AUDIO-VISUAL AUTOMATIC SPEECH RECOGNITION USING DYNAMIC VISUAL FEATURES

Nasir Ahmad[*], David Mulvaney[*], Sekharjit Datta[*] and Omar Farooq[#]
[*]Department of Electronic and Electrical Engineering, Loughborough University UK, LE11 3TU
[#]Department of Electronics Engineering, AMU Aligarh 202 002, India
n.ahmad@lboro.ac.uk, d.j.mulvaney@lboro.ac.uk, s.datta@lboro.ac.uk,
omarfarooq70@gmail.com

ABSTRACT

Human speech recognition is bi-modal in nature and the addition of visual information from the speaker's mouth region has been shown to improve the performance of automatic speech recognition (ASR) systems. The performance of audio-only ASRs deteriorates rapidly in the presence of even moderate noise, but can be improved by including visual information from the speaker's mouth region. The new approach taken in this paper is to incorporate dynamic information captured from the speaker's mouth occurring during successive frames of video obtained during uttered speech. Audio-only, visual-only and audio-visual recognisers were studied in the presence of noise and show that the audio-visual recogniser has more robust performance.

KEYWORDS: Audio-visual speech recognition, motion vectors

1. INTRODUCTION

Research in automatic speech recognition (ASR) has the main purpose of making human-computer interactions more natural and concise. Although successful ASR systems have been developed that are able to perform well under ideal conditions, there remains a substantial challenge in developing solutions that operate in practical situations where multiple sources or noise are present [1]. Under such conditions, the performance of ASR systems that use solely audio information degrade rapidly, whereas human speech recognition, with our ability to supplement audio with visual information, remains less severely affected. A number of recent publications have reported improvements in speech recognition performance by incorporating visual information from a speaker's face or mouth region [2].

To extract suitable visual information, research approaches described in the literature use either low-level appearance features obtained from a suitable transformation of images obtained from the speaker's mouth or face regions, or high-level features based on geometry, such as the length, width or roundness of the mouth. Although the immediate positions of articulators yields useful information about spoken words, these features fail to capture the dynamic information present in speech. For example, the position of the tongue when uttering /l/ or /d/ appears similar, but the phonemes can be perhaps better distinguished by analysing the tongue's motion.

The new approach described in this paper is to incorporate information obtained from dynamics in the mouth region of interest (ROI) that occur in successive frames of video obtained during uttered speech. The new visual features obtained in this work are combined

with audio features derived from Mel-frequency cepstral coefficients (MFCC) and its first and second derivatives. Audio-only, visual-only and audio-visual recognisers have been studied in the presence of noise.

2. BACKGROUND

The operations of an Audio-Visual ASR (AVASR) can be divided into the following three stages.

(a) Identifying, tracking and extracting the visual region of interest (ROI). Appearance-based techniques are often able to make use of an approximate ROI that needs to bound the actual mouth region, but in geometric-based techniques a more accurate mouth contour is needed. In the geometric-based approach, the ROI extraction and feature calculation stages often become amalgamated into a single stage.

(b) Determining suitable visual features. The types of visual features extracted fall broadly into three categories. In low-level or appearance-based techniques, the whole mouth or face region is considered as containing speech information. To reduce the dimensionality, a suitable transformation of the speaker's mouth region is taken followed by Principal Component Analysis (PCA) or Linear Discriminant Analysis (LDA). High-level or geometric-based techniques use geometric parameters of the mouth as a feature set. A third technique uses a combination of above two types of feature.

(c) ***Integration with the audio speech recognition.*** In the literature, three methods of integration have been proposed. The first is early or feature integration where audio and visual streams are combined at the feature level. The second is late integration where the recognition of the audio and visual streams are performed separately and the integration carried out in the decision stage, for example as shown in Fig. 1. A third approach is to perform partial integration in each of the early and late stages.

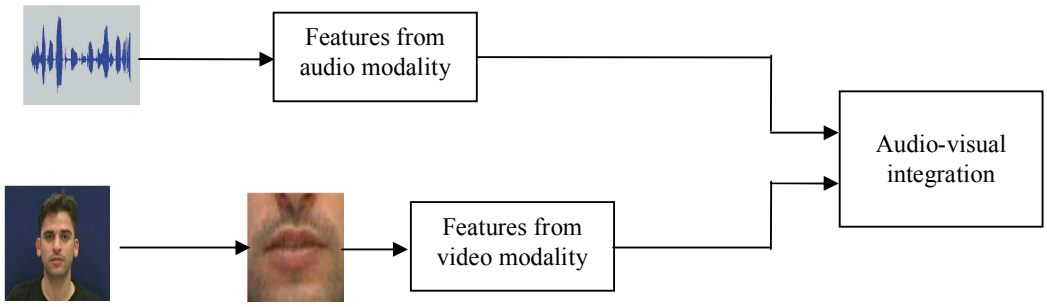

Fig.1 Block diagram of an AVASR system

3. AUDIO-VISUAL DATABASES

Few suitable audio-visual databases are available for speech recognition purposes, perhaps partly due to the large capacity requirements for video and the revealing of the speaker's identity. In addition, some databases contain information more appropriate for a specific approach; for example databases intended for geometric approaches require accurate localisation of lip edges and corners and marking is often added to the speakers' lips. In

contrast to audio-only ASR, no standard database is available for AVASR. There are only two AVASR databases in common usage, namely the audio-visual TIMIT (AVTIMIT) [3] and Vid-TIMIT [4] databases, both of which contain large vocabularies and are suitable for adaptation to other tasks such as phoneme and viseme recognition.

In our experiments, a subset of the Vid-TIMIT database consisting of 32 speakers (16 male and 16 female) is used. Eight different sentences are spoken by each speaker, containing 925 words in total from which 24 speakers with 216 sentences are used for training and the remaining 8 speakers with 40 sentences kept for testing purposes. Video in the database 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.

4. FACE DETECTION AND MOUTH EXTRACTION

Audio features are extracted at a rate of 100 times per second while the original video stream is 25 frames per second. To synchronise the audio and visual streams, the video is up-sampled to 100 frames per second using linear interpolation. Local successive mean quantisation transform (SMQT) features were used to locate the face region in the image [5]. The lower half of the face region is then assumed to contain the mouth region and a bounding box of 100x75 pixels at the centre of these coordinates is extracted to become the visual ROI. Due to the nature of the video streams and to reduce computational cost, this process is only applied in the first frame of the sequence and the same coordinates are used for ROI extraction in the remaining frames. This approach was successful in the vast majority of cases, but occasionally the face region was either not properly located or the mouth region not contained entirely inside the bounding box and so manual correction was applied in such cases, as shown in Figure 2.

(a) (b) (c)

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

5. FEATURE EXTRACTION

The selection of suitable features plays a critical role in the performance of speech recognition systems. Ideally, the features will retain all the relevant information needed from the original signals relating to speech in a vector of small dimensions. Clearly, an audio-visual speech recognition system requires that both audio and visual features are extracted.

5.1 Audio feature extraction

In this work, the standard Mel-frequency cepstral coefficients (MFCC) are used and

Cambridge University's Hidden Markov Model Toolkit (HTK) [6] is used to extract 13 MFCC coefficients along with its first and second derivatives.

5.2 Visual feature extraction

The new motion-based approach used here for visual feature extraction takes into account the dynamics of the mouth region during speech that are not captured by the appearance-based and geometic-based feature methods reported in literature. Motion vectors are calculated between successive frames using the block matching algorithm described in [7]. The ROI extracted in section 4 is resized to 88x72 pixels in order to allow an integral number of macro blocks of size 8x8 to be generated. As the required speech information is mainly present in vertical movements, the 99-dimension motion vector is obtained by reshaping only the vertical components of the obtained motion vectors. PCA is then applied to reduce the number of dimensions to 30 and this vector is used in conjunction with the 13 audio features to provide an early integration approach.

6. RESULTS AND CONCLUSION

In these experiments, HTK has been adopted for training and testing purposes and no use is made of dictionary information or a language model during the recognition process.

Three separate speech recognition systems were trained for audio-only, visual-only and audio-visual speech recognition. Experiments were performed with a range of audio noise levels. As can be seen from Table 1, the audio-only recogniser outperforms both visual-only and audio-visual recognisers when no noise is present, but its relative performance deteriorates as additional noise is introduced. As expected, the performance of the visual-only system is independent of audio noise and the audio-visual recognition system is more robust to noise than the audio-only method.

Table 1 Comparative performance of the speech recognition systems

signal to noise ratio	audio-only	visual-only	audio-visual
clean speech	34.95	26.34	27.69
30 db	34.95	26.34	27.69
20 db	34.68	26.34	27.15
10 db	34.14	26.34	27.15
0 db	23.12	26.34	26.88
-10 db	22.04	26.34	25.54

7. REFERENCES

[1] R. P. Lippmann, 'Speech recognition by machines and humans', *Speech Communication*, vol. 22, Issue. 1, pp. 1-15 (1997)

[2] J. S. Lee, and C. H. Park, 'Adaptive decision fusion for audio-visual speech recognition', *in Speech Recognition, Technology and Applications, I-Tech*, pp.275-296 (2008)

[3] T. J. Hazen, K. Saenko, C. H. La, and J. Glass, 'A segment-based audio-visual speech recognizer: data collection, development, and initial experiments', *in Proc ICMI*, pp. 235-242 (2004)

[4] C. Sanderson, K.K. Paliwal, 'Polynomial features for robust face authentication', *in Proc. IEEE Int. Conf. Image Processing*, vol.3, pp. 997-1000 (2002)

[5] M. Nilsson, J. Nordberg, and I. Claesson, 'Face detection using local smqt features and split up snow

classifier', *in IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 2, pp. II-589-II-592 (2007)
[6] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book*, United Kingdom: Entropic Ltd., (1999)
[7] A. Barjatya, 'Block Matching Algorithms for Motion. Estimation', *DIP 6620, Final Project Paper* (2004)