Using Sparse Classification Outputs as Feature Observations for Noise-robust ASR

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

[#]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

Sparse Classification (SC) is an exemplar-based approach to Automatic Speech Recognition. By representing noisy speech as a sparse linear combination of speech and noise exemplars, SC allows separating speech from noise. The approach has shown its robustness in noisy conditions, but at the cost of degradation in clean conditions. In this work, rather than using the state probability estimates obtained with SC directly in a Viterbi decoding, the probability distributions of SC are modeled by Gaussian Mixture Models (GMMs), for which purpose we introduce a novel transformation. Results on the AURORA-2 task show that our proposed approach is effective in all high SNR conditions in both test set A and B. We achieve a word error rate reduction of 47.4% and 29.9% averaged cross 0-20dB SNR in test set A and B respectively. The reduction rate at clean goes up to 70.6% relative to the SC baseline.

Index Terms: template-based ASR, noise robustness, speech modeling

1. Introduction

Sparse Classification (SC) or Sparse Coding [1, 2, 3], a non-parametric exemplar-based approach to automatic speech recognition, has shown superior robustness in very noisy conditions. Noisy speech is modeled as a linear combination of both clean speech and noise exemplars; when a suitable dictionary of speech and noise exemplars is available, the exemplar-based model is inherently noise robust. In SC, each speech exemplar, which spans multiple frames to model dependencies between neighboring frames, can be labelled with a sequence of sub-word units. In this research, we used state labels from an HMM framework to label subsequent frames of the exemplars. Using the weights of the linear combination of speech exemplars these labels were then used to estimate unscaled likelihoods of the states in an unknown speech segment. These likelihoods can then be normalized to obtain probabilities. In the well-known AURORA-2 task [4], SC with a classical Viterbi backend outperforms traditional GMM-based systems significantly in matched noisy conditions [1], at the cost of a degradation of its performance in cleaner conditions.

In [5] it has been shown that transforming posterior probability estimates for context-independent phones obtained with a multi-layer perceptron (MLP) into a representation that makes them suitable for modeling as Gaussian Mixtures (similar to conventional MFCC features) can lead to improved performance in both clean and noisy conditions. In that approach the phone posterior probabilities estimated by a conventional three-layer MLP are used as base features after a two-step transformation, namely Gaussianization followed by decorrelation. Finally, the resulting so-called "tandem features" are processed by a conventional Gaussian Mixture Model (GMM)-based speech decoding system.

In this work, we investigated a similar approach for the SC output posterior probabilities. This is not evident: unlike the classifiers in [5], in which *all* phones get a non-zero probability mass, the probability vectors resulting from SC contain many "hard" zeros [1]. Moreover, the SC system yields posterior probability estimates for 179 states, instead of for the 18 phones that are relevant for the AURORA-2 task. As a result, it is not possible to use a straightforward log-transform, perhaps followed by a PCA for dimensionality reduction. To alleviate this problem, we here present an alternative way to whiten the SC probabilities into features that are suitable for being modeled by GMMs, by replacing all hard zeros (and all near-zero probability estimates) by samples drawn from an appropriately chosen Gaussian. The transformed SC features are modeled by GMMs in the traditional GMM-HMM based way. Experimental results on AURORA-2 task show that the proposed approach is effective in all conditions, except SNR -5dB in test set B.

The rest of the paper is organized as follows, in Section 2 we review the basic properties of the SC approach. Then we describe our proposed tandem approach and experiments in Section 3 and Section 4, respectively. This is followed by a discussion in Section 5. Finally conclusions and future plan can be found in Section 6.

2. Review of Sparse Classification

In this section, we first provide a brief review of the principle of the exemplar-based sparse representation and estimation of class conditional probabilities. And in Sec-

The research of Yang Sun has received funding from European Community's Seventh Framework Programme [FP7/2007-2013] under grant agreement no. 213850 - SCALE. The research of Jort F. Gemmeke was funded by IWT-SBO project ALADIN contract 100049.

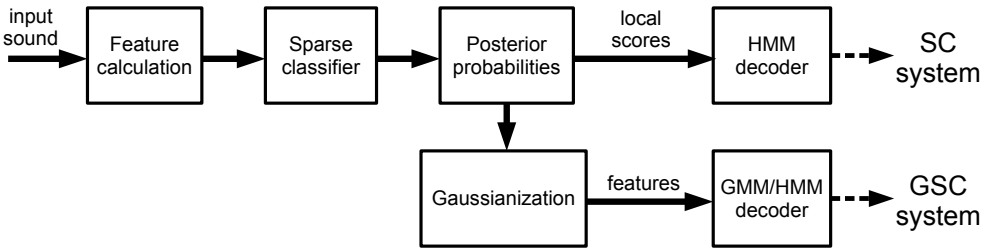

Figure 1: Block diagram of the SC baseline system, which uses posteriors as local scores, and the GSC system, in which posteriors are treated as features and imported to a GMM/HMM decoder.

tion 2.3 we discuss some limitations of non-parametric classifiers, such as SC.

2.1. Linear Combination

The SC system assumes that speech spectrograms can be expressed as a sparse, linear, non-negative combination of the spectrograms of clean speech exemplars $\boldsymbol{a}_j^s$, with $j = 1, \ldots, J$ denoting the exemplar index. We model noise spectrograms (after reshaping all matrices into vectors) as a linear combination of noise exemplars $\boldsymbol{a}_k^n$, with $k = 1, \ldots, K$ the noise exemplar index. This leads to representing noisy speech spectrograms $\boldsymbol{y}$ as a linear combination of both speech and noise exemplars:

$$\boldsymbol{y} \approx \boldsymbol{s} + \boldsymbol{n} \tag{1}$$

$$\approx \sum_{j=1}^{J} \boldsymbol{x}_j^s \boldsymbol{a}_j^s + \sum_{k=1}^{K} x_k^n \boldsymbol{a}_k^n \tag{2}$$

with $\boldsymbol{x}^s$ and $\boldsymbol{x}^n$ sparse representations of the underlying speech and noise, respectively. The sparse representations can be obtained by minimizing a cost function based on the generalized Kullback-Leibler (KL) divergence (For more details see [1]).

2.2. State Probability Estimation

Each exemplar $\boldsymbol{a}_j^s$ in the speech exemplar dictionary is labelled using HMM-state labels obtained from a conventional MFCC-based decoder. Using a frame-by-frame state description of the exemplars in the dictionary, we associate each exemplar $\boldsymbol{a}_j^s$ with a label matrix $\boldsymbol{\mathcal{L}}_j$, of dimensions $Q \times T$, with Q the total number of states in the system and T the number of frames in an exemplar. The matrix $\boldsymbol{\mathcal{L}}_j$ is a binary matrix containing for each frame $\tau \in [1, T]$ a single nonzero value for its corresponding state label. For each observed speech segment, the unscaled likelihood matrix is calculated as:

$$\boldsymbol{L} = \sum_{j=1}^{J} \boldsymbol{\mathcal{L}}_j x_j^s \tag{3}$$

In computing the likelihoods special attention must be paid to proper balancing between the likelihoods of the silence states on the one hand and the speech states on the other. As in [1] we increased the likelihood of the silence states by adding a value based on the estimated speech activity in each segment. Finally, the likelihoods are normalized into probabilities for each frame and thresholded to make sure that the Viterbi search always can find a complete path.

2.3. Drawbacks of Non-parametric Approaches

Like all exemplar (template) based systems, SC is a non-parametric classifier. A problem shared by all exemplar-based systems is that they are crucially dependent on the size and the representativeness of the dictionary [6, 7, 8] and that they may not generalize very well to data characteristics that are not represented in the exemplar dictionary. As a consequence, the posterior probability estimates from an SC system may occasionally be biased to the wrong states. Transforming the SC probabilities into features and *training* GMMs on these features may help to improve the capability of the output of a non-parametric classifier to generalize to unseen conditions. Therefore, we developed a method for transforming SC probability estimates such that they can be modeled by a GMM.

3. SC Probability-based Features

In the tandem approach proposed by [5], Gaussianized and decorrelated phone posterior features have already shown better performance on both clean and noisy speech than a conventional GMM or a hybrid system [9]. Similarly, we propose a GSC system in this work, where the estimated state posterior probabilities from SC are Gaussianized into features that can be used in a GMM/HMM system. The architecture is shown in the flowchart in Figure 1.

As briefly mentioned in Section 1, in each time frame many states receive a zero activation in the Sparse Representation. This results in a large proportion of hard zeros in the probability vectors produced by the SC system. This makes it impossible to Gaussianize the probability estimates by means of a straightforward log-transform. Moreover, especially in noisy cases, states with very small non-zero values in the SC probability vector are probably due to random speech exemplars, used in the linear SC approximation to fill in the gap between the unknown noisy speech and the noise-free speech exemplars. Consequently, states with very low probabilities may not represent useful information.

To alleviate this problem, a modified approach is de-

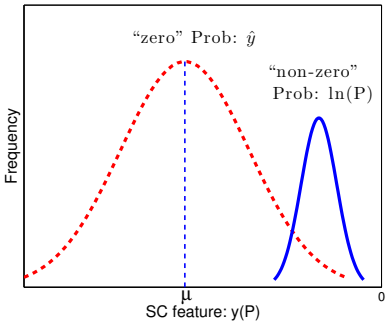

Figure 2: A schematic diagram of two distributions generated by two conditions in Eq. 4.

fined in Eq. (4), where all probabilities above a threshold θ are transformed by the logarithm as in the standard Tandem approach. All hard zeros and all values $< \theta$ (which might not carry useful information) are replaced by samples from a Gaussian distribution with mean μ and variance σ^2. Preliminary experiments have shown that decorrelating the transformed state posterior features does not improve performance.

$$y(P) = \begin{cases} \ln(P) & \text{if } P > \theta \\ \hat{y} \in \mathcal{N}(\mu, \sigma^2) & \text{otherwise} \end{cases} \quad (4)$$

The shape of the distribution obtained after the transformation defined by Eq. (4) is very schematically depicted in Figure 2. The narrow distribution on the right models the transformed estimates of the posterior "nonzwero" probabilities $\geq \theta$ of the states that were activated in the SC system; the wide distribution on the left results from the random values that replace the posterior "zero" probability values below θ. The resulting transformed *SC features* are fed as input to a traditional GMM/HMM system.

4. Experimental Setup and Results

We used the AURORA-2 database for our experiments, where a subset (755 utterances) of the multi-condition training data is set apart as the validation data set, which will be used to find optimal values for the parameters μ and σ in Eq. (4). The utterances in this set are evenly distributed across SNR and noise types. The remaining data are used in training the GMMs. We used test set A (utterances corrupted by the same noise types as in the multi-condition training set) and test set B, containing utterances corrupted by four other noise types. Both test set A and B contain 4004 utterances consisting of a sequence of one to seven digits, 1001 utterances for each noise type. All utterances occur in seven noise levels, viz. clean, and SNR = 20, 15, 10, 5, 0, and -5 dB.

The conventional AURORA-2 setup described in [4] is used in the SC system: 16 states are used for each of the 11 digit words and 3 states for the single silence model. To obtain the posterior probability estimates from the SC system, we used the same configuration as in [1] with a threshold $\theta = 10^{-3}$. In a nutshell, the SC method operates on 23-dimensional Mel-scale magnitude features, and uses a dictionary comprising 4000 clean exemplars, randomly extracted from the speech in the multi-condition training set and 4000 noise exemplars, also randomly selected from the multi-condition training set (by subtracting the corresponding clean speech signals). We used an exemplar size of 300ms (30 frames). For each frame, the output of the SC system is a 179 dimensional vector, corresponding to the posterior probability estimate of each state.

Table 1: *WER in %. Grid search on validation data across μ and σ in Eq. 4.*

		μ							
		-10	-9	-8	-7	-6	-5	-4	-3
σ	1	2.3	1.7	0.9	**0.5**	0.7	1.1	0.9	1.4
	2	1.1	0.9	0.8	0.9	0.9	0.9	0.9	0.9
	3	1.1	1.1	1.1	1.0	0.9	0.9	0.9	0.9
	4	1.2	1.3	1.2	1.1	1.1	1.1	0.9	1.1

Table 2: *WER in %.* SC *refers to the SC baseline and* GSC *refers to Tandem SC system. Relative improvements are given in the %*diff *column.*

	test set A			test set B		
SNR	*SC*	*GSC*	%diff	*SC*	*GSC*	%diff
clean	6.6	1.9	**+70.6**	6.6	1.9	**+70.6**
0-20 dB	11.8	6.7	**+43.2**	15.8	12.7	**+20.0**
-5 dB	42.9	30.4	**+29.1**	63.1	63.5	-0.1

In order to obtain the optimal values of μ and σ, we performed a grid search with μ varying from -10 to -3 and σ varying from 1 to 4 in steps of 1 on the validation data. The results (averaged over the five SNR conditions in the validation data) are shown in Table 1. The best pair of values ($\mu = -7$, $\sigma = 1$) in terms of WER is used in the transformation defined by Eq. (4). In training the diagonal covariance GMMs, three Gaussians are used for each state – this number was expected to be high enough to accurately model the distribution after application of Eq.4. Experiments are conducted in HTK [10]. The word error rates obtained for the test sets A and B are shown in Table 2.

5. Discussion

The purpose of the transformation in Eq. (4) is to obtain a distribution that is suitable for subsequent processing by a GMM/HMM decoding system. This means that we must shape and position the Gaussian that generates the random numbers to replace probability estimates $< \theta$ in such a way that it forms the tail of the distributions of the probability estimates $\geq \theta$, without overlapping too much with the latter distribution. The best performance is found when μ is set to -7, which is close to our threshold value θ in the log domain [$\ln(10^{-3}) = -6.9$]; if μ decreases further, performance on the validation set decreases. This suggests that most of the hard zero's in the original distribution must be considered as being part of the lower end of the tail of the non-zero distribution and

Table 3: *WER in %.* SC *refers to the SC baseline and* GSC *refers to Tandem SC system.* ETSI *refers to the reference system using* ETSI *advanced front-end* [11].

	test set A			test set B		
SNR	*SC*	*GSC*	*ETSI*	*SC*	*GSC*	*ETSI*
clean	6.6	1.9	**0.79**	6.6	1.9	**0.79**
0-20 dB	11.8	**6.7**	7.7	15.8	12.7	**8.2**
-5 dB	42.9	**30.4**	56.5	63.1	63.5	**57.7**

that the transformation needs to retain the continuity of the distribution. This is further supported by the fact that for lower μ the value of σ must be increased. The fact that given $\mu = -7$ optimal performance is found for $\sigma = 1$ (and not $\sigma > 1$) indicates that substantial overlap with the distribution of the non-zero estimates should be avoided.

From Table 2 it can be seen that the GSC system, which uses the SC features in a conventional GMM/HMM system improves the performance of using the 'raw' SC output in a Viterbi search in all SNR conditions, except the -5 dB condition in test set B. In the 0-20dB SNR conditions the average relative WER reduction is 43.2% in test set A and 20.0% in test set B respectively. The relative reduction in the clean condition is as high as 70.6%.

As in [5], we have no completely convincing explanation for the improvement of the GSC system over the SC system. The GSC system uses more parameters, and it is given an extra opportunity to learn the structure in the data. The fact that the overall improvement in test set B is substantially smaller than in test set A is probably due to the fact that learning three-mixtures GMMs from the multi-condition training data does not solve the problem that the noise exemplars in the SC system do not cover the noises in test set B. Thus, it seems that training GMMs does improve the capability of the posterior estimates of the SC system to generalize, but that this is not enough to compensate for the basic problem that the noise exemplar dictionary of the SC system is essentially incomplete.

Comparing to state-of-the-art ETSI [11] in Table 3, the proposed GSC system shows its noise robustness in matched noisy conditions (test set A). Especially at SNR -5dB, SC already performs significantly better than ETSI (42.9% vs. 56.5%) and GSC decreased the WER by another 12% (absolute). However, GSC's advantage cannot be generalized to unmatched noise types (test set B). A dramatic degradation of GSC's WER can be found from 30% to 63% from test set A to B at SNR -5dB. Plus, although a promising boost of the WER at clean can be observed at clean from SC to GSC, GSC is still worse than ETSI.

6. Conclusions and Future Work

In this work, we present a technique for transforming the output of an SC system in such a way that it becomes suitable for being modeled by means of GMMs. The purpose of this operation is to increase the capability of the non-parametric SC system to generalize to data that are not well covered by the exemplars. The results for the AURORA-2 task show that the approach does indeed improve the generalization power of the SC system substantially. However, the generalization is not enough to fully compensate for the fact that the noise exemplars do not cover some realistic noises.

Follow-up research can be done along two lines. First, there are several options for better handling the distributions of the posterior estimates. For example, instead of trying to force the distributions into a Gaussian framework, a linear discriminant analysis (LDA) can be directly used on raw posteriors [12] or the KL divergence based HMM approach [13] can be explored that removes the need for transforming the posterior distributions. The second line is to investigate the effect of the reduction of probability dimensions, for example using LDA or PCA, in order to scale well to large vocabulary tasks.

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] W. J. Smit and E. Barnard, "Continuous speech recognition with sparse coding," *Computer Speech and Language*, vol. 23, no. 2, pp. 200–219, 2009.

[3] Y. C. Cho and S. J. Choi, "Nonnegative features of spectro-temporal sounds for classification," *Pattern Recognition Letters*, vol. 26, no. 9, pp. 1327–1336, Jul. 2005.

[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] H. Hermansky, D. P. W. Ellis, and S. Sharma, "Tandem connectionist feature extraction for conventional HMM systems," in *Proceedings of ICASSP*, 2000, pp. 1635–1638.

[6] V. Maier and R. K. Moore, "Temporal episodic memory model: an evolution of minerva2," in *Proceedings of INTERSPEECH*, 2007, pp. 866–869.

[7] M. D. Wachter, M. Matton, K. Demuynck, P. Wambacq, R. Cools, and D. V. Compernolle, "Template-based continuous speech recognition," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 15, no. 4, pp. 1377–1390, 2007.

[8] J. F. Gemmeke, L. ten Bosch, L.Boves, and B. Cranen, "Using sparse representations for exemplar based continuous digit recognition," in *Proceedings of EUSIPCO*, Glasgow, Scotland, August 24–28 2009, pp. 1755–1759.

[9] H. Bourlard and N. Morgan, "Hybrid HMM/ANN systems for speech recognition: Overview and new research directions," in *International School on Neural Nets: Adaptive Processing of Temporal Information*. Springer Verlag, 1997.

[10] S. J. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book Version 3.4*. Cambridge University Press, 2006.

[11] H. Hirsch and D. Pearce, "Applying the advanced ETSI frontend to the aurora-2 task," in *in version 1.1*, 2006.

[12] T. N. Sainath, D. Nahamoo, B. Ramabhadran, D. Kanevsky, V. Goel, and P. M. Shah, "Exemplar-based sparse representation phone identification features," in *Proceedings of ICASSP*, Prague, Czech Republic, 2011.

[13] G. Aradilla, H. Bourlard, and M. Magimai-Doss, "Using KL-based acoustic models in a large vocabulary recognition task," in *Proceedings of Interspeech*, 2008, pp. 928–931.