A Data Driven Approach to Audiovisual Speech Mapping

Andrew Abel[1], Ricard Marxer[2], Amir Hussain[1], Jon Barker[2], Roger Watt[1], Bill Whitmer[3], and Peter Derleth[4]

[1] University of Stirling, Scotland, FK9 4LA
[2] University of Sheffield, Sheffield, S1 4DP, UK
[3] MRC/CSO IHR - Scottish Section, GRI, Glasgow, G31 2ER, Scotland
[4] Sonova AG, Switzerland, 8712 Staefa
{aka,ahu}@cs.stir.ac.uk,r.j.watt@stir.ac.uk
{r.marxer,j.barker}@sheffield.ac.uk,
william.whitmer@nottingham.ac.uk,peter.derleth@sonova.com,

Abstract. The concept of using visual information as part of audio speech processing has been of significant recent interest. This paper presents a data driven approach that considers estimating audio speech acoustics using only temporal visual information without considering linguistic features such as phonemes and visemes. Audio (log filterbank) and visual (2D-DCT) features are extracted, and various configurations of MLP and datasets are used to identify optimal results, showing that given a sequence of prior visual frames an equivalent reasonably accurate audio frame estimation can be mapped.

Keywords: Audiovisual, Speech processing, speech mapping, ANNs

1 Introduction and Background

There has been much recent research investigating the use of visual information as part of speech processing systems. The relationship between audio and visual aspects of speech production and perception has been heavily investigated in the literature, proving the relationship between audio speech and lip movements [13],[18]. A detailed summary can be found in Abel and Hussain [1]. A number of multimodal approaches have been proposed for speech filtering, including using visual information for beamforming and noise cancellation [3], [1]. There have also been attempts to map audio and visual speech to each other [11] [15] [6].

Recent work such as [12] [4], attempts to use visual data for “lip reading”, by mapping lip information to audio phonemes or words, which is a linguistic basis to work from. For this, they generate visemes and compare them to speech units such as words, syllables, or phonemes. By viseme, as discussed by [4], there is no standard definition, with a range of possible definitions such as ”a set of phonemes that have identical appearance on the lips” [5]. However, there are some limitations with this approach. There is not a complete one-to-one mapping of phonemes to visemes, as one viseme can be mapped to several phonemes

[4], which makes classification challenging. Another issue with using visemes is co-articulation, where a speaker starts to form words before they are spoken, resulting in a phone being pronounced differently due to the effect of adjacent phonemes, which was identified to have a negative effect on lipreading results [4]. Finally, while a linguistic approach (using phonemes and linguistic information) can be particularly useful for speech recognition, in speech filtering technologies, a frame based approach [7] is often used, i.e. processing on a frame-by-frame basis, rather than using discrete linguistic units.

Other research has considered the use of visual information to estimate an audio frame for speech filtering [3] [1], by estimating clean audio from visual information as part of a Wiener filter to remove noise from speech. However, the results are extremely limited, due to the narrow range of data used for the audiovisual speech models, and also the limited evaluation of the mapping performance alone. To develop an improved system, it is considered necessary to first evaluate the mapping process alone. Although this approach may not capture the fine time information required to produce a detailed speech estimation (which would be required for synthesis) [16], it has been argued [3] that this could produce a smoothed estimation which could be feasibly used for filtering.

This paper contributes a new initial investigation of visual to audio mapping. Rather than working on a linguistic basis, it purely considers the data on a frame-by-frame basis, and attempts to identify conditions that produce the best audiovisual mapping. A large multi-speaker dataset (the Grid corpus [8]) and different configurations of a non-linear neural network are used to identify optimal parameters and the best use of data for estimating an audio feature vector, given only visual information as input. This could arguably be considered to be a data driven, rather than a language driven, approach.

2 Dataset and Feature Extraction

2.1 Grid Corpus

For the research in this paper, we used the Grid Corpus[8], an audiovisual dataset which contains 34 speakers, each reciting 1000 command sentences (e.g. "bin blue on red seven now"). Clean audio and video recordings are available. We use five speakers, two white males, two white females, and one black male speaker. This means that there are clear differences between speakers. Two different sets of sentences were used. Firstly, the full fixed length Grid sentences, and also an end pointed set, where the sentences were cut to exclude all silences preceding and following the sentence, which will be referred to as the aligned sentences.

2.2 Audio Feature Extraction

This work uses log filterbank (log-fb) audio features, and assumes a filterbank (fb) of 23 filters, based on other work in the literature [3] [1]. The fb of the audio signal is logarithmically compressed to produce the log-fb signal. This is very similar to the implementation found in previous work [2]. First, given an audio signal S, this is windowed and overlapped using a hamming window to produce 100 vectors per second, which at the chosen sampling rate of 50kHz results in N 16ms frames $s(n)$, where $n = 1...N$, with a 62.5% increment, with each $s(n)$ consisting of 800 samples. This is then Fourier transformed to produce $s_{\Phi}(n)$, and the final log-fb output is generated, following the approach outlined in [2].

2.3 Visual Feature Extraction

Visual data is extracted from video files following the process described fully in [2]. Firstly, the video file is divided into individual image frames. To identify and track the Region of Interest (ROI), a Viola-Jones lip detector [19] is used on the initial frame. This ROI is then tracked from frame-to-frame with an online shape model [17]. This outputs a 2-dimensional lip region for each image frame, which is then used for cropping the original frames to leave only the lip region. This was found in [2] to be an accurate approach, however, as this is an automated process (due to the quantity of data processed), each sentence was validated to ensure that examples of poor lip tracking (such as cases where only a partial lip region was correctly cropped) or camera glitches are removed. This was done by inspecting a number of frames from each sentence, and then deleting sentences where the lip region was not fully identified in any frame.

Both the full and aligned sentence datasets were extracted. The majority of removed sentences were cut because of a small portion of the ROI was not precisely captured. This resulted in the datasets given in table 1.

Table 1. Summary of sentences used from the Grid Corpus, and results of validating fully automated lip tracking, with number of sentences used and removed.

			Full Sents.		Aligned Sents.	
Speaker ID	Grid ID	No. of Sents	Rem.	Used	Rem.	Used
Speaker 1	S1	1000	11	989	11	989
Speaker 2	S15	1000	164	836	164	836
Speaker 3	S26	1000	16	984	71	929
Speaker 4	S6	1000	9	991	9	991
Speaker 5	S7	1000	11	989	11	989

The 2D-DCT (discrete cosine transform) vector $F_v = 2D - DCT\,(v)$ of each cropped lip region in the sequence is then found, a commonly used technique [3], [2],[5]. For a $V_U \quad x \quad V_V$ matrix V_P of pixel intensities (i.e. the cropped lip region), a 1D-DCT is applied to each row of V_P and then to each column of the result. The first 24 2D-DCT components of each image are vectorised in a zigzag order to produce the final frame vector. The resulting 2D-DCT sequence is then interpolated to match the equivalent audio sequence. As the video was recorded at 25 fps, it is upsampled to match the 100 frames per second rate of the audio features by using each visual feature frame for four consecutive audio frames, utilising the same approach as described in previous research [1] [2].

3 Results

3.1 Methodology

A number of Multi-Layer Perceptron (MLP) neural networks are trained. The datasets were divided into training (80%) and test (20%) sets, with the exact content varying due to the dataset being randomly split for each run. The datasets consist of visual frame vectors as inputs, with the equivalent audio log fb vector used as the label. Although there are many network topologies and approaches available that can be used, including deep learning [15], it was decided that the quantify of data available and the preliminary nature of this work justified the use of an MLP. A number of parameters were varied, namely the datasets, the number of visual frames used to estimate an audio frame, and the hidden layer size. The networks were trained in Matlab. For each individual configuration, 10 different networks were initialised, with the training and test datasets permuted randomly before each run. It was decided for consistency to focus on a single hidden layer, rather than considering deep networks with multiple hidden layers.

For evaluation, Mean Squared Error (MSE) was used, as this research takes a data driven approach to estimation, and so rather than trying to predict a complete phoneme or word, the aim is to produce the closest match between an estimated audio fb frame generated by the MLP using visual information, and the actual labelled frame. As the resulting MSE performance of each 10 runs of a network was not always symmetrical, it was best to calculate the median rather than the mean, and also use interquartile range (IQR) to signify the deviation.

3.2 Sentence Length Evaluation

Firstly, two possible datasets were considered, using all frames in each chosen sentence, or considering only aligned data. As discussed, five speakers, and 900 (or the maximum number available, see table 1) sentences from each were used for each configuration. All audio and visual features are extracted, and the resulting vector pairings are shuffled randomly. For full sentences, this resulted in an initial 1,326,364 visual and audio vector pairings to be divided into training and test sets, and for the aligned sentences, 804,160 pairs. Initially, a comparison was made between full and aligned data for MLP hidden layer sizes ranging between 10 neurons and 150 neurons. The results of test data evaluation are given in figure 1, and results are given in table 2.

Considering the initial results, it appears that the overall median full sentence results are better than aligned sentences, and that good results can be achieved with a relatively small hidden layer. However, a closer inspection of the individual results within the datasets showed that because the full sentences contained a relatively large number of frames (522,204) that could be classified as silence, the dataset was over-represented for silence, and this resulted in the network being able to very effectively distinguish between silence and non-speech frames, resulting in a potentially good voice activity detector (VAD). However, the main interest in this paper is speech mapping, and so although there is a higher

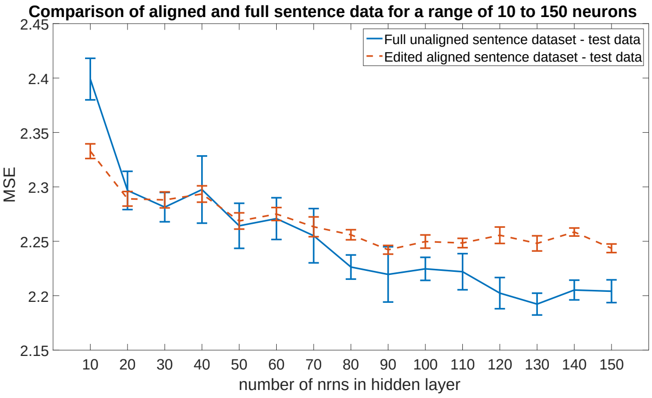

Fig. 1. MSE median test data results of using whole (solid blue line) or aligned (red dashed line) sentences for different MLP hidden layer sizes.

Table 2. Selected median MSE results of test data for aligned and full sentences for different MLP hidden layer sizes, giving the median and the IQR.

		Dataset type			
		Full Sentence		Aligned	
Sentence type	No. nrns	median	IQR	median	IQR
Full Sentence	10	2.399	0.019	2.333	0.007
	30	2.282	0.014	2.288	0.007
	60	2.271	0.019	2.275	0.006
	90	2.220	0.025	2.242	0.004
	120	2.202	0.014	2.256	0.008
	150	2.204	0.011	2.244	0.004

MSE with aligned data, this is of more relevance to this work. As there is an improved balance to the dataset, an inspection of individual results showed an improvement in performance. Rather than learning silence only, it made more attempts to generalise over a greater variety of audio vectors, resulting in a slightly higher MSE. This is reflected in the noticeably smaller IQR, which shows a consistency in results. It can also be seen that a larger hidden layer (i.e. 90+ neurons) does not generate noticeably improved results, with improvement tending to be within the interquartile range, despite the additional time required for training. Accordingly, for the remainder of this paper, aligned data is used.

3.3 Use of Multiple Visual Frames for Estimation

The effect of using a different number of visual frames is considered, by concatenating prior frames into one large input visual vector. This would theoretically allow movement information to be captured. Prior experiments (not reported here for space reasons) showed that using the temporal differences ($\Delta(t)$) was an improvement over only using one frame, but not as good as using multiple visual frames. A range of visual frames, ranging from one frame (as used previously), to a concatenation of the visual frame with the previous 27 visual frames, are used as the input into vector to generate audio estimations. Test data results are shown in figure 2, and in table 3.

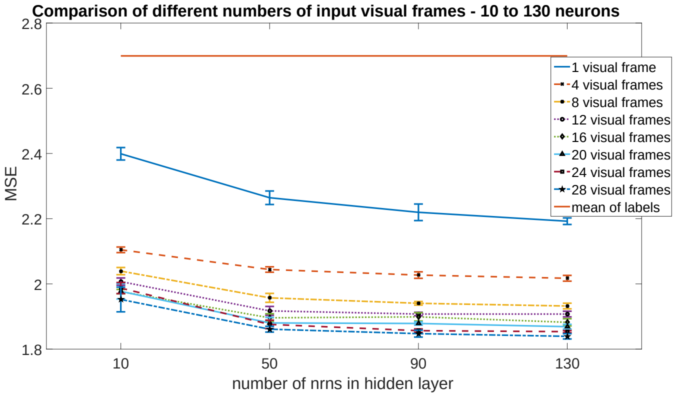

Fig. 2. MSE test data results of using different numbers of visual frames to estimate a single audio frame.

Firstly, the horizontal line at the top of figure 2 shows the result of calculating the mean of all the output labels and using this as the prediction for every frame, giving a baseline average value. There is a large improvement when using an ANN, demonstrating that using visual information has a strong relationship to the audio vector. The solid blue line is the MSE for using a single visual frame. The other lines represent using different numbers of visual frames, i.e. more inputs to the network. Immediately, it is obvious that using multiple visual frames improves performance, with the use of four frames showing a big improvement. This is further reflected with 8 (red dashed line), 12 (yellow dotted line), and 16 frames (purple dotted line), but after this, the improvement is less clear.

Table 3. Selected median and IQR of MSE test data results of using different numbers of visual frames to estimate a single audio frame.

	Hidden Layer Size Median Results							
	10 nrns		50 nrns		90 nrns		130 nrns	
No. vis frames	MSE	IQR	MSE	IQR	MSE	IQR	MSE	IQR
1	2.399	0.019	2.264	0.021	2.220	0.025	2.192	0.010
4	2.105	0.009	2.044	0.008	2.027	0.010	2.017	0.009
8	2.039	0.011	1.957	0.014	1.941	0.006	1.932	0.009
12	2.007	0.017	1.917	0.007	1.908	0.009	1.9075	0.006
16	1.976	0.007	1.896	0.012	1.899	0.013	1.882	0.011
20	1.977	0.024	1.880	0.017	1.879	0.005	1.869	0.007
24	1.987	0.017	1.879	0.013	1.857	0.005	1.854	0.003
28	1.952	0.038	1.861	0.009	1.848	0.011	1.839	0.008

Using more frames continues to improve the results, but only by a small amount each time. Although the IQR (given as error bars) tends to be smaller for multiple frames than for a single visual frame, for larger visual inputs (20,24, and 28 frames), the MSE improvement is sometimes within the IQR of other results, suggesting extremely small improvements. As increasing the input vector size increases training time, this suggests only a limited benefit with very large input vectors. It can also be seen that network performance continues to improve with a larger hidden layer, but only by a small amount, arguably not justifying the much larger required training time. It is clear that using more visual frames is the biggest improvement that can be identified for audio vector prediction. This the benefit of accounting for lip movement, which makes sense, given that there is more information present in the lips than just static information. Other research has identified co-articulation as an issue present in some speakers, which makes phoneme-viseme mapping more challenging. As we use frame based data rather than a viseme, this should be less of an issue.

3.4 Speaker Dependent Training Comparison

Another optimisation approach is to train and test with individual speakers. This results in a smaller dataset being available, but ensures audio and visual greater consistency between vectors. We consider the results for each individual speaker, considering the use of 14 visual frames as input, and only hidden layer sizes up to 100 neurons (chosen based on the results in the previous section). Test data results are shown in figure 3 and table 4. These are compared to the results of training models with all five speakers.

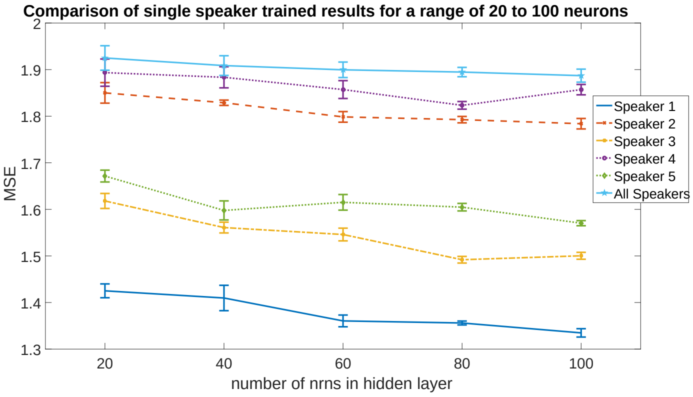

Fig. 3. MSE test data results of using individual speakers.

It can be seen that using a model trained using the combined five speakers produces a consistently higher MSE at all hidden layer sizes than using any single speaker model. This is not unexpected, as there are differences between how individual speakers articulate and sound, and using five speakers does not appear to fully generalise, suggesting that more speakers or more information are required to train a fully robust multi-speaker model. The individual speaker results show an improvement, varying from speaker to speaker.

Speaker 1 (solid blue line) shows a particularly low MSE, signifying a particularly good performance, with speakers 2 and 4 being noticeably worse. However,

Table 4. Selected normalised median MSE results of training and test data for using 14 prior frames, showing the results for individual speaker datasets.

	Hidden Layer Size MSE				
Speaker ID	20 nrns	40 nrns	60 nrns	80 nrns	100 nrns
All	1.925	1.909	1.900	1.895	1.887
S1	1.425	1.410	1.361	1.356	1.335
S2	1.850	1.829	1.799	1.793	1.784
S3	1.618	1.561	1.546	1.492	1.500
S4	1.894	1.884	1.857	1.823	1.857
S5	1.672	1.598	1.615	1.605	1.570

it was not simple to identify why. All speakers seem to articulate clearly, the speech contents are similar, and inspecting example spectrograms for the audio files did not identify any obvious differences. One potential issue could be in the tracking approach, with slight differences in results, depending on the speaker. For example, speaker two tends to have a slightly smaller mouth region with more of the face captured. However, by the same token, speaker 4 tends to have a very close mouth region tracked, with less of the face captured. These could potentially explain the difference, suggesting perhaps some other visual features, such as geometric measurements [14], could be useful.

3.5 Individual Vector Examination

Considering the size of the dataset and the random shuffling involved during the training set, it is challenging to consider individual speech vectors alone, hence the use of MSE as an evaluation metric. For illustrative purposes, six labelled and estimated audio vectors are shown that were generated using the best case model (i.e. a single speaker trained model using 14 prior frames, and 100 neurons in the hidden layer). These are shown in figure 4. The frames shown here are selected pseudo-randomly, although care was taken to include at least one good example and one poor example. In these examples, the y-axis is cropped to focus only on the trained and actual output, with the untrained output in the background (orange line) demonstrating that training has improved the initial network outputs significantly.

Figure 4 shows that in general, the estimation is reasonably good. The actual vector (i.e. the ground truth) is shown with a blue dashed line, and the trained estimation from only visual information is shown with a solid red line. The examples show that the fit is not perfect, which is to be expected, represents a reasonably accurate smoothed estimation. However the bottom and middle right examples in figure 4 show that where the actual data is for consistent low speech energy, the estimate predicts more noise present, particularly at lower frequencies (i.e. the lower fb channels). Overall, it shows that training using only visual information, without any audio input, and with a fairly basic visual feature vector (i.e. DCT) is successful, and the right combination of visual inputs, speaker selection, and network size, can produce a reasonably good estimation, but that there is also still room for improvement.

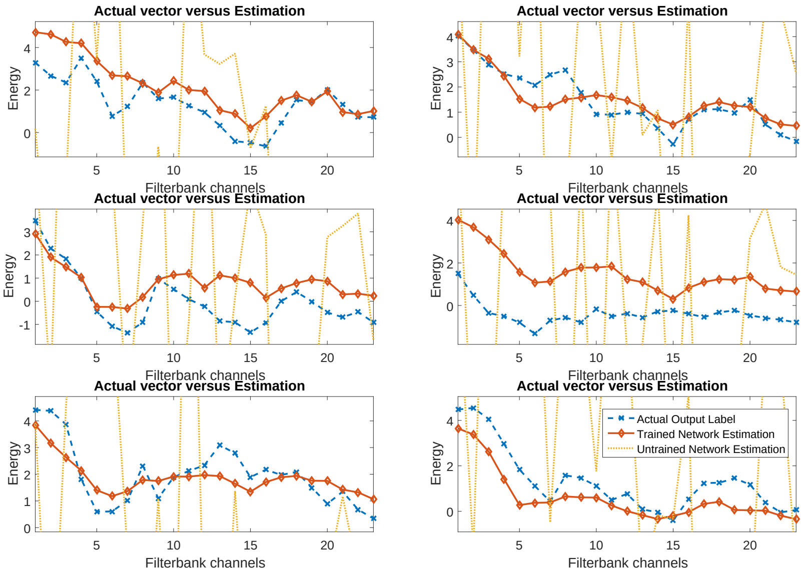

Fig. 4. Examples of individual vector results

4 Conclusions and Future Work

This paper presented a detailed investigation of audiovisual visual to audio speech mapping. A number of different data and neural network configurations were experimented with, showing that optimal results could be gained with speaker specific data, using a large number of prior visual frames, and that a hidden layer of larger than fifty neurons delivers optimal results. These results verify that visual information can feasibly be used to estimate audio features, and serves as a very useful precursor to the development of a more advanced single channel audiovisual speech enhancement system, as outlined in [2]. However, there is still scope for improvement. Other work in the literature [14] uses a range of features to optimise results, including shape biologically inspired barcode features [9], which should be investigated in more depth. There is also scope for using deep neural networks [15], and finally, although this paper focused on data driven mapping, in future work, it will be of interest to investigate the difference in individual parts of speech, such as phonemes and words, both to compare more closely to other research, and for speaker specific analysis.

Acknowledgements

This work was supported by EPSRC Grant EP/M026981/1 (AV-COGHEAR). The authors would also like to gratefully acknowledge the contributions of Prof. Leslie Smith at the University of Stirling and Dr Kristína Malinovská at Comenius University in Bratislava for their help, advice, and contributions.

References

1. Abel, A., Hussain, A.: Novel two-stage audiovisual speech filtering in noisy environments. Cognitive Computation pp. 1–18 (2013)
2. Abel, A., Hussain, A.: Cognitively inspired audiovisual speech filtering: towards an intelligent, fuzzy based, multimodal, two-stage speech enhancement system, vol. 5. Springer (2015)
3. Almajai, I., Milner, B.: Effective visually-derived Wiener filtering for audio-visual speech processing. In: Proc. Interspeech, Brighton, UK (2009)
4. Bear, H., Harvey, R.: Decoding visemes: improving machine lip-reading. In: International Conference on Acoustics, Speech, and Signal Processing (2016)
5. Bear, H.L., Harvey, R.W., Theobald, B.J., Lan, Y.: Which phoneme-to-viseme maps best improve visual-only computer lip-reading? In: Advances in Visual Computing, pp. 230–239. Springer (2014)
6. Cappelletta, L., Harte, N.: Phoneme-to-viseme mapping for visual speech recognition. In: ICPRAM (2). pp. 322–329 (2012)
7. Chung, K.: Challenges and recent developments in hearing aids part i. speech understanding in noise, microphone technologies and noise reduction algorithms. Trends in Amplification 8(3), 83–124 (2004)
8. Cooke, M., Barker, J., Cunningham, S., Shao, X.: An audio-visual corpus for speech perception and automatic speech recognition. The Journal of the Acoustical Society of America 120(5 Pt 1), 2421–2424 (2006)
9. Dakin, S.C., Watt, R.J.: Biological bar codes in human faces. Journal of Vision 9(4), 2–2 (2009)
10. Farkaš, I., Rebrová, K.: Bidirectional activation-based neural network learning algorithm. In: Artificial Neural Networks and Machine Learning–ICANN 2013, pp. 154–161. Springer (2013)
11. Fu, S., Gutierrez-Osuna, R., Esposito, A., Kakumanu, P.K., Garcia, O.N.: Audio/visual mapping with cross-modal hidden markov models. Multimedia, IEEE Transactions on 7(2), 243–252 (2005)
12. Lan, Y., Theobald, B.J., Harvey, R., Ong, E.J., Bowden, R.: Improving visual features for lip-reading. In: AVSP. pp. 7–3 (2010)
13. McGurk, H., MacDonald, J.: Hearing lips and seeing voices. Nature 264, 746–748 (1976)
14. Milner, B., Websdale, D.: Analysing the importance of different visual feature coefficients. FAAVSP 2015 (2015)
15. Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., Ng, A.Y.: Multimodal deep learning. In: Proceedings of the 28th international conference on machine learning (ICML-11). pp. 689–696 (2011)
16. Pahar, M.: A Novel Sound Reconstruction Technique based on a Spike Code (event) Representation. Ph.D. thesis, Computing Science and Mathematics, University of Stirling, Stirling, Scotland (2016)
17. Ross, D.A., Lim, J., Lin, R.S., Yang, M.H.: Incremental learning for robust visual tracking. International Journal of Computer Vision 77(1-3), 125–141 (2008)
18. Sumby, W., Pollack, I.: Visual contribution to speech intelligibility in noise. Journal of the Acoustical Society of America 26(2), 212–215 (1954)
19. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on. vol. 1, pp. I–511. IEEE (2001)