A Comparison of Model Validation Techniques on Audio-Visual Speech Recognition

Thum Wei Seong[1], M.Z. Ibrahim[1*], Nurul Wahidah Binti Arshad[1], D. J. Mulvaney[2]

[1]Faculty of Electrical & Electronic Engineering University Malaysia Pahang,
26600 Pekan, Pahang, Malaysia
weiseong91@hotmail.com, *zamri@ump.edu.my, wahidah@ump.edu.my

[2]School of Electronic, Electrical and Systems Engineering, Loughborough University,
LE11 3TU, United Kingdom
d.j.mulvaney@lboro.ac.uk

Abstract. This paper implements and compares the performance of a number of techniques proposed for improving the accuracy of Automatic Speech Recognition (ASR) systems. As ASR that uses only speech can be contaminated by environmental noise, in some applications it may improve performance to employ Audio-Visual Speech Recognition (AVSR), in which recognition uses both audio information and mouth movements obtained from a video recording of the speaker's face region. In this paper, model validation techniques, namely the holdout method, leave-one-out cross validation and bootstrap validation, are implemented to validate the performance of an AVSR system as well as to provide a comparison of the performance of the validation techniques themselves. A new speech data corpus is used, namely the Loughborough University Audio-Visual (LUNA-V) dataset that contains 10 speakers with five sets of samples uttered by each speaker. The database is divided into training and testing sets and processed in manners suitable for the validation techniques under investigation. The performance is evaluated using a range of different signal-to-noise ratio values using a variety of noise types obtained from the NOISEX-92 dataset.

Keywords: Audio-visual speech recognition, Hidden Markov model, HTK Toolkit, Holdout validation. Leave one out cross validation, Bootstrap validation.

1 INTRODUCTION

A well-establish audio-visual speech recognition (AVSR) system capable be a guide line for other researchers regrading to the techniques of features extractions, frond end process, model integration, classification and validation methods. All the techniques used in their work should clearly and properly stated, so future enhancement could be done by other researchers. Although, combining two modalities achieves better performance than single modalities, but unfortunately, the validation techniques on dataset samples will contribute different performance accuracy. Many researchers been

searching for the best model validation recently. But, the prior studies have arrived at contradictory conclusions on it. There is still no such a commonly agreed the best model validation, which contribute the most consistent and accurate estimation. Some previous work conclude that bootstrap validation is better[1], while others claim that Leave-one-out cross validation (LOOCV) can achieve most accurate result[2].

In this paper, a comparison of the validation techniques (holdout, LOOCV and bootstrap) have been done based on the AVSR system developed. In more detail, Section 2 concentrates on the overview explanation of model validation techniques. Then, Section 3 present the methodology to be adopted to analyze the AVSR system. Performance result and comparison of different types of model validation techniques being addressed in Section 4. Lastly, conclusions are discussed in Section 5.

2 MODEL VALIDATION TECHNIQUES

According to the previous study, there is few common validation methods to estimate the performance, such as holdout method, cross validation and bootstrap validation[3].

2.1 HOLDOUT METHOD

Holdout method can be consider as a basic validation method for result estimation. It simply divide the samples set into two set, one is training set and another one is testing set. Basically, this methods can be perform well if the training set contains no corrupted data. In reality, corrupted data is hard to be detected while having hundreds of samples. If the training set having corrupted data, it will causing a poor performance recognition when evaluated by testing set. Although this method having drawback, but there is still many research work evaluated by using this method[4][5].

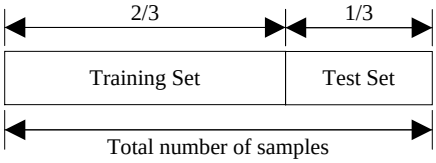

Fig. 1. Example of Hold-out validation distribution ratio

2.2 LEAVE ONE OUT CROSS VALIDATION (LOOCV)

Next, LOOCV is the extreme case of K-fold cross validation, where K represent the total number of samples. It means if the samples dataset having K samples, then the validation process will repeated for K times. Single sample will used for testing purpose, while K-1 samples used for training purpose. According to previous work[2], this techniques proven it having least bias and able to overcome the drawback of holdout method too. However, this work claim a conclusion cannot be made where

which method is superior over other method as the training and testing set during cross validation is quite different. So, LOOCV is used in this work for result comparison, because previous work suggest LOOCV should be the deterministic method for future result comparison.

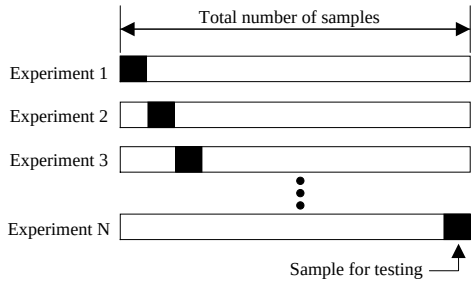

Fig. 2. Illustration diagram of leave-one-out cross validation

2.3 BOOTSTRAP VALIDATION

Bootstrap is a model validation techniques with random of N size number of samples set is picked with replacement from original samples set which is size of N. Those random picked N size samples will used for training, while for those not selected samples will used for testing. That process will keep repeated for M times, and the final performance estimation will be obtain by averaging those M times results.

The figure below shows the example of replacement process for each experiment. The example below showing the complete set sample with X_1, X_2, X_3, X_4 and X_5. For the experiment 2, once X_2 and X_4 are selected as test set, then the rest X_1, X_3 and X_5 will be as train set with 2 samples repeated, where the train set now become X_1, X_5, X_3, X_3 and X_5. The process repeated for K times, then the final result will be averaging from all experiment set.

Table 1. Illustration sample diagram of bootstrap validation

Total samples = X1, X2, X3, X4, X5		
Experiment set	Training set	Testing set
Set 1	X1 X2 X3 X5 X5	X4
Set 2	X1 X3 X3 X5 X5	X2 X4
Set 3	X1 X1 X2 X2 X4	X3 X5
	. . .	
Set K	X1 X3 X3 X3 X3	X2 X4 X5

3 METHODOLOGY

This work is an extension of previous research and mainly focus on result comparison of cross validation techniques. Matlab R2015a is used for simulation testing with open source image processing library. Then, HTK is used to generate and manipulate the 9-states of HMM with speech processing library[6] in this work. HTK was originated from Machines Intelligence Laboratory in Cambridge University Engineering Department[7].

3.1 VISUAL FEATURE EXTRACTION

Firstly, the speaker visual information will extracted using geometrical-based features. It using Viola-Jones face detection algorithm[8] which including mouth face and mouth detection process. HSV colour filter was applied to differentiate the lip region[9], then border following[10] and convex hull techniques used to get the exact complete actual shape of speaker lip. The visual feature extraction techniques was follow the exact steps from previous research[11]. It also shown that this extraction technique is robust to head rotation and illumination changes[12].

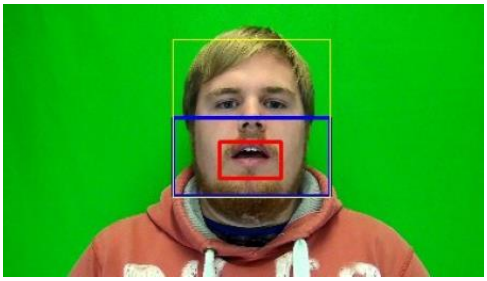

Fig. 3. Example of face and mouth detection

3.2 AUDIO FEATURE EXTRACTION

According to prior study, Mel frequency cepstral coefficient (MFCC) and Linear prediction coefficient (LPC) seen to be famous audio feature extraction technique recently[13]. There is a comparison done to prove that MFCC is typically suitable feature to represent human speech and seen to be outperform than LPC in previous work[14].

In this work, HTK library was applied for MFCC feature extraction and there is total 39 dimension of feature vector, which including the dynamic feature (delta-MFCCs and delta-delta-delta MFCCs). Its dynamic feature is proved that able to improve the performance of speech recognition[15]. The flow of audio-visual speech recognition of this system are shown in the block diagram below.

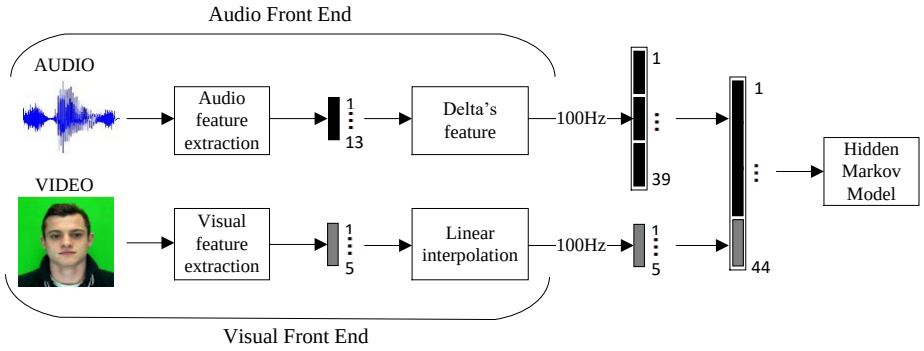

Fig. 4. Block diagram of AVSR

4 EXPERIMENTAL RESULT

This experiments are conducted based on the newly developed database which is Loughborough University Audio-Visual (LUNA-V) speech data corpus[11]. It having high resolution of visual image which is 1280x720 pixels, which higher resolution of image contribute more information and capable to improve performance of AVSR system[16]. The database consisted 10 speakers (9 males and 1 female) with each utterance 5 samples of English digit from 'zero' to 'nine'. Varies noises (white, babble and factory1) with different signal-to-noise ratio (SNR) from NOISEX-92 database are used to test the robustness of the AVSR system.

The accuracy of digit recognition for Holdout, LOOCV and bootstrap validation techniques with different types of noises introduced with SNR value in the interval 25dB to -10dB are listed in the table below.

Table 2. Word accuracy (%) of different types of validation techniques when "white noise" applied. Bold number represent highest accuracy for the selected SNR value.

SNR (dB)	Holdout	LOOCV	Bootstrap
clean	**100**	99.4	98.1
25	95.5	**97.6**	94.2
20	94.0	**94.6**	90.0
15	84.0	**86.2**	80.7
10	72.0	**73.2**	69.0
5	58.5	**60.4**	56.5
0	46.0	**50.0**	47.0
-5	**44.0**	41.4	40.2
-10	**37.0**	36.2	35.6

White noise is a type of noise that acquired by sampling all the different frequency of audible sound. It can highly effect the recognition of word ‘six’ as the pronunciation only required minimum movement of lip. According to the Table 2, LOOCV achieved higher accuracy than other two validation techniques in the interval SNR value of 20dB to 0dB. Bootstrap seen slightly lower than the performance of LOOCV in all SNR value from 25dB to -10dB.

Table 3. Word accuracy (%) of different types of validation techniques when “babble noise” applied. Bold number represent highest accuracy for the selected SNR value.

SNR (dB)	Holdout	LOOCV	Bootstrap
clean	**100**	99.4	98.1
25	**99.5**	99.2	97.4
20	**99.0**	**99.0**	96.5
15	96.5	**97.0**	93.7
10	90.5	**91.0**	87.1
5	79.0	**81.2**	77.7
0	64.0	**67.4**	63.4
-5	49.0	**50.6**	48.8
-10	43.0	**43.5**	42.1

Table 3 showing the result that audio signal corrupted by babble noise in NOISEX-92, which it represent 100 people speaking in a canteen. It affect the recognition of digit ‘seven’. LOOCV almost achieved the highest performance compare to other two validation techniques. It achieved highest value from SNR value 20dB to -10dB.

Table 4. Word accuracy (%) of different types of validation techniques when “factory1 noise” applied. Bold number represent highest accuracy for the selected SNR value.

SNR (dB)	Holdout	LOOCV	Bootstrap
Clean	**100**	99.4	98.1
25	**99.5**	99.2	97.2
20	97.5	**98.8**	95.8
15	92.5	**96.0**	92.2
10	86.5	**89.2**	83.9
5	75.0	**78.2**	73.1
0	59.0	**62.2**	58.8
-5	45.0	**48.6**	46.2
-10	**41.5**	39.4	39.3

In Table 4, the noise used to contaminate the audio signal is factory1 noise. It was recorded near to plate-cutting and electrical. Then, from the SNR value 20dB to -5dB, LOOCV again achieved the highest accuracy compare to holdout and bootstrap validation. Performance of bootstrap validation is slightly lower than holdout method in all range of SNR value.

In overall, bootstrap remain the lowest accuracy in all range of SNR value. Holdout method seen perform very well during the SNR value in the interval clean to 25dB. This may happen due to corrupted samples involved during training session and this will lead to a biased result. Besides that, previous work proved that much data do not involved for training session, this caused insufficient data to train predictive model[17].

5 CONCLUSION

This paper presented result comparison on varies validation techniques on English digit speech recognition system using high definition LUNA-V data corpus to analyze the word accuracy in noisy environments. The validation techniques used are normal hold-out method, LOOCV and bootstrap validation. Based on the experiment result and comparison analysis, LOOCV techniques achieved slightly higher accuracy percentage compare to holdout and bootstrap validation. This technique manage to evaluate quality of every samples and gain the final accuracy by averaging the result from each samples.

6 ACKNOWLEDGMENTS

This work was supported by Universiti Malaysia Pahang and funded by the Ministry of Higher Education Malaysia under FRGS Grant RDU160108.

7 REFERENCES

1. K. Kokkinidis, A. Panagi, and A. Manitsaris, “Finding the optimum training solution for Byzantine Music Recognition - a Max / Msp approach .,” pp. 6–9 (2016).
2. E. Kocaguneli and T. Menzies, “Software effort models should be assessed via leave-one-out validation,” *J. Syst. Softw.*, vol. 86, no. 7, pp. 1879–1890 (2013).
3. R. Kohavi, “A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection,” *Int. Jt. Conf. Artif. Intell.*, vol. 14, no. 12, pp. 1137–1143 (1995).
4. S. Receveur, D. Scheler, and T. Fingscheidt, “A turbo-decoding weighted forward-backward algorithm for multimodal speech recognition,” pp. 179–192 (2014).
5. M. Z. Ibrahim, D. J. Mulvaney, and M. F. Abas, “Feature-fusion based audio-visual speech recognition using lip geometry features in noisy enviroment,” *ARPN J. Eng. Appl. Sci.*, vol. 10, no. 23, pp. 17521–17527 (2015).
6. S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, and others, “The HTK book (for HTK version 3.4),” *Cambridge Univ. Eng. Dep.*, vol. 2, no. 2, pp. 2–3 (2006).

7. G. S. Pawar and S. S. Morade, “Isolated English Language Digit Recognition Using Hidden Markov Model Toolkit,” vol. 4, no. 6, pp. 781–784 (2014).
8. P. Viola and M. Jones, “Rapid object detection using a boosted cascade of simple features,” *Comput. Vis. Pattern Recognit.*, vol. 1, p. I--511--I--518 (2001).
9. 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).
10. 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).
11. Z. Ibrahim, “A novel lip geometry approach for audio-visual speech recognition,” (2014).
12. M. Z. Ibrahim and D. J. Mulvaney, “Robust geometrical-based lip-reading using hidden Markov models,” *IEEE EuroCon 2013*, no. July, pp. 2011–2016 (2013).
13. 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).
14. S. Tripathy, N. Baranwal, and G. C. Nandi, “A MFCC based Hindi speech recognition technique using HTK Toolkit,” *2013 IEEE 2nd Int. Conf. Image Inf. Process. IEEE ICIIP 2013*, no. January 2016, pp. 539–544 (2013).
15. N. S. A. Wahid, P. Saad, and M. Hariharan, “Automatic Infant Cry Pattern Classification for a Multiclass Problem,” vol. 8, no. 9, pp. 45–52 (2016).
16. A. G. Chitu and L. J. M. Rothkrantz, “Building a Data Corpus for Audio-Visual Speech Recognition,” vol. 1, no. Movellan 1995 (2007).
17. C. Tantithamthavorn, S. Mcintosh, A. E. Hassan, and K. Matsumoto, “An Empirical Comparison of Model Validation Techniques for Defect Prediction Models,” *IEEE Trans. Softw. Eng.*, vol. 5589, no. c, pp. 1–16 (2016).