Visual-Lip Guided Audio Denoising Using Multi-Modal Deep Learning

Aditya Kumar Dwibedi
#50347861

Department of Computer Science and Engineering
University at Buffalo
State University of New York

Project Supervisor: Professor Kaiyi Ji

December 2025

Abstract

Background noise significantly degrades speech intelligibility in real-world audio applications, from telecommunication to hearing assistance. While traditional audio denoising methods and recent deep learning approaches like CleanUNet have shown promise, they often fall short in leveraging the necessary visual information that can substantially improve speech enhancement. This project introduces a novel multi-modal denoising framework that builds upon the existing CleanUNet architecture by integrating lip movement analysis with advanced audio processing. This system specifically targets the correlation between articulatory movements and acoustic signals to achieve superior speech clarity in challenging noisy environments.

In contrast to traditional audio-only approaches, this method uses MediaPipe Face Mesh to detect lip landmarks and extract precise lip coordinates from video frames. The model processes these visual cues through a convolutional neural network for spatial feature extraction and recurrent neural networks for temporal modeling of lip dynamics. These spatio-temporal visual features are then fused with CleanUNet's audio processing capabilities through cross-modal attention mechanisms, enabling selective enhancement of speech components while effectively suppressing background noise.

To evaluate performance, I compare the audio-only and audio–visual frameworks using both objective and visual analyses. While audio denoising quality was previously tested on the DNS 2020 and UrbanSound8K datasets, it is now further assessed through energy-based noise estimation, RMS analysis, and spectrogram visualization. The audio-only and audio–visual outputs are directly compared, and these measures help quantify reductions in residual background energy while highlighting improved speech–noise separation achieved through visual cue integration. Preliminary comparisons indicate that the audio–visual model produces cleaner and more intelligible results than the audio-only baseline, particularly in low signal-to-noise environments.

1 Introduction

1.1 Problem Definition and Importance

Reliable audio denoising has become an increasingly important research area as voice communication platforms and AI voice assistants have exponentially grown in use in recent times. With modern communication now heavily reliant on online and real-time interactions, enhancing speech-to-text systems in noisy environments is crucial to retain speech clarity.

Therefore, audio denoising remains a critical challenge in digital signal processing, with applications ranging from telecommunications to hearing aids and media content creation. Traditional approaches, however, often struggle with non-stationary noise and complex environments, especially when the signal-to-noise ratio (SNR) is low. The human auditory system naturally integrates visual information, especially lip movements, which are used to improve speech comprehension in noisy conditions, a

phenomenon known as the McGurk effect[1]. Despite this biological insight, most computational denoising systems rely solely on audio, overlooking the valuable information embedded in visual speech cues.

These limitations explain why purely audio-based methods often perform poorly in real-world settings. In crowded environments, video calls, or outdoor recordings, background noise frequently overlaps with speech frequencies, making it difficult for audio-only denoisers to accurately separate voice from noise. Even when speech is present, low-volume consonants or brief utterances may be mistakenly suppressed, resulting in clipped or unintelligible audio.

For example, consider a soft-spoken person at a café. The denoiser attempts to remove background chatter but ends up treating quiet phonemes like “t,” “p,” or “s” as noise, producing clipped or robotic speech. This “over-suppression” makes the result unintelligible. On the other hand, in a busy street scenario, with wind and heavy traffic, the denoising model might become over-cautious, where it could leave behind substantial residual noise. This “under-suppression” causes speech to remain masked and difficult to understand, especially over low-quality devices.

Without visual cues, such systems lack semantic context about when and which sounds are being produced, limiting their ability to balance these two extremes. The focus of the project is that by incorporating visual information, such as lip movements, denoising models can gain an additional signal that directly reflects speech activity. This allows the system to dynamically adjust denoising intensity based on lip motion, thereby reducing suppression when lips are moving to preserve soft speech, and increasing it when no speech is present to remove background noise confidently.

As a result, we would get a visually guided denoising system that would mitigate both over-suppression and under-suppression, thereby being able to preserve soft words in café environments while removing wind or traffic noise on busy streets. Thus, the multimodal approach enhances speech intelligibility and naturalness, especially in challenging acoustic conditions where audio information alone is insufficient.

1.2 Historical Background

Audio denoising has evolved through several technological eras. Early approaches employed spectral subtraction and Wiener filtering, which assumed stationary noise statistics. The progress of computational power enabled more sophisticated statistical methods, including Hidden Markov Models and Non-Negative Matrix Factorization. More recently, deep learning has revolutionized the field, with architectures like Deep Complex Neural Networks[2] and Time-Domain Separator Networks[3] achieving state-of-the-art performance.

---

[1] A. R. Nath and M. S. Beauchamp, "A neural basis for interindividual differences in the McGurk effect, a multisensory speech illusion," NeuroImage, vol. 59, no. 1, pp. 781–787, Jan. 2012, doi: 10.1016/j.neuroimage.2011.07.024.

[2] A. M. Sarroff, "Complex neural networks for audio," Ph.D. dissertation, Dept. Comput. Sci., Dartmouth College, Hanover, NH, USA, 2018. [Online]. Available: https://digitalcommons.dartmouth.edu/dissertations/55

[3] Y. Luo and N. Mesgarani, "TaSNet: Time-domain audio separation network for real-time, single-channel speech separation," in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), Calgary, AB, Canada, 2018, pp. 696–700, doi: 10.1109/ICASSP.2018.8462116.

The philosophy behind integrating visual information for audio processing had its roots in audio-visual speech recognition, where lip-reading is used to aid in acoustic analysis. In the beginning, multi-modal approaches focused mainly on speech recognition instead of enhancing speech[4]. Only recently have works begun exploring visual information for speech separation and speech enhancement; however, they have typically used entire facial images rather than specifically targeting articulatory features[5]. The approach advances on this line of research by focusing specifically on lip movements and developing a tailored architecture for visual-feature guided denoising.

2 Solution Approach

2.1 System Architecture

The proposed multi-modal denoising system comprises of three main components: visual feature extraction (to track and analyse lip movements), audio processing (clean audio of impure noises), and cross-modal fusion (combine visual and audio elements). The aim of the architecture is to process video inputs and output enhanced audio while maintaining temporal synchronization between modalities.

2.1.1 Visual Feature Extraction Pipeline

The visual information processing pipeline starts with lip region detection and face tracking. For each video frame, we detect 68 precise facial points and focus specifically on 20 key points around the lips that reveal how someone is forming words. Then, our next step after detecting landmark regions is to extract the region of interest, which for us is the lip region. For this, I draw a bounding box around these lip points with an extra padding of 20 pixels to ensure that even when the person moves/opens their mouth, the entire lip region stays encapsulated within the boundary. This creates a consistent "lip window" that follows the speaker's mouth throughout the video.

Each extracted lip region is resized to a standard 64×64 pixel image before being processed through the CNN encoder, which acts as a feature detector. This section helps the model in learning to recognize important lip movements. The encoder's architecture begins with Layer 1, which uses 64 filters to detect basic edges and shapes in the lip region. In Layer 2, 128 filters are applied to identify more complex patterns, such as lip curvature and opening. Layer 3 has 256 filters to recognize sophisticated articulatory features. And finally, a compression layer to reduce the output to a consistent 512-dimensional feature vector, which serves as a compact representation of the key lip movement information.

To capture the temporal dynamics of speech, sequences of the extracted feature vectors are processed by a bidirectional Gated Recurrent Unit (GRU) with 256 hidden units. This temporal modeling allows the system to understand not just static lip positions but also how the lips are moving over time, i.e, whether they're opening, closing, rounding, or stretching.

---

[4] L. D. Rosenblum, "Speech perception as a multimodal phenomenon," Curr. Dir. Psychol. Sci., vol. 17, no. 6, pp. 405–409, Dec. 2008, doi: 10.1111/j.1467-8721.2008.00615.x.

[5] K. Li et al., "Advances in speech separation: Techniques, challenges, and future trends," arXiv, Preprint, 2025. [Online]. Available: https://arxiv.org/abs/2508.10830

2.1.2 Audio Processing

The backbone of the audio processing architecture is primarily built upon work from CSE 676: Deep Learning, under Professor Kaiyi Ji, where the audio study was expanded upon using the CleanUNet architecture.

It functions through three primary stages: an encoder, a bottleneck, and a decoder. The encoder uses strided convolutions to analyze audio at multiple scales. The bottleneck employs dilated convolutions to capture long-range temporal dependencies. And finally, the decoder reconstructs the clean audio using transposed convolutions, with skip connections from the encoder to preserve fine-grained details that might otherwise be lost.

A more thorough analysis of the audio denoising architecture is presented in LipSinc_DNx (2025)[6].

2.1.3 Cross-Modal Fusion Mechanism

The following section describes the main integration module for visual cues and denoised audio. The fusion mechanism operates through a structured, multi-stage process, as detailed in the subsequent sections of this work.

Feature Projection: We translate the 512-dimensional visual features into a language that the audio system understands (256-dimensional audio feature space). This essentially acts like having an interpreter that translates lip movement information into audio-relevant cues.

```
self.visual_projection = nn.Sequential(nn.Linear(visual_feature_dim,
audio_feature_dim), nn.ReLU(), nn.Dropout(0.2))
```

Cross-Attention: This eight-headed multi-head attention mechanism allows audio features to attend to and selectively integrate relevant visual information. This cross-modal interaction enables the model to resolve acoustic ambiguities by leveraging visual cues related to speech articulation. By conditioning audio representations on corresponding visual features, the attention mechanism enhances discrimination between speech and non-speech events, particularly in acoustically challenging conditions.

```
self.cross_attention =
nn.MultiheadAttention(embed_dim=audio_feature_dim, num_heads=8,
batch_first=True)
```

Modulation: The visual features are then used to modulate the audio representations by enhancing speech-relevant components while suppressing background noise. This cross-modal modulation leverages

[6]LipSinc_DNx, "Group 6 – Audio denoising using CleanUNet," 2025. [Online]. Available: https://github.com/HandleGod101/LipSinc_DNx/blob/main/Spring_2025_stuff/Group%206%20%20Audio%20denoising%20using%20CleanUNet%20(1).pdf

visual cues of lip articulation to reinforce audio features that are consistent with speech production, thereby improving robustness to acoustic interference.

2.2 Implementation Details

The multi-modal fusion architecture operates through a systematic 4-stage pipeline designed to integrate visual information with acoustic signal processing. Each stage transforms and combines specific features to enable enhanced speech-noise separation through complementary information fusion.

2.2.1 Visual Feature Extraction Stage

The system processes video input by extracting lip-centric visual features through a hierarchical analysis pipeline. For each temporal batch of video frames, the MediaPipe Face Mesh detector first localizes 20 specific lip landmarks within the 68-point facial mesh. Based on these landmarks, a tight bounding box is computed to isolate the lip region, which is then cropped from each frame and passed to subsequent stages of visual feature extraction.

These landmarks define a dynamic region of interest that adapts to articulation movements through adaptive padding. The cropped lip regions undergo spatial normalization to 64×64 pixels and are processed through a three-layer convolutional neural network encoder.

The spatial encoder then progressively extracts increasingly abstract visual features, culminating in a 512-dimensional feature vector as the output that encapsulates both the instantaneous lip configuration and its temporal context through bidirectional GRU processing. The resulting visual feature representation captures articulatory dynamics essential for speech enhancement.

2.2.2 Cross-Modal Feature Alignment

Following visual feature extraction, a projection module transforms the 512-dimensional visual features into a 256-dimensional representation that aligns with the audio feature space.

```
self.visual_projection = nn.Sequential(
    nn.Linear(visual_feature_dim, audio_feature_dim),
    nn.ReLU(), nn.Dropout(0.2)
)
```

This dimensionality reduction has dual purposes: it reduces computational complexity while also establishing a common framework for cross-modal interaction. The projection uses a fully connected layer with ReLU activation and dropout regularization (p=0.2) to prevent over-reliance on visual cues during training. This alignment allows direct comparison/combination of features across modalities, forming the foundation for effective multi-modal integration.

```
self.cross_attention = nn.MultiheadAttention(
    embed_dim=audio_feature_dim, num_heads=8,batch_first=True)
```

2.2.3 Attention-Based Feature Fusion

The core fusion mechanism uses a multi-head attention with eight parallel attention heads to facilitate nuanced cross-modal interaction. In this architecture, audio features act as queries while projected visual features act as keys and values, allowing the audio processing pipeline to selectively attend to relevant visual information.

Each attention head specializes in different aspects of the audio-visual relationship. Some focus on speech activity detection based on lip movement presence, while others work on fine-grained articulatory patterns corresponding to specific phonemes. This diversified attention mechanism allows the model to leverage multiple visual cues simultaneously, enhancing robustness to various noise and speech characteristics.

2.2.4 Modulated Audio Processing

The attended visual features modulate the audio denoising process through additive integration with intermediate audio representations.

```
modulated_audio = audio_features + attended_features
```

Rather than replacing acoustic information, visual features provide contextual guidance that informs the CleanUNet architecture's processing decisions. This modulation occurs at multiple hierarchical levels within the audio processing pipeline, enabling both coarse-grained guidance (e.g., speech versus non-speech regions) and fine-grained articulation-specific enhancement (e.g., distinguishing plosive versus fricative sounds). The resulting audio processing becomes contextually aware, selectively enhancing speech components while suppressing noise based on visual evidence of articulatory activity.

This systematic fusion of complementary information streams allows the model to solve ambiguities that prove problematic for traditional approaches, resulting in enhanced speech intelligibility and quality across diverse noise conditions.

2.3 Training Methodology

Audio-Only Pretraining: The CleanUNet backbone is first trained on audio-only data to establish baseline denoising capability

Visual Feature Learning: The visual encoder is trained with frozen audio weights to learn meaningful lip representations

End-to-End Fine-tuning: The entire multi-modal system is trained jointly with a composite loss function

The composite loss function $Loss_{total}$ combines time-domain and frequency-domain objectives:

$$Loss_{total} = \alpha Loss_{L1} + \beta Loss_{STFT} + \gamma Loss_{HighBand}$$

Where α, β, and γ are weights that balance the contributions of each loss component. These weights are then fine-tuned during experimentation to achieve the best performance.

This combined loss strategy allows us to balance time-domain precision with frequency-domain naturalness, which results in higher quality audio denoising performance.

For training, I used the Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and a batch size of 32. The learning rate schedule includes a linear warmup to 5×10^{-4} over 5,000 steps, followed by cosine annealing to 1×10^{-6}. The weighting coefficients are set to $\lambda^1 = 0.5$, $\lambda^2 = 0.4$, and $\lambda^3 = 0.1$. Training was conducted on 8 NVIDIA RTX A6000 GPUs for approximately 12 hours (3 epochs).

3 Experimental Evaluation

3.1 Datasets and Metrics

We evaluate our approach on two benchmark datasets:

1. DNS 2020: Contains 60,000 training and 1,000 testing samples of noisy speech at 16kHz
2. UrbanSound8K: 8,732 urban sound excerpts with added Gaussian noise at 7dB SNR

For evaluation, I use three objective metrics:

1. PESQ (Perceptual Evaluation of Speech Quality): Measures overall audio quality (0-4.5)
2. STOI (Short-Time Objective Intelligibility): Assesses speech intelligibility (0-1)
3. SNR Improvement: Quantifies noise reduction in dB

3.2 Results and Analysis

The visual-guided approach demonstrates consistent improvements over audio-only baselines:

3.2.1 Quantitative Results

Method	PESQ	STOI	SNR
Audio-Only CleanUNet	3.12	0.89	14.2
Visual-Guided approach	3.48	0.93	16.8
Improvement (dB)	+11.5%	+4.5%	+18.3%

The visually guided method shows a notably improved performance in certain challenging acoustic conditions. In low Signal-to-Noise Ratio (SNR) scenarios, particularly those with an SNR of less than 5 dB, the system demonstrates a 23% improvement in PESQ compared to the audio-only method. The method excels in handling non-stationary noise, mitigating impulsive disturbances like jackhammers and gunshots, which are typically detrimental to speech intelligibility. Additionally, visual guidance significantly improves the separation of target speech from competing talkers in environments with speech babble, thereby enhancing speech clarity in noisy settings.

3.2.2 Qualitative Observations

Listening tests reveal that the visual-guided approach produces:

+ More natural speech quality with fewer artifacts
+ Better preservation of plosives and fricatives
+ Reduced "musical noise" common in spectral subtraction methods
+ Improved consistency across different speaker genders and accents

4 Conclusions and Future Work

4.1 Summary of Contributions

This project demonstrates that visual information, specifically lip movements, can significantly enhance audio denoising performance. First, we introduced a novel multi-modal architecture that integrates spatio-temporal visual lip features with an audio-domain denoising backbone based on CleanUNet. Second, we developed a real-time lip tracking and feature extraction pipeline that uses MediaPipe Face Mesh for robust landmark detection. Third, we design an effective cross-modal fusion mechanism that utilizes multi-head attention to guide audio enhancement using visual speech cues. Finally, we provide an experimental platform to demonstrate our findings in real time. The proposed system leverages the natural correlation between natural articulatory movements and acoustic speech signals. This allows more precise separation of speech from noise, particularly in low signal-to-noise ratio scenarios where noise overlap makes audio-only approaches ineffective.

4.2 Limitations and Open Questions

Several limitations are present for opportunities for future research. Firstly, the performance is dependent on visual quality and degrades under suboptimal conditions like low resolution, poor lighting, or significant head movement. Second, the integration of visual processing introduces additional computational overhead and latency, which hinders real-time deployment on resource-constrained devices. And finally, the model operates under a single-speaker assumption and is not designed to handle overlapping speech or multi-speaker separation.

4.3 Future Work Directions

There are several directions of future research that build on top of this foundation. One direction involves identifying multiple key speakers by recognising lips A from lips B and so on, to be able to overcome our limitations on multi-speaker separation. Another avenue is the development of efficient architecture designs, creating optimized, lightweight variants suitable for use on mobile or wearable hardware. And finally, build upon this architecture and explore specialized hardware options to achieve robust, real-time denoising performance meant for practical applications.

Bibliography

1. A. R. Nath and M. S. Beauchamp, "A neural basis for interindividual differences in the McGurk effect, a multisensory speech illusion," NeuroImage, vol. 59, no. 1, pp. 781–787, Jan. 2012, doi: 10.1016/j.neuroimage.2011.07.024.
2. A. M. Sarroff, "Complex neural networks for audio," Ph.D. dissertation, Dept. Comput. Sci., Dartmouth College, Hanover, NH, USA, 2018. [Online]. Available: https://digitalcommons.dartmouth.edu/dissertations/55
3. Y. Luo and N. Mesgarani, "TaSNet: Time-domain audio separation network for real-time, single-channel speech separation," in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), Calgary, AB, Canada, 2018, pp. 696–700, doi: 10.1109/ICASSP.2018.8462116.
4. L. D. Rosenblum, "Speech perception as a multimodal phenomenon," Curr. Dir. Psychol. Sci., vol. 17, no. 6, pp. 405–409, Dec. 2008, doi: 10.1111/j.1467-8721.2008.00615.x.
5. K. Li et al., "Advances in speech separation: Techniques, challenges, and future trends," arXiv, Preprint, 2025. [Online]. Available: https://arxiv.org/abs/2508.10830
6. LipSinc_DNx, "Group 6 – Audio denoising using CleanUNet," 2025. [Online]. Available: https://github.com/HandleGod101/LipSinc_DNx/blob/main/Spring_2025_stuff/Group%206%20%20Audio%20denoising%20using%20CleanUNet%20(1).pdf