Combination of Sparse Classification and Multilayer Perceptron for Noise-robust ASR

Yang Sun [#], Mathew M. Doss [*], Jort F. Gemmeke [@], Bert Cranen [#], Louis ten Bosch [#], Lou Boves [#]*

[#]Centre for Language and Speech Technology, Radboud University Nijmegen, the Netherlands
[@]Department ESAT, KU Leuven, Belgium
[*]Idiap Research Institute, 1920, Martigny, Switzerland
[y.sun;b.cranen;l.tenbosch;l.boves]@let.ru.nl, jgemmeke@amadana.nl, mathew@idiap.ch

Abstract

On the AURORA-2 task good results at low SNR levels have been obtained with a system that uses state posterior estimates provided by an exemplar-based sparse classification (SC) system. At the same time, posterior estimates obtained with a multilayer perceptron (MLP) yield good results at high SNRs. In this paper, we investigate the effect of combining the estimates from the SC and MLP systems at the probability level. More precisely, the probabilities are combined by a sum or a product rule using static and inverse-entropy based dynamic weights. In addition, we investigate a modified dynamic weighting approach which enhances the contribution of the SC stream based on the information about static weights and average dynamic weights obtained on cross-validation data. Our study on the AURORA-2 task shows that in all conditions the modified dynamic weighting approach yields a dual-input system that performs better than or equal to the best stand-alone system.

Index Terms: multiple-stream combination, noise robustness, exemplar-based system, multilayer perceptron network

1. Introduction

Over the years, many acoustic features and modeling approaches have been developed for automatic speech recognition (ASR). However, there are no features or modeling approaches that yield a superior performance in all signal-to-noise ratio (SNR) conditions. Some perform well in clean conditions, while others perform best in noisy conditions. For example, the results in [1] show that their Sparse Classification (SC) system provides a higher degree of noise robustness than traditional Gaussian Mixture Model-based (GMM) systems, but under cleaner conditions the GMM systems perform better.

In our previous work [2], we investigated a combination of the likelihoods estimated by a GMM system and the class conditional probabilities estimated by an SC in the framework of dynamic Bayesian networks. The motivation was that the SC system and GMM system had different feature modeling capabilities. More precisely, the SC system stores spectral exemplars and each sliding windom of the test utterances are approximated as a linear combination of speech and noise exemplars from using sparse component analysis. Using the associated state-id labels of the speech exemplars, the posterior state probabilities of the frames of the test utterance are estimated using the weights from the linear combination. At the same time, the GMM system captures the distribution of cepstral features for each state and matches the test feature vectors with a statistical distribution. In addition, as noted above, the GMM and SC systems yield superior performance in different noise conditions. This difference could possibly be exploited and contribute to a dual-input system that performs well across all conditions. Our studies showed that the SC system did indeed provide complementary information to the GMM, especially in noisy conditions. However, under extremely noisy conditions the combined systems could not outperform the stand-alone SC system.

An inherent drawback of the approach in our previous research is that combining likelihoods (GMM) and probabilities (SC) is not trivial. In ASR systems it is common practice to use a language model scaling factor to combine language model probabilities with acoustic model likelihoods, and we applied a similar mechanism for combining the GMM likelihoods and the SC probabilities. This scaling factor is usually estimated by tuning the performance on cross-validation data and may not be the best for the unseen test data. Therefore, we want to develop a dynamic scaling method that is able to adapt to the unknown conditions in a test.

As an alternative to GMM systems, multilayer perceptrons (MLPs) have been proposed as a hybrid HMM/MLP system to ASR [3]. Thanks to the discriminative training mechanism, MLPs can *directly* estimate class conditional probabilities. In a hybrid HMM/MLP system, the estimated class conditional probabilities, usually after scaling them by the priors, are used as HMM state emission probabilities.

In this paper, we replace the HMM/GMM system in [2] by an HMM/MLP system and combine it with the SC system. By doing so, we alleviate the problem of combining probabilities and likelihoods, but we retain the advantage of the HMM/GMM system, i.e. a better performance at high SNRs. MLP and SC probability estimates are combined by the SUM or PRODUCT rule. In this paper we use the AURORA-2 corpus to compare the SUM and PRODUCT combination rules in the context of static and dynamic scaling methods. It is shown that in all noise conditions there is at least one combination method that is at least as good as the best baseline system.

The rest of the paper is organized as follows, in Section 2 we review the basic properties of the SC approach. Then the experimental setup and combining approaches are described in Section 3 and 4 respectively. In Section 5 a detailed discussion is given on the experimental results. Finally conclusions and

The research of Yang Sun has received funding from the European Community's Seventh Framework Programme [FP7/2007-2013] under grant agreement no. 213850 - SCALE. The research of Jort F. Gemmeke was funded by the IWT-SBO project ALADIN, under contract no. 100049. The research of Mathew M. Doss is partly funded by the Swiss National Science Foundation under the National Centre of Competence in Research (NCCR) on Interactive Multimodal Information Management (IM2).

suggestions for follow-up research can be found in Section 6.

2. The Sparse Classification system

As explained in [1], the SC system assumes that noisy speech segments can be expressed as a sparse, linear, non-negative combination of noise and clean speech exemplars from a fixed dictionary. This exemplar-based system is especially robust in very noisy conditions, thanks to the fact that it achieves an effective separation of speech and noise by taking advantage of long temporal contexts. With each frame within each exemplar labeled with the HMM-states for the full-word models proposed in [4], the likelihoods over all HMM-states can be estimated. Finally, likelihoods are converted into posterior probabilities, assuming a uniform prior distribution over all states.

3. Experimental Setup

We perform experiments on the multi-condition setup of AURORA-2 database [4]. Briefly, the training set comprises 8440 utterances in both clean and noisy (SNR$\geq$ 5 dB) conditions. We use two sets of test data, namely *test set A* and *test set B*, which have been created by adding noise to clean recordings. The *test set A* contains utterances corrupted by the same four noise types as the multi-condition training set, while *test set B* contains utterances corrupted by four different noise types. Both *test set A* and *test set B* contain 4004 utterances, each of which consists of a sequence of one to seven digits. All utterances occur at seven noise levels, viz. clean, and SNR = 20, 15, 10, 5, 0, and -5dB. We will report performance on both test sets.

3.1. SC System

The SC dictionary is generated by randomly selecting 4000 noise exemplars from reconstructed noise recordings and 8000 clean speech exemplars from the clean training set. Exemplars are represented as Mel-Frequency energy spectra with 23 bands. Each exemplar spans 300ms. The noise exemplars are extended by the same 23 artificial exemplars as in [5].

3.2. MLP

We trained a three layer MLP using the Quicknet software [6]. The input of the MLP are vectors of 39 dimensional perceptual linear prediction cepstral coefficients ($c_0 - c_{12} + \Delta + \Delta\Delta$) with a context of four preceding and four following frames. The output layer of the MLP consists of 179 classes corresponding to the HMM states of the full-word models (including silence). For training the MLP, an alignments in terms of the HMM states for the 8440 utterances in the multi-condition training data was obtained using the HTK based HMM recognizer proposed in [4].

We used 7685 utterances for training the MLP and the remaining 755 utterances for the cross-validation. The split was speaker-independent, i.e., the speakers in the cross-validation data are not present in the training set. During training, the frame level classification accuracy performance on the cross-validation data was used to guide the training process (i.e., control the learning rate). The optimal number of units in the hidden layer (850) was obtained with the validation data as well.

3.3. Analysis of estimates of SC and MLP Probabilities

We analyzed the properties of the class-conditional probabilities estimated by SC and MLP by computing the average entropy at different SNRs on the cross-validation data. As shown in Table 1, the entropy of the probabilities estimated by the SC system does not change much across different conditions, while the entropy of the probabilities estimated by MLP steadily increases when going from high to low SNRs. This brings to light two aspects. First, the SC system tends to distribute the probability mass over more classes/states than the MLP system. Second, and most importantly, the average entropy of the MLP output seems to be a good indicator of SNR of the signal.

Table 1: Averaged entropy per frame of the SC and MLP systems on the validation data of the AURORA-2 database

	clean	20	15	10	5
MLP	1.07	1.45	1.64	1.85	2.28
SC	2.50	2.61	2.62	2.71	2.76

4. Combination of SC and MLP Probabilities

In the literature, different rules for combining probabilities have been proposed [7]. Among them, the SUM and PRODUCT rule are the most commonly used. If $P^{sc}(k|\mathbf{x}_t)$ denotes the probability estimated by the SC system for class/state k given the feature $\mathbf{x}_t$ and $P^{mlp}(k|\mathbf{x}'_t)$ denotes the probability estimated by MLP classifier for class/state k given the feature $\mathbf{x}'_t$, then the local score/emission probability for HMM state k at time frame t is estimated by Eq. 1 or 2, where w_t^{sc} and w_t^{mlp} are the weights at time frame t for the SC and the MLP stream, respectively. Both w_t^{sc} and w_t^{mlp} are positive, and $w_t^{sc} + w_t^{mlp} = 1$.

sum rule (SUM)

$$s_t^{sum}(k|\mathbf{x}_t, \mathbf{x}'_t) = w_t^{sc} \cdot P^{sc}(k|\mathbf{x}_t) + w_t^{mlp} \cdot P^{mlp}(k|\mathbf{x}'_t) \quad (1)$$

product rule (PROD)

$$s_t^{prod}(k|\mathbf{x}_t, \mathbf{x}'_t) = P^{sc}(k|\mathbf{x}_t)^{w_t^{sc}} \cdot P^{mlp}(k|\mathbf{x}'_t)^{w_t^{mlp}} \quad (2)$$

Decoding is performed after estimating $s_t^{sum}(k|\mathbf{x}_t, \mathbf{x}'_t)$ or $s_t^{prod}(k|\mathbf{x}_t, \mathbf{x}'_t)$ for all states in a frame. Note that the combined estimates could be normalized to yield probabilities. However, the normalization factor at each time frame t is constant for all states k thus does not affect the decoding. The weights w_t^{sc} and w_t^{mlp} can be static or dynamic over time. We investigate both of these weighting techniques and details about how these weights are estimated are given in the following subsections.

4.1. Static Weighting

In the static weighting technique, a fixed weight is assigned to each probability stream over all time frames. In our case, the probability streams are the SC and the MLP probabilities. Weights for the different streams are usually optimized on a cross-validation data set. Since AURORA-2 does not have an explicit cross-validation set, we take the cross-validation data used in training the MLP to determine the static weights w^{sc} and w^{mlp} for the SC and MLP probability streams.

For both the SUM and the PROD combination technique, we performed a grid search of w^{sc} in the range of $[0, 1]$ in steps of 0.05. The optimal weights, $w^{sc} = 0.65$ for SUM and $w^{sc} = 0.7$ for PROD, were chosen such that they minimize the average word error rate in all conditions. Because we consider word error rate in all SNR conditions as equally important, higher weights for SC are to be expected: the largest

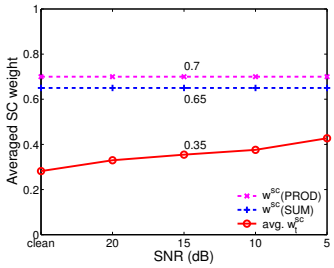


Figure 1: Static weights and averaged inverse-entropy-based dynamic weights of SC, estimated on the cross-validation data. w^{sc} (PROD) and w^{sc} (SUM) refer to the static weights estimated for the PRODUCT or the SUM rule, respectively. $avg\ w_t^{sc}$ denotes the averaged inverse-entropy-based dynamic weights of SC.

performance differences between the MLP and SC systems are found in the lowest SNR conditions, where SC is superior.

4.2. Dynamic Weighting

We use a weighting technique based on the inverse of entropy to estimate weights dynamically [8]. In this technique, the classifier which has the highest entropy at its output gets the lowest weight. If H_t^{sc} and H_t^{mlp} are the entropies of the output of the SC and MLP systems respectively,

$$H_t^{sc} = -\sum_{k=1}^{K} P^{sc}(k|\mathbf{x}_t) \cdot \log(P^{sc}(k|\mathbf{x}_t))$$

$$H_t^{mlp} = -\sum_{k=1}^{K} P^{mlp}(k|\mathbf{x}'_t) \cdot \log(P^{mlp}(k|\mathbf{x}'_t))$$

where K is the number of output classes, then,

$$w_t^{sc} = \frac{1/H_t^{sc}}{1/H_t^{sc} + 1/H_t^{mlp}} \quad (3)$$

$$w_t^{mlp} = 1 - w_t^{sc} \quad (4)$$

4.3. Integrating Static Weight information in the Dynamic Weighting

Our analysis in Section 3.3 showed that the entropy of the SC system output does not change much over different SNRs, contrary to the entropy of the MLP output, which does change across different SNRs. More precisely, the entropy of the MLP increases and approaches the entropy of SC in the conditions with the lowest SNRs.

Figure 1 shows the average of w_t^{sc} estimated by the inverse-entropy based weighting technique across different SNRs as a solid, red curve, together with the static weights w^{sc} obtained for the SUM and PROD combinations (shown as two horizontal lines). It can be observed that on average the SC system gets lower weights than the MLP system at all SNRs. This suggests that the inverse-entropy weighting technique may not fully exploit the complementarity between SC and MLP, neither in high nor in low SNRs. At high SNRs, the MLP on average gets very high weights, so that the SC system will contribute little to the combination. And at low SNRs, despite the fact that the SC system performs very well under noisy conditions, the weighting technique is not able to exploit this advantage, since SC does not obtain high enough weights at low SNRs.

In order to enhance the contribution of SC in both high and low SNRs, we tried a technique where the static weights information can be integrated with the dynamic weights. In Figure 1 we observe that the static weighting technique is assigning a higher weight to the SC. Based on the static weight, we estimated an enhancing factor γ for the dynamic weight of the SC system. This factor was calculated as the ratio between the static weights and the average inverse-entropy weight of 0.35 at SNR 15dB (the median SNR in the cross-validation data): $\gamma = 0.65/0.35 = 1.86$ for SUM, and $\gamma = 0.7/0.35 = 2.0$ for PROD.

The ratio can be seen as an indicator of how much the SC stream must be enhanced on average. Given the enhancing factor γ and the estimate of dynamic weights w_t^{sc} for the SC system, the new dynamic weights for the SC and MLP systems are estimated as

$$*w_t^{sc} = \min(\gamma \cdot w_t^{sc}, 1) \quad (5)$$

$$*w_t^{mlp} = 1 - *w_t^{sc} \quad (6)$$

5. Results and Discussion

Table 2 shows the performance of different systems on both *test set A* and *test set B*.

1. mlp(base): the stand alone HMM/MLP system
2. sc(base): the stand alone SC system
3. stc SUM: the sum rule combination using static weights w^{sc} and w^{mlp}
4. stc PROD: the product rule combination using static weights w^{sc} and w^{mlp}
5. dyn SUM: the sum rule combination using dynamic weights w_t^{sc} and w_t^{mlp}
6. dyn PROD: the product rule combination using dynamic weights w_t^{sc} and w_t^{mlp}
7. stc-dyn SUM: the sum rule combination using the modified dynamic weights $*w_t^{sc}$ and $*w_t^{mlp}$
8. stc-dyn PROD: the product rule combination using the modified dynamic weights $*w_t^{sc}$ and $*w_t^{mlp}$

A comparison of the two stand-alone systems shows that the hybrid HMM/MLP performs better in cleaner conditions and the SC system performs better in noisy conditions. This confirms the results for comparing SC with a traditional HMM/GMM system [1, 2].

The static weighting strategy can yield better word accuracy than either baseline system, except in the SNR -5dB condition in test set *A*. However, it can also be seen that the best performance is sometimes obtained with the SUM and at other times with the PRODUCT rule. Also, a compromise has to be made when using fixed weights for all conditions, while higher MLP weights and higher SC weights are expected to yield better performances at high and low SNRs respectively.

The dynamic weighting strategy allows for a frame-wise adaptation and the inverse entropy was shown to be a good indicator of SNRs. Still, it can be seen that dynamic weighting is not consistently better than static weighting. This is because the averaged dynamic weights of SC are always below the static counterparts (cf. Fig. 1).

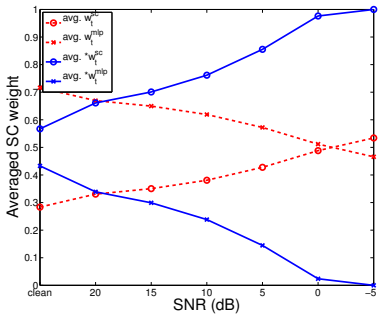
The modified dynamic weights, shown in Figure 2, move closer to the optimized static weights, and keep the adaptation to different SNR conditions. It can be observed that stc-dyn PROD always performs at least the same as, or better than, the

Table 2: Word recognition accuracy (in %) on the AURORA-2 task. The best performing stand-alone system in each condition is marked in Italics and the best performance for each SNR is in bold.

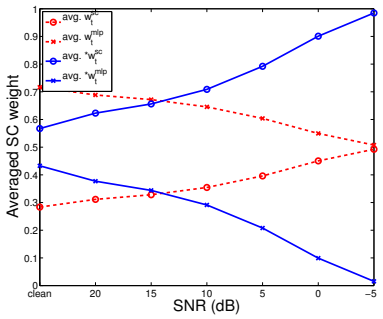
		test set A						test set B					
	clean	20dB	15dB	10dB	5dB	0dB	-5dB	20dB	15dB	10dB	5dB	0dB	-5dB
mlp(base)	*99.08*	*98.89*	*98.45*	*96.89*	91.80	72.80	35.67	*98.58*	*98.02*	*96.39*	*91.04*	71.48	33.41
sc(base)	96.16	95.68	95.22	94.38	*92.14*	*84.82*	*63.60*	95.65	95.18	93.47	88.46	*74.85*	*44.69*
stc SUM	98.90	98.62	98.23	97.19	**94.47**	85.27	58.74	98.56	98.04	96.86	92.92	**79.27**	45.49
stc PROD	99.29	98.83	98.31	97.26	**94.47**	84.39	54.65	98.83	98.23	**97.11**	92.66	78.73	45.75
dyn SUM	99.17	98.83	98.44	97.33	94.42	84.22	55.57	98.66	98.24	96.95	**93.03**	78.82	42.99
dyn PROD	99.30	**98.95**	**98.64**	**97.44**	94.06	82.49	50.07	**98.84**	**98.41**	97.02	92.68	78.06	42.96
stc-dyn SUM	99.05	98.63	98.21	97.13	94.16	85.98	63.92	98.57	97.98	96.38	91.86	77.93	**46.85**
stc-dyn PROD	**99.32**	98.85	98.41	97.17	94.38	**86.11**	**65.78**	98.80	98.06	96.50	91.41	77.68	46.14

best stand-alone system in all conditions, including SNR -5dB in test set *A*. However, it can also be seen in Table 2 that stc-dyn does not achieve a better performance than stc or dyn in the 20, 15, and 10 dB SNR conditions, in both test set A and B. This indicates that the optimal weights for different SNRs are not a linear function of the inverse entropy. Adaptations will be needed for further improvement.

Finally, a comparison between the SUM and PROD rules shows that PROD performs well at higher SNRs and SUM performs well at lower SNRs. It is well known in the literature that it is better to combine probabilities with the PROD rule if the classification performance is high (implying similar scores of the classifiers), which is the expected case in cleaner conditions, while SUM is better if the classification performance is low (implying low agreement between the classifiers) [9].



(a) test set A



(b) test set B

Figure 2: Average dynamic weights for System dyn and stc-dyn, shown as dashed red and solid blue curves, respectively. $avg\ w_t^{sc}$ and $avg\ w_t^{mlp}$ denote the averaged inverse-entropy-based dynamic weights and $avg * w_t^{sc}$ and $avg * w_t^{mlp}$ indicate the syn-dyn weights for each stream.

6. Conclusions and Future Work

In this work, we investigated the combination of an SC system and a hybrid HMM/MLP system at the probability level. Our study shows that although the combinations can yield improvements in all SNR condition, there is no single combination rule or weighting method that consistently achieved the best performance in all conditions. However, the modified dynamic weighting method which which integrates static and dynamic weights to enhance the contribution of the SC stream yields a system that performs better than or equal to the best stand-alone system in all SNR conditions.

In our future work, we will explore ways to estimate the enhancing factor dynamically. In addition, we will investigate and compare other dynamic combination approaches, such as the Dempster-Shafer method [10].

7. References

[1] J. F. Gemmeke, T. Virtanen, and A. Hurmalainen, "Exemplar-based sparse representations for noise robust automatic speech recognition," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 19, no. 7, pp. 2067–2080, 2011.

[2] Y. Sun, J. Gemmeke, B. Cranen, L. ten Bosch, and L.Boves, "Early fusion of sparse classification and gmm for noise robust asr," in *Proceedings of EUSIPCO*, Barcelona, Spain, 2011.

[3] H. Bourlard and N. Morgan, *Connectionist speech recognition: A hybrid approach*. Kluwer, 1994.

[4] H. G. Hirsch and D. Pearce, "The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proceedings of ISCA ITRW ASR2000*, 2000, pp. 29–32.

[5] J. F. Gemmeke, A. Hurmalainen, T. Virtanen, and Y. Sun, "Toward a practical implementation of exemplar-based noise robust ASR," in *Proceedings of EUSIPCO*, 2011, pp. 1490–1494.

[6] "The ICSI Quicknet Software Package [Online]. Available: http://www.icsi.berkeley.edu/speech/qn.html."

[7] C. Genest and J. V. Zidek, "Combining probability distributions: A critique and an annotated bibliography," *Statistical Science*, vol. 1, no. 1, pp. 114–135, 1986.

[8] 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 Acoustics, Speech, and Signal Processing (ICASSP)*, 2003.

[9] D. M. J. Tax, M. van Breukelen, R. P. W. Duin, and J. Kittler, "Combining multiple classifiers by averaging or by multiplying?" *Pattern Recognition*, vol. 33, no. 9, pp. 1475–1485, 2000.

[10] F. Valente and H. Hermansky, "Combination of acoustic classifiers based on Dempster-Shafer theory of evidence," in *Proceedings of ICASSP*, 2007.