UNSUPERVISED STREAM WEIGHT ESTIMATION USING ANTI-MODELS

Eduardo Sánchez-Soto, Alexandros Potamianos[†]*, Khalid Daoudi*

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

ABSTRACT

In this paper, a novel solution to the problem of unsupervised stream weight estimation for multi-stream classification tasks is proposed. Our work is based on theoretical results in [10] for the two-class problem were the optimal stream weights are shown to be inversely proportional to the single stream misclassification error. These two-class results are applied to the multi-class problem by using models and "anti-models" (class-specific background models) thus posing the multi-class problem as multiple two-class problems. A non-linear function of the ratio of the inter- to intra-class distance is proposed as an estimate for single stream classification error and used for stream weight estimation. The proposed unsupervised stream weight estimation algorithm is evaluated on both artificial data and on the problem of audio-visual speech recognition. It is shown that the proposed algorithm achieves results comparable to the supervised minimum-error training approach under most testing conditions.

1. INTRODUCTION

Information fusion methods have been extensively employed for speech processing applications in the literature. In this work, the performance of the automatic speech recognition (ASR) systems is improved by using complementary features that are extracted either from the audio and/or the video streams for audio-visual ASR (AV-ASR). In [1], the authors propose an ASR systems based on the multi-band approach, features from different frequency bands have different reliability are combined and weighted accordingly in a multi-stream speech recognition approach. In [2], features such as fundamental frequency are combined with traditional spectral-based features to improve speech recognition performance. Visual information has also been integrated in combination with audio using the multi-stream approach [3]. In this AV-ASR case, audio and video features contain complementary information. In addition, visual information is not affected by adverse recording conditions significantly improving the robustness of the AV-ASR system in noise.

An important problem found in these systems where multiple feature "streams" are employed is the combination of these sources of information. The integration is characterized by the stage at which the information obtained from the different "modalities" are merged. The basic approach is to work at the feature level. In this technique, called Early Integration (EI), the features are concatenated in a single stream [4], which, in the particular case of audio and visual combination, presents a difficulty due to the lack of synchronicity of the flows of information. Alternatively the classifier scores can be combined in the Late Integration (LI) scenarios [5]. In this case, the synchronicity problem is solved but the temporal dependencies are lost.

For different environments and noise conditions, not all the sources of information are equally reliable. For example, for AV-ASR the audio stream is sensitive to adverse recording conditions which may be varying with time. Depending on the type and level of background noise for example the audio stream should be weighted more or less in the decision process. Therefore a mechanism to *adaptively* weight the contribution of the various information sources (feature streams) in the final decision is needed. In the literature, there are well known methods for computing feature stream weights using minimum error classification in a *supervised* manner: the reliability can be obtained directly from the streams through their training error minimization [6, 7]. Alternatively, and for changing recording conditions the reliability of the streams can be computed for each environmental conditions, typically using the signal-to-noise ratio (SNR); the stream weights are then estimated using the SNR estimate in the field [8, 9].

The algorithms proposed above are either supervised or require specific knowledge of the conditions in the field. In this work, we propose stream weight estimates that can be computed in an unsupervised way (no class labels required or field conditions). The proposed algorithm builds on prior theoretical work on optimal stream weight estimation [10] where it is shown that stream weights should be approximately inversely proportional to the single stream classification error. We propose estimates of the single stream classification error using limited amount of unlabeled data and show that the proposed stream weight estimates perform well for both artificial and real data for an audio-visual speech classification task.

2. MULTI-STREAM CLASSIFICATION

For the two class $\{w_1, w_2\}$ problem the feature pdfs and class prior probabilities are $\{p(x|w_1), p(x|w_1)\}$ and $\{p(w_1, p(w_2)\}$ respectively. See Figure 1 for a 2-D two-class classification problem visualization. Assuming a random variable z_i that follows a normal distribution with mean zero and variance $\sigma_i^2, \mathcal{N}(z; 0, \sigma^2)$ that model the estimation/modeling error

$$p(w_i|x, \lambda) - p(w_i|x) = z_i, \qquad (1)$$

where λ stands for the selected model/estimation method, and the estimated and real distributions are $p(w_i|x, \lambda)$ and $p(w_i|x)$ respectively. As it was explained in [10] the deviation from the optimal boundary value is given by the random variable $z = (z_1 - z_2)p(x)$. For a feature vector broken up into two independent streams $\{x_1, x_2\}$, with dimension $\{d_1, d_2\}$ and stream weights $\{s_1, s_2\}$ and under the assumption of a constant $p(x)$ in the region of interest the total (Bayes, estimation and modeling) error

This work was partially funded by MUSCLE and HIWIRE EU FP6-IST projects. The authors wish to thank to Prof. Gowdy for the database.

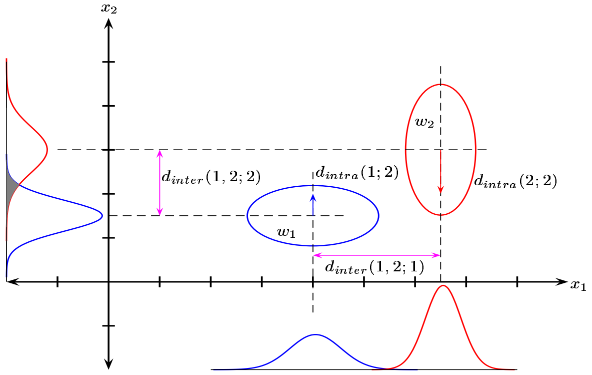

Fig. 1. *2-D two-class classification.*

can be computed and minimized as follows. Assuming that the Bayes error increase is small compared to the decrease in estimation/modeling error and that the estimation error for the i^{th} class and j^{th} stream is a random variable that follows a normal distribution $z_{ij} = \mathcal{N}(z; 0, \sigma_{ij}^2)$, z is:

$$z \approx 2[p(x_1|w_1)p(w_1)]^{s_1}[p(x_1|w_2)p(w_2)]^{s_2} \\ [s_1(z_{11} - z_{21}) + s_2(z_{12} - z_{22})], \tag{2}$$

and its variance is :

$$\sigma^2 \sim p(x_1|w_1)^{2s_1} p(x_2|w_1)^{2s_2} [s_1^2 \sigma_{S_1}^2 + s_2^2 \sigma_{S_2}^2], \tag{3}$$

where $\sigma_{S_j}^2 = \sum_{i=1}^2 \sigma_{ij}^2$ is the total stream variance.

From the last equation it can be observed that the error can be reduced by employing weights if the estimation errors are different and/or if the single-stream classification errors are different. These two factors correspond to the two cases that follow. First, assuming the same Bayes classification error, $p(x_1|w_1) \approx p(x_2|w_1)$, the optimal weights can be computed as:

$$\frac{s_1}{s_2} = \frac{\sum_{i=1}^2 \sigma_{i,2}^2}{\sum_{i=1}^2 \sigma_{i,1}^2}. \tag{4}$$

Second, assuming the same single-stream estimation error, $\sigma_{S_1} = \sigma_{S_2}$, the optimal weights are

$$\frac{s_1}{s_2} \approx \frac{p(x_2|w_1)}{p(x_1|w_1)}. \tag{5}$$

where $p(x_1|w_1)$, $p(x_2|w_1)$ are computed close to the decision boundary. As discussed in [10] the quantities above approximate the misclassification error for streams 1 and 2 respectively, and thus *the stream weights should be approximately inversely proportional to the single stream misclassification error*. Furthermore the weights are constrained as follows:

$$s_1 + s_2 = 1, \quad 0 \leq s_1, s_2 \leq 1. \tag{6}$$

Next we apply these theoretical results to the problem of unsupervised stream weight estimation.

3. UNSUPERVISED STREAM WEIGHT ESTIMATION

The problem at hand is stream weight estimation for multi-stream classification *in the field*. For example, for the problem of audio-visual speech recognition it is common that the recording conditions in the field are both time-varying and different from the conditions under which the acoustic models were trained. In this case, the stream weights for the audio and video streams have to be adapted to their optimal values without knowledge of the transcription or "class labels". Our goal is to devise robust algorithms for estimating the stream weights using small amounts of unlabeled data, i.e., unsupervised stream weight estimation. For the speech recognition example stream weights are estimated at a *per-utterance* basis.

We attack the problem of unsupervised stream weight estimation using the theoretical results summarized in the previous section as our guide. However, these results are not directly applicable to our problem due to two main reasons: (i) only results for the two-class classification problem are available, while in general the multi-class classification problem is of interest, and (ii) knowledge of class membership for each observation vector x is required to compute the likelihoods in equation Eq. 5, i.e., the theoretical results are directly applicable only to the *supervised* stream weight estimation problem.

To resolve the first issue we introduce the concept of *anti-models*[1]. Specifically, during training and for each class we separate the training data into two groups: one containing the training examples of the class of interest and the other containing the rest of the training examples. Models and an "anti-models" are built from the two training sets; anti-models can be though of as a class-specific "background/garbage" models. By creating models and anti-models the multi-class classification problem is reposed as (multiple) two-class classification (problems).

To resolve the second issue the single stream misclassification error has to be estimated in an unsupervised way. It is well known, that for the two class classification problem, when $p(x|w_i)$ follow Gaussian distributions $\mathcal{N}(\mu_i, \sigma^2)$, the Bayes error is a function of $D = |(\mu_1 - \mu_2)|/\sigma$. In general, the quantity D can be estimated in an unsupervised way, by performing k-means classification and then using the inter- and intra-class distances to estimate the quantities in the nominator and denominator respectively. Indeed the intra-class distance is the average distance between the means of each class and the intra-class distance an estimate of the average class variance. In our case, the mean of the model and anti-model are used to initialize the k-means algorithm ($k = 2$) for each class; the estimated D's are then averaged over all classes.

To gain better insight into the use of the inter- to intra-class ratio see Fig. 1. A two-stream two-class classification problem is outlined: axes x_1 and x_2 correspond to the features in the two streams; the (Gaussian) distributions for classes w_1 and w_2 are shown for each stream and jointly. The relationship between the Bayes error (shaded area) and the inter- and intra distances is inversely and directly proportional respectively.

Overall, the stream weights are computed using the inter-class distance $d_{inter}(l, m; j)$ between classes l and m for stream j, normalized by the intra-class distance $d_{intra}(i; j)$ for the class i in each stream. For the two-stream two-class case the stream weights s_1, s_2 are estimated as:

$$\frac{s_1}{s_2} = c\ f\left(\frac{d_{inter}(1,2;j) / \sum_i d_{intra}(i;j)|_{j=2}}{d_{inter}(1,2;j) / \sum_i d_{intra}(i;j)|_{j=1}} \right), \tag{7}$$

where $f(.)$ is a nonlinear function that relates D with the Bayes error (erf function) and c is a constant accounting for the difference in

[1] Anti-digit models have been employed in utterance verification [13].

estimation error in the two streams (see Eq. (4)). For the two-stream multi-class case the quantity in function $f(.)$ becomes

$$\sum_k \left(\frac{d_{inter}(m_k, am_k; j) / \sum_{i=(m_k,am_k)} d_{intra}(i;j)|_{j=2}}{d_{inter}(m_k, am_k; j) / \sum_{i=(m_k,am_k)} d_{intra}(i;j)|_{j=1}} \right), \tag{8}$$

where m_k and am_k are the centroids for the “model” and “anti-model”[2] for class k and $\sum_k$ is over all classes.

Here are the main assumptions underlying the proposed unsupervised stream weight estimation method: (i) Two-class classification error can be approximated as a function of inter- to intra-class distance ratio. (ii) Multi-class classification error can be estimated by the class/anti-class classification error averaged across all classes. (iii) Single stream estimation error variance is approximately constant for each stream under all field conditions. We proceed next to experimentally verify the validity of these assumption both for artificial and real data.

4. ARTIFICIAL DATA EVALUATION

For the artificial data experiments the next table summarizes the employed parameters for the 1-D Gaussian distributions for two class $\{w_1, w_2\}$ problem. A number N of samples was generated using those parameters and the total classification error was computed for different weights. The samples were used to estimate the distributions for the two classes by a clustering process. The k-means process with $k = 2$ was employed to cluster the samples. The estimated clusters were used to compute the distances as it was explained in the previous section. In Figure 2, the results obtained using the parameters specified in the Table. 1 are presented. In each figure, the wither lines (black) represent the estimated distributions and a line (green) connects the k-mean estimated centroids. The thin lines (blue and red) represent the real distributions. In the figure on the left, computing the total error, the optimal value was equal to 0.5 and the computed value for the weight s_1 was 0.47 using the distances approach proposed in the previous section. It can be seen that the true and estimated values are close. In the figure on the right, one of the classes mean was moved from $\mu_{21} = 6$ to $\mu_{21} = 8$. Employing the total error computation the obtained optimal weight s_1 was 0.6. Using the distances approach the obtained value was 0.65. Overall, for artificial data and for the two-class two-stream problem, the proposed approach gives satisfactory results.

	w_1		w_2	
x_1	$\mu_{11} = 4.0$	$\sigma_{11}^2 = 2.0$	$\mu_{21} = 6.0$	$\sigma_{21}^2 = 1.5$
x_2	$\mu_{12} = 1.5$	$\sigma_{12}^2 = 1.5$	$\mu_{22} = 3.5$	$\sigma_{22}^2 = 2.0$

Table 1. *Parameters for the Gaussian distributions; two classes (w_1, w_2) and two streams stream (x_1, x_2).*

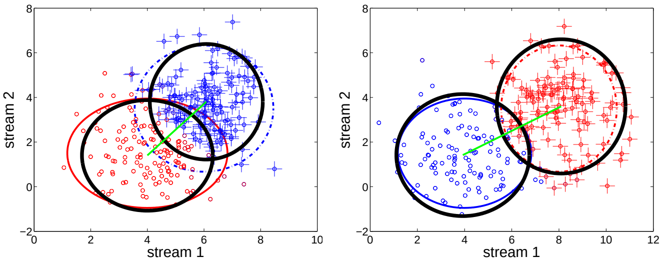

Fig. 2. *Clustering and distance computation representation for the two classes problem (two example cases).*

5. AUDIO-VISUAL SPEECH CLASSIFICATION

To verify our claims a set of experiments using real data were performed. An audio-visual speech classification task was investigated were the two feature streams contain audio and visual information respectively.

For the purposes of this experiment the CUAVE audio-visual speech database was employed [11]. The subset of the CUAVE database used in these experiments consists of videos of 36 persons each uttering 50 connected digits. The training set is made up of 30 speakers (1500 utterances) and the test set contains 6 speakers (300 utterances). The audio signal was corrupted by additive babble noise at various SNR levels; the video signal was clean in all the experiences. The audio features used were the “standard” Mel-Frequency Cepstrum Coefficients (MFCC) for frames with a duration of 20 ms extracted every 10 ms. The acoustic vectors, dimension $d_A = 39$, consist of 12-dimensional Mel-frequency cepstral coefficients (MFCCs), energy, and their first and second order derivatives. The visual features were extracted from the month region of each video frame by gray-scaling, down-sampling and finally performing a 2-D Discrete Cosine Transform (DCT). The first 13 most energetic DCT coefficients within the odd columns were kept [12] resulting in a video feature vector of dimension $d_V = 39$ including the first and second order derivatives. Hidden Markov Models (HMMs) were used for both acoustic and video model training. Context-independent whole-digit models with 8 states per digit and a single Gaussian continuous density distribution per state were used. The HTK HMM toolkit was used for training each stream, audio and video, and for also for testing (using HTK's built-in multi-stream capabilities).

An important part of the training process is the generation of “anti-models” [13]. The class and anti-class models are both built in the training phase using only “clean” data. The class model for each stream is built following the traditional training process. The anti-class models are trained using all the data that does not belong to the corresponding class. For example, the model for the digit *one* is created using all training data labeled as *one*, while the anti digit *one* is learned using all the data not labeled as *one*. At the end of this process 20 model are obtained for each stream, ten models for the digits (0-9) and ten anti-digits all with the same number of parameters.

During the test phase these class and anti-class models are used to initialize the k-means classification. Specifically, the means of the Gaussian distribution in the class and anti-class model are used as the initial k-mean centroids[3]. Given that *a-priori* it is not known to which class each utterance belongs, the features in each utterance are split into two classes ($k = 2$) in ten different ways one for each digit and anti-digit model. The stream weights are estimated using Eqs. (7),(8). The inter- d_{inter} and intra-class d_{intra} distance is com-

[2] m_k and am_k are computed in an unsupervised way using k-means initialized from the “mis-matched” model and anti-model means; thus a more appropriate term might be “adapted” model and anti-model centroids.

[3] It is important to remark that these anti-class models are only used to initialize the clustering process and that the models are trained using data recorded in “clean”conditions very different than the conditions in the field.

puted for each of the ten splits and the resulting inter- to intra-class ration is averaged over the ten splits. Note that the stream weights are estimated for *each utterance*.

In Figure 3, the digit classification results are shown for various stream weight estimation algorithms. The wider curve (green) represents the results obtained searching by hand for the optimal values for the weights. The lower curve (black) in the left uses equal weights in both streams. These two curves represent the reference used to compare our approach. In a first case, with the dashed curve (blue), it can be observed that the weights give an improvement in the results when the SNR is high. To take into account the estimation error a constant c (see Eq. (7)) can be estimated on held-out data and used to improve the results; this is represented with the dashed-dotted curve (red). As seen in Eq. (7), the optimal weights are a non-linear function $f(.)$ of the distances. The dotted curve (magenta) shows the results obtained using a nonlinear transformation of the weights, a parabola in this case (similar to the erf function). This last curve provides a good match between the D value and the Bayes error and results in performance comparable to the hand-picked optimal stream weight values.

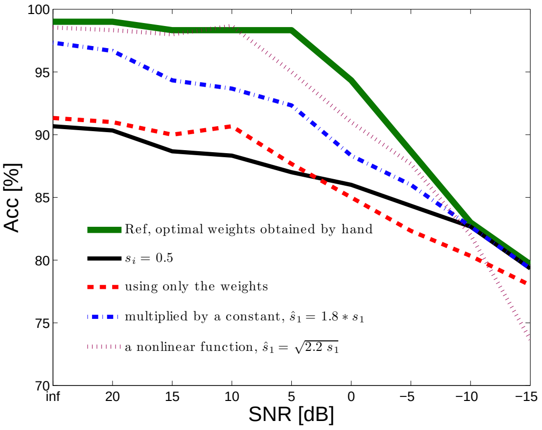

Fig. 3. *Digit recognition accuracy as a function of the SNR for the proposed stream weights computation.*

6. CONCLUSIONS

In this paper, we proposed a stream computation method for a multi-class classification task based on theoretical results obtained for a two classes classification problem and making use of an anti-model technique. The proposed method employs only the information contained in the trained models and requires a single utterance to compute the stream weights. Therefore the obtained results are of interest for the problem of unsupervised estimation of optimal stream weights for multi-streams classification and recognition problems. The proposed method achieved comparable performance with supervised minimum error estimation of the weights. In future work, the problem of unsupervised weight estimation for statistical recognition tasks will be addressed, as well instantaneous stream weight estimation.

7. REFERENCES

[1] H. Bourlard and S. Dupont, “A new ASR approach based on independent processing and recombination of partial frequency bands,” Philadelphia, USA, October 1996, ICSLP, vol. 1.

[2] T.A. Stephenson, M. Mathew, and H. Bourlard, “Modeling Auxiliary Information in Bayesian Network Based ASR,” Scandinavia, 2001, Eurospeech.

[3] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A.W. Senior, “Recent Advances in the Automatic Recognition of Audiovisual Speech,” *Proceedings of the IEEE*, vol. 91, no. 9, September 2003.

[4] C. C. Chibelushi, J.S. Mason, and F. Deravi, “Integration of acoustic and visual speech for speaker recognition,” Berlin, Germany, September 1993, Eurospeech, pp. 157–160.

[5] M. Heckmann, F. Berthommier, and K. Kroschel, “Noise adaptive stream weighting in audio-visual speech recognition,” *EUROASIP, Journal of Applied Signal Processing*, vol. 1, pp. 1260–1273, November 2002.

[6] G. Potamianos, J. Luettin, and C. Neti, “Hierarchical discriminant features for audio-visual LVCSR,” Salt Lake City, USA, May 2001, ICASSP, vol. 1, pp. 165–168.

[7] S. Tamura, K. Iwano, and S. Furui, “A Stream-Weight Optimization Method for Multi-Stream HMMs Based on Likelihood Value Normalization,” Philadelphia, USA, March 18-23 2005, ICASSP.

[8] A. Rogozan, P. Deléglise, and M. Alissali, “Adaptive determination of audio and visual weights for automatic speech recognition,” Rhodes, 1997, Proc. Europ. Tut. Works. Audio-Visual Speech Process, pp. 61–64.

[9] H. Glotin, D. Vergyri, C. Neti, G. Potamianos, and J. Luettin, “Weighting Schema for Audio-Visual Fusion in Speech Recognition,” Salt Lake City, USA, May 2001, ICASSP.

[10] A. Potamianos, E. Sánchez-Soto, and K. Daoudi, “Stream Weight Computation for Multi-Stream Classifiers,” Toulouse, France, April 2006, ICASSP.

[11] E.K. Patterson, S. Gurbuz, Z. Tufekci, and J. N. Gowdy, “CUAVE: A new audio-visual database for multimodal human-computer interface research,” Orlando, Florida, May 13-17 2002, ICASSP.

[12] G. Potamianos and P. Escanlon, “Exploiting Low Face Symmetry in Appearance-Based Automatic Speechreading,” British Columbia, Canada, July 24-27 2005, Auditory-Visual Speech Processing, AVSP.

[13] M.G. Rahim, C.H. Lee, and B.H. Juang, “Discriminative Utterance Verification for Connected Digits Recognition,” *IEEE Transactions on Speech and Audio Processing*, vol. 5, no. 3, May 1997.