Robust Audio-Visual Speech Recognition Under Noisy Audio-Video Conditions

Darryl Stewart, Rowan Seymour, Adrian Pass, and Ji Ming, *Member, IEEE*

***Abstract*—This paper presents the maximum weighted stream posterior (MWSP) model as a robust and efficient stream integration method for audio-visual speech recognition in environments, where the audio or video streams may be subjected to unknown and time-varying corruption. A significant advantage of MWSP is that it does not require any specific measurements of the signal in either stream to calculate appropriate stream weights during recognition, and as such it is modality-independent. This also means that MWSP complements and can be used alongside many of the other approaches that have been proposed in the literature for this problem. For evaluation we used the large XM2VTS database for speaker-independent audio-visual speech recognition. The extensive tests include both clean and corrupted utterances with corruption added in either/both the video and audio streams using a variety of types (e.g., MPEG-4 video compression) and levels of noise. The experiments show that this approach gives excellent performance in comparison to another well-known dynamic stream weighting approach and also compared to any fixed-weighted integration approach in both clean conditions or when noise is added to either stream. Furthermore, our experiments show that the MWSP approach dynamically selects suitable integration weights on a frame-by-frame basis according to the level of noise in the streams and also according to the naturally fluctuating relative reliability of the modalities even in clean conditions. The MWSP approach is shown to maintain robust recognition performance in all tested conditions, while requiring no prior knowledge about the type or level of noise.**

***Index Terms*—Automatic speech recognition, human computer interaction, speech recognition.**

I. Introduction

A WEAKNESS of most modern ASR systems is their inability to cope well with signal corruption, and there are many ways in which this may occur. There may be other sound sources (e.g., background noise, other people speaking), wave reflections (e.g., reverberation or echoes), or transmission channel distortions caused by the hardware (usually the microphone) used to capture the speech signal. Thus, one of the main challenges in the ASR domain is how to develop systems that are more robust to the kinds of noise that are typically encountered in real-world situations.

Manuscript received May 11, 2012; revised September 22, 2012; accepted February 12, 2013. Date of publication April 8, 2013; date of current version January 13, 2014. This work was supported in part by the U.K. EPSRC under Grants EP/G034303/1 and EP/G001960/1. This paper was recommended by Associate Editor C.-T. Lin.

D. Stewart and J. Ming are with the Queen's University of Belfast, Belfast BT3 9DT, U.K. (e-mail: dw.stewart@qub.ac.uk; j.ming@qub.ac.uk).

R. Seymour is with the International Training and Education Center for Health, University of Washington, Seattle, WA 98105 USA (e-mail: rseymour01@qub.ac.uk).

A. Pass is with Pace plc, Saltaire BD18 3LF, U.K. (e-mail: apass01@qub.ac.uk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCYB.2013.2250954

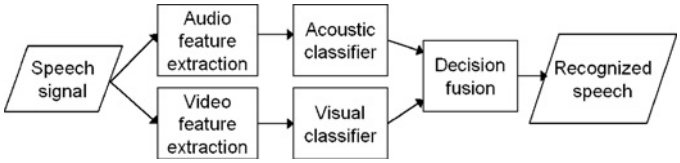

Fig. 1. Decision-fusion based audio-visual speech recognition.

One approach to this problem is to introduce another modality to complement the acoustic speech information, and usually this is a video recording of the speaker's lips. It is well known that humans use both acoustic and visual information when communicating with each other when both modalities are available. Indeed, this has been usefully demonstrated by the McGurk effect [1].

An audio-visual speech recognition (AVSR) system uses visual speech information in addition to the acoustic information used by a standard ASR system. Audio and video information can be integrated by feature fusion or by decision fusion.

Feature fusion means that the information is combined at the feature level, and a single combined feature vector is passed to a single classifier. This is generally simple to implement and allows modeling of the correlation between audio and video. The simplest method of feature fusion is feature concatenation [2], where the audio and video feature vectors are simply concatenated. Other approaches also include feature weighting [3], the hierarchical linear discriminant feature extraction method described in [4], and audio feature enhancement [5]. Feature fusion based techniques lack the ability to explicitly model the relative reliability of each feature stream. This is important as the reliability of either stream may vary significantly even within the duration of an utterance because of constant or instantaneous background noise or channel degradations.

In contrast, decision fusion systems assume independence between the two streams and instead combine the results of separate classifiers for audio and video. Fig. 1 shows the general process of an AVSR system that uses decision fusion.

Unlike feature fusion approaches, decision fusion offers a mechanism for modeling the reliabilities of each feature stream by using separate classifiers for audio and video. The most common types of decision fusion methods are those that combine a parallel classifier architecture, such as a multistream

This work is licensed under a Creative Commons Attribution 3.0 License. For more information, see http://creativecommons.org/licenses/by/3.0/

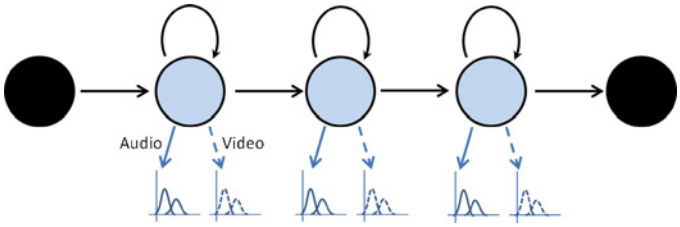

Fig. 2. Example of a multistream HMM with two streams (audio and video) and three states per stream (inspired by [11]).

HMM (MSHMM) with fixed or adaptive combination weights [6]–[11]. These methods calculate the most likely class (state, sub-word, or word) by combining the scores or log-likelihoods of the individual classifiers for the audio and video streams. Some alternatives to MSHMMs include the product or coupled HMM [12], [13], discriminative model combination [2], artificial neural networks (ANNs) [14], ANN/HMM hybrids [15], or dynamic Bayesian networks (DBNs) [16].

In an MSHMM, each observation is modeled as a single-stream HMM that emits two likelihoods for each state, as illustrated in Fig. 2. Typically, the class conditional observation likelihoods of an MSHMM are calculated as the product of the observation likelihoods of the individual streams. Weights may be applied to stream likelihoods to capture the relative reliabilities of the streams. Such MSHMM architectures have been used for audio-only ASR to combine the results of subband features or static and dynamic feature streams. For the task of audio-visual integration, MSHMMs with two streams are used. These have been used in small vocabulary AVSR systems [12], as well as large vocabulary systems [6].

In this model, if the two streams are A and V for audio and video, respectively, and X_{A_t} and X_{V_t} represent their feature vectors for a frame at time t, then the log-likelihood of observing both X_{A_t} and X_{V_t} given state s is calculated as

$$log[P(X_t|s)] = \gamma_A log[P(X_{A_t}|s)] + \gamma_V log[P(X_{V_t}|s)] \quad (1)$$

where γ_A and γ_V are the weights applied to the audio and video streams, respectively. These may be fixed [6], [17] or calculated to reflect the actual reliabilities of the streams. Constraints such as $\gamma_A + \gamma_V = 1$ are also often used. Thus, the class conditional state emission probability of the MSHMM given the feature vector X_t is

$$P(X_t|s) = \prod_{b\in\{A,V\}} [\sum_{m=1}^{M_b} c_{smb}\mathcal{N}_b(x_{b_t}, \mu_{smb}, U_{smb})]^{\gamma_b} \quad (2)$$

where M_b is the number of mixtures per state for stream $b \in \{A, V\}$, c_{smb} is the mixture weight for the mth mixture component in state s for stream b. $\mathcal{N}_b$ represents a Gaussian density with mean vector μ_{smb} and covariance matrix U_{smb}. The parameters for the single stream HMMs may be estimated individually using the EM algorithm in the same way as for a traditional HMM. However, this approach means that the HMMs are trained asynchronously and will use different forced alignments. Alternately, they can be trained together using an *a priori* choice of stream weights.

As was stated earlier, a major advantage of the decision fusion approach is the ability to apply weights during the fusion process to capture the relative reliabilities of the audio and video feature streams. The weights may be set globally to fixed values that are calculated from testing the system to find the weights that produce optimal speech recognition [6], [9], [17].

However, in many real-world environments where there is a possibility of noise or corruption entering the video or audio signal the reliabilities of each modality will be fluctuating. For instance for applications where the speaker is in an acoustically hostile environment with many potential sources of background noise the reliability of the audio stream will deteriorate and change over time. Similarly, the reliability of the video stream will be affected by various factors, such as camera position, head pose, head and mouth tracking, and video compression. Therefore, to model this variance in reliability, frame or utterance level weights are required, which are usually calculated from an estimation of environmental conditions or stream reliability. This has been done in a variety of ways, including the following.

1) Reference [2] calculated weights as a function of the audio channel signal-to-noise ratio (SNR).
2) Reference [18] minimized a smooth function of the minimum classification error (MCE) using generalized probabilistic descent.
3) Reference [19] minimized the frame misclassification rate, by using the maximum entropy criterion.
4) Reference [20] calculated frame dependent weights using the maximum conditional likelihood and the MCE.

An ideal audio-visual integration approach should satisfy two criteria. First, it should allow performance that is greater than either stream on its own for low levels of corruption in either stream, and second, in cases where one stream is highly corrupted the recognition performance should remain similar to the most reliable stream. These criteria should be met even in conditions where the corruption in either stream may be time varying and involves *a priori* unknown types of corruption.

In this paper, we propose a novel approach called the maximum weighted stream posterior (MWSP) method that aims to meet these criteria. MWSP offers a smooth integration of the two modalities, making best use of the available reliable information and remaining simple to implement without *a priori* knowledge of the environmental conditions, in which it will be deployed or tested. As it does not entail any measurements of the signal in either stream when calculating the stream weights it is essentially a modality-independent approach to fusion that could be applied to a range of problems where multiple modalities or feature streams are being combined in the potential presence of unpredictable corruption in any of the streams. A preliminary study of MWSP was reported in [21]. This paper extends the study with a more in-depth description of the method and an increased experimental investigation and analysis of the performance of MWSP for AV speech recognition in both clean and noisy conditions for both modalities. In this paper, we compare MWSP with

another dynamic modality weighting approach and also an Oracle fixed-weighted model to gain an insight into its relative performance and stability. We also expand our experimental test conditions to include video data that is corrupted by MPEG-4 video compression. While the vocabulary used in our test data (digits) is not intended to be application specific, these new experiments are significant in showing the relative robustness of the MWSP approach for application domains, where cloud-based speech recognition may be preferred, e.g., on mobile platforms. In such applications, it is probable that the AV data would be streamed over the network in a compressed format to the cloud service that would then carry out the feature extraction and recognition process before returning the results, as described in [22].

The rest of this paper is organized as follows. In Section II, we describe the development of the MWSP approach in detail. In Section III, we present the experimental results and finally a summary and concluding remarks are given in Section IV.

II. Maximum Weighted Stream Posterior Model

The simplest form of integration using decision fusion would be the *Product Model* that takes a product of the individual likelihoods of the audio and video streams as in

$$p(X_{AV_t}|s) = p(X_{A_t}|s) \cdot p(X_{V_t}|s). \tag{3}$$

However, to ensure the values produced by each stream are on the same scale, it is usually necessary to normalize each likelihood based on the dimensionality of the stream as follows (please note that in all further equations found here the t has been dropped to improve clarity):

$$p(X_{AV}|s) = p(X_A|s)^{\frac{1}{d_A}} \cdot p(X_V|s)^{\frac{1}{d_V}} \tag{4}$$

where d_A and d_V are the dimensions of the audio and video feature vectors respectively. Furthermore, as was explained in the previous section, it is often useful to apply stream weights to account for the expected relative reliabilities of each stream, as shown in

$$p(X_{AV}|s) = p(X_A|s)^{\frac{\gamma_A}{d_A}} \cdot p(X_V|s)^{\frac{\gamma_V}{d_V}} \tag{5}$$

where the stream exponents γ_A and γ_V denote the relative reliability values, and typically, a constraint such as $\gamma_A+\gamma_V = 1$ is applied. For instance, in [6] and [17], values of $\gamma_A = 0.7$ and $\gamma_V = 0.3$ were used.

These static weights are useful if the recognition conditions are stable and match the training conditions. However for real world conditions where there may be time varying noise/corruptions affecting either stream then static weights are often sub optimal. The challenge in these conditions, therefore, is to develop a system which can adapt quickly to changing reliabilities.

If we do not have any knowledge of the level of corruption in the streams during recognition and we must calculate a likelihood for each frame which uses the best available clean information then we must choose between $p(X_{AV}|s)$, $p(X_V|s)$ and $p(X_A|s)$. In order to directly compare these values, which may be on different scales, we can normalize by converting to posterior probabilities. One method of selecting the optimal stream probability is the maximum stream posterior (MSP) method [23] that is expressed formally as follows:

$$P(s|X) = max[P(s|X_A), P(s|X_V), P(s|X_{AV})] \tag{6}$$

where $P(s|X)$ denotes the optimal posterior probability of state s given frame X, where X could be X_A, X_V or the combination X_{AV}.

Using Bayes' theorem the individual posterior probabilities for each stream can be written as (note that we consider AV to be a stream rather than multiple streams)

$$P(s|X_A) = \frac{p(X_A|s)P(s)}{\sum_{s'} p(X_A|s')P(s')} \tag{7}$$

$$P(s|X_V) = \frac{p(X_V|s)P(s)}{\sum_{s'} p(X_V|s')P(s')} \tag{8}$$

$$P(s|X_{AV}) = \frac{p(X_A|s)p(X_V|s)P(s)}{\sum_{s'} p(X_A|s')p(X_V|s')P(s')} \tag{9}$$

where $p(X_A|s)$ and $p(X_V|s)$ are the likelihood functions of X_A and X_V and independence is assumed between them, $P(s)$ is the prior probability of state s, and the summation in the denominators for s is over all possible states within the search beam. The optimal posterior $P(s|X)$ will be incorporated into a HMM as an approximation of the state-based emission probability.

Assuming that the least corrupted stream will produce the maximum likelihood ratio between correct and incorrect states, then selecting the maximum of the posteriors $P(s|X_A)$, $P(s|X_V)$ and $P(s|X_{AV})$ is likely to obtain the least corrupt stream. This can be shown by rewriting the posterior probabilities in a form of likelihood ratios between the states. For example, in the case of X_A we can rewrite (7) as

$$P(s|X_A) = \frac{P(s)}{P(s) + \sum_{s' \neq s} P(s') \frac{p(X_A|s')}{p(X_A|s)}}. \tag{10}$$

For correct state s the likelihood ratio in the denominator given by

$$p(X_A|s) / \sum_{s' \neq s} p(X_A|s')$$

and hence the posterior probability $P(s|X_A)$ is likely to be maximized when X_A is the least corrupt stream. Therefore, the MSP method (6) represents a method for choosing the best feature stream from (A, V or AV) for recognition without assuming prior information of the corruption.

A potential weakness of the MSP method is that an individual stream either contributes equally (or at a single fixed weighting) to the final posterior probabilities, or is ignored completely. It is intuitive, however, that if one stream is clean and one has moderate corruption, the latter can still contribute useful discriminatory information, but that less confidence should be placed on it, i.e., a weighting against it. To this end, the MWSP method seeks to find a softer and optimal weighting for the combination of the two streams by examining a set of weightings that cover the full range of relative stream confidences. This range includes equal confidence (equivalent to (9) of the MSP), and absolute bias toward either stream

(equivalent to (7) and (8) of the MSP). Thus, the MSP can be considered a special case of the MWSP. The MWSP for AV fusion is described by the following formula that gives the posterior probability of state s for a given weighting w

$$P_w(s|X) = \frac{p(X_A|s)^w p(X_V|s)^{1-w} P(s)}{\sum_{s'} p(X_A|s')^w p(X_V|s')^{1-w} P(s')} \tag{11}$$

where it is assumed that $w \in [0, 1]$. The MSP method can be obtained approximately by setting $w = 1$ (7), $w = 0$ (8), or $w = 0.5$ [(9), approximately]. The optimal posterior probability used for recognition is the maximum across all of the possible weightings, that is

$$P(s|X) = \max_w \; P_w(s|X). \tag{12}$$

For our implementation of the above approach, a suitable finite set of values for w must be evaluated, which will cover the potential relative reliabilities of each stream. This forms part of the experimental investigation in the following sections of the paper.

A. Posterior Union Model

For our experiments in the following section, we have included a comparison between MWSP, MSP, the Product model and the posterior union model (PUM). The PUM, which is based on probability theory for the union of random events, can be viewed as an alternative way of implementing a dynamic weighting scheme within a multi-stream system. Previous research has demonstrated the PUM to be a robust method of combining feature streams in unknown noise conditions. In [24], the PUM was used in a subband approach to noisy speech and speaker recognition where the aim was to base the recognition as much as possible on the cleanest frequency bands without any prior knowledge of the noise in any subband. In [25], the PUM was used to combine separate feature streams from image segments for face recognition where the images were subject to various forms of corruption. As both these problems are analogous to the problem under investigation in this paper and the fact that the PUM performs under the same operating assumptions as MWSP, i.e., no prior knowledge of the noise in any stream is assumed, we chose the PUM as a suitable baseline method for comparison with MWSP.

In our experiments involving audio-visual integration a PUM with two possible combinations of streams is used. These correspond to the conjunction and disjunction of the two streams, that is

$$P_\wedge(s|o) = \frac{p(o^A|s)p(o^V|s)P(s)}{\sum_{s'} p(o^A|s')p(o^V|s')P(s')} \tag{13}$$

$$P_\vee(s|o) = \frac{(p(o^A|s) + p(o^V|s))P(s)}{\sum_{s'}(p(o^A|s') + p(o^V|s'))P(s')}. \tag{14}$$

The maximum of $P_\wedge(s|o)$ and $P_\vee(s|o)$ is selected as the state-based emission probability within the MSHHM.

III. Experiments

A. Database

In this paper, we have used the XM2VTS database [26] to carry out tests on speaker-independent isolated digit recognition. XM2VTS was chosen for this paper as it is one of the largest audio-visual databases available and as such it has been widely used for many significant studies by the AV speech community in the past. The database contains 295 speakers, roughly balanced between genders. Each speaker was recorded in four different sessions in a quiet environment. The following three sentences are repeated twice in each of a speaker's four sessions.

1) "0 1 2 3 4 5 6 7 8 9."
2) "5 0 6 9 2 8 1 3 7 4."
3) "Joe took father's green shoe bench out."

For this paper we used only the digit utterances. Messer *et al.* [26] provide the speech data in two sets: the 3A set that contains 200 subjects for training a speaker independent system, and the 3B set that contains the remaining 95 subjects for testing such a system. Thus, there were 3200 training occurrences of each digit and the test data includes 1520 test tokens for each digit. The data is supplied as continuous digit sequences with only sentence level transcriptions. However, for this paper we decided to carry out isolated digit recognition experiments, so a forced alignment procedure was initially carried out on all utterances using the hidden Markov toolkit (HTK) [27] to obtain word boundary positions.

The database was supplied with some lip tracking results, using the colour based approach described by Ramos [28]. These were used to localize the mouth region of interest (ROI) in each video frame, eliminating the need for mouth tracking.

B. Data Corruptions

To examine the effect of corruption in the audio stream, additive full-band white noise was added individually to each isolated digit in the audio test data at different average SNR levels (−20 dB to 30 dB). We chose full-band white noise to corrupt the audio stream in our experiments as it represents the most severe scenario where all frequencies of the audio are corrupt to some level. In these conditions, the audio stream becomes increasingly unreliable, and we can therefore see how the multimodal approaches cope in these conditions. It should be noted that although we add the noise to each utterance at an average SNR, this does not mean that the noise statistics are static over the entire utterance. The local SNR for each frame in an utterance will be different due to the non-stationary nature of the speech. The weights calculated for each stream by the MWSP approach are influenced only by the local emphframe SNR, and therefore, the weights may not be constant over an utterance; instead, they fluctuate on a frame-by-frame basis. Similarly, to examine the effects of corruption in the video stream, two different types of noise were used: MPEG-4 video compression and simulated camera jitter. MPEG-4 compression was applied at various levels (ranging from 512 kb/s to 4 kb/s) and as mentioned earlier in Section I, these tests are useful in showing the relative robustness of the tested approaches for application domains

Fig. 3. (a) Original lip ROI image. (b) Original image with Jitter applied (level 12).

where cloud-based speech recognition may be used. Jitter is a novel form of corruption developed to simulate situations where either the camera or the speaker is moving, preventing smooth tracking of the mouth. It is applied by adding a random variation to the coordinates and orientation of the mouth ROI. Different levels of jitter are generated by scaling the random variation. For example, jitter level 10 corresponds to a random rotation in the range $[-10\,\text{deg}, 10\,\text{deg}]$ and separate random translations along the x and y axes, both in the range $[-10, 10]$ pixels. Six jitter levels (2, 4, 6, 8, 10, 12) were used on the test video data. An example video frame with jitter corruption added can be seen in Fig. 3.

C. Feature Extraction

Numerous approaches to visual feature extraction have been presented in the literature, each with potential advantages and disadvantages [29]–[31]. In this paper, we chose to use the DCT (type II) transform to extract features from each ROI. It is not our intention to present these features as optimal in our experimental setup as that is not the focus of our research. Instead we chose these features as they are a commonly used feature type in a variety of other important studies [29], [32], [33]. Furthermore, we found in previous research using the XM2VTS database [30] that DCT features are efficient to extract and reasonably robust in the presence of several types of image corruption. As our focus is primarily on investigating how the MWSP approach responds to degradation in either the audio or video streams we do not view the specific choice of features used in either stream as critical in this study.

Hence, for each frame of video, the mouth ROI was cropped and subsampled to 16×16 before applying the DCT, as shown in Fig. 4. The transformation results in a 16×16 array of coefficients for each frame of video. The visual features were selected from these coefficients using a triangle mask (Fig. 4), as this reflects how the coefficients are arranged with the lowest frequencies at the origin. Finally, the visual feature stream was formed by interpolating these features to 100 Hz to match the sampling rate of the audio feature stream. Cubic splines were used to interpolate the features, which enabled easy calculation of dynamic features. Again, based on our previous work [30], we chose to use 15 static features along with 15 Δ dynamic features for our final feature vector.

All of the features were normalized because there is a lot of variation between speakers and sessions. This was done by analyzing each utterance and performing mean subtraction and variance normalization over that utterance. Experiments using

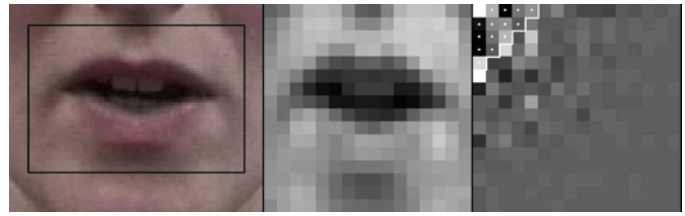

Fig. 4. From left to right: original lip image, subsampled 16×16 ROI, and DCT output showing 5×5 triangle coefficient selection.

only visual features for digit recognition showed that with the feature set described above, the word error rate (WER) was 26.15% without any normalization, 15.01% with just mean subtraction, and 12.11% with both mean substraction and variance normalization.

We used mel-frequency cepstral coefficients (MFCCs) to represent the features in the audio stream. The samples in the audio signal were grouped into 25 ms frames, with a between-frame overlap of 10 ms. For each frame the features were extracted using a 30 channel filter bank from which ten MFCCs were taken along with the energy feature. The corresponding Δ and $\Delta\Delta$ dynamic features of these were also calculated resulting in a feature vector of 33 coefficients.

D. Digit Modeling

Each digit type was modeled with an MSHMM model with ten states. The states were represented by Gaussian mixture models with four mixtures. As these parameters are heavily dependent on the domain and data being modeled, we tested a range of parameter values in a preliminary set of experiments and found that ten states and four mixtures gave almost perfect recognition results in clean audio experiments, and no improvement was gained from adding extra states or mixtures beyond this point.

MSHMMs can be trained separately, jointly, or using a fused approach, as described in [31]. In this paper, the MSHMMs were trained jointly. For all experiments, the models were trained using corruption-free audio and video data from the 200 subjects in the 3A set, and tested using both the clean and corrupt versions of the data from the remaining 95 subjects in the 3B set. In all experiments the prior probabilities $P(s)$ for all states are assumed equal.

E. Results Using Fixed Weights

The following experiments show how relative weighting of the audio and video streams effects recognition performance at different levels of corruption in the audio and video streams. The results indicate what the ideal weighting should be at every level of corruption, and shows the recognition performance that the MWSP model should aim to achieve if it successfully selects the appropriate weighting at each level of corruption.

For each weighting w, the likelihood of observing feature vector X_{AV} given state s is

$$p(X_{AV}|s) = p(X_A|s)^w \cdot p(X_V|s)^{1-w}. \tag{15}$$

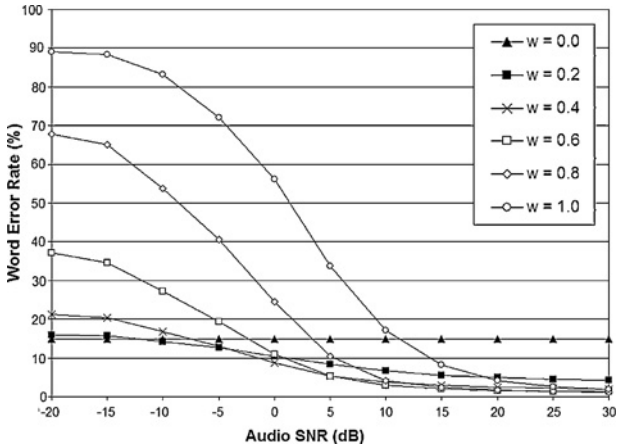

Fig. 5. Recognition performance using fixed stream weightings at different audio SNRs.

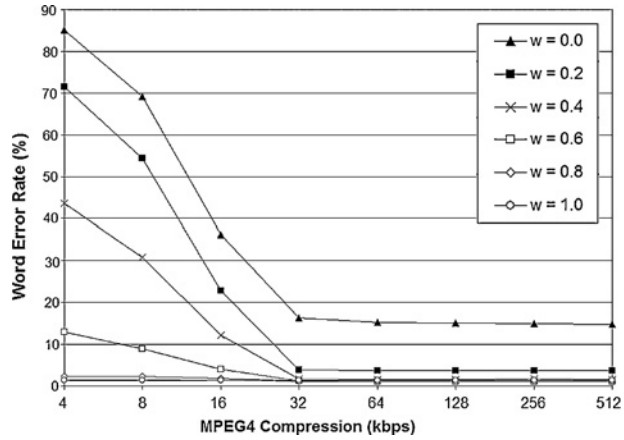

Fig. 6. Recognition performance using fixed stream weightings at different levels of video compression.

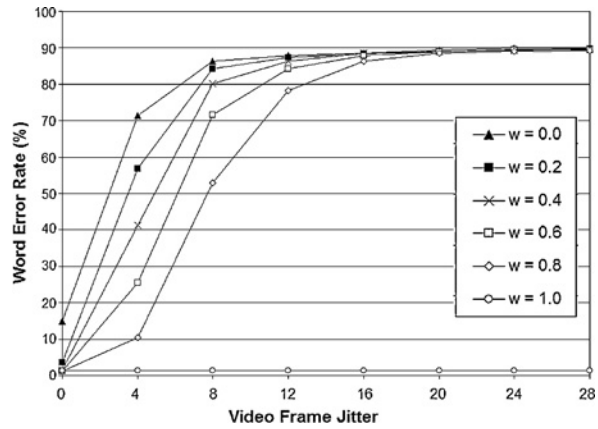

Fig. 7. Recognition performance using fixed stream weightings at different levels of video jitter.

Fig. 5 shows the recognition performance of each weighting for corruption in the audio stream. As one would expect, $w = 1.0$ is equivalent to the audio stream on its own and $w = 0.0$ is equivalent to the video stream on its own. For the lowest level of audio corruption (30 dB SNR) the WER of the audio features alone was 1.89% and the WER of the Product model was 1.50%. The weighting $w = 0.8$ achieved a WER of 1.27% that is better than either of those. The only SNR at which the Product model gave a lower WER than all of the six weightings is 5 dB. One assumes that at this SNR, the optimal weighting would be somewhere between 0.4 and 0.6 (i.e., close to 0.5, which is approximately equivalent to the Product model). At -10 db, it can be seen that the $w = 0.2$ weighting outperformed the video stream on its own ($w = 1.0$), which shows that even at this very low SNR, the audio stream can still contribute useful information.

Fig. 6 shows how each of the weightings performed when the video stream is compressed. As in the previous case of audio corruption, at the lowest level of corruption (the 512 kb/s compression bitrate), the weighting of $w = 0.8$ gave a lower WER (1.14%) than the audio features alone (1.43%) or the Product model (1.29%). Only at 16 kb/s did the audio stream on its own (i.e., $w = 1.0$) achieve the lowest WER. The effect of adding jitter to the video data is shown in Fig. 7. Clearly jitter is a much more destructive form of corruption than compression with regard to visual speech recognition when using DCT type features. For all levels of jitter the lowest WER was achieved with $w = 1.0$.

These results demonstrate that an optimized soft weighting scheme should outperform a fixed weighting scheme such as the Product model when the noise conditions in either modality can vary.

F. Selecting Number of Weights for MWSP

An important variable for the MWSP model is the number of values of w, i.e., the number of weightings. Fig. 8 shows how increasing the number of weightings affects speech recognition performance at different SNR levels of full-band Gaussian noise. In these experiments, the set of weights in each case are evenly distributed between (0.0,1.0) and (1.0,0.0). The differences are small between the models, although there is a significant difference between three weightings and other values. It can be seen that three weightings perform best at the most corrupted level (-20 dB SNR), but not as well for moderate levels of corruption. This may be due to the fact that with only three weightings, it is more likely to select the clean video stream on its own. For the least corrupted audio data (20 dB and 30 dB SNR), the difference between different numbers of weightings is negligible.

The optimal number of weightings for all our subsequent experiments was decided to be 6 as the model with six weightings performed strongly at all SNRs. The number of weightings is proportional to the processing time required to execute the Viterbi algorithm part of the speech recognition system. The marginal recognition performance gains of higher numbers of weightings do not justify the significant increases in processing time.

G. Noisy Audio Conditions

Having selected the optimal number of weightings, the MWSP can now be tested in various noise conditions to assess its stability and robustness. For all our experiments we have included results for the Product model, PUM, MSP model and the MWSP model for comparison. In this experiment, we also

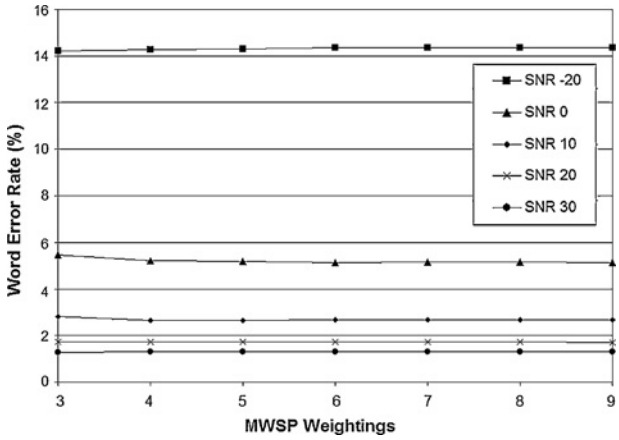

Fig. 8. MWSP recognition performance with different numbers of weightings at different audio SNRs.

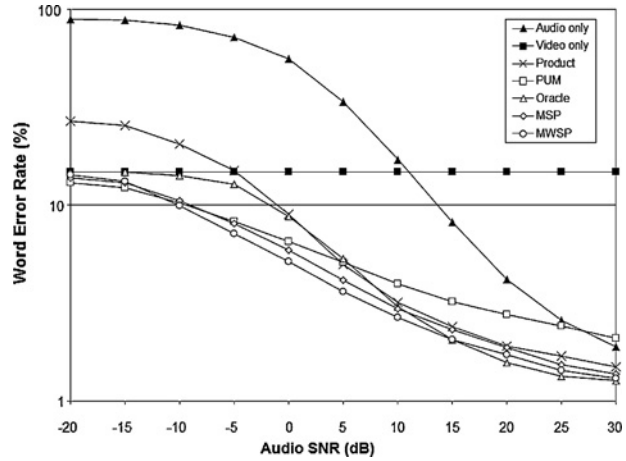

Fig. 9. Recognition performance (WER) of the MWSP compared to other models at different audio SNRs.

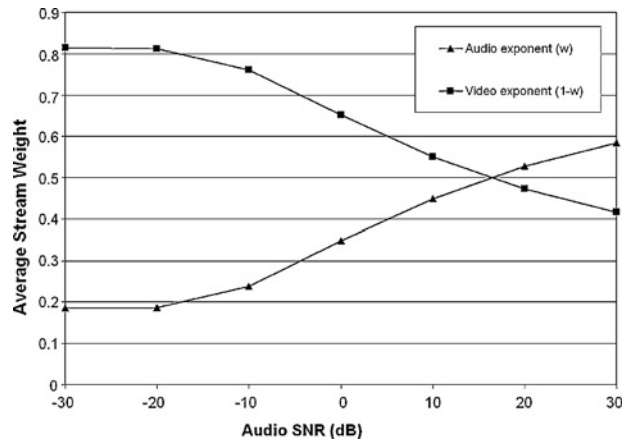

Fig. 10. Average recorded weighting values used by MWSP for different levels of audio SNR.

show the results obtained when the best fixed-weighted model for each noise level is used. We call this the Oracle model as it assumes full *a priori* knowledge of the noise in the streams, and hence, uses the most suitable weight. Fig. 9 shows how the methods compared in the case of corrupted audio. The log scale has been used on the y-axis to improve clarity between the lines.

It can be seen that all of the models except the Product model meet our criteria for ideal integration in that they perform better than either stream on their own in clean conditions and also remain at least as effective as the remaining clean modality in extreme noise conditions. The Product model gives good performance in relatively low noise conditions but becomes very unreliable as the noise level increases. Interestingly, both the MWSP and MSP models outperformed the Oracle model on average in these tests. This indicates that the ability of these models to select appropriate weights on a frame-by-frame basis is of benefit compared to static weighted approaches. The relative reliability of the streams is clearly not only affected by the level of corruption that is present in the stream, but also depends on the information content of the current frame, i.e., in some frames audio may naturally provide more discriminant information than video and vice versa. These results provide a strong justification for the need for a dynamic stream weighting approach even in clean conditions.

Both the MWSP and MSP performed significantly better than the PUM in clean conditions and moderately noisy conditions and performed similarly well in the most corrupt conditions. The MWSP approach was on average and in most test cases the best model that shows that it's softer weighting scheme is of benefit compared to MSP.

In order to verify that the MWSP model was selecting appropriate values of w for different levels of corruption, an experiment was performed to measure the average w value used at different levels of audio SNR. After the Viterbi component of the speech recognition system had selected the most likely state for each frame in an utterance, the system retrieved the value of w that generated the highest value of $P_w(s|X)$ for that state (i.e., the value of w that generated the final $P(s|X)$). These w values were transcribed for every single frame in the test data for a given level of corruption, and finally the average value of w was calculated over the entire test data set.

Fig. 10 shows the average recorded values of w (the weight applied to the audio stream) for each level of audio SNR, plotted against $1 - w$ that is the average weight applied to the video stream. For the cleanest audio (30-dB SNR) the average values of w and $1 - w$ were 0.58 and 0.42, respectively. At the most corrupted level, these values become 0.19 and 0.81. Overall, the graph shows that the value of w smoothly adapted to different levels of corruption in one stream that is the desired outcome for the MWSP approach. Fig. 11 further illustrates this by showing the actual distribution of weights used by MWSP for different levels of audio SNR. It can be seen that the value of w is not static at each SNR for all test frames. w takes on the full range of possible values even though the noise remains static over all the frames. The audio stream is dominant in cleaner audio conditions, but the video stream becomes progressively more dominant as the audio noise level increases. This highlights the fact that the weights used in the MWSP are affected by the noise level in the streams in combination with the naturally fluctuating relative reliability of the modalities.

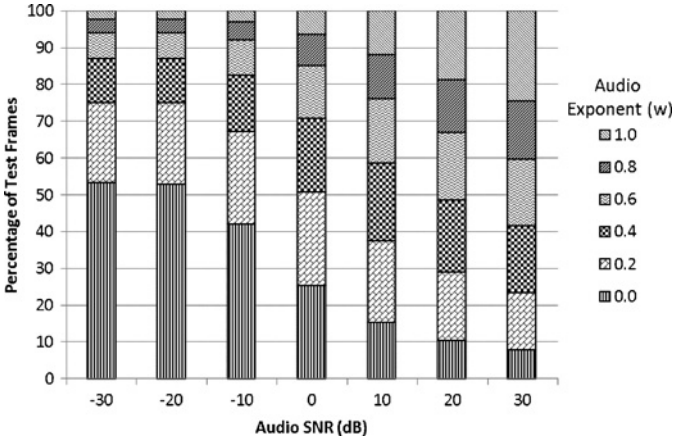

Fig. 11. Distribution of recorded audio weight values used by MWSP for different levels of audio SNR.

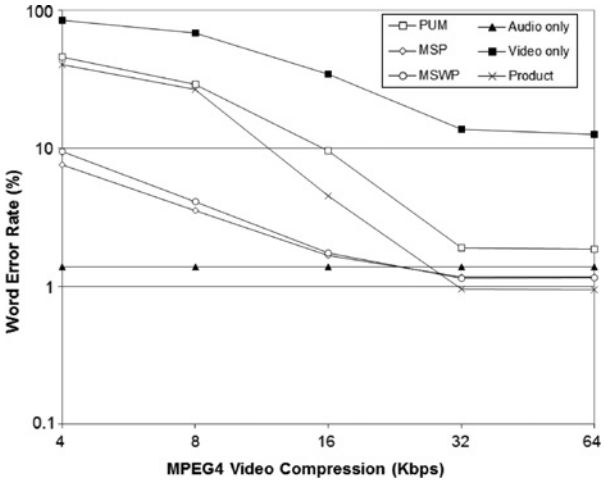

Fig. 12. Recognition performance (WER) of the MWSP compared to other models at different levels of video compression.

H. Noisy Video Conditions

The effect of compressing the video data is shown in Fig. 12. There was a negligible difference between the MSP and MWSP for all compression levels above 8 kb/s. At the high levels of compression (4 kb/s and 8 kb/s), the MSP performed slightly better than the MWSP. In all conditions, they both perform significantly better than the PUM. They also maintain performance much closer to the remaining clean modality than with the Product model at high compression levels. Therefore, in applications where the video data would be streamed in a compressed form before recognition is performed, i.e., for cloud-based speech recognition, the MWSP or MSP approach would allow much greater compression rates to be applied, while allowing high recognition rates to be maintained.

Fig. 13 shows the effect of adding jitter to video data along with the clean audio data. For all levels of jitter there was a negligible difference between the MWSP and MSP approaches, which maintain excellent performance in line with the remaining clean audio stream. They perform better than both the Product model and PUM in all conditions.

A further experiment was carried out to assess the effect of jitter corruption along with audio data that was corrupted to 15-dB SNR. The results are shown in Fig. 14. It can be seen

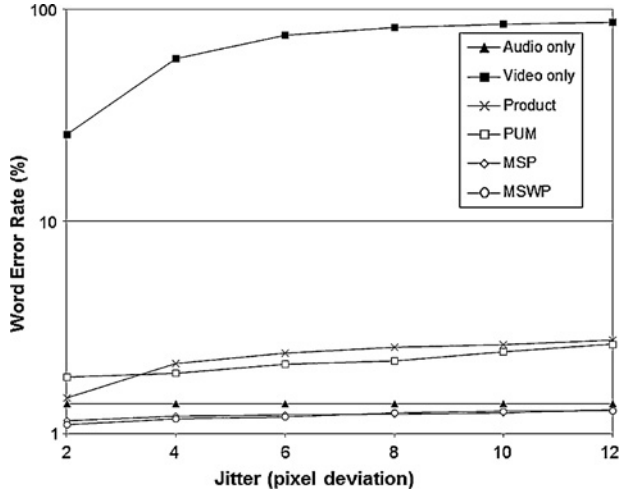

Fig. 13. Recognition performance (WER) of the MWSP compared to other models at different levels of video jitter.

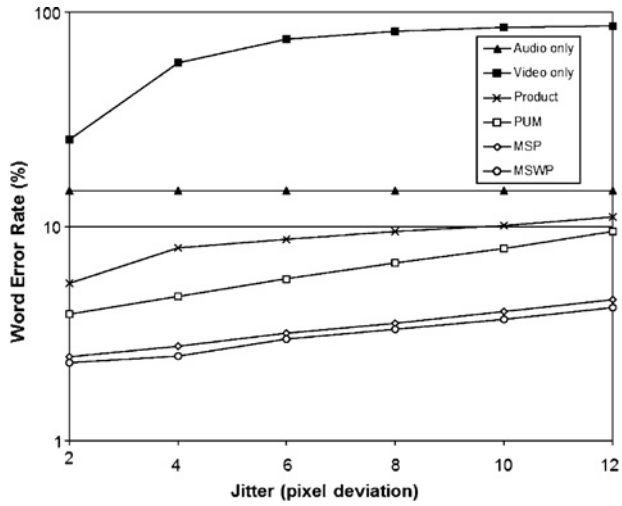

Fig. 14. Recognition performance (WER) of the MWSP compared to other models at different levels of video jitter along with corrupt audio at 15-db SNR.

that when the reliability of the audio stream is decreased in this way the performance of the MWSP and MSP approaches are separated slightly. The MWSP performs slightly better in all conditions than the others and the Product model is the least robust although it still maintains performance above the audio only approach.

The time taken for each recognition test was recorded to enable comparisons of the system performance of each model. These showed that compared to an equivalent Product model, the MSP required on average 18% more processing time, and the MWSP (with six weightings) required approximately 50% more. This shows that most of the processing time is spent calculating $p(X_A|s)$ and $p(X_V|s)$, rather than calculating the combinations used by the MSP and MWSP models.

IV. Summary and Conclusion

This paper dealt with the problem of audio-visual speech integration given that the relative reliabilities of the two modalities may fluctuate due to corruption in either modality,

e.g., due to environmental audio noise or perhaps due to video compression. We started by proposing a set of ideal operating characteristics of a robust integration strategy. These stated, first, that it should allow performance that is greater than either stream on its own for low levels of corruption in either stream, and second, in cases where one stream is highly corrupted the recognition performance should remain similar to the most reliable stream. These criteria should be met even in conditions where the corruption in either stream may be time varying and involves *a priori* unknown types of corruption.

Previous methods of tackling this problem based on decision fusion usually require an estimation of the noise level in each modality that can be difficult to measure accurately in real-world conditions. In this paper, we described a novel approach called the MWSP model that does not require any specific measurement of the noise level in the signal of either modality. As MWSP is modality-independent it complements the previous research in this area and could be used as an alternative or perhaps alongside other approaches. MWSP has been shown in experiments to offer a smooth integration of the two modalities, making best use of the available reliable information on a frame-by-frame basis and remaining simple to implement without *a priori* knowledge of the environmental conditions in which it will be deployed or tested.

The extensive experiments using the large XM2VTS database showed that the MWSP model provided significantly improved performance compared to another well-known dynamic weighting approach and also outperformed even an Oracle fixed-weighted approach where the noise levels and ideal weights are known *a priori*. This was demonstrated for both clean and noisy conditions in either or both modalities. Our experiments include tests on MPEG-4 compressed video data in a bid to simulate some of the effects of performing AVSR in a cloud-based framework where the data would be compressed and streamed before being processed. Our results showed that the MWSP approach would allow greater compression rates to be applied, while maintaining significantly better recognition performance. From our experiments, we would suggest that the MWSP model offers a suitable integration approach for robust audio-visual speech recognition in many real-world operating conditions.

REFERENCES

[1] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, no. 5588, pp. 746–748, 1976.

[2] A. Adjoudani and C. Benôit, "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. Berlin, Germany: Springer, 1996, pp. 461–471.

[3] T. Chen, "Audiovisual speech processing. lip reading and lip synchronization," *IEEE Signal Process. Mag.*, vol. 18, no. 1, pp. 9–21, Jan. 2001.

[4] G. Potamianos, J. Luettin, and C. Neti, "Hierarchical discriminant features for audio-visual LVCSR," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, May 2001, pp. 165–168.

[5] R. Goecke, G. Potamianos, and C. Neti, "Noisy audio feature enhancement using audio-visual speech data," in *Proc. Int. Conf. Acoust. Speech Signal Process.*, 2002, pp. 2025–2028.

[6] J. Luettin, G. Potamianos, and C. Neti, "Asynchronous stream modeling for large vocabulary audio-visual speech recognition," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, May 2001, pp. 169–172.

[7] M. Wollmer, F. Eyben, B. Schuller, and G. Rigoll, "A multi-stream ASR framework for BLSTM modeling of conversational speech," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, May 2011, pp. 4860–4863.

[8] J. Huang and K. Visweswariah, "Improved decision trees for multi-stream HMM-based audio-visual continuous speech recognition," in *Proc. Workshop IEEE Autom. Speech Recognit. Understanding*, Nov. 2009, pp. 228–231.

[9] R. Rajavel and P. S. Sathidevi, "A novel algorithm for acoustic and visual classifiers decision fusion in audio-visual speech recognition system," *Signal Process. Int. J.*, vol. 4, no. 1 pp. 23–37, 2010.

[10] V. Estellers, M. Gurban, and J. Thiran, "On dynamic stream weighting for audio-visual speech recognition," *IEEE Trans. Audio Speech Language Process.*, vol. 20, no. 4, pp. 1145–1157, May 2012.

[11] D. B. Dean, P. J. Lucey, S. Sridharan, and T. J. Wark, "Fused HMM-adaptation of multi-stream HMMS for audio-visual speech recognition," in *Proc. 8th Annu. Conf. Int. Speech Commun. Assoc.*, 2007, pp. 666–669.

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

[13] A. V. Nefian, L. Liang, X. Pi, X. Liu, and K. Murphy, "Dynamic Bayesian networks for audio-visual speech recognition," *EURASIP J. Appl. Signal Process.*, pp. 1274–1288, Nov. 2002.

[14] M. Heckmann, F. Berthommier, and K. Kroschel, "Noise adaptive stream weighting in audio-visual speech recognition," *EURASIP J. Appl. Signal Process.*, vol. 11, pp. 1260–1273, Nov. 2002.

[15] M. Heckmann, F. Berthommier, and K. Kroschel, "Optimal weighting of posteriors for audio-visual speech recognition," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, vol. 1. May 2001, pp. 161–164.

[16] L. Terry, D. Shiell, and A. Katsaggelos, "Feature space video stream consistency estimation for dynamic stream weighting in audio-visual speech recognition," in *Proc. 15th IEEE Int. Conf. Image Process.*, Oct. 2008, pp. 1316–1319.

[17] X. Zhang, C. C. Broun, R. M. Mersereau, and M. A. Clements, "Automatic speechreading with applications to human-computer interfaces," *EURASIP J. Appl. Signal Process.*, vol. 11, pp. 1228–1247, Nov. 2002.

[18] S. Nakamura, H. Ito, and K. Shikano, "Stream weight optimization of speech and lip image sequence for audiovisual speech recognition," in *Proc. Int. Conf. Spoken Language Process.*, vol. 3. 2000, pp. 20–23.

[19] G. Gravier, S. Axelrod, G. Potamianos, and C. Neti, "Maximum entropy and MCE based HMM stream weight estimation for audio-visual ASR," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, vol. 1. May 2002, pp. 853–856.

[20] A. Garg, G. Potamianos, C. Neti, and T. S. Huang, "Frame-dependent multi-stream reliability indicators for audio-visual speech recognition," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, vol. 1. Apr. 2003, pp. 24–27.

[21] R. Seymour, D. Stewart, and J. Ming, "Audio-visual integration for robust speech recognition using maximum weighted stream posteriors," in *Proc. Interspeech*, 2007, pp. 654–657.

[22] J. Schalkwyk, D. Beeferman, F. Beaufays, B. Byrne, C. Chelba, M. Cohen, M. Kamvar, and B. Strope, "'Your word is my command': Google search by voice: A case study," in *Advances in Speech Recognition: Mobile Environments, Call Centers and Clinics*, 2010, ch. 4, pp. 61–90.

[23] R. Seymour, J. Ming, and D. Stewart, "A new posterior based audio-visual integration method for robust speech recognition," in *Proc. Interspeech-Eurospeech*, Sep. 2005, pp. 1229–1232.

[24] J. Ming, J. Lin, and F. J. Smith, "A posterior union model with applications to robust speech and speaker recognition," *EURASIP J. Applied Signal Process.*, Apr. 2006, pp. 1–12.

[25] J. Lin, J. Ming, and D. Crookes, "Robust face recognition using posterior union model based neural networks," *Comput. Vision, IET*, vol. 3, no. 3, pp. 130–142, Sep. 2009.

[26] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre, "XM2VTSDB: The extended M2VTS database," in *Proc. Audio video-Based Biometric Person Authentication*, Mar. 1999, pp. 72–77.

[27] S. Young. (2000). *The HTK Book (for HTK Version 3.0)*, Microsoft Corporation [Online]. Available: http://htk.eng.cam.ac.uk/docs/docs.shtml

[28] M. U. R. Sanchez, "Aspects of facial biometrics for verification of personal identity," Ph.D. dissertation, Univ. Surrey, Guilford, U.K., 2000.

[29] G. Potamianos, H. P. Graf, and E. Cosatto, "An image transform approach for HMM based automatic lipreading," in *Proc. Int. Conf. Image Process.*, vol. 3. 1998, pp. 173–177.

[30] R. Seymour, D. Stewart, and J. Ming, "Comparison of image transform-based features for visual speech recognition in clean and corrupted videos," *EURASIP J. Image Video Process.*, vol. 2008, article 14, Apr. 2008.

[31] D. Dean and S. Sridharan, "Dynamic visual features for audio-visual speaker verification," *Comput. Speech Language*, vol. 24, no. 2, pp. 136–149, 2010.

[32] M. Heckmann, K. Kroschel, C. Savariaux, and F. Berthommier, "DCT-based video features for audio-visual speech recognition," in *Proc. Int. Conf. Spoken Language Process.*, Denver, CO, USA, Sep. 2002, pp. 1925–1928.

[33] D. B. Dean, T. J. Wark, and S. Sridharan. (2006). "An examination of audio-visual fused HMMS for speaker recognition," in *Proc. 2nd Workshop Multimodal User Authentication*, Toulouse, France [Online]. Available: http://eprints.qut.edu.au/5343/

Darryl Stewart received the B.Tech. degree in mechanical engineering from the University of Ulster, Ulster, U.K., in 1995, and the M.Sc. and Ph.D. degrees in computer science from the Queen's University of Belfast, Belfast, U.K., in 1996 and 2000, respectively.

He became a Lecturer at the Queen's University of Belfast in 2000. His current research interests include multimodal speech and speaker recognition.

Rowan Seymour received the B.Sc. and Ph.D. degrees in computer science from the Queen's University of Belfast, Belfast, U.K., in 2003 and 2007, respectively.

He is currently a Consultant with the International Training and Education Center for Health, University of Washington, Seattle, WA, USA.

Adrian Pass received the B.Eng. degree in acoustics from the University of Salford, Salford, U.K., in 2008, and the Ph.D. degree in computer science from the Queen's University of Belfast, Belfast, U.K., in 2012.

He is currently an Embedded Software Engineer with Pace plc, Saltaire, U.K.

Ji Ming (M'97) received the B.Sc. degree from Sichuan University, Chengdu, China, in 1982, the M.Phil. degree from the Changsha Institute of Technology, Changsha, China, in 1985, and the Ph.D. degree from the Beijing Institute of Technology, Beijing, China, in 1988, all in electronic engineering.

He was an Associate Professor with the Department of Electronic Engineering, Changsha Institute of Technology, from 1990 to 1993. Since 1993, he has been with the Queen's University Belfast, Belfast, U.K., where he is currently a Professor with the School of Electronics, Electrical Engineering, and Computer Science. From 2005 to 2006, he was a Visiting Scientist at the Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA, USA. His current research interests include speech processing, image processing, signal processing, and pattern recognition.