Interspeech 2024
1-5 September 2024, Kos, Greece

Edged based Audio-Visual Speech enhancement demonstrator

Song Chen[1,2], *Mandar Gogate*[2], *Kia Dashtipour*[2], *Jasper Kirton-Wingate*[2], *Adeel Hussain*[2], *Faiyaz Doctor*[4], *Tughrul Arslan*[3], *Amir Hussain*[2]

[1] College of Mechanical and Electrical Engineering, Anhui Jianzhu University, Hefei, China
[2] School of Computing, Engineering and the Built Environment, Edinburgh Napier University, UK
[3] School of Engineering, The University of Edinburgh, UK
[4] School of Computer Science and Electronic Engineering, University of Essex, U.K.

chen.ahjzedu@gmail.com, m.gogate@napier.ac.uk, k.dashtipour@napier.ac.uk, jasper.kirton-wingate@napier.ac.uk, adeel.hussain@napier.ac.uk, fdocto@essex.ac.uk, tughrul.arslan@ed.ac.uk, a.hussain@napier.ac.uk

Abstract

Difficulty understanding speech in noisy environments presents a significant challenge for individuals with hearing loss and is a primary factor contributing to non-adherence to hearing aid use. Recent technological advancements integrating artificial intelligence, machine learning, and smartphone technology hold promise in advancing and customizing hearing healthcare. A proposed solution is a portable hearing assistive system designed for speech enhancement in noisy settings. We anticipate that this system will enhance the auditory experience of hearing aid users. The system leverages a mobile phone's camera, microphone, and speaker, ensuring ease of portability. Raw video and audio data are stored locally on the phone and processed by the device's processor alongside an audio-visual speech enhancement algorithm. This algorithm is capable of identifying voice signals and lip movements using a lightweight deep neural network model, thereby optimizing memory efficiency required for real-time processing

Index Terms: Speech Enhancement, Deep Learning, Audio-Visual

1. Introduction

With the advancement of the information age, the use of smartphones is becoming increasingly popular, and voice communication in daily life is expanding significantly. However, in practical applications, ambient noise often impairs the accuracy of speech recognition [1, 2, 3]. Therefore, it is crucial to develop a voice noise reduction program tailored for mobile phones. This paper introduces an optional voice noise reduction program designed specifically for mobile phone usage. The program efficiently reduces voice noise through a designated noise reduction algorithm, thereby offering users a clearer and more accurate voice communication experience.

2. System design

2.1. System structure

This program is designed to reduce audio noise in video recordings made on a mobile phone. The system first captures environmental information using the mobile phone's audio acquisition device, distinguishing whether only the audio changes, or both video and audio synchronize changes. It then selects the appropriate noise reduction algorithm based on the user's environment. The system's structure diagram is illustrated in Figure 1. Upon launching the software, users are prompted to grant access to the phone's camera and microphone, followed by video and audio recording, and subsequent data processing. The program's flowchart is depicted in Figure 2.

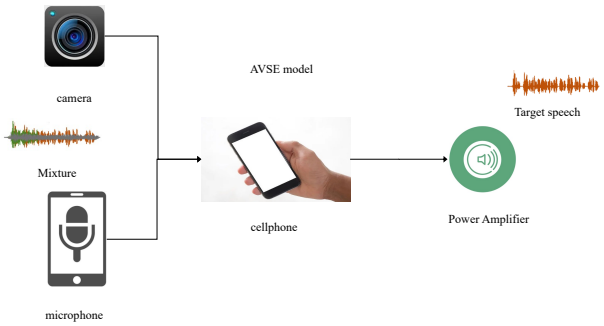

Figure 1: *System structure diagram*

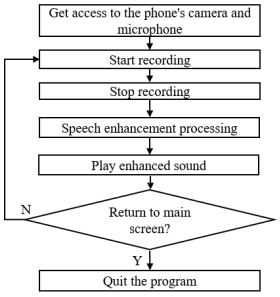

Figure 2: *System structure diagram*

2.1.1. *User interface design*

The user interface of this program consists of three modules: start recording, stop recording, and run the noise reduction program to output enhanced voice. Users can utilize the AVSE

(Audio-Visual Speech Enhancement) voice noise reduction program.

3. Algorithm Design

3.1. Data preprocessing

First the mobile phone collect data and preprocess it to extract lip information, identify the lip area in each frame, and remove background noise in video or audio.

3.2. Feature extraction and selection

This algorithm employs a deep learning model to extract features from videos, leveraging spatio-temporal information within the video data. Additionally, attentional intelligence is utilized to select features and enhance the algorithm's focus on key information.

3.3. Model construction

The model consists of a three-layer three-dimensional convolutional neural network (CNN) and Rectified Linear Unit (ReLU) is used as the activation function. Each CNN layer has a maximum pooling layer behind it, which helps to downsample the spatial dimension of the output feature map. After the CNN layer, the output is flattened into a one-dimensional vector. Next, two layers of bidirectional Long short-term memory (LSTM) are added to the model. The LSTM layer is a recurrent neural network (RNN) layer that captures time dependencies in sequence data [4].

3.4. Test

The mobile phone camera collects real-time image data, and the microphone collects voice data. Mobile APP provides users with operation interface and data processing functions.

- Record several audio and video clips on your phone with different background noises.
- Stop recording.
- Use noise reduction algorithm. The mobile phone uses the software's AVSE program to denoise the audio, and finally obtains the target voice after denoising, and the voice signal is output by the player.

3.5. Result Analysis

In this paper, 1000 groups of speech signals were collected for testing, and the test results were analyzed using two objective speech evaluation methods: STOI (Short-Term Objective Intelligibility) and PESQ (Perceptual Evaluation of Speech Quality). The test results demonstrate that the average values of PESQ and STOI are higher with AVSE compared to audio-only reduction. Therefore, it can be concluded that the AVSE algorithm achieves better noise reduction effectiveness and higher speech quality than the audio-only algorithm, as shown in Table 1.

4. Conclusion

This paper introduces a voice noise reduction program designed for mobile phones. The program utilizes the AVSE (Audio-Visual Speech Enhancement) algorithm to effectively reduce background noise and improve voice quality. Test results demonstrate that the program achieves good noise reduction effectiveness and performance stability.

Table 1: *Test results*

Index	values
STOI_original	0.631
STOI_A-only	0.743
STOI_AVSE	0.892
PESQ_original	1.152
PESQ_A-only	1.266
PESQ_AVSE	2.2
PESQ_original	1.152
PESQ_A-only	1.266
PESQ_AVSE	2.2

No video data is transmitted to any remote web server. Local processing preserves privacy and reduces the time required for audio upload and download. The optimized system is capable of providing clearer sound for individuals with hearing loss and can be beneficial in challenging noisy environments such as restaurants, train stations, meeting rooms, and other similar settings. We anticipate that future iterations of this technology will be developed for use with smart glasses or other accessories to enhance usability.

5. ACKNOWLEDGEMENTS

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

6. References

[1] K. Tan, X. Zhang, and D. Wang, "Deep learning based real-time speech enhancement for dual-microphone mobile phones," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 29, pp. 1853–1863, 2021.

[2] M. Gogate, K. Dashtipour, and A. Hussain, "Robust real-time audio-visual speech enhancement based on dnn and gan," *IEEE Transactions on Artificial Intelligence*, 2024.

[3] A. L. A. Blanco, C. Valentini-Botinhao, O. Klejch, M. Gogate, K. Dashtipour, A. Hussain, and P. Bell, "Avse challenge: Audio-visual speech enhancement challenge," in *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2023, pp. 465–471.

[4] M. Gogate, K. Dashtipour, A. Adeel, and A. Hussain, "Cochleanet: A robust language-independent audio-visual model for real-time speech enhancement," *Information Fusion*, vol. 63, pp. 273–285, 2020.