Improved Audio-visual Speaker Recognition Via the Use of a Hybrid Combination Strategy

Simon Lucey and Tsuhan Chen

Advanced Multimedia Processing Laboratory
Department of Electrical and Computer Engineering
Carnegie Mellon University
Pittsburgh PA 15213, USA
slucey@ieee.org and tsuhan@cmu.edu

Abstract. In this paper an in depth analysis is undertaken into effective strategies for integrating the audio-visual modalities for the purposes of text-dependent speaker recognition. Our work is based around the well known hidden Markov model (HMM) classifier framework for modelling speech. A framework is proposed to handle the mismatch between train and test observation sets, 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 audio-visual speaker recognition 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. Results are presented on the M2VTS database.

1 Introduction

Text-dependent applications for the task of speaker recognition typically out perform their text-independent counter parts due to the simplification of the recognition task. 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.

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 generalised 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 [1]. These confidence errors should not be confused with Bayesian error [2], which is inherent to the classification task. It has been well documented by Kittler [1] that when combining independent classifiers, where such confidence errors are *not* present, the product rule is optimal, under the assumption of conditional independence. 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* [3].

Based on the assumption that poor performance in most audio-visual speech processing (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*. In this paper we assume, when train/test mismatches do occur in each modality, it is better to integrate the audio-visual modalities at the decision level. Two different combination functions for decision level combination are investigated, namely the weighted product and 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.

2 Speaker recognition

Speaker recognition encompasses 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.

3 Audio-visual database and feature extraction

The M2VTS database [4] 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. This database has been used in previous multimodal speaker recognition experiments [5]. 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 [6].

The mouth 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) [2]. Linear discriminant analysis (LDA) [2] 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 [7].

4 Hidden Markov Model Training

All HMMs were trained using the Baum Welch algorithm via the HTK [7] 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 initialising 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.

For the acoustic and visual modalities, an utterance was modelled using a 3 state, left to right, HMM with 3 mixtures per state and diagonal covariance matrices. The likelihood scores $p(\mathbf{O}|\boldsymbol{\lambda}_i)$ from each HMM $\boldsymbol{\lambda}_i$ were used to gain the a posteriori probability estimates, assuming equal priors, using Bayes rule,

$$\hat{Pr}(\omega_i|\mathbf{O}) = \frac{p(\mathbf{O}|\boldsymbol{\lambda}_i)}{\sum_{n=1}^{N} p(\mathbf{O}|\boldsymbol{\lambda}_i)} \tag{3}$$

Shots 1-3 of the M2VTS database were used for training the HMMs with shot 4 being used for testing.

5 Integration Strategies

5.1 Weighted product rule

Excellent results in AVSP have been received through integrating the confidence scores received from the acoustic and visual classifiers via the weighted product rule. The weighted product rule can be expressed as,

$$\zeta(\omega_i|\mathbf{O}^{\{av\}})_{\times} = \hat{Pr}(\omega_i|\mathbf{O}^{\{a\}})^{\alpha} \times \hat{Pr}(\omega_i|\mathbf{O}^{\{v\}})^{(1-\alpha)} \tag{4}$$

where $\hat{Pr}(\omega_i|\mathbf{O}^{\{m\}})$ is the a posteriori *estimate* of utterance $\mathbf{O}^{\{m\}}$ coming from subject class ω_i for modality $\{m = a \text{ or } v\}$. It must be emphasised that $\zeta(\omega_i|\mathbf{o})$ is

a confidence score (not necessarily between zero and one), *not* a probability, but is equivalent to the audio-visual a posteriori probability estimate $\hat{Pr}(\omega_i|\mathbf{O}^{\{av\}})$ in terms of the class decision boundaries it realises.

Bayesian theory dictates [1] that the weighted product rule should be optimal, given conditional independence between modalities, when $\alpha = 0.5$ (i.e. normal product rule); if one is combining *error free* a posteriori class probabilities. In practice however, one can rarely use the normal product rule due to the differing decision boundaries realised from the mismatch between train and test utterances. This mismatch results in a confidence error,

$$\hat{Pr}(\omega_i|\mathbf{O}^{\{m\}}) = Pr(\omega_i|\mathbf{O}^{\{m\}}) + \epsilon_i(\mathbf{O}^{\{m\}}) \tag{5}$$

When combining a posteriori probability estimates from both modalities the compounding effect of these confidence errors must be taken into account when selecting a suitable combination strategy. In some circumstances the magnitude of this confidence error can be diminished for certain types of mismatches through the use of an exponential weighting as found in Equation 4, where there is an approximate isotropic shrinking between the train and test set distributions. The exponential weighting has no effect on the order of scores in each modality individually. However, through the judicious choice of an appropriate exponential weighting in the application of weighted product rule, improved combined performance can be witnessed.

This type of "shrinking" has been shown to occur in acoustic cepstral features in the presence of additive noise [8]. Dupont and Luettin [9] were able to establish an empirical relationship between the exponential weighting and additive acoustic noise, although the effectiveness of the weighting does decrease in large amounts of acoustic noise. When addressing the isotropic shrinking of distributions the exponential weighting in the weighted product rule can be thought of as acting in an adapting capacity; such that it tries to remove the confidence error from the distribution shrinkage completely. It must be mentioned that the exponential weighting in the weighted product rule does also aid, to some degree, with other types of mismatch, other than isotropic shrinkage, that may be present (such as in the visual modality). In this capacity the exponential weighting tries to transform, rather than remove, the confidence errors into Bayesian error. Although, the ability of the weighting to act in this alternate dampening capacity is quite limited. A more thorough discussion of this topic can be found in [10] with respect to AVSP. For the weighted product rule an $\alpha = 0.9$ was found, through an exhaustive search, to perform best in clean to medium acoustic conditions (i.e. 40db - 20db).

5.2 Sum rule

The product rule, although optimal in the theoretical case, is effectively a severe rule when confidence 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. The effect of this mismatch can sometimes be lessened, with respect to the product rule, through the use of an exponential weighting. However, this weighting can only address certain types of mismatches (i.e. isotropic shrinking) and requires intimate knowledge on the degree of mismatch (e.g. signal to noise ratio (SNR)). If this knowledge is not known or mistaken the incorrect selection of a weighting can have dire consequences on recognition performance.

Alternatively, the sum rule in Equation 6 is a benevolent combination rule, as errors in one classifier have a smaller effect on the final result. The sum rule makes the assumption that the error free a posteriori class probabilities for each modality do not deviate greatly from the priors [1].

$$\zeta(\omega_i|\mathbf{O}^{\{av\}})_+ = 0.5\hat{Pr}(\omega_i|\mathbf{O}^{\{a\}}) + 0.5\hat{Pr}(\omega_i|\mathbf{O}^{\{v\}}) \tag{6}$$

Note a 0.5 scaling factor was placed out the front of the a posteriori probability estimates of both modalities in Equation 6. This was done to try and scale the resultant confidence scores in $\zeta(\omega_i|\mathbf{O}^{\{av\}})_+$ to be in the same range as $\zeta(\omega_i|\mathbf{O}^{\{av\}})_\times$, which is of particular importance with respect to verification using the hybrid rule outlined in the next section.

5.3 A hybrid between product and sum rules for robust recognition

Kittler [1] hypothesised that a non-linear combination rule may in fact give superior performance over those previously mentioned. In our experimental work, we have devised a hybrid combination scheme using both the weighted 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,

$$\zeta(\omega_i|\mathbf{O}^{\{av\}})_{\times/+} = \begin{cases} \zeta(\omega_i|\mathbf{O}^{\{av\}})_\times, & \sigma_{\zeta^{\{a\}}} < \theta \\ \zeta(\omega_i|\mathbf{O}^{\{av\}})_+, & \sigma_{\zeta^{\{a\}}} \geq \theta \end{cases} \tag{7}$$

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

$$\boldsymbol{\zeta}^{\{a\}} = \{\log p(\mathbf{O}^{\{a\}}|\boldsymbol{\lambda}_1) - \log p(\mathbf{O}^{\{a\}}|\boldsymbol{\lambda}_{bck}), \ldots, \log p(\mathbf{O}^{\{a\}}|\boldsymbol{\lambda}_N) - \log p(\mathbf{O}^{\{a\}}|\boldsymbol{\lambda}_{bck})\} \tag{8}$$

The decision rule is based purely on the normalised 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 [11] to be a reasonable indicator of acoustic noise, but start failing in high levels of noise. As shown in Equation 8 the normalisation of the log likelihoods is performed by subtracting the background model scores for a particular digit from the speaker dependent models for that digit. This was done so that a common reliable threshold θ could be found for all digits.

The threshold θ used in Equation 7 was determined empirically to optimise performance across all acoustic noise levels, and in this scenario was chosen to be $\theta = 12$. 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).

5.4 Results and discussion

The results in Figure 1 show that our proposed hybrid technique is of some benefit across all tested configurable acoustic noise conditions. However, Figure 2 for the verification task depicts that the hybrid approach is similar, if not slightly worse, with the weighted sum rule performing best across most tested configurable acoustic noise conditions. This disparity in performance can be partly 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; although the scaling of the sum rule tries to address this difference. 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 1 and 2 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 audio-visual recognition system.

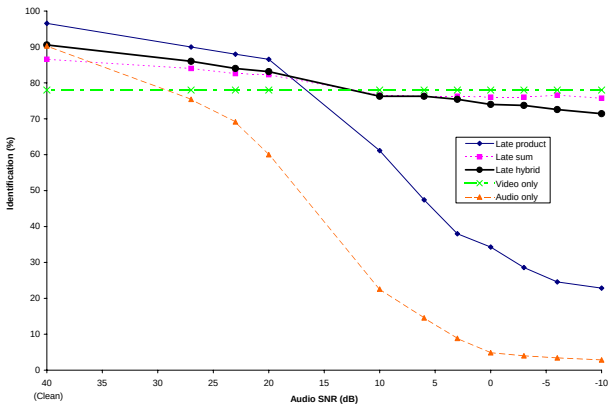

Fig. 1. Identification rates over various additive acoustic noise conditions.

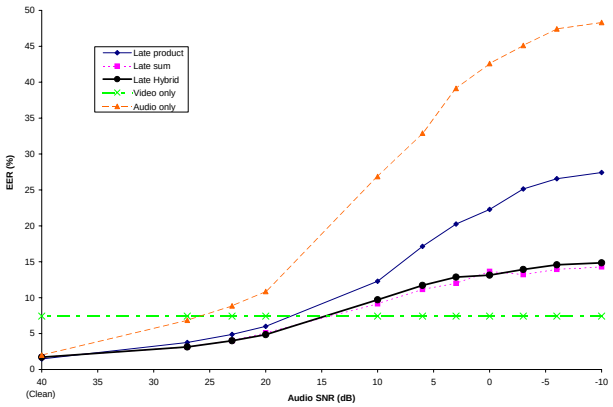

Fig. 2. Equal error rates (EER) over various additive acoustic noise conditions.

References

1. J. Kittler, "Combining classifiers: A theoretical framework," *Pattern Analysis and Applications*, vol. 1, no. 1, pp. 18–27, 1998.
2. K. Fukunaga, *Introduction to statistical pattern recognition.* 24-28 Oval Road, London NW1 7DX: Academic Press Inc., 2nd ed., 1990.
3. J. R. Movellan and P. Mineiro, "Modularity and catastrophic fusion: A bayesian approach with applications to audio-visual speech recognition," Tech. Rep. 97.01, Departement of Cognitive Science, USCD, San Diego, CA, 1997.
4. S. Pigeon, "The M2VTS database," (Laboratoire de Telecommunications et Teledection, Place du Levant, 2-B-1348 Louvain-La-Neuve, Belgium), 1996.
5. P. Jourlin, J. Luettin, D. Genoud, and H. Wassner, "Acoustic-labial speaker verification," *Pattern Recognition Letters*, 1997.
6. S. Lucey, S. Sridharan, and V. Chandran, "Improved facial feature detection for AVSP via unsupervised clustering and discriminant analysis," *EURASIP Journal on Applied Signal Processing*, March 2003.
7. S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book (for HTK version 2.2).* Entropic Ltd., 1999.
8. R. J. Mammone, X. Zhang, and R. P. Ramachandran, "Robust speaker recognition: A feature-based approach," *IEEE Signal Processing Magazine*, vol. 13, pp. 58–70, September 1996.
9. S. Dupont and J. Luettin, "Audio-visual speech modeling for continuous speech recognition," *IEEE Transactions on Multimedia*, vol. 2, pp. 141–151, September 2000.
10. S. Lucey, S. Sridharan, and V. Chandran, "A link between cepstral shrinking and the weighted product rule in audio-visual speech recognition," in *International Conference on Spoken Language Processing*, (Denver, Colorado), September 2002.
11. A. Adjoudani and C. Benoit, "Audio-visual speech recognition compared across two architectures," in *EUROSPEECH'95*, (Madrid Spain), pp. 1563–1566, September 1995.