UNSUPERVISED STREAM WEIGHT COMPUTATION IN A SEGMENTAION TASK: APPLICATION TO AUDIO-VISUAL SPEECH RECOGNITION

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

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

ABSTRACT

We propose an efficient algorithm for unsupervised stream weight estimation in a segmentation task. Our method uses only the information carried by the test signal and the trained models. The work is based on results presented in [1, 2] for the classification problem where it is indicated that the optimal stream weights are inversely proportional to the single stream misclassification error. We approximate this error relation by the intra- and inter-class distance ratio over the measured class distributions. This approach is then generalized to the segmentation problem by computing the distances among all the concerned classes. The proposed unsupervised estimation algorithm is evaluated on a an audio-visual speech recognition task. The obtained performances are comparable to the supervised minimum error training approach, up to a certain SNR level.

***Index Terms*—** Fusion Methods, Stream Weight Estimation, Audio-Visual Speech Recognition.

1. INTRODUCTION

Although common sens indicates that use of all the complementary information, which can come from different sources, should produce the best performance, in real applications information fusion is not an easy problem. Based on this idea fusion methods have been employed in many scientific areas. In speech processing applications, for example, the literature shows Automatic Speech Recognition (ASR) systems been improved by using complementary features: prosodic, and visual information [3, 4]. Each extra information palliate an insufficiency in a given circumstance. In the case of Audio Visual-ASR systems the visual information, lips and oral cavity movements, is not affected by adverse acoustic recording conditions significantly improving the robustness of the ASR system.

An important problem found in this kind of systems is the procedure for combining all the sources of information. If this process is done in incorrect context and conditions, because of the incorrect assumptions, the obtained fused system can obtains catastrophic results. The integration strategy 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 unique feature vector. In the particular case of combining audio and visual information this approach presents a difficulty that is due to the lack of synchronicity between the audio and the visual flows. To overcome this problem the integration can be done at the decision level, called Late Integration (LI). Although, all the temporal dependencies between the sources of information are lost.

Unfortunately, not all the sources of information are equally reliables for different environments and noise conditions. Therefore a mechanism for weighting their contribution to the final decision is needed. This reliability can be obtained directly from the streams through their performances minimizing the likelihood or using a Generalized Probabilistic Descent (GPD) algorithm [5]. Another approach compute the value related to the reliability of the streams from the environmental conditions based, in general, on the SNR. This value is estimated in a direct or indirect manner. It can be said that the systems, in this approach, make an adaption of the modalities to a given context. Degree of voicing presented on the utterance, [6], the difference of probability among the first N candidates [7], or the minimization of the misclassification error on a held-out data set [8] are used as a measure of the audio stream reliability.

The proposed algorithms above are either supervised or require specific knowledge of the conditions in the field. Therefore, in [2], to overcome the lack of knowledge in real applications, we proposed stream weights estimates in an unsupervised manner (no class labels or field conditions) by using an anti-models approach for the multi-class classification problem based on theoretical results presented in [1]. In the present paper we extend this work to the segmentation problem. Estimates of the stream

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

weights are obtained by generalizing the computed distances among all the concerned classes. We show that the proposed approach performs well on real data for an audio-visual speech recognition task.

The organization of the remainder of this paper is as follows: Section 2 describes the theoretical results to compute the optimal stream weights. Section 3 and 4 describe the practical method employed to computed the weights in real conditions. Section 5 describes the corpora and results for the Audio-Visual recognition experiments and Section 6 presents conclusions and proposals for future work.

2. STREAM WEIGHTS COMPUTATION

In [1] it was presented a theoretical approach to compute the stream weights $\{s_1, s_2\}$ that reduce the total error for a two class $\{w_1, w_2\}$, two streams $\{x_1, x_2\}$ classification problem (see Figure 1).

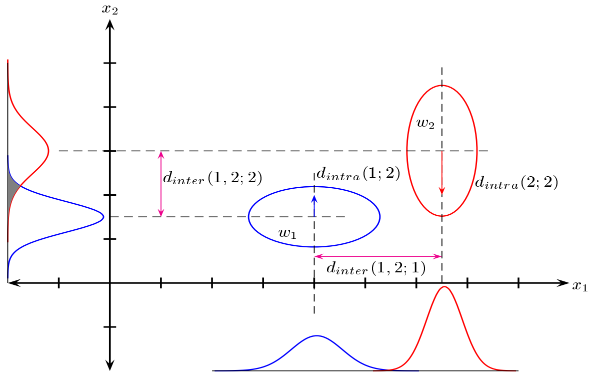

Figure 1. *2-D two-class classification problem representation.*

Stream weights $\{s_1, s_2\}$ are used to "equalize" the probability in each stream, i.e,

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

Estimation/modeling error, z_{ij}, for the i^{th} class and j^{th} stream, is computed as being the difference between the estimated $p(w_i|x_j, \lambda)$ and the real $p(w_i|x_j)$ distributions:

$$p(w_i|x_j, \lambda) - p(w_i|x_j) = z_{ij}, \tag{2}$$

where λ designates the selected estimation/modeling method. The random variable modeling this difference is assumed to follows a normal distribution with zero mean $\mu = 0$ and variance σ_i^2, $\mathcal{N}(z; 0, \sigma^2)$. Therefore, a reduction on the variance σ_i^2 will reduce the error. It was shown that the deviation from the optimal decision boundary is given by the random variable z in Eq. (3).

$$\prod_{j=1} 2\,[p(x_j|w_1)p(w_1)]^{s_j} - \prod_{j=1} 2\,[p(x_j|w_2)p(w_2)]^{s_j} + z \overset{\geq}{<} 0, \tag{3}$$

Assuming that the posterior probabilities are equal for both classes in the decision region, the random variable z is given by the next equation:

$$\begin{aligned} z \approx\ & 2[p(x_1|w_1)p(w_1)]^{s_1}[p(x_2|w_1)p(w_1)]^{s_2} \\ & [s_1(z_{11} - z_{21}) + s_2(z_{12} - z_{22})]. \end{aligned} \tag{4}$$

The variance of z is, in this case, expressed as follows:

$$\sigma 2 \sim p(x_1|w_1)^{2s_1} p(x_2|w_1)^{2s_2}[s_1^2\sigma_{S_1}^2 + s_2^2\sigma_{S_2}^2], \tag{5}$$

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

Two main factors can be remarked on the last equation (Eq. 5). The first factor is related to the estimation error and the second factor is related to the single-stream classification errors. In the first case, assuming the same Bayes classification error, $p(x_1|w_1) \approx p(x_2|w_1)$, the optimal weights are:

$$\frac{s_1}{s_2} = \frac{\sum_{i=1}^{2} \sigma_{i,2}^2}{\sum_{i=1}^{2} \sigma_{i,1}^2}. \tag{6}$$

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

$$\frac{s_1}{s_2} \approx \frac{p(x_2|w_1)}{p(x_1|w_1)}. \tag{7}$$

As discussed in [1] the quantities in the previous equation 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*.

3. UNSUPERVISED WEIGHTS COMPUTATION FOR CLASSIFICATION

In real-world applications several problems have to be faced. For instance it is frequent that recording conditions, which can be time-varying during the training process, are different to those present during the test process. In such case, the stream weights should be adapted to the optimal values without any prior knowledge using only the small quantity of unlabeled data available in each test utterance.

In addition to this problem, the theoretical results above are not directly applicable to classification problems for 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 and/or the single-stream modeling/estimation error is required to compute the weights.

To solve the first problem, we used the principle of anti-models in [2]. By creating models and anti-models, the multi-class classification problem is transformed to multiple two-class classification problems.

To solve the second problem, we approximated in [2] the ratio in Eq. (7) by measurable quantities that can be computed in an unsupervised manner. Indeed, it is well known that, if $p(x|w_i)$ follow a Gaussian distributions $\mathcal{N}(\mu_i, \sigma 2)$, the Bayes error is a function of $D = |(\mu_1 - \mu_2)|/\sigma$, which can be computed in an unsupervised manner using the inter- and intra-class distances (see Figure 1). The intra-class distance is an estimate of the average class variance and the inter-class distance is the average distance between the means of each class. These distances are used to measure the separation and the overlapping between the class distributions, and they are directly dependent of the centroids of each class. In our approach, the Gaussian-means of the (HMMs) model and anti-model are used to initialize the k-means algorithm ($k = 2$) for each class to obtain the centroids and then compute the inter- and intra-distances (see Figure 1).

Finally, stream weights are computed using the intra- and inter-class distance, $d_{inter}(k,l;j)$, for the classes k and l in the stream j normalized by the intra-class distance, $d_{intra}(i;j)$ for the class i in the same stream j as it is expressed in the next equation :

$$\frac{s_1}{s_2} = c\, f\left(\frac{d_{inter}(1,2;j)/\sum_{i=1}^{2} d_{intra}(i;j)|_{j=2}}{d_{inter}(1,2;j)/\sum_{i=1}^{2} d_{intra}(i;j)|_{j=1}}\right), \tag{8}$$

It is important to remember that this process involves the creation of one anti-model for each class. In consequence, in the previous process one k-means algorithm have to be performed for each couple model-anti-model in order to obtain the desired weights.

4. UNSUPERVISED WEIGHTS COMPUTATION FOR SEGMENTATION

In this section, we extend the multi-class classification approach presented in the previous section to the segmentation problem in order to apply it to audio-visual speech recognition. Although segmentation is in principle a more difficult problem than classification, our extension makes it actually easier to implement stream-weights computation in segmentation, and without the need of anti-models.

We do so based on the observation that larger separation between class distributions in a given stream increase its discriminative power. Concretely, the inter-class distance can be computed among all the classes by measuring the inter-class distance for each pair of classes in order to measure their separation. The intra-class distances, as before, is measured to obtain an estimate of their variance. Altogether, the inter- and intra-class distances, computed in this manner, yields an estimate of the missclassification error over all the classes without need of any associated anti-model.

As delineated in the previous section the initial centroids are obtained directly from the (HMM) models learned in training. This time the k-means algorithm is performed over all the classes only one time (k this time is the number K of classes). The total inter-classes distance, $d_{inter}(j)$ is computed by adding the inter-class distance over all the possible combinations of two classes:

$$d_{inter}(j) = \sum_{k=1}^{K} \sum_{l=k+1}^{K} d_{inter}(k,l;j), \tag{9}$$

where K is the total number of classes.

Finally, the stream weights are computed as in Eq. (8) but using the total inter-class distance $d_{inter}(j)$ normalized by the sum of the intra-class distances $d_{intra}(i;j)$ in the corresponding stream j as follows:

$$\frac{s_1}{s_2} = c\, f\left(\frac{d_{inter}(j)/\sum_{i}^{K} d_{intra}(i;j)|_{j=2}}{d_{inter}(j)/\sum_{i}^{K} d_{intra}(i;j)|_{j=1}}\right). \tag{10}$$

The algorithm for weights computation, displayed in Figure 2, can be summarized as follows:

- Provide initial centroids for k-means.
- Perform k-means over all classes using test data.
- Compute inter- and intra-class distances.
- Estimate stream weights (Eq. (10)).

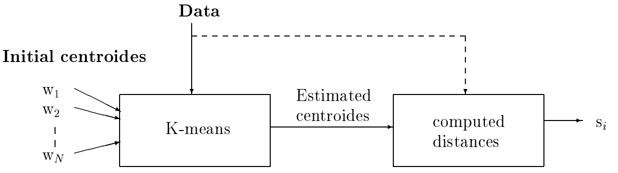

Figure 2. *General stream weights estimation process.*

5. AUDIO-VISUAL SPEECH RECOGNITION

A set of experiments using real data were performed to verify our affirmations. A two streams, audio and visual, recognition task was investigated. For the purposes of this experiment the CUAVE audio-visual speech database was employed [9]. 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 experiments. The audio features used were the standard Mel-Frequency Cepstrum Coefficients (MFCCs) computed for frames with duration 20 ms, extracted every 10 ms. The acoustic vectors, dimension $d_A = 39$, consist of 12-dimensional MFCCs, energy, and their first and second order derivatives. The visual features were extracted from the mouth 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 [10] 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 also for testing (using HTK's built-in multi-stream capabilities). In all the experiments the results are given as a function of the audio SNR ({infinity 20 15 10 5 0 -5 -10 -15} all are given in dB). Only two steps of the K-means algorithm are performed to compute the estimated centroids. It is important to remark that only models trained with clean data are used.

Figure 3 presents digit recognition results for various stream weight estimation methods.

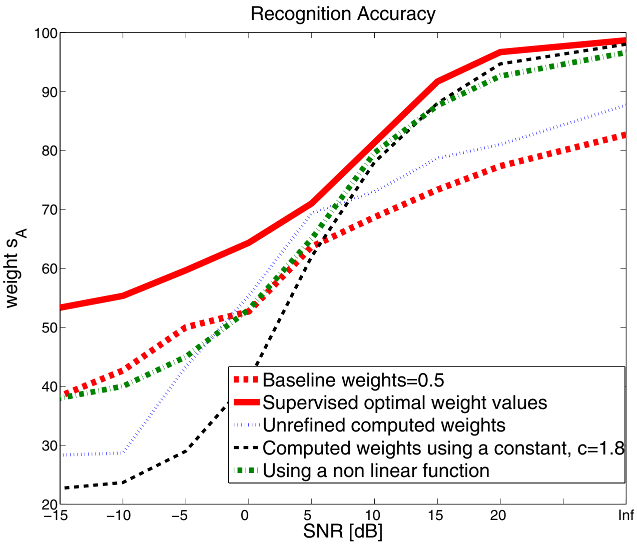

Figure 3. *Speech Recognition Accuracy: Optimal supervised, Baseline and obtained using the proposed method.*

The width solid curve (red) represents the results obtained searching (by hand) in a supervised manner the optimal weight values. The width-dashed curve (red) uses equal weights in both streams ($s_1 = s_2 = 0.5$). These two curves serve as reference and are used to evaluate our approach. The first (and crudest) stream weight estimate is shown with the dotted curve (blue) and corresponds to Eqs. (10) with $c = 1$ and $f(.)$ being the identity function. Even this crude estimate improves the equal weighting scheme (over most SNR ranges). To take into account the estimation error a constant c (see Eq. (10)) can be estimated on held-out data and used to improve the results; this is represented with the dashed-dotted curves (black). This curve, where $c = 1.8$, shows that the performance is close to the optimal supervised one up to a 10dB SNR. However, below 10dB the performance is poor which shows that, as suggested by the theory (and as observed in classification), the relationship (the function f) is non-linear in low SNR regions. This is confirmed by the dash-dotted (green) curve, where a parabola function f is used (as in classification [2]). One can see that the performance improves in low SNR, still the results are not as satisfactory as the ones we obtained in the classification problem [2].

6. CONCLUSIONS

In this paper, we proposed an efficient unsupervised stream weights computation algorithm for a segmentation task. This new algorithm is based on previous results in a classification problem that make the use of distances, inter and intra, among the concerned classes. 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 stream weights for multi-streams segmentation/recognition problems. We applied the new algorithm to an audio-visual speech recognition task at different (audio) SNR levels. The proposed method achieved comparable performance to the supervised minimum error estimation of the weights, up to a certain SNR level. In future work, the search of the "optimal" nonlinear function $f(.)$ will be addressed, as well instantaneous stream weight estimation.

7. REFERENCES

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

[2] E. Sánchez-Soto, A. Potamianos, and K. Daoudi, "Unsupervised Stream Weight Estimation Using Anti-Models." Honolulu, Hawai'i, U.S.A.: ICASSP, April 2007.

[3] T. Stephenson, M. Mathew, and H. Bourlard, "Modeling Auxiliary Information in Bayesian Network Based ASR." Scandinavia: Eurospeech, 2001.

[4] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. Senior, "Recent Advances in the Automatic Recognition of Audio-visual Speech," *Proceedings of the IEEE*, vol. 91, no. 9, September 2003.

[5] G. Potamianos and H. Graf, "Discrimative Training of HMM Stream Exponents for Audio-Visual Speech Recognition." ICASSP 1998.

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

[7] A. Adjoudani and C. Benoit, "On the integration of Auditory and Visual Parameters in an HMM-based ASR." Speechreading by humans and Machine, Springer 1996.

[8] G. Potamianos and C. Neti, "Stream Confidence Estimation for Audio-Visual Speech Recognition," vol. III. Beijing: ICSLP 2000, 2000, pp. 746–749.

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

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