Multi-stream ASR: An Oracle Perspective

Hemant Misra, Jithendra Vepa, Hervé Bourlard

IDIAP Research Institute, Martigny
and
Ecole Polytechnique Federale de Lausanne (EPFL)
Switzerland
{misra, vepa, bourlard}@idiap.ch

Abstract

Multi-stream based automatic speech recognition (ASR) systems are usually shown to outperform single stream systems, specially in noisy test conditions. And, indeed, there is a trend today in ASR towards using more and more acoustic features combined at the input (early integration, possibly preceded by some linear or non-linear transformation) or later in the recognition process (e.g., at the level of likelihoods, then referred to as late integration). However, to guarantee optimal exploitation of such multi-stream systems, we need to use features that are as much complementary as possible, while also using the best combination method for those streams. In practice, it is never clear whether we fully exploit the potential of the available streams. This present paper investigates an 'oracle' test to provide some insight in these issues. Although not providing us with an absolute performance upper bound, oracle is shown to indicate the complimentary of the feature streams used, and to provide a reasonable reference target to evaluate combination strategies. The oracle analysis is supported by results obtained on Numbers95 database using different feature streams and entropy based combination method.

Index Terms: Speech recognition, multi-stream, spectral entropy features, oracle analysis

1. Introduction

Multi-stream systems in ASR [1, 2] are known to yield better performance as compared to single stream systems. The important issues in multi-stream systems are to find feature streams that carry complementary information and to combine the outputs of the classifiers trained on feature streams such that outputs of the classifiers get importance according to their respective reliability. However, the potential of a multi-stream system is not exploited fully when we use some statistical measures for combining the outputs of different classifiers.

In a multi-stream system, if at every time instant an oracle can select the stream which is the "best" among all the streams considered for combination, the performance thus obtained is referred to as oracle performance. Such oracle tests have been reported in the literature to find out the oracle performance in pattern recognition tasks [3]. However, oracle studies have been restricted to finding the oracle performance.

In this paper, we propose an alternative interpretation of oracle test to analyze the issue of complementarity of feature streams in a multi-stream system. Also, we investigate how well the oracle selection can be described by entropy at the output of the classifiers, which is a statistical measure. The aim of the oracle test presented in this paper is to find the answers to the following questions:

1. What is the oracle performance that can be achieved by frame level weighting for a given set of feature streams?
2. Whether the streams considered for combination are carrying any complementary information?
3. How well the minimum entropy weighting proposed in [2] corresponds with oracle selection?

When compared with single-stream systems, the oracle analysis also indicates the potential of multi-stream systems which is not realized completely by employing different statistical measures for weighting [4, 2].

The rest of the paper is organized as follows: In Section 2, we present the proposed oracle test and explain its properties. The experimental setup and the database have been explained in Section 3. The performance of the oracle test is presented in Section 4, and in the same section we analyze various characteristics of the oracle test. The results of multi-stream systems using different feature streams and entropy based weighting are given in Section 5. The conclusions of the paper are presented in Section 6.

2. Oracle test

2.1. Oracle performance in multi-stream

In the frame-level oracle setup, at every time instant (frame), we choose the outputs of the multi-layered perceptron (MLP) classifier[1] that has the highest posterior for the correct class [3]. In essence, the oracle does 1/0 weighting, that is, the outputs of the "best classifier" get the weight of 1 while the outputs of rest of the classifiers get the weight of 0. This test can let us know the oracle performance that can be achieved by frame level weighting for a given set of feature streams in a multi-stream system. The oracle also indicates the gain that can be achieved by moving from single-stream systems to multi-stream systems.

2.2. Complementarity of feature streams

Apart from the typical oracle performance often shown in pattern recognition tasks, an alternative interpretation of the oracle test can indicate the complementarity of the feature streams. The proposed interpretation is based on the following argument: If two streams carry exactly the same information, combining those

[1] In hybrid HMM/ANN ASR systems used in this paper, we train MLP as a classifier.

two streams we cannot improve the accuracy of the system. If two streams carry complementary information, combining them we can achieve an improvement in the performance. In essence, more the complementary information between the two streams used for the combination, more we can gain by combining those two streams.

This interpretation of the oracle test can help in finding whether the feature streams considered for combination carry any complementary information. We may drop the feature streams that give less improvement even in the ideal case (oracle selection). This could be a quick method to check whether the streams considered for combination will yield any improvement when combined by sub-optimal methods [1, 2]. In practice, the improvement achieved by oracle may not be reached by statistical combination methods which rely on the average behavior of the streams.

3. Experimental setup

In the experiments reported in this paper, Numbers95 database of US English connected digits telephone speech [5] was used. There are 30 words in the database represented by 27 phonemes. Training is performed on clean speech utterances and testing data (which is different from the training data) is either clean or corrupted by factory noise from the Noisex92 database [6] added at different signal-to-noise ratios (SNRs) to the Numbers95 database. The baseline perceptual linear prediction (PLP) [7] features in this study were 13 dimensional static features appended by their first and second order time derivatives. There were 3330 utterances for training and 2250 utterances were used for testing the system.

The studies were carried out in the framework of hybrid hidden Markov model/artificial neural network (HMM/ANN) system. In the setup, the ANNs used were multi-layer perceptron (MLP) with one hidden layer. The input layer was fed by 9 consecutive data frames. The HMM used for decoding had fixed state transition probabilities of 0.5. Each phoneme had a single state model for which emission likelihoods were supplied as scaled-likelihoods. The minimum duration for each phoneme was modeled by forcing 1 to 3 repetitions of the same state for each phoneme.

3.1. Full-combination multi-stream (FCMS)

We have used FCMS [1] framework illustrated in Fig. 1 to carry out the experiments. In FCMS, more than one type of feature representation is extracted from the speech signal and every possible combination of the feature representations is treated as a separate feature stream. In hybrid HMM/ANN approach, one MLP is trained for each such feature stream and posterior estimates at the output of the MLPs are weighted and combined. The combined posteriors are divided by prior probabilities and the scaled-likelihoods thus obtained are used for decoding.

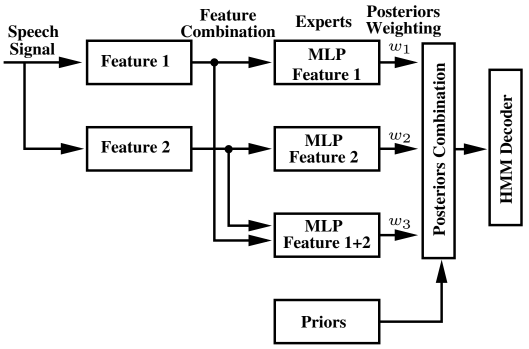

Figure 1: *Full-combination multi-stream: All possible combinations of the two features are treated as separate streams. An MLP expert is trained for each stream. The posteriors at the output of the MLPs are weighted and combined. The combined posteriors thus obtained are passed to an HMM decoder.*

This paper investigates the following two setups:

- PLP and CJ-Rasta-PLP [8] features in the FCMS setup leading to 3 feature streams (PLP, CJ-Rasta-PLP and the concatenation of the two features).
- PLP and 24-Mel band spectral entropy features [9] (briefly described in Section 3.2) in the FCMS setup, again leading to 3 feature streams (PLP, 24-Mel band spectral entropy and the concatenation of the two features).

3.2. Spectral entropy features

Spectral entropy can indicate the flatness/peakiness of a spectrum and was used in [10] for speech/silence detection. The entropy of the spectrum is computed by converting the spectrum into a probability mass function (PMF) by normalizing it.

However, full-band spectral entropy feature can capture only the gross peakiness of the spectrum but not the position of the formants. In [9], we suggested multi-band spectral entropy features to capture the peakiness of the sub-bands. The spectrum was divided into sub-bands and entropy of each sub-band was computed. The sub-band spectral entropies were concatenated and used as a feature vector for ASR task. In [9], we obtained the best results by dividing the normalized full-band spectrum into 24 overlapping sub-bands defined on Mel-scale and computed entropy from each sub-band. Further, we appended the first and second order time derivatives to include temporal information.

4. Oracle performance

In this section, we present the oracle performance. This performance is not the upper bound 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 multi-stream systems.

4.1. Number of streams

The experiments reported in this sub-section are for clean test condition. In the first experiment, we used the following 3 feature streams: PLP, CJ-Rasta-PLP and the combination of the two features by concatenation. One MLP was trained for each feature stream. Out of the 3 MLPs, outputs of n MLPs were considered for combination, n varied from 1 to 3. Fig. 2 shows the average word-error-rates (WER) for n streams chosen out of 3 possible streams[2]. For $n = 1$, we have the possibility of 3 single stream systems in the present setup (PLP or CJ-Rasta-PLP or PLP concatenated with CJ-Rasta-PLP). WER was obtained for each single stream system and the average WER was computed from the 3 experiments. When $n = 2$, we again have 3 possibilities to choose 2 streams out of 3 possible streams (PLP and CJ-Rasta-PLP or PLP

[2] We have $C_n^N = \frac{N!}{n!(N-n)!}$ possibilities to choose n streams for combination out of N streams. We considered all the possible combinations to compute the average WERs.

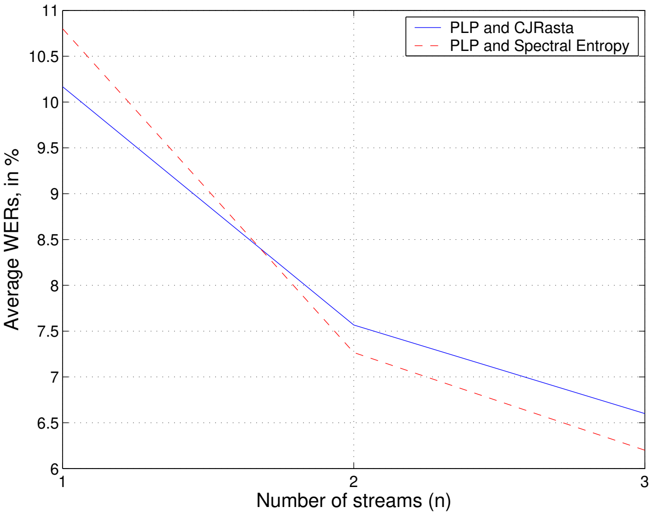

Figure 2: *Oracle performance: Number of streams combined vs average WER. The plots are for two set of feature streams: (a) PLP, CJ-Rasta-PLP and the combination of the two features by concatenation, and (b) PLP, spectral entropy and the combination of the two features by concatenation.*

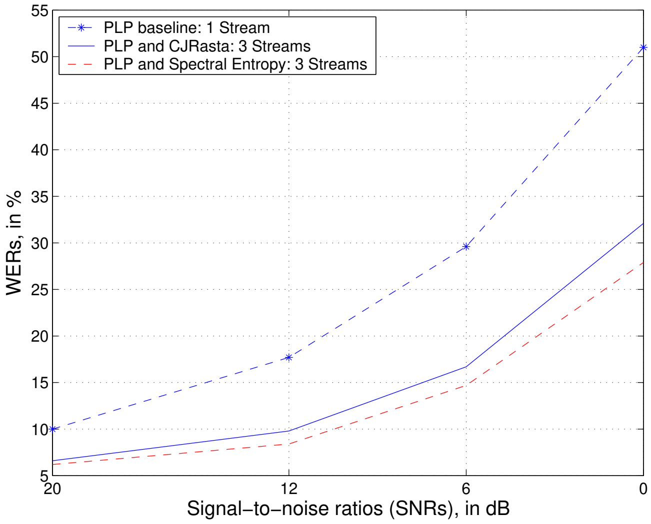

Figure 3: *The oracle performance in the FCMS setup to find out complementarity of feature streams. The performance is compared for: (a) PLP features with CJ-Rasta-PLP features in FCMS, and (b) PLP features with spectral entropy features in FCMS. The PLP performance is for comparison.*

and PLP concatenated with CJ-Rasta-PLP or CJ-Rasta-PLP and PLP concatenated with CJ-Rasta-PLP). WERs of these 3 experiments were used to compute the average WER. When all the 3 streams are combined (FCMS: $n = 3$), we get a WER of 6.6%.

Fig. 2 also shows the plot for PLP, 24-Mel band derived spectral entropy and the combination of the two features by concatenation used in a similar setup. n is again varied from 1 to 3. When all the 3 streams are considered, we achieve a WER of 6.2%.

The oracle demonstrates that the performance of a multi-stream ASR system ($n = 2, 3$) is significantly better than a single-stream system ($n = 1$). Another important observation from Fig. 2 is, as the number of streams increases, the performance of oracle improves. However, the slope of the curve decreases when more streams are added, indicating that the additional streams bring less complementary information into the system.

4.2. Complementarity of streams

The property of oracle test that indicates the complementarity of feature streams is shown in Fig. 3. The figure shows performance for different noisy test conditions (additive factory noise at several SNRs). By combining PLP features with CJ-Rasta-PLP features in the FCMS setup (3 streams) using the oracle, we obtain a significant improvement in the performance over the baseline. When we combine PLP features with 24-Mel band derived spectral entropy features in the FCMS setup (3 streams) using the oracle, the improvement is more as compared to the one observed by adding the CJ-Rasta-PLP streams. This supports our earlier studies [11] and indicates that 24-Mel band derived spectral entropy features bring more complementary information into the system, and are a good candidate for multi-stream combination.

4.3. Relationship with minimum entropy

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

The entropy computed from the output posterior probabilities of an MLP classifier indicates the confidence of the classifier. A classifier with equal probabilities for all the classes has high entropy and does not convey any information. In contrast, a classifier with high posterior probability for one class and low posterior probabilities for rest of the classes has low entropy and indicates that the classifier has high confidence. Therefore, entropy at the output of a classifier can be used as a measure to weight the outputs of a classifier. The output posteriors of a classifier with high entropy should be given less weight and vice-a-versa. In [4] and [2], similar approaches were suggested for multi-band and multi-stream combinations, respectively. In minimum entropy weighting, which is again a 1/0 weighting method, at every time instant, the outputs of the classifier that has the least entropy are selected and sent for decoding.

In the simple setup, we computed the entropy of the stream selected by the oracle at each time step, and compared it with the entropy of all the other streams. Interestingly, in case of PLP and CJ-Rasta-PLP features being used for combination in the FCMS framework, in clean speech, 79.9% of the times oracle selection was the same as the selection done by minimum entropy weighting. That is, 79.9% of the times, minimum entropy stream was selected by the oracle. In case of multi-stream combination of PLP features with 24-Mel band derived spectral entropy features in the FCMS setup, oracle selected the minimum entropy stream 79.2% of the times.

Fig. 4 shows how many times (frames) oracle selected the minimum entropy stream for different noise levels (additive factory noise at several SNRs). We notice that as the noise level increases, the preference for the minimum entropy frames decreases, but still the minimum entropy frames enjoy a majority in oracle selection (random selection is 33% in case of 3 streams). This suggests that entropy at the output of a classifier is a reasonable choice for weighting, as suggested in our previous work [2].

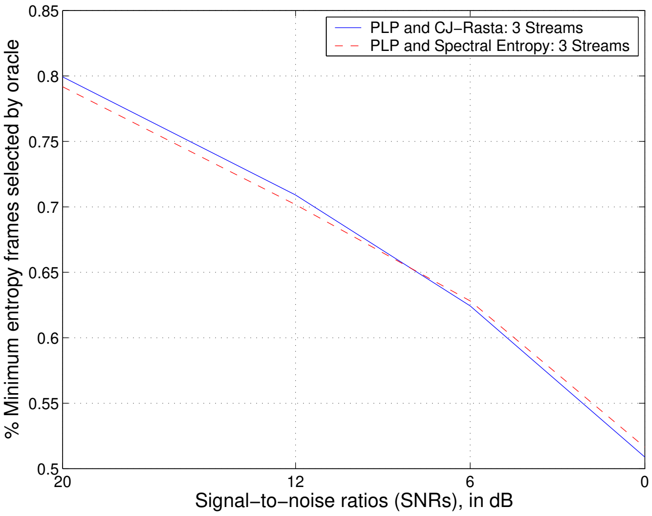

Figure 4: *Number of times (in percentage of frames) the oracle selected the stream with minimum entropy in FCMS hybrid system.. The plot is for clean as well as noisy (additive factory noise) test conditions.*

5. Results of single and multi-stream systems

In this section, the results for the combination of CJ-Rasta-PLP with PLP features and 24-Mel band derived spectral entropy with PLP features are presented. In these experiments, FCMS framework was used and the entropy weighting method suggested in [2] was used for combining the outputs of the MLP classifiers. Table 1 gives the results for different feature representations and the combination of CJ-Rasta-PLP and spectral entropy features individually with PLP features in the FCMS framework [9]. The Table

	Clean	**SNR12**	**SNR6**	**SNR0**
PLP	10.0	17.7	29.6	51.0
CJ-Rasta-PLP	10.6	17.1	27.9	48.6
Spectral Entropy	12.8	18.3	27.0	45.1
PLP,CJ-Rasta-PLP	9.4	15.3	26.4	46.8
PLP,Spectral Entropy	9.2	15.0	24.5	45.5

Table 1: *WER in % for different individual feature representations, PLP baseline with CJ-Rasta-PLP features in FCMS framework (***PLP,CJ-Rasta-PLP***), and PLP baseline with spectral entropy features in FCMS framework (***PLP,Spectral Entropy***).*

supports the results that were obtained by oracle. The combination of spectral entropy features with PLP baseline yields better improvements in the performance as compared to the improvements obtained by combination of CJ-Rasta-PLP and PLP features. This supports the oracle analysis that spectral entropy features bring more complementary information as compared to CJ-Rasta-PLP features when used along with PLP features.

6. Conclusions

In this paper, we presented a 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. In a multi-stream system, this property of the oracle may be used as an efficient method to select feature streams carrying high complementary information. The results obtained by oracle selection and entropy based weighting showed that when combined with PLP features, spectral entropy features were having more complementary information as compared to CJ-Rasta-PLP features. Also, in a multi-stream setup, we observed that the oracle tends to choose the MLP classifiers (trained on feature streams) that had the minimum output entropy. This further supports our previously proposed method of entropy based weighting for combining the outputs of the classifiers.

7. 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).

8. References

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

[2] Hemant Misra, Hervé Bourlard, and Vivek Tyagi, "New entropy based combination rules in HMM/ANN multi-stream ASR," in *Proceedings of ICASSP*, Hong Kong, Apr. 2003.

[3] Ludmila I. Kuncheva, "A theoretical study on six classifier fusion strategies," *IEEE Trans. Pattern Analysis and Machine Intelligence*, , no. 2, pp. 281–286, Feb. 2002.

[4] Shigeki Okawa, Enrico Bocchieri, and Alexandros Potamianos, "Multi-band speech recognition in noisy environments," in *Proceedings of ICASSP*, Seattle, Washington, May 1998, pp. 641–644.

[5] Richard Cole, M. Noel, T. Lander, and T. Durham, "New telephone speech corpora at CSLU," in *Proceedings of EuroSpeech*, 1995, vol. 1, pp. 821–824.

[6] 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.

[7] Hynek Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *J. Acoust. Soc. Amer.*, vol. 87, no. 4, pp. 1738–1752, 1990.

[8] Hynek Hermansky and Nelson Morgan, "RASTA processing of speech," *IEEE Trans. Speech, Audio Processing*, vol. 2, no. 4, pp. 578–589, 1994.

[9] Hemant Misra, Shajith Ikbal, Sunil Sivadas, and Hervé Bourlard, "Multi-resolution spectral entropy feature for robust ASR," in *Proceedings of ICASSP*, Philadelphia, U.S.A., Mar. 2005.

[10] Jia-lin Shen, Jeih-weih Hung, and Lin-shan Lee, "Robust entropy-based endpoint detection for speech recognition in noisy environments," in *Proceedings of ICSLP*, Sydney, Australia, 1998.

[11] Hemant Misra and Hervé Bourlard, "Spectral entropy feature in full-combination multi-stream system for robust ASR," in *Proceedings of EuroSpeech*, Lisbon, Portugal, Sept. 2005.