Proc. of Intl. Conf. Acoustic, Speech and Signal Proc., Toulouse, France, May 2006.

STREAM WEIGHT COMPUTATION FOR MULTI-STREAM CLASSIFIERS

Alexandros Potamianos, Eduardo Sanchez-Soto, Khalid Daoudi†

Dept. of ECE, Technical Univ. of Crete, Chania 73100, Greece
† IRIT-UPS, Toulouse 31062, France
{potam,esanchez}@telecom.tuc.gr daoudi@irit.fr

ABSTRACT

In this paper, we provide theoretical results on the problem of optimal stream weight selection for the multi-stream classification problem. It is shown, that in the presence of estimation or modeling errors using stream weights can decrease the total classification error. The stream weights that minimize classification estimation error are shown to be inversely proportional to the single-stream pdf estimation error. It is also shown that under certain conditions, the optimal stream weights are inversely proportional to the single-stream classification error. We apply these results to the problem of audio-visual speech recognition and experimentally verify our claims. The applicability of the results to the problem of unsupervised stream weight estimation is also discussed.

1. INTRODUCTION

A common practice for combining information sources in a statistical classification framework is the use of "feature streams". A fundamental assumption behind streams is that the information sources/features are independent of each other and that the probability distribution functions (pdfs) of the two streams can be multiplied to obtain the global observation pdf. However, often this independence assumption does not hold or the reliability (estimation/modeling error) of each stream is different. In these cases, it has been empirically shown that stream weights (exponents weighting the contribution of each stream pdf) can reduce the total classification error.

In the speech recognition literature, multi-stream recognizers have been used to combine feature streams of different reliability [4] or different information content [1]. Multi-stream recognition is also a popular method for combining the audio and visual information in audio-visual automatic speech recognition (AV-ASR) [7]. The problem of supervised stream weight computation for these recognition scenarios is well studied: minimum error (discriminative) training can be used to select the best combination of stream weights during model training [8]. Recently there has been some interest in investigating unsupervised algorithms for estimating stream weights during recognition [9]. Unsupervised estimation of stream weights is an especially important problem when "mismatch" exists between the training and test data [1] or when supervised training of weights is not possible [5].

In this paper, we provide analytical results for the selection of stream weights as a function of single-stream estimation and misclassification errors. Optimality is investigated in terms of multi-stream classification error minimization for the two class problem. The analytical results are verified for an AV-ASR multi-stream application.

2. TOTAL CLASSIFICATION ERROR

Consider the two class w_1, w_2 classification problem with feature pdfs $p(x|w_1), p(x|w_2)$ and class priors $p(w_1), p(w_2)$ respectively. Lets assume that the estimation/modeling error is a random variable z_i that follows a normal pdf with variance σ_i^2, i.e.,

$$p(w_i|x,\lambda) - p(w_i|x) = z_i \text{ and } z_i \sim \mathcal{N}(z;0,\sigma_i^2) \quad (1)$$

where λ denotes the selected model/estimation method and thus $p(w_i|x,\lambda)$ is the estimated value of the true distribution $p(w_i|x)$. Then the Bayes classification decision [2] using the estimated pdfs becomes

$$\begin{aligned} p(w_1|x,\lambda) - p(w_2|x,\lambda) &\gtrless 0 \Leftrightarrow \\ p(w_1|x) - p(w_2|x) + (z_1 - z_2) &\gtrless 0 \Leftrightarrow \\ p(x|w_1)p(w_1) - p(x|w_2)p(w_2) + (z_1 - z_2)p(x) &\gtrless 0 \end{aligned} \quad (2)$$

where $z = (z_1 - z_2)p(x)$ is a random variable that determines the deviation of the decision boundary from the optimal value $p(x|w_1)p(w_1) - p(x|w_2)p(w_2) = 0$. To simplify our computations lets assume that $p(x)$ is constant in the region of interest (close to the decision boundary) and that z follows a normal pdf, i.e.,

$$z \sim \mathcal{N}(z;0,\sigma^2) \text{ where } \sigma^2 = p(x)^2(\sigma_1^2 + \sigma_2^2). \quad (3)$$

Given that the classification decision is now a function of the random variable z to compute the actual (total) classification error we proceed as follows:

$$\begin{aligned} p(error|x) &= \sum_{i=1}^{2} p(error, \Omega_i|x) \\ &= \sum_{i=1}^{2} p(error|\Omega_i, x)\, p(\Omega_i|x) \\ &= p(w_2|x)p(\Omega_1|x) + p(w_1|x)p(\Omega_2|x) \end{aligned} \quad (4)$$

where Ω_1, Ω_2 are the decision regions for class w_1, w_2 respectively and $p(\Omega_1|x)$, $p(\Omega_2|x)$ are the probabilities of taking classification decision w_1, w_2 respectively for feature sample x. The decision probabilities $p(\Omega_i|x)$ can be expressed as a function of z as follows:

$$\begin{aligned} p(\Omega_1|x) &= p(p(x|w_1)p(w_1) - p(x|w_2)p(w_2) + z > 0) \\ &= \int_{f(x)}^{+\infty} \mathcal{N}(z;0,\sigma^2)dz \end{aligned} \quad (5)$$

where $f(x) = p(x|w_2)p(w_2) - p(x|w_1)p(w_1)$. Thus the total (Bayes and estimation/modeling) error can be computed as

$$\begin{aligned}P(error) &= \int_{\mathcal{R}^d} P(error,x)dx = \int_{\mathcal{R}^d} P(error|x)p(x)dx \\ &= \int_{\mathcal{R}^d}[p(w_2|x)p(\Omega_1|x) + p(w_1|x)p(\Omega_2|x)]p(x)dx \\ &= \int_{\mathcal{R}^d}[p(x|w_2)p(w_2)p(\Omega_1|x)+p(x|w_1)p(w_1)p(\Omega_2|x)]dx \\ &= \int_{\mathcal{R}^d}\int_{f(x)}^{+\infty}\mathcal{N}(z;0,\sigma^2)dz\; p(x|w_2)p(w_2)dx + \\ &\quad \int_{\mathcal{R}^d}\int_{-\infty}^{f(x)}\mathcal{N}(z;0,\sigma^2)dz\; p(x|w_1)p(w_1)dx \quad (6)\end{aligned}$$

where d is the dimension of the feature space $\mathcal{R}^d$.

3. MULTI-STREAM CLASSIFICATION

Next we consider the case where the feature vector x is broken up into two independent streams x_1, x_2 of dimension d_1 and d_2 respectively. Stream weights s_1, s_2 are used to "equalize" the probability in each stream, i.e.,

$$p(x|w_i) = \prod_{j=1}^{2} p(x_j|w_i)^{s_j} \quad (7)$$

given that $\sum_j s_j = 1$. Let us also assume that the estimation/modeling error in the Bayes decision is given by the random variable z that follows the normal pdf $\mathcal{N}(z;0,\sigma^2)$, i.e.,

$$\prod_{j=1}^{2}[p(x_j|w_1)p(w_1)]^{s_j} - \prod_{j=1}^{2}[p(x_j|w_2)p(w_2)]^{s_j} + z \gtrless 0 \quad (8)$$

The total error can be computed as outlined in the previous section; the only change in the total error estimate formula in Eq. (6) is in the decision function $f(x)$ is now defined as

$$\prod_{j=1}^{2}[p(x_j|w_2)p(w_2)]^{s_j} - \prod_{j=1}^{2}[p(x_j|w_1)p(w_1)]^{s_j} \quad (9)$$

It is interesting to note that when using stream weights the total error is higher than the Bayes error (because we have moved the decision boundary defined by $f(x)$). However, stream weights can decrease the estimation/modeling error. In general, stream-weight estimation is the process of finding the optimal values that minimize the total expected error; in this process, Bayes error will increase and estimation error will decrease by a larger amount. However, selecting weights that minimize the total error is a hard problem. Instead we assume that the Bayes error increase due to stream weights is small and focus on minimizing the estimation error.

3.1. Estimation error minimization

In this section, we investigate the problem of stream weight selection that minimizes the estimation error, i.e., the variance σ^2 of the random variable z in Eq. (8). Lets assume that the estimation error for the ith class and jth stream is a random variable that follows the normal pdf $z_{ij} \sim \mathcal{N}(z;0,\sigma_{ij}^2)$, i.e.,

$$p(w_i|x_j,\lambda) - p(w_i|x_j) = z_{ij} \quad (10)$$

Then the Bayes classification discriminant function can be expressed as

$$\begin{aligned}&\prod_{j=1}^{2}[p(x_j|w_1,\lambda)p(w_1)]^{s_j} - \prod_{j=1}^{2}[p(x_j|w_2,\lambda)p(w_2)]^{s_j} \\ &= \prod_{j=1}^{2}[p(x_j|w_1)p(w_1) + z_{1j}p(x_j)]^{s_j} - \\ &\quad \prod_{j=1}^{2}[p(x_j|w_2)p(w_2) + z_{2j}p(x_j)]^{s_j} \\ &\approx (\prod_{j=1}^{2}[p(x_j|w_1)p(w_1)]^{s_j} - \prod_{j=1}^{2}[p(x_j|w_2)p(w_2)]^{s_j}) + \\ &[\frac{s_1 z_{11}p(x_1)}{p(x_1|w_1)} + \frac{s_2 z_{12}p(x_2)}{p(x_2|w_1)}][p(x_1|w_1)^{s_1}p(x_2|w_1)^{s_2}] - \\ &[\frac{s_1 z_{21}p(x_1)}{p(x_1|w_2)} + \frac{s_2 z_{22}p(x_2)}{p(x_2|w_2)}][p(x_1|w_2)^{s_1}p(x_2|w_2)^{s_2}]\end{aligned}$$

assuming that $z_{ij} << p(w_i|x_j)$ and the quadratic z^2 terms can be ignored. Note that the second part of the equation above is z (as in Eq. (8)). Making the further assumption that in the decision region the posterior probabilities for the two classes are equal, i.e., $p(w_1|x_j) \approx p(w_2|x_j)$, we get

$$\begin{aligned}z \approx\; & 2[p(x_1|w_1)p(w_1)]^{s_1}[p(x_2|w_1)p(w_1)]^{s_2} \\ & [s_1(z_{11}-z_{21}) + s_2(z_{12}-z_{22})]\end{aligned}$$

and for the variances

$$\begin{aligned}\sigma^2 &\approx 4p(w_1)^2 p(x_1|w_1)^{2s_1}p(x_2|w_1)^{2s_2}\sum_{i=1}^{2}\sum_{j=1}^{2}s_j^2\sigma_{ij}^2 \Rightarrow \\ \sigma^2 &\sim p(x_1|w_1)^{2s_1}p(x_2|w_1)^{2s_2}[s_1^2\sigma_{S1}^2 + s_2^2\sigma_{S2}^2] \quad (11)\end{aligned}$$

where $\sigma_{Sj}^2 = \sum_{i=1}^{2}\sigma_{ij}^2$ is the total stream variance. From the equation above it is easy to see that stream weights may reduce estimation error only when either the pdf estimation errors of the single-stream (stand-alone) classifiers are different, i.e., one feature stream is more reliable than the rest, and/or the Bayes errors of the single-stream classifiers are different, i.e., one stream contains more information pertinent to the classification problem than the rest. Next we investigate these two cases:

- **Equal Bayes classification error**: We assume that the each of the single-stream classifiers have the same Bayes classification error but different estimation errors. In this case, we can make the assumption that in the decision region $p(x_1|w_1) \approx p(x_2|w_1)$, provided that the features x_1, x_2 follow a similar parametric distribution (e.g., Gaussian) and are variance-normalized.
- **Equal pdf estimation error variance**: We assume that the (stand-alone) single-stream classifiers have the same pdf estimation error variance but different classification errors, i.e., $\sigma_{S1} = \sigma_{S2}$.

3.1.1. Equal Bayes Error

If $p(x_1|w_1) \approx p(x_2|w_1)$ the variance σ^2 of the random variable z is proportional to

$$\sigma^2 \sim \sum_{i=1}^{2}\sum_{j=1}^{2} s_j^2 \sigma_{ij}^2 \tag{12}$$

and it is easy to show that the weights s_j that minimize the variance (and the estimation error) are

$$\frac{s_1}{s_2} = \frac{\sum_{i=1}^{2}\sigma_{i,2}^2}{\sum_{i=1}^{2}\sigma_{i,1}^2} = \frac{\sigma_{S2}^2}{\sigma_{S1}^2} \tag{13}$$

i.e., *the stream weights are inversely proportional to the variance of the pdf estimation error for each stream*. If the pdf estimation error variance in the two stream is equal then stream weights are equal, i.e., no stream weights should be used.

3.1.2. Equal Estimation Error

Minimization of Eq. (11) with respect to s_1 yields

$$D(\sigma_{S1}^2 + \sigma_{S2}^2)s_1^2 + (\sigma_{S1}^2 + \sigma_{S2}^2 - 2D\sigma_{S2}^2)s_1 + \sigma_{S2}^2(D-1) = 0$$

where $D = \ln \frac{p(x_1|w_1)}{p(x_2|w_1)}$. For $\sigma_{S1} = \sigma_{S2}$

$$2Ds_1^2 + 2(1-D)s_1 + (D-1) = 0 \tag{14}$$

i.e., *the optimal stream weights are not a function of the estimation error if the pdf stream estimation variances are equal*; the optimal stream weighs are only a function of D. Note that $\exp(D)$ can be seen as a crude estimate of the ratios of the Bayes errors of the two single-stream classifiers. The solution of the second order equation that minimizes σ^2 is $s_1 = \frac{D-1+\sqrt{1-D^2}}{2D}$ which gives

$$\frac{s_1}{s_2} \approx \frac{p(x_2|w_1)}{p(x_1|w_1)} \quad \text{for} \quad -1.5 \leq \frac{p(x_1|w_1)}{p(x_2|w_1)} \leq 1.5 \tag{15}$$

i.e., in the region of interest *the stream weights should be inversely proportional to the classification error of the single-stream classifiers*. Note that for $|\exp(D)| \geq 2.72$ the estimation error is minimized by setting one of the two stream weights to zero, i.e., if $p(x_2|w_1) >> p(x_1|w_1))$ then $s_1 = 1$ and $s_2 = 0$. These results agree with our intuition and the results from experiments using supervised discriminative algorithm for estimating stream weights.

Note that the choice of weights provided above minimizes the estimation error but not the total error (since the Bayes error increases when using weights). Direct minimization of equation Eq. (6) with respect to s_j is required to find the value of stream weights that minimize the total error. However, the results above hold in the region of interest, i.e., when the classification errors, the estimation variance and the feature dimensions are comparable for the two streams.

3.2. Multi-class Multi-stream Classification

The results presented above can be readily generalized to the multi-class case by considering a class of discriminant functions $f_{ij}(x)$ for each pair of classes w_i and w_j and expressing $P(error) = 1 - P(correct)$. For the multi-class multi-stream case, the analysis in the previous section holds. Specifically, if we assume that the pdf estimation error random variable z_{ij} is independent of class w_i and only dependent on stream j the results in Eq. (13) and Eq. (15) are also valid for the multi-class case. More work is needed to show that these equations hold when the feature vector consists of more than two streams.

4. APPLICATION TO AUDIO-VISUAL RECOGNITION

In this section, the theoretical results of the previous section are applied to the audio-visual automatic speech recognition problem (AV-ASR) using a hidden-Markov model multi-stream recognizer. Note that the dynamic recognition problem is different that the static classification problem: in addition to the classification errors (misrecognitions) there are also insertion and deletion errors that affect the recognition accuracy; the ratio of insertion to deletions errors is controlled by an empirically determined parameter known as the "word insertion penalty". In the experiments that follow, the insertion penalty was selected to minimize the total recognition error (maximize word accuracy).

In the typical AV-ASR evaluation scenario [7], we artificially inject noise at various signal to noise ratio (SNR) levels to the audio signal, thus reducing the performance of the stand-alone single-stream audio recognizer. The (clean) visual feature stream is then combined with the (corrupted) audio stream and two-stream (audio and visual) HMM models λ_{AV}^{SNR} are trained at each audio SNR level. The stream weights for the audio and visual streams, s_A^{SNR} and s_V^{SNR} respectively, are selected to maximize word accuracy of the AV-ASR system on the training set at each SNR level (provided that $s_A^{SNR} + s_V^{SNR} = 1$). Single-stream HMM models are also build from the audio features λ_A^{SNR} at various SNR levels and from the visual features λ_V. Note that models are retrained at each SNR level, thus models and data are "matched".

For the "matched" scenario, we can assume that the pdf estimation error variance for the audio stream is approximately fixed for different levels of additive noise, i.e., the ratio of estimation variance for the audio and visual stream $\sigma_A^{SNR}/\sigma_V = \sigma_A/\sigma_V$ is not a function of SNR. Potentially σ_A and σ_V could be different (especially if the feature stream dimensions are very different). All in all we expect that the optimal stream weights should be given by a combination of Eq. (13) and Eq. (15), i.e.,

$$\frac{s_A^{SNR}}{s_V^{SNR}} \approx \frac{\sigma_V^2}{\sigma_A^2} \frac{p(x_V|w,\lambda_V)}{p(x_A|w,\lambda_A^{SNR})} \tag{16}$$

where x_A and x_V are the audio and visual stream vectors. Note that the approximation holds only in the region where the single-stream classifiers errors are comparable according to Eq. (15). As discussed in the previous section the ratio in Eq. (15) can be approximated by the single-stream classification error ratio, i.e.,

$$\frac{s_A^{SNR}}{s_V^{SNR}} \approx \frac{\sigma_V^2}{\sigma_A^2} \frac{100 - \text{WACC}(\lambda_V, D_{TEST})}{100 - \text{WACC}(\lambda_A^{SNR}, D_{TEST})} \tag{17}$$

where $\text{WACC}(\lambda, D)$ is the percent word accuracy of the model λ evaluated of the data set D. The accuracy of this approximation for continuous HMM-based recognition remains to be tested (see next section).

5. EXPERIMENTAL RESULTS

For the purposes of this experiment we have used CUAVE audio-visual speech database [6]. The subset of the CUAVE database

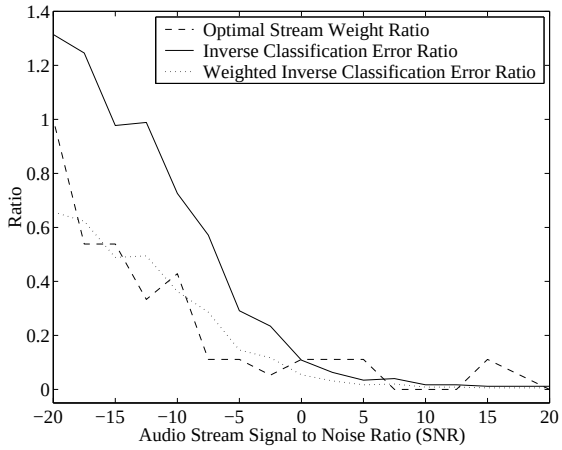

Fig. 1. Optimal audio and visual stream weights ratio s_A/s_V for AV-ASR (dashed line) vs. inverse single-stream word recognition error ratio $\text{WACC}_V/\text{WACC}_A$ (solid line). The weighted (times 0.5) inverse classification error ratio is also shown (dotted line).

used for this experiment consists of videos of 36 persons each uttering 50 connected digits. The training set consists of 30 speakers (and 1500 utterances) and the test set consists of 6 speakers (and 300 utterances). The audio features were the "standard" mel-cepstrum coefficients (MFCCs) and the audio stream dimension was $d_A = 39$ (12 MFCCs, energy, first and second derivatives). The visual features were extracted from the mouth region of each video frame by gray-scaling, down-sampling and performing 2D-Discrete Cosine Transform (DCT). A total of 35 DCT coefficients were kept resulting in $d_V = 105$ (35 DCT coefficients, first and second derivatives). The HMM models were context-independent whole-digit models with 8 states per digit and a single Gaussian per state. The HTK HMM toolkit was used for training and testing. Note that forced alignment (from clean audio single-stream data) was used to train the multi-stream models (no embedded training allowed).

The audio signal was corrupted by additive white noise at various SNR levels; the single-stream audio and two-stream audio-visual models were re-trained at each SNR level. The word accuracy of the single-stream visual recognizer was 42%, while the word accuracy of the single-stream audio recognizer ranged from 23% at -20db SNR to 99% at 20db SNR. The audio and visual recognizers had equal error rates at (approximately) -15 db SNR. The stream weights for the audio-visual recognition system were selected to obtain the best word accuracy on the training set. The ratio of the stream weights s_A/s_V ranged from 1 at -20 db SNR to 0 at 20 db SNR.

The relation between the optimal stream weight ratio and the inverse single-stream recognition error is shown in Fig. 1. It is clear that in the region where the assumption in Eq. (17) holds, i.e., between -20 and -8 db SNR, the two curves are approximately proportional to each other. Indeed, the correlation coefficient between the two curves is 0.96 showing that the linearity assumption holds well in this experiment. The value of $\sigma_V^2/\sigma_A^2 \approx 2$ gives a good match between the two curves in the region of interest. The weighted (times 0.5) inverse classification error ratio is also shown in Fig. 1 (dotted line). Further experimentation is necessary to better understand the validity of Eq. (17) for multi-stream HMM classifiers.

6. CONCLUSIONS

We have presented theoretical and experimental results on the problem of optimal stream weight computation for multi-stream recognition. The optimal stream weights were shown to be inversely proportional to the single-stream classification errors in most practical cases. This result wields much interest for the problem of unsupervised estimation of the optimal stream weights. We are currently working on obtaining estimates of single-stream classification error from test data to address the unsupervised stream weight estimation problem. More work is underway to help us better understand the applicability of the optimal stream weight results to multi-stream recognition using HMM models.

Acknowledgments

This work was partially funded by the EU FP6-IST projects "HIWIRE" and "MUSCLE". The authors wish to thank Prof. Gowdy for the use of the CUAVE database, and Prof. Petros Maragos, Dr. Gerasimos Potamianos and Dr. Dimitris Dimitriadis for many helpful discussions.

7. REFERENCES

[1] D. Dimitriadis, P. Maragos, and A. Potamianos, "Robust AM-FM features for speech recognition," *IEEE Signal Processing Letters*, vol. 12, pp. 621–624, Sept. 2005.

[2] R. O. Duda and P.E. Hart, *Pattern Classification and Scene Analysis.* John Wiley and Sons, New York, 2001.

[3] J. Kittler, M. Hatef, R.P.W. Duin, and J. Matas, "On combining classifiers," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 20, no. 3, pp. 226–239, 1998.

[4] S. Okawa, E. Brocchieri, and A. Potamianos, "Multi–band speech recognition in noisy environments," in *Proc. ICASSP*, 1998.

[5] A. Pangos, *Combining Semantic Similarity Measures for Automatic Induction of Semantic Classes*, M.Sc. Thesis, Technical Univ. of Crete, 2005.

[6] E. K. Patterson, S. Gurbuz, Z. Tufekci, and J. N. Gowdy, "CUAVE: A new audio-visual database for multimodal human-computer interface research," in *Proc. ICASSP*, 2002.

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

[8] G. Potamianos and H. P. Graf, "Discriminative training of HMM stream exponents for audio-visual speech recognition," in *Proc. ICASSP*, 1998.

[9] S. Tamura, K. Iwano, and S. Furui, "A Stream-Weight Optimization Method for Multi-Stream HMMs Based on Likelihood Value Normalization," in *Proc. ICASSP*, 2005.