INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Visual Speech In Real Noisy Environments (VISION): A Novel Benchmark Dataset and Deep Learning-based Baseline System

Mandar Gogate, Kia Dashtipour, Amir Hussain

School of Computing, Edinburgh Napier University, Scotland, UK
m.gogate@napier.ac.uk , k.dashtipour@napier.ac.uk , a.hussain@napier.ac.uk

Abstract

In this paper, we present VIsual Speech In real nOisy eNvironments (VISION), a first of its kind audio-visual (AV) corpus comprising 2500 utterances from 209 speakers, recorded in real noisy environments including social gatherings, streets, cafeterias and restaurants. While a number of speech enhancement frameworks have been proposed in the literature that exploit AV cues, there are no visual speech corpora recorded in real environments with a sufficient variety of speakers, to enable evaluation of AV frameworks' generalisation capability in a wide range of background visual and acoustic noises. The main purpose of our AV corpus is to foster research in the area of AV signal processing and to provide a benchmark corpus that can be used for reliable evaluation of AV speech enhancement systems in everyday noisy settings. In addition, we present a baseline deep neural network (DNN) based spectral mask estimation model for speech enhancement. Comparative simulation results with subjective listening tests demonstrate significant performance improvement of the baseline DNN compared to state-of-the-art speech enhancement approaches.

Index Terms: Speech Enhancement, Audio-Visual Fusion, VISION Corpus, Deep Learning, Multi-modal Speech Processing, Listening Tests

1. Introduction

Approximately 360 million people in the world currently suffer from a debilitating hearing loss [1]. By 2030, these numbers are expected to rise by 50%. The most common age-related and noise-induced hearing losses are progressive and neither curable nor reversible. People with serious hearing-issues often find themselves socially isolated leading to depression and a range of other negative consequences. Hearing aids and cochlear implants are the most widely used devices for compensating hearing loss. However, even sophisticated listening devices cause huge problems for the hearing impaired, as they often make the speech more audible but do not always restore intelligibility in noisy social situations [2]. Human beings in such settings are known to exploit the audio-visual nature of speech to contextually suppress background noise and focus on the target speech.

In addition, it is well known that visual information help disambiguate the phonological ambiguity. For example, in speech recognition, people integrate AV cues in order to better perceive speech. This phenomenon was observed in the McGurk effect [3] where a visual $/ga/$ with a voiced $/ba/$ is perceived as $/da/$ by most subjects. In particular, the visual cues provide information on the place of articulation [4] and muscle movements that can often aid to differentiate between speech with similar acoustic sounds (e.g., the unvoiced consonants $/p/$ and $/k/$).

In the literature, extensive research has been carried out, inspired by the unique human hearing ability, to develop single channel and multi channel AV speech enhancement (SE) frameworks [5, 6, 7, 8, 9, 10, 11]. Most of these frameworks use a synthetic mixture of clean speech and noises to evaluate the enhancement quality and intelligibility. However, a synthetic mixture does not depict everyday noisy settings, since in real mixtures the speech is often reverberantly mixed with multiple competing background sources and the Lombard effect is observed.

Although, there exist a number of controlled AV speech corpora [12, 5, 13, 14, 15] with limited vocabulary and noise types, there is need for more realistic AV speech data comprising a wide variety of speakers, competing noises and visual imperfections. Recently, Gogate et al. [10] introduced ASPIRE, a AV speech corpus recorded in real noisy settings such as restaurants and cafeterias. However, the main limitation with the corpus is the lack of speakers variety as it only consists of three speakers' recorded uttering sentences from the limited vocabulary Grid corpus[16]. Further, the noisy visual recording conditions are ideal with uniform lighting and minimum speaker movements. We envisage future AV hearing devices will not only be expected to generalise on a large variety of speakers, including non-native English speakers, but also work with noisy visual data including speaker movement, imperfect lighting and various levels of background noises. To the best of our knowledge, there currently exists no medium vocabulary AV corpus that comprises a sufficient number of speakers along with a variety of acoustic and visual noises.

In this study, we introduce VISION, a first of its kind medium vocabulary, binaural AV corpus comprising 2500 utterances from 209 speakers, recorded in real noisy settings including social gatherings, streets, cafeterias and restaurants. The corpus can serve as a test or validation set for the development of AV speech enhancement/separation, speech recognition and lip reading systems. In addition, we present a baseline deep neural network (DNN) based spectral mask estimation model for speech enhancement. The DNN model integrates convolutional feature extraction with long short-term memory (LSTM) to take into account the temporal dynamics and long-term contextual dependencies of AV data, and is trained on a synthetic mixture of GRID [12] and CHiME3 [19] corpus. The DNN learns the correlation between noisy AV cues and the ideal binary mask (IBM) to estimate noise and speech dominant regions. Finally, the enhanced speech is re-synthesised by combining the processed signal across frequency channels. We exploit the VISION corpus to demonstrate superior speech quality resulting from application of our proposed baseline over state-of-the-art A-only speech enhancement approaches, including spectral subtraction (SS) and linear minimum mean square error (LMMSE), as well as recent DNN based AV speech enhancement models, including the new CochleaNet[10]).

The rest of the paper is organised as follows: Section 2 introduces the VISION corpus. Section 3 presents the baseline

http://dx.doi.org/10.21437/Interspeech.2020-2935

Figure 1: *Sample Frames from the VISION Corpus*

Dataset	Modality	Speakers	Real Noisy Environment	Noise types
COSINE [17]	A-only	133	Yes	Cafeteria, Streets
VOICES [18]	A-only	300	No	Television, Speech
GRID [12]	AV	34	-	No noise
Mandarin Sentences [5]	AV	1	-	No noise
AVSPEECH [13]	AV	-	-	No noise
BANCA [14]	AV	208	Yes	Speech noise only
AVICAR [15]	AV	100	Yes	Car noise only
ASPIRE [10]	AV	3	Yes	Cafeteria, Restaurant, Speech
VISION	**AV**	**209**	**Yes**	**Social gathering, Street, Cafeteria, Speech**

Table 1: *Comparison of VISION with state-of-the-art A-only and AV Corpora*

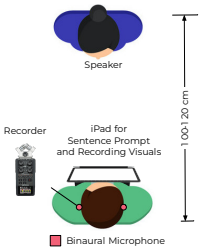

Figure 2: *Recording Setup*

DNN based AV SE model. Section 4 discusses comparative experimental results and finally, Section 5 concludes this work and proposes some future directions.

2. Vision Corpus

In the literature, as illustrated in Table 1, extensive research has been carried out to develop A-only and AV corpora for speech enhancement. It can be seen that previous AV corpora recorded in real noisy environments consist of limited noise types (visual and acoustic) and speakers. In this section, we present VISION, a first of its kind medium vocabulary AV speech corpus recorded in real noisy environments to support evaluation of AV SE frameworks. Fig. 1 shows a few sample frames from the VISION corpus.

2.1. Sentence Design

The VISION corpus follows the same sentence format as IEEE 'Harvard' sentences [20]. The sentences consist of 720 phonetically balanced sentences. Most of the words are monosyllabic (e.g. cat, break, bus), with exception of a few words that are longer (e.g. shimmered, friendly). The sentences are selected to represent various phonemes of English in accordance with their frequency of occurrence. The IEEE sentences were used because of the low word context predictability, standardised sentence structure and length.

2.2. Speaker Population

209 speakers (105 male and 104 female) contributed in the corpus. The speakers age ranged between 18 to 55. Most of the participants spoke English as their first language. A limited number of participants were recorded with English as their second language. The corpus comprises a total 2500 utterances (around 3 hours) recorded in a range of real world noisy environments. The distribution of the noisy environment is depicted in Table. 2.

2.3. Collection

The VISION corpus has been recorded in real noisy environments including busy cafeterias, restaurants, streets and social gatherings. The recorded setup is depicted in Fig. 2. An Apple iPad mini was used to prompt the sentences to the user and record the video. The listener was holding the iPad opposite to the speaker at an approximate distance of 100 cm. In addition, a high quality Sennheiser binaural microphone was used to record the audio. A custom iOS application was developed to simultaneously prompt the sentence and to record the video (consisting of speaker's face and background surrounding) and audio from the iPad and binaural microphone respectively. This ensured synchronisation between the binaural audio and video.

The purpose of the research was first explained in detail to the speakers prior to recording. Initially, the participants were trained with a few utterances. They were allowed to repeat sentences if any mistake identified by the speaker or listener. In total, 2500 utterances were collected in real noisy environments and around 5% of the utterances were re-recorded.

3. Proposed Baseline DNN-driven AV Speech Enhancement System

This section describes the proposed baseline DNN architecture, inspired by CochleaNet[10], depicted in Fig. 3.

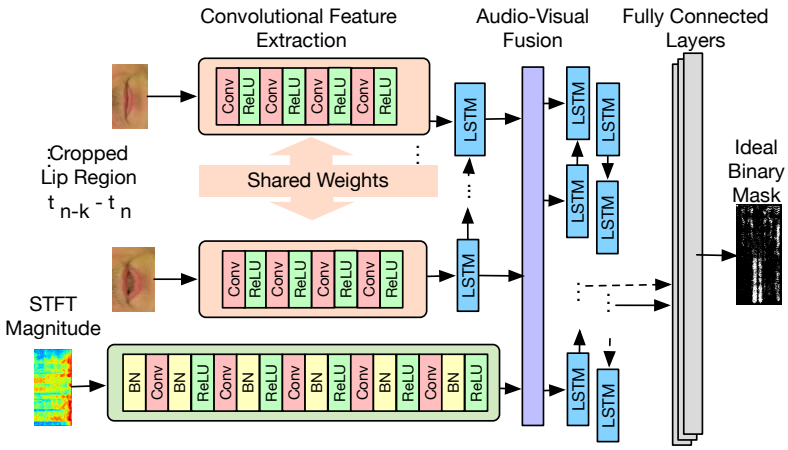

Figure 3: *Proposed DNN based Speech Enhancement Baseline Model*

Noisy Environment	# utterances
University	734
Street	633
Social gathering	677
Cafeteria/Restaurant	456

Table 2: *VISION corpus: Noisy environments distribution*

3.1. Audio Feature Extraction

The input temporal short-time fourier transform (STFT) magnitude features are first fed into a batch normalisation layer. The batch normalisation output is then fed into the audio-feature extraction part of the network with 5 convolutional layers. The first four convolutional layers consist of 64 filters and the last convolutional layer consists of 8 filters. Each filter is of size 3 x 3. After each convolutional layer, batch normalisation and ReLU activation is applied. The output of the last convolutional layer is fed into the AV fusion part of the framework as shown in Fig. 3.

3.2. Video Feature Extraction

The cropped temporal lip images are fed into the visual-feature extraction part of the framework with four convolutional layers including 32, 48, 64 and 96 filters respectively. Each convolutional filter is of size 3 x 3. ReLU activation is used after each convolutional layer. The output of the last convolutional layer is fed into a LSTM layer with 512 units. The visual features at 25 frames-per-second (fps) are upsampled by a factor of 3 to match the audio feature sampling rate i.e. 75 vector-per-second (VPS). The output of the LSTM layer is fed into the AV fusion part of the framework. Note that the convolutional weights are shared across the temporal dimension.

3.3. Audio-Visual Fusion

The features extracted from audio and video streams are concatenated across the time dimension and fed into a bidirectional LSTM layer with 512 units. Note that the visual features are upsampled to match the audio VPS. The output of the bidirectional LSTM is concatenated and fed into 2 fully connected layers with 526 neurons and ReLU activation. Finally, the output of the last connected layer is fed into an output layer with sigmoid activation. Note that the A-only baseline model is constructed by removing the video-feature extraction part of the network.

4. Experimental Results

4.1. Synthetic AV Dataset

For training and evaluation, a widely used benchmark Grid corpus [16] was randomly mixed with non-stationary noises from the 3rd CHiME challenge (CHiME 3)[19], consisting of bus, cafeteria, street, and pedestrian noises, for SNRs ranging from -12dB to 12dB with a step size of 3dB. The dataset was divided into 21, 4 and 8 speakers for training, development and evaluation respectively. All utterances from each speaker were used in the training, development and evaluation set. To avoid the dominance of audio modality during multimodal training, 25% of the utterances were mixed with speech from the same speaker. Note that the VISION corpus cannot be used for training SE systems, since a corpus mainly consisting of noisy recordings and a clean reference signal is required for computing the IBM and training the DNN. Hence, a synthetic AV dataset (GRID + CHiME 3) was used here.

4.2. Preprocessing

Audio: The audio signals were resampled at 16 kHz and a mono-channel used for processing. The resampled audio signals were segmented into 65 milliseconds (ms) frames and 20% increment rate. A hanning window and STFT were applied to the segmented audio to produce a 526-bin magnitude spectrogram.
Video: The speakers lip images were extracted out of the 25 fps Grid corpus video using a minified dlib [21] model optimised for extracting the lip landmarks. A lip-centred region of aspect ratio 1:2 was extracted using lip landmark points. The extracted region was converted to grey scale and resized to 40 pixels x 80 pixels. Note that the lip sequences were extracted at 25 fps while the audio features were extracted at 75 VPS.

4.3. Experimental Setup

The DNN was trained using TensorFlow library and NVIDIA Titan Xp GPUs. A subset of speakers from Grid CHiME 3 corpus (as described in section 4.1) were used for training/validation of the neural network. When a missing visual

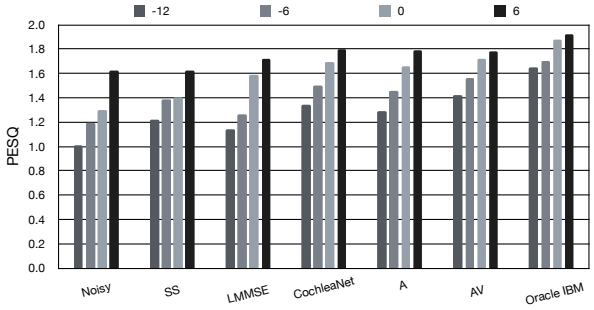

Figure 4: *PESQ scores for SS, LMMSE A-only, AV and Oracle IBM for the synthetic AV test set*

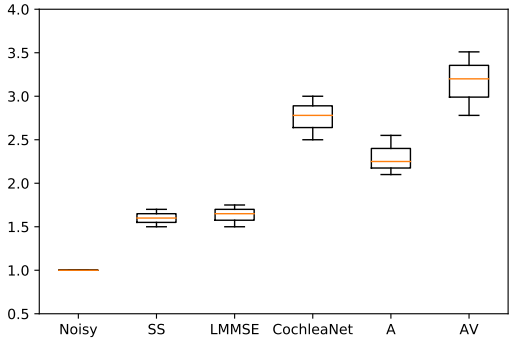

Figure 5: *Results of MOS Subjective listening tests*

frame was encountered, an array of zeros was used. The network was trained to minimise binary cross-entropy using back-propagation with the Adam optimiser [22] for 50 epochs. Early stopping was used if the validation error stopped decreasing after 5 epochs. Note that no thresholding was applied to the sigmoidal outputs of the network i.e. the sigmoidal outputs were considered as the predicted mask.

4.4. Objective testing using a Synthetic AV Dataset

The Perceptual Evaluation of Speech quality (PESQ) method [23] was used for an objective evaluation of the resynthesised speech. Specifically, PESQ was used to evaluate the speech quality computationally. A linear combination of the average disturbance value and the average asymmetrical disturbance values between a reference signal and modified signal were used to calculate PESQ scores ranging from $[-0.50, 4.50]$, indicating the minimum and maximum reconstruction quality. The PESQ results for spectral subtraction (SS), linear minimum mean square error (LMMSE), CochleaNet [10], A-only, AV and Oracle mask are depicted in Fig. 4. It can be seen that the AV model significantly outperformed SS, LMMSE, A-only, and CochleaNet model at low SNRs. Specifically, at -12dB, the AV model achieved a PESQ score of 1.42 compared to 1.14, 1.22, 1.29, and 1.34 achieved by the SS, LMMSE, A-only and CochleaNet models respectively. On the other hand, at 6dB, A-only and AV models achieved PESQ scores of 1.78 and 1.79 compared to 1.62, 1.72 and 1.81 achieved by SS, LMMSE and CochleaNet respectively. It can be seen that, at low SNR, the AV model performs better than the A-only model and other state-of-the-art approaches. However, at higher SNRs the performance of A-only and AV models is similar to other state-of-the-art approaches. Note that the VISION corpus cannot be used for objective evaluation of SE systems, since a corpus mainly consisting of noisy recordings and a clean reference signal is required for objective evaluation.

4.5. Subjective listening tests using the VISION Corpus

In order to assess the effectiveness of the proposed AV framework, subjective listening tests were conducted with self-reported normal-hearing listeners in terms of MOS, using the real noisy VISION corpus. The listeners were presented with 25 randomly selected, enhanced speech utterances and were asked to rate the resynthesised speech on a scale of 1 to 5. The rating choices were: (1) - Very Annoying (Bad), (2) - Annoying (Poor), (3) - Slightly Annoying (Fair) (4) Perceptible but annoying (Good), (5) - Perceptible (Excellent). The proposed AV baseline was compared with A-only DNN, SS and LMME methods. A total of 12 listeners took part in the subjective evaluation sessions. Fig. 5 shows the box plot of listeners' MOS ratings when the noisy speech from VISION is enhanced using the SS, LMMSE, CochleaNet [10], A-only DNN and AV DNN. It can be seen that the AV baseline outperforms the A-only model, SS, and LMMSE based SE methods. The results demonstrate the ability of the proposed baseline to generalise in real noisy settings including reverberations caused by multiple competing background sources. In addition, the subjective test results reveal that an AV model trained on a synthetic mixtures of clean speech and noise generalises well to a real noisy corpus.

5. Conclusions

The VISION corpus[1] provides real noisy binaural AV data consisting of speech signals reverberantly mixed with multiple competing noise sources. The corpus will enable development of robust speech enhancement systems that can generalise to a large number of speakers, multiple competing noises and imperfect visual inputs. The corpus can serve as a test and development benchmark for AV signal processing applications including speech enhancement/separation, speech recognition, and lip reading. By making this corpus publicly available, we aim to promote AV and multimodal speech processing research and applications. In addition, we have shown that the proposed baseline DNN trained on a synthetic mixture of GRID and CHiME 3 noises generalises well on the VISION corpus even though there is not much overlap in the GRID and VISION vocabularies. In future, we intend to extend the VISION corpus with a greater variety of speakers and noisy environments. In addition, we intend to further evaluate the VISION corpus for binaural speech enhancement applications.

6. Acknowledgements

This work was funded by a Doctoral Research Award from Edinburgh Napier University, UK. Prof A. Hussain would like to acknowledge the support of the UK Engineering and Physical Sciences Research Council (EPSRC) grants: EP/M026981/1, EP/T021063/1 and EP/T024917/1. Finally, we would like to acknowledge all the volunteers and support staff involved in collection of the VISION corpus.

[1]The Vision Corpus is available on the project website: vision-corpus.github.io

7. References

[1] A. C. Davisa and H. J. Hoffmanb, "Hearing loss: rising prevalence and impact," *Bull World Health Organ*, vol. 97, pp. 646–646A, 2019.

[2] N. A. Lesica, "Hearing aids: Limitations and opportunities," *The Hearing Journal*, vol. 71, no. 5, pp. 43–46, 2018.

[3] L. D. Rosenblum, M. A. Schmuckler, and J. A. Johnson, "The mcgurk effect in infants," *Perception & Psychophysics*, vol. 59, no. 3, pp. 347–357, 1997.

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

[5] J.-C. Hou, S.-S. Wang, Y.-H. Lai, Y. Tsao, H.-W. Chang, and H.-M. Wang, "Audio-visual speech enhancement using multimodal deep convolutional neural networks," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 2, no. 2, pp. 117–128, 2018.

[6] A. Gabbay, A. Shamir, and S. Peleg, "Visual speech enhancement," in *Interspeech*. ISCA, 2018, pp. 1170–1174.

[7] M. Gogate, A. Adeel, R. Marxer, J. Barker, and A. Hussain, "Dnn driven speaker independent audio-visual mask estimation for speech separation," *Proc. Interspeech 2018*, pp. 2723–2727, 2018.

[8] A. Adeel, M. Gogate, A. Hussain, and W. M. Whitmer, "Lipreading driven deep learning approach for speech enhancement," *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2019.

[9] A. Adeel, M. Gogate, and A. Hussain, "Towards next-generation lip-reading driven hearing-aids: A preliminary prototype demo," in *International Workshop on Challenges in Hearing Assistive Technology (CHAT-2017), Stockholm University, August 19th, Collocated with Interspeech 2017*, 2017.

[10] M. Gogate, K. Dashtipour, A. Adeel, and A. Hussain, "Cochleanet: A robust language-independent audio-visual model for speech enhancement," *Information Fusion*, 2020.

[11] M. Gogate, K. Dashtipour, P. Bell, and A. Hussain, "Dnn driven binaural audio visual speech separation," in *2020 International Joint Conference on Neural Networks (IJCNN 2020)*, 2020.

[12] 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.

[13] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. T. Freeman, and M. Rubinstein, "Looking to listen at the cocktail party: a speaker-independent audio-visual model for speech separation," *ACM Transactions on Graphics (TOG)*, vol. 37, no. 4, p. 112, 2018.

[14] E. Bailly-Bailliére, S. Bengio, F. Bimbot, M. Hamouz, J. Kittler, J. Mariéthoz, J. Matas, K. Messer, V. Popovici, F. Porée *et al.*, "The banca database and evaluation protocol," in *International conference on Audio-and video-based biometric person authentication*. Springer, 2003, pp. 625–638.

[15] B. Lee, M. Hasegawa-Johnson, C. Goudeseune, S. Kamdar, S. Borys, M. Liu, and T. Huang, "Avicar: Audio-visual speech corpus in a car environment," in *Eighth International Conference on Spoken Language Processing*, 2004.

[16] 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.

[17] A. Stupakov, E. Hanusa, J. Bilmes, and D. Fox, "Cosine-a corpus of multi-party conversational speech in noisy environments," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2009, pp. 4153–4156.

[18] C. Richey, M. A. Barrios, Z. Armstrong, C. Bartels, H. Franco, M. Graciarena, A. Lawson, M. K. Nandwana, A. Stauffer, J. van Hout *et al.*, "Voices obscured in complex environmental settings (voices) corpus," *arXiv preprint arXiv:1804.05053*, 2018.

[19] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'chime'speech separation and recognition challenge: Dataset, task and baselines," in *Automatic Speech Recognition and Understanding (ASRU), 2015 IEEE Workshop on*. IEEE, 2015, pp. 504–511.

[20] E. Rothauser, "Ieee recommended practice for speech quality measurements," *IEEE Trans. on Audio and Electroacoustics*, vol. 17, pp. 225–246, 1969.

[21] D. E. King, "Dlib-ml: A machine learning toolkit," *Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

[22] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[23] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221)*, vol. 2. IEEE, 2001, pp. 749–752.