Context-Aware Audio-Visual Speech Enhancement Based on Neuro-Fuzzy Modelling and User Preference Learning

Context-Aware Audio-Visual Speech Enhancement Based on Neuro-Fuzzy Modelling and User Preference Learning

Song Chen; Jasper Kirton-Wingate; Faiyaz Doctor; Usama Arshad; Kia Dashtipour; Mandar Gogate; Zahid Halim; Ahmed Al-Dubai; Tughrul Arslan; Amir Hussain

***Abstract*— It is estimated that by 2050 approximately one in ten individuals globally will experience disabling hearing impairment. In the presence of everyday reverberant noise, a substantial proportion of individual users encounter challenges in speech comprehension. This study introduces a novel application of neuro-fuzzy modelling that synergizes and fuses audio-visual speech enhancement (AV SE) with an initial user preference learning based framework. Specifically, our approach uniquely integrates multimodal AV speech data with innovative SE methods and fuzzy inferencing techniques. This integration is further enriched by incorporating a user-preference learning model that adapts to environmental and user-specific contexts, including signal-to-noise ratios, sound power, and the quality of visual information. The proposed framework facilitates the incorporation of clinical measures such as user cognitive load (or listening effort) with real-world uncertainty to steer the system outputs. We employ an adaptive fuzzy neural network to derive the most effective Sugeno fuzzy inference model, employing particle swarm optimization to ensure optimal SE by considering sound power, ambient noise levels, and visual quality. Experimental results utilise our new benchmark AV multi-talker Challenge dataset to demonstrate the superiority of our user preference-informed, context-aware AV SE approach in enhancing speech intelligibility and quality in challenging noisy conditions, marking a significant advancement over conventional methods while reducing energy consumption. The conclusion supports the ecological scalability of our approach and its potential for real-world applications, setting a new benchmark in AV SE research, paving the way for future assistive hearing and communication technologies.**

***Index Terms*— Fuzzy inference, deep neural networks, preference learning, speech enhancement.**

Song Chen is with College of Mechanical and Electrical Engineering, Anhui Jianzhu University, China (e-mail: chensong_aju@ahjzu.edu.cn)

Jasper Kirton-Wingate; Kia Dashtipour; Mandar Gogate; Ahmed Al-dubai; Amir Hussain are with School of Computing, Engineering and the Built Environment, Edinburgh Napier University, Edinburgh, United Kingdom (e-mail: jasper.kirton-wingate@napier.ac.uk, K.Dashtipour@napier.ac.uk, M.Gogate@napier.ac.uk, A.Al-Dubai@napier.ac.uk, A.Hussain@napier.ac.uk).

Faiyaz Doctor is with School of Computer Science and Electronic Engineering, University of Essex, Colchester, United Kingdom (e-mail: fdocto@essex.ac.uk).

Usama Arshad; Zahid Halim are with the Faculty of Computer Science & Engineering, Ghulam Ishaq Khan Institute, Topi, KPK, Pakistan (e-mail: usamajanjua9@gmail.com, zahid.halim@giki.edu.pk).

Tughrul Arslan is with School of Engineering, The University of Edinburgh, Edinburgh, United Kingdom (e-mail: T.Arslan@ed.ac.uk).

I. INTRODUCTION

In the realm of modern communication, speech enhancement (SE) plays an important role in improving speech clarity and intelligibility within noisy environments [1]. The effectiveness of SE technologies is impacted by a number of challenges, mainly the difficulty of accurately separating speech from background noise without distorting the speech signal itself. Many conventional methods rely on assumptions about the noise characteristics that may not apply in all real-world scenarios. While recent deep learning (DL) based audio-visual (AV) SE techniques have made substantial improvements in this area, they often struggle to adapt to the dynamic nature of communication scenarios, where speech signals are excessively masked in the audio and/or visual domain. This results in a further compromise between noise reduction and speech distortion [2, 3]. In light of this, recent studies have proposed generative modelling approaches [4], [5]. However, these approaches still face significant computational complexity constraints, which makes them impractical for real-time Hearing Aids (HA) or other devices with limited processing power. Furthermore, these methods may not fully consider listener's preferences for the noise reduction within the specific context of communication [6]. Additionally, HA devices have limited battery capacity, which requires power sensitive optimisation of SE processing. Overall, while SE is essential for clear communication in a noisy world, overcoming the limitations of traditional and modern techniques requires innovative approaches that are adaptable, efficient, and responsive to the needs of diverse users and situations [7].

In the SE literature, AV and audio-only (A-only) based methods advocate significant differences in their approach and outcome. AV enhancement leverages both sound and visual cues, such as lip movements, to improve speech intelligibility. These methods have been shown to result in superior speech clarity compared to A-only approaches, especially in challenging noisy environments or when the audio signal is weak [8]. However, the integration of visual data requires more complex processing algorithms, which can increase the time and energy required for enhancement. A-only enhancement, on the other hand, relies solely on auditory signals. While it may be less effective in isolating speech from noise compared to AV methods, its simpler processing demands make it faster and more energy-efficient. This trade-off between enhancement quality and resource efficiency is crucial in applications where

processing power or battery life is limited [9]. The choice between AV and A-only enhancement methods thus depends on the specific requirements of the target application, including the need for speed, energy conservation, and the level of speech clarity desired.

Adaptive SE strategies in speech and hearing technologies are essential due to the diverse needs of users and the dynamic environments in which they communicate [10]. Traditional approaches often prove inadequate as they fail to consider individual listener preferences and specific contextual challenges. By incorporating user preferences and environmental context, adaptive strategies can customise SE to the specific situation and individual: thereby optimising battery life and overall sound quality while improving clarity and comprehension. This personalization ensures a more effective communication experience, overcoming the limitations of standard methods that may not suit all scenarios or meet the needs of individual users [11].

Fuzzy inference systems (FIS) offer a robust framework for decision-making in environments characterised by uncertainty and imprecision, which are inherent in ecologically valid SE scenarios [12]. Historically, FIS has been applied to SE tasks to effectively manage the variability and ambiguity inherent in real-world audio signals [13][14]. The advantage of employing FIS lies in its inherent flexibility and ability to dynamically adapt to complex, uncertain inputs [15]. In this first of its kind study, we explore the significance of considering user preferences and environmental context within our proposed FIS based AV SE framework for future hearing assistive technologies.

A. Objectives of the study

By understanding factors such as sound power, noise levels, and video quality, adaptive systems can be tailored to personalise the user experience. These elements are crucial for developing a system that not only enhances speech but also aligns with the user's environment and personal preferences. This paper introduces an adaptive neuro-fuzzy model for AV SE that integrates user preference learning for improving speech intelligibility and quality for hearing-impaired users. This entails creating a system capable of adjusting its behaviour based on the specific needs and conditions of its user, such as adjusting for background noise or focusing on visual cues when audio quality is poor. Through these adaptations, the proposed model aims to deliver a more effective and personalised communication experience across diverse scenarios. This paper makes the following key contributions:

- Develop a cutting-edge neuro-fuzzy model integrating AV SE with user preference learning for personalised, context-aware communication improvements.
- Implement fuzzy logic to dynamically adapt SE strategies based on user-specific contexts, including sound power and environmental noise, marking a significant advancement in adaptable communication technologies.
- Utilise an initial predefined set of user preferences in order to respond to individual needs, enhancing user experience in diverse auditory environments. The preference informed rules can be further substantiated using cognitive load (or listening effort) data from clinical speech-in-noise tests, using e.g. pupillometry sensing [16].
- Simulations and results clearly represent the effectiveness of the proposed model in contrast to relevant traditional models. Tests comparing different fuzzy models show that the Sugeno fuzzy model, when trained with our adaptive fuzzy neural network, provides the best balance between SE quality and system efficiency. It fine-tunes the enhancement based on real-world conditions like noise level and video quality. The results also highlight the system's ability to reduce noise effectively while conserving energy. This means our model not only makes conversations clearer but also works well on devices with limited battery life, like portable hearing aids or smartphones.
- By analysing performance evaluation metrics such as STOI, PESQ, MOS and power consumption (W), we demonstrate that our proposed fuzzy system based approach leads to more adaptive and efficient SE across a variety of noisy recordings. This confirms our model's potential to improve communication, especially for those with hearing challenges.

The rest of this paper is structured as follows. Section II provides a comprehensive literature review of recent developments in AV SE focusing on the application of state-of-the-art DL techniques. In section III we introduce the novel neuro-fuzzy modelling approach that combines AV SE with an initial user preference framework for hearing assistive devices. Section IV details the simulation setup and hardware design of the proposed AV SE system where its performance is evaluated using benchmark multitalker data from an international Challenge dataset. This evaluation utilises both subjective and objective metrics and includes comparison with state of the art approaches. Finally, section V provides conclusions and future work.

II. LITERATURE REVIEW

The emerging field of AV SE has witnessed significant advancements, driven by the integration of multimodal representation learning, self-supervised machine learning, and innovative feature extraction techniques. A notable contribution in this domain is from the recent work reported in [17], which developed an AV canonical-correlated Graph Neural Network model. This model distinguishes itself through its capacity to optimise the canonical association between audio and visual data, presenting a more context-aware approach to SE. Self-supervised machine learning has emerged as a promising approach in speech processing, leveraging temporal data tracking and multimodal information fusion [18]. This approach facilitates the learning of temporal speech dynamics and fusion of AV cues without relying on manually annotated labelled datasets. These methods have demonstrated effectiveness in identifying and isolating speech signals from noisy backgrounds. By deriving Ideal Binary Masks, these approaches can selectively enhance the speech component of an audio signal, significantly improving clarity and intelligibility [19].

Through the integration of AV data, machine learning, and advanced signal processing, researchers are advancing the development of sophisticated and user-friendly hearing and communication aids. These innovations promise to enhance

the listening experience of users in diverse settings, ranging from private to public spaces, thereby increasing accessibility to clear and intelligible speech for all individuals, particularly those with hearing loss.

Recent advancements in deep neural networks (DNNs) have led to the development of more sophisticated A-only and AV based SE algorithms that can further improve the speech clarity and intelligibility. In the context of A-only based SE advancements, techniques such as sparse low-rank decomposition combined with multi-objective DNNs trained on acoustic features have been pivotal in achieving accurate phase and magnitude approximation [20]. These methods refine speech signals, ensuring that the enhanced speech is clearer and more intelligible. Additionally, the integration of band-pass and Wiener filters has been explored to further enhance noise reduction capabilities, offering more comprehensive approaches to improve speech quality [21]. More recently, an efficient recurrent DNN based architecture has been reported that can enhance both speech enhancement and recognition [22].

Despite the notable advancements above, a gap exists in the adaptability of SE systems to dynamically accommodate user preferences across varying environmental conditions. Furthermore, traditional SE algorithms often exhibit limited flexibility to contextually adjust their noise reduction strategies in real-time, which can compromise their effectiveness in dynamic settings. To address these challenges, some researchers have explored the application of fuzzy logic, inspired by cognitive models, to create more adaptable and context-aware SE systems [12].

Our preliminary approach [23] and others too have explored the integration of fuzzy logic in multimodal SE applications, which represented a significant step towards creation of AV SE devices that are not only more intelligent in their processing capabilities but also more attuned to the complexities of human communication and environmental variability. In our other related pilot work, we demonstrated the potential of SE technologies to offer a personalised listening experience [6]. This proof of concept study underscored the importance of user-centred design in the development of future SE technologies that cater to the diverse needs of their users in a highly personalised manner [10].

In this paper, we explore the novel integration of fuzzy neural network based user preference learning within SE systems to deliver more personalised technologies. A classical approach in fuzzy systems research involves the use of geometric functions to define fuzzy membership for inputs, allowing for a more responsive adaptation to gradual changes in the environment. This method can enhance the precision of SE by ensuring the system output becomes more sensitive to subtle variations in input values. Here, we propose the application of genetic optimization libraries and density-based fuzzy rule interpolation methods [24][25] that can allow for customization of SE systems to individual user profiles, considering their unique preferences and environmental contexts [10][11].

Recent research on neuro-fuzzy models has explored the extraction of fuzzy rule-based systems from trained artificial neural networks for classification purposes [26]. These methods offer a way to leverage the power of neural networks while maintaining the interpretability and adaptability of fuzzy rule-based systems [27][28]. Such approaches have been optimised for pattern classification, enhancing system capability to differentiate speech and noise under varying conditions [13][14]. However, the customization of fuzzy neural network-based approaches to meet individual comfort levels remains an ongoing challenge [7], requiring careful consideration of a wide range of contextual input data. This often involves the acquisition of extensive, carefully labelled datasets in order to train systems effectively.

We hypothesis that the integration of neuro-fuzzy and DNN based AV SE techniques with clinical user data, such as cognitive load or listening effort, which refers to the mental effort required to process information, could lead to a significant step forward in creating SE systems that are not only more effective but also more aligned with the needs and preferences of individual users [10].

In summary, this paper proposes a novel framework that optimises a neuro-fuzzy based AV SE model towards an initial set of linguistically defined user preferences that are intuitively informed by cognitive load or listening effort that is typically experienced by users in noisy environments. This model then acts as a personalisation scheme for an individual with the capability of further generalising towards fully personalizable multimodal hearing-aids and communication systems of tomorrow.

III. PROPOSED MODEL

Although there are several SE methods that have demonstrated good results in noisy scenarios, current state-of-the-art approaches have not effectively addressed the challenge of personalisation and adaptability of SE systems across diverse real-world environments. This entails multi-objective multi-constraint optimisation of SE systems to minimise power efficiency, enhance user comfort through personalising noise reduction settings and managing associated trade-offs (e.g. between intelligibility, power consumption and system latency) to meet individual user preferences. Here, we attempt to address this formidable challenge and additionally consider the smoothness of switching between multiple SE models for optimised functioning in challenging noisy environments.

Specifically, we propose a novel application of neuro-fuzzy modelling for personalised AV speech processing. Our proposed AV SE model aims to significantly improve speech clarity and intelligibility by contextually leveraging audio and visual information in a context aware manner. This involves dynamic adaptation to the listener's environment and preferences and utilisation of neuro-fuzzy modelling to optimise SE processing. The proposed system is designed to be adaptable and capable of learning from user feedback to enhance the listening experience.

This section presents the fundamental principles of the fuzzy inference model for our proposed context-aware AV SE system. This includes a description of fuzzy neural network integration and the training and optimization stages.

A. Fuzzy Inference Model Foundations

1) Sugeno Fuzzy Model Characteristics: The Sugeno fuzzy model, known for its efficiency in computational applications, differs notably from its counterparts by defining the output membership function as either 'constant' or 'linear' in

Equation (1). This distinction facilitates precise quantity outputs, making the model especially well suited for applications in engineering and control systems.

$$f(x) = \sum_{i=1}^{n} w_i \cdot f_i(x) \quad (1)$$

where w_i are the weights, and $f_i(x)$ are the output membership functions.

2) Gaussian Function Application: The Gaussian membership function is utilised within the model to handle the fuzziness of input variables. It is defined in Equation (2) as:

$$G(x) = ae^{-\frac{(x-b)^2}{2c^2}} \quad (2)$$

where x denotes the input variable, ae represents the peak of the Gaussian function. b signifies the peak position of the function, and c denotes the standard deviation that controls the width of the Gaussian bell curve.

B. Principles of Fuzzy Inference Systems

1) Completeness: Completeness ensures that for any input value, there exists at least one fuzzy rule that can be applied, forming the basis for a functional fuzzy inference system.

2) Intersectionality: To maintain continuity and smoothness in the input-output relationship of the system, adjacent fuzzy rules exhibit intersectionality. This overlap signifies the conceptual ambiguity and enhances the system's robustness through Equation (3):

$$Intersectionality = \mu_{A\cap B}(x) \quad (3)$$

where $\mu_{A\cap B}(x)$ represents the degree of membership of element x in the intersection of fuzzy sets A and B.

3) Barycentric Method: The barycentric method, a common defuzzification strategy and computes a crisp value from a fuzzy set as shown in Equation (4):

$$y^* = \frac{\sum_{i=1}^{n} u_i y_i}{\sum_{i=1}^{n} u_i} \quad (4)$$

where y^* is the crisp value, and u_i is the membership function of the fuzzy set.

C. Input and Output Variables

The input and output variables used by the fuzzy inference process are summarised in Table I. The fuzzy input variables are defined as sound power, SNR, and visual quality, which are derived from real-world observations or recorded audio and video data. Sound power is quantified as the ratio of the sound energy to the maximum volume scaled by 10 (ranging from 0 to 10) and categorised into linguistic fuzzy sets of high, medium and low. SNR is the ratio of signal power to noise power. SNR levels are objectively assessed and based within a range from -20 to 20 dB, and partitioned into High, Medium and Very Low linguistic terms. Finally, the Visual quality is evaluated based on image sharpness and clarity under various lighting conditions measured in the range (0 to 800). This is partitioned into '*Good*', '*Poor*' and '*Very Poor*' fuzzy sets.

Note that centre values of '*High*', '*Medium*' and '*Low*' ranges are shown for the input variables in Table I. Specifically, these High, Medium and Low are initially set to a vague interval, and after training, a definite intermediate value can be obtained that define centres of the gaussian fuzzy membership functions representing these linguistic quantifiers. The range determined by the Gaussian function is infinite however, defined by their standard deviation parameters where the support of each fuzzy set can be approximated to 3 × standard deviation parameter. Only intermediate values are fixed. The fuzzy neural network is obtained from the training approach as described in Section 3E.

The output variable from the fuzzy inference process represents SE processing decisions as the following three constants: AV, A-only and No Enhancement. Both input and output variables are shown in the user preference-informed rule base in Table I, which aim to maximise the speech quality and intelligibility for users. For example, for the case of 'good' quality visual inputs, 'low' sound power and 'high' environmental SNR detection, the input speech signal is passed through the fuzzy inference based AV SE system with 'no enhancement' to avoid distorting the naturalness of speech for the listener. Conversely, for the case of a 'very low' SNR, 'high' input sound power and 'good' quality visual input, the energy-consuming AV SE mode is selected to maximise intelligibility gain and reduce user cognitive load.

The total sound power (Equation 5), environmental SNR (Equations 6,7,8), and visual image quality/sharpness Equation (9) are calculated from the input signals.

$$energy_{mix} = \frac{\sqrt{\sum_{i=1}^{n} mix_i^2}}{energy_{max}} \times 10 \quad (5)$$

$$energy_S = \sqrt{\sum_{i=1}^{n} Signal_i^2} \quad (6)$$

$$energy_n = \sqrt{\sum_{i=1}^{n} Noise_i^2} \quad (7)$$

$$SNR = 10 * \log_{10}\left(\frac{energy_S}{energy_n}\right) \quad (8)$$

During test-time for SE processing in hearing assistive devices, we do not have access to the speech signal, therefore a non-intrusive SNR method via Deep Learning such as in [6] can be integrated into a full SE system for ecologically valid inference.

When processing images in video, the horizontal and vertical gradient values are used to obtain the edge strength, which is expressed in Equation (9) as follows:

$$G(x,y) = \sqrt{G_x^2(x,y) + G_y^2(x,y)} \quad (9)$$

Table I categorises SE strategies based on the dynamic interaction between sound power, SNR, and visual quality and the initial median values are given. With higher sound and noise levels, and good quality visual inputs (where the camera can capture the speaker's lips) the system tends to use AV enhancement. On the other hand, for the case of low sound and noise levels, and poor quality visuals, the system tends not to perform SE. Lighting conditions and camera obstacles determine visual quality. The input fuzzy sets are initialised with a set of prior parameters which following training, are optimised.

These decisions were substantiated through our new benchmark 'in-the-wild' AV SE Challenge (AVSEC) dataset [29]. This showcases highly challenging real-world application scenarios across over 3000 multi-talker TED and TEDx talk videos, and underlines our proposed system's adaptability to varied AV contexts to improve speech intelligibility. Each video lasts 3-30 seconds, with 40% female and 60% male speakers. The SNR is varied between -20 to 20 dB.

TABLE I
INPUT VARIABLES (WITH MEDIAN VALUES) AND OUTPUT VARIABLES

Input			Output
Sound Power (0~10)	SNR (−20~20dB)	Visual Quality (0~800)	SE Output
High (8)	Low (-10 dB)	Good (562)	AV
Medium (5)	Medium (0 dB)	Poor (360)	A-only
Low (2)	High (10 dB)	Very Poor (157)	No Enhancement

Our innovative neuro-fuzzy modeling based approach ensures that the SE is both flexible and responsive to varying environmental conditions and user needs. The SE system's input management is determined by the fuzzy inference module's output, categorising it into three distinct modes: no enhancement, A-only input, and combined AV input, in Figure 1. Inputs x, y and z representing sound power, noise level (or SNR), and image quality respectively are analysed through the fuzzy inference system to decide the appropriate SE strategy based on the initial linguistically informed rules described in Table I. Following this selection, we assess the system's effectiveness by calculating perceptual evaluation metrics like PESQ (Perceptual Evaluation of Speech Quality) and STOI (Short-Time Objective Intelligibility), ensuring the chosen mode optimally enhances speech intelligibility and quality. This is illustrated in our proposed fuzzy-inference based context-aware AV SE framework in Figure 2, which utilises our baseline, STOI-loss optimised AV SE model [30].

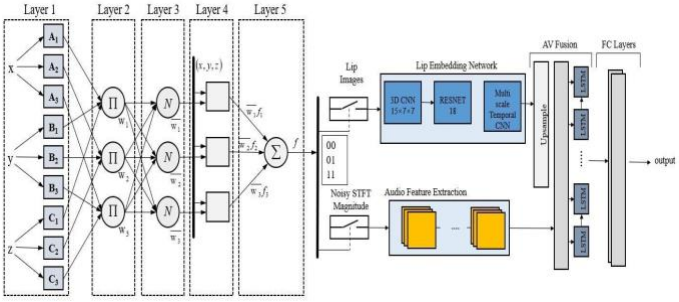


Fig. 1. Proposed Fuzzy-inference based AV SE model structure

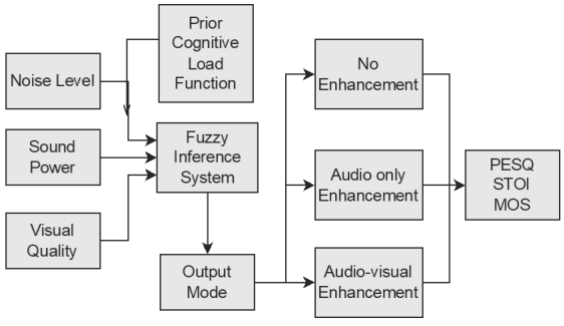


Fig. 2. Proposed fuzzy-inference based context-aware AV SE framework

D. Proposed Speech Enhancement Algorithm

1) Audio-Visual Speech Enhancement (AV SE) Model: The proposed model is built upon our widely used baseline AV SE methods reported in [30][9] that integrate AV cues, to complement conventionally used audio-only cues, in order to further enhance speech intelligibility. The AV SE model utilises audio features and a *lip-embedding network* alongside *feature fusion techniques* to improve the quality of speech signals. The model is pre-trained on over 34,000 AVSEC Challenge videos, totalling over 113 hours, with 605 target speakers (see [29] for details). It processes audio and visual data through a series of convolutional and temporal layers to extract meaningful features for SE, as shown in the right half of Figure 1. The number and size of filters for the first four convolutional layers are 64 and 5×5, and the number and size of filters for the fifth convolutional layer are 4 and 1×1. More details can be found in [30].

2) Lip-Embedding Network and Feature Fusion: The lip-embedding network employs a 3D-convolutional approach, utilising RESNET-18 and temporal CNN layers to capture dynamic visual cues from lip movements. These visual features are sampled at 25 frames per second, while audio features are processed at a rate of 75 vectors per second. The fusion of acoustic and visual features is achieved through an LSTM layer, optimising the SE process.

3) Voice Synthesis: The AV SE model predicts noise spectra and time-frequency binary masks after the lip is embedded in the input network. The output enhanced speech is obtained by multiplying the predicted amplitude mask with the noise amplitude spectrum.

E. Fuzzy Neural Network Integration for AV SE

We propose the integration of fuzzy logic with neural networks, termed an Adaptive Neuro-Fuzzy Inference System (ANFIS), illustrated in the left half of Figure 1. This leverages the learning capabilities of neural networks and the reasoning capabilities of fuzzy logic to handle complex and uncertain environments in the context of AV SE. The system aims to learn optimal parameters for the Gaussian membership functions (median and standard deviation of the Gaussian fuzzy sets) and Sugeno fuzzy inference model rules. This enables accurate mapping of various input conditions to output SE modes to be efficiently utilised. These learnt relationships can be based on data collected or labelled by real users such that the learnt fuzzy model can then adaptively enhance speech intelligibility in various noise conditions, improving the effectiveness of the fuzzy inference model in dynamically adjusting to user preferences and environmental contexts.

1) ANFIS Structure: The structure of ANFIS includes several layers, each responsible for specific operations such as fuzzification, rules, normalisation, defuzzification, and output. The generated rules adhere to user specified definitions such as those given in Table I.

2) ANFIS Layer Functions:

Layer 1: Fuzzification of input variables using membership functions.
Layer 2: Application of fuzzy rules.
Layer 3: Normalisation of the rule strengths.
Layer 4: Generation of output functions.
Layer 5: Summation of all incoming signals to produce the final output.

Given the input variables x, y, and z, the ANFIS model applies the following functions across its layers. The fuzzy sets for the input sound power and visual quality (see Table I) are defined by their respective Gaussian membership functions where each member is initialised with reasonable prior parameters based on Equation 2 as the following. For Sound Power, $a = 1, b_1 = 2, b_2 = 5, b_3 = 8$ and $c_1 = c_2 = c_3 = 2.3$. For Visual Quality, $a = 1, b_1 = 157.7, b_2 = 360$, $b_3 = 562.2$ and $c_1 = c_2 = c_3 = 112.03$. For SNR, $a = 1, c_1 = c_2 = c_3 = 12.5$.

We assume, as depicted in Figure 2, that user preferences elicited in terms of their clinically assessed listening effort or Cognitive Load (CL) can effectively influence the input SNR parameters of the FIS, and the output MOS during the training process. Specifically, we use a Sigmoid function (Equation 10) to represent the individual's CL for varying input SNR.

In this paper, since all participants had normal hearing, utilising the same prior sigmoid function for all was deemed to be acceptable as a proof of concept. We take the mean values b_i for the SNR Gaussian membership functions from the CL Sigmoid (Equation 10) at suitable intervals, in order to represent Low, Medium and High CL categories respectively.

$$\sigma(x) = \frac{10}{1+e^{kx}} \tag{10}$$

Where x is the listening SNR, and k is an individual hyper-parameter (empirically set to $k = -0.22$) that depends on the user's functional hearing capability, representing their Cognitive Load or listening effort. The analytical hyper-parameter *k* allows our approach to generalise to different users. Equation (11) defines the Gaussian membership functions for the input SNR.

$$\mu_{A_i}(x) = a \times e^{-\frac{(x-b_i)^2}{2c_i^2}} \tag{11}$$

Where b_i represents the initialisation values for the input SNR means, with $b_1 = \sigma^{-1}(1)$ representing 'Low' CL, $b_2 = \sigma^{-1}(5)$ representing 'Medium' CL, $b_3 = \sigma^{-1}(9)$ represents 'High' CL.

Equations (12) and (13) define Gaussian membership functions for the input sound power and visual quality, respectively.

$$\mu_{B_i}(y) = a \times e^{-\frac{(y-b_i)^2}{2c_i^2}} \tag{12}$$

$$\mu_{C_i}(z) = a \times e^{-\frac{(z-b_i)^2}{2c_i^2}} \tag{13}$$

Rule Firing Strengths: Equation (14) calculates the firing strength of the applied rules as:

$$w_i = \mu_{A_i}(x)\mu_{B_i}(y)\mu_{C_i}(z) \tag{14}$$

Normalisation: Equation (15) normalises the firing strengths:

$$\overline{w_i} = \frac{w_i}{\sum_{i=1}^{n} w_i} \tag{15}$$

Consequent Calculation: Equation (16) calculates the output generated for each rule:

$$\overline{w_i} f_i = \overline{w_i}(m_i x + p_i y + q_i z + r_i) \tag{16}$$

Overall Output: The final output is the summation of all rule outputs, as shown in Equation (17):

$$O = \sum_{i=1}^{n} \overline{w_i} f_i \tag{17}$$

3) Training process: ANFIS employs a hybrid learning algorithm to fine-tune the system parameters that combines a backpropagation gradient descent and least-squares approach for modelling the training data. Backpropagation is used to learn the rule input membership function parameters while least square estimation is used to determine the rule consequent parameters of the modelled FIS. The parameter optimization during training aims to minimise the error between the target (mode type output from the user's preferences) and the actual mode type output of the FIS to determine the model's training accuracy.

4) Optimization: The training process involves optimising the membership function parameters to minimise a predefined error function. This can be represented mathematically as in Equation (18):

$$\min E = \frac{1}{2}\sum_{j=1}^{N}\left(y_j - \hat{y}_j\right)^2 \tag{18}$$

where E is the error, y_j is the actual output, and $\hat{y}_j$ is the predicted output by the ANFIS for the j-th data point.

We optimise ANFIS using a genetic algorithm (GA), where the fitness value is calculated as the absolute error. The selection operation chooses the strategy according to the fitness ratio and obtains the reciprocal of the fitness value. The crossover operation uses real number encoding to the individual [24]. The mutation operation selects genes in the code for mutation. We also utilise and compare an alternative benchmark approach to optimise ANFIS using particle swarm (PSO) [31] Here, the velocity and position of the particle are initialised, and fitness function values of particles are calculated and used to update the speed and position of each particle. The iterations stop when reaching a maximum specified number.

IV. Simulation Experiments and Comparative Results

This section discusses the performance evaluation metrics, simulation setup, the training and test process and comparative results. The evaluation of our proposed AV SE system's performance with state-of-the-art approaches utilises various metrics to assess the quality of SE, with a particular focus on the PESQ and STOI score analysis and Curve smoothness of the fuzzy model.

A. Performance Evaluation Metrics

The system's performance is quantitatively evaluated using several metrics, detailed as follows:

1) Speech Enhancement Quality Assessment: The quality of SE is assessed by measuring the clarity and intelligibility of speech in noisy environments. This involves a comparative analysis of the enhanced speech signals against the clean, original speech recordings to identify any distortions or improvements in the sound quality.

2) PESQ and STOI Score Analysis: The PESQ score provides an objective measurement of the speech quality as perceived by human ears, with scores ranging from -0.5 to 4.5. Higher scores indicate better speech quality. Conversely, the STOI score, ranging from 0 to 1, evaluates the intelligibility of speech. Higher STOI scores signify better speech intelligibility, showcasing the effectiveness of SE in adverse conditions. These metrics offer a comprehensive evaluation of the system's

ability to enhance speech quality and intelligibility, which is crucial for effective communication in noisy environments.

3) Curve smoothness: By observing the function curve between the input and output of the fuzzy system, and calculating the slope between adjacent points, we obtain the smoothness of the curve.

4) Hardware power consumption and latency: The average current is used to evaluate Hardware power consumption. The latency is the time difference between program initialization and output of enhanced speech.

5) Mean Opinion Score: Mean opinion score (MOS) is a commonly used subjective method for evaluating speech quality. We gather a diverse group of people, ask them to listen to recordings, and rate the quality on a scale of 1 to 5.

B. Simulation Setup

This subsection describes the setup used for the simulations to evaluate the performance of the proposed AV SE system.

1) Computational Environment and Dataset for Training and Testing: The simulations were conducted within a controlled computational environment equipped with high-performance GPUs (NVIDIA A10) to facilitate the processing of complex neural network models. The dataset comprises AV recordings collected under various noisy conditions, ensuring a comprehensive evaluation across different scenarios.

For generating the training and test data, we selected 216 sets of video and audio files from our benchmark 'in-the-wild' multi-talker AVSEC Challenge (TED talks) dataset [29] ensuring representation across a spectrum of sound power, SNRs or noise levels, and image clarity rated by users. Specifically, five young individuals including 2 males and 3 females, with normal hearing assessed each of the three enhancement modes on a 1-5 scale, while viewing and listening to the video samples. Based on the score, the optimal SE approach was selected and integrated into the initial rule table (Table I). Through a trial and error approach, we allocated 120 randomly selected sets for training purposes and the remaining 96 sets for testing the proposed algorithms. The fuzzy rules of the ANFIS model were initially defined based on the existing rule base (Table I) and membership functions informed by the cluster-based learning (Equations. 11-13) during the training phase of the ANFIS model (see Section III.E for details). After ANFIS training, the final Gaussian membership functions and parameters are obtained. The accuracy rate of the model positively correlated with the degree of resemblance between the output mode generated by the ANFIS model and derived from the user-defined rule table.

2) Preprocessing and Input Normalisation: Prior to inputting the data into the enhancement model, preprocessing steps such as noise reduction, signal amplification, and synchronisation between audio and visual components are executed. Additionally, input normalisation is applied to standardise the data, thereby enhancing the model's learning efficiency and prediction accuracy.

3) Hardware: The AV SE system operates on a Raspberry PI using a Linux operating system. It utilises a USB interface to connect a microphone and camera to receive voice and video input, and output enhanced audio from the SE system. The system includes an LCD screen for user-friendly operation. Compared to conventional desktop computers, the Raspberry Pi-based system offers portability and low power consumption.

The hardware structure diagram of the SE system is shown in Figure 3. The hardware connection is shown in Figure 4.

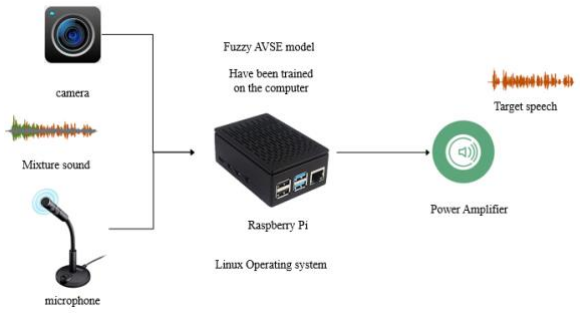


Fig. 3. Hardware structure diagram

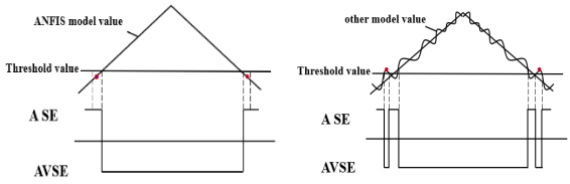
C. Fuzzy Inference System Performance

This subsection explores the performance of the fuzzy inference system implemented in the AV SE framework, focusing on the system's output and the visualisation of inference patterns.

1) Output Analysis and Mode Impact: The performance is evaluated based on the system's ability to adaptively enhance speech intelligibility in various noise conditions, highlighting the effectiveness of the FIS in dynamically adjusting to user preferences and environmental contexts.

Fig. 4. Hardware diagram of Raspberry Pi implementing the AV SE model

2) Surface diagram: Visualisation techniques are employed to illustrate the inference patterns generated by the fuzzy system. These visualisations provide insight into how the system processes input data and makes decisions, offering a clear view of the relationship between different input variables and the enhancement process. This analysis aids in understanding the system's behaviour and optimising its performance for improved SE outcomes.



(a) ANFIS (b) other model

Fig. 5. The process of switching modes

The user's interaction with the system, particularly in terms of mode transitions, is crucial for a positive user experience. When a variable gradually changes in one direction, the ANFIS model can be used for stable transition to another speech mode,

unlike other models that may produce unstable jumps, as depicted in Figure 5. It is essential to maintain consistency in speech enhancement especially during minor environmental changes to avoid unnecessary fluctuations that could disrupt user engagement.

Therefore, a detailed examination of the relationship between the input variables and the system's output—the function curve—is necessary to ensure smooth transitions and an optimised user experience. Gaussian membership functions are used for input variables and 3D surface plots of input variables (x, z axes) against the output decision (y axis) for both Mamdani and Sugeno generated rules are shown in Figures 6 and 7 respectively.

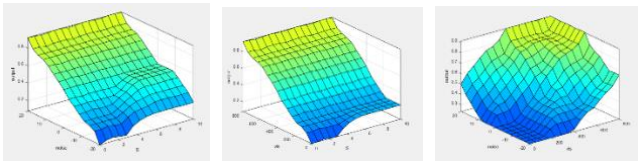


Fig. 6. 3D graph of variables using Mamdani rules

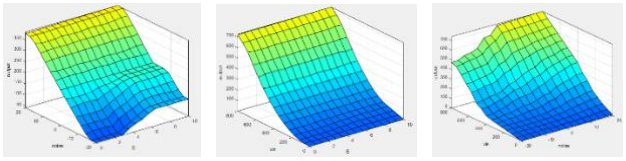


Fig. 7. 3D graph of variables using Sugeno rules

In the PSO-ANFIS model, we use the Sugeno model and again Gaussian membership functions are used for the input parameters. The 3D graphs of variables using Sugeno rules are shown in Figure 8.

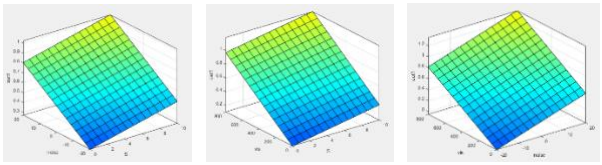
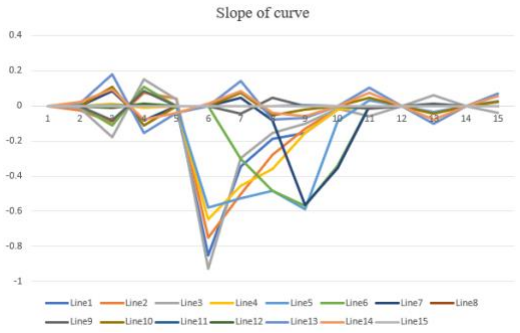
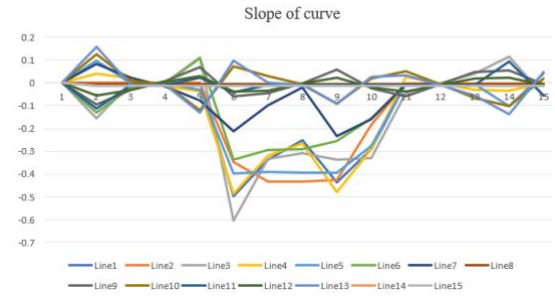


Fig. 8. 3D graph of variables using PSO-ANFIS

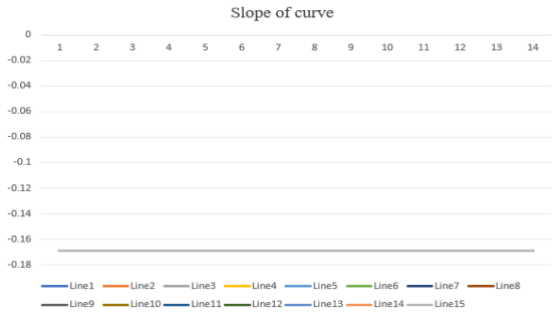
As can be seen from Figure 6-8, with the improvement of sound power, SNR and visual quality, the value of output mode gradually increases, and it tends to use audio-video SE mode.



(a)



(b)

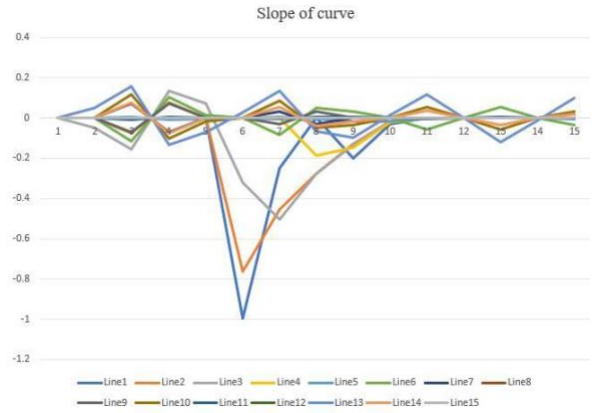


(c)

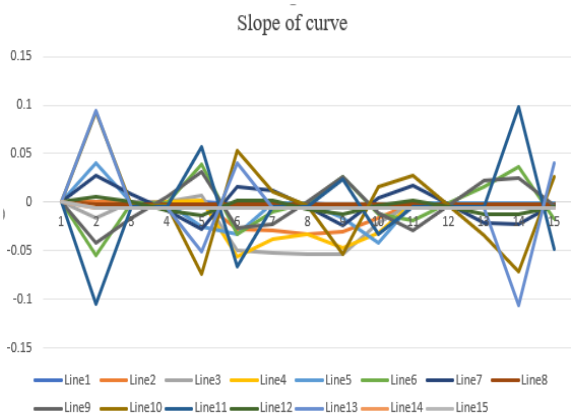
Fig. 9. The relationship between output values (as gradient on y axis) and constant sound power (x axis) for (a) Mamdani model, (b) Sugeno model and (c) PSO-ANFIS

In the context of maintaining a constant noise level, we examine the relationship between the output value and sound power to analyse the changes in slope, as illustrated in Figure 9. Each line in the graph represents the change in the slope of a curve in the 3D surface plot figure. A smaller change in slope correlates with a flatter curve, whereas greater variations in slope lead to more pronounced curvature. The ideal output pattern should exhibit gradual changes, favouring models characterised by minimal slope alteration to reduce fluctuations in SE modes. Conversely, rapid changes in output modes can induce oscillations between different modes, leading to a suboptimal user experience.

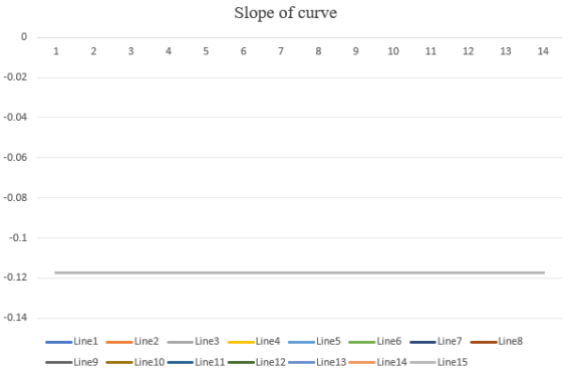
Here, we observe that Sugeno model performs slightly better than the Mamdani model in terms of output fluctuations of sound power changes. The fluctuations in PSO-ANFIS model are too small to be observed on a line chart; therefore, we can consider the input-output functions to be linear.



(a)



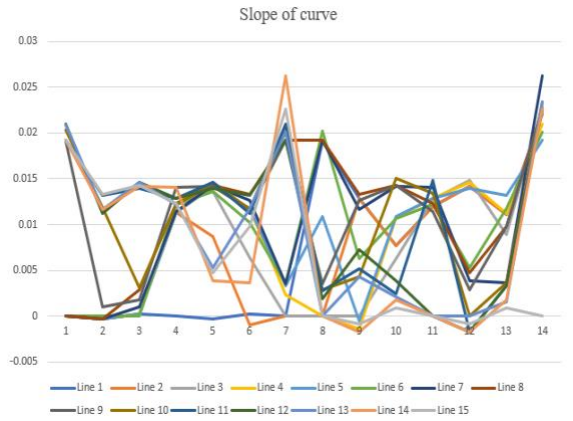
(b)



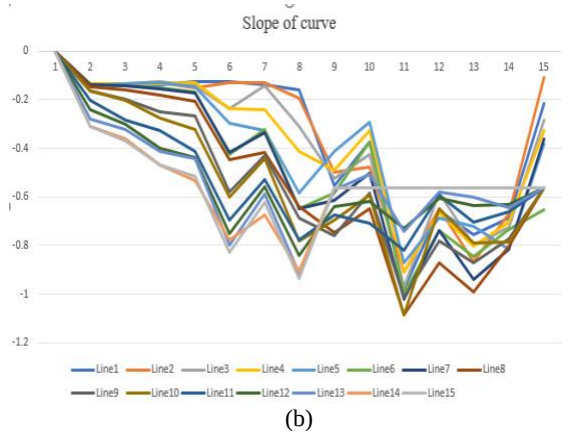
(c)

Fig. 10. The relationship between output values and noise value for (a) Mamdani model, (b) Sugeno model and (c) PSO-ANFIS

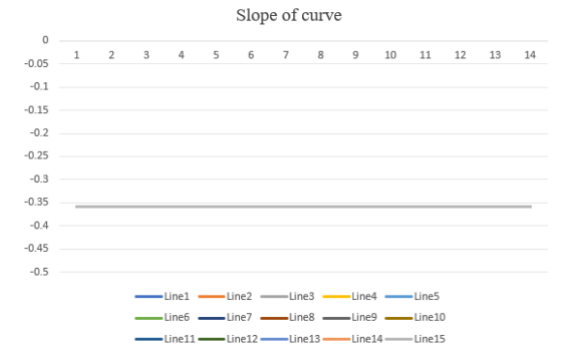
Keeping the video quality value constant, the relationship between the output value and the noise value is observed to determine the change in slope, as shown in Figure 10. In Figure 11, we further illustrate the slope of curve plots where the noise value is kept constant, and the relationship between the output value and the video quality is observed.



(a)



(b)



(c)

Fig. 11. The relationship between output values and constant visual quality value for (a) Mamdani model, (b) Sugeno model and (c) PSO-ANFIS

The standard deviation of the slope of the FIS output is shown in Table II for the various inputs and fuzzy models. It can be seen from the data that when using PSO-ANFIS model, the standard deviation of the slope is much smaller, and the slope changes are reduced resulting in a smoother transition between SE modes.

Theoretically, due to the insignificance of the changes observed in Figure 11c, they can be considered negligible, suggesting that the relationship between inputs and outputs can be approximated linearly. Therefore, a simpler linear predictive model could be constructed in future to fulfil the requirements of smoothness and model accuracy, while incorporating fuzzy membership input functions and effectively de-fuzzifying the output. A user's individual model parameterisation may benefit from simple interpretability; this is scope for future work.

TABLE II
STANDARD DEVIATION (SD) OF SLOPE OF FUZZY INFERENCE SYSTEM (FIS) OUTPUT FOR VARIOUS FUZZY MODELS

	Mamdani	Sugeno	PSO-ANFIS
FIS output slope SD for input noise	0.00269	0.00103	1.36481E-15
FIS output slope SD for input sound power	0.07060	0.01525	1.16358E-15
FIS output slope SD for input visual quality	0.00735	0.00294	9.7652E-16

D. Noise Reduction Effect

Performance evaluation was carried out on over 1000 videos comprising clean sounds, multi-speaker noises, speaker images, and denoised sounds sourced from the benchmark AVSEC Challenge dataset [29]. Objective evaluation metrics were

calculated both for A-only noise reduction and after AV SE noise reduction. The system operated based on output values: if the value ranged from 0 to 0.3, SE was not applied; for values between 0.3 and 0.6, A-only SE was used; and for values between 0.6 and 1, AV SE was applied. The average STOI and PESQ values are presented in Table III. The noise reduction effects of GA ANFIS and PSO ANFIS models demonstrate improved performance compared to the standard non-optimized ANFIS model. Comparatively the proposed fuzzy based AV SE model gives enhanced or similar results to state-of-the-art AV SE methods. This is reflected in high STOI and PESQ scores which are also attributable to the overall high quality of videos in the benchmark Challenge dataset. However, in a real-life situation where visual input quality may be poor due to camera obstructions, poor points of view or low lighting conditions, our proposed approach has the advantage of determining the optimal SE mode to utilise, instead of providing suboptimal enhancement due to the poor quality (or absence of) visual features.

E. System Efficiency Analysis

Computational Load and Energy Consumption: This section undertakes an evaluation of the system's efficiency through an in-depth analysis of its computational load and energy or power consumption. The goal is to ensure not only the effectiveness but also resource efficiency of the SE system, thereby rendering it suitable for deployment in portable devices and platforms. The power measurement of the hardware in each of the three noise reduction modes is acquired by a power sensor yielding values of: 2.875W, 3.965W, and 4.315W respectively. The fuzzy rules are used to select the noise reduction mode and calculate the average power. However, the system needs to consider the noise reduction effect and energy consumption. Compared with the AV SE model, the three fuzzy inference models can reduce the energy consumption under the premise of better noise reduction effect. The latency of the hardware in three noise reduction modes is measured as: 0s, 6s, and 18s respectively. Compared with the standalone benchmark AV SE method (without fuzzy inferencing), our proposed use of fuzzy-inference based AV SE can greatly shorten the delay while obtaining a sufficient enhancement effect as seen in Table III. In addition, the accuracy rates of Sugeno, ANFIS, GA-ANFIS and PSO-ANFIS were found to be 73.9%, 77.1%, 80.2% and 84.4%, respectively

F. Mean Opinion Score (MOS)

We recruited a cohort of five young volunteers (2 males and 3 females with normal hearing), to assess the quality of recordings using a 1 to 5 rating scale for subjective evaluations. The box plot diagram in Figure 12 reveals that the MOS of the PSO-ANFIS model outperforms other fuzzy models. This suggests that participants assigned higher subjective ratings to this model during testing. Moreover, PSO-ANFIS provides enhancement that is comparable to conventional AV enhancement, and certainly improves over A-only, whilst providing the flexibility to select the most effective SE modes in response to uncertain real-life conditions, user preferences and energy constraints of HA devices.

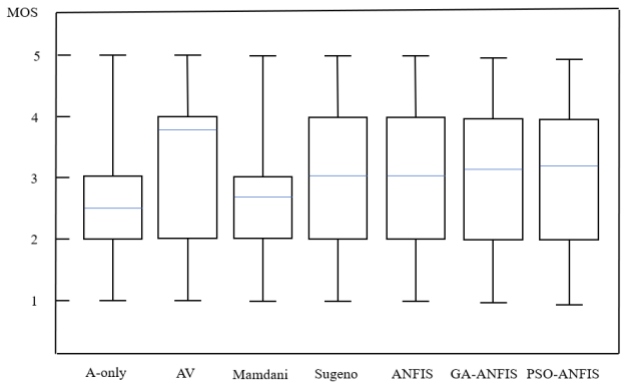


Fig. 12. Comparative MOS evaluation of state-of-the-art vs our proposed model

G. Comparative Analysis

This section conducts a comparative analysis aimed to evaluate the proposed system in relation to existing methods. Through the examination of various performance metrics, user satisfaction rates, and system efficiency, this analysis provides a comprehensive understanding of the advantages and potential limitations of our approach. It highlights the enhancements in SE quality and intelligibility achieved by our system, positioning it as a notable advancement in the field of AV speech processing.

Compared to the AV SE model, the three fuzzy inference models demonstrate the capability to reduce energy consumption and latency while achieving superior noise reduction effect. The comparison of these models alongside the proposed approach across various factors is detailed in Table III. Specifically, the Mamdani model exhibits lower accuracy and limited noise reduction effectiveness. Conversely, the PSO-ANFIS model showcases smooth performance, minimal recognition error, and high objective and subjective evaluation scores for SE.

Additionally, we explore the Back-Propagation (BP) neural network approach, which is characterised by a single fully-connected perceptron layer linking input features to the output value thus determining the nature of the enhancement processing.

In terms of speech quality measures such as STOI, PESQ and MOS, the AV SE system exhibits superior performance compared to our proposed model. However, in the context of delay and power metrics, our model demonstrates significant improvement over AV SE. Additionally, it's important to note that MOS scores do not consider the real-time implementation of the model, thereby overlooking the impact of delay on the user's ratings.

To this end, we believe that our PSO-ANFIS model represents a compelling balance between enhancement performance and computational efficiency.

Future endeavours will focus on conducting robust real-time evaluations and ecologically valid user testing to substantiate these assertions.

TABLE III
PERFORMANCE COMPARISON OF STATE-OF-THE-ART MODELS VS OUR PROPOSED MODEL

Algorithm	Accuracy rate	Gradient	Delay (s)	Power (W)	STOI	PESQ	MOS
ASE[30]	-	-	6	3.965	0.744	1.266	2.5
AV SE[30]	-	-	18	4.315	0.892	2.2	3.8
BP	83.3%	-	9.054	3.93	0.786	1.513	2.9
Mamdani [23]	68%	0.00269	9.28	3.98	0.781	1.503	2.8
Sugeno [23]	73.9%	0.00103	9.72	4.02	0.786	1.510	3.05
ANFIS [24]	77.1%	1.36E-15	9.606	3.995	0.786	1.513	3.05
GA-ANFIS [24]	80.2%	1.36E-15	9.288	3.985	0.788	1.522	3.17
PSO-ANFIS [31]	84.4%	1.36E-15	9.54	3.945	0.792	1.541	3.2

V. CONCLUSIONS AND FUTURE WORK

This work centred around the novel application of an adaptive neuro-fuzzy model by integrating AV SE with personalised learning to enhance the intelligibility and quality of speech in challenging noisy environments. By exploiting fuzzy logic and user cognitive load based preference rules, we successfully developed a context-aware system that adapts SE by considering the surrounding noise and visual conditions, pushing the boundaries of future adaptable assistive hearing and communication technologies. The inclusion of a genetic algorithm and particle swarm optimization methods proved very effective, fine-tuning our system to recognize environmental nuances. The proposed ANFIS based AV SE system was shown to result in less energy consumption and delay than conventional non-fuzzy based AV SE methods, whilst improving performance compared to A-only methods. This balance is particularly important in portable devices where energy efficiency is crucial. Moreover, the introduction of smoothness analysis emerged as a vital component, ensuring the enhanced audio remains stable over prolonged use, thereby prioritising user comfort alongside clarity. Finally, by implementing our model into a Raspberry PI, we bridged the gap between theory and practice, demonstrating the model's real-world viability. This confirms the potential of our neuro-fuzzy based AV SE system to be implemented in existing custom hardware technology.

A simpler linear predictive model for the SE type preference as mentioned in Section IV C presents an interesting avenue for future work, whereby a user's individual AV SE system parameterisation tailored to their preferences may benefit from enhanced interpretability and facilitate statistical comparisons of intelligibility and quality gains. Additionally, the streamlined system would be simpler to implement and optimise.

Further, as part of our future work, we are exploring the use of clinical cognitive load data to comprehensively evaluate personalised fuzzy based AV SE systems. Additionally, we will investigate ways to reduce the system running time, including through integration of our baseline lightweight and real-time AV SE models [9][32] for real-time SE whilst assessing and optimising system energy and computational cost. Joint optimization of both user-defined preferences and objective metrics such as latency/power consumption and PESQ/STOI would further make the system highly adaptable in different real world operating conditions Ideally, the system would learn from user feedback to continuously enhance the listening experience for various environments in real-time. The assumption is that a smooth model with the lowest 'switching frequency' whilst maintaining high accuracy of user preferred SE processing is logical. However, more tests are required to be conducted under a range of realistic situations.

Given the dataset is generated from the benchmark AVSEC [29] (derived from the LRS3 dataset based on TED talks), AV SE delivers consistently good subjective and objective performance due to the overall high visual quality of the dataset. In a real-life situation where, visual quality may be poor, for example, due to camera obstructions, unclear points of view or low lighting conditions, our proposed approach has the advantages of providing 1) improved power efficiency and 2) optimal enhancement due to the poor quality (including potential absence of) visual features. Future work will endeavour to prove our fuzzy AV SE system's efficacy in more ecologically valid situations and also compare its performance more comprehensively with a range of state-of-the-art A-only and AV based approaches.

The simplicity of crisp fuzzy based AV SE system outputs here is useful for this initial work to progress towards our final goal of having an efficient and personalised SE system for hearing-aid users. However, it may be of interest to incorporate more sophistication by modulating the set of DNN parameters at a more granular level, for example, through a modular DNN architecture replacing the current multiple SE models, to enable smooth switching between various SE processing modes. This could further decrease power consumption according to the combination of environmental conditions and user preferences. Additionally, by systematically considering different types of noises, acoustic environments, and SNRs [33], we could gain a further degree of personalisation. Another required investigation would be the clinical validation of our ongoing user Cognitive Load integration, and assessment of related AV SE system performance enhancement and potential training time reduction for individuals with hearing loss.

ACKNOWLEDGEMENTS

This work is supported by the UK Engineering and Physical Sciences Research Council (EPSRC) funded COG-MHEAR programme (under Grant EP/T021063/1) and the NatGEN project (Grant EP/T024917/1). S. Chen is supported by the Scientific research project of Colleges and Universities in Anhui Province of China (under Grant 2022AH040044). For the purpose of Open Access, the authors have applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript (AAM) version arising from this submission.

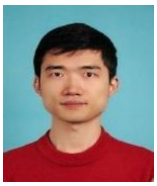
REFERENCES

[1] C. Zheng, H. Zhang, W. Liu, X. Luo, A. Li, X. Li, and B. C. Moore, (2023) "Sixty years of frequency-domain monaural speech enhancement: From traditional to deep learning methods," Trends in Hearing, vol. 27, p. 23312165231209913

[2] A. Azarang and N. Kehtarnavaz, (2020) "A review of multi-objective deep learning speech denoising methods". Speech Communication, vol. 122, pp. 1-10

[3] P. Gonzalez, T. S. Alstrøm, and T. May, "Assessing the generalisation gap of learning-based speech enhancement systems in noisy and reverberant environments," IEEE/ACM Trans. Audio, Speech, and Lang. Process., 2023.

[4] J. Richter, S. Welker, J. M. Lemercier, B. Lay, and T. Gerkmann, (2023) "Speech enhancement and dereverberation with diffusion-based generative models," IEEE/ACM Trans. Audio, Speech, and Lang. Process.
[5] Z. Qiu et al., (2023) "Artnet: Time domain speech enhancement via stochastic refinement," in ICASSP 2023 - 2023 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), pp. 1–5.
[6] J. Kirton-Wingate, S. Ahmed, M. Gogate, Y. Tsao, and A. Hussain (2023), 'Towards Individualised Speech Enhancement: An SNR Preference Learning System for Multi-Modal Hearing Aids', in 2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSP), pp. 1–5.
[7] A. H. Andersen, S. Santurette, M. S. Pedersen, E. Alickovic, L. Fiedler, J. Jensen, T. Behrens., (2021), August. Creating clarity in noisy environments by using deep learning in hearing aids. In Seminars in hearing (vol. 42, No. 03, pp. 260-281). Thieme Medical Publishers, Inc..
[8] R. Mira et al. (2023), "LA-VocE: Low-SNR Audio-visual Speech Enhancement using Neural Vocoders," in ICASSP 2023 - 2023 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), pp. 1–5.
[9]. M. Gogate, K. Dashtipour, and A. Hussain, "Robust Real-time Audio-Visual Speech Enhancement based on DNN and GAN," IEEE Trans. on Artificial Intelligence, vol. 1, no. 01, pp. 1–10, 2024.
[10] J. B. Nielsen, J. Nielsen, B. S. Jensen, and J. Larsen, (2013) 'Hearing Aid Personalization', in 27th Annual Conference on Neural Information Processing Systems (NIPS 2013
[11] A. Pasta, M. K. Petersen, K. J. Jensen, N. H. Pontoppidan, J. E. Larsen, and J. H. Christensen, (2022) 'Measuring and modelling context-dependent preferences for hearing aid settings', User Modeling and User-Adapted Interaction, vol. 32, no. 5, pp. 977–998
[12] Y. Dong and Q. Ye, (2023) "Neural network-based speech fuzzy enhancement algorithm for smart home interaction," J. of Computational Methods in Sciences and Engineering, (Preprint), pp. 1–12
[13] C.-C. Yao and M.-H. Tsai, (2010) 'Adaptive fuzzy filter for speech enhancement', in Computational Science and Its Applications--ICCSA 2010: International Conference, Fukuoka, Japan, March 23-26, 2010, Proceedings, Part III 10, 2010, pp. 511–525.
[14] M. A. Ben Messaoud, A. Bouzid, and N. Ellouze, (2016) 'A new biologically inspired fuzzy expert system-based voiced/unvoiced decision algorithm for speech enhancement', Cognitive computation, vol. 8, pp. 478–493
[15] D. Bozanic et al., (2023) "Ranking challenges, risks and threats using Fuzzy Inference System," Decision Making: Applications in Management and Engineering, vol. 6, no. 2, pp. 933–947
[16] M.-B. Neagu, A. A. Kressner, H. Relaño-Iborra, P. Bækgaard, T. Dau, and D. Wendt (2023), 'Investigating the Reliability of Pupillometry as a Measure of Individualized Listening Effort', *Trends in Hearing*, vol. 27, p. 23312165231153288
[17] L. A. Passos, J. P. Papa, A. Hussain, and A. Adeel, (2023) "Canonical cortical graph neural networks and its application for speech enhancement in audio-visual hearing aids," Neurocomputing, vol. 527, pp. 196–203
[18] I.-C. Chern et al. (2023), 'Audio-visual speech enhancement and separation by utilizing multi-modal self-supervised embeddings', in 2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW), 2023, pp. 1–5.
[19] S. Balasubramanian, R. Rajavel, and A. Kar, (2023) "Estimation of Ideal Binary Mask for Audio-Visual Monaural Speech Enhancement," Circuits, Systems, and Signal Processing, pp. 1–25
[20] M. I. Khattak et al., (2022) "Regularised sparse features for noisy speech enhancement using deep neural networks," Computers and Electrical Engineering, vol. 100, p. 107887
[21] R. Liu, (2022) "Speech noise reduction system based on combined filter," in 2022 IEEE Conf. on Telecommunications, Optics and Computer Science (TOCS), pp. 611–616.
[22] J. Wang, N. Saleem, T. S. Gunawan (2024) Towards Efficient Recurrent Architectures: A Deep LSTM Neural Network Applied to Speech Enhancement and Recognition. Cogn Comput, vol. 16, pp. 1221–1236.
[23] A. Abel, A. Hussain, and B. Luo, (2014) "Cognitively inspired speech processing for multimodal hearing technology," in 2014 IEEE Symposium on Computational Intelligence in Healthcare and e-health (CICARE), pp. 56–63.
[24] S. V. Wong and A. M. S. Hamouda, (2000) "Optimization of fuzzy rules design using genetic algorithm," Advances in Engineering Software, vol. 31, no. 4, pp. 251–262
[25]. F. Gao, (2023) "Density-based approach for fuzzy rule interpolation," Applied Soft Computing, vol. 143, p. 110402
[26] C. J. Mantas, J. M. Puche, and J. M. Mantas, "Extraction of similarity-based fuzzy rules from artificial neural networks," International Journal of Approximate Reasoning, vol. 43, no. 2, pp. 202–221, 2006
[27] Z. -X. Wei, F. Doctor, Y. -X. Liu, S. -Z. Fan and J. -S. Shieh, (2020) "An Optimized Type-2 Self-Organizing Fuzzy Logic Controller Applied in Anesthesia for Propofol Dosing to Regulate BIS," in *IEEE Transactions on Fuzzy Systems*, vol. 28, no. 6, pp. 1062-1072.
[28] L. Jara et al., (2022) "Efficient inference models for classification problems with a high number of fuzzy rules," Applied Soft Computing, vol. 115, p. 108164
[29] A. L. A. Blanco et al., (2023) "AVSE Challenge: Audio-Visual Speech Enhancement Challenge," in 2022 IEEE Spoken Language Technology Workshop (SLT), Jan. 2023, pp. 465–471.
[30] T. Hussain, M. Gogate, K. Dashtipour, A. Hussain (2021). Towards intelligibility-oriented audio-visual speech enhancement. *arXiv preprint arXiv:2111.09642*.
[31] H. Basser, H. Karami, S. Shamshirband, S. Akib, M. Amirmojahedi, R. Ahmad, A. Jahangirzadeh, H. Javidnia., (2015) "Hybrid ANFIS–PSO approach for predicting optimum parameters of a protective spur dike" Applied Soft Computing, vol. 30, pp. 642-649.
[32] Cogmhear. 2024. Lightweight Audio-Visual Speech Enhancement. GitHub. Retrieved from https://github.com/cogmhear/lightweight-AV-SE
[33] J. Kirton-Wingate, S. Ahmed, A. Hussain, M. Gogate, K. Dashtipour, J-C. Hou, T. Hussain, Y. Tsao, A. Hussain (2024). Towards Environmental Preference Based Speech Enhancement For Individualised Multi-Modal Hearing Aids. *arXiv preprint arXiv:2402.16757*.



Song Chen received his Ph. D from the University of Science and Technology of China. He is working as a Lecturer in the School of Mechanical and Electrical Engineering at Anhui Jianzhu University since 2011. Dr Chen has been an academic visitor at Edinburgh Napier University, UK since 2023. His research interests are centred on deep neural networks, pattern recognition and real-world applications.

Jasper Kirton-Wingate is a PhD student in Computing since 2021, at Edinburgh Napier University, UK. His research is focussed on multimodal speech enhancement and personalised machine learning for hearing assistive technologies.

Faiyaz Doctor (M'08-SM'21) received his PhD. degree in computer science in 2006 from the University of Essex. He is currently a Senior Lecturer and head of the Intelligent Connected Societies Group at the same University. He serves as an Associate Editor for the IEEE Transactions on Fuzzy Systems and has published over 100 peer-reviewed articles.

Usama Arshad is currently completing his Ph.D. in Computer science at the Ghulam Ishaq Khan Institute of Engineering Sciences and Technology in Pakistan. He is working on the integration of fuzzy systems with Machine Learning and other technologies to enhance system optimization in terms of scalability, cost-effectiveness, security, privacy, and robustness.

Kia Dashtipour is a Lecturer in Computing at Edinburgh Napier University since 2021. He has published over 120 peer-reviewed research papers including numerous highly-cited works in the areas of multimodal signal and image processing with a focus on audio-visual speech enhancement and sentiment and opinion mining.

Mandar Gogate is a Principal Research Fellow in the School of Computing at Edinburgh Napier University, UK. He has been an invited visiting research fellow at MIT (Synthetic Intelligence Lab), and University of Oxford (Computational Neuroscience Lab). His research interests include: real-time audio-visual enhancement and multimodal fusion for assistive hearing and healthcare applications.

Zahid Halim earned his Ph.D. degree from the National University of Computer and Emerging Sciences, Pakistan in 2010. His current research interests include neuro-fuzzy systems and their applications. He is an Associate Editor for the IEEE Transactions on Artificial Intelligence.

Ahmed Al-Dubai (Senior Member, IEEE) received the Ph.D. degree in computing from the University of Glasgow in 2004. He is currently a Professor in Computing at Edinburgh Napier University, U.K. His research interests include communication algorithms, Edge Computing, and cognitive Internet of Things. He is Associate Editor of the IEEE Transactions on Sustainable Computing.

Tughrul Arslan is a Professor in the School of Engineering at Edinburgh University, UK. His research interests include developing low-power radio frequency sensors for wearable and portable biomedical applications. He has authored over 500 refereed papers and more than 20 patents. He is an Associate Editor of the IEEE Transactions on Very Large Scale Integration (VLSI) Systems.

Amir Hussain (Senior Member IEEE) is Founding Director of the Centre of AI and Robotics at Edinburgh Napier University, U.K. His research interests are focused on trustworthy AI and cognitive data science technologies to engineer the smart healthcare and industrial systems of tomorrow. He has authored several patents and over 600 publications, including around 300 journal papers and over 25 Books/monographs. He is founding Chief Editor of Springer's Cognitive Computation journal and currently serves as Associate Editor for the IEEE Transactions on Artificial Intelligence, IEEE Transactions on Emerging Topics in Computational Intelligence and the IEEE Transactions on Systems Man and Cybernetics: Systems.