A Novel Criterion for Classifiers Combination in Multi-Stream Speech Recognition

Fabio Valente, *Member, IEEE*

***Abstract*—In this paper we propose a novel information theoretic criterion for optimizing the linear combination of classifiers in multi stream automatic speech recognition. We discuss an objective function that achieves a trade-off between the minimization of a bound on the Bayes probability of error and the minimization of the divergence between the individual classifier outputs and their combination. The method is compared with the conventional inverse entropy and minimum entropy combinations on both small and large vocabulary automatic speech recognition tasks. Results reveal that it outperforms other linear combination rules. Furthermore we discuss the advantages of the proposed approach and the extension to other (non-linear) combination rules.**

***Index Terms*—Multi-stream speech recognition, Classifiers combination.**

I. Introduction

Multi-band and multi-stream [1], [2] speech recognition are based on the combination of information obtained from different feature streams, and are typically used for increasing the robustness of Automatic Speech Recognition (ASR) systems in noisy or mismatched conditions. The rationale behind multi-stream approaches is that, in adverse conditions, different streams will be affected in different ways. The combination method should be able to select dynamically the streams that are least affected . This work builds on the same framework proposed in [1], [2] in which several Multi Layer Perceptron (MLP) classifiers are trained in order to discriminate between phonemes using different input features. The MLP output consists of phoneme posterior probabilities that can be combined according to probabilistic rules. The combination involves two tasks:

1 Determining a confidence measure for each feature stream.
2 Defining a rule for combining the different streams according to their confidence measure.

Typical rules for classifiers combination are linear weighting, product, majority voting, maximum and minimum rules (see [3]). We will focus here on the case of *linear* classifier combination. An effective approach for determining the confidence of each stream is based on the use of the entropy of the MLP output [4]. For example *inverse entropy* combination sets the weights of the linear combination inversely proportional to the value of the entropy.

In this paper we propose a criterion that models the trade-off between the linear averaging of the posterior probabilities (i.e. the sum rule [3]) and the minimization of the Bayes probability of error.

F. Valente is with IDIAP Research Institute, Rue Marconi 19 Case Postale 592 , CH-1920, Martigny, Switzerland email: fabio.valente@idiap.ch. Paper submitted 23 January 2009, revised 25 February 2009.

II. Motivations

Let us denote two different feature streams by X_a and X_b and a set of k phonetic targets by $\Theta = \{\theta_i\}$. In the following, we will consider the combination of only two sets of features without loss of generality. Let us train two MLPs according to [5] using X_a and X_b as input features; they will produce phoneme posterior probabilities $\{p_{i\,a} = p(\theta_i|X_a)\}$ and $\{p_{i\,b} = p(\theta_i|X_b)\}$ with $i = 1, ..., k$.

The linear combination of posterior estimates $p_{i\,a}$ and $p_{i\,b}$ can be written as:

$$p_{i\,c} = \omega_a\, p_{i\,a} + \omega_b\, p_{i\,b} \;\text{ with }\; \omega_a + \omega_b = 1 \tag{1}$$

where $\omega_a, \omega_b \geq 0$, $\omega_a = p(X_a)$ and $\omega_b = p(X_b)$. If $\omega_a = \omega_b = 0.5$ i.e. X_a and X_b receive equal weights, the combination is simply the linear average of the two posterior estimates, i.e., the sum rule [3]. In [4], it was observed that the value of the entropy of the MLP output $H(p) = -\sum_i p_i log(p_i)$ increases with the SNR, meaning that the posterior estimate $p(\Theta|X)$ converges towards a uniform, non-informative distribution over the phonemes. Thus entropy values $H(p_a) = -\sum_i p_{i\,a} log\, p_{i\,a}$ and $H(p_b) = -\sum_i p_{i\,b} log\, p_{i\,b}$ can provide a confidence measure related to how feature streams X_a and X_b are affected by the noise. Those findings inspired two weighting schemes referred as *minimum entropy* and *inverse entropy* combination [4].

In *minimum entropy* combination, the stream with the minimum entropy receives weight one i.e.

$$\begin{aligned} &\omega_a = 1\,,\, \omega_b = 0 \;\text{ if }\; H(p_a) < H(p_b) \\ &\omega_a = 0\,,\, \omega_b = 1 \;\text{ if }\; H(p_a) > H(p_b) \end{aligned} \tag{2}$$

This is equivalent to selecting the feature stream with the lowest entropy thus the more confident. If $H(p_a) = H(p_b)$, the method randomly select one of the streams. In *inverse entropy* combination, the weights are set inversely proportional to the value of the entropy i.e.

$$\omega_a = \frac{1/H(p_a)}{1/H(p_a) + 1/H(p_b)}\,,\, \omega_b = \frac{1/H(p_b)}{1/H(p_a) + 1/H(p_b)} \tag{3}$$

In contrast to *minimum entropy* which operates an "hard" decision, *inverse entropy* gives highest weight to low entropy distributions in a "soft" way.

In [4] it was noticed that typically *inverse entropy* combination performs better then *minimum entropy* combination when streams have comparable performances. However if one of the

feature streams is non-informative or completely corrupted by noise, *minimum entropy* yields better results.

Although *inverse entropy* combination has been proven effective in both small and large vocabulary tasks [6], it is based on empirical observations of the behavior of the MLP output in noisy conditions. We can identify the following problems:

1 The use of the inverse value of the entropy is not theoretically motivated or justified. The weighting scheme (3) does not arise as optimization of an objective function.
2 *Inverse entropy* does not properly handle non-informative posterior distributions. To understand the problem, let us consider a non-informative uniform distribution $\{p_{ia}\} = 1/k$ (i.e. $H(p_a) = H_{max}$ where H_{max} is the maximum entropy value) and an informative distribution p_b such that $H(p_b) \neq H_{max}$ and $H(p_b) \neq 0$. Given that p_a does not contain information on Θ, we would expect $\omega_a = 0$. Inverse entropy weighting will provide $\omega_a \neq 0$. In [4] the problem is tackled comparing $H(p)$ with a threshold (static or dynamic); if $H(p)$ exceeds the threshold, the weight ω_a is set to an arbitrary small value.

Inverse entropy can be considered as a trade-off between the linear averaging of p_a, p_b and the *minimum entropy* solution. In the following we propose an information theoretic interpretation of the linear averaging and the *minimum entropy*. In section III we show that linear average can be obtained from the minimization of a weighted sum of KL divergences. In section IV, we show that *minimum entropy* can be obtained as minimization of a bound on the Bayes probability of error. The proposed criterion is a trade-off between the two quantities and it is discussed in section V.

III. Linear average as minimization of distance function

Let us consider $\{p_{i\,a} = p(\theta_i|X_a)\}$ and $\{p_{i\,b} = p(\theta_i|X_b)\}$ and let us denote with $\pi_a = p(X_a)$ and $\pi_b = p(X_b)$ the prior probabilities of feature streams X_a and X_b (with $\pi_a + \pi_b = 1$) . Assuming the linear combination (1), we can write the following function:

$$\begin{aligned} D(p_c) &= \pi_a\, KL(p_a||p_c) + \pi_b\, KL(p_b||p_c) = \\ &= -\pi_a H(p_a) - \pi_b H(p_b) + \\ &- \sum_i (\pi_a p_{ia} + \pi_b p_{ib}) log(\omega_a\, p_{ia} + \omega_b\, p_{ib}) \end{aligned} \quad (4)$$

where $KL(.||.)$ denotes the Kullback-Leibler divergence between two distributions. $D(p_c)$ is the weighted sum of KL divergences between the individual posteriors p_a, p_b and their linear combination p_c. Minimizing $D(p_c)$ is equivalent to minimizing the sum of cross entropies $-\sum p_{ia} \log p_{ic}$ and $-\sum p_{ib} \log p_{ic}$ weighted by priors π_a and π_b. It follows directly from the Gibbs inequality ($-\sum p_i\, logp_i \leq -\sum p_i\, logq_i$) that $(\omega_a^*, \omega_b^*) = argmin\, D(\omega_a, \omega_b) = (\pi_a, \pi_b)$.

If the streams have equal prior probability i.e. $\pi_a = \pi_b = 0.5$, the distribution p_c that minimize $D(p_c)$ is the average of p_a, p_b, i.e., $p_c = \frac{1}{2}(p_a + p_b)$.

In summary, the average of two posterior estimates can be the obtained as the minimum of the function (4) under equal prior $\pi_a = \pi_b = 0.5$. In the following, we will make the assumption of equal prior probability for feature streams X_a and X_b.

IV. Minimum Entropy Solution as minimization of Bayes Error Bound

Let us assume a classification problem between a set of k classes denoted by $\Theta = \{\theta_i\}$ with $i = 1, ...k$. Given posterior probabilities $\{p_i = p(\theta_i|X)\}$ where X is an observation vector, in [7], it has been shown that a bound on the Bayes probability of error $P_e(\Theta)$ is given by:

$$P_e(\Theta) \leq \frac{1}{2}\, H(\Theta|X) \quad (5)$$

In other words, the minimization of the entropy $H(\Theta|X)$ corresponds to the minimization of an upper bound on the Bayes probability of error. Given p_a and p_b, the linear combination p_c that minimizes the bound (5) is obtained as:

$$(\omega_a^*, \omega_b^*) = argmin\, H(p_c) = argmin\, H(\omega_a\, p_{i\,a} + \omega_b\, p_{i\,b}) \quad (6)$$

with $\omega_a + \omega_b = 1$. Because of the concavity of the entropy function, we have:

$$H(\omega_a\, p_a + \omega_b\, p_b) \;\geq\; \omega_a\, H(p_a) + \omega_b\, H(p_b) \quad (7)$$

Thus the minimum of $H(p_c)$ is achieved for $\omega_a^* = 1, \omega_b^* = 0$ if $H(p_a) < H(p_b)$ and for $\omega_a^* = 0, \omega_b^* = 1$ if $H(p_a) > H(p_b)$. If $H(p_a) = H(p_b)$, $H(p_c)$ has two minima, thus the method randomly selects one of them. This is equivalent to the *minimum entropy* solution. Expression (5) is an upper bound, minimizing the entropy $H(\Theta|X)$ does not guarantee the minimization of the error.

V. Information theoretic trade-off

Inverse entropy combination can be considered as a trade-off in between the average of the two distribution p_a and p_b and the *minimum entropy* solutions. $D(p_c)$ and $H(p_c)$ are minimized by the sum and the minimum entropy rules respectively and they have different (complementary) solutions. Thus we propose the use of the following objective function in order to obtain the desired trade-off between the two solutions:

$$\begin{gathered} J(p_c) = \alpha\, [\tfrac{1}{2}\, H(p_c)] + D(p_c) = \\ = \alpha\, [\tfrac{1}{2}\, H(p_c)] + [\pi_a\, KL(p_a||p_c) + \pi_b\, KL(p_b||p_c)] = \\ = -\pi_a H(p_a) - \pi_b H(p_b) + \\ -\textstyle\sum_i [(\pi_a + \alpha\frac{\omega_a}{2}) p_{ia} + (\pi_a + \alpha\frac{\omega_b}{2}) p_{ib}] log(\omega_a\, p_{i\,a} + \omega_b\, p_{i\,b}) \end{gathered} \quad (8)$$

The minimization of $J(p_c)$ can be considered as the minimization of $D(p_c)$ under the constraint of minimum entropy of $H(p_c)$. Dually it can be interpreted as the minimization of the entropy $H(p_c)$ (thus the bound on the Bayes probability of error) under the constraints of minimum divergence between p_c and the distributions p_a and p_b. The parameter α is the trade-off factor between the two quantities.

Let us consider $(\omega_a^*, \omega_b^*) = argmin\, J(\omega_a, \omega_b)$.

- For $\alpha \to 0$, $J(p_c) = D(p_c)$ thus the minimum of $J(p_c)$ is achieved for $(\omega_a^*, \omega_b^*) = (\pi_a, \pi_b)$. If $\pi_a = \pi_b = 0.5$ this corresponds to the linear average of p_a and p_b.

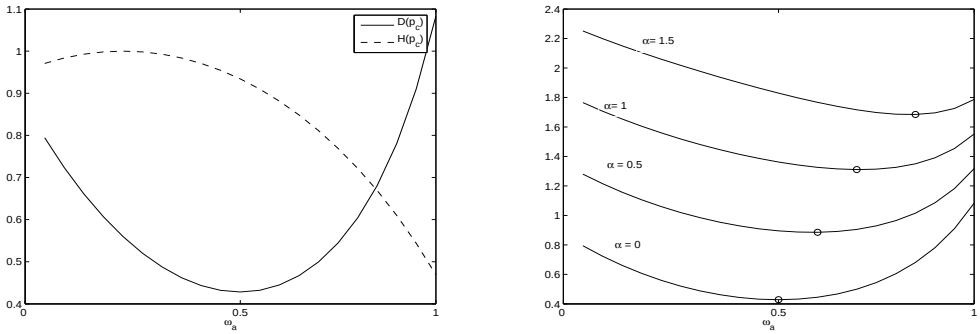

Fig. 1. (Left plot) Example of function $D(p_c)$ with $\pi_a = \pi_b = 0.5$ and $H(p_c)$ for $p_a = [0.9\ \ 0.1]$, $p_b = [0.4\ \ 0.6]$ as function of ω_a. $D(p_c)$ is a convex function with a minimum for $\omega_a = 0.5$ while $H(p_c)$ is a concave function with a minimum in $\omega_a = 1$. (Right plot) $J(p_c) = \alpha\, H(p_c)/2 + D(p_c)$ as function of ω_a for different values of α; for increasing α the minimum moves from $\omega_a = \pi_a = 0.5$ towards the minimum entropy solution. $J(p_c)$ is a trade-off between its parts $D(p_c)$ and $H(p_c)$.

- For $\alpha \rightarrow \infty$, $J(p_c) \rightarrow H(p_c)$ thus the minimum of $J(p_c)$ is achieved for $\omega_a^* = 1$ if $H(p_a) < H(p_b)$ and $\omega_a^* = 0$ if $H(p_a) < H(p_b)$ i.e. the *minimum entropy* solution.

For other values of $\alpha > 0$, (ω_a^*, ω_b^*) will be included between the *minimum entropy* solution and the average combination i.e.

$$\begin{aligned} \omega_a^* \,\epsilon\, [\pi_a, 1] \ \ \text{if} \ \ H(p_a) < H(p_b) \\ \omega_a^* \,\epsilon\, [0, \pi_a] \ \ \text{if} \ \ H(p_a) > H(p_b) \end{aligned} \tag{9}$$

with $\omega_b^* = 1 - \omega_a^*$ and $\pi_a = 1 - \pi_b$. If $H(p_a) = H(p_b)$, the number of minima in $J(p_c)$ depends on the value of α. If $\alpha\, H(p_c)$ is larger then $D(p_c)$, $J(p_c)$ has two minima, in the other case just one minimum. Figure 1 shows an example of functions $H(p_c)$ and $D(p_c)$ w.r.t. the weight ω_a and $J(p_c)$ for different values of α. The solution arises from the optimization of the informtion theorethic trade-off.

(ω_a^*, ω_b^*) do not have an analytic form. We used a standard gradient descent technique to find the root of the equation $\partial\, J(\omega_a)/\partial\omega_a = 0$ in the range of values defined by the expressions (9). If no root is available, the minimum is at one extreme of the range and is determined by the sign of the derivatives.

A. The trade-off factor

The trade-off factor α can be statically set (i.e., independent of the current values of p_a and p_b) and determined by cross validation experiments. We propose to set it dynamically as a function of p_a and p_b. According to the discussion of section II point 2, we would like to obtain a weight equal to zero in the case of non-informative uniform posterior distributions. Let us define:

$$\alpha(p_a, p_b) = \frac{1}{KL(p_a||p_u) \times KL(p_b||p_u)} \tag{10}$$

where $\{p_{ui} = 1/k\}\ \forall i = 1, .., k$ is a uniform distribution. α is set inversely proportional to the divergence between p_a,p_b and p_u. If p_a is, for instance, non-informative (i.e., a uniform distribution) and $H(p_b) \neq H_{max}$ then $KL(p_a||p_u) = 0$ and $\alpha = \infty$. Thus, minimizing $J(p_c)$ is equivalent to minimizing $H(p_c)$, which gives $(\omega_a = 0, \omega_b = 1)$. The non-informative distribution has a weight equal to zero.

In general, if p_a and p_b are low entropy distributions (i.e., far from the uniform distribution, which means that the classifiers are confident about the decision), the value of α will be small. Thus the optimization of $J(p_c)$ will mainly focus on the term $D(p_c)$. On the other hand, when p_a or p_b are high entropy distributions (i.e. close to the uniform distribution which means that the classifiers are not confident on the decision), the value of α will be large. Thus the optimization of $J(p_c)$ will mainly focus on the term $H(p_c)$, which only selects the most confident stream.

VI. Experiments

In the following, we investigate the use of the $J(p_c)$ function for combining phoneme posterior probabilities obtained using different input streams. Experiments aim at comparing the proposed approach with other linear frame-based combination rules like the *inverse-entropy*, *minimum-entropy* and uniform weighting. The combination happens at the frame level. The experimental setting is the following: two MLPs are trained using different temporal context: a short temporal context (9 frames PLP [8]) and a long temporal context (one second critical band energy pre-processed with a set of zero mean filters a.k.a. as MRASTA [9]). Those two different posterior estimates are then combined together using sum, inverse entropy, minimum entropy or the J function. Combined posteriors are transformed according to TANDEM processing [8] (i.e. using a log/KLT transform) and used as features in a conventional HMM/GMM system.

A. Small Vocabulary

The database used for recognition experiments consists of the *OGI-Numbers 95* while MLPs are trained using 3 hours of hand-labeled speech from the *OGI-Stories* database in order to discriminate between phonemes. We add noises from the NOISEX database (babble, factory, F16) at different SNR to the test set. Training of MLPs and HMM/GMM is done on clean data. Results are reported in table V. For SNRs equal to 20 and 15 dB *inverse entropy* and J function have comparable results. For SNRs equal to 10, 5 and 0 dB, the J function outperforms the *inverse entropy* combination, the

Features	20dB	15dB	10 dB	5dB	0dB
9frames-PLP	8.7	15.7	30.6	52.1	74.0
MRASTA	5.9	10.3	22.5	51.4	78.7
Sum	5.6	9.8	21.8	48.8	77.1
Min-entropy	5.6	9.5	21.5	45.8	73.1
Inv-entropy	**5.1**	**9.0**	20.5	48.1	77.0
J	**5.1**	**9.0**	**19.7**	**43.3**	**72.6**
$< w >$ (Inv-entropy)	0.55	0.57	0.58	0.58	0.40
$< w >$ (J)	0.56	0.67	0.67	0.68	0.30
$< \alpha >$	0.25	0.30	0.37	0.44	0.47

Features	TOT	AMI	CMU	ICSI	NIST	VT
9frames-PLP	46.6	41.4	43.7	31.3	54.5	64.9
MRASTA	45.9	48.0	41.9	37.1	54.4	48.8
Sum	41.5	41.1	37.6	30.4	50.2	49.8
Min-entropy	41.3	40.4	37.9	29.6	49.1	52.3
Inv-entropy	40.4	39.8	37.0	29.6	48.3	**48.7**
J	**39.8**	**39.5**	**36.7**	**28.8**	**47.5**	**48.7**
$< w >$ (Inv-entropy)	-	0.29	0.63	0.16	0.48	0.63
$< w >$ (J)	-	0.23	0.74	0.10	0.45	0.68
$< \alpha >$	-	0.23	0.24	0.20	0.26	0.26

TABLE I
WER FOR NOISY NUMBERS AT DIFFERENT SNR (RIGHT TABLE) AND FOR RT05 EVALUATION DATA (LEFT TABLE). WER REPORTED FOR INDIVIDUAL STREAMS AND COMBINATION (SUM,MINIMUM ENTROPY, INVERSE ENTROPY AND J CRITERION). THE AVERAGE VALUES OF THE MRASTA STREAM WEIGHT $< w >$ AND THE AVERAGE VALUE OF THE TRADE-OFF $< \alpha >$ ARE REPORTED AS WELL.

improvements being larger at lower dB. It is interesting to notice that at 0 dB, *minimum entropy* outperforms *inverse entropy*. However the J function still produces lower WER than *minimum entropy*. Although the weights and the trade-off α are computed at the frame level, we report in table I the average value of α and the average weight of the MRASTA stream both for inverse entropy and J function. The value of α increases (as expected) with the SNR level. Furthermore the J function weights more the stream with lower WER respect to inverse entropy, the difference being larger at low SNRs.

B. Large Vocabulary

Experiments were run on a meetings transcription task. The training data for this system comprises individual headset microphone (IHM) data of four meeting corpora; the NIST (13 hours), ISL (10 hours), ICSI (73 hours) and a preliminary part of the AMI corpus (16 hours). Those data are used for training MLPs and HMM/GMM models. Acoustic models are phonetically state tied triphones models trained using standard HTK maximum likelihood training procedures. The recognition experiments were conducted on the NIST Rich Transcription 05 (RT05) evaluation data. We use the reference speech segments provided by NIST for decoding. The pronunciation dictionary is the same as the one used in the AMI NIST RT05 system [10]. The challenge of this data set is the variety of acoustic environments in which data have been collected. Results are reported in table I. *Inverse entropy* combination achieves a WER of 40.4% while the J function achieves a WER of 39.8%. The improvements are verified on 4 of the 5 meeting rooms in the RT05 evaluation data set. Table I also reports the average value of α and the average weights of the MRASTA stream both for inverse entropy and J function. Conclusions are similar to those obtained in the previous section.

VII. CONCLUSIONS AND DISCUSSION

In this work we proposed an objective function for the linear combination of classifiers in multi-stream ASR. In contrast to other methods like *inverse entropy*, weights are obtained as minimization of an objective function $J(p_c)$ (9). $J(p_c)$ can be considered as a trade-off between the linear average of posterior distributions and the distribution that minimize the bound on the Bayes probability of error. Furthermore we discuss how to set the trade-off in order to deal with non-informative distributions. In contrary to inverse entropy combination, non-informative distributions receive zero weight without the use of any heuristic threshold. Experiments on small and large vocabulary tasks reveal that the $J(.)$ function outperforms inverse entropy, minimum entropy and uniform weighting. The analysis of the weights average values shows that in case of mismatch the $J(.)$ function provides an higher weight for the most confident stream respect to inverse entropy.

Preliminary experiments on larger amount of data (approximatively 1500 hours of speech) show that the improvements scale-up as long as the MLP features and the HMM/GMM are trained on the same amounts of data.

We limited the discussion to only two streams. The $J(.)$ function can be easily extended to N streams. Assuming the linear combination $p_c = \sum_j^N \omega_j p_j$, it is straightforward to obtain $H(p_c)$ and $D_{p_c} = \sum_j^N \pi_j KL(p_j||p_c)$, thus $J(p_c)$. Furthermore the same principle can also be applied to combinations that are not linear (e.g. log-linear combinations or product rules) given that the criterion is completely general[1].

REFERENCES

[1] Bourlard H. and Dupont S., "A new asr approach based on independent processing and re-combination of partial frequency bands.," *Proc. ICSLP 96*.

[2] Hermansky H., Tibrewala S., and Pavel M., "Towards asr on partially corrupted speech," *Proc. ICSLP*, 1996.

[3] Kittler J. et al, "On combining classifiers," *IEEE Transactions on PAMI*, vol. 20, 1998.

[4] Misra H., Bourlard H., and Tyagi V., "Entropy-based multi-stream combination," in *Proceedings of ICASSP*, 2003.

[5] Bourlard H. and Morgan N, *Connectionist Speech Recognition - A Hybrid Approach*, Kluwer Academic Publishers, 1994.

[6] Morgan N., Chen B., Zhu Q., and Stolcke A., "Trapping conversational speech: Extending trap/tandem approaches to conversational telephone speech recognition," *Proceedings of ICASSP 2004*.

[7] Hellman M.E. and Raviv J., "Probability of error, equivocation, and the chernoff bound," *IEEE Trans. on Information Theory*, vol. 16(4), 1970.

[8] Hermansky H., Ellis D., and Sharma S., "Connectionist feature extraction for conventional hmm systems.," *Proceedings of ICASSP*, 2000.

[9] Hermansky H. and Fousek P., "Multi-resolution rasta filtering for tandem-based asr.," in *Proceedings of Interspeech 2005*, 2005.

[10] Hain T. et al, "The 2005 AMI system for the transcription of speech in meetings," *NIST RT05 Workshop Edinburgh, UK.*, 2005.

[1]This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR0011-06-C-0023 and by the European Union under the integrated project AMIDA. The author thanks the anonymous reviewer for their comments.