WADA-W: A Modified WADA SNR Estimator for Audio-Visual Speech Recognition

Thum Wei Seong, M. Z. Ibrahim, and D. J. Mulvaney

***Abstract*—One of the main challenges in speech recognition is developing systems that are robust to contamination by intrusive background noise. In audio-visual speech recognition (AVSR), audio information is augmented by visual information in order to help improve the performance of speech recognition, particularly when the audio modality is so significantly corrupted by background noise and it becomes hard to differentiate the original speech signal from the noise. The signal-to-noise ratio (SNR) can be used to identify the level of noise in original speech signal and one widely used method for SNR estimation is waveform amplitude distribution analysis (WADA), which is based on the assumption that the speech and noise signals have Gamma and Gaussian amplitude distributions respectively. Based on previous approaches, this work uses a precomputed look-up table as a reference for SNR estimation. In this study, WADA-white (WADA-W) has been developed, which rebuilds the precomputed look-up table using a white noise profile in combination of our own AVSR database. This new data corpus, namely the Loughborough University Audio-Visual (LUNA-V) dataset that contains recordings of 10 speakers with five sets of samples uttered by each speaker is used for this experimental work. We evaluate the performance of WADA-W on this database when it is corrupted by noise generated from three profiles obtained from the NOISEX-92 database included at varying SNR values. Evaluation of performance using the LUNA-V database shows that WADA-W performs better than the original WADA in terms of SNR estimation.**

***Index Terms*—Audio visual speech recognition, LUNA-V, SNR estimator, WADA.**

I. Introduction

In speech signal processing, the signal-to-noise ratio (SNR) is used to measure the power content of speech relative to the noise. The larger the SNR value the greater the wanted speech content relative to the presence of noise. SNR estimation has long been an important topic in speech recognition and it is still an active field of research [1]-[3]. This can be a challenging task, since some form of environmental noise is almost always present and there is often no prior information about the type of noise that will affect the original speech signal when it is being analyzed.

Under most conditions, it is less reliable to attempt to recognize speech using video information than audio information, but under low SNR conditions, the video stream may be able to provide useful data to support the speech recognition process. Consequently, by implementing a reliable SNR estimator, the degree to which the video information should be used to support the audio information can be determined, so providing an audio-visual speech recognition (AVSR) system that is adaptable to the prevailing presence of noise.

A number of SNR estimation methods are available. One approach is to attempt to differentiate between the frequency spectra of speech and noise; noise spectrum estimation or spectra subtraction both fall into this category and they have been used by a number of researchers [4]-[6]. Another approach is to use energy measurements and, in one widely-adopted SNR estimation method, the US National Institute of Standards and Technology (NIST) [7] used an energy histogram computed over the entire file to estimate the signal and noise energy distributions. The waveform amplitude distribution analysis (WADA) approach [8] assumes that the amplitude of speech and noise conform to Gamma and Gaussian distributions respectively A specific parameter is employed to uniquely represent the SNR by referring to a pre-computed look-up table.

The structure of this paper is as follows. Section II introduces related published work, Section III discusses the feature extraction and classification techniques used in the current work and Section IV describes the proposed WADA-W technique and the implementation of the new look-up table built from the AVSR database and added white noise. The experimental setup and results are discussed in Section V and the conclusions are presented in Section VI.

II. Related Work

The previous work involving WADA [9], [10] has demonstrated that the amplitude distribution of speech waveforms can be characterized using a Gamma distribution based on a shaping parameter value between 0.4 and 0.5. WADA makes the assumption that the audio speech is totally independent of the background noise, that clean speech always has a gamma distribution with a fixed shaping parameter and that the background noise has a Gaussian distribution. Based on a published evaluation in [11], the WADA approach has been shown to be more a more accurate estimator of SNR than is the NIST algorithm. Even in the presence of background music or interfering speech, WADA was still able to provide better results than the NIST algorithm.

SNR is estimated by referring to a pre-computed look-up

Manuscript received December 19, 2018; revised June 17, 2019. This work was supported by Universiti Malaysia Pahang and funded by the Ministry of Higher Education Malaysia under Fundamental Research Grant Scheme (FRGS) RDU160108.

Thum Wei Seong and M. Z. Ibrahim are with Faculty of Electrical and Electronics Engineering, University Malaysia Pahang, 26600 Pekan, Pahang, Malaysia (e-mail: weiseong91@hotmail.com, zamri@ump.edu.my).

D. J. Mulvaney is with School of Electronic, Electrical and Systems Engineering, Loughborough University, LE113TU, United Kingdom (e-mail: d.j.mulvaney@lboro.ac.uk).

doi: 10.18178/ijmlc.2019.9.4.824

table constructed by the WADA research group. The table has been used by later researchers [12], [13], and the WADA approach was tested using the DARPA Resource Management database and using three different types of noise: addictive white Gaussian noise; musical segments from the DARPA HUB 4 Broadcast News database and noise from a single interfering speaker.

In this paper, the researchers propose to use a modified form of the original WADA technique in which a new pre-computed look-up table is developed from a new database, the Loughborough University Audio-Visual (LUNA-V) data corpus. A new look-up table was computed using LUNA-V database audio information corrupted with white noise to provide a reference for SNR estimation, giving rise to the name WADA-White (WADA-W). The effectiveness of this approach was compared with the performance obtained from the original WADA pre-computed look-up table on a number of test speech examples corrupted by four different forms of noise.

III. Methodology

This research is an extension of previously-reported work [14] and here focuses on adapting the method used to provide adaptive weight distribution of the original WADA so that it applies to the newly-proposed WADA-W SNR estimation. The experiments were implemented using MATLAB R2015a and utilized the Hidden Markov Model Toolkit (HTK) speech processing library [15].

A. Visual Feature Extraction

During pre-processing, speaker visual information is extracted the using the Viola-Jones object recognizer [16], which performs face detection followed by mouth detection. The visual feature extraction techniques are carried out according to the approach reported in previous research work [14], which involved applying an HSV color filter [17], border following [18] and convex hull techniques [19]. The outcome is an approximate complete contour of the lips (an example is shown in Fig. 1), from which a five-dimensional visual feature vector is produced that describes height, width, ratio (height/width), area and perimeter. Details and performance of these techniques can be found in [20], [21].

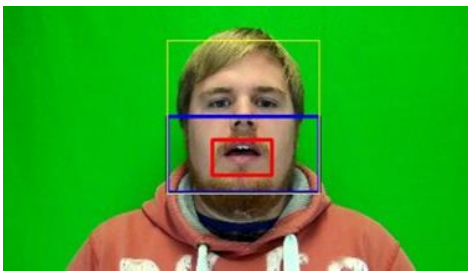


Fig. 1. Example of face and mouth detection [14].

B. Audio Feature Extraction

The Mel frequency cepstral coefficient (MFCC) and linear prediction coefficient (LPC) are among the most commonly-used feature extraction techniques [22]. MFCCs have been shown to perform better than LPC in previous work and so were used in the current research [23]. The HTK library was used for MFCC feature extraction and a 39-dimensional feature vector was generated that includes the dynamic features of MFCCs (including delta-MFCCs and delta-delta-MFCCs). Based on published research [24], the combination of dynamic features and cepstral coefficients is able to provide high performance speech recognition.

C. Classification

There have been a lot of studies regarding speech recognition that utilized hidden Markov model toolkit (HTK) [15]. HTK is one of the most widely used tools for speech recognition and it has also been extensively used in other applications such as speech synthesis, character recognition, and DNA sequencing. HTK originated from the Machine Intelligence Laboratory in Cambridge University Engineering.

The main function of HTK is to control the sets of HMM. HMM vectors can be separated into single or multiple independent data streams and each stream has its own stream weight. For audio-only speech recognition, there will be only one stream of HMM created. A two-stream HMM is required for an AVSR system that utilized both audio and visual modalities.

The multi-stream HMM (MSHMM) is a highly popular classifier in AVSR, as it can be trained automatically and is computationally easily to use. In AVSR, there will normally be two HMM streams, namely the audio and video streams. Weights are applied to provide the stream likelihoods which capture the relative reliabilities of the streams. In MSHMM, if the two streams are represented by *A* and *V* for audio and video respectively, $X_{A_t}$and X_{V_t} represent their feature vectors for a frame at time *t*, then the log-likelihood of observing both $X_{A_t}$and X_{V_t}, given a state *s*, can be calculated from the equation

$$\log[P(X_t|s)] = \lambda_a \log[P(X_{A_t}|s)] + \lambda_v \log[P(X_{V_t}|s)] \quad (1)$$

where λ_a and λ_v are the weights applied to the audio and video streams respectively.

In this work, WADA-W will be used to provide an estimated value of SNR to allocate values for the weights λ_a and λ_v. The parameter λ_a is mapped to the interval [0, 1] by using the tuneable sigmoid function below

$$\lambda_\alpha = \frac{1}{1+exp^{-\alpha(x-\beta)}} \quad (2)$$

$$\lambda_v = 2 - \lambda_a \quad (3)$$

where *x* is the SNR value, α=0.55 and β is a trade-off threshold value that needs to be obtained for individual speakers in the database.

IV. Waveform Amplitude Distribution Analysis - White (WADA-W)

The current work proposes that the performance of the WADA approach can be improved if the SNR look-up table is generated using examples from the target database, but which

have been corrupted with general white noise as the noise reference. The premise is that this modified approach, termed WADA-W, is more likely to be robust and perform better than the original WADA approach since it has been developed to support a specific application.

The WADA-W look-up table was re-generated using the LUNA-V speech data corpus with added white noise. Since the parameter G_z that is obtained from the look-up table uniquely represents a specific SNR η_x value, the formulae below can be used to re-construct the table of parameters G_z corresponding to specific SNR values [8].

$$z[n] = x[n] + v[n] \quad (4)$$

$$G_z = \ln\left(\frac{1}{N}\sum_{n=0}^{N-1}|z[n]|\right) - \frac{1}{N}\sum_{n=0}^{N-1}\ln(|z[n]|) \quad (5)$$

where $x[n]$ is clean speech, $v[n]$ is Gaussian white noise, $z[n]$ is the corrupted speech signal and N is the number of samples used for each G_z value.

For the look-up table, all the samples of the LUNA-V database were corrupted by white noise with SNR values in the range -10dB to 30dB in steps of 5dB. Each corrupted signal with a specific SNR value requires the use of Equation (5) to calculate the corresponding value of G_z. This list of values of G_z was then interpolated to form a more complete list of parameters G_z for all SNR values in the range -10dB to 30dB in steps of 1dB. The list of parameters G_z was saved in a text format, which then act as the pre-defined lookup table for the SNR estimation. Testing samples were excluded during the creation of the new look-up table to prevent overfitting. The structure of WADA-W SNR estimation is shown in Fig. 2.

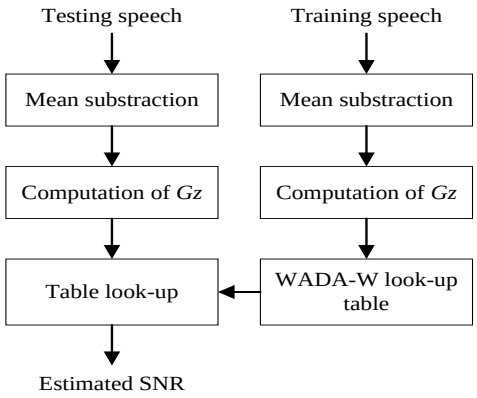


Fig. 2. The structure of WADA-W SNR estimation system.

This work is an extension of the previous work [14] and mainly focused on the adaptive stream weight distribution to audio modality and visual modality based on the original WADA and WADA-W SNR estimation techniques by using a MS-HMM as a classifier. The overall architecture AVSR system using proposed SNR estimation processes used in this work is shown in Fig. 3.

During pre-processing of the experiment, both audio and visual features were extracted and ready for the training process. To achieve feature integration, the visual and audio feature extraction rates must be equalized. In the current work, the video frame rate is 29.97Hz, whereas the audio MFCC feature rate is 100Hz. Equalization involved linear interpolation of the visual features to match the audio frame rate.

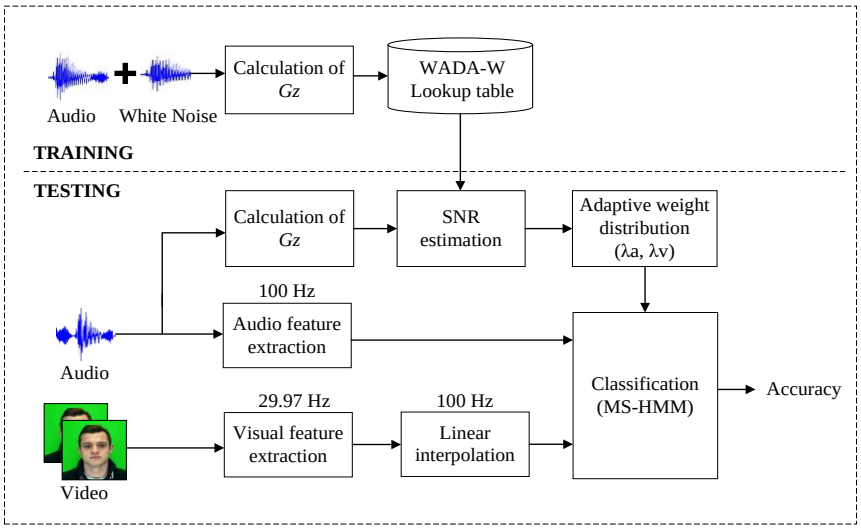


Fig. 3. Architecture of the proposed AVSR system used in this work.

V. Results

A. Experiment Setup

The experiments were conducted using the LUNA-V speech database corpus that consists of 10 subjects (nine male and one female) with five samples being provided for each subject. This experiment focused on the recognition of the spoken English digits for the words 'zero' to 'nine' only. Noise from NOISEX-92 database (white, babble, factory1 and factory2) with different SNR values (-10dB, -5dB, 0dB, 5dB, 10dB, 15dB, 20dB, 25dB and clean) were used to corrupt the speech signals. For the MSHMM classifier, seven states in a HMM were used for the combined audio and visual streams with a speech processing library implemented using the HTK tool. The work used K-1 sample files (adding noise to yield predefined SNR levels) to train the HMM classifier, where K is the total number of sample files.

All the experiments result used the 'leave-one-out' cross validation (LOOCV) method to cross check the result. Based

on previous research work [25], LOOCV techniques have been able to achieve a slightly improved accuracy compared to the more common Holdout and bootstrap validation approach. Furthermore, the LOOCV technique is able to allow an evaluation of every sample and obtain the final accuracy value by averaging the results from all samples.

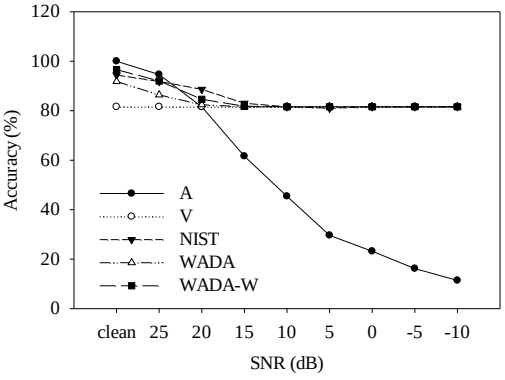
TABLE I: WADA-W LOOK-UP TABLE FOR SPEAKER V07M

Signal Noise Ratio (dB)	Parameter *Gz*
-10	0.400944
-5	0.424390
0	0.459340
5	0.536998
10	0.647825
15	0.772863
20	0.884583
25	0.962893

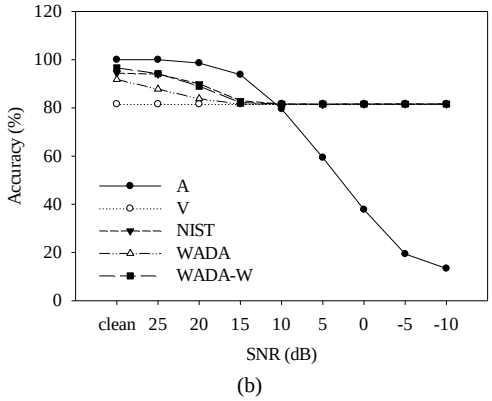
Table I shows some data of the WADA-W re-computed look-up table starting from -10dB to 25dB with the step size of 5dB. The value of parameter Gz was smaller towards low SNR condition and it increased gradually with SNR value.

B. Experimental Result

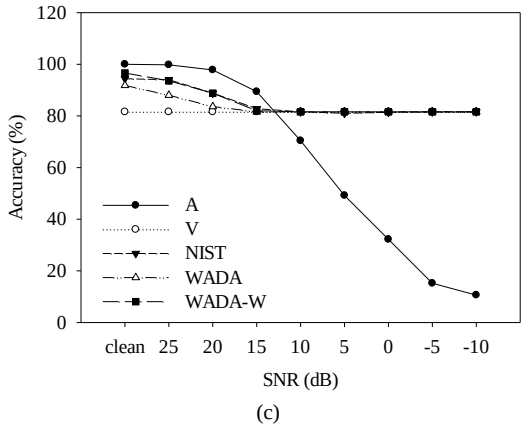
Fig. 4 and Fig. 5 show the word accuracy results obtained when implementing audio visual speech recognition using audio-only inputs (A), visual-only inputs (V) and audio-visual inputs using three methods, namely WADA, WADA-W and NIST for SNR estimation. The experiments were evaluated using a number of different types of noise each introduced with SNR values in the range -10dB to 25dB. Four types of noise were used to generate the results, namely white noise, 'babble noise' of 100 people talking in a canteen, 'factory 1 noise' recorded near to industrial plate cutting and 'factory 2 noise' in a vehicle production plant.



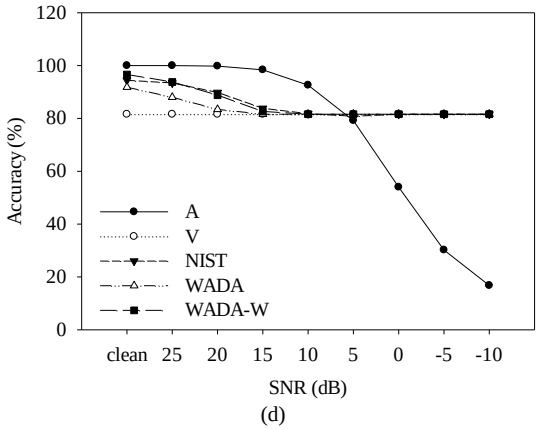
(a)



(b)



(c)



(d)

Fig. 4. Performance of AVSR system testing using four different types of noise (a) white, (b) babble, (c) factory1, (d) factory2.

From the results in Fig. 4, it can be seen that audio-only (A) speech recognition perform the best when SNR values are high, but performed very poorly when SNR values were low. Visual-only (V) speech recognition maintained an accuracy of 81.6% regardless of the noise as it is not affected by audible noise. By comparing all the performances that utilized SNR estimation techniques, the maximum accuracy of 96.6% was achieved by using WADA-W under a clean condition.

WADA-W performed slightly better compared to the original WADA technique, but NIST was found to perform similarly to WADA-W but better than the WADA technique, in contrast with the results found in previous work [8]. Further experiments were carried out using babble, factory1, and factory2 noises from the NOISEX-92 dataset to test the robustness of the AVSR system towards different types of noise. Based on the overall result, the proposed WADA-W technique outperformed the original WADA technique by about 20%. Although the NIST technique overestimated the SNR value by about 5dB, but it could still provide an approximately similar recognition result as the WADA-W technique.

Fig. 5 shows the results of an investigation into the SNR values estimated by WADA, WADA-W and the NIST algorithm and their comparison with the actual (ideal) SNR values. The WADA-W SNR estimation technique as outlined in Section IV has one main unique parameter Gz, which represents the SNR of noisy speech. To rebuild the SNR look-up table, the list of Gz was extracted from the audio speech from the LUNA-V database. Since this is a speaker-dependent speech recognition, thus each speaker will have their specific look-up table generated from their own

audio samples.

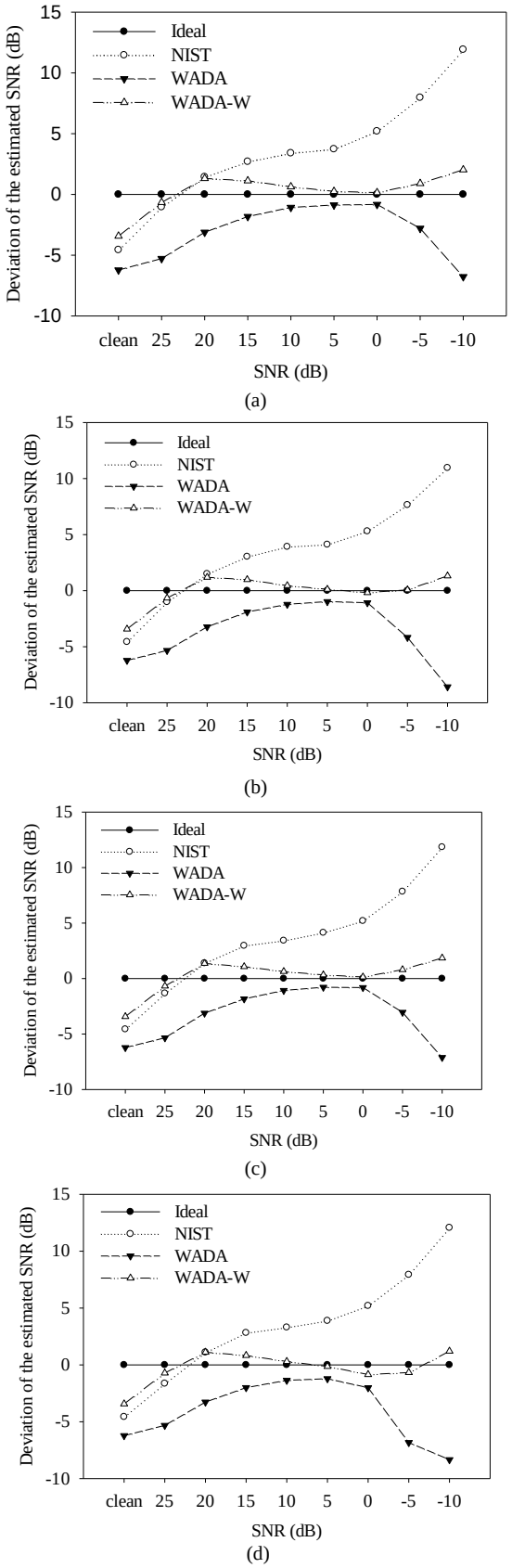


Fig. 5. Comparison of the standard deviation of NIST, WADA and WADA-W corrupted by different types of noise (a) white, (b) babble, (c) factory1, (d) factory2.

From Fig. 5, it can be seen that WADA significantly underestimated the SNR value in all the tests. The NIST technique consistently wrongly estimated noise as clean speech when the SNR was less than 0dB, but otherwise generally differed from the SNR by about 5dB. The proposed WADA-W technique consistently exhibited the best SNR accuracy for all the types of noise. The average bias of the proposed WADA-W technique was only around 2dB. But, the average bias of NIST and WADA were both greater than 5dB.

VI. Conclusion and Future Work

This paper has concentrated on the proposed WADA-W SNR estimation technique and allocated accordingly the values of the weights determining the relative contents of audio and visual features for use in speech recognition. This new technique rebuilds the WADA look-up table by calculating the parameter to uniquely represent individual SNR values when white noise is added to the samples obtained from the LUNA-V speech database. The experiments were performed using spoken English digits from the LUNA-V database, contaminated using four different types of noise (white, babble, factory1 and factory2) obtained from the NOISEX-92 database. The results show that the proposed WADA-W technique was able to achieve better and more consistent estimations of SNR values. WADA-W is able to adapt to individual speech recognition applications and it has been demonstrated that WADA-W forms a good basis for AVSR techniques that require a high accuracy in their SNR estimation.

As for future work, we are currently analyzing the performance of the proposed WADA-W SNR estimation technique with other AVSR database such as CUAVE, XM2VTS, AVletters, and VidTIMIT database. We also investigating how different weights can be applied to each state rather than stream of HMM model for both audio and visual streams by using the Coupled Hidden Markov Model (CHMM) to achieve better performance.

References

[1] J. Tchroz and B. Kollmeier, "SNR estimation based on amplitude modulation analysis with applications to noise suppression," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 3, pp. 184–192, 2003.

[2] C. Plapous, C. Marro, and P. Scalart, "Improved Signal-to-Noise Ratio Estimation for Speech Enhancement," *IEEE Trans. Audio, Speech Lang. Process*., 2010.

[3] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," IEEE Trans. Speech Audio Process., vol. 9, no. 5, pp. 504–512, 2001.

[4] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1979, vol. 4, pp. 208–211.

[5] C. Cole, M. Karam, and H. Aglan, "Noise removal in speech processing using spectral subtraction," *J. Signal Inf. Process.*, vol. 5, pp. 32–41, 2014.

[6] K. Paliwal, K. Wójcicki, and B. Schwerin, "Single-channel speech enhancement using spectral subtraction in the short-time modulation domain," *Speech Commun*., vol. 52, no. 5, pp. 450–475, 2010.

[7] The NIST Speech Signal to Noise Ratio Measurements. (2017). [Online]. Available: https://www.nist.gov/information-technology-laboratory/iad/mig/nist-speech-signal-noise-ratio-measurements

[8] R. M. S. Chanwoo Kim, "Robust signal-to-noise ratio estimation based on waveform amplitude distribution analysis," *Interspeech*, pp. 2598–2601, 2008.

[9] M. Paez and T. Glisson, "Minimum mean-squared-error quantization in speech PCM and DPCM systems," *IEEE Trans. Commun*., vol. 20, no. 2, pp. 225–230, 1972.

[10] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*, Prentice Hall, 1978.
[11] A. Papadopoulos, P Tsiartas, J. Gibson, and S. Narayanan, "A supervised signal-to-noise ratio estimation of speech signals," in *Proc. 2014 IEEE Int. Conf. Acoust. Speech Signal Process*., 2014, pp. 8287–8291.
[12] F. Liu and A. Demosthenous, "Variance of Spectral Entropy (VSE): An Snr estimator for speech enhancement in hearing aids," *Int. Congr. Sound Vib.*, pp. 1–8, 2017.
[13] A. Ziaei, A. Sangwan, and J. H. L. Hansen, "A speech system for estimating daily word counts," in *Proc. Fifteenth Annual Conference of the International Speech Communication Association*, 2014.
[14] M. Z. Ibrahim, "A novel lip geometry approach for audio-visual speech recognition," Ph.D. thesis, Loughborough University, 2014.
[15] S. Young, G. Evermann, M. Gales *et al*., "The HTK book (for HTK version 3.4)," Cambridge Univ. Eng. Dep., vol. 2, no. 2, pp. 2–3, 2006.
[16] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," *Comput. Vis. Pattern Recognit.*, vol. 1, pp. 511-518, 2001.
[17] P. Kakumanu, S. Makrogiannis, and N. Bourbakis, "A survey of skin-color modeling and detection methods," *Pattern Recognit*., vol. 40, no. 3, pp. 1106–1122, 2007.
[18] H. Li and M. Greenspan, "Model-based segmentation and recognition of dynamic gestures in continuous video streams," *Pattern Recognit*., vol. 44, no. 8, pp. 1614–1628, 2011.
[19] J. Sklansky, "Finding the Convex Hull of a Simple Polygon," *Pattern Recogn. Lett*., vol. 1, no. 2, pp. 79–83, Dec. 1982.
[20] M. Z. Ibrahim and D. J. Mulvaney, "A lip geometry approach for feature-fusion based audio-visual speech recognition," in *Proc. IEEE 6th International Symposium on Communications, Control and Signal Processing (ISCCSP)*, 2014, pp. 644–647.
[21] M. Z. Ibrahim and D. J. Mulvaney, "Geometry based lip reading system using Multi Dimension Dynamic Time Warping," in *Proc. IEEE Visual Communications and Image Processing (VCIP 2012)*, 2012, pp. 1–6.
[22] K. Chauhan and S. Sharma, "A review on feature extraction techniques for CBIR system," *Signal Image Process. An Int. J.*, vol. 3, no. 6, pp. 1–14, 2012.
[23] S. Tripathy, N. Baranwal, and G. C. Nandi, "A MFCC based Hindi speech recognition technique using HTK Toolkit," in *Proc. 2013 IEEE 2nd Int. Conf. Image Inf. Process*, 2013, pp. 539–544.
[24] N. S. A. Wahid, P. Saad, and M. Hariharan, "Automatic infant cry pattern classification for a multiclass problem," *Journal of Telecommunication, Electronic and Computer Engineering (JTEC)*, vol. 8, no. 9, pp. 45–52, 2016.
[25] T. W. Seong, M. Z. Ibrahim, N. W. B. Arshad, and D. J. Mulvaney, "A comparison of model validation techniques for audio-visual speech recognition," *IT Convergence and Security 2017*, vol. 1, pp. 112–119, 2018.

Thum Wei Seong received the B.E degree in electrical and electronic engineering from the University Malaysia Pahang, Campus Pekan, Pahang, Malaysia, in 2016.

He is currently working towards the M.S degree from the same university which mainly focuses audio-visual speech recognition. His published research areas are in work on modal validation techniques, signal processing, and noise estimation. His research interests are in image analysis, computer vision, signal-to-noise estimation and multimodal fusion for audiovisual speech recognition.

M. Z. Ibrahim obtained both B.Eng and M.Eng in electrical engineering from Universiti Teknologi Malaysia, Malaysia and a PhD in electrical and electronics engineering from Loughborough University, United Kingdom. His research interests are in the area of computer vision, internet of thing, embedded system programming, brain computer interaction, image processing, intelligent system and speech recognition.

He is a senior lecturer at the Faculty of Electrical and Electronics Engineering of University Malaysia Pahang, Malaysia. He had received 7 awards at national and international research exhibition and currently a principle investigator on TERAJU RM500K project on the Portable Vein Finder imaging device.

Before joining University Malaysia Pahang, He was a procurement engineer at Hewlett-Packard (HP) Malaysia where he worked as main technical interface with suppliers and HP design center (Vancouver and San Diego) to drive cost reduction, quality improvement and assurance of supply.

D. J. Mulvaney obtained a BSc in electrical and electronic engineering and a PhD in mechanical engineering, both from the University of Leeds, UK. His main research interests are in hardware-based AI solutions and in the development of design and simulation tools for distributed real-time applications.

He is a senior lecturer at Loughborough University and is currently an investigator on the EPSRC/JLR £1.5M project on the Analysis of the Vehicle as a Complex System. He is a founder of Axilica, a spin-out company specializing in novel embedded system tool flows.

Before joining Loughborough, Dr Mulvaney worked as a senior software engineer at Cambridge Consultants where he developed a number of real-time AI systems for military applications. Dr Mulvaney has also delivered commercial training in embedded real-time systems and has undertaken consultancy work for BP, Otis, Cadbury-Schweppes and GE Lighting.