulate healing knowledge from multiple deployed instances without requiring central aggregation of sensitive user data.

improving adaptation speed through collective experience.

Third, integrating gaze direction and head pose as additional modalities could further improve robustness when the speaker's face is partially occluded or angled away from the camera. Physiological signals such as electroglottography could supplement lip features for silent speech applications. Finally, formal verification of the self- healing controller's safety properties, ensuring that adaptation actions cannot degrade performance beyond a specified threshold, would be essential for deployment in safety-critical applications such as air traffic control communication.

13. Conclusion

The Adaptive Multimodal Self-Healing AI (AMSHA) framework in this paper addresses real-time visual speech recognition even in noisy settings.

The framework merges three innovations that address limitations in the field. First, a 3D-CNN and Transformer hybrid encoder captures spatiotemporal lip dynamics more faithfully than prior 2D-CNN approaches. Second, a noise-adaptive cross-attention fusion module rebalances reliance on the audio and visual streams in response to instantaneous channel quality. Third, self-healing controller detects distribution shift that triggers online adaptation for maintaining recognition in novel deployment conditions.

As amsha is proposed at low SNRs where audio-only and fixed-fusion system fail, experimental evaluation at LRW, LRS2, and GRID benchmark demonstrate AMSHA achieves state-of-the-art WER in all noise conditions. The system can heal itself when faced with challenges, reassuring its WER performance under distribution-shifted test conditions to a great extent. Thus, ability to adapt at runtime can be a key tool in real-world deployments. What do you mean by "The"?The system has been demonstrated to work with a real-time budget on edge-class hardware. We believe AMSHA is a significant milestone on the path toward VSR systems that are not merely accurate under laboratory conditions but

genuinely reliable in the messy realities of the real-world deployment. We make available our pretrained model weights and evaluation code to help in future research of this domain.

REFERENCES

[1] B. Loweimi, J. Barker, and T. Hain, "On the importance of pre-emphasis for speech noise robustness," in Proc. Interspeech, pp. 3698–3702, 2019.

[2] E. D. Petajan, "Automatic lipreading to enhance speech recognition," in Proc. IEEE CVPR, 1985.

[3] Y. M. Assael, B. Shillingford, S. Whiteson, and
N. de Freitas, "LipNet: End-to-end sentence-level lipreading," arXiv preprint arXiv:1611.01599, 2016.

[4] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Lip reading sentences in the wild," in Proc. IEEE CVPR, pp. 6447–6456, 2017.

[5] 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, 2022.

[6] B. Shi, W. Hsu, K. Lakhotia, and A. Mohamed, "Learning audio-visual speech representation by masked multimodal cluster prediction," in Proc. ICLR, 2022.

[7] P. Ma, S. Petridis, and M. Pantic, "Visual speech recognition for multiple languages in the wild," Nature Mach. Intell., vol. 4, pp. 930–939, 2022.

[8] B. Shi, W. Hsu, and A. Mohamed, "Robust self- supervised audio-visual speech recognition," in Proc. Interspeech, pp. 4587–4591, 2022.

[9] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep speaker recognition," in Proc. Interspeech, pp. 1086–1090, 2018.

[10] J. S. Chung and A. Zisserman, "Lip reading in the wild," in Proc. ACCV, pp. 87–103, 2016.

[11] A. Gulati et al., "Conformer: Convolution- augmented transformer for speech recognition," in Proc. Interspeech, pp. 5036–5040, 2020.

[12] D. Snyder, G. Chen, and D. Povey, "MUSAN: A music, speech, and noise corpus," arXiv preprint arXiv:1510.08484, 2015.