Integration Strategies for Audio-visual Speech Processing: Applied to Text Dependent Speaker Recognition

Simon Lucey+, Tsuhan Chen*, Sridha Sridharan+ and Vinod Chandran+
+Speech Research Laboratory, RCSAVT
School of Electrical and Electronic Systems Engineering
Queensland University of Technology
GPO Box 2434, Brisbane QLD 4001, Australia
*Advanced Multimedia Processing Laboratory
Department of Electrical and Computer Engineering
Carnegie Mellon University
Pittsburgh PA 15213, USA
slucey@ieee.org, tsuhan@cmu.edu, s.sridharan@qut.edu.au and v.chandran@qut.edu.au

ABSTRACT

In this paper an in depth analysis is undertaken into effective strategies for integrating the audio-visual speech modalities with respect to two major questions. Firstly, at what level should integration occur? Secondly, given a level of integration how should this integration be implemented? Our work is based around the well known hidden Markov model (HMM) classifier framework for modelling speech. A novel framework for modelling the mismatch between train and test observation sets is proposed, so as to provide effective classifier combination performance between the acoustic and visual HMM classifiers. From this framework, it can be shown that strategies for combining independent classifiers, such as the weighted product or sum rules, naturally emerge depending on the influence of the mismatch. Based on the assumption that poor performance in most AVSP applications can be attributed to train/test mismatches we propose that the main impetus of practical audio-visual integration is to dampen the *independent* errors, resulting from the mismatch, rather than trying to model any bimodal speech *dependencies*. To this end a strategy is recommended, based on theory and empirical evidence, using a hybrid between the weighted product and weighted sum rules in the presence of varying acoustic noise for the task of text-dependent speaker recognition.

Keywords: a udio-visual speech processing, speaker recognition, integration strategies, multistream HMM, classifier combination.

©Corresponding author Simon Lucey

Contents

List of Tables

List of Figures

1. INTRODUCTION

The automatic recognition of a claimant speaker over a set of impostor speakers, has been traditionally an application of great importance. Text dependent applications for the task of speaker recognition typically outperform their text independent counter parts due to the simplification of the recognition task [1]. In a text-dependent application, the recognition system has prior knowledge of the text to be spoken and it is expected that the user will cooperatively speak this text. In this paper the usefulness of the visual speech modality, particularly the mouth, is investigated for the task of isolated word, text dependent, speaker recognition paying special attention to strategies for effectively integrating the acoustic and visual modalities.

The usefulness of the visual modality in human speech is now well understood [2–5] and plays a very important role in both speech perception and production as demonstrated by the McGurk effect [4]. However, the effective integration of the acoustic and visual modalities of speech has still remained an open question in audio-visual speech recognition and text dependent speaker verification. Two questions are posed in this paper. The first, is at what level should the acoustic and visual speech modalities be integrated, for isolated word text-dependent speaker recognition? The second, given a level of integration, how best can one combine those modalities taking into account the practical limitations of the classifiers being used for the recognition task? Unfortunately, these two questions cannot be answered separately as the result of one heavily influences the result of the other.

The first question concerns itself with how the human brain takes advantage of the complementary nature of audio-visual speech and integrates the two information sources. There is some contention over terminology in AVSP [6], pertaining to the different levels of integration possible. It is widely agreed [2, 6, 7] however, that the acoustic and visual modalities can be combined either at the feature or the decision level. These two differing integration paradigms commonly go under the guise of early integration (EI) and late integration (LI), but the interpretation of what EI and LI strictly are can vary markedly depending on perspective and the task at hand. In this paper we try to dispel some of this confusion by trying to define the different levels of audio-visual integration in terms of the methods used to train and test the practical audio-visual classifier, with the aim that different integration levels will naturally emerge. Multi-stream [6–9] hidden Markov models (HMM) are employed in this paper to evaluate a third level on integration, namely middle integration (MI). MI allows for a varying degree of temporal dependence between the acoustic and visual modalities during testing, whilst still allowing the benefits of training the modalities independently. Although distinct from EI and LI, MI can be thought equivalent to LI for some specific cases.

The second question, of how best to integrate given a pre-defined level of integration, delves into practical dilemmas associated with the combination of classifiers and how different combination strategies have to be undertaken depending on the train/test mismatches occurring in either modality. Hidden Markov model (HMM) classifiers are used for our experimental work due to their ability to stochastically model the temporal fluctuations present in both modalities of speech. HMMs are also able to naturally incorporate different levels of audio-visual integration.

Throughout this paper the term *train/test mismatch* will be used extensively. The difference between the train and test sets is referred to as a train/test mismatch. The measure of train/test mismatch is not the physical difference between the train and test observation sets but a measure of how generalized the knowledge (i.e. ability to make a correct decision) of the classifier gained from the train set is, with reference to the unknown test set. When a mismatch occurs in the testing set that differs from what has been seen in the training set this uncertainty should be represented in the confidence score, otherwise a confidence error will occur [10]. These confidence errors should not be confused with Bayesian error [11], which is inherent to the classification task. It has been well documented by Kittler et al. [10, 12] that when fusing the confidence scores of conditionally independent classifiers, where such confidence errors are *not* present, the product rule is optimal *if* the scores are representative of a posteriori probabilities. However, when confidence errors are present the compounding effect of these errors, when classifiers are combined, must be taken into account as the blind application of the product rule may result in *catastrophic fusion*. Movellan [13] defines catastrophic fusion as the combination of an ensemble of classifiers that results in performance that is worse than the performance of those classifiers individually.

There are two options available to us to try and lessen the compounding effect of these confidence errors. Ideally, one can try and adapt the classifier to the test utterances, thus removing the confidence error and allowing for

optimal combination through the normal product rule. Generally, this is impossible in practice as it requires a violation of causality (i.e. access to the test utterances before testing). In some circumstances it is possible to adapt a classifier to a specific mismatch given quantitative knowledge of the mismatch (e.g. acoustic cepstral features can be adapted to changes in additive acoustic noise [14] without a violation of causality), but often times it is impossible to adapt to all possible sources of mismatch (e.g. undertrained classifier). Alternatively, one can try and dampen, rather than adapt to, the effects of these confidence errors upon combination, through the judicious choice of combination strategies. This approach has a lot more appeal in practice, as it can be implemented without violating causality and may not require quantitative knowledge of the mismatch. From within this type framework the sum rule naturally emerges, which has been shown theoretically and empirically to be a more benevolent rule than the product rule when confidence errors are present [10]. Borrowing on classifier combination terminology from Kamel and Wanas et. al [15], we will be employing data dependent and independent combination functions; specifically the weighted product and sum rules respectively. Data dependent functions, like the weighted product rule, require some training and are of use in situations where one has prior quantitative knowledge of the environment that it will be tested under. Data independent functions, like the sum rule, do not require any training and are of use where there is minimal knowledge about testing conditions.

In this paper we aim to show that in a practical scenario, when train/test mismatches do occur in each modality, it is better to integrate the audio-visual modalities at the confidence score level. Based on the assumption that poor performance in most AVSP applications can be attributed to train/test mismatches we propose that the main impetus of such integration is to dampen these *independent* errors rather than trying to model any bimodal speech *dependencies*. Two different combination functions are investigated, namely the weighted product and weighted sum rules. A hybrid approach between the weighted product and sum rules is shown to give robust results in identification when being tested across a number of broad acoustic noise conditions.

This paper is broken into a number of sections. Section 3 formally describes the speaker recognition tasks of identification and verification as well as quantitative ways of measuring their performance. Section 4 describes the audio-visual database used for our experiments as well as the feature extraction techniques used for both modalities. In Section 5 an in depth look is taken at the hidden Markov model (HMM) classifiers being used for the identification and verification task. Differences in HMM topology and training strategies are described for the different integration levels as well an equivalence that exists between MI and LI in some circumstances. Combination strategies for the LI strategy are discussed in Section 6 with a firm theoretical grounding being set for the use of the weighted product and sum rules when train/test mismatches are encountered. Finally in Section 7 results and discussion are presented.

2. AUDIO-VISUAL INTEGRATION

There is some contention over terminology in AVSP [6] pertaining to the different levels of integration possible. It is widely agreed [2, 6, 7] however, that the acoustic and visual speech modalities can be combined either at the feature (i.e. EI) or the score level (i.e. LI). However, strict definitions for EI and LI still remain unclear in AVSP literature.

For example, in continuous audio-visual speech applications Dupont and Luettin [7] interpret LI as combining scores at the sentence level. However, in earlier literature [2, 16, 17], specifically for the tasks of isolated word recognition, LI is interpreted as combining scores at the word unit level. This is the interpretation adhered to in this paper, as it seems a moot point to consider combination at any level higher than this.

Similar ambiguity exists in defining EI. One hypothesis [2] for EI suggests that visual speech information is converted to a vocal tract function, where then the acoustic and visual transfer functions are averaged during integration. Alternatively EI can be interpreted [2, 7, 18] as the concatenation of acoustic and visual stimuli for processing as a single observation.

For the purposes of clarity this paper defines three broad levels of integration,

1. Early integration (EI), in which acoustic and visual speech stimuli are synchronized and merged in some manner (e.g. concatenation or averaging of vocal tract functions) for joint learning and classification. This

approach assumes there is direct dependence between the acoustic and visual modalities at the lowest levels of human speech perception.

2. Middle integration (MI), attempts to integrate the acoustic and visual speech modalities at a slightly higher level than EI. MI attempts to learn and classify acoustic and visual speech cues *independently*. However, during the classification of an utterance there may be *temporal* dependence between modalities.

3. Late integration (LI), assumes complete independence between the acoustic and visual speech modalities. During the classification process there is *no* interaction between the modalities with only the final classifier confidence scores being combined. In this approach temporal information between speech modalities is lost, except at the anchor point at which the decisions are being combined (i.e. word unit level).

A graphical depiction of these integration levels can be seen in Figure 1.

The distinction between EI, MI and LI is made in this paper specifically on how the audio-visual classifier is trained and tested. Dupont and Luettin [7] arrived at three similar levels of integration, although they referred to them as *model 0, 1 and 2*. It may be argued that there is a degree of overlap between our interpretation of MI and LI. This is a valid point, as there is an actual equivalence between MI and LI for some specific cases. However, the sometimes subtle differences between MI and LI, and their benefits, shall be elucidated further in this paper.

EI techniques for audio-visual speech [2, 16] and speaker [18] recognition have been previously used, and are of benefit as they model the dependencies between acoustic and visual speech modalities directly. However, EI approaches suffer in two respects. First, if the acoustic or visual speech modalities are corrupted then the entire speech modality is corrupted due to classification occurring at such a low level. Second, there is an assumption that the acoustic and visual speech modalities are synchronized at the state level when HMM classifiers are being employed.

Lavagetto [3] demonstrated that acoustic and visual speech stimuli are *not* synchronous, at least at a feature based level. It was shown that visible articulators, during an utterance, start and complete their trajectories asynchronously, exhibiting both forward and backward coarticulation with respect to the acoustic speech wave. Intuitively this makes a lot of sense, as visual articulators (i.e. lips, tongue, jaw) have to position themselves correctly before and after the start and end of an acoustic utterance. This time delay is known as the voice-onset-time (VOT) [7], which is defined as the time delay between the burst sound, coming from the plosive part of a consonant, and the movement of the vocal folds for the voiced part of a voiced consonant or subsequent vowel. McGrath et al. [19] also found an audio lead of less than 80ms or lag of less than 140ms could not be detected during speech. However, if the audio was delayed by more than 160ms it no longer contributed useful information, signifying the importance of *some* degree of asynchrony and synchrony in continuous audio-visual speech perception.

LI is able to largely circumvent these problems. For automated isolated word applications LI strategies have reported superior results to EI for speech [2, 16, 17, 20] and speaker recognition [1] tasks. LI allows for the asynchronous classification of speech and can emphasize or deemphasize the importance of a modality in classification depending on the corruption present. However, any static or temporal dependencies occurring between modalities is lost. As previously mentioned LI has not proven as effective in continuous speech applications [7] where integration is attempted at greater than the word unit level. Waiting until the end of the spoken utterance before combining modalities, as the LI strategy was perceived by Dupont and Luettin [7], introduces an undesirable time delay. To this end some form of synchrony is required.

The question remains at what level should audio-visual speech be synchronized. MI allows for such synchrony whilst still providing a framework for guarding against corruption in either modality. MI based approaches have been used to great success in continuous audio-visual speech applications [7, 8, 20]. However the benefit of MI over LI, if LI is constrained to be synchronized at the word unit level, is still not clear and is the topic of further investigation in this paper.

3. SPEAKER RECOGNITION

Speaker recognition encapsulates two tasks, namely identification and verification. Speaker identification is the task of selecting the most likely speaker ω_{i^*} from a group of N known speakers for an observation utterance $\mathbf{O}$ such that,

$$i^* = \arg\max_{i=1}^{N} \zeta(\omega_i|\mathbf{O}) \tag{1}$$

where $\zeta(\omega_i|\mathbf{O})$ is the confidence score describing how likely the utterance $\mathbf{O}$ belongs to speaker ω_i. Speaker identification performance is normally evaluated in terms of identification rate, the ratio of correct classifications over total classifications, in a given test set.

The speaker verification task is the binary process of accepting or rejecting the identity claim made by a subject under test. The verification process can be expressed simply as the decision rule,

$$\zeta(\omega_{claim}|\mathbf{O}) \underset{\text{accept}}{\overset{\text{reject}}{\lessgtr}} Th \tag{2}$$

where $\zeta(\omega_{claim}|\mathbf{O})$ is the confidence score describing how likely utterance $\mathbf{O}$ belongs to the claimant speaker ω_{claim}. A threshold Th needs to be found so as to make the decision. Speaker verification performance is evaluated in terms of two types of error being false rejection (FR) error, where a true client speaker is rejected against their own claim, and false acceptance (FA) errors, where an impostor is accepted as the falsely claimed speaker. The FA and FR errors increase or decrease in contrast to each other based on the decision threshold Th set within the system. A simple measure for overall performance of a verification system is found by determining the equal error rate (EER) for the system. This is the operating point where the FA and FR error rates are equal. A detection error tradeoff (DET) [21] curve, similar to a receiver operating characteristic (ROC) [22] curve, can also be used to represent the trade off between the two errors for a varying threshold.

4. AUDIO-VISUAL DATABASE, MOUTH DETECTION AND FEATURE EXTRACTION

The M2VTS database [23] was used for experiments in this paper. Out of the possible 37 subjects in the database the subject 'pm' was excluded from testing, due to his beard which was thought to unfairly skew the verification results. This was due to the bearded subject never getting incorrectly identified in the visual modality, as his appearance was completely different from the other 36 subjects. The M2VTS database has been used in previous multimodal speaker recognition experiments [1, 24]. The database used for our experiments consisted of 36 subjects (male and female) speaking four repetitions (shots) of ten French digits from *zero* to *nine*. The database was separated into train and test sets for audio-visual classifier training and testing. Shots one to three were used for training with shot four being used for testing. A subject's mouth was tracked through a video sequence by first segmenting the face from its background using chromatic segmentation. Through a multi-scale search the eyes are then detected, to gain a measure of face scale. Finally the mouth is detected and tracked throughout the visual sequence. The tracked mouth coordinates are then smoothed using a median filter to remove any spurious detection results. Across the entire M2VTS database the mouth was tracked accurately to within a couple of pixels of its true position. The algorithm used to detect the eyes and mouth was based on an unsupervised intra-class clustering approach using discriminant analysis. More details on our facial feature detection/tracking approach can be found in [25].

The mouth region of interest (ROI) chosen for tracking was based on the subject's eye separation distance d_{eye}, with a $(3d_{eye}) \times (4d_{eye})$ box centered at the mouth center. Visual features were extracted by first obtaining the first 50 principal components of the mouth ROI images from the training set of all speakers, in the train set, using principal component analysis (PCA) [11]. Linear discriminant analysis (LDA) [11] was then employed to further reduce the dimensionality of the visual feature set down to the 10 most linear discriminating components (using all 36 speaker classes in the train set). Delta coefficients were included for the visual features thus expanding the final visual feature vector to 20 dimensions. For the acoustic features we used mel-frequency cepstral coefficients (MFCC) with mean cepstral subtraction and delta coefficients to create a 26 dimensional feature vector [26].

5. HIDDEN MARKOV MODELS, TRAINING AND INTEGRATION STRATEGIES

Hidden Markov models (HMMs) were used to model audio-visual utterances using HTK ver 2.2 [26]. The M2VTS database was employed for all experiments. The first three shots of the M2VTS database were used to train the audio-visual HMMs with shot four being used for testing. HMMs are excellent for modelling bimodal speech as they provide a natural way to stochastically capture the temporal fluctuations of speech in each modality and are able to naturally incorporate the three levels of integration (ie. EI, MI and LI) mentioned previously into their topology.

All HMMs were trained using the Baum Welch algorithm via the HTK [26] package. Two models were acquired for each digit: the speaker dependent model $p(\mathbf{O}|\boldsymbol{\lambda}_i)$, and the background model $p(\mathbf{O}|\boldsymbol{\lambda}_{bck})$. The latter, which is common to all subjects, captures the variability of the uttered sound. Due to the relatively small size of the M2VTS database and the requirement for separate speaker dependent digit HMMs all speaker dependent HMM digit models were trained by initializing training with the previously found speaker independent or background digit model. This approach prevented variances in each model becoming too small and allows each model to converge to sensible values for the task of speaker recognition.

Training for the EI strategy involved the synchronization of the acoustic and visual features. Both acoustic and visual features were concatenated into one feature vector, which was used to train a single joint audio-visual HMM. Via an exhaustive search a 3 state, 2 mixture HMM topology received best results for both modalities. For the MI and LI strategies, two separate independent acoustic and visual HMMs were trained. Again via an exhaustive search, a 3 state, 2 mixture topology was selected for the independently trained acoustic and visual HMMs.

The EI, MI and LI integration strategies require the computation of a likelihood. Considering the general case when one wishes to classify an utterance $\mathbf{O}$ given by,

$$\mathbf{O} = \{\mathbf{o}_1, \ldots, \mathbf{o}_T\} \tag{3}$$

where T is the number of observations and $\mathbf{o}_t$ denotes the feature vector for observation t. The likelihood $p(\mathbf{O}|\boldsymbol{\lambda}_i)$, for HMM parametric class (digit/speaker) model $\boldsymbol{\lambda}_i$ given the utterance $\mathbf{O}$, is found canonically by expanding all possible state paths,

$$p(\mathbf{O}|\boldsymbol{\lambda}_i) = \sum_{\text{all } \mathbf{q}} p(\mathbf{O}, \mathbf{q}|\boldsymbol{\lambda}_i) \tag{4}$$

where the sum extends over all possible paths $\mathbf{q}$. This full likelihood is estimated in practice using the Viterbi approximation [26] which only requires a single path,

$$\log p(\mathbf{O}|\boldsymbol{\lambda}_i) \approx \frac{1}{T} \log p(\mathbf{O}, \mathbf{q}^*|\boldsymbol{\lambda}_i) \tag{5}$$

where $\mathbf{q}^* = \{q^*(1), \ldots, q^*(T)\}$ is the optimal state path. The optimal path $\mathbf{q}^*$ is found in practice via the Viterbi decoding algorithm [26, 27].

An a posteriori probability estimate for modality $m \in \{a, v, av\}$, referring to the acoustic, visual and audio-visual modalities respectively, can be found through Bayes rule [11], assuming equal priors, from the estimated likelihoods such that,

$$\hat{Pr}(\omega_i|\mathbf{O}^{\{m\}}) = \frac{p(\mathbf{O}^{\{m\}}|\boldsymbol{\lambda}_i^{\{m\}})}{\sum_{n=1}^{N} p(\mathbf{O}^{\{m\}}|\boldsymbol{\lambda}_n^{\{m\}})} \tag{6}$$

It must be remembered that Equation 6 gives only an *estimate* of the a posteriori probability. This is due to the conditional class models, used in the evaluation of Equation 6, describing observations drawn from the train set $\mathcal{S}_{trn}$ *not* the test set $\mathcal{S}_{tst}$. Both the EI and MI strategies employ an audio-visual HMM during evaluation so Equation 6 can be used directly to obtain a audio-visual confidence score $\zeta(\omega_i|\mathbf{O}^{\{av\}}) = \hat{Pr}(\omega_i|\mathbf{O}^{\{av\}})$. The EI strategy obtains its audio-visual HMM through training directly from synchronized audio-visual features. The

MI strategy however, employs independently trained HMMs from the acoustic and visual modalities from which a composite audio-visual HMM is created known as a *multistream* HMM.

The LI strategy uses two independently trained HMMs from the acoustic and visual modalities for evaluation, so that it cannot employ Equation 6 directly to obtain an audio-visual confidence score. The confidence score for the recognition process can be expressed as a function of each modality's a posteriori probability estimate,

$$\zeta(\omega_i|\mathbf{O}^{\{av\}}) = F(\hat{Pr}(\omega_i|\mathbf{O}^{\{a\}}), \hat{Pr}(\omega_i|\mathbf{O}^{\{v\}})) \tag{7}$$

Throughout this correspondence we will be using $F()$ as compact notation for $F(\hat{Pr}(\omega_i|\mathbf{O}^{\{a\}}), \hat{Pr}(\omega_i|\mathbf{O}^{\{v\}}))$.

5.1. Multistream HMMs

Multistream HMMs use two separate independently trained streams (ie. HMMs) and combines them into a single HMM in such a way that one stream may have some temporal dependence on the other during decoding, without having to train both sequences together. There are two main ways to build a multistream HMM, namely synchronously or asynchronously [7], which we will refer to as multistream synchronous HMMs (MSHMMs) and multistream asynchronous HMMs (MAHMMs) respectively. Both methods can be thought of as a 2-D HMM as depicted in Figure 3.

Varga and Moore [9] first explored asynchronous multistream HMMs, or product HMMs as they were originally referred to, for the task of improving acoustic speech recognition through signal decomposition. Tomlinson et al. [28] first used asynchronous multistream HMMs for isolated word audio-visual speech recognition by modelling the acoustic and visual modalities via independently trained streams. In this work they showed there was considerable benefit in allowing the acoustic and visual streams to traverse the virtual 2-D state trellis asynchronously, as the acoustic stream had a tendency to cause misalignment.

Multistream HMMs can be used to naturally model the MI integration strategy as they provide relative independence between streams statically with a loose temporal dependence dynamically. To date there has been a limited amount of research done into the effect multistream HMMs have on isolated word text-dependent audio-visual speaker recognition. Additionally an equivalence between MI, specifically MAHMMs, and LI, using the product rule, can also be established based on previous work by Varga and Moore [9].

6. LI COMBINATION STRATEGIES

According to [10, 11] and Bayesian theory the optimal combination function $F()$ for conditionally independent error free a posteriori probabilities is the product rule,

$$\zeta^*_{pr}(\omega_i|\mathbf{O}^{\{av\}}) = F_{pr}() \doteq Pr(\omega_i|\mathbf{O}^{\{a\}})Pr(\omega_i|\mathbf{O}^{\{v\}})P(\omega_i)^{-1} \tag{8}$$

where $P(\omega_i)$ is the class a priori probability. It must be emphasized that $\zeta^*_{pr}(\omega_i|\mathbf{O})$ is a confidence score (not necessarily between zero and one) *not* a probability, but is equivalent to the true probability $Pr(\omega_i|\mathbf{O}^{\{a\}}, \mathbf{O}^{\{v\}})$ in terms of the class decision boundaries it realizes. For conditionally independent classifiers $\zeta^*_{pr}(\omega_i|\mathbf{O})$ gives an upper bound in classifier combination performance.

Equation 8 holds if one has access to the true a posteriori probabilities from conditionally independent observation modalities. In practice one can rarely apply this rule due to true a posteriori probabilities as the mismatch between train and tests sets results in a confidence error.

$$\hat{Pr}(\omega_i|\mathbf{O}^{\{m\}}) = Pr(\omega_i|\mathbf{O}^{\{m\}}) + \epsilon_i(\mathbf{O}^{\{m\}}) \tag{9}$$

One can only ever apply the product rule $F_{pr}()$ to the a posteriori probability estimates $\hat{Pr}(\omega_i|\mathbf{O}^{\{m\}})$, as the decision boundaries realized by the classifier is based on the train not the test set, resulting in the confidence score estimate $\zeta_{pr}(\omega_i|\mathbf{O}^{\{m\}})$. In the presence of large confidence errors the estimated confidence score $\zeta_{pr}(\omega_i|\mathbf{O})$ may realize different decision boundaries to the optimal $\zeta^*_{pr}(\omega_i|\mathbf{O})$ resulting in suboptimal classifier combination performance. We must again note that the notion of optimality only holds for $\zeta^*_{pr}(\omega_i|\mathbf{O})$ if we assume the true a

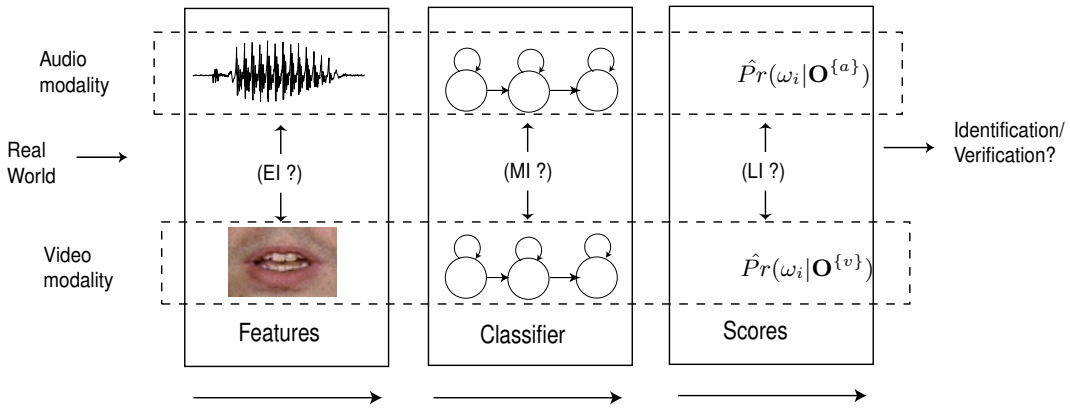

Figure 1. Depiction of possible levels of integration.

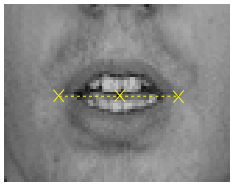

Figure 2. Example of how the center of the mouth is found from the bisection of the left and right corners of the mouth.

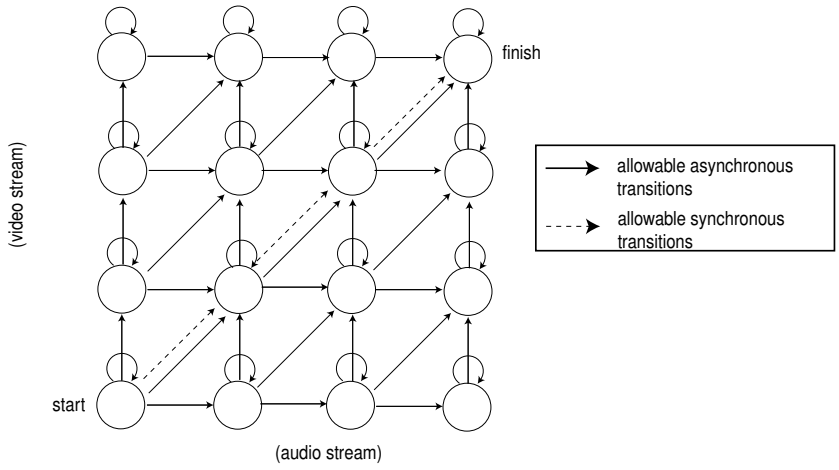

Figure 3. Example of 2D left to right HMM state lattice for asynchronous and synchronous decoding.

posteriori probabilities are conditionally independent. The task of effective classifier combination is to judiciously choose combination strategies $F()$ that alleviate the compounding effects of confidence errors. If due care is not taken to choose an appropriate combination function $F()$ catastrophic fusion may occur.

6.1. Modelling train/test mismatch

The idea of a train/test mismatch can be formally described if one analyzes the problem of determining an a posteriori probability in terms of sets. For ease of notation in this section the observation sequence $\mathbf{O}$ is deemed to stem from a single modality. The observation sequence $\mathbf{O}$ exists in the set $\mathcal{S}_{all}$ such that $\mathbf{O} \in \mathcal{S}_{all}$. At any given time one only has at their disposal observations existing in the subset $\mathcal{S}_{trn} \subset \mathcal{S}_{all}$ or $\mathcal{S}_{tst} \subset \mathcal{S}_{all}$, representing training and testing observation sets respectively. When one has to gain an a posterior probability estimate $\hat{Pr}(\omega_i|\mathbf{O})$ for class i of observation sequence $\mathbf{O} \in \mathcal{S}_{tst}\{i\}$, one has to make a decision based on knowledge gained from observations lying in $\mathcal{S}_{trn}\{i\}$ even though $\mathbf{O}$ may not. A depiction of this situation is shown in the Venn diagram in Figure 4(b) where $\mathcal{S}_{trn}$ and $\mathcal{S}_{tst}$ are subsets of $\mathcal{S}_{all}$. Within a Bayesian framework, one has to allow for the possibility that $\mathbf{O} \notin \mathcal{S}_{trn}$ even though $\mathbf{O} \in \mathcal{S}_{tst}$.

Using Bayes [11] rule, when different train/test conditions are encountered, one would ideally use likelihoods $p(\mathbf{O}|\mathcal{S}_{tst}\{i\})$[†] derived from our knowledge of the test set to gain an a posteriori probability,

$$Pr(\mathcal{S}_{tst}\{i\}|\mathbf{O}) = \frac{P(\omega_i)p(\mathbf{O}|\mathcal{S}_{tst}\{i\})}{\sum_{n=1}^{N} P(\omega_n)p(\mathbf{O}|\mathcal{S}_{tst}\{n\})} \tag{10}$$

However, due to causality the classifier's knowledge is always restricted to $\mathbf{O} \in \mathcal{S}_{trn}$ which should be reflected in the model thus giving,

$$p(\mathbf{O}|\omega_i) = \underbrace{P(\Omega)p(\mathbf{O}|\mathcal{S}_{trn}\{i\})}_{\mathcal{S}_{tst} \subseteq \mathcal{S}_{trn}} + \underbrace{P(\overline{\Omega})p(\mathbf{O}|\overline{\Omega})}_{\mathcal{S}_{tst} \nsubseteq \mathcal{S}_{trn}} \tag{11}$$

Equation 11 can be understood by using the concept of set dependent knowledge. There are two terms in Equation 11. The first term $P(\Omega)p(\mathbf{O}|\mathcal{S}_{trn}\{i\})$ represents the classifier's knowledge of discerning between classes when one is within the known train set (i.e. $\mathbf{O} \in \mathcal{S}_{trn}$), where $P(\Omega)$ is the prior of the observation coming from that known train set. The second term $P(\overline{\Omega})p(\mathbf{O}|\overline{\Omega})$ represents our knowledge for discerning between classes outside the train set (i.e. mismatch set). This term is the same for all classes, as the unadapted classifier has no knowledge for discerning between classes in the mismatch set. $P(\overline{\Omega})$ and $p(\mathbf{O}|\overline{\Omega})$ is the prior and likelihood of the observation sequence coming from the mismatch set respectively. Using this equivalence one can gain estimates of the error free a posteriori probabilities, using Bayes rule, by taking into account the likelihoods of all classes simultaneously,

$$Pr(\omega_i|\mathbf{O}) = \frac{P(\omega_i)\left[P(\Omega)p(\mathbf{O}|\mathcal{S}_{trn}\{i\}) + P(\overline{\Omega})p(\mathbf{O}|\overline{\Omega})\right]}{\sum_{n=1}^{N} P(\omega_n)\left[P(\Omega)p(\mathbf{O}|\mathcal{S}_{trn}\{n\}) + P(\overline{\Omega})p(\mathbf{O}|\overline{\Omega})\right]} \tag{12}$$

Under similar train/test conditions one can make the assumption,

$$P(\Omega)p(\mathbf{O}|\mathcal{S}_{trn}\{n\}) \gg P(\overline{\Omega})p(\mathbf{O}|\overline{\Omega}) \qquad 1 \leq n \leq N \tag{13}$$

which leads to the commonly used estimate,

$$\begin{aligned} \hat{Pr}(\omega_i|\mathbf{O}) &= \frac{P(\omega_i)p(\mathbf{O}|\mathcal{S}_{trn}\{i\})}{\sum_{n=1}^{N} P(\omega_n)p(\mathbf{O}|\mathcal{S}_{trn}\{n\})} \\ &\doteq Pr(\mathcal{S}_{trn}\{i\}|\mathbf{O}) \end{aligned} \tag{14}$$

Unfortunately, in practice it is infeasible to gain a model of $P(\overline{\Omega})$ and $p(\mathbf{O}|\overline{\Omega})$, as one requires intimate knowledge of $\mathcal{S}_{trn}$ and $\mathcal{S}_{tst}$ a priori. However, one can see that if Equation 14 is applied when Equation 13 does not hold

[†]To aid in the development of the train/test mismatch framework the HMM likelihood function $p(\mathbf{O}|\boldsymbol{\lambda}_i)$ shall be referred to as $p(\mathbf{O}|\mathcal{S}_{trn}\{i\})$.

(ie. in the case of external noise or an under trained classifier) the resultant a posteriori probabilities will be erroneous, due to the mismatch class being ignored. This results in a confidence error $\epsilon_i(\mathbf{O})$ as first mentioned in Equation 9.

Nothing can be done about this error $\epsilon_i(\mathbf{O})$ when dealing with a single modality with respect to classification error. However, by allowing for a confidence error in $\hat{Pr}(\omega_i|\mathbf{O})$ one can dampen the aggregate effect of these errors when used in a classifier combination scheme, whilst not violating causality.

This differs markedly to the confidence error free a posteriori probability $Pr(\mathcal{S}_{tst}\{i\}|\mathbf{O})$ realized from adapting the train set to the test set. This difference is apparent in the handling of the confidence error which is completely removed, as opposed to being transformed into Bayesian error when trying to dampen the compounding effects of the error when introducing the mismatch likelihood.

6.2. Weighted product rule

The use of an exponential weighting on the acoustic and visual a posteriori probability estimates, when using the product rule, has been shown [1, 2, 6, 7, 16] empirically to be of use in audio-visual speech and speaker recognition, especially in the presence of varying additive acoustic noise. This formulation, known as the weighted product rule, can be seen below in Equation 15,

$$F_{wpr}() \doteq \hat{Pr}(\omega_i|\mathbf{O}^{\{a\}})^{\alpha}\hat{Pr}(\omega_i|\mathbf{O}^{\{v\}})^{(1-\alpha)} \tag{15}$$

Dupont and Luettin [7] recently demonstrated a clear empirical relationship between the amount of additive acoustic noise and effective values for the exponential weighting α. This relationship can largely be attributed to the approximate isotropic shrinking that occurs in the distribution of acoustic cepstral speech features upon the introduction of additive noise [14]; as the cepstral distributions tend to maintain their orientation and position but the overall variance shrinks when corrupted with additive noise. In previous work [29] we have demonstrated that these exponential weightings can be directly related to the phenomena of cepstral shrinking in the acoustic modality.

Results using the weighted product rule in Equation 15 can be seen in Figure 5 for various acoustic noise conditions. In all cases the weighted product rule out performs the normal product rule, acoustic only and visual only identification results. In practice the exhaustive search of weighting α^* is infeasible as one needs a priori knowledge of the test set to get α^*, but results received can be considered as the upper limit of performance for known additive acoustic noise train/test mismatches.

The weighted product rule can be thought of as a data dependent combination function, again borrowing on terminology from [15], as it requires the training of α^* to match conditions in the acoustic test set.

6.3. Sum rule

The product rule, although optimal in the theoretical case [10], is effectively a severe rule when errors are present, as a single classifier can inhibit a particular class by outputting a probability that is close to zero. The weighted product rule can alleviate the influence of these errors to some degree, but must have quantitative knowledge of the train/test mismatch in both modalities. If this knowledge is not known or mistaken the incorrect selection of weighting α^* can have dire consequences on recognition performance. Alternatively, the sum rule is a benevolent combination rule, as errors in one classifier have a smaller effect on the final result. More importantly the sum rule is a static combination function that does not require any training. The sum rule makes the assumption that the error free a posteriori class probabilities for each modality $\{m\}$ do not deviate greatly from the priors [10] when there is a train/test mismatch,

$$Pr(\omega_i|\mathbf{O}^{\{m\}}) = P(\omega_i)(1 + \delta_i^{\{m\}}) \tag{16}$$

where $\delta_i^{\{m\}}$ satisfies $\delta_i^{\{m\}} \ll 1$. Kittler [12] called this a strong assumption and only offered heuristic insights into why it is valid. However, this assumption is naturally validated using our framework for train/test mismatch

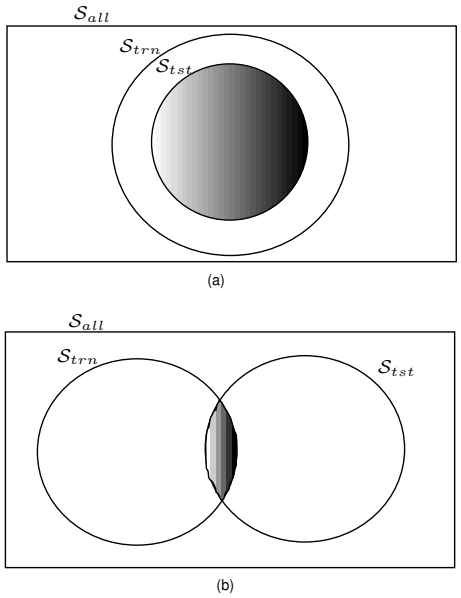

Figure 4. Venn diagram of changes in train/test conditions, (a) $\mathcal{S}_{tst} \subseteq \mathcal{S}_{trn}$ (similar train/test conditions), (b) $\mathcal{S}_{tst} \nsubseteq \mathcal{S}_{trn}$ (different train/test conditions).

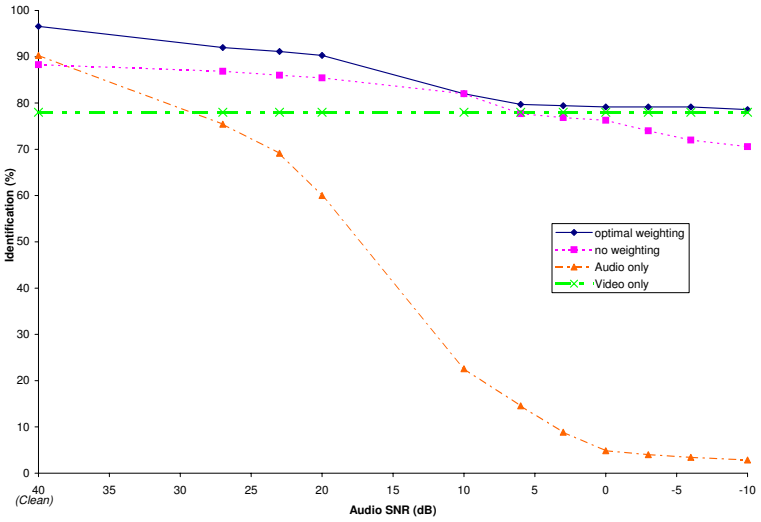

Figure 5. Benefit of finding optimal weighting α^* for a given acoustic noise condition.

established in Equations 12 and 13. Since these errors can be attributed to a train/test mismatch, the mismatch density which is common to all classes, naturally becomes dominant skewing the error free a posteriori class probabilities to similar values; thus validating the assumption made in Equation 16.

Under this assumption[‡] the sum rule can approximate the optimal product rule,

$$F_{pr}() \approx F_{sr}() \doteq Pr(\omega_i|\mathbf{O}^{\{a\}}) + Pr(\omega_i|\mathbf{O}^{\{v\}}) - P(\omega_i) \tag{17}$$

The sum rule is far more robust to the effects of train/test mismatches than the weighted product rule, when there is minimal quantitative knowledge of the mismatch. For the purposes of verification using a global threshold, it is useful to scale the resultant confidence scores from the sum rule to operate in a similar range to the weighted product rule. Using this scale factor we arrive at the practical sum rule employing our a posteriori probability estimates for each modality

$$F_{sr}() = sc\left[\hat{Pr}(\omega_i|\mathbf{O}^{\{a\}}) + \hat{Pr}(\omega_i|\mathbf{O}^{\{v\}})\right] \tag{18}$$

where a scale factor $sc = 0.5$ was employed.

6.4. A hybrid between product and sum rules for robust recognition

Kittler [10] hypothesized that a non-linear combination rule may in fact give superior performance over those previously mentioned. In our experimental work [30], we have devised a hybrid combination scheme using both the sum and weighted product rules based on a theoretical, empirical and heuristic understanding of where they work effectively. The hybrid combination scheme is defined as,

$$F_{hyb}() = \begin{cases} F_{sr}(), & \text{if } \sigma_{\zeta^{\{a\}}} < \theta \\ F_{wpr}(), & \text{otherwise} \end{cases} \tag{19}$$

The scheme uses the standard deviation $\sigma_{\zeta^{\{a\}}}$ of the vector $\boldsymbol{\zeta}^{\{a\}}$ of N acoustic log-likelihoods to dictate when the sum or weighted product rule should be used, where

$$\boldsymbol{\zeta}^{\{a\}} = \{\log p(\mathbf{O}^{\{a\}}|\boldsymbol{\lambda}_1), \ldots, \log p(\mathbf{O}^{\{a\}}|\boldsymbol{\lambda}_N)\} \tag{20}$$

The decision rule is based purely on the acoustic log-likelihoods as our experiments were concerned with additive acoustic noise. Dispersion measures of log likelihoods from an acoustic classifier have been shown empirically [16] to be a reasonable indicator of acoustic noise, but start failing in high levels of noise.

The threshold θ and weighting factors α^* used in Equation 19 were determined empirically to optimize performance across all acoustic noise levels. In this scenario they were chosen to be $\theta = 12$ and $\alpha^* = 0.9$. The technique was devised under the assumption that better results would be achieved with the weighted sum rule when there is minimal variation in scores (high acoustic noise), while the more severe but optimal weighted product rule would be used where there is large variation (low acoustic noise).

7. RESULTS AND DISCUSSION

In this paper two broad operational scenarios are entertained for the task of audio-visual speaker recognition.

Case I: defines an upper bound for performance, where there is quantitative knowledge of the train/test mismatch. Although unrealistic for a practical application, due to the violation of causality, this case gives an indication of the upper bound of performance for an audio-visual speaker recognition application. This case is investigated to gain empirical insights into what level of integration (EI, MI or LI) is most appropriate. Case I presents experiments for HMMs trained and evaluated under acoustically clean conditions.

[‡]Kittler et. al [10] has the full derivation of the approximate equivalence of the product and sum rules under this assumption.

Case II: defines a causal measure of performance, where there is only qualitative knowledge of possible train/test mismatches (e.g. that acoustic noise may be present, but the amount is unknown). This case represents the far more practical scenario of an AVSP application, where the mismatch is unknown and may vary. Insights can be gathered for this case pertaining to what combination functions perform best in a practical scenario assuming a certain level of integration (e.g. LI). Case II contains experiments for HMMs trained under clean acoustic conditions but evaluated over a variety of acoustic conditions.

Results are presented in Figure 6 and Table 1 for Case I, when the train/test match for a clean test set is known, showing the superiority of the weighted product LI strategy over all other integration strategies when an exhaustive search for the optimal α^* has been undertaken for both the identification and verification tasks. For the verification task in Cases I and II, since there were 36 speakers being evaluated, the verification task involved 36 claimant matches and 36×35 impostor tests. Results are also presented in Figures 7 and 8 for the identification and verification tasks respectively for Case II where there is very limited knowledge of the train/test conditions. For the identification task it can be seen in Figure 7 our hybrid approach gives best overall results. Identification results were above acoustic only results in low noise. The hybrid results were also just below the visual only results in high noise. For the verification task however, the weighted sum rule received best results in terms of EER, falling just below those received for the weighted product rule in clean conditions and giving results closest to that of the visual only results in high noise.

It should be noted that MI's MSHMMs perform better than EI's HMMs in both the identification and verification tasks. Even though both techniques enforce state transitions between modalities synchronously (ie. allowing movement only along the 2-D state lattice diagonally) the MSHMMs superior performance can be attributed to two characteristics. First, since the acoustic and visual modalities have been trained separately for MSHMMs the resultant classifiers are much less likely to be undertrained, due to the "curse" of dimensionality [22], to the same degree as an EI HMM. Second, the EI implementation we used had no ability to dampen errors unlike MSHMMs which used a α^* factor in a similar manner to its MAHMM cousin. For Case I results it must be emphasized that optimal weighting factors were found for all integration strategies through an exhaustive search.

7.1. Case I

Results for Case I demonstrate the benefit of treating the acoustic and visual modalities independently, as done in LI, in terms of classifier complexity and its ability to dampen errors occurring in both modalities.

The necessity for the extra classifier complexity found in LI, over EI or synchronous MI, becomes apparent if one inspects the 2-D state histograms (SH) of various digits for the test set in clean conditions. A 2-D SH can be defined as the number of times in an observation sequence $\mathbf{O}$ a given observation $\mathbf{o}_t$ has been in state (i, j) in the 2-D state lattice. These SHs can be averaged across many utterances of the same digit so as to gain an idea of the amount of time spent in each state for a given digit utterance. The primary condition of an SH is for its sum to be equal to one. Typical 2-D SHs can be seen in Table 2 for the digits five and eight with the most likely trajectory, travelling from the bottom left to the top right, being highlighted for both cases. Note, since these SHs were found for individual digits across all speakers the background HMM digit models were used.

If one inspects both examples (a) and (b) in Table 2 one can see that a lot of time is spent in the off diagonals of the 2-D SH. This result highlights two things. First, there is obvious benefit of asynchronously traversing the 2-D state lattice as it allows for higher degrees of freedom during the classification process. Second, strategies using a left to right EI type HMM or MI's synchronous MSHMMs do not allow for such flexibility as they can only traverse the 2-D state lattice along the diagonal.

As expected the weighted product rule obtained very good performance as long as a suitable weighting factor α is chosen to account for the effects of cepstral shrinking. Performance is very sensitive to the correct selection of α^* for a given acoustic train/test mismatch. This class dependency manifests most noticeably for LI's weighted product rule in the performance improvement received for the identification and verification tasks in comparison to other integration strategies. For the identification task the weighted product rule clearly out performs the next closest LI's weighted sum rule. The verification task however, receives much closer results in terms of EER.

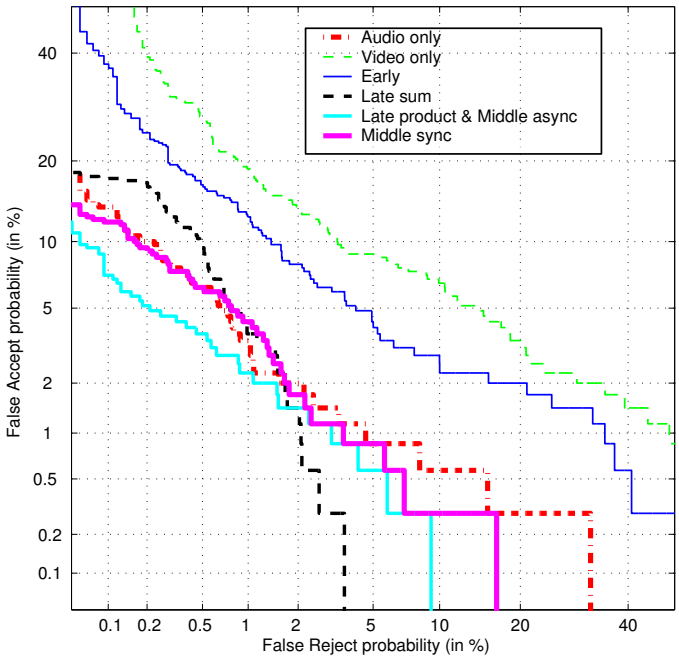

Figure 6. Case I: DET curves various integration strategies under clean conditions.

modality	integration	EER (%)	Id (%)
audio	none	2.00	90.28
video	none	7.43	78.00
av (sum)	late	1.43	92.28
av (product)	late	1.15	96.57
av	early	4.85	83.43
av (async)	middle	1.15	96.57
av (sync)	middle	1.74	91.14

Table 1. Case I: Equal error rates (EER) and identification rates (Id) for integration strategies under clean conditions using optimal α^* (best strategies are highlighted).

Video States	Audio States 1	2	3	Finish
3	0.0117	0.0864	0.2325	
2	0.1624	0.1238	0.035	
1	0.2897	0.0409	0.0175	
Start				

(a)

Video States	Audio States 1	2	3	Finish
3	0.0125	0.0815	0.1486	
2	0.0882	0.1457	0.0853	
1	0.187	0.1908	0.0604	
Start				

(b)

Table 2. 2-D state histograms taken from M2VTS verification set for digits (a) FIVE and (b) EIGHT.

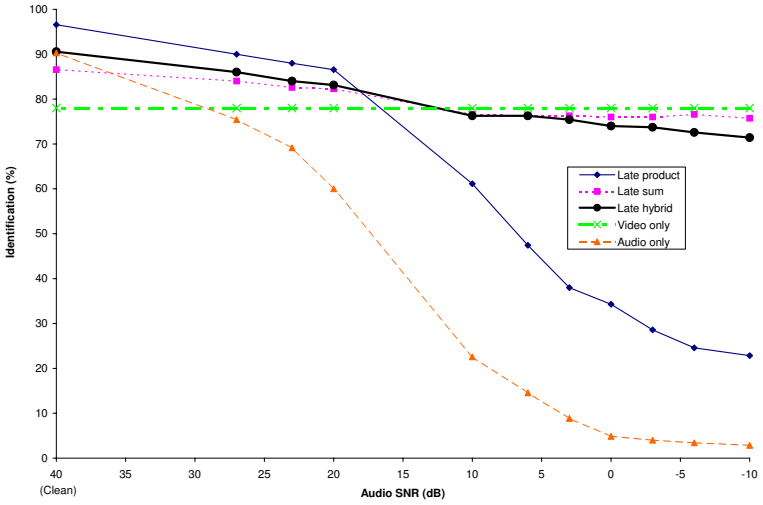

Figure 7. Case II: Identification rates for various LI strategies over various additive acoustic noise conditions.

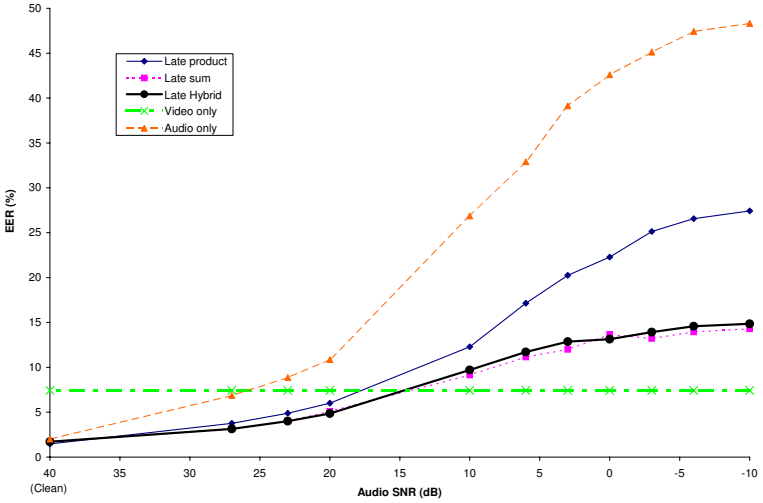

Figure 8. Case II: Equal error rates (EER) for various LI strategies over various additive acoustic noise conditions.

This can be attributed to the fundamental difference in the nature of tasks for identification and verification. Identification is concerned with selecting the speaker with the highest ranking score for a single observation. The verification task is heavily dependent on the value of the score, not the rank, as a threshold Th has to be found to differentiate between claimants and impostors across many observations. The weighted product rule, while giving the correct rank of speakers, is a very extreme rule and may not give score values that generalize well across many observations and speakers for the verification task. The sum rule however, is a more stable rule, as it is much less sensitive to errors than the product rule [10], and gives scores that provide better generalization across observations and speakers.

7.2. Case II

The results in Figure 7 show that our proposed hybrid technique for Case II is of some benefit across all tested configurable acoustic noise conditions. However, Figure 8 for the verification task shows that the hybrid approach is similar, if not slightly worse, with the weighted sum rule performing best across most tested configurable acoustic noise conditions. The disparity in performance can be attributed to the switch that occurs between the weighted product and sum rules in the hybrid approach making the calculation of a satisfactory general threshold θ difficult. However, the difference in performance between our proposed technique and the weighted sum rule is negligible. For all cases in high noise the verification performance in terms of EER is very poor in comparison to the visual only classifier. The obvious benefit of our hybrid approach is its ability to be tunable, in terms of the threshold θ, to the conditions it is to be used under. For instance in the results presented in Figures 7 and 8 a threshold θ was chosen to ensure that identification results and verification results were above the catastrophic fusion boundary in clean conditions while receiving reasonable results in higher noise environments. The tunable characteristic is of considerable use if one knows the what upper and lower performance limits one wants in their AVSP system.

8. CONCLUSION AND FUTURE WORK

In this paper we have shown empirically and theoretically the benefit of LI over other integration strategies in terms of classifier flexibility and its ability to dampen independent errors coming from either modality. An equivalence has also been established between MI and LI strategies under certain circumstances in terms of a HMM classifier framework. The weighted product rule has been shown to perform best for *Case I* mode of AVSP operation, where one has quantitative knowledge of the train/test mismatch. The benefits of a hybrid combination scheme have been highlighted for addressing *Case II* mode of AVSP operation. Finally, a framework has been devised for dealing with the natural train/test mismatches that occur when classifiers are used practically via the concept of set based knowledge. Within this framework it has been shown that the sum rule naturally results as an approximation to the product rule when confidence errors are present, provided the true a posteriori probabilities are conditionally independent.

Our future work shall try to address the use of 2-D state histograms as a possible avenue for further identification and verification performance in an AVSP application. A soft transition between the weighted product and sum rules shall also be investigated, for our hybrid combination scheme, rather than the hard transition implementation currently in use. The equivalence of MI and LI will also be further investigated to allow for combination functions other than the weighted product rule. The extended M2VTS (XM2VTS) [31] audio-visual database, with 295 speakers, could also be used for more comprehensive recognition results over a greater number of speakers.

9. ACKNOWLEDGEMENTS

The authors would like to thank the M2VTS Project for use of their database. Part of our work is supported by an Intel research grant.

REFERENCES

1. T. Wark, S. Sridharan, and V. Chandran, "Robust speaker verification via fusion of speech and lip modalities," in *Int. Conf. Acoustics, Speech and Signal Processing (ICASSP'99)*, vol. 6, 1999, pp. 3061–3064.
2. T. Chen and R. Rao, "Audio-visual integration in multimodal communication," *Proc. IEEE*, vol. 86, no. 5, pp. 837–852, May 1998.
3. F. Lavagetto, "Converting speech into lip movements: A multimedia telephone for hard hearing people," *IEEE Trans. Rehab. Eng.*, vol. 3, no. 1, pp. 90–102, March 1995.
4. H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, pp. 746–748, December 1976.
5. D. G. Stork and M. E. Hennecke, Eds., *Speechreading by Humans and Machines*, ser. NATO ASI Series F: Computer and Systems Sciences. Springer-Verlag, 1996, vol. 150.
6. C. C. Chibelushi, F. Deravi, and J. S. D. Mason, "A review of speech-based bimodal recognition," *IEEE Trans. Multimedia*, vol. 4, no. 1, pp. 23–37, March 2002.
7. S. Dupont and J. Luettin, "Audio-visual speech modeling for continuous speech recognition," *IEEE Trans. Multimedia*, vol. 2, no. 3, pp. 141–151, September 2000.
8. G. Potamianos and H. P. Graf, "Discriminative training of HMM stream exponents for audio-visual speech recognition," in *Int. Conf. Acoustics, Speech and Signal Processing (ICASSP'98)*, vol. 6, 1998, pp. 3733–3736.
9. A. P. Varga and R. K. Moore, "Hidden markov model decomposition of speech and noise," in *Int. Conf. Acoustics, Speech and Signal Processing (ICASSP'90)*, vol. 2, 1990, pp. 845–848.
10. J. Kittler, M. Hatef, R. Duin, and J. Matas, "On combining classifiers," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 20, no. 3, pp. 226–239, March 1998.
11. K. Fukunaga, *Introduction to statistical pattern recognition*, 2nd ed. Academic Press Inc., 1990.
12. J. Kittler, "Combining classifiers: A theoretical framework," *Pattern Analysis and Applications*, vol. 1, no. 1, pp. 18–27, 1998.
13. J. R. Movellan and P. Mineiro, "Modularity and catastrophic fusion: A bayesian approach with applications to audio-visual speech recognition," Departement of Cognitive Science, USCD, San Diego, CA, Tech. Rep. 97.01, 1997.
14. D. Mansour and B. H. Juang, "A family of distortion measures based upon projection operation for robust speech recognition," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 37, no. 11, pp. 1659 –1671, November 1989.
15. M. S. Kamel and N. M. Wanas, "Data dependence in combining classifiers," in *Multiple Classifier Systems*, T. Windeatt and F. Roli, Eds., 2003, pp. 1–14.
16. A. Adjoudani and C. Benoit, "Audio-visual speech recognition compared across two architectures," in *European Conf. Speech Communication and Technology (Eurospeech'95)*, 1995, pp. 1563–1566.
17. S. Cox, I. Matthews, and J. A. Bangham, "Combining noise compensation with visual information in speech recognition," in *Auditory-Visual Speech Processing (AVSP'97)*, Rhodes, 1997.
18. C. C. Chibelushi, J. S. Mason, and F. Deravi, "Integration of acoustic and visual speech for speaker recognition," in *European Conf. Speech Communication and Technology (Eurospeech'93)*, 1993, pp. 157–160.
19. M. McGrath and Q. Summerfield, "Intermodal timing relations and audio-visual speech recognition," *J. Acoust. Soc. Amer.*, vol. 77, no. 2, pp. 678–685, February 1985.
20. J. Luettin, G. Potamianos, and C. Neti, "Asynchronous stream modeling for large vocabulary audio-visual speech recognition," in *Int. Conf. Acoustics, Speech and Signal Processing (ICASSP'01)*, vol. 1, 2001, pp. 169 –172.
21. A. Martin, G. Doddington, T. Kamm, M. Ordowski, and P. Przybocki, "The DET curve in assessment of detection task performance," in *European Conf. Speech Communication and Technology (Eurospeech'97)*, vol. 4, 1997, pp. 1895–1898.
22. R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. John Wiley and Sons, Inc., 2001.
23. S. Pigeon and L. Vandendorpe, "The M2VTS multimodal face database," in *Int. Conf. Audio and Video-based Biometric Person Authentication (AVBPA'97)*, 1997.
24. P. Jourlin, J. Luettin, D. Genoud, and H. Wassner, "Acoustic-labial speaker verification," *Pattern Recognition Letters*, 1997.

25. S. Lucey, V. Chandran, and S. Sridharan, "Improved facial-feature detection for AVSP via unsupervised clustering and discriminant analysis," *EURASIP Journal on Applied Signal Processing*, no. 3, pp. 264–275, 2003.
26. S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book (for HTK version 2.2)*. Entropic Ltd., 1999.
27. L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, February 1989.
28. M. J. Tomlinson, M. J. Russell, and N. M. Brooke, "Integrating audio and visual information to provide highly robust speech recognition," in *Int. Conf. Acoustics, Speech and Signal Processing (ICASSP'96)*, 1996, pp. 821–824.
29. S. Lucey, V. Chandran, and S. Sridharan, "A link between cepstral shrinking and the weighted product rule in audio-visual speech recognition," in *Int. Conf. Spoken Language Processing (ICSLP'02)*, 2002, pp. 1961–1964.
30. S. Lucey, S. Sridharan, and V. Chandran, "An investigation of HMM classifier combination strategies for improved audio-visual speech recognition," in *European Conf. Speech Communication and Technology (Eurospeech'01)*, 2001, pp. 1185–1188.
31. K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre, "XM2VTSDB: The extended M2VTS database," in *Int. Conf. Audio and Video-based Biometric Person Authentication (AVBPA'99)*, 1999, pp. 72–77.