IDIAP RESEARCH REPORT


# Multi-stream ASR: Oracle Test and Embedded Training

Hemant Misra [a] [b] Jithendra Vepa [a]
Hervé Bourlard [a] [b]

IDIAP–RR 05-62

Dalle Molle Institute for Perceptual Artificial Intelligence • P.O.Box 592 • Martigny • Valais • Switzerland

phone +41 – 27 – 721 77 11
fax +41 – 27 – 721 77 12
e-mail secretariat@idiap.ch
internet http://www.idiap.ch

[a] IDIAP Research Institute, Martigny, Switzerland
[b] EPFL - Swiss Federal Institute of Technology, Lausanne, Switzerland

IDIAP Research Report 05-62

# Multi-stream ASR: Oracle Test and Embedded Training

Hemant Misra Jithendra Vepa Hervé Bourlard

**Abstract.** Multi-stream based automatic speech recognition (ASR) systems outperform their single stream counterparts, especially in the case of noisy speech. However, the main issues in multi-stream systems are to know a) Which streams to be combined, and b) How to combine them. In order to address these issues, we have investigated an 'Oracle' test, which can tell us whether two streams are complimentary. Moreover, the Oracle test justifies our previously proposed inverse entropy method for weighting various streams. We have carried out experiments on two multi-stream systems and results indicate that in clean speech around 80% of the time Oracle selected the stream which had the minimum entropy. In this paper, we have also presented an embedded iterative training for multi-stream systems. The results of the recognition experiments on Numbers95 database showed that we can improve the performance significantly by multi-stream iterative training, not only for clean speech but also for various noise conditions.

# 1 Introduction

Multi-stream systems in ASR [1, 2] are known to yield better performance as compared to single stream systems. Specially in presence of noise, if the streams carry complementary information, their combination can lead to an improved performance.

In a multi-stream system, if at every time instant an oracle can tell us which stream output is the "best" among all the streams considered for combination, that is the best performance we can achieve by frame level weighting techniques. The aim of the Oracle test presented in this paper is to find the answers to the following questions:

1. What is the best performance that can be achieved by frame level weighting for a given set of feature streams?

2. Whether the streams considered for combination have complementary information?

3. How well the inverse entropy weighting[1] studied in [2] corresponds with Oracle choice?

It will also help us in understanding the positive attributes and potential of a multi-stream system which are not fully realized by employing different statistical weighting strategies [3, 4, 2].

The work reported here has taken place in the framework of hybrid HMM/ANN ASR system. The embedded iterative training exists for HMM/ANN systems [5, 6], and it is known to yield improved performance. However, in practice, it is avoided because of high computational cost involved in training the ANNs. With the present day faster machines, processing is not a constraint and it is easier to train the ANNs. In the second part of the paper, we suggest an embedded training procedure for multi-stream systems. We show that the advantages of multi-stream also apply to embedded multi-stream system. Finally, we show that we can reduce the gap between Oracle weighting and inverse entropy weighting approach by multi-stream embedded training.

The rest of the paper is organized as follows: in Section 2, we present the proposed Oracle test and explain its advantages. The performance of the test on Numbers95 [7] database is presented in Section 3. In the same section, we discuss various characteristics of Oracle test. In Section 4, the suggested multi-stream method is presented along with the results. Results are presented on clean as well noise conditions. Noise conditions were simulated by adding factory noise at different levels from Noisex92 [8] database.

# 2 Oracle Test

## 2.1 Oracle performance in multi-stream

In the simple Oracle experiment, at every time instant (frame), we chose the output of the MLP expert[2] that has the highest posterior for the right class [9, 10]. In essence, the Oracle does the 1/0 weighting, that is, the "best expert" gets the weight of 1 while rest of the experts get a weight of 0. The right class was obtained by Viterbi forced alignment of the test data by the baseline perceptual linear prediction (PLP) derived cepstral coefficients (13 static features appended by their first and second order time derivatives). This Oracle test can let us know the best performance that can be achieved by frame level weighting for a given setup in multi-stream combination. In the absence of Oracle, it might not be possible to achieve the same performance. Nevertheless, in the later part of this paper, we analyze the Oracle performance and show that we indeed moved in the right direction by inverse entropy weighting method investigated in [2].

[1] In inverse entropy weighting, the posterior outputs of multi-layered perceptron (MLP) experts for various streams are weighted inversely proportional to their respective output entropies and combined.

[2] In HMM/ANN ASR system, we train MLP as a classifier

### 2.2 Complementarity of feature streams

Apart from the best frame level performance, the Oracle test can also give us an indication about the complementarity of the streams. If two streams carry exactly the same information, using those two streams we cannot improve the accuracy of the combined system. If two streams carry complementary information, combining them we can achieve an improvement in performance. In essence, more the complementary information between two streams, better the gains we can attain by combining those two streams.

This property of Oracle test can help in finding whether the feature streams considered for combining carry any complementary information. This could be a fast method to check whether the streams considered for combination will yield any improvement when combined by sub-optimal methods [1, 9, 2]. In practice, the improvement achieved by Oracle might not be reached by statistical combination methods which rely on the average behavior of the streams. Nevertheless, Oracle can stop us from looking at streams which do not give improvement even in the ideal case, and to begin with we can consider only those feature streams which give better improvements when combined by Oracle.

## 3 Oracle Performance

In this section, we present the performance obtained by Oracle. This performance is not absolute because the "goodness" of Viterbi forced-aligned data itself depends on the posteriors used for finding the alignment. We have used the output of the baseline PLP system to obtain the forced alignment.

We demonstrate the performance for two systems. The first system uses 7 PLP streams: static PLP, delta PLP, delta-delta PLP and their all possible combinations in full-combination multi-stream (FCMS) [1] setup. The second system investigates the 3 streams: baseline PLP, multi-band spectral entropy features [11] derived from sub-bands defined by Mel-scale and PLP features concatenated with spectral entropy features. The spectral entropy feature is obtained by dividing the normalized full-band spectrum into sub-bands and estimating the entropy in each sub-band [11].

### 3.1 Number of streams

In the first setup, we increased the number of streams considered for combination from 1 to 7 for the PLP only system. We observe that as the number of streams increase, the performance of the Oracle system improves. Fig. 1 (a) shows the average word-error-rates (WER) for number of streams chosen out of 7 possible PLP streams[3]. The circles ($o$) in the figure show the standard-deviation around the average WERs. When all the 7 streams are considered, we get a WER of 5.6%. Similarly, Fig. 1 (b) shows the plot for PLP and 24-Mel band derived spectral entropy features used in full-combination multi-stream setup (3 possible streams). When all the 3 streams are considered, we achieve a WER of 6.5%.

An important observation from the figures is that, as the number of streams increases, and assuming that the streams carry complementary information, the performance of Oracle improves. Further, the curve starts getting flat when more streams are added indicating the additional streams are not bringing in much new (and complementary) information into the system.

### 3.2 Complementarity of streams

The property that Oracle can shed information about the complementarity of the feature streams is depicted in Fig. 2. In the figure, we start with the baseline PLP system and start adding other streams to it. When we combine another PLP stream (choosing one from the six remaining streams) to the baseline PLP stream, we see an improved performance. When we combine the spectral entropy feature

[3]We have $C_i^I = \frac{I!}{i!(I-i)!}$ combinations to choose $i$ streams out of $I$ streams.

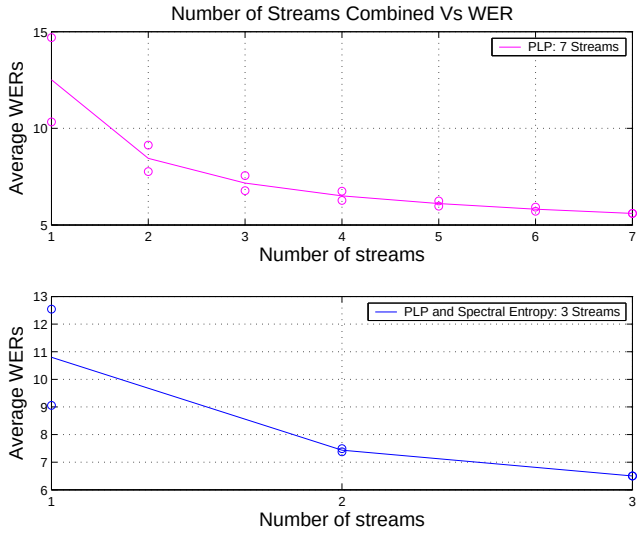


Figure 1: *Performance of the Oracle for multi-stream combination. The streams considered for combination are: (a) All possible combinations of the static PLP features and their first and second order time derivatives (7 streams). (b) PLP features with first and second order time derivatives, spectral entropy features derived from 24-Mel band with their first and second order time derivatives and concatenation of the two features (3 streams).*

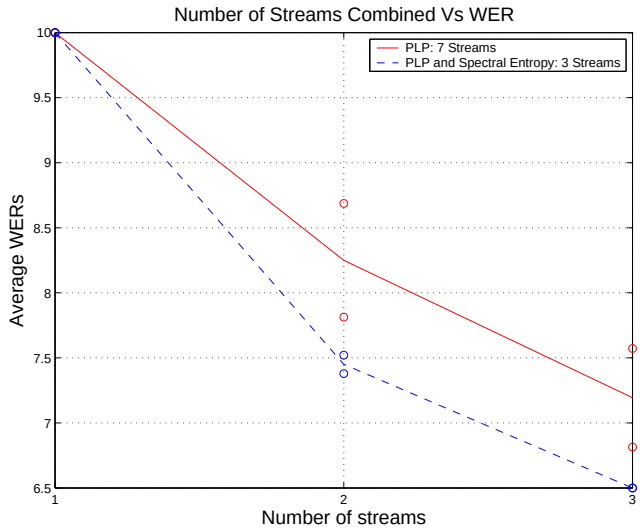


Figure 2: *Oracle performance to find out complementarity of streams used in the multi-stream combination. The performance is compared for PLP features (static, delta and delta-delta) in FCMS (7 streams: —-) and PLP features along with spectral entropy features in FCMS (3 streams: - - -)*

streams, the improvement is more as compared to what we observed by adding the PLP streams. This confirms our intuition and indicates that spectral entropy features bring more complementary information into the system. The circles ($o$) in the figure show the standard-deviation around the average WERs.

It is noticeable that similar but less pronounced trend was observed when we considered the 7 PLP streams for combination and compared the results with 3 PLP and spectral entropy streams for combination in the weighting techniques investigated in [2, 11]. This result indicates that spectral

entropy features had information complementary to the PLP features, and it was worth investigating them for multi-stream combination.

### 3.3 Relationship with minimum entropy

In this section, we analyze how the Oracle chooses a particular stream among all the streams. We restrict our studies to analyze the relationship between Oracle selection and the entropy at the output of MLPs trained on their respective feature streams.

In the simple setup, we computed the entropy of the stream selected by Oracle at each time step, and compared it with the entropy of all the other streams. Interestingly, in case of 7 streams PLP features used for combination, in clean speech, 75.7% of the times Oracle selection was the same as the selection we would have made if we would have considered the stream with the minimum entropy. That is, 75.7% of the times, minimum entropy stream was selected by Oracle. In case of multi-stream combination of PLP features along with spectral entropy features in full-combination, Oracle selected the minimum entropy stream 79.2% of the times.

Fig. 3 shows how many times (frames) Oracle selected the minimum entropy stream for different

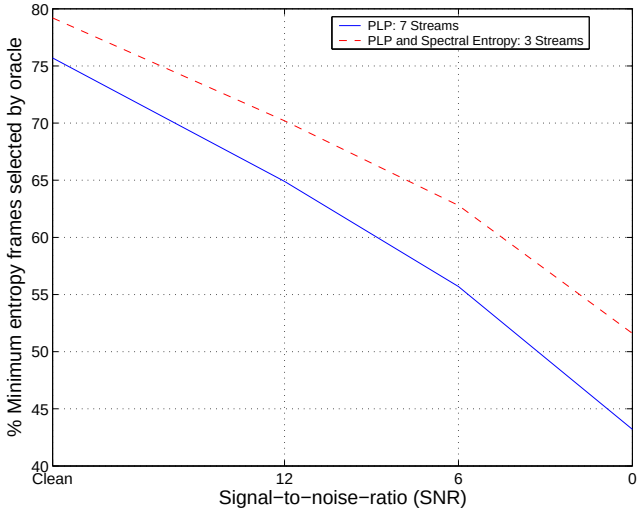


Figure 3: *Number of times (in percentage of frames) the Oracle selected the stream with minimum entropy in FCMS hybrid system.. The plot is for clean as well as noisy (additive factory noise) test conditions.*

noise levels (additive factory noise at several SNRs). We notice that as the noise level increases, the preference for the minimum entropy frames diminishes, but still the minimum entropy frames enjoy a majority in Oracle selection. This suggest that entropy at the output of a classifier is a reasonable choice for weighting, as done in our previous work [2]. So we can conclude that entropy at the output of a classifier is a good measure for selection, and correlates well with the Oracle selection.

## 4 Multi-stream Embedded Iterative Training

There are many methods to do embedded iterative training in multi-stream combination. For example, we can do separate embedded training for each stream [5] and combine the outputs of all the streams at the time of testing. Another approach can be to train the models for each stream but the labels of all the streams are same and are obtained by Viterbi forced alignment of the combined posteriors. We investigated both the methods and observed that they yield similar results. The steps of the iterative training for the later approach are:

1. As in single stream embedded iterative training, we started with hand-labelled frame segmentation and trained one MLP for each feature stream.

2. The training data of each feature stream was passed as test data through the corresponding MLP and the posteriors were obtained at the output of the respective MLPs.

3. The posteriors from different streams were combined using inverse entropy weighting [3, 2].

4. New frame-level segmentation was obtained by Viterbi forced alignment of the combined posteriors obtained on training data.

5. New MLPs with the same initialization were trained from scratch for every feature stream using the new segmentation.

6. Steps 2 and 3 were repeated several times (four iterations in the present setup). In the end, we had one trained MLP for each feature stream and each iteration.

7. To test the MLPs at each iteration, test data for each feature stream was passed through the respective MLP to obtain the posteriors. The posteriors from different MLPs were combined by inverse entropy weighting and used for recognition.

The feature streams we considered for multi-stream embedded iterative training were PLP features, spectral entropy features from 24-Mel bands and the concatenation of these two features.

In the bar plot (Fig 4), the performance of the PLP baseline, PLP trained with embedded training

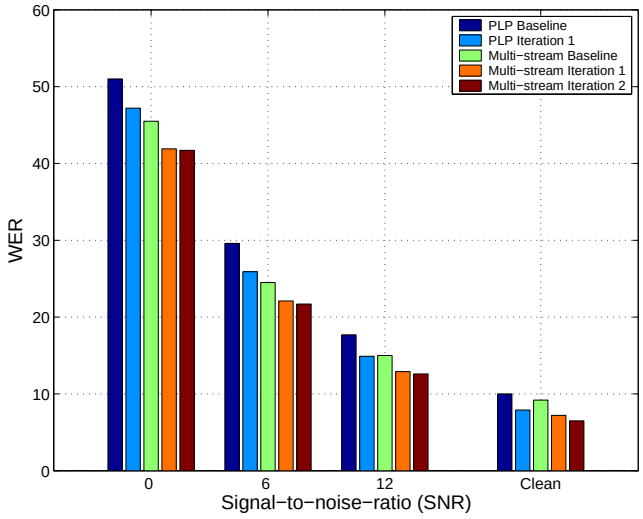


Figure 4: *Performance in terms of WER for for PLP features with hand segmentation, PLP features with segmentation obtained by forced alignment during embedded iterative training, PLP and spectral entropy features in FCMS with inverse entropy weighting and hand segmented labels, PLP and spectral entropy features in FCMS with segmentation obtained by forced alignment during embedded iteration training. Different SNR conditions were tested.*

(first iteration), multi-stream baseline and multi-stream system with embedded training (first two iterations) are shown for comparison. Embedded iterative training helps in improving the baseline PLP performance as well as the performance of the multi-stream system. The improvement is consistent and generalizes for different noise levels studied.

The results of iterative training give an impression that we have achieved the performance of the Oracle (6.5% WER) by iterative training, but it is not entirely true. This fallacy can be explained

by the following reasoning: In presence of better segmentation and better modelling, the Oracle itself improves its performance from 6.5% to 4.5% as shown in Table 1.

| Segmentation | Baseline PLP | Multi-stream | Oracle |
|---|---|---|---|
| **Hand** | 10.0 | 9.1 | 6.5 |
| **Forced-alignment** | 7.9 | 6.5 | 4.5 |

Table 1: *WER in % for training with hand-segments and segments obtained by iterative embedded training (best result for second iteration is shown). a) PLP baseline features, b) multi-stream combination of PLP features with spectral entropy features in FCMS, and c) Oracle. Testing on clean conditions only.*

# 5 Summary

In this paper, we presented frame level Oracle test for multi-stream systems and analyzed its characteristics. We showed that the Oracle test can be used to investigate the complementary properties of new feature streams. Also, we found that Oracle tends to choose the output of the MLP experts (trained on feature streams) that had the minimum entropy at their output. This further strengthens our proposed method of inverse entropy weighting for combining the outputs of the classifiers.

In the second part of the paper, we proposed an embedded iterative training procedure for hybrid multi-stream systems. We observed that multi-stream iterative training can lead to improved performance, not only in clean test conditions but also for noisy test conditions. We achieved an improvement of 2% absolute over the baseline PLP system by employing iterative training to PLP features. We further gained a WER drop of close to 1.5% absolute on clean conditions by multi-stream iterative training over the single-stream iterative training applied to PLP baseline features.

**ACKNOWLEDGEMENTS**

This work was supported by the Swiss National Science Foundation through the National Centre of Competence in Research (NCCR) on "Interactive Multimodal Information Management (IM2)" and the EU 6th FWP IST Integrated Project AMI (Augmented Multi-party Interaction, FP6-506811).

# References

[1] A. C. Morris, A. Hagen, H. Glotin, and H. Bourlard, "Multi-stream adaptive evidence combination for noise robust ASR," *Speech Communication*, vol. 34, no. 1–2, pp. 25–40, 2001.

[2] H. Misra, H. Bourlard, and V. Tyagi, "New entropy based combination rules in HMM/ANN multi-stream ASR," in *Proceedings of IEEE International Conference on Acoustic, Speech, and Signal Processing*, (Hong Kong), Apr. 2003.

[3] S. Okawa, E. Bocchieri, and A. Potamianos, "Multi-band speech recognition in noisy environments," in *Proceedings of IEEE International Conference on Acoustic, Speech, and Signal Processing*, (Seattle, Washington), pp. 641–644, May 1998.

[4] M. Heckmann, F. Berthommier, and K. Kroschel, "Noise adaptive stream weighting in audio-visual speech recognition," *To be published in Journal on Applied Signal Processing (special issue on Audio-Visual Processing, 2002)*, vol. 2, no. 11, 2002.

[5] S. Renals, N. Morgan, H. Bourlard, M. Cohen, and H. Franco, "Connectionists probability estimators in HMM speech recogntion," *IEEE Trans. Speech, Audio Processing*, vol. 2, no. 1, pp. 161–174, 1994.

[6] N. Mirghafori and N. Morgan, "Transmissions and transitions: A study of two common assumptions in multi-band ASR," in *Proceedings of IEEE International Conference on Acoustic, Speech, and Signal Processing*, (Munich, Germany), 1997.

[7] R. Cole, M. Noel, T. Lander, and T. Durham, "New telephone speech corpora at CSLU," in *Proceedings of European Conference on Speech Communication and Technology*, vol. 1, pp. 821–824, 1995.

[8] A. Varga, H. Steeneken, M. Tomlinson, and D. Jones, "The NOISEX-92 study on the affect of additive noise on automatic speech recognition," technical report, DRA Speech Research Unit, Malvern, England, 1992.

[9] M. L. Shire and B. Y. Chen, "Data-driven RASTA filters in reverberation," in *Proceedings of IEEE International Conference on Acoustic, Speech, and Signal Processing*, (Istanbul, Turkey), 2000.

[10] L. I. Kuncheva, "A theoretical study on six classifier fusion strategies," *IEEE Trans. Pattern Analysis and Machine Intelligence*, pp. 281–286, Feb. 2002.

[11] H. Misra and H. Bourlard, "Spectral entropy feature in full-combination multi-stream system for robust ASR," in *Proceedings of European Conference on Speech Communication and Technology*, (Lisbon, Portugal), Sept. 2005.