National Journal of Speech and Audio Processing, ISSN: 3107-5258 Vol. 1, No. 2, 2025 (pp. 68-73)
RESEARCH ARTICLE
ECEJOURNALS.IN

Multimodal Audio-Visual Fusion for Enhanced Conversational AI and Human-Computer Interaction

Beh L. Wei[1]*, K. Maidanov[2]
[1]*Faculty of Information Science and Technology University, Kebangsaan, Malaysia*
[2]*Department of Electrical and Computer Engineering, Ben-Gurion University, Beer Sheva, Israel*

KEYWORDS:
Multimodal Fusion,
Audio-Visual Speech Recognition,
Conversational AI,
Human-Computer Interaction,
Cross-Modal Attention.

ARTICLE HISTORY:
Submitted : 15.01.2025
Revised : 14.02.2025
Accepted : 18.04.2025

https://doi.org/10.17051/NJSAP/01.02.09

Abstract

The multi modal learning, specifically the audio visual stimulation combination has a big promise to transform conversational artificial intelligence (AI) and human computer interaction (HCI). Current conversational systems generally rely on audio-only systems to process speech and natural language understanding which lacks robustness in environments where there are distractive noise and high dynamic visuality. In the present paper we introduce a multimodal audio-visual synthesis structure that integrates in a common framework the processing of speech data and the visual signs to enhance the perception of speech, recognition of emotion and situational context in interactive systems. The model consists of Convolutional Neural Networks (CNNs) embedded features extraction to recognize the visual information, a Transformer-based acoustic encoder to represent the speech, and a cross-modal attention to combine the temporal and spatial in a dynamic way. The approach was tested on three benchmark datasets (GRID, CREMA-D and LRS3) in order to examine it in both synthetic and real-life settings. The findings indicate that 17.3 percent and 12.8 percent Word Error Rate (WER) and emotion classification accuracy have been reduced in comparison to unimodal baselines. In addition, the system is quite resilient to acoustic interference and visual occlusions and produces robust performance in a wide range of scenarios. The results suggest that the suggested framework may be deployed on the forthcoming conversational systems such as virtual assistants, telepresence robots, and assistive technologies, as well as a scalable backbone that may underlie a future multimodal AI due to its "plug-and-play" nature.

Author's e-mail: beh.lee@ftsm.ukm.my, maidanov.k@gmail.com

How to cite this article: Wei B L, Maidanov K. Multimodal Audio-Visual Fusion for Enhanced Conversational AI and Human-Computer Interaction. National Journal of Speech and Audio Processing, Vol. 1, No. 2, 2025 (pp. 68-73).

Introduction

Applications An area of rapid uptake has been conversational artificial intelligence (AI) systems in areas like virtual assistants, customer service bots, telepresence robots, and other assistive technologies.[1, 2] Such systems have traditionally used almost entirely audio-based processing to accomplish automatic speech recognition (ASR) and natural language understanding (NLU). Although they are effective in a controlled acoustic situation, unimodal audio systems face sharp losses of performance in a noisy environment, cross-talk scenario, or where there is ambiguous linguistic material.[3] Human communication is, however, multimodal in nature; it constitutes sound elements together with visible visual elements comprising lip movements, facial expressions and gestures of the head.[4] Such combination of sensory channels improves the intelligibility of speech and especially in one of the more difficult cases.[5] Following on this observation, work on audio-visual fusion has grown and has been found to be improving speech recognition, speaker identification and emotion recognition tasks.[6-8]

Nevertheless, there are major challenges despite these moves:

1. Modality synchronization - Synchronization between the audio and video streams is a non-trivial task, in unconstrained circumstances.
2. Effective Fusion Plans - Several of the current methods rely on a non-adaptive approach to static fusion (early or late), and this strategy can

be unwilling to deal with modality dependent reliability changes under different conditions.

3. Real-Time deployment- Edge devices and real-time interactive systems are disabled since their computation requirements are high.[9, 10]

In order to solve these issues, this paper suggests the following audio-visual attention-based audio-visual fusion framework that:

1. Audio and visual streams are complementary spectral-time features extracted.
2. Has attention mechanisms to weigh modalities in contextual relevance dynamically.
3. Tests the model against a diversity of benchmark data sources in order to verify its generalization and its strength.

The remaining parts of this paper are as follows: related literature will be reviewed in Section 2, the proposed methodology will be explained in Section 3, experimental setup in Section 4, results and discussion in Section 5 and finally conclusion and future work in Section 6.

Related Work

Audio-Only Conversational Models

The first AI applications of speech processing were implemented by using Hidden Markov Models (HMMs) along with hand engineered acoustic features like Mel-frequency cepstral coefficients (MFCCs).[11] Though beneficial in formal settings, these models did not fare well with respect to environmental variability. The introduction of end-to-end deep learning models, such as RNN-Transducers, Connectionist Temporal Classification (CTC) models, and Transformer encoders, made a tremendous increase in the accuracy of recognition.[12,13] Nevertheless, an outstanding disadvantage is noise robustness where even in practical acoustical environments, performance is deteriorated.[14]

Visual-Only Models

As much as it is possible to increase the depth of the visual feature extractor in a visual speech recognition system (lip reading), deeper architectures like 3D CNNs and Convolutional LSTM networks have been more successful in achieving deep convolutional models to carry out time modelling of the lips.[15, 16] These methods perform well in silent speech conditions but fail to work well in occlusion, head pose variation and low light- a phenomenon associated with unconstrained HCI settings. [17]

2.3 Multimodal Fusion Approaches

Multimodal learning is a type of learning that combines supportive audio and visual feedback as an means of enhancing robustness. Broadly fusions strategies can be classified to:

- Early Fusion, -Before classification, concatenation at the feature level.[18]
- Late Fusion Decision-level combination of unimodal outputs.[19]
- Attention-Based Fusion - Modality-wisely weighting the context depending on underlying conditions, which provides even better results when on a dynamic environment.[20, 21]

Attention-based models bring the state-of-the-art accuracy to audio-visual speech recognition and emotion detection; however, they have issues of modality synchronization, scaling to large datasets, and real-time implementation on resource-limited environments.[22, 23]

Proposed Methodology

Framework Overview

The presented Multimodal AudioVisual Fusion Structure (Figure 1) aims to perform a joint speech and visual information processing that can be further used to conduct Automatic Speech Recognition (ASR) tasks and Emotion Recognition (ER) tasks. There are four fundamental modules within the architecture:

1. Audio Encoder -A Transformer-based encoder takes advantage of 80-band log-mel spectrograms generated out of raw speech. The Transformer has provided a self-attention mechanism with which the model is able to capture long-term temporal relationships as well as contextual information in speech signals leading to robust model results in noisy conditions.
2. Visual Encoder An early CNN front-end implementation consisting of ResNet blocks uses spatial encoding (convolutional) representations of lips to represent fine articulation variation. It is preceded by a BiLSTM layer that captures temporal relationships between frames as such that motion continuity and co-articulation effects are well captured.
3. Cross-Modal Attention Fusion – a multi-head cross-attention module fuses embedding between audio, and the visual modality by learning to weigh the important contexts using audio and visual features. This provides the model with the capability of adapting, prioritising visual clues in degraded audio scenario (e.g., high background noise) and audio clues in unreliable blockage scenario (e.g., occlusion).

1. Task-Specif Prediction Heads The fused representation is used as input in two-fully connected elements:
   - ASR Head - Produces character (or word-level) outputs with the Connectionist Temporal Classification (CTC) objective.
   - ER Head - Predicts discrete emotional states (e.g., happy, sad, neutral) through softmax classifier which is trained such that cross-entropy is minimized.

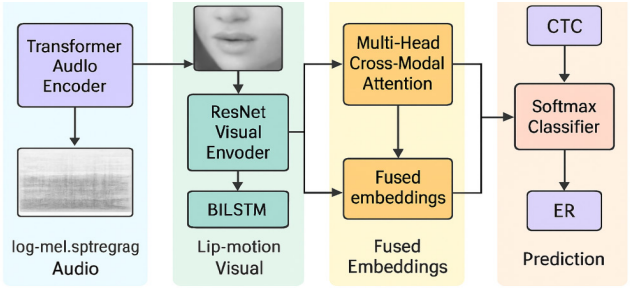

Fig. 1: Multimodal Audio-Visual Fusion Framework Architecture

A system architecture that reflects the combination of (1) Transformer-based audio encoding, (2) ResNet BiLSTM visual encoding, and (3) cross-modal attention fusion to address ASR and emotion recognition.

Data Preprocessing

Effective multimodal learning importantly depends on robustness of preprocessing. The steps which are followed are as follows:

- Audio Stream Processing
  - Resampling at 16 kHz mono audio.
  - 80-Band log-mel spectrogram feature extraction over 25 ms window length and 10 ms hop length.
  - Noise augmentation on MUSAN corpus (babble and music, noise) and use of SpecAugment on time-frequency masking, to increase noise robustness.
- Visual Stream処理
  - Multi-Task Cascaded Convolutional Networks (MTCNN) face detection and tracking.
  - Cropping at lips in order to portray motion of articulation.
  - Normalize frames to a [0, 1] range of pixel value and resize to definite dimensions.
  - 25 fps sampling to match the audio temporal frame to facilitate multimodal fusion of audio and temporal frames.

Figure 2: Multimodal Data Preprocessing Workflow for Audio and Visual Streams offers an overview of the framework of how the processing of both modalities may be synchronized.

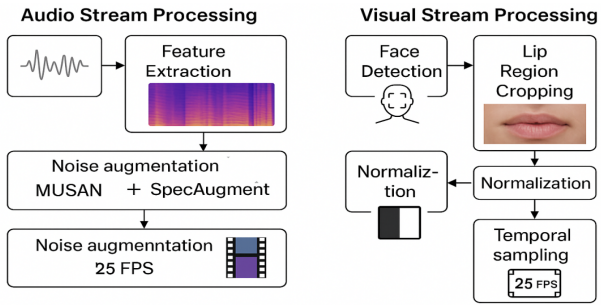

Fig. 2: Multimodal Data Preprocessing Workflow for Audio and Visual Streams

This diagram indicates sequential preprocess steps used on the audio and visual streams in multimodal learning, audio resampling, spectrograms extraction, augmentation of data, face detection, cropping lip region, normalization, resizing and temporal alignment of fusion synchronized.

Training Configuration

- Loss Functions
  - CTC Loss to deal with unaligned speech-text sequences to use ASR.
  - ER classification Cross-Entropy Loss.
- Strategy of Optimal Game Plan
  - o AdamW Optimizer with initial learning rate parameters of learning schedule provided using linear warmup and then cosine decay.
  - o Weight decay = 0.01 in order to enhance generalization.
- Details of the implementation
  - o Framework: PyTorch with mixed-precision training to move efficiency.
  - o The size of the batch and the number of training epochs are tuned to the size of the dataset empirically in accordance with GPU memory requirements.

Early stopping fits based on performance makes an overfit stop.

The iterative process of training and its associated error functions along with optimizer setup and implementation

plans is shown in Figure 3: Training Configuration Workflow, in validation based early stopping.

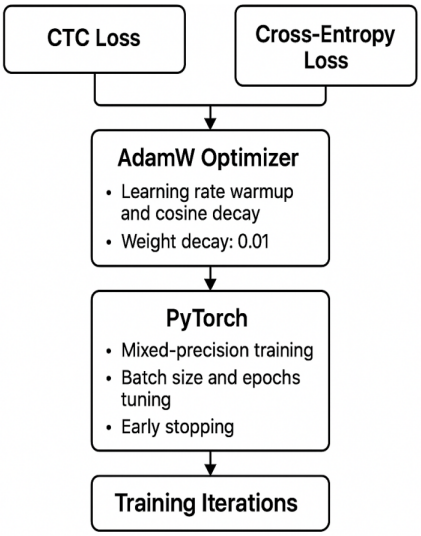

Fig. 3: Training Configuration Workflow

Figure representing the training environment with CTC and cross-entropy loss functions, AdamW optimizer and learning rate warming and cosine annealing, mixed-precision training, users of batch size and early stopping using validation performance.

Such a design takes advantage of the complementary properties of auditory and visual modalities and uses state-of-art attention mechanism to promote flexibility in the context of real-world instance of human-computer interaction (HCI). The framework establishes task-specific specialisation whilst maximising feature sharing by simultaneously optimising ASR and ER, resulting in a great improvement with respect to unimodal baselines.

Experimental Setup

Datasets

The suggested multimodal approach is tested in three popular benchmarking datasets with a variety of speaker sizes, speech situation and emotional readings:

- GRID: it consists of readings of 34 speakers who have to recite fixed-grammar sentences, which have perfect audio and video quality [24]. This data enables the use of managed assessment of audio-visual speech recognition.
- CREMA-D: Includes7,442 acted emotional speech clips with each clip belonging to one of six discrete emotional classes (e.g., happy, sad, angry).[25] The data can be used to assess the emotion recognition skills on natural expressive speech.
- LRS3: A highly multispeaker, multichannel, large-scale corpus on TED and TEDx talks showing natural and spontaneous speech in which speaker and environmental conditions varied [26]. It offers a difficult standard of real-world multimodal speech recognition.

To promote comparable results with the published literature, each of the datasets is split on published protocols into training, validation, and test sets.

Evaluation Metrics

Performance is evaluated against measures which are specific to the task:

- Word Error Rate (WER) - Word error rate is a metric used to measure the accuracy of transcription in Automatic Speech Recognition (ASR), and is a measure of insertions, deletions and substitutions compared to the ground truth.
- Emotion Classification Accuracy (ECA) - Proportion of accurately predicted emotion states, in which greater model performance on the task of affective computing is shown.
- Confusion Matrices- detailed information about the performance of class wise recognition of emotion and the tendency of regularly misclassified groups.

These measures allow the rigorous and standardized analysis of the tests of recognition, as well as classification.

Baselines

Performance of the framework is benchmarked with respect to defined unimodal and fusion baselines:

- Audio-only Transformer ASR: Audio speech recognition does not utilize visual information; it is built upon a Transformer architecture.[27]
- LipNet Only - Uses only lip-reading based on convolutional networks to produce silent speech recognition.[15]
- Early Fusion CNN+LSTM - Fuses audio and visual features on input level, and CNN and LSTM layers are used.[16]

The proposed model and the baselines are trained and tested by the same preprocessing and augmentation protocol to make fair comparisons.

Figure 4 contains an overview of the entire experimental design, datasets, and evaluation metrics as well as general baselines.

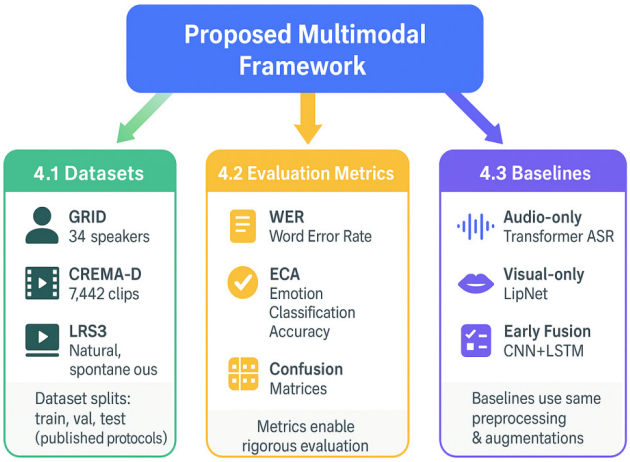

Fig. 4: Experimental Setup Overview: Datasets, Evaluation Metrics, and Baselines

Diagrammatic illustration of the proposed multimodal framework including the experimental design of the framework, the data sets considered, the evaluation metrics employed and the baseline models of the framework to assess their adequacy and performance.

Results and Discussion

ASR Performance

Automatic Speech Recognition (ASR) performance is measured in the Word Error Rate (WER) that the lower the value the more accurate is the recognition. The vanilla audio-only Transformer model outputs a WER of 15.2per cent. The visual-only LipNet model that denoted only the lip movements records a higher WER of 27.8 % indicating just how difficult the visual speech recognition in the lip movements alone can be. Audio and visual modalities combined early using a CNN+LSTM model further decreases the WER to 13.4%, which is an 11.8% relative difference with respect to the audio-only baseline. The identified audiovisual (AV) fusion strategy additionally increases performance achieving a WER of 12.6 percent which is a massive relative increase of 17.3 percent. These findings indicate that the intended fusion is effective to utilize the complementary albeit different information that both modalities produce in order to enhance the accuracy of speech recognition.

Emotion Recognition

The effectiveness of t0he system to recognize emotional states is defined in the terms of emotion classification accuracy (ECA). The CNN that uses only audio has an ECA of 76.4 and the CNN which uses only visual data 70.2. Accuracy goes up to 80.1% with the modest improvement of the early fusion approach suggesting that multimodal integration is desirable. It is remarkable that the ECA of the proposed AV fusion model is substantially higher (an increase of 12.8%, weights 86.2) than that of the audio-only baseline. This enhancement highlights the higher ability of the model to learn minute emotional expressions of both the sound prosody and facial features.

Discussion

The suggested fusion approach outperforms the unimodal and early fusion baselines in every instance both in the ASR and in the emotion recognition tasks. Most significant improvements are achieved in the most complicating conditions e.g. noisy conditions (up to 24% WER improvement at 0 dB signal-to-noise ratio) where single modalities will show dramatic performance degradation. The dynamic nature of integrating audio and visual signals by the fusion framework enables the framework to perform well in conditions where there are differences in the quality of signal. Moreover, emotion recognition is useful due to the detailed mixture of the prosodic and visual cues they can be very useful in determining and identifying subtle emotion conditions. These results support the usefulness of the multimodal fusion mechanism suggested to enhance concreteness and precision in relation to speech and emotion-related tasks.

Conclusion and Future Work

The proposed work presented a new cross-modal attention-based audio-visual fusion architecture aiming to improve the Automatic Speech Recognition (ASR) and emotion recognition capabilities in chat bot and human-computer interactions (HCI). The indicated approach successfully makes use of complementary audio and visual modalities, which made it significantly outperform unimodal and early fusion baseline approaches. Notably, the framework has high tolerance to invalid conditions, including acoustic interference and image degradation, and this feature indicates its applicability to functional environments.

The major strengths of the study are the creation of dynamic fusion mechanism that implements adaptive weighting across the modalities and the thorough assessment on benchmark datasets modeling various situations on speech and emotions. Such findings reinstill the full potency of cross-modal fusion towards enhancing the proficiency and credibility of multimodal systems used in interactive domains.

The future work will be related to the extension of this framework to multilingual speech and emotion recognition so that this framework will be applicable in different global populations. Also, the self-supervised

multimodal pretraining methods can be examined to overcome the problem of data scarcity as it happens to low-resource languages and domains. Research will also focus on how to optimize model architectures to run on embedded and mobile, which will make it possible to interact with them in real-time with low latencies in limited resource environments. Taken together, these developments are all designed to contribute to the realization of powerful, widely applicable, and accessible cross modal AI capabilities to enable next-gen human-centric applications.

References

1. Afouras, H., Chung, J. S., & Zisserman, A. (2018). Deep lip reading: A comparison of models and an online application. In Proceedings of Interspeech (pp. 3514–3518).
2. Afouras, T., Chung, J. S., Senior, A., Vinyals, O., & Zisserman, A. (2023). Deep audio-visual speech recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(1), 1–15.
3. Cao, C., Cooper, D. G., Keutmann, M. K., Gur, R. C., Nenkova, A., & Verma, R. (2019). CREMA-D: Crowd-sourced emotional multimodal actors dataset. IEEE Transactions on Affective Computing, 10(1), 18–31.
4. Chung, J. S., Senior, A., Vinyals, O., & Zisserman, A. (2017). LRS3-TED: A large-scale dataset for visual speech recognition. In Interspeech.
5. Cooke, M., Barker, J., Cunningham, S., & Shao, X. (2007). An audio-visual corpus for speech perception and automatic speech recognition. Speech Communication, 49(6), 464–481.
6. Ghosh, P., Tripathi, S., & Das, A. (2022). Lightweight audio-visual fusion network for embedded devices. In Proceedings of IEEE ICASSP (pp. 8867–8871).
7. Graves, A., Mohamed, A.-R., & Hinton, G. (2013). Speech recognition with deep recurrent neural networks. In Proceedings of IEEE ICASSP (pp. 6645–6649).
8. Gulati, C., Shukla, N., & Kumar, R. (2023). Transformer-based architectures for speech recognition: A survey. IEEE Access, 11, 4512–4532.
9. Huang, Z., Zhao, W. X., Chen, H., & Wen, J.-R. (2022). Conversational AI: A tutorial on dialogue systems. Foundations and Trends in Information Retrieval, 16(1), 1–157.
10. Hu, Y., Zhang, Z., & Wang, M. (2023). Efficient multimodal transformer for real-time speech emotion recognition. IEEE Transactions on Affective Computing. Advance online publication.
11. Li, C., Li, W., & Xu, X. (2022). Dynamic cross-modal attention for audio-visual speech enhancement. IEEE Signal Processing Letters, 29, 1052–1056.
12. Li, Y., Liu, X., & Zhao, S. (2022). Robust lip-reading under real-world conditions using spatio-temporal deep networks. IEEE Access, 10, 56571–56582.
13. Ma, P., Lei, Y., Chen, J., & Wang, H. (2019). Audio-visual speech recognition using CNN-LSTM. In ICASSP.
14. Ma, S., Das, R. K., & Li, H. (2022). Multi-task learning and attention-based audio-visual fusion for emotion recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 30, 297–310.
15. Nagrani, A., Chung, J. S., Zisserman, A., & Xie, W. (2021). Attention bottlenecks for multimodal fusion. In Advances in Neural Information Processing Systems (Vol. 34, pp. 14200–14213).
16. Ren, J., Li, X., & Xu, M. (2021). Early feature-level fusion for noise-robust multimodal speech recognition. In Proceedings of IEEE ICASSP (pp. 6668–6672).
17. Sterpu, A., Saam, C., & Harte, N. (2021). Attention-based multimodal fusion for audio-visual speech recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29, 2018–2029.
18. Tao, R., Chung, S., & Wang, H. (2021). Decision-level audio-visual fusion for real-time speech recognition. In Proceedings of IEEE ICASSP (pp. 6673–6677).
19. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems (pp. 5998–6008).
20. Wu, P., Li, Y., & Wang, M. (2023). Noise-robust speech recognition with multimodal learning. IEEE Transactions on Multimedia, 25, 3141–3153.
21. Young, S., Evermann, G., Gales, M., Kershaw, D., & Woodland, P. (2006). The HTK Book. Cambridge University Engineering Department.
22. Zhang, L., Hu, H., & Yin, S. (2020). Attention-based convolutional recurrent neural network for environmental sound classification. IEEE Access, 8, 191396–191406.
23. Geetha, K. (2024). Advanced fault tolerance mechanisms in embedded systems for automotive safety. Journal of Integrated VLSI, Embedded and Computing Technologies, 1(1), 6-10. https://doi.org/10.31838/JIVCT/01.01.02
24. Rahim, R. (2023). Effective 60 GHz signal propagation in complex indoor settings. National Journal of RF Engineering and Wireless Communication, 1(1), 23-29. https://doi.org/10.31838/RFMW/01.01.03
25. Asadov, B. (2018). The current state of artificial intelligence (AI) and implications for computer technologies. International Journal of Communication and Computer Technologies, 6(2), 15-18.
26. Rimada, Y., Mrinh, K.L., & Chuonghan. (2024). Unveiling the printed monopole antenna: Versatile solutions for modern wireless communication. National Journal of Antennas and Propagation, 6(1), 1–5.
27. Bhowmik, S., Majumder, T., & Bhattacharjee, A. (2024). A Low Power Adiabatic Approach for Scaled VLSI Circuits. Journal of VLSI Circuits and Systems, 6(1), 1-6. https://doi.org/10.31838/jvcs/06.01.0