A Comparison of Model Validation Techniques for 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 be 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 models, HTK toolkit, holdout validation, leave-one-out cross validation, bootstrap validation.

1 INTRODUCTION

This work adopts an established audio-visual speech recognition (AVSR) system that uses a range of modern techniques for feature extraction, frond-end processing, model integration, classification approaches and validation methods. Although, it would initially appear that combining two modalities (audio and visual) is likely to result in better overall system performance, many AVSR researchers have found this not to be the case in practice and this is at least partly due to poor selection of a validation technique to apply to dataset samples. Although a large body of literature exists that confirms that researchers are aware of the need to identify a suitable validation technique that provides the most consistent and accurate estimation, no consensus has been reached. For example, one recent study found that the bootstrap validation approach was the best [1], while another claimed that leave-one-out cross validation (LOOCV) achieves the most accurate classification results [2].

In this paper, a comparison of three validation techniques (holdout, LOOCV and bootstrap) for an AVSR system is carried out. Section 2 concentrates explains the model validation techniques, section 3 presents the methodology to be adopted to analyze the AVSR system and performance results of the different types of validation techniques are addressed in Section 4. The conclusions are discussed in Section 5.

2 MODEL VALIDATION TECHNIQUES

This section describes the most popular validation methods for estimating AVSR recognition performance, namely the holdout method, LOOCV and bootstrap validation [3].

2.1 HOLDOUT METHOD

The holdout method can be considered to be one of the most basic validation methods for result estimation. Its operation involves simply dividing the sample set into two; the first is used as a training set and the second is used as a test set, see Fig. 1. The bootstrap method performs well if the training set contains no corrupted data, but in practice corrupted data are often hard to detect among a large set of samples and, if they are not removed, poor performance results when evaluated using the testing set. Despite such drawbacks, there remains a number of applications to which the approach is well suited and there is a considerable body of research that has exploited this method [4][5]'.

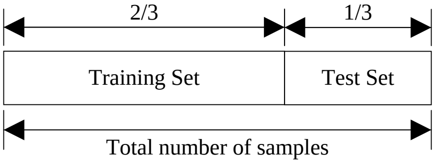

Fig. 1. Example of holdout validation distribution ratio

2.2 LEAVE ONE OUT CROSS VALIDATION (LOOCV)

LOOCV is an extreme case of *k*-fold cross validation, where *k* represents the total number of samples. In *k*-fold cross validation, the validation process is carried out *k* times. In LOOCV, *k*-1 samples are used for training purposes and only a single sample is used for testing. According to Kocaguneli and Menzies [2], this techniques has been shown to have low bias and is able to overcome the drawback of the holdout method in having poor performance in the presence of corrupted data. However, Kocaguneli and Menzies also found that there is no definitive solution regarding whether the holdout approach or the LOOCV method perform the better as the training and test sets used during validation are very different.

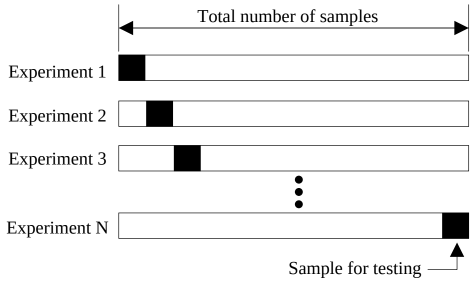

Fig. 2. Illustration diagram of leave-one-out cross validation

2.3 BOOTSTRAP VALIDATION

In the bootstrap model validation technique, assuming that there are *N* samples in the data set, then a number of samples are selected at random and are used for training, while those not selected are used for testing. The process is carried out *M* times and the final performance estimation is obtained by averaging the *M* sets of results.

Table 1 shows an example of replacement process in which the set of selected samples is given by X_1, X_2, X_3, X_4 and X_5, assuming *N*=5 in this case. For example, in experiment set 2, then once X_2 and X_4 are selected as the test set, the training set contains X_1, X_3 and X_5, but, as two of the entries are repeated, the actual entries in the training set become X_1, X_3, X_3, X_5 and X_5. This process is carried out *M* times and the final validation outcome is averaged from all the experiment sets.

Table 1. Example of bootstrap validation, where the samples available are X_1, X_2, X_3, X_4, X_5.

experimental set number	training set	test set
1	X_1, X_2, X_3, X_4, X_5	X_4
2	X_1, X_3, X_3, X_5, X_5	X_2, X_4
3	X_1, X_1, X_2, X_2, X_4	X_3, X_5
⋮	⋮	⋮
M	X_1, X_3, X_3, X_3, X_3	X_2, X_4, X_5

3 METHODOLOGY

This AVSR implementation has been carried in previous research and the contribution of this paper is principally the results of a comparison of cross validation techniques. In the previous work, Matlab R2015a together with the OpenCV open source image processing library was used for simulation and testing and the hidden Markov model toolkit (HTK) was used to generate and manipulate the nine states of a hidden Markov model [6]. HTK originates from the Machines Intelligence Laboratory at Cambridge University’s Engineering Department [7].

A system diagram of the AVSR processes used in this work are shown in Fig. 3.

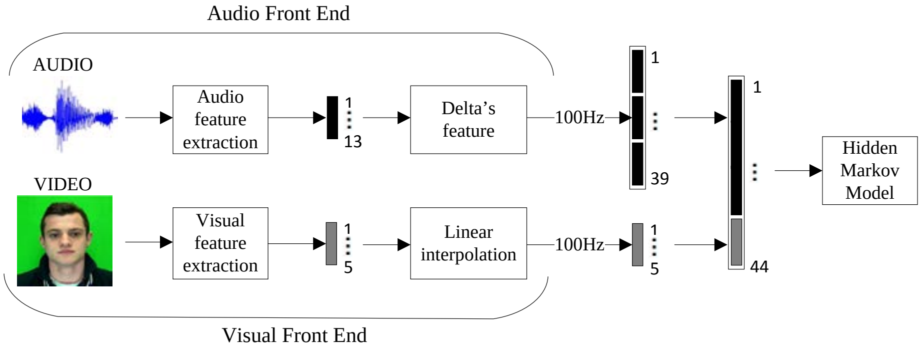

Fig. 3 AVSR processes carried out in this work

3.1 VISUAL FEATURE EXTRACTION

The visual feature extraction techniques followed the steps from previous research [11]. It was also shown that this extraction technique is robust to head rotation and illumination changes[12]. The process is as follows. Firstly, visual information from the speaker is extracted in the form of geometrical-based features. A Viola-Jones face detection algorithm [8] was applied in which face and then mouth detection processes were carried out, as can be seen in Fig. 4. An HSV color filter was applied to differentiate the lip region [9], then border following [10] and finally convex hull techniques were used to extract the actual shape of speaker’s lip.

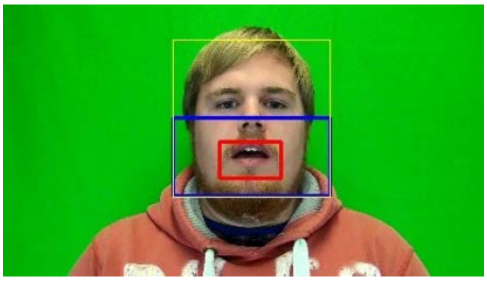

Fig. 4 Example of face and mouth detection

3.2 AUDIO FEATURE EXTRACTION

In the literature, the mel-frequency cepstral coefficient (MFCC) and the linear prediction coefficient (LPC) are currently popular audio feature extraction techniques [13], although a recent investigation has suggested that MFCC may be the better approach to providing human speech features [14].

In this work, the HTK library was employed for MFCC feature extraction and a feature vector of 39 dimensions was obtained. The vector includes dynamic feature (delta-MFCCs and delta-delta MFCCs) as these were shown in previous work to improve the performance of speech recognition systems [15].

4 EXPERIMENTAL RESULTS

This experiments were conducted using the newly developed database known as the Loughborough University audio-visual (LUNA-V) speech data corpus [11]. Compared to other existing databases, the video recordings have a relatively high resolution of 1280x720 pixels, making more detailed information available to the recognition process and so perhaps enabling improvements in the performance of AVSR systems [16]. The database has contributions from 10 speakers (9 male and 1 female) with each speaker providing five separate samples of uttering the English digits from 'zero' to 'nine'. Varies types of noise were applied at a number of different signal-to-noise ratios (SNRs) in order to test the robustness of the AVSR system.

For each of the holdout, LOOCV and bootstrap validation techniques, a range of noise types with SNR values in the interval 25dB to -10dB relative to the speech signals were added. In the results presented here, NOISEX-92 was used to supply the noise signals and the types of noise used are known as 'white', 'babble' and 'factory1' in the database archive

The results of the 'white' noise experiments are shown in Table 2. White noise contains contributions for all frequencies in the audible sound range and is known to have a more profound effect on the perceived audibility of certain words, including 'six' which is not strongly sounded. Furthermore, the word is often difficult for AVSR systems to detect as its production requires only minimal lip movements. As can be seen in Table 2, LOOCV achieved better accuracy in the AVSR tests than other two validation techniques when operating in the SNR range from 20dB to 0dB and holdout only performed well on clean audio and when the strength of the noise signal was greater than that of the speech. Bootstrap consistently performed the worst of the three methods.

Table 2 Word accuracy of the validation techniques when 'white noise' is added to the speech signals. Figures in bold type show the technique producing the best result at each SNR value.

SNR (dB)	holdout (%)	LOOCV (%)	bootstrap (%)
clean	**100.0**	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

Table 3 shows the recognition results when the speech signals were corrupted by ‘babble’ noise, which was captured from 100 people talking in a canteen. The digit ‘seven’ was found to be the word most adversely affected in the recognition results. Apart from at very low noise levels where its performance was only slightly worse than holdout, LOOCV achieved the best performance. Again, bootstrap performed the worst of the three methods.

Table 3 Word accuracy of the validation techniques when ‘babble noise’ is added to the speech signals. Figures in bold type show the technique producing the best result at each 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

In Table 4, the noise used to contaminate the audio signal was ‘factory1’ noise, recorded in the proximity of plate-cutting and electrical equipment. Again, except in cases where the noise content was very low or very high, LOOCV achieved the greatest accuracy compared to the holdout and bootstrap validation methods. The performance of the bootstrap validation method was again somewhat worse across the full range of SNR values.

Table 4 Word accuracy of the validation techniques when ‘factory1 noise’ is added to the speech signals. Figures in bold type show the technique producing the best result at each 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

Overall, the bootstrap methods exhibited the worst accuracy across the full range of SNR values. The holdout method performed particularly well when there was no noise contamination and when little noise was present. It is known that the holdout method is particularly susceptible to the presence of corrupted samples and if any were present during training, this could have led to a biased result. Furthermore, from previous work, the holdout method is known to be more sensitive to the quantity of data used in training, and if the number of values used was insufficient this may have also affected the accuracy available from the predictive model [17].

5 CONCLUSION

This paper has presented a comparison of the speech recognition results generated by a range of validation techniques when tested on the word accuracy of an AVSR operating in noisy environments. The work used an existing AVSR system that attempted to recognize English digits using a combination of speech and high-definition video sequences from the LUNA-V data corpus. Based on the experiment results, the LOOCV technique achieved a slightly better performance compared to the holdout and bootstrap validation methods.

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,” *5th Int. Conf. on Modern Circuits and Systems Technologies*, 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,” *PhD thesis, Loughborough University* (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,” *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,” *Proc Euromedia*, pp.88-92 (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).