DEVELOPMENT ON SNR ESTIMATOR FOR AUDIO-VISUAL SPEECH RECOGNITION BASED ON WAVEFORM AMPLITUDE DISTRIBUTION ANALYSIS

THUM WEI SEONG

Master of Science

UNIVERSITI MALAYSIA PAHANG

SUPERVISOR'S DECLARATION

We hereby declare that we have checked this thesis and in our opinion, this thesis is adequate in terms of scope and quality for the award of the degree of Master of Science.

____

(Supervisor's Signature)

Full Name : DR. MOHD ZAMRI BIN IBRAHIM

Position : SENIOR LECTURER

Date : 10 DECEMBER 2018

____

(Co-supervisor's Signature)

Full Name : NURUL WAHIDAH BINTI ARSHAD

Position : LECTURER

Date : 10 December 2018

STUDENT'S DECLARATION

I hereby declare that the work in this thesis is based on my original work except for quotations and citations which have been duly acknowledged. I also declare that it has not been previously or concurrently submitted for any other degree at Universiti Malaysia Pahang or any other institutions.

____

(Student's Signature)

Full Name : THUM WEI SEONG

ID Number : MEL16013

Date : 10 December 2018

DEVELOPMENT ON SNR ESTIMATOR FOR AUDIO-VISUAL SPEECH RECOGNITION BASED ON WAVEFORM AMPLITUDE DISTRIBUTION ANALYSIS

THUM WEI SEONG

Thesis submitted in fulfillment of the requirements
for the award of the degree of
Master of Science

Faculty of Electrical & Electronics Engineering

UNIVERSITI MALAYSIA PAHANG

DECEMBER 2018

ACKNOWLEDGEMENTS

First of all, I am very grateful and thankful to the God who give me the strength and wisdom to discover and complete my graduate program, under His guidance and blessing, I am able to complete my graduate study smoothly.

I would like to thank my excellent supervisor, Dr. Zamri bin Ibrahim, of the Faculty of Electrical and Electronic Engineering at University Malaysia Pahang. The door to Dr. Zamri's office was always open whenever I ran into a trouble spot or faced any questions about my work or thesis writing. His invaluable help of constructive comments and suggestion throughout the experimental and thesis works have contributed to the success of this work. I am also very grateful to Mrs Nurul Wahidah Binti Arshad as the co-supevisor for many valuable lesson and encouragements.

Besides that, I would like to thank Universiti Malaysia Pahang for granting me research funding to complete my research. I would also like to express my sincere gratitude towards to the staff and lecturers of both Institutes of Postgraduate Studies and Faculty of Electrical and Electronic Engineering for providing help directly or indirectly to complete my studies. My special gratitude also goes to the Institute of Postgraduate Studies and research and Innovation Department for the financial support, the Master Research Scheme (MRS) and funded by the Ministry of Higher Education Malaysia under Fundamental Research Grant Scheme (FRGS) with Grant Nos. RDU160108.

My deepest gratitude goes to my beloved parents, Mr. Thum Seik Mun and Mrs. Chai Choon Fah for providing me with their endless love, emotional support and continuous encouragement throughout my years of study. I am also grateful to my other family members who have supported me along the way.

Last but not least, I would like to thank my friends for their kindness and moral support during my study. Thanks for the friendship and memories. To those who are involved indirectly in this work, your kindness means a lot to me. Thank you very much.

ABSTRAK

Prestasi sistem pengecaman pertuturan boleh diperbaiki untuk pengecaman pertuturan audio-visual (*audio-visual speech recognition*, AVSR) yang menggunakan modaliti audio digabungkan dengan modaliti visual, terutamanya apabila beroperasi dalam persekitaran yang hingar. Modaliti audio amat mudah terganggu oleh hingar ambien, dan ini menyebabkan kesukaran dalam membezakan isyarat pertuturan sebenar dengan isyarat hingar dengan betul. Nisbah isyarat-hingar (*signal-to-noise ratio*, SNR) ialah nisbah asas pengukur isyarat kuasa isyarat terhadap kuasa hingar dalam unit desibel (dB). Salah satu daripada teknik anggaran SNR yang terkenal ialah analisis agihan amplitud bentuk gelombang (*waveform amplitude distribution analysis*, WADA) yang mengandaikan bahawa amplitud pertuturan dan hingar mengikut taburan gamma dan Gaussian. Teknik ini telah digunakan dalam kerja-kerja penyelidikan sebagai tanda aras untuk perbandingan keputusan. Walau bagaimanapun, tiada arahan yang jelas mengenai cara untuk membina jadual carian. Dalam kajian ini, pembangunan dan pembinaan semula jadual carian menggunakan pangkalan data sendiri yang terganggu dengan hingar putih umum sebagai rujukan hingar dicadangkan. Pembinaan semula teknik jadual carian WADA, yang dikenali sebagai analisis agihan amplitud bentuk gelombang-putih (WADA-W), mampu untuk mempertingkatkan anggaran SNR dengan merujuk kepada jadual carian WADA-W terbina semula, dan bukan jadual carian prahitung WADA umum. Teknik anggaran WADA-W SNR yang dicadang dinilai dengan membangunkan satu sistem AVSR yang menggunakan ciri-ciri pekali cepstral mel-frekuensi (*mel-frequency cepstral coefficients*, MFCC) dan ciri-ciri visual berasaskan bentuk daripada dua pangkalan data: LUNA-V dan CUAVE. Menurut keputusan eksperimen, dengan merujuk kepada jadual carian WADA-W, anggaran SNR yang tekal boleh dilaksanakan dengan lebih tepat dan kurang berat sebelah berbanding dengan teknik WADA asal di bawah empat jenis hingar iaitu hingar putih, hingar bebel, hingar factory1, dan hingar factory2 daripada set data NOISEX-92. Sisihan keseluruhan anggaran SNR bagi pangkalan data LUNA-V menggunakan teknik WADA-W yang dicadangkan adalah hanya kira-kira 9.6dB, manakala sisihan teknik NIST dan WADA adalah kira-kira 42.3dB dan 67.3dB masing-masing. Dengan menggunakan teknik cadangan yang sama untuk pangkalan data CUAVE, sisihan keseluruhan anggaran SNR itu hanya 13.3dB, manakala sisihan teknik NIST dan WADA masing-masing adalah 50.6dB dan 62.3dB. Pengklasifikasian telah dilakukan dengan menggunakan model tersembunyi Markov pelbagai aliran (*Multi-stream Hidden Markov Model*, MSHMM) dengan teknik kebenaran-satu keluar merentas pengesahan (*leave-one-out cross validation*, LOOCV). Berdasarkan ekperimen, ia menunjukkan bahawa system AVSR yang dicadangkan dapat mencapai ketepatan tertinggi pada 96.6% menggunakan pangkalan data LUNA-V dan 95.2% untuk pangkalan data CUAVE di dalam keadaan ketiadaan hingar. Kesimpulannya, teknik anggaran WADA-W SNR yang dicadangkan dapat meningkatkan keupayaan sebanyak 4.5% dan 12.7% berbanding dengan teknik WADA asal dengan menggunakan pangkalan data LUNA-V dan CUAVE masing-masing.

ABSTRACT

For audio-visual speech recognition (AVSR) that uses audio modality combined with visual modality, the performance of speech recognition system can be improved, particularly when operating in a noisy environment. Audio modality can be easily corrupted by ambient noise, and this causes difficulty in distinguishing the actual speech signal with noise signal correctly. Signal-to-noise ratio (SNR) is a fundamental measuring ratio of signal power over noise power, which is expressed in decibels (dB). One of the most famous SNR estimation techniques is the waveform amplitude distribution analysis (WADA), where it assumes that the amplitude of speech and noise follows gamma and Gaussian distributions. It has been used in some research works as a benchmark for result comparison. However, there is no clear instruction on how to build the look-up table. In this work, the development and rebuild of the look-up table using the own database corrupted with general white noise as the noise reference has been proposed. The reconstruction of WADA look-up table technique, which is known as the waveform amplitude distribution analysis-white (WADA-W), is able to enhance the SNR estimation by referring to the reconstructed WADA-W look-up table instead of a general WADA precomputed look-up table. The proposed WADA-W SNR estimation technique was evaluated by developing an AVSR system that utilised mel-frequency cepstral coefficients (MFCC) features and shape-based visual features from two speech databases: LUNA-V and CUAVE. According to the experimental result, it showed that by referring to the WADA-W look-up table, it is capable of performing a consistent SNR estimation with more accurate and less bias result compared to the original WADA technique under four types of noises, which are white, babble, factory1, and factory2 noises from the NOISEX-92 dataset. The overall deviation of the SNR estimation of the LUNA-V database using the proposed WADA-W technique was just approximately 9.6dB, whereas the deviation of NIST and WADA techniques was approximately 42.3dB and 67.3dB respectively. By using the same proposed technique for CUAVE database, the overall deviation of the SNR estimation was only 13.3dB, whereas the deviation of NIST and WADA techniques was 50.6dB and 62.3dB respectively. The classification was done using the multi-stream hidden Markov model (MSHMM) with leave-one-out cross-validation (LOOCV) technique. From the experiments, it showed that the proposed AVSR system able to achieve the highest accuracy at 96.6% using LUNA-V database and 95.2% for CUAVE database under clean condition. In conclusion, the proposed WADA-W SNR estimator able to improve by 4.5% and 12.7% compared to the original WADA technique by using the LUNA-V and CUAVE database respectively.

TABLE OF CONTENT

DECLARATION

TITLE PAGE

ACKNOWLEDGEMENTS ii

ABSTRAK iii

ABSTRACT iv

TABLE OF CONTENT v

LIST OF TABLES viii

LIST OF FIGURES ix

LIST OF SYMBOLS xii

LIST OF ABBREVIATIONS xiii

CHAPTER 1 INTRODUCTION 1

1.1 Background of Work 1

1.2 Problem Statements 3

1.3 Objectives 5

1.4 Scopes 5

1.5 Contribution of the Work 6

1.6 Thesis Outline 7

CHAPTER 2 LITERATURE REVIEW 8

2.1 Introduction 8

2.2 AVSR Architecture 8

2.3 Multi-Stream Hidden Markov Model using Hidden Markov Model Toolkit 12

2.4 SNR Estimation 15

2.5 Audio-Visual Data Corpus 18

2.6 Sampling Methods 21

2.6.1 Random Subsampling 22

2.6.2 K-Fold Cross-Validation 23

2.6.3 Leave-One-Out Cross-Validation 24

2.7 Critical Review 25

2.8 Summary 29

CHAPTER 3 METHODOLOGY 30

3.1 Introduction 30

3.2 Software, Database and Experimental Setup 30

3.3 Overview of Architecture 31

3.4 SNR Estimation Techniques 34

3.4.1 SNR Estimation using NIST 34

3.4.2 SNR Estimation using WADA 35

3.4.3 SNR Estimation using WADA-W 37

3.5 Adaptive Multi-Stream HMM 38

3.6 Summary 41

CHAPTER 4 RESULTS AND DISCUSSION 42

4.1 Introduction 42

4.2 Results using LUNA-V Database 42

4.2.1 WADA-W SNR Estimation 43

4.2.2 Performance under Noise Conditions 45

4.2.3 Audio-Visual Weight Distribution 47

4.3 Results using CUAVE Database 50

4.3.1 WADA-W SNR Estimation 50

4.3.2 Performance under Noise Conditions 52

4.3.3 Audio-Visual Weight Distribution 54

4.4 Summary 57

CHAPTER 5 CONCLUSION 59

5.1 Introduction 59

5.2 Summary of the Work 59

5.3 Future Work 61

REFERENCES 62

APPENDIX A Speech recognition of some LUNA-V database speakers using NIST, WADA, and WADA-W techniques simulated with babble noise 71

APPENDIX B Speech recognition of some CUAVE database speakers using NIST, WADA, and WADA-W techniques simulated with factory1 noise 76

LIST OF TABLES

Table 2.1	Performance of recent works using single-stream HMM and MS-HMM	15
Table 2.2	The sentences collected in LUNA-V database	19
Table 2.3	Summary of audio-visual speech databases	21
Table 2.4	Summary of recent works on speech recognition	26
Table 2.5	Comparison of different sampling methods	28
Table 3.1	Precomputed look-up table of WADA	37
Table 3.2	WADA-W look-up table for speaker v07m	38

LIST OF FIGURES

Figure 2.1	Block diagram of an AVSR system	10
Figure 2.2	Hidden Markov model toolkit processing stages	12
Figure 2.3	Definition of a simple single-stream HMM	13
Figure 2.4	A definition of two-stream HMM	14
Figure 2.5	HXR-MC2000E video camera	20
Figure 2.6	Sony ECM-PS1 stereo microphone	20
Figure 2.7	Illustration diagram of random subsampling	22
Figure 2.8	Illustration diagram of four-fold cross-validation	23
Figure 2.9	Illustration diagram of leave-one-out cross-validation	24
Figure 3.1	Architecture of the proposed AVSR in this work	31
Figure 3.2	Face and mouth detection of the speaker	32
Figure 3.3	Five visual features of the lips	33
Figure 3.4	RMS power histogram of NIST SNR estimation	35
Figure 3.5	The structure of WADA SNR estimation system	36
Figure 3.6	The structure of WADA-W SNR estimation system	38
Figure 3.7	Integration of audio and visual streams using 7-states MS-HMM model	39
Figure 3.8	Multi-stream HMM structure using HTK tools	40
Figure 4.1	Bias of SNR estimation performance using NIST, WADA, and WADA-W techniques for application of white noise	44
Figure 4.2	Bias of SNR estimation performance using NIST, WADA, and WADA-W techniques for application of babble noise	45
Figure 4.3	Bias of SNR estimation performance using NIST, WADA, and WADA-W techniques for application of factory1 noise	45
Figure 4.4	Bias of SNR estimation performance using NIST, WADA, and WADA-W techniques for application of factory2 noise	45
Figure 4.5	AVSR system performance using NIST, WADA, and WADA-W techniques for application of white noise	46
Figure 4.6	AVSR system performance using NIST, WADA, and WADA-W techniques for application of babble noise	47
Figure 4.7	AVSR system performance using NIST, WADA, and WADA-W techniques for application of factory1 noise	47
Figure 4.8	AVSR system performance using NIST, WADA, and WADA-W techniques for application of factory2 noise	47

Figure 4.9	Bias of SNR estimation performance of a LUNA-V speaker, v10m, using NIST, WADA, and WADA-W techniques for application of babble noise	48
Figure 4.10	Stream weight distribution of a LUNA-V speaker, v10m, using NIST as the SNR estimation technique	49
Figure 4.11	Stream weight distribution of a LUNA-V speaker, v10m, using WADA as the SNR estimation technique	49
Figure 4.12	Stream weight distribution of a LUNA-V speaker, v10m, using WADA-W as the SNR estimation technique	49
Figure 4.13	AVSR system performance of a LUNA-V speaker, v10m, using NIST, WADA, and WADA-W techniques for application of babble noise	50
Figure 4.14	Bias of overall SNR estimation performance on the CUAVE database using NIST, WADA, and WADA-W techniques for application of white noise	51
Figure 4.15	Bias of overall SNR estimation performance on the CUAVE database using NIST, WADA, and WADA-W techniques for application of babble noise	52
Figure 4.16	Bias of overall SNR estimation performance on the CUAVE database using NIST, WADA, and WADA-W techniques for application of factory1 noise	52
Figure 4.17	Bias of overall SNR estimation performance on the CUAVE database using NIST, WADA, and WADA-W techniques for application of factory1 noise	52
Figure 4.18	Overall AVSR system performance on the CUAVE database using NIST, WADA, and WADA-W techniques for application of white noise	53
Figure 4.19	Overall AVSR system performance on the CUAVE database using NIST, WADA, and WADA-W techniques for application of babble noise	54
Figure 4.20	Overall AVSR system performance on the CUAVE database using NIST, WADA, and WADA-W techniques for application of factory1 noise	54
Figure 4.21	Overall AVSR system performance on the CUAVE database using NIST, WADA, and WADA-W techniques for application of factory2 noise	54
Figure 4.22	Bias of SNR estimation performance of a CUAVE speaker, s28f, using NIST, WADA, and WADA-W techniques for application of factory1 noise	55
Figure 4.23	Stream weight distribution of a CUAVE speaker, s28f, using NIST as the SNR estimation technique	56
Figure 4.24	Stream weight distribution of a CUAVE speaker, s28f, using WADA as the SNR estimation technique	56

Figure 4.25	Stream weight distribution of a CUAVE speaker, s28f, using WADA-W as the SNR estimation technique	56
Figure 4.26	AVSR system performance of a CUAVE speaker, s28f, using NIST, WADA, and WADA-W techniques for application of factory1noise	57

LIST OF SYMBOLS

A	Audio modality
α	Shaping parameter
β	Trade-off threshold values
G_z	Unique parameter to represent SNR
λ_a	Weight of audio modality
λ_v	Weight of visual modality
N	Total number of observations
$o_{av,t}$	Audio-visual observation
S	Modality
S_t	HMM state at time *t*
t	Time
V	Visual modality
$v[n]$	Noise
$x[n]$	Clean speech
$z[n]$	Noisy speech

LIST OF ABBREVIATIONS

ASR	Automatic speech recognition
AVSR	Audio-visual speech recognition
DWT	Dynamic time wrapping
GSNR	Global signal-to-noise ratio
HMM	Hidden markov model
HOG	Histogram of oriented gradients
HSV	Hue saturation value
HTK	Hidden markov model toolkit
ICA	Independent component analysis
LDA	Linear discriminant analysis
LOOCV	Leave-one-out cross validation
LPC	Linear predictive cepstral coefficient
LTSV	Long-term signal variability
MAE	Mean absolute error
MFCC	Mel-frequency cepstrum coefficient
MS-HMM	Multi-stream HMM
NIST	National institute of standards and technology
PCA	Principal component analysis
RMS	Root mean squared
ROI	Region-of-interest
SNR	Signal-to-noise ratio
STS-SNR	Short-time silence SNR
WADA	Waveform amplitude distribution analysis
WADA-W	Waveform amplitude distribution analysis – white

REFERENCES

A, I., & E, C. (2015). Noise Estimation Using Standard Deviation of the Frequency Magnitude Spectrum for Mixed Non-Stationary Noise. *ICTACT Journal on Communication Technology*, *6*(4), 1218–1222. http://doi.org/10.21917/ijct.2015.0178

Abdelaziz, A. H., Steffen Zeiler, & Kolossa, D. (2013). Twin-Hmm-Based Audio-Visual Speech Enhancement. *Icassp*, 3726–3730.

Abdelaziz, A. H., Zeiler, S., & Kolossa, D. (2015). Learning dynamic stream weights for coupled-HMM-based audio-visual speech recognition. *IEEE/ACM Transactions on Audio Speech and Language Processing*, *23*(5), 863–876. http://doi.org/10.1109/TASLP.2015.2409785

Abdi, H., & Williams, L. J. (2010). Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, *2*(4), 433–459.

Abdulaziz, A. S., & Këpuska, V. Z. (2016). The Short-Time Silence of Speech Signal as Signal-To-Noise Ratio Estimator, *6*(8), 99–103.

Ahmed Hussen Abdelaziz, Steffen Zeiler*, D. K. (2013). Twin-HMM-based audio-visual speech enhancement. *Digital Signal Processing*, 3726–3730.

Alam, M. J., Kenny, P., & O'Shaughnessy, D. (2012). Robust Feature Extraction for Speech Recognition by Enhancing Auditory Spectrum. *Interspeech*, 3–6. Retrieved from http://20.210-193-52.unknown.qala.com.sg/archive/archive_papers/interspeech_2012/i12_1360.pdf

Bengio, S. (2004). Multimodal speech processing using asynchronous Hidden Markov Models. *Information Fusion*, *5*(2), 81–89. http://doi.org/10.1016/j.inffus.2003.04.001

Berouti, M., Schwartz, R., & Makhoul, J. (1979). Enhancement of speech corrupted by acoustic noise. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'79.* (Vol. 4, pp. 208–211).

Bursztein, E., Beauxis, R., Paskov, H., Perito, D., Fabry, C., & Mitchell, J. (2011). The failure of noise-based non-continuous audio captchas. *Proceedings - IEEE Symposium on Security and Privacy*, 19–31. http://doi.org/10.1109/SP.2011.14

Chanwoo Kim, R. M. S. (2008). Robust signal-to-noise ratio estimation based on waveform amplitude dstribution analysis. *Interspeech*, 2598–2601.

Chen, T. (2001). Audiovisual speech processing. *IEEE Signal Processing Magazine*, *18*(1), 9–21.

Chitu, A. G., & Rothkrantz, L. J. M. (2007). Building a Data Corpus for Audio-Visual Speech Recognition, *1*(Movellan 1995). Retrieved from http://mmi.tudelft.nl/pub/alin/APTEC-04.pdf

Cole, C., Karam, M., & Aglan, H. (2014). Noise Removal in Speech Processing Using Spectral Subtraction. *Journal of Signal and Information Processing*, *5*, 32–41. http://doi.org/10.1109/ITNG.2008.86

DARPA-ISTO. (1990). The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus (TIMIT). *Speech Disc cd1- 1.1 Edition*.

Dave, N. (2013). Feature Extraction Methods LPC , PLP and MFCC In Speech Recognition. *International Journal for Advance Research in Engineering and Technology*, *1*(Vi), 1–5.

Efron, B., & Gong, G. (1983). A Leisurely Look at the Bootstrap , the Jackknife , and. *American Statistician*. http://doi.org/10.1080/00031305.1983.10483087

Feng, L., & Hansen, L. K. (2005). *A new database for speaker recognition*. IMM, Informatik og Matematisk Modelling, DTU.

Gastpar, M., & Vetterli, M. (2003). Information Processing in Sensor Networks. *Information Processing in Sensor Networks*, *2634*(January 2003), 553–553. http://doi.org/10.1007/3-540-36978-3

Ghadage, Y. H., & Shelke, S. D. (2016). Speech to Text Conversion for Multilingual Languages, 236–240. http://doi.org/10.1109/ICCSP.2016.7754130

Gurban, M., Thiran, J.-P., Drugman, T., & Dutoit, T. (2008). Dynamic Modality Weighting for Multi-stream HMMs in Audio-Visual Speech Recognition. *10th International Conference on Multimodal Interfaces*, 237–240. http://doi.org/10.1145/1452392.1452442

Guri, M., Solewicz, Y., Daidakulov, A., & Elovici, Y. (2016). SPEAKE(a)R: Turn Speakers to Microphones for Fun and Profit. Retrieved from http://arxiv.org/abs/1611.07350

H. B. Chauhan, P. B. A. T. (2015). Comparative Study of MFCC And LPC Algorithms for Gujrati Isolated Word Recognition. *IJIRCCE*, *3*(2), 822–826. http://doi.org/10.15680/ijircce.2015.0302056

Hongbing Hu, S. A. . Z. (2010). Dimensionality reduction methods for HMM phonetic recognition. *Training*, 4854–4857.

Hu, Y., & Loizou, P. C. (2007). Subjective comparison and evaluation of speech enhancement algorithms. *Speech Communication*, *49*(7–8), 588–601. http://doi.org/10.1016/j.specom.2006.12.006

Hydari, M., & Karami, M. R. (2009). Speech Signals Enhancement Using LPC Analysis based on Inverse Fourier Methods, *2*(1), 1–15.

Ibrahim, M. Z. (2014). *A novel lip geometry approach for audio-visual speech recognition*. Loughborough University.

Ibrahim, M. Z., & Mulvaney, D. J. (2013). Robust geometrical-based lip-reading using hidden Markov models. *IEEE EuroCon 2013*, (July), 2011–2016. http://doi.org/10.1109/EUROCON.2013.6625256

Ibrahim, M. Z., & Mulvaney, D. J. (2014). A lip geometry approach for feature-fusion based audio-visual speech recognition. *ISCCSP 2014 - 2014 6th International Symposium on Communications, Control and Signal Processing, Proceedings*, 644–647. http://doi.org/10.1109/ISCCSP.2014.6877957

Ibrahim, M. Z., Mulvaney, D. J., & Abas, M. F. (2015a). Feature-fusion based audio-visual speech recognition using lip geometry features in noisy enviroment. *ARPN Journal of Engineering and Applied Sciences*, *10*(23), 17521–17527.

Ibrahim, M. Z., Mulvaney, D. J., & Abas, M. F. (2015b). Feature-fusion based audio-visual speech recognition using lip geometry features in noisy enviroment, *10*(23), 17521–17527.

Ittichaichareon, C. (2012). Speech recognition using MFCC. *... Conference on Computer ...*, (September), 135–138. http://doi.org/10.13140/RG.2.1.2598.3208

Kakumanu, P., Makrogiannis, S., & Bourbakis, N. (2007). A survey of skin-color modeling and detection methods. *Pattern Recognition*, *40*(3), 1106–1122. http://doi.org/10.1016/j.patcog.2006.06.010

Kambiz Rahbar. (2010). Independent-Speaker Isolated Word Speech Recognition Based on Mean-Shift Framing Using Hybrid HMM/SVM Classifier, 156–161.

Kocaguneli, E., & Menzies, T. (2013). Software effort models should be assessed via leave-one-out validation. *Journal of Systems and Software*, *86*(7), 1879–1890. http://doi.org/10.1016/j.jss.2013.02.053

Kulkarni, D. S., Deshmukh, R. R., Shrishrimal, P. P., Waghmare, S. D., & Science, C. (2016). HTK Based Speech Recognition Systems for Indian Regional languages : A Review.

Li, H., & Greenspan, M. (2011). Model-based segmentation and recognition of dynamic gestures in continuous video streams. *Pattern Recognition*, *44*(8), 1614–1628. http://doi.org/10.1016/j.patcog.2010.12.014

Liu, F., & Demosthenous, A. (2017). Variance of Spectral Entropy (Vse): an Snr Estimator for Speech Enhancement in Hearing Aids. *International Congress on Sound and Vibration*, 1–8.

Lucey, S., Chen, T., Sridharan, S., & Chandran, V. (2005). Integration strategies for audio-visual speech processing: applied to text-dependent speaker recognition. *IEEE Transactions on Multimedia*, *7*(3), 495–506. http://doi.org/10.1109/TMM.2005.846777

Luettin, J., Thacker, N. a., & Beet, S. W. (1996). Visual speech recognition using active shape models and hidden Markov models. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, *2*(95), 817–820. http://doi.org/10.1109/ICASSP.1996.543246

Maio, D., Maltoni, D., Cappelli, R., Wayman, J. L., & Jain, A. K. (2002). FVC2002: Second fingerprint verification competition. In *Pattern recognition, 2002. Proceedings. 16th international conference on* (Vol. 3, pp. 811–814).

Marcheret, E., Chu, S. M., Goel, V., Potamianos, G., Watson, I. B. M. T. J., & Heights, Y. (2004). Efficient Likelihood Computation in Multi-Stream HMM based Audio-Visual Speech Recognition. *Word Journal Of The International Linguistic Association*, (1).

Matthews, I. (1998). *Features for audio-visual speech recognition*. Citeseer.

Matthews, I., Potamianos, G., Neti, C., & Luettin, J. (2001). A Comparison Of Model And Transform-Based Visual Features For Audio-Visual LVCSR. In *ICME*.

Mohamed, A., Sainath, T. N., Dahl, G., Ramabhadran, B., Hinton, G. E., Picheny, M. a, … Heights, Y. (2011). Deep belief networks using discriminative features for phone recognition. *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, 5060–5063.

Motlicek, P. (2002). *Feature extraction in speech coding and recognition*.

Movellan, J. R. (1995). Visual speech recognition with stochastic networks. In *Advances in neural information processing systems* (pp. 851–858).

Mporas, I., & Ganchev, T. (2007). Comparison of speech features on the speech recognition task. *Journal of Computer Science*, *3*(8), 608–616. http://doi.org/10.3844/jcssp.2007.608.616

Navarathna, R., Dean, D. B., Lucey, P. J., Sridharan, S., & Fookes, C. B. (2010). Recognising audio-visual speech in vehicles using the AVICAR database. *Proceedings of the 13th Australasian International Conference on Speech Science and Technology*, (December), 110–113.

Paez, M., & Glisson, T. (1972). Minimum mean-squared-error quantization in speech PCM and DPCM systems. *IEEE Transactions on Communications*, *20*(2), 225–230.

Paleček, K. (2016). Lipreading using spatiotemporal histogram of oriented gradients. *European Signal Processing Conference*, *2016–Novem*, 1882–1885. http://doi.org/10.1109/EUSIPCO.2016.7760575

Paliwal, K., Wójcicki, K., & Schwerin, B. (2010). Single-channel speech enhancement using spectral subtraction in the short-time modulation domain. *Speech Communication*, *52*(5), 450–475.

Pan, Y., Shen, P., & Shen, L. (2012). Speech emotion recognition using support vector machine. *International Journal of Smart Home*, *6*(2), 101–108. http://doi.org/10.5120/431-636

Papadopoulos, P., Travadi, R., & Narayanan, S. (2017). Global SNR estimation of speech signals for unknown noise conditions using noise adapted non-linear regression. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, *2017–Augus*, 3842–3846. http://doi.org/10.21437/Interspeech.2017-230

Papadopoulos P Tsiartas, A., Gibson, J., & Narayanan, S. (2014). A supervised signal-to-noise ratio estimation of speech signals. *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 8287–8291. http://doi.org/10.1109/ICASSP.2014.6855207

Patterson, E. K., Gurbuz, S., Tufekci, Z., & Gowdy, J. N. (2002). CUAVE: A new audio-visual database for multimodal human-computer interface research. *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, *2*, II-2017-II-2020. http://doi.org/10.1109/ICASSP.2002.5745028

Pawar, G. S. (2014). Realization of Hidden Markov Model for English Digit Recognition, *98*(17), 98–101.

Pawar, G. S., & Morade, S. S. (2014). Isolated English Language Digit Recognition Using Hidden Markov Model Toolkit, *4*(6), 781–784.

Pei, Y., Kim, T.-K., & Zha, H. (2013). Unsupervised random forest manifold alignment for lipreading. In *Computer Vision (ICCV), 2013 IEEE International Conference on* (pp. 129–136).

Potamianos, G., Neti, C., Luettin, J., & Matthews, I. (2004). Audio-Visual Automatic Speech Recognition : An Overview. *Issues in Visual and AudioVisual Speech Processing*, (January), 1–30.

Price, P., Fisher, W. M., Bernstein, J., & Pallett, D. S. (1988). The DARPA 1000-word resource management database for continuous speech recognition. In *Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on* (pp. 651–654).

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, *77*(2), 257–286.

Rabiner, L. R., & Juang, B.-H. (1993). *Fundamentals of speech recognition* (Vol. 14). PTR Prentice Hall Englewood Cliffs.

Rabiner, L. R., & Schafer, R. W. (1978). *Digital processing of speech signals*. Prentice Hall.

Rajavel, R., & Sathidevi, P. S. (2012). Adaptive reliability measure and optimum integration weight for decision fusion audio-visual speech recognition. *Journal of Signal Processing Systems*, *68*(1), 83–93. http://doi.org/10.1007/s11265-011-0578-x

Receveur, S., Scheler, D., & Fingscheidt, T. (2014). A turbo-decoding weighted forward-backward algorithm for multimodal speech recognition, 179–192. http://doi.org/10.1007/978-3-319-21834-2

RecordPad Sound Recording Software - NCH Software. (n.d.). Retrieved August 27, 2018, from http://www.nch.com.au/recordpad

Rekik, A., Ben-Hamadou, A., & Mahdi, W. (2014). A new visual speech recognition approach for RGB-D cameras. In *International Conference Image Analysis and Recognition* (pp. 21–28).

S.M., A., & P.S., S. (2012). Optimal Score Level Fusion using Modalities Reliability and Separability Measures. *International Journal of Computer Applications*, *51*(16), 1–8. http://doi.org/10.5120/8123-1953

Sainath, T. N., Ramabhadran, B., & Picheny, M. (2009). An exploration of large vocabulary tools for small vocabulary phonetic recognition. In *Automatic Speech Recognition & Understanding, 2009. ASRU 2009. IEEE Workshop on* (pp. 359–364).

Sawakare, P. A., Deshmukh, R. R., & Shrishrimal, P. P. (2015). Speech Recognition Techniques: A Review, *6*(8), 1693–1698.

Seong, T. W., & Ibrahim, M. Z. (2018). A Review of Audio-Visual Speech Recognition. *Journal of Telecommunication, Electronic and Computer Engineering (JTEC)*, *10*(1–4), 35–40.

Seong, T. W., Ibrahim, M. Z., Arshad, N. W. B., & Mulvaney, D. J. (2018). A Comparison of Model Validation Techniques for Audio-Visual Speech Recognition. In K. J. Kim, H. Kim, & N. Baek (Eds.), *IT Convergence and Security 2017: Volume 1* (pp. 112–119). Singapore: Springer Singapore. http://doi.org/10.1007/978-981-10-6451-7_14

Shah, D., Han, kyu j., & Narayanan, shrikanth s. (2010). Robust Multimodal Person Recognition Using Low-Complexity Audio-Visual Feature Fusion Approaches. *International Journal of Semantic Computing*, *4*(2), 155–179. http://doi.org/10.1142/S1793351X10000985

Sheng, X., & Hu, Y.-H. (2005). Maximum likelihood multiple-source localization using acoustic energy measurements with wireless sensor networks. *Signal Processing, IEEE Transactions on*, *53*(1), 44–53. http://doi.org/10.1109/TSP.2004.838930

Shrawankar, U., & Thakare, V. (2010). Feature Extraction for a Speech Recognition System in Noisy Environment: A Study. *Computer Engineering and Applications (ICCEA), 2010 Second International Conference on*, *1*, 358–361. http://doi.org/10.1109/ICCEA.2010.76

Signal Processing Information Base (SPIB) Noise Data. (n.d.). Retrieved August 27, 2018, from http://spib.linse.ufsc.br/

Sklansky, J. (1982). Finding the Convex Hull of a Simple Polygon. *Pattern Recogn. Lett.*, *1*(2), 79–83. http://doi.org/10.1016/0167-8655(82)90016-2

Souza, P. E., Boike, K. T., Witherell, K., & Tremblay, K. (2007). Prediction of Speech Recognition from Audibility in Older Listeners with Hearing Loss: Effects of Age, Amplification, and Background Noise. *Journal of the American Academy of Audiology*, *18*(1), 54–65. http://doi.org/10.3766/jaaa.18.1.5

Stewart, D., Seymour, R., Pass, A., & Ming, J. (2014). Robust audio-visual speech recognition under noisy audio-video conditions. *IEEE Transactions on Cybernetics*, *44*(2), 175–184. http://doi.org/10.1109/TCYB.2013.2250954

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *The Journal of the Acoustical Society of America*, *26*(2), 212–215.

Taal, C. H., Hendriks, R. C., Heusdens, R., & Jensen, J. (2011). An algorithm for intelligibility prediction of time-frequency weighted noisy speech. *IEEE Transactions on Audio, Speech and Language Processing*, *19*(7), 2125–2136. http://doi.org/10.1109/TASL.2011.2114881

Terry, L. H., Shiell, D. J., & Katsaggelos, A. K. (2008). Feature space video stream consistency estimation for dynamic stream weighting in audio-visual speech recognition. *Proceedings - International Conference on Image Processing, ICIP*, 1316–1319. http://doi.org/10.1109/ICIP.2008.4712005

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

The NIST Speech SNR Measurement. (n.d.). Retrieved August 27, 2018, from https://www.nist.gov/information-technology-laboratory/iad/mig/nist-speech-signal-noise-ratio-measurements

Tripathy, S., Baranwal, N., & Nandi, G. C. (2013). A MFCC based Hindi speech recognition technique using HTK Toolkit. *2013 IEEE 2nd International Conference on Image Information Processing, IEEE ICIIP 2013*, (January 2016), 539–544. http://doi.org/10.1109/ICIIP.2013.6707650

Varga, A., & Steeneken, H. J. M. (1993). Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems. *Speech Communication*, *12*(3), 247–251.

VimalaC. (2012). A Review on Speech Recognition Challenges and Approaches. *World of Computer Science and Information Technology Journal*, *2*(1), 2221–741.

Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. *Computer Vision and Pattern Recognition (CVPR)*, *1*, I--511--I--518. http://doi.org/10.1109/CVPR.2001.990517

Wu, Z., Cai, L., & Meng, H. (2005). Multi-level Fusion of Audio and Visual Features for Speaker Identification. *Advances in Biometrics*, 493–499. http://doi.org/10.1007/11608288

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., … others. (2002). The HTK book. *Cambridge University Engineering Department*, *3*, 175.

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., … others. (2006). The HTK book (for HTK version 3.4). *Cambridge University Engineering Department*, *2*(2), 2–3.

Zhao, G., Barnard, M., & Pietikainen, M. (2009). Lipreading with local spatiotemporal descriptors. *IEEE Transactions on Multimedia*, *11*(7), 1254–1265.

Ziaei, A., Kaushik, L., Sangwan, A., Hansen, J. H. L., & Oard, D. (2014). Speech activity detection for NASA apollo space missions: Challenges and solutions. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, (September), 1544–1548.