HAL
open science

Stream weight estimation for multistream audio-visual speech recognition in a multispeaker environment

Xu Shao, Jon Barker

To cite this version:

Xu Shao, Jon Barker. Stream weight estimation for multistream audio-visual speech recognition in a multispeaker environment. Speech Communication, 2008, 50 (4), pp.337. 10.1016/j.specom.2007.11.002 . hal-00499201

HAL Id: hal-00499201

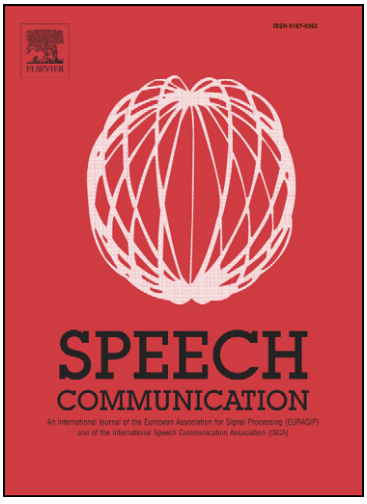
https://hal.science/hal-00499201v1

Submitted on 9 Jul 2010

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Accepted Manuscript



Stream weight estimation for multistream audio-visual speech recognition in a multispeaker environment

Xu Shao, Jon Barker

PII: S0167-6393(07)00186-0
DOI: 10.1016/j.specom.2007.11.002
Reference: SPECOM 1676

To appear in: *Speech Communication*

Received Date: 11 December 2006
Revised Date: 9 November 2007
Accepted Date: 12 November 2007

Please cite this article as: Shao, X., Barker, J., Stream weight estimation for multistream audio-visual speech recognition in a multispeaker environment, *Speech Communication* (2007), doi: 10.1016/j.specom.2007.11.002

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

ACCEPTED MANUSCRIPT

Stream weight estimation for multistream audio-visual speech recognition in a multispeaker environment [1]

Xu Shao *, Jon Barker

The University of Sheffield, Department of Computer Science, Regent Court, 211 Portobello Street , Sheffield, S1 4DP, UK.

Abstract

The paper considers the problem of audio-visual speech recognition in a simultaneous (target/masker) speaker environment. The paper follows a conventional multistream approach and examines the specific problem of estimating reliable time-varying audio and visual stream weights. The task is challenging because, in the two speaker condition, signal-to-noise ratio (SNR) – and hence audio stream weight – cannot always be reliably inferred from the acoustics alone. Similarity between the target and masker sound sources can cause the foreground and background to be confused. The paper presents a novel solution that combines both audio and visual information to estimate acoustic SNR. The method employs artificial neural networks to estimate the SNR from hidden Markov model (HMM) state-likelihoods calculated using separate audio and visual streams. SNR estimates are then mapped to either constant utterance-level (global) stream weights or time-varying frame-based (local) stream weights.

The system has been evaluated using either gender dependent models that are specific to the target speaker, or gender independent models that discriminate poorly

ACCEPTED MANUSCRIPT

between target and masker. When using known SNR, the time-varying stream weight system outperforms the constant stream weight systems at all SNRs tested. It is thought that the time-vary weight allows the automatic speech recognition system to take advantage of regions where local SNRs are temporally high despite the global SNR being low. When using estimated SNR the time-varying system outperformed the constant stream weight system at SNRs of 0 dB and above. Systems using stream weights estimated from both audio and video information performed better than those using stream weights estimated from the audio stream alone, particularly in the gender independent case. However, when mixtures are at a global SNR below 0 dB, stream weights are not sufficiently well estimated to produce good performance. Methods for improving the SNR estimation are discussed. The paper also relates the use of visual information in the current system to its role in recent simultaneous speaker intelligibility studies, where, as well as providing phonetic content, it triggers 'informational masking release', helping the listener to attend selectively to the target speech stream.

Key words: audio-visual speech recognition, multistream, multispeaker, likelihood, artificial neural networks

1 Introduction

Speech communication in its most natural face-to-face form is an audio-visual experience. Participants in a conversation both *hear* what is being said and *see* the corresponding movements of the speaker's face. The human speech percep-

* Corresponding author.
Email address: {x.shao,j.barker}@shef.ac.uk.

[1] This project was supported by grants from the UK Engineering and Physical Research Council (GR/T04823/01).

ACCEPTED MANUSCRIPT

tion system exploits this multimodality, integrating both the visual and audio streams of information to form a robust coherent percept. The central role of visual information has been described in recent accounts of speech perception [1,2]. The visual speech information is particularly valuable when there are competing sound sources in the environment. When this is the case, the audio signal becomes unreliable, and observation of the speaker's face greatly improves speech intelligibility. Sumby and Pollack [3] demonstrated that the visual speech signal could confer an increase in intelligibility equivalent to that produced by reducing the noise level by about 16 dB. More recent studies have demonstrated similarly dramatic benefits, particularly in cases where speech is masked by a competing speaker [4–6].

Demonstrations of the robustness of audio-visual speech have inspired much recent research in audio-visual automatic speech recognition (AVASR). There are two major research questions in this field. First, how best to represent the visual speech information [7–9]. Second, how to combine the audio and visual information so as to optimise recognition performance [10–12]. Although these questions are no doubt subtly connected they are usually addressed in isolation. It is the latter question, that of audio-visual integration, that forms the focus of the work presented here.

Although visual speech is inherently ambiguous with many words having an identical appearance (e.g. 'pop', 'bop', 'bob'), its value can be clearly seen by noting its ability to disambiguate acoustically confusable word pairs. For example, consider the words 'met' and 'net'. Although they are acoustically similar, being distinguished by subtle differences in the nasal consonants $/m/$ and $/n/$, they are visually distinct, with the lips closing for $/m/$ but not for $/n/$. In situations where the acoustic differences are masked by additive

ACCEPTED MANUSCRIPT

noise, the visual features may be all that is available to distinguish the two words. Although it is clear that the visual stream carries valuable information, making optimal use of this information has proved very difficult in practise.

AVASR systems typically incorporate a parameter that controls the relative influence of the audio and visual streams – in the common multistream formalism, that is used in the current study, this parameter is known as the *stream weight*. Much recent research has been devoted to developing schemes for estimating suitable values for the stream weight. A standard approach is to base the stream weight on an estimate of the signal to noise ratio (SNR) [12–14]. SNR-based stream weights have proven to be very successful in a number of scenarios, particularly in cases where the background noise is approximately stationary. These systems typically employ a fixed stream weight for the duration of an utterance. In situations where the background noise is highly *non-stationary* a time-varying stream weight based on local SNR estimates, or local audio reliability measures, is more appropriate (e.g. [13,15]). However, estimating local SNR is itself a challenging problem.

In the current work we search for solutions to the stream weighting problem in a particularly challenging condition: the case where the background noise source is itself a speech signal (which we will refer to as the '*masking speaker*'). This situation is of particular interest because it is one that occurs frequently in everyday life, and it is a situation in which human performance has been closely studied. Listening tests have shown that speech intelligibility is remarkably robust to the effects of background speech. This is particularly true if the target speaker is sufficiently 'unlike' the masker speaker that it can be selectively attended to with minimal distraction [16]. However, background speech causes particular problems for ASR. The difficulty here is twofold. First, as

ACCEPTED MANUSCRIPT

already mentioned, the noise signal is highly non-stationary. This means that the SNR is changing from instant to instant, and the reliability of the audio signal is rapidly varying. Second, the noise is not easily distinguished from the speech signal. Acoustically, a region of the signal dominated by the masking speaker (low SNR), may be little different to a region dominated by the target speaker (high SNR). So the non-stationarity demands a time varying stream weight, but the weight is hard to estimate reliably from local properties of the acoustic signal.

We attempt to solve the weight estimation problem by involving both the audio and video streams in the local SNR estimation. In outline, we employ a standard state-synchronous multistream system [17] in which the underlying generative model for the audio-visual speech is a hidden Markov model (HMM) in which each state generates both audio and visual observation drawn from different distributions (i.e. audio and video observations are modelled as being independent given the state). The system can be trained on noise-free audio-visual speech. To make it robust to acoustic noise the audio and visual likelihood are combined using exponential weights based on a measure of their relative reliability. In the current work the stream reliability is estimated from HMM state-likelihood information using artificial neural networks (ANN)[18]. The performance of ANNs trained either on audio-likelihoods, or on combined audio- and visual-likelihoods is compared. The hypothesis is that the pattern of audio and visual likelihoods should distinguish between, i) local SNRs close to 0 where the acoustics match the models poorly; ii) positive SNRs where the acoustics match the models well and are correlated with the visual information; and iii) negative SNRs where the (masker) acoustics may match the models well, but will be poorly correlated with the target's visual features.

ACCEPTED MANUSCRIPT

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 cases ii) and iii), i.e. regions of high SNR where the *target* matches the models and regions of low SNR where the *masker* matches the models.

The paper compares the role of video information in stream weight estimation in two different scenarios. In the first, male target utterances are mixed with female masker utterances, and recognition is performed using models trained on male speech. In this task the acoustic models are specific to the target. In the second scenario, male or female target utterances are mixed with either male or female maskers (but not from the same speaker), and recognition is performed using models that have been trained on a mixture of male and female speech. In this case the acoustic models are not specific to the target, and could equally well match the masking utterance.

The remainder of this paper is arranged as follows. Section 2 reviews the multistream approach to AVASR that we use to integrate the audio and visual feature streams. Section 3 details the AVASR system employed in this work, and in particular, the proposed method for estimating a time-varying stream weight. Section 4 details experiments that evaluate the behaviour of the system when using both same-gender and mixed-gender utterance pairs. Finally, Section 5 presents conclusions and discusses possible directions for future work.

ACCEPTED MANUSCRIPT

2 Background

2.1 Integration of audio and visual features

AVASR systems can be broadly classified according to the manner in which they combine the incoming audio and visual information streams – see Lucey et al. for a recent review [10]. Most systems can be described as performing either feature fusion or decision fusion. In feature fusion systems – also termed 'early integration' (EI) – the audio and visual feature vectors are combined, typically by simple concatenation, and the classifier learns the statistics of the joint audio-visual observation. By contrast, in decision fusion systems – also termed 'late integration' (LI) – separate classifiers are constructed for the audio and visual features, and it is the classifier outputs, rather than the feature vectors themselves, that are combined. A third class of techniques that combine the audio and visual signal before the feature extraction stage has largely been abandoned.

Most recent systems, including the work presented here, use some form of LI (e.g. [19,20,17]). Although EI allows the modelling of the complete audio-visual observation (and hence can model detailed correlation between features of the audio and video observations) it suffers to the extent that corruption of either the audio or the visual data stream can lead to incorrect decisions being made. In contrast, late integration systems can easily be made robust to known corruption of either stream by simply weighting the audio and visual classifier decisions during combination.

The design of LI systems is very flexible and many variations on the theme

ACCEPTED MANUSCRIPT

exist, but they can be roughly sub-classified according to the the lexical level at which decision fusion occurs. At the lowest level decisions are fused on a frame by frame basis. If the streams are modelled using HMMs then this equates to combining the likelihoods of corresponding audio and visual HMM model states – often termed 'state-synchronous decision fusion'. Combining decisions at higher lexical levels – such as the phoneme or word level – is usually achieved by using a parallel HMM design. Within each unit (phoneme or word) the model progresses independently through the states of the separate audio and visual HMMs under the constraint that the boundaries between units occur synchronously in the audio and visual domains. This technique can model some of the asynchrony within the modelling units that is observed between the audio and visual speech streams. For example, visual evidence for the onset of a phoneme often precedes the audio evidence of the same event [19,21]. This extra modelling power can produce small performance improvements when recognising spontaneous speech. In the current work we are employing a small vocabulary 'read speech' task and precise acoustic modelling of audio visual asynchrony is considered of secondary importance to the design of techniques for reliably estimating stream weights. Accordingly we employ the simpler state-synchronous decision fusion technique.

In the state-synchronous decision fusion AVASR systems, the HMM can be considered to be a generative model which produces observations for both the audio and the visual streams that are independent given the state. So the likelihood of state q given the observed audio and visual features, $o_{a,t}$ and $o_{v,t}$ respectively, is computed as,

$$P(o_{a,t}, o_{v,t}|q) = P(o_{a,t}|q) \times P(o_{v,t}|q) \tag{1}$$

ACCEPTED MANUSCRIPT

In order to make the system robust to acoustic noise, at recognition time the state likelihoods are replaced with a score based on a weighted combination of the audio and visual likelihood components,

$$S(o_{a,t}, o_{v,t}) = P(o_{a,t}|q)^{\lambda_{a,t}} \times P(o_{v,t}|q)^{\lambda_{v,t}} \tag{2}$$

where the exponents $\lambda_{a,t}$ and $\lambda_{v,t}$ are the audio and visual stream weights. Typically, they are constrained such that $\lambda_{a,t}, \lambda_{v,t} \geq 0$ and $\lambda_{a,t} + \lambda_{v,t} = 1$. These scores are usually computed in the log domain,

$$\log S(o_{a,t}, o_{v,t}) = \lambda_t \log P(o_{a,t}|q) + (1 - \lambda_t) \log P(o_{v,t}|q) \tag{3}$$

where λ lies in the range $[0, 1]$. Values of λ close to 0 give emphasis to the video stream and are used when the audio stream is believed to be unreliable – and values close to 1 give emphasis to the audio stream.

2.2 Stream weight estimation

The stream weighting parameter is related to the relative reliability of the audio and visual modalities, which in turn is dependent on the SNR. If the acoustic signal has high SNR at time t, the audio stream is reliable and a high value of λ can be used. In contrast, during regions of low SNR the stream weight parameter should be reduced so that the visual information is emphasised. In previous studies various audio-based stream weight estimation strategies have been employed. For example, Glotin et al. [15] used *voicing* as a measurement of audio reliability. Garg et al. [20] employed the N-best log-likelihood as an SNR indicator to measure the modality reliability. Tamura, Iwano and Furui [22] estimated the stream weight from the normalised like-

ACCEPTED MANUSCRIPT

lihoods. Gurbuz et al. [23] chose the stream confidence value from a lookup table according to the noise type and SNR. These earlier techniques have all tried to estimate stream weight from the acoustic signal. However, in the two speaker problem, the acoustic properties of regions dominated by the masking speech source may be similar to those of regions dominated by the target. This ambiguity makes it difficult to achieve accurate stream weight estimates using purely acoustic measurements. As a solution to this problem, the current study proposes an audio-visual stream weighting method (see Section 3.3.2).

Another design choice is the time interval over which to estimate the stream weight. Many previous systems have kept the stream weight fixed over the duration of an utterance [23,24]. In such systems the stream weight is being set according to some average value of the SNR. Even when the additive noise is stationary the SNR computed over a short time window will vary widely. For example, during voiced regions the speech signal may dominate the noise making a high value of λ appropriate, whereas during an unvoiced fricative the noise may mask the speech audio meaning that a low value of λ should be used to give temporary emphasis to the video signal. In the speech-plus-speech case the non-stationarity of the masker makes these effects even larger. In the data we employ, the local (frame-based) SNR can vary by as much as 30 dB above and below the global SNR. In the current work we attempt to capture this variation by allowing the stream weight to vary from frame to frame.

Dynamic stream weighting techniques have previously been proposed by Meier, Hurst and Puchnowski [13] and evaluated with non-speech maskers at SNRs down to 8 dB, and subsequently evaluated by Glotin et al. on a large vocabulary clean-speech task. However, the stream weighting schemes proposed by Meier et al. are not appropriate for the two speaker problem. Their SNR-based

ACCEPTED MANUSCRIPT

weight estimation technique employs noise spectrum estimation techniques due to Hirsch [25]. These do not work reliably in speech plus speech situations. They also proposed a stream weight based on the entropy of posterior phoneme and viseme probabilities. This technique is closer in spirit to our system, but crucially such a technique would fail in the two speaker situation as it relies on the phoneme entropy being high when the SNR is low, whereas in fact the opposite happens: the more the SNR is reduced, the more the masker dominates the target, and the more the acoustics become like those of a clean speech source. In our system we attempt to overcome these problems by basing the stream weight estimate on the complete pattern of audio and visual speech-unit probabilities, using an ANN to learn the mapping. By so doing, information about the *correlation* between the audio and visual data can be learnt – if the audio appears 'clean' but does not 'fit' with the video then the target is being dominated by the masker and the audio stream should have a low weight.

3 System Description

An overview of the AVASR system is shown in Figure 1 illustrating the system's three components: feature extraction, audio-visual HMM training and audio-visual HMM testing.

In brief, audio and visual feature vectors are extracted from the acoustic and visual data respectively and concatenated at a 100 Hz frame rate to form audio-visual feature vectors (Section 3.1). The multistream HMMs are then trained (Section 3.2). The audio and visual HMMs are first trained separately using clean audio and visual features. After independent training the

ACCEPTED MANUSCRIPT

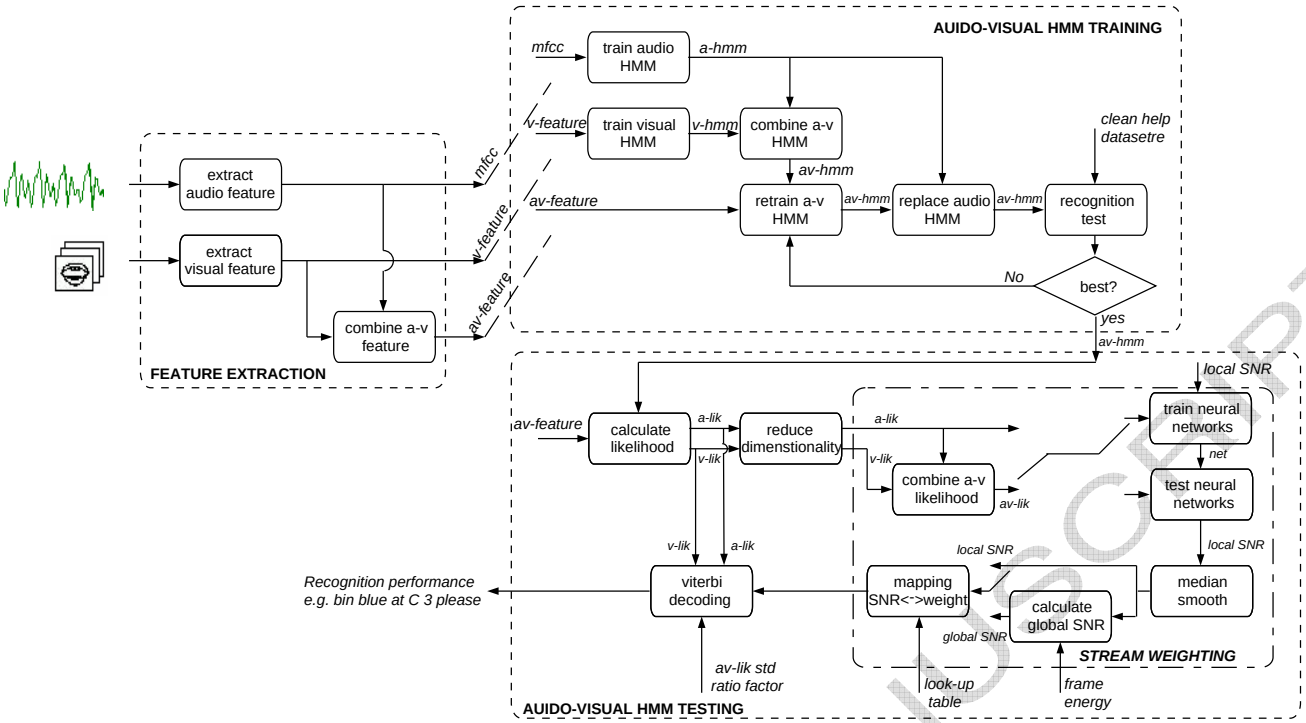


Fig. 1. The block diagram for the audio-visual speech recognition system.

state alignments implied by the audio and video HMMs are not necessarily consistent. A joint AV training step is employed to make the models compatible. The optimised AV HMMs are then passed to the testing stage (Section 3.3). First, the unweighted audio and video state likelihoods, $P(o_a|q)$ and $P(o_v|q)$, are computed. An ANN is used to learn how to map these likelihoods onto frame-based SNR estimates. Then, using a hand optimised mapping, the frame-based SNR estimates are either i) mapped directly onto a time-varying stream weight, or ii) first converted into an utterance level SNR (global SNR) estimate and then mapped onto a stationary stream weight. The audio and visual state likelihoods are then weighted and combined to produce scores for each HMM state. The recogniser output is obtained using a standard Viterbi decoder to find the best scoring state sequence. The details of these procedures are described in the following sections.

ACCEPTED MANUSCRIPT

3.1 Feature extraction

The acoustics are represented using standard MFCC features [26]. The 25 KHz acoustic signal is passed through a high pass preemphasis filter ($1 - 0.97z^{-1}$). 13 MFCC coefficients are computed using 25 ms Hamming windows at a 10 ms interval. The MFCC coefficients are supplemented by their temporal differences computed using linear regression over a 3 frame window, to produce a 26-dimensional feature vector.

The video features are based on a 2D-DCT representation of a rectangular region encompassing the speakers lips extracted using a technique similar to that of Patterson et al. [7]. For each speaker, six video frames are randomly selected. For each frame the outer lip contour is hand labelled. These hand segmented images are then used to train separate three-component Gaussian mixture models (GMM) for the distribution of pixel RGB values in i) the lip region, and ii) the region surrounding the lips. Then in each frame of a video sequence, a Bayes's classification of the pixels in a search region around the estimated centre of the mouth is performed such that each pixel is labelled as either 'lip' or 'skin' (i.e. a binary image). Opening and closing morphological operations are then applied to remove 'salt and pepper' noise from the binary image. The largest connected region bearing the 'lip' label is identified. Then the centre of gravity of this region is taken as the new position of the mouth centre. In this way the centre of the mouth can be tracked from frame to frame (this simple tracking system is similar in principle to the CamShift algorithm [27]). The region of interest is then taken as a rectangular box positioned at the mouth centre with an area proportional to that of the area of the estimated lip region. This procedure was applied to all 34 speakers in the database. The

ACCEPTED MANUSCRIPT

results were checked to ensure that the box reliably tracked the lips throughout each utterance.

For each video frame the image in the box surrounding the lips is downsampled to 32×32 pixels, and then projected into feature space using a 2-D DCT from which the 36 (6 by 6) low-order coefficients are extracted as visual features. These are supplemented with their dynamic features to produce a 72-dimensional visual feature vector. Linear interpolation is then employed to upsample the visual stream from 25 frames per second to the rate of 100 frames per second employed by the audio stream.

Note, in this work we wish to examine the issue of audio-visual integration assuming the presence of high quality video features without regard to how these features have been produced. The reliability of the visual features is being ensured by training speaker specific colour models that are also specific to the lighting conditions of the particular corpus recording session. Furthermore, results of the feature extraction are being monitored and poorly represented speakers are being rejected. By using artificial but high quality video features we hope to be able to focus the study on the problem of stream integration, without having to contend with the problems of varying reliability in the video features. Whether such reliable video features can be produced in a real application is a separate research question.

3.2 Audio and video HMM training

The construction of the multistream AV HMM commences by first training independent audio and video word-level HMMs using clean audio and visual

ACCEPTED MANUSCRIPT

features. These two models are then used to initialise the parameters of the state-synchronous multistream AVASR systems described in Section 2.1. The Gaussian mixture models (GMMs) describing the observation distributions for the states of the audio and visual stream are taken directly from the corresponding states in the independently trained models. The transition matrix for the multistream HMM is initialised to be that of the transition matrix of the audio HMM. Unfortunately, the portions of the signal that are presented by corresponding HMM-states of the independently trained audio and visual HMMs are not necessarily the same. Generally there is an audio-visual asynchrony with the onset of lip movement often preceding the acoustic evidence for the phoneme [19,21]. This lack of synchrony leads to poor performance as the AV HMM model has been made by combining corresponding audio and visual HMM states under that assumption of state synchrony. The parameters of the AV HMM have to be retrained so that the audio and visual components are compatible. As the clean speech recognition performance suggests that the audio parameters are far more informative than their video counterparts, the retraining stage is performed in such a way that the audio parameters and the transition matrices (which were adopted from the audio HMM) are held constant and only the GMMs of the visual stream are adapted.

The retraining step (shown in the block diagram, Figure 1) proceeds as follows: The full set of AV-HMM parameters are retrained using the Baum-Welch algorithm [28,29] and the clean data training set. During training the stream weight is set to 0.5. After each Baum-Welch iteration, the emission distribution for the audio component, and the state transition probability matrices, are reset to the value they had before reestimation. This forces the AV model to adopt the segmentation that was inferred by the audio-only model. If the audio

ACCEPTED MANUSCRIPT

parameters are not held constant in this way it was found that the training unacceptably reduces the performance of the system in clean conditions where the video stream weight is close to 0. It is possible that the clamping of the audio parameters would not have been necessary if a higher audio stream weight was used during training. However, the current procedure produces a model that performs well in both the video-only and audio-only condition, and seems to be appropriate for the small vocabulary task employed in which the video parameters are essentially redundant in the clean speech condition.

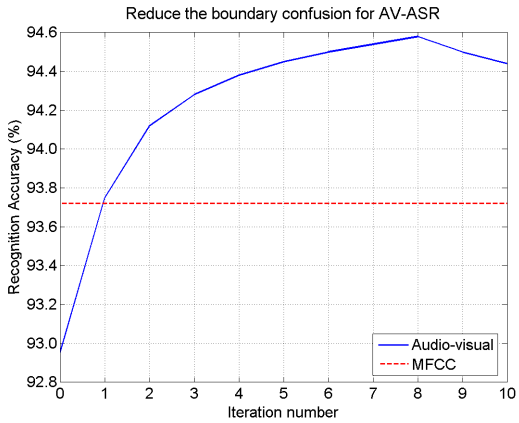


Fig. 2. The development of the audio-visual speech recognition performance measured on the clean cross-validation set during joint AV HMM reestimation. Accuracy is plotted after each iteration of parameter reestimation.

After each training iteration, performance is tested on a cross-validation data set and the procedure is repeated until the performance reaches a maximum. Figure 2 shows the development of the recognition performance during the joint training of the audio-visual HMMs. Performance for the clean speech cross-validation set is plotted against the number of reestimation iterations. When the audio and visual models are initially combined, the performance is not as good as that obtained using the audio model alone – presumably because of the mismatch in the audio and visual models described above. As

ACCEPTED MANUSCRIPT

the retraining iterations are repeated the performance of the system improves. The state-alignment implied by the parameters of the visual HMMs becomes consistent with that of the audio HMMs. At a particular point – 8 iterations, in this case – the performance reaches a peak after which performance starts to reduce. This pattern is consistent with the effects of over-fitting. The model trained with 8 iterations is the one that is selected to be applied to the test set.

Interestingly, it was also noted that, in our experiments, the video-only performance gained using the retrained visual stream HMMs was slightly better than that obtained using the visual stream HMMs before joint AV training. This can possibly be understood by analogy to supervised training – the audio parameters, which are generally more powerful than the visual parameters, act like labels to the unknown state indicated by the visual parameters. The fact that there is scope for the audio parameters to help in this way may indicate that the initial video-only training was not robust and suffered from poor initialisation.

3.3 Audio visual HMM testing

The recognition process involves three stages, i) calculation of state log-likelihoods for both the audio and visual components of the HMM, i.e. $P(o_{a,t}|q)$ and $P(o_{v,t}|q)$ in Equation 2, ii) estimation of the stream weights, i.e. λ_t and $1-\lambda_t$, by first estimating either local or global SNR as a function of the state log-likelihoods, iii) Viterbi-decoding the HMM state scores, $S(o_{a,t}, o_{v,t})$, computed from the stream weights and the log-likelihoods according to Equation 2. To analyse the performance of the system, two pairs of conditions (audio-

ACCEPTED MANUSCRIPT

likelihood *vs.* audio-visual-likelihood and local SNR *vs.* global SNR) are combined to give four possible sets of results to compare.

3.3.1 Audio and video state-likelihood computation

At recognition time, for each frame of observed audio and video features the log-likelihood of each HMM state is computed. Hence for an N-state HMM, N audio-based likelihoods and N video-based likelihoods are computed and stored as a pair of N-dimensional likelihood vectors. For the small-vocabulary Grid task [30] used for the current evaluations, word-level HMMs are employed with a total of 251 states (see Section 4.1 for details). Some of the states have very similar emission distributions. For example, considering the audio stream, the states corresponding to the phoneme /*iy*/ in the English letters, *E, D*, and the digit, *three* will have similar distributions. Likewise, in the video stream, the states corresponding to the plosive visemes at the starts of the Grid corpus words *bin* and *place* will also be similar (see Section 4.1 for the full set of words used in the Grid task). These states will have very similar likelihoods regardless of the observation. Hence, it is possible to reduce the dimensionality of the likelihood vector without loss of information by representing the likelihood for such groups of states with a single value. This is implemented by first employing a state-clustering technique [29] to identify similar states. The clustering is performed separately for the audio and video based HMMs and the degree of clustering is determined by that which gives the best recognition performance when testing on clean speech. The dimensionality of the likelihood vector is then reduced by replacing the elements corresponding to the members of a cluster with a single value computed by averaging the likelihoods of those states. The reduced likelihood vectors are

ACCEPTED MANUSCRIPT

then used as the basis for the stream weight estimation described in the next section.

3.3.2 ANN-based stream weight estimation

In this work we wish to see whether the use of video information may solve a problem specific to the speech-plus-speech case: in regions of the signal where the target talker dominates, the patterns of log-likelihood can be similar to the patterns seen in regions where the masker dominates – in general there is a symmetry around 0 dB local SNR. This symmetry will exist to the extent that the target and masker speaker fit equally well to the speech models. For example, this problem should be particularly apparent when using speaker independent models, but less apparent if using gender-dependent models and mixed gender utterance pairs. Potentially, this confusion can be disambiguated using the visual likelihoods. 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). 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.

Based on these considerations, our system operates in two stages: first, SNR is estimated by considering the match between the data and the clean speech models, and second, SNR is mapped to stream weight using a hand optimised mapping (see dash-dot box in Figure 1.) The difficulty arises in producing reliable estimates of either local or global SNR. Following previous work, the SNR

ACCEPTED MANUSCRIPT

estimates are based on the HMM-state likelihoods. However, unlike previous approaches which have based the estimates on specific features of the pattern of likelihoods, such as entropy [31,32] or dispersion [32,33], we attempt the more general approach of trying to learn the SNR directly from the complete likelihood data using artificial neural networks (ANN). Furthermore, in our experiments, as well as employing mappings based on the audio-only stream likelihoods, we also consider mappings that are based on the state likelihoods of both the audio and visual stream.

Multi-layer perceptrons (MLPs) with three layers were employed with an input unit for each element of the likelihood vector and a single output unit representing local SNR. The network employs a single hidden layer. Units in the hidden layer and output layer have sigmoid and linear activation functions respectively. The network is trained using conjugate gradient descent [34]. The number of hidden units is optimised by observing the errors between the MLP output and the target SNR over a validation data set. The MLP topology that produces the minimum error is selected. The network is trained using either the log-likelihoods of the audio stream, or the concatenated log-likelihoods of both the audio and visual streams.

The sequence of MLP outputs is median smoothed to remove outlying SNR estimates. The time-varying local SNR can either be used directly to produce a time-varying stream weight estimate, or the local SNR estimates are first converted into a single global SNR estimate and are then used to produce a single utterance level stream weight. The global SNR is calculated according to,

$$SNR_g = 10\log_{10}\left(\sum_{i=1}^{I}\frac{E_i}{1+B_i}/\sum_{i=1}^{I}\frac{E_i\cdot B_i}{1+B_i}\right) \tag{4}$$

ACCEPTED MANUSCRIPT

where $B_i = 10^{\frac{SNR_i}{10}}$ and E_i and SNR_i represent the ith frame energy and SNR respectively. SNR_g denotes the global SNR and I is the total number of frames in the utterance.

3.3.3 Optimising the lookup table and likelihood scaling factor

It has been documented in previous studies (e.g. [35]) that normalising the log-likelihood of both streams to have equal standard deviation can improve the performance of multistream based systems. This normalisation can be expressed as,

$$b'_{vi} = \frac{Std_a}{Std_v} \cdot b_{vi} \tag{5}$$

where b_{vi} and b'_{vi} are log-likelihoods of the ith frame of the visual stream, $\log P(o_{v,t}|q)$, before and after scaling respectively. Std_a and Std_v denote the original standard deviation of the log-likelihoods for the audio and visual stream respectively. The normalisation is performed on a per-utterance basis, i.e. a separate scaling factor is computed for each utterance by pooling log-likelihoods across all states and across all frames of the utterance.

In the current work it was further noted that, for the time-varying stream weight system, a further increase in performance could be gained by optimising a global scaling constant, η, applied to the normalised video stream likelihood,

$$b''_{vi} = \eta \cdot \frac{Std_a}{Std_v} \cdot b_{vi} \tag{6}$$

The inclusion of η allows the ratio of the standard deviation of the audio and visual streams to be set to an arbitrary value.

ACCEPTED MANUSCRIPT

Note that varying η will have no effect on the global SNR system. The effect of any multiplicative constant applied to one set of likelihoods can be equally well introduced as part of the SNR to stream weight mapping. As the mapping is optimised for the global SNR system after the application of Equation 6, scaling the video likelihoods will change the mapping but will not change the recognition result. However, scaling the likelihoods does have an effect on the performance of the local SNR-based system.

As noted above, the SNR to stream weight lookup table is dependent on the value of η. In other words, the lookup table and scaling factor must be jointly optimised. This joint optimisation is performed in such a way as to maximise recognition performance for both the global and local SNR-based systems. First, the value of η is chosen from a number of predefined values varying from [0.35...1...25]. When $\eta = 1$ the standard deviation of both audio and visual likelihood are the same. When $\eta > 1$ the standard deviation of audio likelihood is greater than that of visual likelihood and *vice versa*. The selected value of η and the training dataset (the same dataset that was used for training the neural network) are both passed to the exhaustive search block. This block performs an exhaustive search to optimise the lookup table which maps between the global SNR and stream weight, i.e. at each value of global SNR a series of different stream weights is tested and the stream weight that produces the best recognition performance is recorded. This table is then employed to lookup dynamic stream weights for recognition tests using the same training dataset and the same likelihood scaling factor, η. The lookup table and likelihood scaling factor, η, which lead to the best performance for both the global SNR and the local SNR system are chosen as those to be used for evaluation of the final test.

ACCEPTED MANUSCRIPT

4 Experimental Results and Analysis

This section describes the evaluation of the system. It will commence by describing the audio-visual speech data employed and the specifics of the system set-up. Following this, three sets of ASR experiments are presented. First, experiments are performed using *known* SNR (Section 4.2). Results using either known local or known global SNR are compared against audio-only and video-only baselines. These results establish an upper limit for the performance of the system. Section 4.3 presents a direct evaluation of the MLPs ability to predict SNR. These estimated SNRs are then mapped onto the stream weights used in the second set of ASR experiments (Section 4.4). The use of SNR estimates based on audio-only versus audio-visual state likelihoods are compared. The final set of ASR experiments (Section 4.5) illustrates the impact of errors in the estimation of the *sign* of the SNR. These errors result from foreground/background ambiguity. The structure of the training/testing data and details of the HMM-models employed which are common to all the experiments are described in Section 4.1.

4.1 Database and feature extraction

All experiments have employed the audio-visual Grid corpus [30] which consists of high quality audio and video recordings of small vocabulary 'read speech' 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"*. Of the 34 speakers, 20 (10 male and 10 female) are employed in the experiments reported here.

ACCEPTED MANUSCRIPT

Table 1 Structure of the sentences in the GRID corpus [30].

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

Figure 3 shows a representative selection of lip regions extracted from the corpus following the procedures described in Section 3.1. The upper half of the figure shows neutral positions for 8 of the 34 speakers illustrating the very large inter-speaker variability in lip appearance. The lower half shows a selection of lip positions for one of the male speakers. In order to give an indication of the high quality of the original video data the images are shown before the downsampling to 32×32 pixels that occurs during feature extraction. It can be observed that the images are evenly illuminated and have a high level of detail.

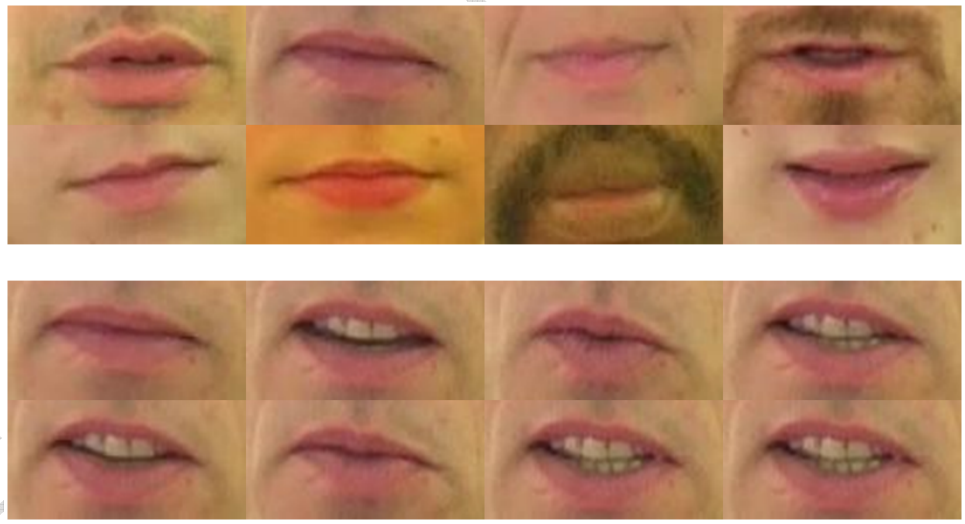


Fig. 3. Example lip regions extracted from the Grid corpus. The images in the top panel show eight different speakers in a resting lip position illustrating the inter-speaker variability. The images in the lower panel are examples of a single speaker with lips in different positions to illustrate the intra-speaker variability.

ACCEPTED MANUSCRIPT

In all the following experiments the behaviour of two different model configurations has been separately considered: i) a gender-dependent configuration in which models of male speech are used to recognise a male target masked by a female speaker; ii) a more challenging gender-independent configuration where gender-independent models are used to recognise a target of unknown gender mixed with a masker that is also of arbitrary unknown gender.

In the gender-dependent condition, 3500 utterances chosen from the 10 different male speakers (350 utterances from each speaker) have been employed to train a set of male (gender-dependent) HMMs. A collection of simultaneous speech mixtures was generated to be used variously for training the SNR estimator and evaluating the recognition system. 3100 utterances that have not been used during training were randomly chosen from the 10 male speakers to mix with 3100 utterances selected from 10 female speakers (310 utterances from each speaker). A Viterbi forced alignment was used to detect the initial and final silences which were then removed. The shorter utterance of each pair was zero-padded to the length of the longer one. The two signals were then artificially mixed at global SNRs of -10, 0, 5, 10, 15 and 20 dB. These mixed utterances were then randomly divided into three sets: 1000 utterances were employed to train the ANNs; 100 utterances were used as cross validation data to prevent the ANNs overfitting; the remaining 2000 utterances were used for the final recognition test set.

Preparation of the gender-independent condition was similar except 4000 utterances, taken from the 10 male and 10 female speakers, were used to train a set of gender-independent HMMs. Again 3100 simultaneous speech examples were constructed, but this time pairs of utterances were randomly chosen from the 20 speakers with the only restriction being that the target and masker are

ACCEPTED MANUSCRIPT

never the same speaker.

For each utterance, the 26-dimensional MFCC-based audio feature vectors and the 72-dimensional DCT-based video feature vectors were extracted according to the procedures described in Section 3.1. The 98-dimensional audio-visual feature vectors were constructed by simple concatenation of the audio and video components.

Word-level HMMs were employed to model the 51 words in the recognition task's vocabulary. The models contained between four and ten states per word determined using a rule of 2 states per phoneme. In each state both the audio and video emission distributions were modelled using a 5-component Gaussian mixture model with each component having diagonal covariance.

The audio and visual HMMs were trained independently using audio-only and visual-only features. The AV model was produced from the independent audio and visual HMM according to the procedure detailed in Section 3.

4.2 ASR Experiment 1: Known SNR

In the first experiments the known SNR is mapped onto a stream weight using the mapping – which takes the form of a global-SNR-to-weight lookup table – that has been previously optimised so as to maximise recognition performance as described in Section 3.3.3. The mapping is either applied to the global SNR to produce a single stream weight to use for all frames in the utterance, or the mapping is applied to the local (frame-based SNRs) to produce a time varying stream weight. In the latter case, the mapping is linearly interpolated to handle local SNRs that do not necessarily match the six global SNRs in

ACCEPTED MANUSCRIPT

the look up table.

At the same time, the likelihoods for the audio and visual streams are scaled appropriately for use in either the global SNR or the local SNR systems using the likelihood scaling factor optimised using the techniques described in Section 3.3.3. It was found, in both our tasks, that the local SNR-based performance was the greatest if the likelihood scaling factor was set such that the standard deviation of audio likelihood was 7.5 times that of the visual likelihood.

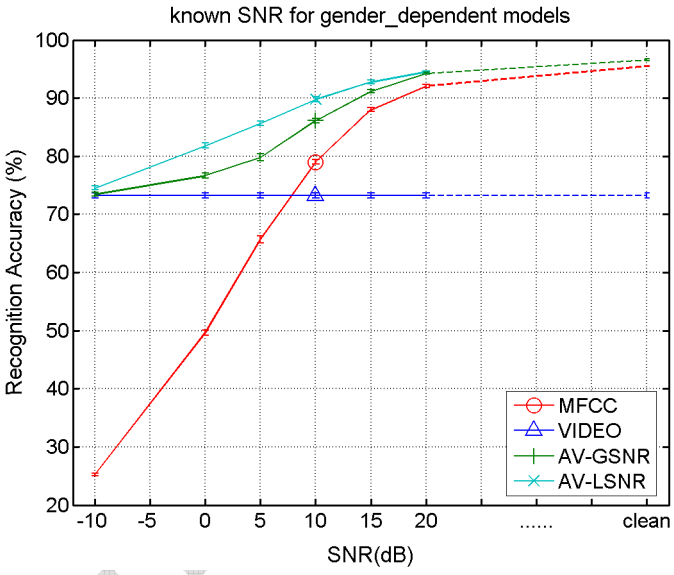


Fig. 4. A comparison of performance on the *gender-dependent* task for audio-only, visual-only and audio-visual ASR. The AV system employs stream weights estimated using *known SNR*. The audio-visual recognition accuracy is plotted against global SNR and is shown for both a constant stream weight (estimated from the global SNR) and a time-varying stream weight (estimated from the local SNR).

Figure 4 shows the results for speech recognition performance using gender-dependent models where the line marked with the circle and the line marked with the triangle are the traditional audio-only and visual-only speech recognition performance respectively. These results are in agreement with previous

ACCEPTED MANUSCRIPT

studies in that the visual stream produces significantly poorer results than the audio in low noise conditions (above an SNR of 8 dB). The video-only data is inherently ambiguous as many phonemes have similar visual appearance [36,37]. However, the video-only result is obviously not affected by the level of the acoustic noise and remains at 73.1% for all SNRs.

The line marked with the '+' symbol and the line indicated by the 'x' show the results of the multistreams model when using either the mapping from global SNR to fixed stream weight (AV-GSNR), or local SNR to time-varying stream weight (AV-LSNR) respectively. The audio-visual results are better than both the audio-only and visual-only baselines across all SNRs tested. The system using local SNRs outperforms that using global SNR in all SNR categories. The maximum gain is at an SNR of 5 dB where the recognition accuracy of the local SNR system is 5.8% (absolute) higher than that of the global SNR system. This is presumably because the time-varying stream weight allows the system to achieve better performance by exploiting the audio information in regions where the local SNR is temporarily favourable.

Figure 5 is the same as Figure 4 except that it shows speech recognition performance for the *gender-independent task*. Note that the pattern of results is very similar, though the average performance using the gender-independent models is clearly worse than that using the gender-dependent models. The speech recognition performance of the visual-only system falls from 73.1% for the gender-dependent system to 58.4% for the gender-independent system. As before, the local SNR information provides a relative improvement over using the single global SNR, with a maximum gain in accuracy of 9.9% occuring at the SNR of 5 dB.

ACCEPTED MANUSCRIPT

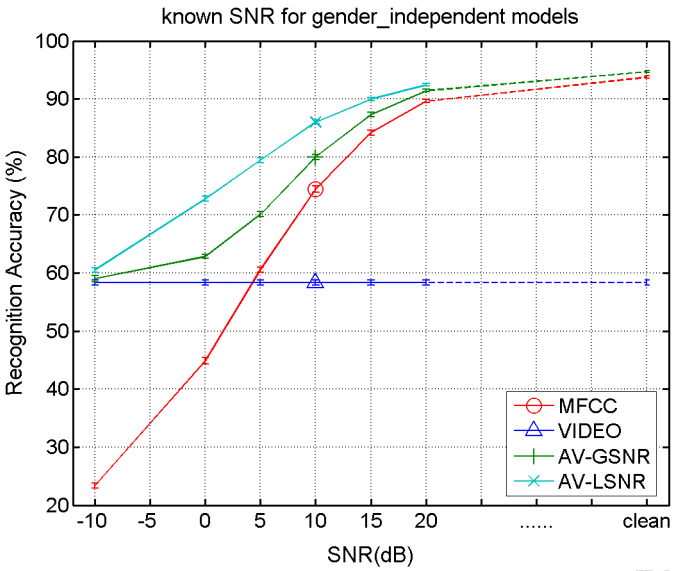


Fig. 5. A comparison of performance on the *gender-independent* task for audio-only, visual-only and audio-visual ASR. The AV system employs stream weights estimated from *known SNR*. The audio-visual recognition accuracy is plotted against global SNR and is shown for both a constant stream weight (estimated from the global SNR) and a time-varying stream weight (estimated from the local SNR).

4.3 Evaluation of SNR estimation

The previous section demonstrated the potential of the system when using *a priori* SNR values. The success of the real system will depend on the degree of reliability with which SNR can be estimated from the likelihood streams. This section presents a direct evaluation of this component of the system.

Separate MLPs of the structure described in Section 3.3.2 were trained for the audio-only and audio-visual SNR estimation. The MLPs had an input unit for each element of the likelihood feature vector. Using the state-clustering techniques described in Section 3.3.1, it was found that the 251-dimensional likelihood vectors could be reduced down to 54 and 18 dimensions for the audio and video stream respectively without reducing the audio-only and video-

ACCEPTED MANUSCRIPT

only recognition performance. Hence, it was decided that this would be an appropriate degree of clustering to apply to the likelihood vectors to be used in the SNR estimation. So, the audio-only MLP has 54 input units, whereas the audio-visual MLP has 72 (54 + 18) input units. The MLPs were trained using around 600,000 frames of data randomly drawn from the training data set. The target SNR at each frame is computed using *a priori* knowledge of the target and masker (i.e. knowledge of the clean target and masker utterances prior to mixing). The number of units in the hidden layer was optimised using a randomly drawn validation set consisting of 60,000 frames. In the gender-dependent task, the best performance for the audio-only MLP was achieved with 46 hidden units, while 30 hidden units gave best performance for the audio-visual MLP. In the gender-independent task, 11 hidden units gave the best performance for both audio-only and audio-visual MLPs.

The more challenging gender-independent system has been evaluated using a selection of 100 utterances taken from the test set mixed at the range of global SNRs used for the ASR experiments. The average magnitude of the difference between the SNR estimate and the target was computed. Also, the global SNR for each utterance was estimated from the sequence of local SNR estimates – using Equation 4 – and compared to the known global SNR. Table 2 shows the size of the local and global SNR error for both the audio-only and audio-visual estimates. It can be seen that other than for the local SNR at 20 dB, the audio-visual system consistently outperform the audio-only system. The error reduction is around 5 dB for the global SNR estimate with the largest gains at the lowest SNRs. The gain is particularly large at -10 dB. This large increase is presumably because at -10 dB there are many frames which appear to be clean speech but are in fact frames that are dominated by the masker. It

ACCEPTED MANUSCRIPT

is precisely this condition that can be disambiguated by the inclusion of video evidence.

Table 2 Average magnitude of local and global SNR estimation error (LSNR and GSNR respectively) for estimations based on either audio-only or audio-visual evidence. SNR estimation errors are shown in dB and are reported separately for utterances mixed at each global SNR.

	Audio-only		Audio-visual	
SNR (dB)	LSNR	GSNR	LSNR	GSNR
20	16.1	15.6	18.2	12.9
15	15.8	16.1	15.4	11.2
10	16.1	16.9	13.7	11.3
5	17.8	17.1	13.3	11.8
0	22.4	18.4	15.2	12.1
-10	34.7	28.9	21.5	17.1

To get an impression of what the figures in Table 2 mean in practice, Figure 6 compares the true and estimated local SNR over the course of a single pair of male utterances mixed at a global SNR of 5 dB. It can be seen that the audio-only system is prone to occasional regions of gross error (e.g. around frames 95 and 125). These gross errors are largely repaired by introducing visual information, and although the estimates are seldom very precise, the trajectory of the estimate is a reasonable match to that of the true SNR.

It may be considered that the errors in Table 2 seem disappointingly large. It would seem that state-likelihoods of a single frame do not contain sufficient evidence to consistently estimate SNR – possible reasons will be discussed in Section 4.6. However, large errors in the SNR estimation do not in themselves mean that the ASR results will be poor. First, it should be remembered that the SNR to stream weight mapping can be fairly flat over large SNR regions. For example, below 0 dB the optimum audio stream weight becomes very close

ACCEPTED MANUSCRIPT

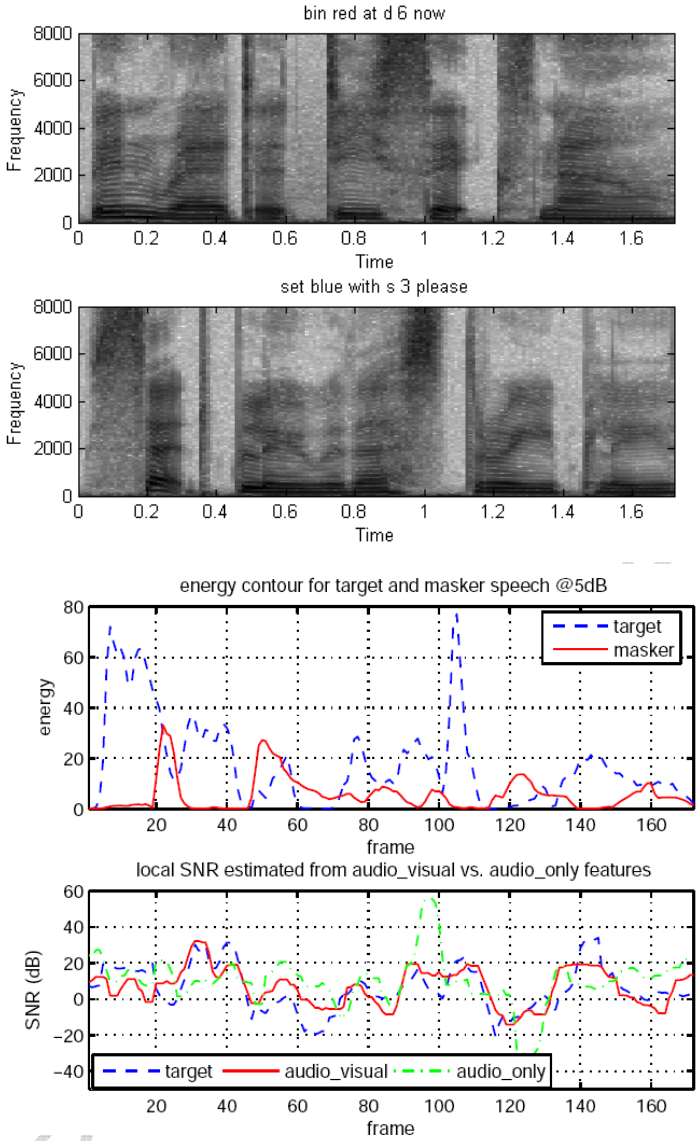


Fig. 6. The figure illustrates the quality of the SNR estimation for a typical utterance pair: The upper panels show the spectrogram of a target (top) and a masker utterance. Energy contours for the target and masker are shown for a global SNR of 5 dB. The bottom panels compares the true local SNR (dashed line) with the estimates produced by the MLP using audio evidence only (solid) and audio-visual evidence (dot-dash). Note that although the errors can be quite large, the audio-visual curve shows a similar trend to the true SNR, and makes fewer gross errors than the audio-only system.

ACCEPTED MANUSCRIPT

to 0 and above 20 dB it is very close to 1. So a large SNR estimation error in these regions – for example, estimating SNR as -30 dB rather than -10 – will not necessarily have a large impact on the ASR result. Second, for some frames, and even some utterances, there may be larger tolerance to error in the stream weighting parameter. If the errors are not occurring in the sensitive regions then they may not lead to recognition errors. Finally, the impact of the errors will depend not just on their average size, but also on their distribution. Although the average error can be large, it can be seen from Figure 6 that there can also be regions in which the SNR is well estimated. It is possible that a small number of poorly estimated stream weights can be tolerated by the HMM decoder. In short, although the direct evaluation of the MLP is informative and may be helpful in the development of system improvements, an evaluation via ASR results is the only fair test of the system as a whole.

4.4 ASR Experiment 2: Estimated SNR – Audio vs. Audio-visual estimators

In the second set of experiments the recognition systems are retested but this time using the SNRs that have been estimated from the likelihood data. Again, performance of both the global and local SNR-based systems is compared. A comparison is also made between the performance of the SNR estimates obtained using audio-only likelihoods and that of the SNR estimates obtained using the combined audio and visual streams (see Section 3.3.2).

Recognition accuracies for the gender-dependent task are shown in Figure 7. Results using the neural networks trained from either audio-only likelihood or both audio and visual likelihood are labelled with (A-*) and (AV-*) respectively. Both neural networks were employed to estimate either global SNR

ACCEPTED MANUSCRIPT

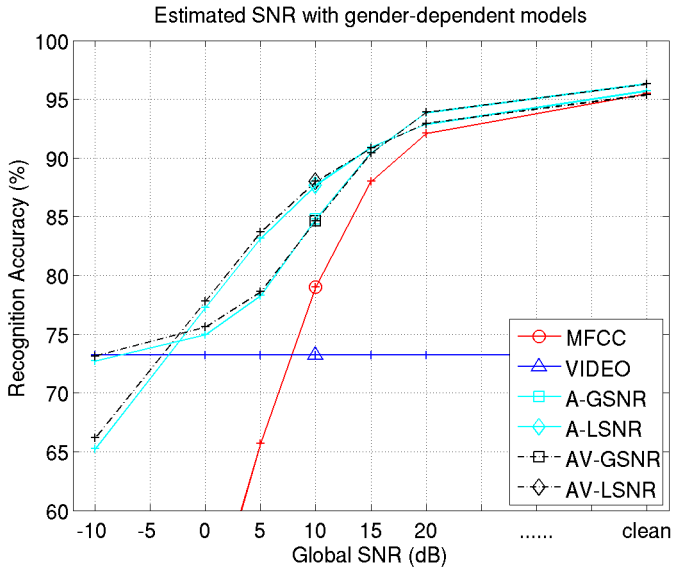


Fig. 7. Recognition accuracy obtained at each global SNR when using *gender-dependent* models and using SNR estimated from the log-likelihoods of either the audio-only or the audio-visual stream (A or AV) and when estimating either global or local SNR (GSNR or LSNR).

(*-GSNR) or local SNR (*-LSNR). It can be seen that the proposed stream weighting method leads to better results than both audio-only and visual-only baselines at all SNRs except -10 dB. However, comparison with Figure 4 indicates that recognition performances using *estimated* global and local SNR are worse than corresponding results using the *known* SNR. This figure also shows that the recognition results achieved using estimated local SNR are better than those for estimated global SNR at noise levels in the SNR range from 15 dB down to 0 dB, despite the fact that local SNR is harder to estimate robustly.

Comparing the pair of local SNR results, A-LSNR with AV-LSNR, or the pair of global SNR results, A-GSNR with AV-GSNR shows that the audio-only and the audio-visual based estimators provide broadly similar performance. At high noise levels – i.e. SNRs less than 10 dB – the audio-visual estimate

ACCEPTED MANUSCRIPT

provides a small but consistent benefit.

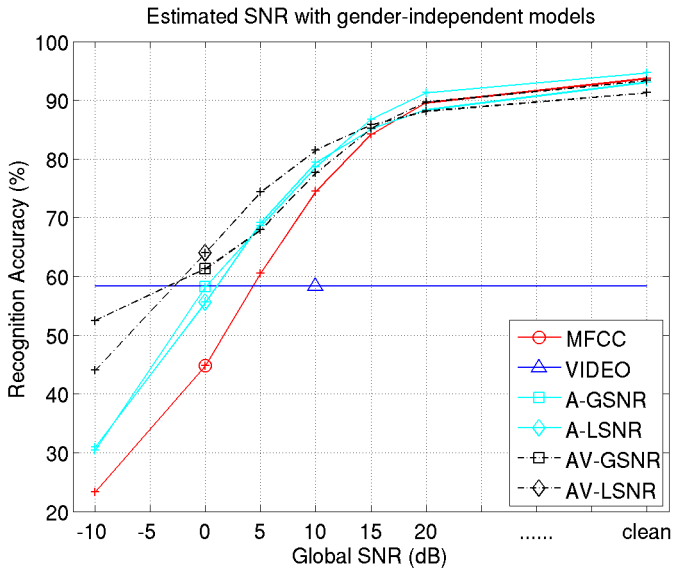


Fig. 8. Recognition accuracy obtained at each global SNR when using *gender-independent* models and using SNR estimated from the log-likelihoods of either the audio-only or the audio-visual stream (A or AV) and when estimating either global or local SNR (GSNR or LSNR).

Figure 8 shows the recognition accuracies for the gender-independent task. As with Figure 7, Figure 8 also indicates that the performance based on the local SNR (*-LSNR) is better than that based on the global SNR (*-GSNR) across a range of SNRs, i.e. 15 dB down to 0 dB. However, at very low SNRs the performance of the global SNR system is superior to that of the local SNR system. This is possibly because although in these conditions local SNR estimates have a significant mean-square error, they are generally unbiased, so the per-frame errors tend to be averaged out during the global SNR estimation.

Again, comparing these results with those in Figure 5 shows that the recognition performance using the proposed stream weighting method achieves better results than using either the audio or visual stream alone, but as expected, the performance using estimated SNR is somewhat less than that achieved using

ACCEPTED MANUSCRIPT

the known SNR.

Comparison of either A-GSNR with AV-GSNR, or A-LSNR with AV-LSNR illustrates that using visual information during the stream weight estimation leads to better recognition performance. This is as expected – the visual information is reducing the effects of '*informational masking*' (which is large in the gender independent condition) and helping to disambiguate regions of positive and negative SNR.

It can also be noted from Figure 8 that the performance of the audio stream (A-*) is slightly better than that of the audio-visual system (AV-*) in the high SNR region. Particularly in the local SNR case. The audio stream weight appears to be too low, an indication that the high SNRs are being underestimated.

Comparing the audio-visual (AV-*) results in Figure 8 with those in Figure 7 it is seen that the visual stream provides greater benefit in the gender-independent task than in the gender-dependent task. This is also as expected. In the gender-dependent task the acoustic models specifically match the gender of the target. The fact that the models match the acoustics of the target but not those of the gender, enables target and masker to be distinguished using acoustics alone, i.e. there is less role for the visual information in the gender-dependent task as there is less acoustic informational masking.

Both the above experiments show that the visual stream can help to improve speech recognition accuracies across a wide range of SNRs. However, the performance is still limited by an inability to form reliable estimate of *local* SNR, particularly when the target and masker are mixed at a low global SNR.

ACCEPTED MANUSCRIPT

4.5 ASR Experiment 3: Estimated SNR magnitude with known sign

In the simultaneous speech condition the foreground and background are acoustically similar. So, 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 experiments investigated the impact of this specific problem on the performance of the recognition system. The audio-only and audio-visual MLPs were retrained using the same dataset, parameters and optimising method as in the last experiment except that the signs of the local SNR targets 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.

Figure 9 and Figure 10 show the results for gender-dependent task and the gender-independent task respectively.

Considering first the gender-dependent system, it was noted that the results obtained using the audio-only estimator were not significantly different from those obtained using the audio-visual estimator (because they are not significantly different only the AV estimator results are shown in Figure 9). Furthermore, it can be noted that the results are close to those obtained by the known-SNR system (Figure 4). This would suggest that the poor performance seen in the previous systems that estimated both the sign and magnitude of the SNR (Figure 7) are arising due to an inability to estimate the sign. Furthermore, if only the magnitude of the SNR needs to be estimated then the audio likelihoods are sufficient. It also suggests that the previous small advantage provided by the visual stream when estimating both sign and magnitude

ACCEPTED MANUSCRIPT

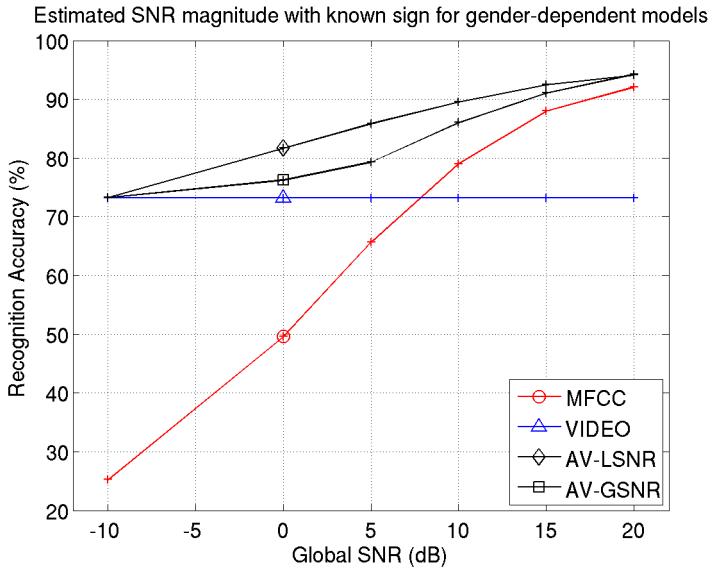


Fig. 9. Comparison of speech recognition performance obtained at each global SNR when combining the estimated SNR magnitude and the *a priori* SNR sign when using either global or local SNR (*-GSNR or *-LSNR) and *gender-dependent* models. Whether the SNR magnitude is estimated with audio-only likelihoods or audio-visual likelihoods makes no significant difference, so for the sake of clarity only the results using the audio-visual estimate are shown.

(Figure 7) is due to an improvement in the estimation of the sign of the SNR rather than its magnitude. This is expected as ambiguity in the sign of the SNR arises due to ambiguity in the determination of which source is the target and which the masker – it is precisely to reduce this ambiguity that the visual information is introduced

Figure 10 shows results for the gender independent task. As before, the structure of this figure is the same as that in Figure 8. It can be seen that the behaviour of the results is similar to that of the gender-dependent task (Figure 9) in that they are now very similar to the results obtained in when the SNR sign and magnitude are both known (Figure 5). The large advantage of using visual information for SNR estimation that is observed in the real

ACCEPTED MANUSCRIPT

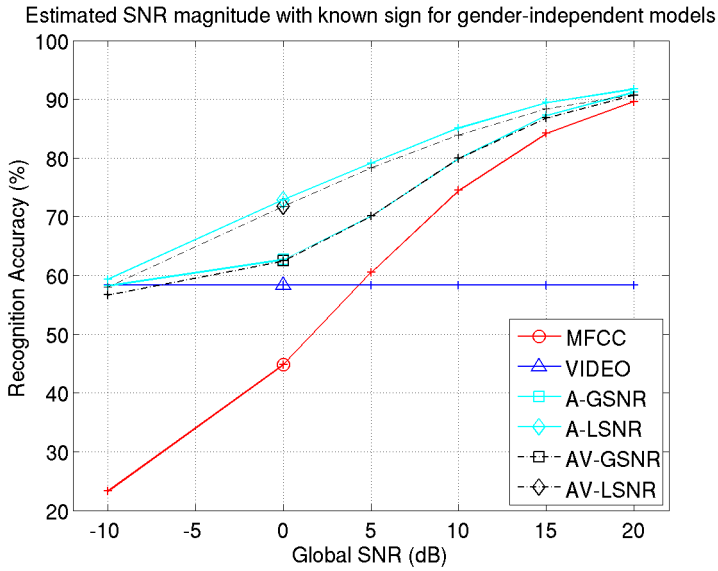


Fig. 10. Comparison of speech recognition performance obtained at each global SNR when combining the estimated SNR magnitude and the *a priori* SNR sign when using either audio-only likelihoods or audio-visual likelihoods (A-* or AV-*) and when using either global or local SNR (*-GSNR or *-LSNR) and *gender-independent* models.

system (Figure 8) is again reduced when the true SNR sign is known *a priori*. Again, this suggests that the main contribution of visual formation is in the estimation of this sign.

4.6 Discussion

4.6.1 The multiple roles of visual information

Considering the audio-visual speech recognition task, it is possible to identify three different levels at which the visual information may be employed. First, visual information may play an early role in mitigating the effects of energetic masking. For example, there may exist multimodal mechanisms that are similar to the auditory mechanisms that operate to provide comodulation masking

ACCEPTED MANUSCRIPT

release [38]. Second, audio visual integration may operate at a later stage to reduce the impact of informational masking (IM), that is, to help the listener selectively attend to the target while ignoring the masker. This particular role of visual speech has been the focus of recent perceptual studies, such as the work of Helfer and Freyman [6] and Wightman, Kistler and Brungart [39]. Finally, visual information may play a role in the speech unit classification task that underlies speech recognition. This is the role that has been central in the great majority of audio-visual ASR research to date. At this level, the visual information is useful to the extent that it provides extra features to help discriminate between partially masked, and otherwise ambiguous, speech sound classes.

A major contribution of the current work lies in emphasising the role of visual information as a cue for reducing 'informational masking', i.e. the 2nd of the three roles described above. In the current system, as in the study of Helfer and Freyman [6], the visual speech signal is helping to discriminate between the acoustic foreground and background, providing greatest benefit in the situation where the target and masker are most greatly confusable. In many everyday listening situations the release from informational masking afforded by the visual signal may be very significant. Brungart has demonstrated large informational masking effects on small vocabulary recognition tasks when the masker and target have the same gender and have a target-masker ratio in the range -3 dB to +3 dB [16]. For conversational speech where the perplexity of the recognition task is greater there is perhaps even more scope for target/masker confusion. A mechanism that allows the listener to reliably extract the target from the background is invaluable.

From the earliest audio-visual speech perception studies it has been noted that

ACCEPTED MANUSCRIPT

the audio-visual error rate is usually significantly lower than both the audio-only and video-only error rates. This is usually explained in terms of the complementarity of the phonetic/visemic cues. Some speech units are highly discriminable acoustically, while others are highly discriminable visually. However, this result may also be in part due to the role visual features play in reducing IM. Even if visual features carried no visemic information, and video-only ASR performance was no more than chance, it would still be possible for them to help drive auditory attention toward the acoustic foreground hence improving AV performance over that achieved using audio alone. Schwartz, Berthommier and Savariaux [40] have demonstrated exactly this point using carefully controlled intelligibility tests in which target words are visually identical and are masked by a speech background. This observation raises interesting possibilities. For example, consider a visual signal that is so low in quality (e.g. having a very poor resolution) that by itself it affords no more than chance recognition performance. It is possible that this signal may usefully reduce IM. In terms of the current system, this would be a case where the visual signal is sufficient to improve the stream weight estimation, but not to provide any visemic information at the classification stage.

4.6.2 Improving audio-visual stream weight estimation

Comparison of the results achieved using *known* SNR (Figures 4 and 5) with those achieved using *estimated* SNR (Figures 7 and 8) highlights the fact that performance of the system is limited by the extent to which SNR can be estimated. Good performance can be achieved if the models are sufficiently specific to the target (e.g. mixed genders and gender-dependent models), but when the background and foreground are statistically similar local SNR estimates are

ACCEPTED MANUSCRIPT

poor. In this condition visual information can help the SNR estimate (Figure 8) but, even then, AV recognition performance falls below that achieved using video alone at SNRs below 0 dB. The poor estimates can be largely overcome by averaging over an utterance to produce a global SNR, but this sacrifices the responsiveness of the time-varying stream weight which Figures 4 and 5 demonstrate to be important for the speech plus speech task.

The current SNR estimation technique bases its judgements on a single frame of data. To be effective the audio-visual system needs to judge whether the audio and visual states correspond. Despite the use of temporal difference features to capture local audio and visual dynamics, there may often be periods where there is insufficient context to reliably judge audio-visual correspondence. 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. The temporal granularity is too small. A potential solution would be to train the MLP using feature vectors formed by concatenating several frames of data. There is some precedence for such an approach in speech recognition. Feature vectors that include even up to half a second of context have been shown to be useful in improving phonetic discrimination [41]. However, a temporal window that is too long, if not employed with care, may lead to oversmoothing of the rapidly varying SNR, reintroducing the compromise that exists in the per-utterance SNR system employed in the current work. Therefore experiments would be needed to carefully optimise the window design, and the compression applied to the larger likelihood vector that would be generated by a larger window.

A further shortcoming of the existing system is that the *local* SNR to stream weight mapping is performed using a table that has been optimised for *global*

ACCEPTED MANUSCRIPT

SNR. Training a mapping using global SNR is convenient as each SNR point in the mapping can be independently optimised using a separate training set. However, it is not clear that this mapping should produce the best performance for a local SNR-based system. Thus, the results achieved with known local SNR, which are presented as an upper limit of the performance given perfect SNR estimation, could potentially be improved upon if an SNR to weight mapping more suitable for *local* SNR were found. One approach would be to use a parameterised curve to estimate the mapping. For example, given that stream weight is likely to be monotonically increasing with increasing SNR, and that is is constrained to lie in the range -1 to +1, a sigmoid might be a reasonable approximation. If the number of parameters is small enough it would be possible to locate the curve that maximises the ASR performance by a straightforward search of the parameter space.

4.6.3 Relation to other robust ASR approaches

The standard multistream approach to AV-ASR – of which the current work is an example – treats the acoustic feature vector as a single monolithic stream that has a single reliability weight. For the combination of additive noise and cepstral features employed in the current system this may be appropriate. Additive noise, whether broadband or narrowband, will effect the entire cepstrum. However, if the acoustics are represented in the spectral domain, then additive noise, at any particular time, will generally have a local effect. In the same way that some time frames may be relatively free of the masker, even in time frames that are heavily corrupted, some frequency bands will contain more masker energy than others.

ACCEPTED MANUSCRIPT

Many robust ASR techniques have been proposed to take advantage of local spectro-temporal regions of high SNR that can be found in noisy speech signals: multi-band systems have been developed which apply separate reliability weights to independent frequency bands [42]; soft-mask missing data system generalise this idea by attempting to judge the SNR at each spectro-temporal point [43]; speech fragment techniques attempt to piece together a partial description of the clean speech spectrum from reliable spectro-temporal fragments [44]. Such techniques might make a better starting point for developing robust audio-visual ASR systems. For example, using both acoustic and *visual* features to estimate multi-band reliability weights, or missing data soft-mask values would be a possibility.

Another general approach to robust ASR is to attempt to remove the noise from the mixture prior to recognition. For example, spectral subtraction techniques attempt to remove estimates of the noise spectrum in order to recover the clean speech spectrum [45,46]. This class of techniques may be considered complementary to the multistream approach described in this paper. Any technique that removes noise from the mixture to leave a cleaner version of the speech representation could be added to the current system as a preprocessing stage.

A common approach to the simultaneous speaker task is to exploit continuities in the speech spectrum, (primarily pitch), to track and separate the target and masker speakers. This was the strategy of several systems competing in the Pascal Speech Separation Challenge held at Interspeech 2007 [2] . However,

[2] Details of the Speech Separation Challenge can be found at http://www.dcs.shef.ac.uk/ martin/SpeechSeparationChallenge.htm

ACCEPTED MANUSCRIPT

apart from in exceptional circumstances such approaches are normally unable to offer a single unambiguous interpretation of the acoustic scene. For example, pitch tracks of competing sound sources can cross in ambiguous ways, or they may contain breaks leading to problems of sequentially grouping discontinuous pitch track segments. Pitch tracking is however sufficient to locate local regions of spectral-temporal dominance (e.g. vowel formants). So a potential strategy could be to first use such tracking technique to form a number of spectro-temporal acoustic fragments of unknown origin, and then to use both audio and visual evidence to judge the source identity of each spectro-temporal fragment. The target versus masker ambiguities that arise in the current system when considering a single acoustic frame would be much reduced when considering an extended spectro-temporal speech fragment. Integration of audio and visual features at this level is likely to lead to systems with greater robustness.

5 Conclusion

This paper has examined the problem of applying multistream audio-visual speech recognition techniques in a challenging simultaneous speaker environment using both gender-dependent and gender-independent hidden Markov models. It has been shown that in this condition, either a static stream weight parameter based on a global SNR estimate, or a dynamic stream weight parameter based on a local SNR estimate can be used to successfully integrate audio and visual information. The dynamic stream weight parameter leads to better overall performance. A technique for estimating either the local or global SNR from audio and visual HMM state likelihoods has been presented.

ACCEPTED MANUSCRIPT

Despite a lack of precision in the SNR estimates, the estimates were sufficiently reliable to lead to recognition results that are better than both the audio-only and visual-only baseline performance across a wide range of SNRs in both a gender-dependent and a gender-independent task.

Experiments using *a priori* SNR estimates have shown that a time-varying stream weight based on local SNR has the potential to greatly outperform a per-utterance stream weight based on a per-utterance SNR. The performance difference was most marked for global SNRs of around 0 dB. However, in practise, the difficulty of making accurate local SNR estimates means that real systems cannot match this performance level. Particular problems occur due to the difficulty in distinguishing between the acoustic foreground and background in local regions. Basing SNR estimates on a combination of both audio and visual likelihoods went some way to reducing this problem.

The paper has demonstrated the potential for a time-varying stream-weighting approach for AV speech recognition in multispeaker environments. However, it has also highlighted the difficulty in achieving results that match up to those that can be achieved using *a priori* SNR information. Possibilities for improving SNR estimation have been discussed, as have possibilities for combining the current approach in a complementary fashion with existing robust ASR approaches.

Finally, and most importantly, the paper has highlighted the potential for using visual information at multiple stages of the recognition process. In particular, there appears to be great potential for developing the use of visual speech information in the separation of acoustic sources, and in the disambiguation of the foreground/background confusions that occur when speech

ACCEPTED MANUSCRIPT

targets are mixed with acoustically similar maskers. More complete, multi-level integration of visual information into recognition systems may lead to future AV-ASR technology that comes closer to exhibiting the robustness of human speech processing.

References

[1] D. W. Massaro, *Perceiving talking faces: from speech perception to behavioral principle*, MIT press, Cambridge, 1998.

[2] L. D. Rosenblum, "The primacy of multimodal speech perception," in *Handbook of speech perception*, D.Pisoni and R.Remez, Eds., pp. 51–78. Blackwell, Malden, MA, 2005.

[3] W. H. Sumby and I. Pollack, "Visual contribution to speech intelligibility in noise," *Journal of the Acoustical Society of America*, vol. 26, pp. 212–215, 1954.

[4] A. Q. Summerfield, "Use of visual information in phonetic perception," *Phonetica*, vol. 36, pp. 314–331, 1979.

[5] D. S. Rudmann, J. S. McCarley, and A. F. Kramer, "Bimodal displays improve speech comprehension in environments with multiple speakers," *Human Factors*, vol. 45, pp. 329–336, 2003.

[6] K. S. Helfer and R. L. Freyman, "The role of visual speech cues in reducing energetic and informational masking," *Journal of the Acoustical Society of America*, vol. 117, no. 2, pp. 842–849, 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.

ACCEPTED MANUSCRIPT

[8] I. Matthews, *Features for audio-bisual speech recognition*, Ph.D. thesis, School of Information Systems, University of East Anglia, Norwich, 1998.

[9] G. Potamianos, H. P. Graf, and E. Cosatto, "An image transform approach for HMM based automatic lipreading," in *Proc. IEEE International Conference on Image Processing*, 1998.

[10] S. Lucey, T. Chen, S. Sridharan, and V. Chandran, "Integration strategies for audio-visual speech processing: Applied to text dependent speaker recognition," *IEEE Transactions on Multimedia*, vol. 7, no. 3, pp. 495–506, June 2005.

[11] C. C. Chibelushi, F. Deravi, and J. S. D. Mason, "A review of speech-based bimodal recognition," *IEEE Transactions on Multimedia*, vol. 4, no. 1, pp. 23–37, Mar. 2002.

[12] S. Dupont and J. Luettin, "Audio-visual speech modeling for continous speech recognition," *IEEE Transactions on Multimedia*, vol. 2, no. 3, pp. 141–151, Sept. 2000.

[13] U. Meier, W. Hurst, and P. Duchnowski, "Adaptive bimodal sensor fusion for automatic speechreading," in *Proc. ICASSP 1996*, Atlanta, GA, May 1996.

[14] E. K. Patterson, S. Gurbuz, Z. Tufekci, and J. N. Gowdy, "Noise-based audio-visual fusion for robust speech recognition," in *Proc. Audio-visual speech processing (AVSP'01)*, Scheelsminde, Denmark, 2001.

[15] H. Glotin, D. Vergyri, C. Neti, G. Potamianos, and J. Luettin, "Weighting schemes for audio-visual fusion in speech recognition," in *Proc. ICASSP 2001*, Salt Lake City, May 2001, pp. 173–176.

[16] D. S. Brungart, "Informational and energetic masking effects in the perception of two simultaneous talkers," *Journal of the Acoustical Society of America*, vol. 109, no. 3, pp. 1101–1109, 2001.

ACCEPTED MANUSCRIPT

[17] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. W. Senior, "Recent advances in the automatic recognition of audiovisual speech," *Proc. IEEE*, vol. 91, no. 9, pp. 1306–1326, 2003.

[18] C. M. Bishop, *Neural networks for pattern recognition*, Clarendon Press, Oxford, 1995.

[19] J. Luettin, G. Poaminanos, and V. Neti, "Asynchronous stream modeling for large vocalbulary audio-visual speech recognition," in *Proc. ICASSP 2001*, Salt Lake City, May 2001.

[20] A. Garg, G. Potamianos, C. Neti, and T. S. Huang, "Frame-dependent multi-stream reliability indicators for audio-visual speech recognition," in *Proc. ICASSP 2003*, Hong Kong, Apr. 2003, pp. 24–27.

[21] D. W. Massaro and D. G. Stork, "Speech recognition and sensory integration," *American Scientist*, vol. 86, no. 3, pp. 236–244, 1998.

[22] S. Tamura, K. Iwano, and S. Furui, "A stream-weight optimization method for multi-stream HMMs based on likelihood value normalisation," in *Proc. ICASSP 2005*, Philidelphia, PA, Mar. 2005.

[23] S. Gurbuz, Z. Tufekci, E. Patterson, and J. N. Gowdy, "Multi-stream product model audio-visual integration strategy for robust adaptive speech recognition," in *Proc. ICASSP 2002*, Orlando, FL, May 2002.

[24] S. Cox, I. Matthews, and J. A. Bangham, "Combining noise compensation with visual information in speech recognition," in *Proc. Audio-visual speech processing (AVSP'97)*, Rhodes, Greece, 1997.

[25] H. G. Hirsch, "Estimation of noise spectrum and its application to SNR-estimation and speech enhancement," Tech. Rep. TR-93-012, International Computer Science Institute, Berkeley, CA, 1993.

ACCEPTED MANUSCRIPT

[26] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, pp. 357–366, 1980.

[27] G. R. Bradski, "Computer video face tracking for use in a perceptual user interface," *Intel Technology Journal*, vol. Q2, 1998.

[28] L. R. Rabiner, "A tutorial on HMMs and selected applications in speech recognition," *Proc. IEEE*, vol. 12, no. 2, pp. 267–296, 1989.

[29] S. Young, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, "The HTK book," http://htk.eng.cam.ac.uk/, 1995.

[30] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, 2006.

[31] S. Okawa, T. Nakajima, and K. Shirai, "A recombination strategy for multiband speech recognition based on mutual information criterion," in *Proc. European Conference on Speech Communication Technology*, Budapest, 1999, pp. 603–606.

[32] G. Potamianos and C. Neti, "Stream confidence estimation for audio-visual speech recognition," in *Proc. ICSLP 2000*, Beijing, China, Oct. 2000.

[33] A. Adjoudani and C. Benoit, "On the integration of auditory and visual parameters in an HMM-based ASR," in *Speechreading by Humans and Machines*, D.G. Stork and M.E. Hennecke, Eds., pp. 461–471. Springer, Berlin, 1996.

[34] J. R. Shewchuk, "An introduction to the conjugate gradient method without the agonizing pain," Tech. Rep. CMU-CS-94-125, School of Computer Science, Carnegie Mellon University, 1994.

ACCEPTED MANUSCRIPT

[35] G. Gravier, S. Axelrod, G. Potamianos, and C. Neti, "Maximum entropy and MCE based HMM stream weight estimation for audio-visual ASR," in *Proc. ICASSP 2002*, Orlando, FL, May 2002.

[36] S. Fu, R. Gutierrez-Osuna, A. Esposito, P. K. Kakumanu, and O. N. Garcia, "Audio/visual mapping with cross-modal hidden Markov models," *IEEE Transactions on Multimedia*, vol. 7, no. 2, pp. 243–252, Apr. 2005.

[37] A. J. Goldschen, *Continuous Automatic Speech Recognition by Lipreading*, Ph.D. thesis, Engineering and Applied Science, George Washington University, 1993.

[38] J. W. Hall, M. P. Haggard, and M. A. Fernandes, "Detection in noise by spectro-temporal pattern analysis," *Journal of the Acoustical Society of America*, vol. 76, pp. 50–56, 1984.

[39] F. Wightman, D. Kistler, and D Brungart, "Informational masking of speech in children: Auditory-visual inte gration," *Journal of the Acoustical Society of America*, vol. 119, no. 6, pp. 3940–3949, 2006.

[40] J. L. Schwartz, F. Berthommier, and C. Savariaux, "Seeing to hear better: evidence for early audio-visual interactio ns in speech identification," *Cognition*, vol. 93, pp. B69–B78, 2004.

[41] H. Hermansky and S. Sharma, "TRAPs - classifiers of temporal patterns," in *Proc. ICSLP 1998*, Sydney, Australia, Nov. 1998.

[42] H. Bourlard and S. Dupont, "A new ASR approach based on independent processing and recombination of partial frequency bands," in *Proc. ICSLP '96*, Philadelphia, PA, Oct. 1996.

[43] J. P. Barker, L. B. Josifovski, M. P. Cooke, and P. D. Green, "Soft decisions in missing data techniques for robust automatic speech recognition," in *Proc. ICSLP'00*, Beijing, China, 2000, pp. 373–376.

ACCEPTED MANUSCRIPT

[44] J. P. Barker, M. P. Cooke, and D. P. W. Ellis, "Decoding speech in the presence of other sources," *Speech Communication*, vol. 45, pp. 5–25, 2005.

[45] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 27, no. 2, pp. 113–120, 1979.

[46] P. Lockwood and J. Boudy, "Experiments with a non-linear spectral subtractor (NSS) Hidden Markov Models and the projection, for robust speech recognition in cars," in *Eurospeech'91*, 1991, vol. 1, pp. 79–82.

ACCEPTED MANUSCRIPT