TOWARDS INDIVIDUALISED SPEECH ENHANCEMENT: AN SNR PREFERENCE LEARNING SYSTEM FOR MULTI-MODAL HEARING AIDS

Jasper Kirton-Wingate[†], *Shafique Ahmed*[⋆], *Mandar Gogate*[†], *Yu Tsao*[⋆], *Amir Hussain*[†]

[†]Edinburgh Napier University, [⋆]Academia Sinica

ABSTRACT

Since the advent of deep learning (DL), speech enhancement (SE) models have performed well under a variety of noise conditions. However, such systems may still introduce sonic artefacts, sound unnatural, and restrict the ability for a user to hear ambient sound which may be of importance. Hearing Aid (HA) users may wish to customise their SE systems to suit their personal preferences and day-to-day lifestyle. In this paper, we introduce a preference learning based SE (PLSE) model for future multi-modal HAs that can contextually exploit audio and visual information to improve listening comfort (LC). The proposed system estimates the Signal-to-noise ratio (SNR) as a basic objective speech quality measure which quantifies the relative amount of background noise present in speech, and directly correlates to the intelligibility of the signal. This is used alongside a preference elicitation framework which learns a predictive function to determine the target SNR. The system is novel, scaling the output of an Audio-Visual (AV) DL-based SE model to provide HA users with individualised SE. Preliminary results support the hypothesis of improving the overall subjective LC, without significantly impeding the speech intelligibility.

Index Terms— Audio-visual speech enhancement, hearing aids, individualisation, preference learning.

1. INTRODUCTION

Speech enhancement (SE) models are typically evaluated by criteria that objectively measure both the speech quality and intelligibility (i.e. PESQ, STOI). However, the ideal SE model is not necessarily ‘one size fits all’. In the literature, it has been established that preference for noise reduction (NR) strength for different signal-to-noise ratios (SNRs) varies amongst hearing aid (HA) users [1]. Given the scientific description of the highly individualised and non-linear pathology that constitutes hearing loss [2, 3], there have been relatively few attempts to personalise modern, non-linear SE algorithms with respect to the hearing impaired listeners preferences. This gap has been highlighted recently in [4].

In terms of established HA personalisation, there are many studies and models for HA fine-tuning, particularly in personalising the parameters in digital multi-band dynamic range compression algorithms that exist in most modern HAs, by adjusting the patient’s audiogram based prescription and/or compression parameters [5]. Other studies have focused on personalising SE models with respect to the listener’s preference by fine-tuning NR parameters [6] , or with respect to the audiogram with spectral change enhancement [7]. However, these approaches do not utilise recent advances in DL technologies, which have significantly improved metrics such as PESQ and STOI compared to other approaches [4]. More recently, Drakopoulos et al. [8] proposed to effectively invert an entire auditory pathology by decreasing the error between Normal Hearing (NH) and Hearing Impaired (HI) simulated auditory nerve responses using DL.

In terms of individualised SE, Bhat et al. [6] proposed a formant based SE framework to customise the noise suppression and speech distortion according to the user preference elicited via a smartphone based elicitation system. The model exploits the formant frequency information to control the HA output while maintaining speech intelligibility. However, to the best of our knowledge there are no attempts in the literature to solve the aforementioned issue for audio or audio-visual (AV) DL based SE. Moreover, the differential in preferences that are shown for varying types of noise experienced in the real world, e.g. accounting for the trade off between preference for noise level and naturalness [9, 10], have not been explicitly accounted for in adaptive DL based SE.

In this paper, we propose a framework for individualised AV speech enhancement, as shown in Fig. 1, that controls the output of DL-based AV SE models according to the user preference of noise reduction at different levels of background noise to improve the overall LC without compromising on speech intelligibility. Specifically, a DL-based SNR estimation model is used to describe the individualised preference for AV SE.

The SE, or perhaps more accurately, Noise Reduction (NR) algorithm employed in this study is a class of DL based AV SE [11]. Intelligibility-Oriented AV SE (IOAVSE) [12] in-particular has been chosen because of it’s high performance in comparison with other NR algorithms at low SNR, as well as it’s Ideal Ratio Mask (IRM) based output which makes adjustment of the SE target SNR (SNR^*) simple via the modulation of the activation function of the final output

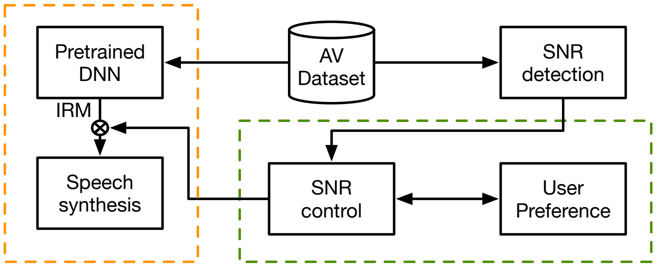


Fig. 1. An overview of the Preference Learning Based Audio-Visual Speech Enhancement Model

layer. Once the user's preference function has been estimated via preference elicitation, we hypothesise that it can be utilised to achieve better overall LC than using a static, 'one-size-fits-all' SE model in HA, without significantly impeding the intelligibility.

We hypothesise that several factors influence the inter-individual contextual preferences (given environmental SNR) for the amount of enhancement required: functional hearing ability (i.e. the ability to discriminate sounds from noise), and the level of discomfort caused by distortion or unnaturalness. For example, someone with high functional hearing ability may wish to be aware of surrounding noise which may contain cues that compete for attentional resources, such as train announcements, mechanical / car noise, or other conversational speakers. This preference may be further motivated by a high level of discomfort caused by the distortions introduced with SE. On the other hand, someone with low functional hearing ability may wish for maximum NR, to mitigate the cognitive load associated with high listening effort, particularly if less discomfort is caused by any unnaturalness introduced by the algorithm.

2. ENVIRONMENTAL SEGMENTAL SIGNAL-TO-NOISE RATIO ESTIMATION

Signal-to-noise ratio (SNR) is a basic objective speech quality measure which quantifies the relative amount of background noise present in a speech clip, usually in terms of sound pressure, measured in decibels (dBs). It is defined as the ratio of signal intensity to noise intensity. In contrast to working directly on the entire signal in our experiments, we used the Segmental Signal-to-Noise Ratio (SNRseg or SSNR), which calculates the average of the SNR values (in dB) of short segments (15 to 20 ms) given as equation 1:

$$\text{Seg.SNR} = \frac{10}{M}\sum_{m=0}^{M-1}\log_{10}\left(\frac{\sum_{n=N_m}^{N_m+N-1} s^2(n)}{\sum_{n=N_m}^{N_m+N-1}\{s(n)-\hat{s}(n)\}^2}\right) \quad (1)$$

The user's preferred SNR essentially refers to the SNR that a user prefers when listening to a target, whilst experiencing a given level (amount) of background noise. Several DL-based assessment tools have been developed utilising a range of model architectures, e.g., BiLSTM [13], CNN [14], and CNN-BiLSTM [15]. Additionally, attention mechanisms [16] and multitask learning [17] have also been employed to enhance assessment abilities. For a non-intrusive measure of SSNR, we have built a SSNR prediction model for our proposed framework. We have chosen power spectral features as input and CNN-BiLSTM with self-attention as a model architecture after experimentation. The CNN-BiLSTM+AT model architecture has 3 convolution blocks each consisting of 4 convolutional layers (16, 32, 64, and 128 filters), followed by single-layered BiLSTM (128 units), a fully connected layer (128 units) and a self attention layer. The output of the attention layer is fed to a fully connected layer (1 unit) a global average operation was then used to produce the prediction score as illustrated in Fig. 2.

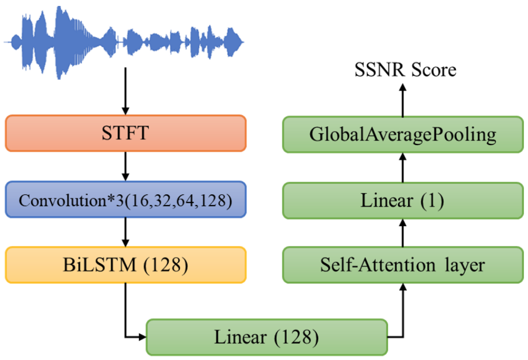


Fig. 2. SSNR prediction model

Table 1. SSNR Prediction Model Results

Model Type	LCC	SRCC	MSE
S	0.934	0.927	0.538

2.1. Evaluation

To evaluate the SSNR prediction model, three evaluation metrics were used: linear correlation coefficient (LCC), Spearman rank correlation coefficient (SRCC) and mean squared error (MSE) [16]. Higher LCC and SRCC scores show that the predicted scores are of higher correlations to the ground truth assessment scores, whilst a lower MSE score indicates that the predicted scores are closer to the ground-truth assessment scores. The experimental results of the SSNR prediction model for the GRID corpus with the 'Pedestrian' CHIME-3 noise type, at a range of -12 to 9 db SNR is shown in Table 1.

3. PREFERENCE LEARNING SYSTEM

This section presents the individual components present in the proposed PLSE framework.

3.1. Pretrained AV SE model

The model employed for this study is IO-AVSE [12]. The DL architecture consists of a deep fully convolutional network-based U-Net style architecture and a concatenation style fusion between the AV modalities after visual feature extraction. It uses the Short-time Objective Intelligibility (STOI) as a loss function.

3.2. SNR Control / Update Mechanism

The basic idea here is to utilise the user preferences and the environmental SSNR prediction in order to control in real-time the target SNR^* of the SE system.

$$SNR^* \,\alpha\, f(\widehat{SNR}, A)\; and\; A = \widehat{SNR}\,\beta + \beta_0 \tag{2}$$

$$A\,\alpha\,(\sum p_{1:n}\;/10) + 0.5\; where\;\; p_{1:n} \in (-1, +1) \tag{3}$$

where $\widehat{SNR}$ is the predicted environmental SNR. The elicitation sequence begins at 50% enhancement and $A\{(0, 1)$ (for input into equation 4). p is the preference vector from the elicitation phase, and β and β_0 are learnt from inputs p and $\widehat{SNR}$.

Generalised Logistic Activation Function (Richards Curve)

$$SNR^* \,\alpha\, A + \frac{(K - A)}{(C + Qe^-Bt)^1/v} \tag{4}$$

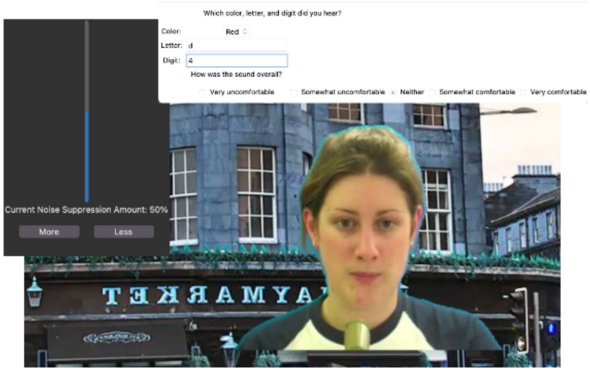
Where A, the lower asymptote (which essentially acts as a SNR^* noise floor) is to be inferred from the learned preference function. Equation 4 scales the activation and thus the SNR^* for the final output layer of magnitude spectrogram estimation from the IRM-SE model. This approach should generalise to other types of mask, which will be demonstrated in future work.

3.3. Preference Elicitation Method

During preference learning, the user's preference for SE given the environmental SNR is elicited by means of a traditional and widely employed 'volume up and down' interface. The interface controls the target SNR^* instead of the volume. This could be thought of as controlling the relative volume of the noise. Because this is a simple, familiar interface, the user can elicit their preference which will in turn adjust the SNR^* of the SE model in real-time, whilst listening to the resultant sound and sequentially adjusting, or not adjusting, their preference.

3.4. Experimental Setup

The data used, $\mathbf{Y}$, for this study is the GRID-CHIME3, augmented dataset [18, 19]. For each of the experimental phases, 6 unique sentences from 4 target speakers from the corpus are overlaid onto each of 5 noise levels sampled from the 'pedestrian' noises of the CHIME3 dataset, cut and volume adjusted. This leads to 5 SNRs (-9,3,0,3,9) dB



Which color, letter, and digit did you hear?

Color: Red

Letter: d

Digit: 4

How was the sound overall?

○ Very uncomfortable ○ Somewhat uncomfortable ● Neither ○ Somewhat comfortable ○ Very comfortable

Current Noise Suppression Amount: 50%

Fig. 3. Audio-visual presentation of GRID-CHIME3 dataset and User interface for preference elicitation (top left) and likert evaluation (top right)

The experiments are split up into 3 phases:
1. Preference Elicitation Phase: Here the participant engages with the 'up, down or no change' interface.
2. Max SNR Evaluation Phase: SE with maximum SNR^* , i.e. sigmoid-scaled final output layer from the pre-trained model.
3. Pref SNR Evaluation Phase: SE with the inferred SNR^* from the learned preference function, i.e. output layer scaled according to equation 2.

During the Evaluation phases, the participants are asked to recall 3 keywords from the sentence (colour, letter and digit), and to rate the LC on a 5 point likert scale. The results of interest for the main hypothesis are then the differential between Phases 2 and 3, which rely on successful elicitation in Phase 1.

4. PRELIMINARY RESULTS AND DISCUSSION

To get preliminary insights as a proof-of-concept test, 4 subjects were tested. The subjects ages ranged from 27 to 35 and all were male. There was one native British English speaker (participant 1) and the rest were non-native English speakers (2-4).

4.1. Results

4.1.1. Preference Functions

In Figure 4, the participant's preference functions, learned from the elicitation phase, are shown. The functions are also maximised by a ceiling, this limits the output scaling (parameter A in equation 4) where $f(\widehat{SNR}) > 1$, such as not to introduce distortions to the audio. Both participant 1 and 2 exhibit a negatively sloping function, showing preference for less SE at higher SNRs. This makes sense according to the expected challenge of the listening situations. Participant 3

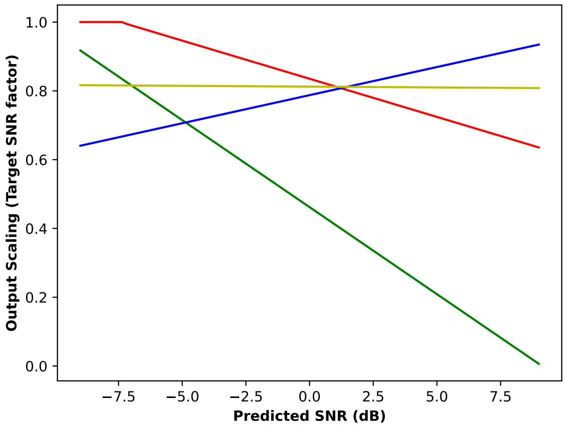


Fig. 4. Visualisation of participant 1-4's preference functions. Participant 1, red (with ceiling); 2, green; 3, blue; 4, yellow

Table 2. Subjective Testing Results

Participant (max/pref)	Intelligibility (%)	LC (Mean)
1 (max)	90	1.8
1 (pref)	92	3.8
2 (max)	66	3.5
2 (pref)	83	4.1
3 (max)	72	3.0
3 (pref)	69	3.1
4 (max)	83	3
4 (pref)	83	3.8

shows a positive sloping preference function, it is thought that this is why the intelligbility is impeded (slightly) after preference learning as there is less SE at lower SNR's. This warrants further consideration to investigate if this is intended or a product of a flaw in the elicitation framework. Paricipant 4 shows a flat preference function, though showing preference for 0.8 scaling as opposed to the default 1 (max). For this participant, allowing a small amount of background noise across SNRs is beneficial for LC.

4.1.2. *Change in Perceived Intelligibility and Quality*

In the table 2, preliminary results for the experimental set up given in Section 3.4 are shown.

On average, there was a 4% increase in intelligibility, however, due to a possible training effect (particularly for participant 2 who showed some initial confusion with the sentence-keyword-recall set up), there may be some bias between phases for intelligibility due to this. To avoid this in future experiments, a practice round for the subjective evaluation will be given prior to the 'max' and 'pref' SE evaluation phases.

In terms of LC, there was an average increase of 0.9 in likert scale (2 s.f.) or 18%, with every participant showing some increase in LC. This is suspected to be a more significant result.

4.2. Limitations

Whilst these preliminary results are promising and warrant further investigations, there are a number of limiting factors to consider. The first of which is the fact that in real-life scenarios where a HA is worn, there is expected to be some degree of noise which is not occluded actively or passively by the HA. To overcome this, it may be necessary to obtain realistic estimates for this 'leak-in' sound pressure, to simulate and incorporate into the experimental design. Additionally, the speech-in-noise data is synthetic, where as in the real world natural effects would be observed such as Lombard and reverberations. As N participants here is quite small, more experiments will be required to show the efficacy of the system (for NH and HI).

4.3. Conclusion and Future Work

This paper presented a proof-of-concept of an individualised speech enhancement framework that utilise user preferences to dynamically change the SE output in order to improve the user's listening comfort (LC). It is to be noted that, the preliminary results are inline with the original hypothesis that there would be no significant decrease in intelligibility, whilst the LC would increase significantly.

As a main priority, although the market for 'hearables' is growing (particularly for normal hearing), this work will be extented to increase the LC and quality specifically for the hearing impaired which suffer pathologically to understand speech in noise. This could in turn help increase HA uptake, which is estimated to be low in proportion to those who need HA [2]. Therefore, extensive and robust testing, including measurements of audiogram, along with normal hearing results for comparison, will be carried out with hearing impaired participants, in order to test statistical significance of any changes in intelligibility, LC and/or sound quality.

It is also of interest to investigate the differential of different noise types on preference for target SNR of the SE system, which would extend the SNR prediction model presented here to a hierarchical model. This may further increase the LC for multimodal HA users. All of these considerations are undergoing further implementations and experiments, including integration with other HA signal processing to deliver multi-modal HA demonstrators.

Acknowledgements

This work is supported by the UK Engineering and Physical Sciences Research Council (EPSRC) programme grant: COG-MHEAR (Grant reference EP/T021063/1).

5. REFERENCES

[1] Tobias Neher and Kirsten Wagener, "Investigating differences in preferred noise reduction strength among hearing aid users," *Trends in Hearing*, vol. 20, 09 2016.

[2] Nicholas A. Lesica, "Why do hearing aids fail to restore normal auditory perception?," *Trends in Neurosciences*, vol. 41, no. 4, pp. 174–185, 2018.

[3] Brian C. J. Moore, David A. Lowe, and Graham Cox, "Guidelines for diagnosing and quantifying noise-induced hearing loss," *Trends in Hearing*, vol. 26, 2022.

[4] Asger Andersen, Sébastien Santurette, Michael Pedersen, Emina Alickovic, Lorenz Fiedler, Jesper Jensen, and Thomas Behrens, "Creating clarity in noisy environments by using deep learning in hearing aids," *Seminars in Hearing*, vol. 42, pp. 260–281, 08 2021.

[5] Jens Nielsen, Jakob Nielsen, and Jan Larsen, "Perception-based personalization of hearing aids using gaussian processes and active learning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, 01 2014.

[6] Gautam S Bhat, Chandan KA Reddy, Nikhil Shankar, and Issa MS Panahi, "Smartphone based real-time super gaussian single microphone speech enhancement to improve intelligibility for hearing aid users using formant information," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2018, pp. 5503–5506.

[7] Jing Chen, Brian Moore, Thomas Baer, and Xihong Wu, "Individually tailored spectral-change enhancement for the hearing impaired," *The Journal of the Acoustical Society of America*, vol. 143, pp. 1128–1137, 02 2018.

[8] Fotios Drakopoulos and Sarah Verhulst, "A Differentiable Optimisation Framework for The Design of Individualised DNN-based Hearing-Aid Strategies," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, May 2022, pp. 351–355.

[9] Inge Brons, Rolph Houben, and Wouter Dreschler, "Perceptual effects of noise reduction with respect to personal preference, speech intelligibility, and listening effort," *Ear and hearing*, vol. 34, 08 2012.

[10] Aleksandra M. Kubiak, Jan Rennies, Stephan D. Ewert, and Birger Kollmeier, "Relation between hearing abilities and preferred playback settings for speech perception in complex listening conditions," *International Journal of Audiology*, vol. 61, no. 11, pp. 965–974, Nov. 2022.

[11] Daniel Michelsanti, Zheng-Hua Tan, Shi-Xiong Zhang, Yong Xu, Meng Yu, Dong Yu, and Jesper Jensen, "An overview of deep-learning-based audio-visual speech enhancement and separation," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 29, pp. 1368–1396, Mar. 2021.

[12] Tassadaq Hussain, Mandar Gogate, Kia Dashtipour, and Amir Hussain, "Towards intelligibility-oriented audio-visual speech enhancement," in *The Clarity Workshop on Machine Learning Challenges for Hearing Aids (Clarity-2021)*, 11 2021.

[13] Szu wei Fu, Yu Tsao, Hsin-Te Hwang, and Hsin-Min Wang, "Quality-net: An end-to-end non-intrusive speech quality assessment model based on blstm," in *Proc. Interspeech 2018*, 2018, pp. 1873–1877.

[14] Yong Feng and Fei Chen, "Nonintrusive objective measurement of speech intelligibility: A review of methodology," *Biomedical Signal Processing and Control*, vol. 71, pp. 103204, Jan. 2022.

[15] Chen-Chou Lo, Szu-Wei Fu, Wen-Chin Huang, Xin Wang, Junichi Yamagishi, Yu Tsao, and Hsin-Min Wang, "MOSNet: Deep Learning-Based Objective Assessment for Voice Conversion," in *Proc. Interspeech 2019*, 2019, pp. 1541–1545.

[16] Ryandhimas E Zezario, Szu-Wei Fu, Chiou-Shann Fuh, Yu Tsao, and Hsin-Min Wang, "Stoi-net: A deep learning based non-intrusive speech intelligibility assessment model," in *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2020, pp. 482–486.

[17] Ryandhimas E Zezario, Szu-Wei Fu, Fei Chen, Chiou-Shann Fuh, Hsin-Min Wang, and Yu Tsao, "Deep learning-based non-intrusive multi-objective speech assessment model with cross-domain features," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 54–70, 2022.

[18] Martin Cooke, Jon Barker, Stuart Cunningham, and Xu Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, Nov. 2006.

[19] Emmanuel Vincent Jon Barker, Ricard Marxer and Shinji Watanabe, "The third 'chime' speech separation and recognition challenge: Dataset, task and baselines.," in *2015 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU 2015)*, Dec 2015.