# STREAM WEIGHTS OPTIMISATION FOR AUDIO-VISUAL AUTOMATIC SPEECH RECOGNITION

Nasir Ahmad*, David Mulvaney*, Sekharjit Datta* and Omar Farooq#

n.ahmad@lboro.ac.uk, d.j.mulvaney@lboro.ac.uk, s.datta@lboro.ac.uk, omarfarooq70@gmail.com

*Department of Electronics and Electrical Engineering, Loughborough University, UK

#Department of Electronics Engineering, AMU Aligarh 202 002, India

**Abstract**

Humans use visual speech information from speaker's mouth, especially in presence of noise. Incorporating visual information in automatic recognition of speech by machine has also shown to improve the performance of Automatic Speech Recognition (ASR) systems. As the audio speech is affected severely by even moderate level of noise, the visual speech information can be used to compensate for the audio information loss. It is yet not known as how human combine audio and visual information and the relative weights used with the variation of noise, the weighting of two sources in ASR design is an active research topic. This work investigates the affects of stream weights variation on the performance of Audio-Visual ASRs. The video and audio streams weights are varied at a number of noise levels and the best weights selected at each noise level. The results shows that using the appropriate weights according to noise level could improve the performance of AV-ASR compared to using fixed weights for audio and visual steams.

**KEYWORDS: Audio-visual speech recognition, multimodal integration, stream weights**

## 1. INTRODUCTION

Automatic recognition of speech by computers is potentially the most suitable way of natural man-machine interaction with large number of practical applications. Sumby and Pollack [1] found in 1954 that seeing the talker

face can contribute significantly to human speech perception but its use in automatic speech recognition was introduced three decades later [2]. Since then AV-ASRs has attracted interest of the researchers and a number of areas have been identified including audio-visual front end design, features extraction and audio-visual integration. The research attempts to combine audio and visual speech information in ways similar to human speech recognition.

Human have commendable ability to recognise different voices even in very challenging environments, however very little is known about the exact mechanism of human speech recognition. That is why it has always been a challenging task to construct ASR systems that can resemble human speech recognition [3].

Although successful ASR systems have been developed that are able to perform well under ideal conditions, developing solutions that operate in practical situations where multiple sources or noise are present is yet a substantial challenge [4]. Performance of automatic speech recognition systems, utilising only audio signal for speech recognition degrades severely in presence of noise. Video of the speaker face is not affected by noise and therefore can be used for speech recognition in degraded conditions [5].

Though combining audio and visual speech information has shown to improve the performance of ASR, the relative weighing of audio and visual streams at different noise level is not yet fully explored. Some work on streams weighing is reported in literature [6], [7] but is mostly task specific and therefore lacks generality. This paper presents an investigation of stream weighing on large vocabulary continuous speech recognition task. The stream weights for the video modality are varied from zero (audio-only) to one (video-only) at different noise levels.

## 2. AUDIO-VISUAL DATABASE

Only a few databases suitable for research on audio-visual speech recognition are available. This is partly due to the larger space requirements for video and issues related to the speaker's identity. In addition some databases are developed for specific task and therefore focus on information required for a particular approach. For example, in databases intended for geometric feature based approaches some sort of marking is often added to the speakers' lips for accurate lip contour estimation. This makes these databases unsuitable for general purpose research. To the best of author's knowledge, there are only two databases that contain large vocabularies with large number of speakers and are suitable for common AVASR research. These are audio-visual TIMIT (AVTIMIT) [5] and Vid-TIMIT [8] databases.

In this work, a subset of the Vid-TIMIT database consisting of 32 speakers (16 male and 16 female) is used. Each speaker utters eight different sentences containing a total of 925 words, from which 24 speakers with 216 sentences are used for training and the remaining 8 speakers with 40 sentences kept for testing purposes. Video is supplied at a rate of 25 frames per second and at a resolution of 512x385. Audio is stored at 32 kHz at a depth of 16 bits.

## 3. EXPERIMENTAL SETUP

The operation of an Audio-Visual ASR (AVASR) is described in Figure 1. The task involved in AVASR design can be divided into region of interests (ROI) extraction, features extraction and audio-visual integration. In following these tasks are explained in context of work described in this paper.

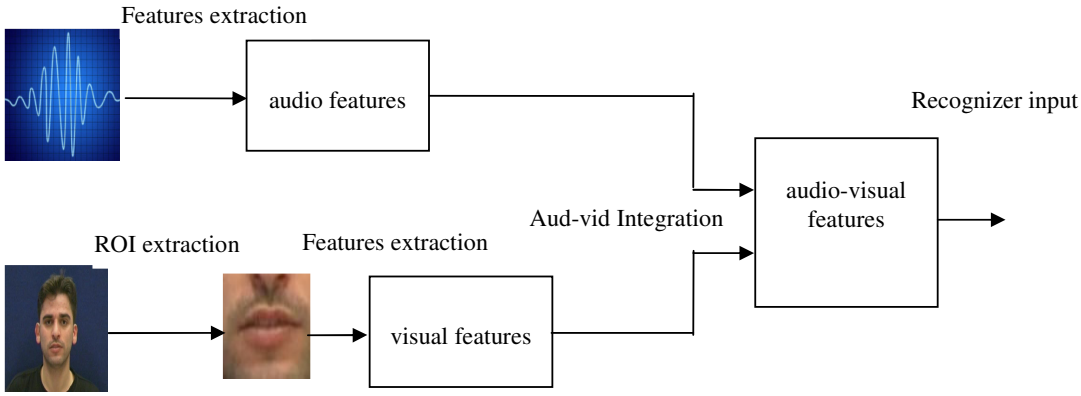
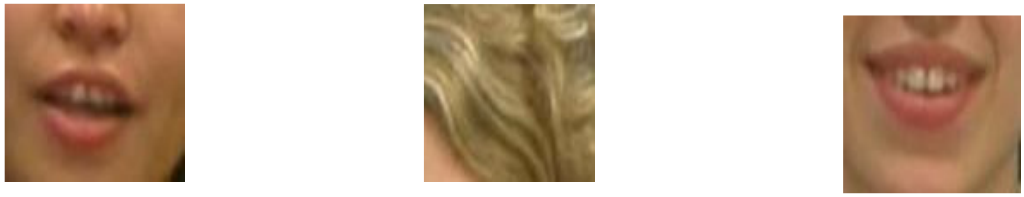



Figure1 Block diagram of an AVASR system

### *3.1* Face detection and mouth region of interest (ROI) extraction

The first step in design of visual front end is face detection and extraction of a visual region of interest (ROI). The choice of region of interest depends mainly on the choice of feature extraction approach used. For appearance-based techniques an approximate ROI bounding the mouth region are often sufficient while in geometric-based techniques a more accurate mouth contour is needed. For geometric-based approach, the ROI extraction and feature extraction stages often integrate into a single stage. In this work, local successive mean quantisation transform (SMQT) features [9] were used to locate the face region in the image. The lower half of the face region is assumed to contain the mouth region and a bounding box of 96x72 pixels at the centre of these coordinates is extracted to form the visual ROI. To reduce computational cost, coordinates for ROI were calculated for the first frame of the utterance and the same coordinates were used for ROI extraction in the remaining frames. This

approach worked well in majority of cases, but in a small number of cases the face region was either missed or the mouth region was not contained entirely inside the bounding box and so manual correction was applied in such cases, as shown in Figure 2.



(a) (b) (c)

Figure 2 Region of interest (ROI) extraction, (a) accurately extracted ROI, (b) missed ROI, (c) manually corrected ROI

Features are extracted from audio at a rate of 100 times per second while the original video stream is at 25 frames per second. The video stream is up-sampled to synchronise with the audio stream. Here linear interpolation is used to up-sample video to 100 frames per second.

## *3.2* Features extraction

The selection of suitable features plays a critical role in the performance of ASR systems. Features extraction is a dimensionality reduction process and ideally, the extracted features will retain all speech information present in original signal in a reasonably small number of dimensions. The most commonly used audio features are Mel-frequency cepstral coefficients (MFCC) where as visual features used in literature could be grouped into categories of low-level or appearance-based features, high-level or geometric-based features and hybrid features formed by combining above two types of features.

In this work Mel-frequency cepstral coefficients (MFCC) are used as audio features while appearance based features based on discrete cosine transform (DCT) are used as visual features. Cambridge University's Hidden Markov Model Toolkit (HTK) [10] is used to extract 13 MFCC coefficients along with its first and second derivatives. For visual stream a 108-dimension vector is obtained from the DCT transform of ROI and its dimensionality reduced to 30 dimensions using linear discriminant analysis (LDA). Delta and delta-delta coefficients are concatenated with static features to form a final feature vector of 39 and 90 dimensions for audio and visual streams respectively.

## *3.3* Audio-visual Integration

Three methods of audio-visual integration have been proposed in literature. They are early or feature integration, late or decision integration and hybrid integration. In early integration the two streams are combined at the feature level before training/recognition. In late integration the recognition is performed separately on two recognisers one for each modality and the results of the two individual recognisers are combined at later stage, while in hybrid integration approach the results of the two recognisers are combined somewhere in between these two extremes.

In this work Multi-Stream Hidden Markov Model (MS-HMM) is used which deals with the two streams independently and the results are combined at state level. Audio and visual features are passed onto the recognisers with pre-defined weights for each stream. The weight for the streams are varied from 0 to 1 in steps of 0.1 giving rise to audio-only, audio-visual (variable weights) and video-only recognisers.

## 4. RESULTS AND DISCUSSION

In these experiments, HMM toolkit (HTK) has been used for training and testing purposes. Acoustic only scores were used for recognition without any language information. Experiments were performed on clean speech and different audio noise level varying from SNR of 30dB to -10dB. As can be seen from Figure 3, the audio-visual recogniser with optimised weights outperforms all the other recognisers. Audio-visual ASR with constant weights performs well compared to audio-only in presence of noise but its recognition too false below the video-only recogniser due to equal audio weight. As visual modality is unaffected by the audio noise, larger weight for visual stream can improve the performance of ASR in these situations.

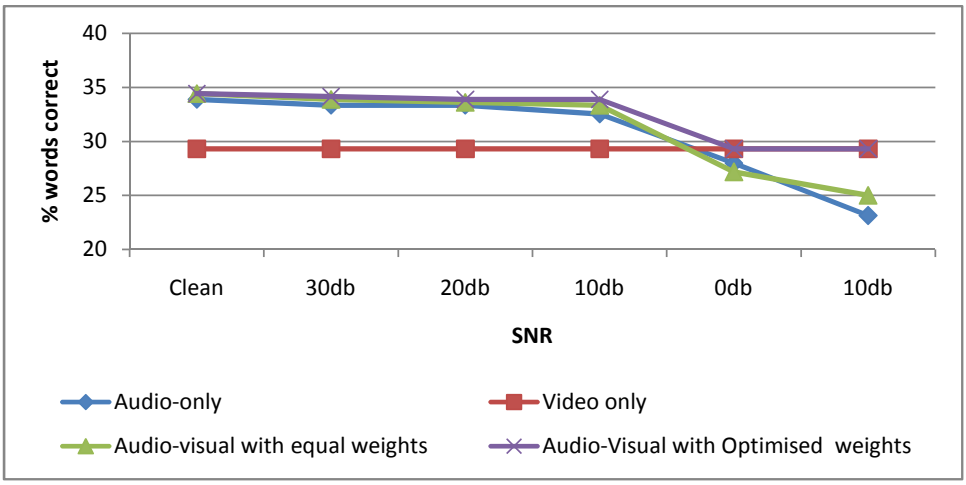



Figure 3 Recognition results for audio-only, video-only and, audio-visual ASR for equal and optimised weights

## 5. REFERENCES


[1] W.H. SUMBY, AND I. POLLACK, 'Visual Contribution to Speech Intelligibility in Noise', Journal of the Acoustical Society of America, Vol. 26, No. 2, pp. 212-215, 1954.

[2] E.D. PETAJAN, 'Automatic Lipreading to Enhance Speech Recognition', Proceedings of the IEEE Communication Society Global Telecommunications Conference, Atlanta, Georgia, 1984.

[3] J. S. LEE, AND C. H. PARK, "Adaptive decision fusion for audio-visual speech recognition", Speech Recognition, Technology and Applications, I-Tech, pp. 275-296, 2008.

[4] R. P. LIPPMANN, "Speech recognition by machines and humans", Speech Communication, vol. 22, no. 1, pp. 1-15, 1997.

[5] T. J. HAZEN, K. SAENKO, C. H. LA, AND J. GLASS, "A segment-based audio-visual speech recognizer: data collection, development, and initial experiments", Proceeding of ICMI, pp. 235-242, 2004.

[6] M. GURBAN, AND J. P. THIRAN, "Using entropy as a stream reliability estimate for audio-visual speech recognition" Proceedings of the 16th European Signal Processing Conference 2008.

[7] K. SAENKO, AND K. LIVESCU, "An asynchronous DBN for audio-visual speech recognition", IEEE Workshop on Spoken Language Technologies, pp. 154-157, 2006.

[8] C. SANDERSON, AND K. K. PALIWAL, "Polynomial features for robust face authentication", Proceedings of IEEE International Conference on Image Processing, vol. 3, pp. 997-1000, 2002.

[9] M. NILSSON, J. NORDBERG, AND I. CLAESSON, "Face detection using local SMQT features and split up snow classifier", IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 2, pp. II-589-II-592, 2007.

[10] S. YOUNG, D. KERSHAW, J. ODELL, D. OLLASON, V. VALTCHEV, AND P. WOODLAND, (1999) "The HTK Book", United Kingdom: Entropic Ltd.