Towards Next-Generation Lip-Reading Driven Hearing-Aids: A preliminary Prototype Demo

Ahsan Adeel, Mandar Gogate, Amir Hussain

Department of Computing Science and Mathematics, Faculty of Natural Sciences, University of Stirling, UK

E-mail: {aad, mgo, ahu}@cs.stir.ac.uk

Abstract

Speech enhancement aims to enhance the perceived speech quality and intelligibility in the presence of noise. Classical speech enhancement methods are mainly based on audio only processing which often perform poorly in adverse conditions, where overwhelming noise is present. This paper presents an interactive prototype demo, as part of a disruptive cognitively-inspired multimodal hearing-aid being researched and developed at Stirling, as part of an EPSRC funded project (COG-AVHEAR). The proposed technology contextually utilizes and integrates multimodal cues such as lip-reading, facial expressions, gestures, and noisy audio, to further enhance the quality and intelligibility of the noise-filtered speech signal. However, the preliminary work presented in this paper has used only lip-reading and noisy audio. Lip-reading driven deep learning algorithms are exploited to learn noisy audio-visual to clean audio mappings, leading to enhanced Weiner filtering for more effective noise cancellation. The term context-aware signifies the device's learning and adaptable capabilities, which could be exploited in a wide-range of real-world applications, ranging from hearing-aids, listening devices, cochlear implants and telecommunications, to need for ear defenders in extreme noisy environments. Hearing-impaired users could experience more intelligible speech by contextually learning and switching between audio and visual cues. The preliminary interactive Demo employs randomly selected, real noisy speech videos from YouTube to qualitatively benchmark the performance of the proposed contextual audio-visual approach against a state-of-the-art deep learning based audio-only speech enhancement method.

Index Terms: Speech Enhancement, Cognitively-Inspired, Multimodal, Lip-Reading, Sentiment Features, Deep Learning

1. Introduction

The extensive speech enhancement requirement in wide-range of real-world applications and the advent of advanced signal processing methods have opened new ways to explore and develop more efficient and advanced speech processing technologies. Over the past few decades, several speech enhancement methods have been proposed, ranging from the state-of-the-art statistical, analytical, and classical optimization approaches, to advanced deep learning based methods. The classic speech enhancement methods are mainly based on audio only processing [1][2][3][4]. Recently, researchers have also proposed deep learning based advanced speech recognition [5] and enhancement [6] methods. However, most of the speech enhancement methods are based on single channel (audio only) processing, which often perform poorly in adverse conditions [7]. In this research, we aim to leverage the audio-visual (AV) nature of speech which is inherently multimodal and capable of improving intelligibility in noise [8][9][10] [11][12][13] and have modelled lip-reading as a regression problem for speech enhancement. Specifically, we envision developing a multimodal hearing device that significantly improves both speech quality and intelligibility in everyday and extreme listening environments.

The inherent multimodal nature of the speech is well established in the literature and it is well understood that how speech is produced by the vibration of vocal chords with respect to the articulatory organs configuration. The correlation between the articulatory organs (visible properties) and speech has been shown in several ways in literature using biological, psychological, and mathematical experiments [8][11][10][14]. Therefore, the clear visibility of some of the articulatory organs such as lips, teeth, and tongue could be effectively utilized to extract the clean speech out of the noisy audio signal. In addition, the visual features such as facial expressions and body language also play a vital role in speech perception. The major advantage of using visual cues for generating clean audio feature is their natural noise immunity (i.e. visual speech representation always remains unaffected by the acoustic noise) [15].

In the literature, most of the proposed lip-reading approaches have modelled lip-reading as a classification problem for speech recognition. However, limited work has been conducted to model lip-reading as a regression problem for speech enhancement. In this research, we envision cognitively-inspired, context-aware multimodal speech processing technology based on lip-reading regression model. The technology is aimed at helping users in noisy environments, by contextually learning and switching between audio and visual cues. The initial aim of this research is to develop a proof of concept prototype of the audio-visual hearing-aid technology. An early prototype demo has been developed for online evaluation and feedback. The prototype demo is benchmarked against the state-of-the-art audio-only approach (reported in IEEE Spectrum Magazine 2017) that applied cutting-edge machine learning based on deep neural networks [16]. The preliminary objective and subjective testing has revealed the potential and reliability of the proposed technology as compared to the state-of-the-art audio only speech enhancement techniques.

The rest of the paper is organized as follows: Section II presents an overview of the proposed context-aware multimodal speech processing technology, including multimodal feature extraction, audiovisual mapping, and noisy audio filtering methods. Section III presents the qualitative speech enhancement testing of the proposed technology. Finally, Section IV concludes this work.

2. Proposed Next-Generation Context-Aware Multimodal Technology

The proposed novel cognitively-inspired, multimodal approach aims to contextually exploit and integrates multimodal cues, such as lip-reading and audio features. The term context-aware signifies the technology's contextual learning and adaptable capabilities, which can be employed in next-generation multi-modal applications, including assistive technology such as hearing-aids, cochlear implants, and listening devices. The disruptive technology is capable of contextually enhancing speech intelligibility in extreme noisy environments, so can also be useful for users in situations where ear defenders are worn, such as emergency and disaster response and battlefield environments. In applications such as teleconferencing, video signals could be used to filter and enhance acoustic signals arriving at the receiver-end. People with visual impairment who are unable to see visual cues can also benefit from the proposed technology, particularly in emergency situations. Preliminary simulation results, including an interactive online prototype, demonstrate the potential of the proposed multimodal speech enhancement technology for enabling transformative applications in extreme environments. An abstracted processing of the proposed technology is depicted in Fig. 1, where the multimodal (audio-visual) system has integrated the aforementioned cues for speech processing. The proposed (audio-visual) system extracts the available multimodal features contextually to estimate the clean audio features and then exploits them for real-time speech enhancement. More technical details are comprehensively presented in [17]

2.1. Dataset

For preliminary analysis, the widely used Grid [18] and CHiME corpuses [19] are used to extract lip-reading and noisy audio features. The visual features are extracted using Grid Corpus, whereas CHiME2 is used for extracting audio features. The work could easily be extended to include other visual features such as gestures, facial expressions, body language etc., which is a part of future work and will be presented in upcoming publications. From both the corpuses, an audiovisual (AV) dataset is built by preprocessing the utterances and extracting the audio and visual features. The preprocessing includes sentence alignment and incorporation of prior visual frames. The sentence alignment is used to remove the silence time from the video to restrict the model from learning redundant or insignificant information. The sentence alignment process enforced the model to learn the correlation between the spoken word and corresponding visual representation, rather than over learning the silence. Secondly, the prior visual frames are used to incorporate the temporal information, which ultimately helped the learning engine to better correlate the visual features to corresponding speech features. The audio and visual features extraction procedure is shown in Fig. 2. The audio feature extraction procedure includes sampling, segmentation, Hamming windowing, Fourier transformation, and FB audio features calculation. The visual feature (lip-reading) extraction procedure includes frames extraction, viola-jones lip detector, object tracker, lip cropping, 2D-DCT/convoluted features. Once the dataset is built, it could then be fed into a deep learning model such as LSTM to learn the correlation between audio and visual features.

2.2. Multimodal Features Extraction

2.2.1. Audio Features

The audio features are extracted using the widely used log-filterbank (FB) vectors and Mel-frequency cepstral coefficients (MFCC). The input audio signal is sampled at 50kHz and segmented into N 16ms frames with 800 samples per frame and 62.5% increment rate. Afterwards, a Hamming window and Fourier transformation are applied to produce 2048-bin power spectrum. Finally, a 23-dimensional log-FB is applied, followed by the logarithmic compression to produce 23-D log-FB signal. For MFCC calculation, DCT of the log-auditory-spectrum is obtained.

2.2.2. Visual Features

The visual features include only lip movements in this preliminary work. The lip movement features are extracted using 2D-DCT based standard and widely used visual feature extraction method. Firstly, the video files are processed to extract a sequence of individual frames. Secondly, the Viola-Jones lip detector [20] is used to identify the Region-of-Interest (ROI) in terms of a bounding box. Finally, the object tracker [21] is used to track the lip regions across the sequence of frames. The visual extraction procedure produced a set of corner points for each frame, where the lip regions are then extracted by cropping the raw image for desired visual features, followed by 2D-DCT calculation. More details are comprehensively presented in [22].

2.3. Audiovisual Mapping and Clean Audio Features Estimation

For successful implementation of the proposed technology, one of the essential steps include the estimation of clean audio power spectrum (i.e. audiovisual speech mapping). The audiovisual speech mapping aims to approximate the audio features given only visual information. In the proposed approach, multimodal features (i.e. lip movements) are mapped to the clean audio features using long-short-term memory network. Once the deep learning models are successfully trained and validated, the model predicts the audio log-FB vectors given only the noisy audio and visual features, which are then exploited by the proposed noisy speech filtering framework for speech enhancement. More detail on AV mapping is comprehensively presented in [22], where multilayer perceptron (MLP) was used for AV mapping.

2.4. Enhanced visually derived Wiener filtering

In signal processing, Wiener Filter is the state-of-the-art filter that helps to produce an estimate of a clean audio signal by linear time-invariant (LTI) filtering of an observed noisy audio signal. In the proposed approach, an enhanced visually derived Wiener filtering (EVWF) for speech enhancement has been used. The EVWF effectively exploited the estimated low dimensional clean audio features (through lip-reading) to estimate the high dimensional clean audio power spectrum. Specifically, the EVWF transformed the estimated low dimensional clean audio features into high dimensional clean audio power spectrum using inverse FB transformation. Afterwards, the Wiener filter is calculated (using the estimated audio features) and applied to the magnitude spectrum of the noisy input audio signal, followed by the inverse fast Fourier transform (IFFT), overlap, and combining processes to produce the enhanced magnitude spec-

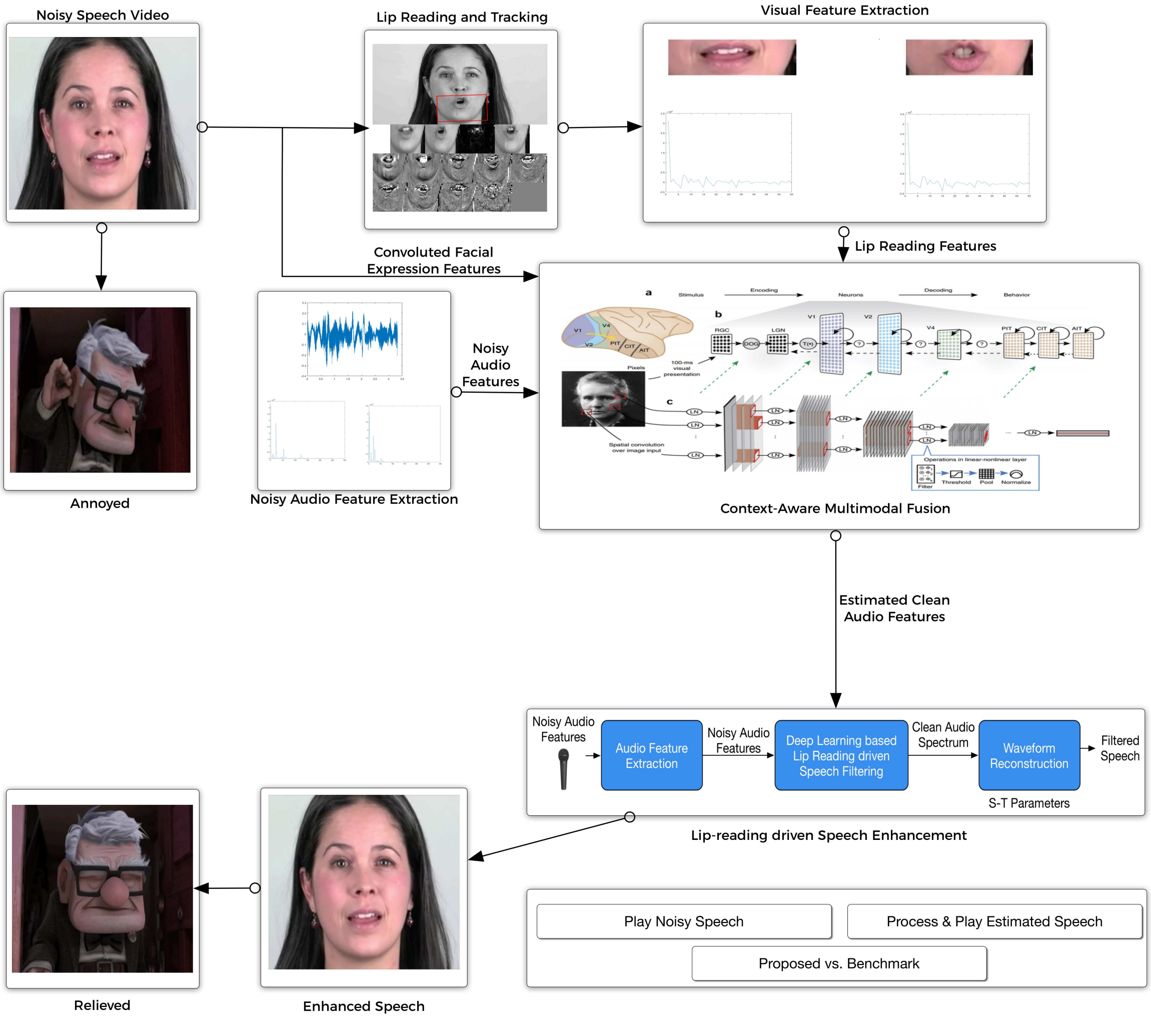

Figure 1: *Prototype Demo: Lip-reading driven deep learning approach for speech enhancement: The device is capable of helping users in noisy environments to experience intelligible clean speech, by contextually learning and switching between audio and visual cues. Audio cues include noisy speech only, whereas visual cues include lip movements.*

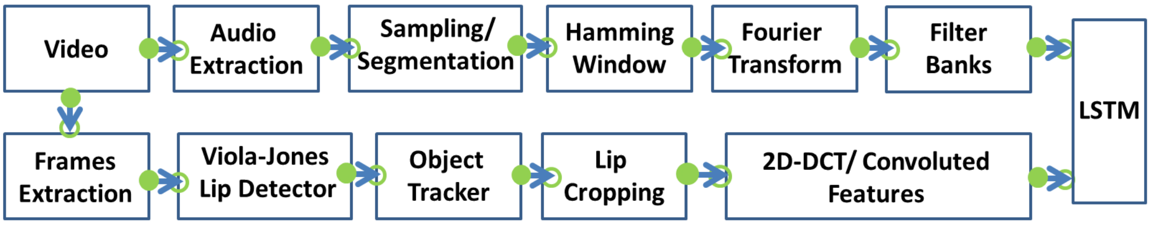

Figure 2: *Audio-Visual Dataset Generation Procedure*

trum.

3. Qualitative Speech Enhancement Testing

The speech enhancement quality of the proposed technology is tested by feeding the randomly selected, real noisy speech videos from YouTube, into the audio-visual speech enhancement system. In the online Demo, the user can listen to the noisy audio, by clicking the *PLAY NOISY SPEECH* button. The visually-enhanced/estimated speech can be perceived by clicking the *PROCESS & PLAY ESTIMATED SPEECH* button. Users can also get a sneak pre-view of behind the scene speech enhancement processing, by clicking individual processing components, such as lip-reading, visual feature extraction, and noisy audio features extraction. The preliminary interactive Demo and the processed multimodal speech demonstrates the potential of context-aware audio-visual hearing-aids, based on big data and deep learning technology.

The second part of the Demo invites listeners to qualitatively compare the new audio-visual approach with a latest audio-only, deep learning benchmark system, recently reported in the IEEE Spectrum Magazine, 2017 [16]. The authors in [16] proposed a DNN-based supervised speech segregation system. The samples of speech before and after enhancement are available at ($http : //cogbid.cs.stir.ac.uk/cogavhearingdemo$) for both DNN-based supervised speech segregation and our proposed multimodal approach. The demo has also demonstrated that how the hearing-aid users perceive the speech. In the sample speech utterances, the proposed multimodal approach has shown better and consistent speech enhancement as compared to the DNN-based supervised speech segregation system. In addition, the proposed multimodal system has recovered both the pitch and clean speech as compared to the DNN-based supervised speech segregation approach, preserving the naturalness of the speech among male/female/infant voices.

4. Conclusion and Future Work

In this paper, an interactive prototype demo is presented as part of a disruptive cognitively-inspired multimodal speech processing technology. In the online demo, the preliminary speech enhancement results and comparisons with the state-of-the-art deep learning based audio-only method have demonstrated the potential and reliability of the proposed speech processing technology. We believe that the disruptive technology is capable of contextually enhancing speech intelligibility in everyday life and even in extreme noisy environments such as emergency and disaster response, and battlefield environments. In addition, the technology could also be utilized in applications such as teleconferencing, where video signals could be used to filter and enhance acoustic signals arriving at the receiver-end. In future, we intend to investigate the performance of the proposed speech processing technology in more realistic real-world scenarios.

5. Acknowledgements

This work was supported by the UK Engineering and Physical Sciences Research Council (EPSRC) Grant No. EP/M026981/1 (CogAVHearinghttp://cogavhearing.cs.stir.ac.uk). In accordance with EPSRC policy, all experimental data used in the project simulations is available at http://hdl.handle.net/11667/81. The authors would also like to gratefully acknowledge Dr Andrew Abel at Xi'an Jiaotong-Liverpool University for building and providing the audio-visual dataset. Lastly, we gratefully acknowledge the support of NVIDIA Corporation for donating the Titan X Pascal GPU for this research.

6. References

[1] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 27, no. 2, pp. 113–120, 1979.

[2] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[3] W. K. Pratt, "Generalized wiener filtering computation techniques," *IEEE Transactions on Computers*, vol. 100, no. 7, pp. 636–641, 1972.

[4] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 33, no. 2, pp. 443–445, 1985.

[5] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[6] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 23, no. 1, pp. 7–19, 2015.

[7] J. Benesty, J. Chen, and E. A. Habets, *Speech enhancement in the STFT domain*. Springer Science & Business Media, 2011.

[8] W. H. Sumby and I. Pollack, "Visual contribution to speech intelligibility in noise," *The journal of the acoustical society of america*, vol. 26, no. 2, pp. 212–215, 1954.

[9] N. P. Erber, "Interaction of audition and vision in the recognition of oral speech stimuli," *Journal of Speech, Language, and Hearing Research*, vol. 12, no. 2, pp. 423–425, 1969.

[10] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," 1976.

[11] Q. Summerfield, "Use of visual information for phonetic perception," *Phonetica*, vol. 36, no. 4-5, pp. 314–331, 1979.

[12] A. MacLeod and Q. Summerfield, "Quantifying the contribution of vision to speech perception in noise," *British journal of audiology*, vol. 21, no. 2, pp. 131–141, 1987.

[13] Q. Summerfield, "Lipreading and audio-visual speech perception," *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, vol. 335, no. 1273, pp. 71–78, 1992.

[14] M. L. Patterson and J. F. Werker, "Two-month-old infants match phonetic information in lips and voice," *Developmental Science*, vol. 6, no. 2, pp. 191–196, 2003.

[15] I. Almajai and B. Milner, "Visually derived wiener filters for speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 6, pp. 1642–1651, 2011.

[16] J. Chen, Y. Wang, S. E. Yoho, D. Wang, and E. W. Healy, "Large-scale training to increase speech intelligibility for hearing-impaired listeners in novel noises," *The Journal of the Acoustical Society of America*, vol. 139, no. 5, pp. 2604–2612, 2016.

[17] A. Adeel, M. Gogate, B. Whitmer, and R. Watt, "A novel lip-reading driven deep learning approach for speech enhancement," *Emerging Topics in Computational Intelligence, IEEE Transactions on*, 2017.

[18] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, 2006.

[19] H. Christensen, J. Barker, N. Ma, and P. D. Green, "The chime corpus: A resource and a challenge for computational hearing in multisource environments." in *Interspeech*. Citeseer, 2010, pp. 1918–1921.

[20] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, vol. 1. IEEE, 2001, pp. I–I.

[21] D. A. Ross, J. Lim, R.-S. Lin, and M.-H. Yang, "Incremental learning for robust visual tracking," *International Journal of Computer Vision*, vol. 77, no. 1-3, pp. 125–141, 2008.

[22] A. Abel, R. Marxer, J. Barker, R. Watt, B. Whitmer, P. Derleth, and A. Hussain, "A data driven approach to audiovisual speech mapping," in *Advances in Brain Inspired Cognitive Systems: 8th International Conference, BICS 2016, Beijing, China, November 28-30, 2016, Proceedings 8*. Springer, 2016, pp. 331–342.