Audio-Visual Speech Recognition in the Presence of a Competing Speaker

Xu Shao and Jon Barker

Department of Computer Science, The University of Sheffield
{X.Shao,J.Barker}@dcs.shef.ac.uk

Abstract

This paper examines the problem of estimating stream weights for a multistream audio-visual speech recogniser in the context of a simultaneous speaker task. The task is challenging because signal-to-noise ratio (SNR) cannot be readily inferred from the acoustics alone. The method proposed employs artificial neural networks (ANNs) to estimate the SNR from HMM state-likelihoods. SNR is converted to stream weight using a mapping optimised on development data. The method produces an audio-visual recognition performance better than that of both the audio-only and the video-only baselines across a wide range of SNRs. The performance using SNR estimates based on audio state-likelihoods is compared to that obtained using both audio and visual likelihoods. Although the audio-visual SNR estimator outperforms the audio-only SNR estimator, the recognition performance benefit is small. Ideas for making fuller use of the visual information are discussed.

Index Terms: audio-visual speech recognition, multistream, stream weighting, SNR estimation, artificial neural networks

1. Introduction

Automatic speech recognition systems can achieve good performance in noise-free conditions but this performance degrades dramatically in the presence of noise. Speech recognition is particularly challenging when the background noise is another speaker. The difficulty of this condition arises for two reasons. First, speech is highly non-stationary. When the target speaker is masked by another speaker (the masker) the frame-based signal-to-noise ratio (local SNR) varies over a wide range, e.g. from over +60 dB to -60 dB. Even at high global SNRs local regions of the target speech may be *energetically* masked. On the contrary, at low global SNRs some frames may be dominated by the target speaker. Second, the noise will be statistically similar to the target speech, making it hard to distinguish between the two. Region of the signal that are dominated by the masker will have similar acoustics to those that are dominated by the target. This foreground/background confusion contributes to what is known in the perceptual literature as 'informational masking' [1] which leads to difficulty in estimating the local or global SNR for the utterance. We attempt to deal with both forms of masking using techniques based on the conventional multistream approach to audio-visual automatic speech recognition (AV-ASR).

As is well known, lip movements provide evidence of the phoneme being spoken and hence they are associated with acoustic signal. Although visual speech may be ambiguous (e.g. /b/ and /p/ appear identical), in noisy conditions the visual information may help to disambiguate acoustically confusable phoneme pairs (e.g. /s/ and /f/). Recognition performance can be made more robust by the integration of both the audio and visual information.

In the current work we use a state-synchronous multistream approach to AV-ASR. Here, audio and visual features are treated as synchronous streams. A hidden Markov model (HMM) is employed in which each state generates both audio and video observations drawn from different distributions (i.e. audio and video observations are modelled as being independent given the state). The system can be trained on clean audio-visual speech. To make it robust to acoustic noise, at recognition time, the audio and visual likelihoods are combined using weights based on a measure of their relative reliability. In the current work we attempt to estimate the stream weightings from HMM state-likelihood information using artificial neural networks (ANN). We compare the performance of ANNs trained on audio likelihoods and audio-visual likelihoods. The hypothesis is that the pattern of audio and visual likelihoods should distinguish between: local SNRs close to 0 where the acoustics match the models poorly; positive SNRs where the acoustics match the models and are correlated with the visual information; and negative SNRs where the (masker) acoustics may match the models, but will correlate poorly with the target's visual features. The audio-visual SNR estimator is expected to outperform the audio-only SNR estimator because the acoustic likelihoods alone are not sufficient for distinguishing between regions where the target matches the models (high SNR) and those where the masker matches the models (low SNR).

The remainder of this paper is arranged as follows. Section 2 describes the estimation of the audio and visual stream weights. Section 3 presents the results of AV-ASR experiments based on a small vocabulary simultaneous speaker task. Section 4 presents conclusions and discusses possible directions for future work.

2. Estimation of acoustic stream reliability

In state-synchronous multistream AV-ASR systems, an HMM is trained that generates observations for both audio and the visual streams. The integrated state emission score of the two-stream HMM, $P(o_t|c)$, based on speech unit class, c, at time, t, is modelled as,

$$P(o_t|c) = P(o_{a,t}|c)^{\lambda_{a,t}} \times P(o_{v,t}|c)^{\lambda_{v,t}} \quad (1)$$

where o_t is the concatenated observation of audio and visual features, $o_t = [o_{a,t} \quad o_{v,t}]$. The exponents $\lambda_{a,t}$ and $\lambda_{v,t}$, where $\lambda_{a,t}, \lambda_{v,t} \geq 0$ and $\lambda_{a,t} + \lambda_{v,t} = 1$, represent the weighting of the audio and visual components respectively.

The weighting parameter is related to the relative reliability of the audio and visual modalities, which in turn is dependent on the

This research was supported by grants from the University of Sheffield Research Fund and the EPSRC (GR/T04823/01).

SNR for whole utterance (global SNR). In previous studies, Glotin *et al.* [2] used voicing as a measurement of audio reliability. Garg *et al.* [3] employed the N-best log-likelihood as an SNR indicator to measure the modality reliability. Tamura *et al.* [4] estimated the stream weight from the normalised likelihoods. However, in the two speaker problem, all of these techniques will predict a high audio reliability at times when the mixture is dominated by the masker - times in which the audio reliability is actually low.

We too use likelihoods to estimate SNR, but we use likelihoods from both audio and visual HMMs in the estimation. The estimation is based on the fact that in regions of high local SNR, the acoustic features are likely to be good matches to a small number of HMM states, whereas in places where the local SNR is closer to 0 dB no single state will be a particularly good match and the likelihoods will be 'spread between states' - i.e. the pattern of log likelihoods across states will not have clear significant peaks. There is thus an association between the pattern of likelihoods and local SNR. However, in the speech-plus-speech case, a problem arises if only the acoustic information is considered. In regions of the signal where the masking talker dominates, the patterns of log-likelihood can be similar to the patterns seen in clean conditions – to the extent that the target and masker speaker fit equally well to the speech models, the pattern of likelihoods is symmetrical around 0 dB local SNR. This confusion can *potentially* be disambiguated using the visual likelihoods. We anticipate that, at positive local SNRs, the visual and audio likelihoods will be concentrated in corresponding HMM states (e.g. if the state representing an audio 'f' has a high likelihood then the state representing a visual 'f' should also have a high likelihood). Whereas, at negative local SNRs, the audio and visual likelihoods will generally be concentrated in different HMM states because the masker's speech is not correlated with the target speaker's lip movements.

Stream weight estimation proceeds as follows. First, two sets of independent word-level HMMs are trained using synchronised audio and visual features respectively. Both the audio and visual HMMs output a total of N log-likelihoods per frame, where N is the total number of states in the model sets (in our case N is 251). Using standard word model HMMs some of these states will have similar likelihood because the same phoneme (or viseme) may appear in different words, e.g. our task employs the alphabet words in which the phoneme */iy/* occurs frequently. This could make the difference between the distributions of likelihoods across states for the clean and noisy cases less easy to characterise. To reduce this problem likelihoods from similar states are averaged. To determine which state to average across a standard state-clustering technique is employed [5]. The appropriate degree of clustering is determined by that which gives good recognition performance on clean speech. This is done separately for the audio and the video based HMMs. In the experiments reported in Section 3 the likelihoods are averaged with 54 clusters for the audio HMMs and 18 for the video HMMs.

The relationship between the vectors of frame-based log-likelihoods and local SNR is then learnt using a multi-layer perceptron (MLP) with a single hidden layer. Target local-SNRs are computed using prior information of the unmixed speech signals. The mapping from log-likelihoods to local SNR is trained using conjugate gradient descent. The audio-based MLP has 54 input nodes (one for each averaged likelihood) and the AV-based MLP has 72. The number of hidden units is optimised by observing the errors between the calculated output and the target of a validation data set. The MLP topology that produces the minimum error is selected. Training proceeds using either the log-likelihoods of the audio stream, or the concatenated log-likelihoods of both the audio and visual streams.

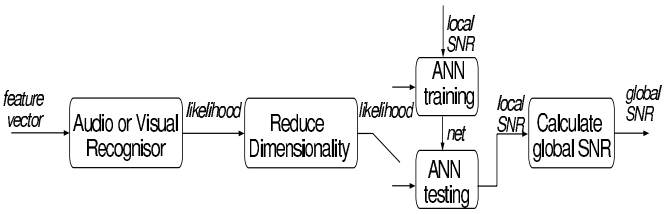

Figure 1: The flow diagram to measure the stream reliability using the artificial neural networks

The MLPs are trained to estimate the *local* SNRs given the log-likelihoods. In theory these local SNRs could be mapped on to a time-varying stream weight. However, in practise they are not likely to be sufficiently reliable for this to produce a good result. Instead, local-SNR estimates for an utterance are combined with a measure of frame-based energy to produce a single global SNR estimate, SNR_g

$$\text{SNR}_g = 10\log_{10}\left(\sum_{i=1}^{n}\frac{E_i}{1+B_i}\Big/\sum_{i=1}^{n}\frac{E_i\cdot B_i}{1+B_i}\right) \qquad (2)$$

where $B_i = 10^{\text{SNR}_{i/10}}$. The E_i and SNR_i represent the i^{th} frame energy and SNR respectively and *n* is the total number of frames in the utterance. Finally, the estimated global SNR is mapped onto a stream weight using an SNR to stream weight mapping that has been optimised using a set of development data.

3. Experimental results

3.1. Audio-visual speech corpus and feature extraction

Experiments have been performed using the audio-visual Grid corpus [6] which consists of high quality audio and video recording of utterances of the form indicated in Table 1 spoken by each of 34 speakers (sixteen female speakers and eighteen male speakers). An example sentence is "*bin red in c 3 again*". A total of 3500 utterances were randomly selected from 10 different male speakers (350 utterances from each speaker) to train a gender dependent HMM model. For each utterance, 13 MFCC features were extracted from the audio stream at a 100 Hz frame rate and these were supplemented by their dynamic components to form 26-dimensional audio feature vectors.

Table 1: *Structure of the sentences in the GRID corpus.*

VERB	COLOUR	PREP.	LETTER	DIGIT	ADVERB
bin	blue	at	a-z	1-9	again
lay	green	by	(no 'w')	and zero	now
place	red	on			please
set	white	with			soon

In the current work we wished to study the the problem of audio-visual feature integration in isolation from problems of visual feature extraction. So to this end, prior speaker information and semi-automatic processes were employed to ensure that the video features were of a consistent high quality. Visual feature extraction was performed using a technique similar to that of Patterson et al. [7]. For each speaker, 10 hand segmented video frames

were used to train separate 3-component Gaussian mixture models for the pixel RGB values in the lip region, and in the region surrounding the lips. Then in each frame of a video sequence, a Bayes' classification of the pixels in the mouth region was performed such that each pixel was labelled as either 'lip' or 'skin'. After some noise removal, the centre of gravity of the largest connected region labelled as lip was computed. A box with an area proportional to that of the estimated lip region was centred on the lips. The image within this box was downsampled to 32×32 pixels, and then projected into feature space using a 2-D discrete cosine transform (DCT) from which the 36-dimension low-order coefficients were extracted as visual features. These were supplemented with their dynamic features to produce a 72 dimensional visual feature vector. Linear interpolation was employed to upsample the visual stream from 25 fps to 100 fps to match the frame rate of the audio stream.

Multistream word-level HMMs were employed to model the 41 words in recognition task's vocabulary. These models contain between two and ten states per word determined using a rule of 2 states per phoneme. A single state short pause model brought the total number of states to 251. Each state is modelled using a 5-component Gaussian mixture model for both the audio and visual streams. Another 3100 utterances were also randomly selected from the same set of ten male speakers and are mixed with 3100 masking utterances selected from ten female speakers at a range of global SNRs. A forced-alignment was used to remove the initial and final silence before mixing, and the shorter utterance of each pair was zero-padded to the length of the longer one. These mixed utterances were then randomly divided into three sets; 2000 were used for the final performance test; 1000 were used to train the MLPs, and the remaining 100 utterances acted as the validation set to avoid overfitting.

3.2. *A priori* global SNR versus *a priori* local SNR

The mapping between stream weight and global SNR was optimised using an exhaustive search, with results shown in Fig 2.

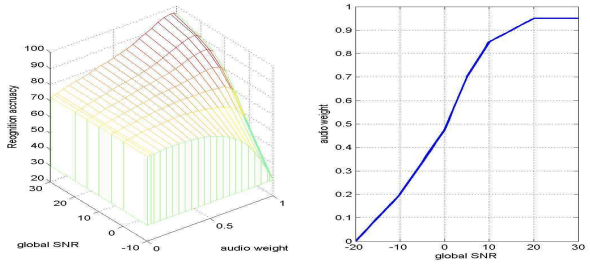

Figure 2: (a) the results of the exhaustive search (left) and (b) the optimised mapping between global SNR and stream weighting component (right)

The relationship between acoustic weighting components, global SNRs and speech recognition accuracy is illustrated in Figure 2-a. The mapping between acoustic weighting components and global SNRs (Figure 2-b) is extracted from Figure 2-a by looking for the weight which maximise speech recognition performance at each SNR. It can be observed that the weighting component of the audio stream is reduced as global SNR decreases, as expected.

In the first experiment this mapping was applied to the *a priori* global SNR to compute a single stream weight per utterance, or to the *a priori* local SNRs to compute a separate stream weight for each frame. The speech recognition performance achieved using these weights is compared in Table 2.

Table 2: Comparison of speech recognition performance from global SNR and from local SNR (%)

	clean	SNR20	SNR15	SNR10	SNR5	SNR0	SNR-10
global SNR	96.40	94.16	91.17	86.11	79.87	76.72	73.96
local SNR	-	92.57	89.66	85.81	81.74	77.82	73.32

The recognition performance obtained using global SNR is better than that obtained using local SNR at all SNR levels except SNR5 and SNR0. The recognition accuracy obtained using the global SNR is 1.07% higher than that obtained from the local SNR when averaged across all noise levels. The expectation is that a time-varying stream weight should outperform a static one. Failure to see a significant benefit is perhaps an indication that the optimised mapping from global SNR to stream weight is not optimal for local SNR. The results from *a priori* global SNR represent an upper limit against which to compare results using estimated SNR.

3.3. Audio versus audio-visual SNR estimates

The second experiment attempted to estimate global SNRs from HMM state-likelihoods using the methods described in Section 2. Two multi-layer perceptrons were trained. One was trained using the state-likelihoods of the audio-based HMM as input, while the other used concatenated likelihoods from both the audio-based and video-based HMMs. Both neural networks were trained using around 600,000 frames of data with 60,000 frames employed for validation. Results for the AV systems are shown in Table 3 compared against audio-only and video-only baselines.

Table 3: Comparison of speech recognition performance (%)

	clean	SNR20	SNR15	SNR10	SNR5	SNR0	SNR-10
MFCC	95.49	92.09	88.02	79.03	65.70	49.63	25.24
VIDT36D	73.09	73.09	73.09	73.09	73.09	73.09	73.09
ANN-MFC	95.51	92.94	89.87	85.27	79.33	76.17	73.62
ANN-AV	95.54	93.07	90.00	85.36	79.81	76.44	73.58

The first two rows in Table 3 show the baseline audio-only (MFCC) and video-only (VIDT36D) recognition performances respectively. These baselines are in agreement with previous studies with the visual stream producing significantly poorer results than the audio in low noise conditions. The video-only data is inherently ambiguous as many phonemes have similar visual appearance. However, the video-only result is obviously not affected by the level of the acoustic noise and remains at 73.09%.

The last two rows are the recognition accuracy from the AV system using the neural networks trained from either audio stream likelihood (ANN-MFC) or both audio and visual likelihood (ANN-AV). It can be seen that, as expected, recognition performance of the AV systems is better than that of both the audio-only and video-only systems across all SNRs. However, neither AV system quite matches the performance obtained using the *a priori* global SNR (Table 2). The results obtained using the AV SNR estimate (ANN-AV) are better than those obtained using the audio based SNR estimate (ANN-MFC). This is as expected as the AV system can learn

that acoustic confidence does not necessarily indicate high SNR - i.e. if acoustic likelihoods appear confident, but the highest scoring acoustic and visual states do not correspond, then we are probably seeing the target but hearing the masker i.e. very low SNR. However, somewhat surprisingly, the performance gained by using the AV weight estimation is small (compare ANN-MFC and ANN-AV in Table 3).

3.4. Analysis of the estimated SNRs

As the foreground and background are acoustically similar although the magnitude of the SNR may be well estimated, the sign of the SNR may be ambiguous and hard to estimate correctly. The final experiment investigated the extent of this effect. The two MLP neural networks were retrained using the same dataset and parameters as in the last experiment except the signs of the local SNR were removed. During the testing stage, the estimated SNR magnitude for each frame is combined with the *a priori* SNR sign obtained using knowledge of the unmixed signals.

Table 4: Comparison of speech recognition performance between different MLPs using *a priori* signs and estimated values (%)

	SNR20	SNR15	SNR10	SNR5	SNR0	SNR-10
AV-EVCS	93.95	91.02	86.17	79.87	76.67	73.88
MFC-EVCS	93.92	91.07	86.19	79.85	76.65	73.91

The first row in the Table 4 is the recognition performance from MLPs trained using both audio and visual likelihood and the second row is the recognition accuracy from MLPs trained using audio likelihood only. In both cases the estimated SNR magnitude was combined with the known prior SNR sign. These results are very close to the results from optimised global SNR and both of them are better than that obtained without using the prior sign. The small advantage obtained using the AV SNR estimates in Table 3 is no longer present - this is expected as the video information in the AV SNR estimator is being used to distinguish +ve SNR (target dominating) and -ve SNR (masker dominating), but in this experiment the sign of the SNR has been provided *a priori* so video information has a reduced role.

3.5. Discussion

The current work has been successful to the extent that, in the context of a simultaneous speaker task, it has shown that, i) a static stream weight can be used to combine audio and visual evidence to produce an AV-ASR system whose performance is better that of the audio or video alone, ii) it is possible to estimate the static stream weight directly from the audio-visual data. However, at a global SNR of 0 dB the AV speech recognition performance is only marginally superior to the performance of the visual-only system. A better integration would be afforded by a time-varying stream weight that allowed the recogniser to fully utilise regions of the target utterance where the SNR is temporarily high.

Providing a time varying stream weight requires accurately estimating local SNR. The estimates produced by the current system are too noisy to be used directly. The largest problem has been encountered in the estimation of the sign of the SNR. Although the audio-visual based estimator was designed to solve this problem it only confers a small advantage. There is inherent ambiguity. Many audio speech units have the same visual appearance, so a masking phoneme may differ from the target phoneme, but still be consistent with the target's lip movements. One solution would be to base the estimates on longer time windows. However, this solution is problematic as the local SNR varies rapidly as a function of time, so the likelihoods generated in neighbouring frames may not be indicative of the local SNR at the frame being considered. A better approach may be to use acoustic constraints to identify brief periods that appear to be dominated by a single speaker - these regions are typically voiced and can be located using pitch tracking techniques (see for example [8]). Then, for each 'single source' period, to estimate whether the acoustics are correlated with the video signal and should therefore be treated as reliable, or are uncorrelated and should therefore be ignored. These decisions could be made reliably for any segment of sufficient duration.

4. Conclusions

The paper has examined the problem of applying multistream audio-visual speech recognition techniques in a challenging simultaneous speaker environment. It has been shown that in this condition a static stream weight parameter based on a global SNR estimate can be used to successfully integrate audio and visual information. A technique for estimating the global SNR from audio and visual HMM state-likelihoods has been presented that produces results similar to those obtained using the *a priori* global SNR. However, it seems likely that better results could be achieved using a time-varying stream weight determined from an estimate of local (frame-based) SNR. Research is needed to develop techniques for reliably estimating local SNR from the the audio and video streams. Future work will look at estimating local SNR based on a pre-segmentation of the acoustic signal.

5. References

[1] N.I. Durlach, C.R. Mason, G. Kidd, Jr, Arbogast, H.S. T, L Colburn, and B.G. Shinn-Cunningham, "Note on informational masking," *JASA*, vol. 113, pp. 2984–2988, 2003.

[2] H. Glotin, D. Vergyri, C. Neti, G. Potamianos, and J. Luettin, "Weighting schemes for audio-visual fusion in speech recognition," *Proc. of ICASSP*, vol. 1, pp. 173–176, 2001.

[3] A. Garg, G. Potamianos, C. Neti, and Huang T.S., "Frame-dependent multi-stream reliability indicators for audio-visual speech recognition," *Proc. of ICASSP*, vol. 1, pp. 24–27, 2003.

[4] S. Tamura, K. Iwano, and S. Furui, "A stream-weight optimization method for multi-stream hmms based on likelihood value normalisation," *Proc. of ICASSP*, 2005.

[5] S. Young, J. Odell, V. Valtchev, and P. Woodland, "The htk book," *Cambridge*, 1995.

[6] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *submitted to JASA*, 2005.

[7] E.K. Patterson, S. Gurbuz, Z. Tufekci, and J.N. Gowdy, "Moving-talker, speaker-independent feature study and baseline results using the cuave multimodal speech corpus," *EURASIP Journal on Applied Signal Processing, vol. 11, pp. 1189-1201.*, 2002.

[8] N. Ma, P. Green, and A. Coy, "Exploiting dendritic autocorrelogram structure to identify spectro-temporal regions dominated by a single sound source.," in *Proc. Interspeech 2006*, submitted.