INTERSPEECH 2009 BRIGHTON

Trimmed KL Divergence between Gaussian Mixtures for Robust Unsupervised Acoustic Anomaly Detection

Nash Borges and Gerard G. L. Meyer

Human Language Technology Center of Excellence
Department of Electrical and Computer Engineering
The Johns Hopkins University, Baltimore, MD 21218
{nashborges,gglmeyer}@jhu.edu

10.21437/Interspeech.2009-673

Abstract

In previous work [1], we presented several implementations of acoustic anomaly detection by training a model on purely normal data and estimating the divergence between it and other input. Here, we reformulate the problem in an unsupervised framework and allow for anomalous contamination of the training data. We focus exclusively on methods employing Gaussian mixture models (GMMs) since they are often used in speech processing systems. After analyzing what caused the Kullback-Leibler (KL) divergence between GMMs to break down in the face of training contamination, we came up with a promising solution. By trimming one quarter of the most divergent Gaussians from the mixture model, we significantly outperformed the untrimmed approximation for contamination levels of 10% and above, reducing the equal error rate from 33.8% to 6.4% at 33% contamination. The performance of the trimmed KL divergence showed no significant dependence on the investigated contamination levels.

Index Terms: anomaly detection, speech activity detection, unsupervised learning, Kullback-Leibler divergence.

1. Introduction

One important aspect of human intelligence that speech processing systems often lack is the ability to recognize the unknown. State-of-the-art speaker, language, and word recognition systems perform quite well when the training data and test data are similar. Large corpora of speech labeled for such systems have been carefully constructed to simulate real-world problems. However, the data is often homogeneous with respect to noise and channel conditions and it is usually culled of any gross anomalies deemed irrelevant to the task at hand. In the interest of robust speech processing, our goal was to develop an anomaly detector to supplement existing classifiers by recognizing novel data that should be processed differently or flagged for review.

Supervised machine learning algorithms predict an output $y \in \mathcal{Y}$ for each input $x \in \mathcal{X}$ using a set of manually labeled training data $\{(x_1, y_1), \ldots, (x_n, y_n)\}$. In a binary classification context we can define $\mathcal{Y} = \{0, 1\}$ and use examples from both classes to develop a generative model $p(x, y)$ or a discriminative classifier $p(y|x)$.

If we label each normal example with $y = 0$ and each anomaly with $y = 1$, anomaly detection falls naturally into this binary classification framework. Research on the detection of outliers and anomalies has been conducted in other fields such as network intrusion detection [2] and fraud detection [3]. In Hodge and Austin's survey of the subject [4], they categorized the research into three methods: those that required labeled examples of both classes, those that learned from only normal data, and those that were completely unsupervised. A subset of the unsupervised methods used a strategy of accommodation, incorporating some outliers into the model and using a robust classification method to detect them later on.

The term "robustness" was first coined in the field of statistics by Box when he noted the insensitivity to non-normality for tests of equal means [5]. Robust statistics provide alternatives to classical statistical methods without being severely impacted by outliers or inaccurate assumptions. For robust estimates of central tendency, Tukey [6] and others proposed the trimmed and Winsorized means, which remove a fraction of the smallest and largest samples and either discard them or replace them with the maximal remaining values. Shortly thereafter, Huber observed that classical estimators were not robust due to their inherent normality assumptions and reliance on least squares estimation [7] . He developed a general theory of robust statistics and showed the mean, median, and maximum likelihood (ML) estimates all to be special cases of M-estimators that minimize some function of the error between the samples and the estimator.

In order to quantify robustness, Hampel defined the "breakdown point" of an estimator to be the smallest amount of contamination that can cause it to take on "arbitrarily large aberrant values" [8]. The mean is not robust with breakdown point 0 since a single outlier can arbitrarily effect the estimate, whereas the α-trimmed mean is robust with breakdown point α, and the median with breakdown point $\frac{1}{2}$. In this work, we sought to use robust statistics to make our previous GMM approaches to anomaly detection less dependent on purely normal training data.

2. Data

2.1. Syllable Rate Features

One challenge in speech processing is dealing with a large quantity of data. While many speech tasks use spectral features computed every 10 ms resulting in approximately 50 dimensions, we used two features from a syllable rate speech activity detector (SRSAD) [9] computed every 100 ms. Since speech has a syllable rate of approximately 5 Hz, the frequency of its envelope modulation is different from that of white noise. Using a sliding half-second window of audio, SRSAD computes both the expected value of this modulation frequency and an estimate of its power. We modeled the distribution of this two-dimensional sequence as a set of independent observations.

2.2. Anomalous Data

We developed a set of synthetic anomalies known to be problematic to SRSAD, such as tones and noises of short duration and certain kinds of muzak [10]. The anomalous set was comprised of 50 examples of each of the following: DTMF sequences, Morse code, MIDI tones, MIDI songs, and various telephony noises. Using varying tone lengths from 25 ms to 1.25 seconds and a set of always-on MIDI tones centered at frequencies from 10 Hz to 300 Hz, Audacity [11] plug-in effects were used to generate random DTMF and Morse code sequences. The MIDI songs were downloaded from the MIDI Database [12] and telephony noises (busy signal, modem, etc) were obtained from FindSounds [13]. Since some of these noises were of short duration, the audio was cycled prior to feature generation until each was at least 5 minutes long. Half of the examples of each anomaly type were randomly selected for testing and the other half were reserved for possible use as training contamination.

2.3. Normal Data

The CallHome English corpus [14] of unscripted conversational speech between family and friends was selected to represent normal audio. Each conversation side in the *train* set was divided into 5 minute segments and 250 were randomly resampled from the total 918 to create ten training sets. This enabled us to estimate the variance in performance and allowed for up to 33% contamination when using all 125 anomalous segments. The English *eval* set was similarly divided into 5 minute segments yielding 226 for testing. We performed all our testing on contiguous 30 second segments, which were randomly selected from each 5 minute segment at test time.

3. Mostly Normal Model

We began by building a GMM to characterize the training data that we assumed to be mostly normal. We experimented with GMMs using up to 16 components with full covariance matrices.

The probability of an input $x \in \mathbb{R}^d$ for a single Gaussian is

$$\mathcal{N}(x;\mu;\Sigma) = \frac{\exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right)}{(2\pi)^{d/2}|\Sigma|^{1/2}} \quad (1)$$

with mean vector μ and covariance matrix Σ. Several of these are combined to form a mixture of m Gaussians,

$$p(x|\theta) = \sum_{i=1}^{m} w_i \cdot \mathcal{N}(x;\mu_i;\Sigma_i) \quad (2)$$

each with its own mean μ_i, covariance matrix Σ_i, and weight w_i, such that $w_i > 0$ for $i = 1,\ldots,m$ and $\Sigma_{i=0}^{m} w_i = 1$.

Unlabeled data often comes at a low cost, so we were not concerned with using all of it. Whenever training the mostly normal model (MNM), we randomly assigned 66% of the data to an *initial* set, keeping the remainder in a *heldout* set. Training GMMs using the Expectation-Maximization (EM) algorithm can be a delicate process and we wanted to avoid local maxima and overfitting to the training data. To deal with the former we trained eight separate initial models. Each initial model was initialized using k-means clustering on 1000 samples randomly chosen without replacement from the *initial* set. After a maximum of 10 iterations of k-means clustering, we performed ML estimation using the EM algorithm until the parameters converged. The initial model with the highest log likelihood for the *heldout* set was then selected. Using this model and all of the data in the *initial* set, we continued to perform EM iterations while the log likelihood of the *heldout* set increased to ensure that the model would generalize.

4. Methods

4.1. Average Log Likelihood Baseline

We first present a baseline anomaly detector using the average log likelihood of an input sequence. Here, $X = \{x_{n+1},\ldots,x_{n+t}\}$ was labeled as anomalous if

$$\frac{1}{t}\sum_{i=n+1}^{n+t} \log p(x_i|\theta) < \lambda, \quad (3)$$

where the threshold λ was set so that the false alarm rate was equal to the miss rate. This equal error rate (EER) allowed us to summarize detection performance with a single number.

4.2. Distributional Anomaly Detection

The distribution of syllabic rate features is noticeably different between anomalies and normal audio [1]. To obtain a model of a test sequence, we initialized the parameters to those of the MNM and performed ML estimation using the EM algorithm on the test data. While it is common to use MAP adaptation in such a scenario, we felt that deriving prior probabilities for the parameters using mostly normal data could not be justified when adapting to data that might be severely anomalous.

The comparison of two distributions $p(x)$ and $q(x)$ is often done using the Kullback-Leibler (KL) divergence [15],

$$KL(p\|q) = \int p(x)\log\frac{p(x)}{q(x)}dx. \quad (4)$$

We used the MNM for $p(x)$ and the test segment model for $q(x)$, which was found to perform better than the alternative.

4.2.1. KL Divergence Approximation for GMMs

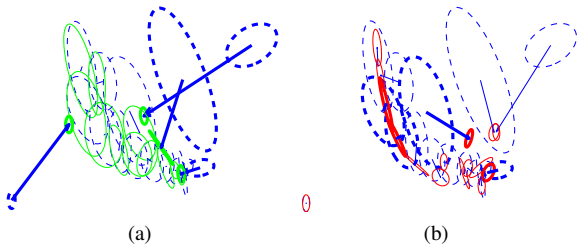
For single Gaussians, $p(x) = \mathcal{N}(x;\mu_p;\Sigma_p)$ and $q(x) = \mathcal{N}(x;\mu_q;\Sigma_q)$, the KL divergence can be computed directly [16],

$$KL_{\mathrm{G}}(p\|q) = \frac{1}{2}\log\left(\frac{|\Sigma_p|}{|\Sigma_q|} + \mathrm{Tr}|\Sigma_p^{-1}\Sigma_q| - d + (\mu_p-\mu_q)^T\Sigma_q^{-1}(\mu_p-\mu_q)\right). \quad (5)$$

However, there is no such closed form expression between two GMMs. We used the approximation suggested by Goldberger et al. [17],

$$\widehat{KL}_{\mathrm{GMM}}(p\|q) = \sum_{i=1}^{m} w_{p,i}\left(KL_{\mathrm{G}}\left(p_i\|q_{\pi(i)}\right) + \log\frac{w_{p,i}}{w_{q,\pi(i)}}\right) \quad (6)$$

and substituted our own trivial mapping function $\pi(i) = i$, since our adaptation strategy resulted in a correspondence between the Gaussians of each mixture model.



(a) (b)

Figure 1: GMM ML adaptation from the MNM (dashed blue) with 33% training contamination to (a) CallHome English cut 4829 side A (solid green) and (b) a random sequence of DTMF tones keyed on and off every 700 ms (solid red). The four most divergent Gaussians that would be trimmed are shown with thicker lines.

4.2.2. Trimming Gaussians for Robustness

With sufficient contamination some Gaussians in the MNM would inevitably model anomalous regions of the feature space. When adapting to normal data, changes in these Gaussians could lead to false alarms. With labelled training data we could have estimated which Gaussians were modeling anomalous data and then discarded them before estimating the KL divergence. In our unsupervised setting we did not have such labels, and our aim was to exploit the mostly normal data by discarding a fraction of the most divergent Gaussians. Our approach begins by treating the summands of Equation 6,

$$d_i = w_{p,i}\left(KL_{\mathrm{G}}\left(p_i \| q_{\pi(i)}\right) + \log\frac{w_{p,i}}{w_{q,\pi(i)}}\right) \tag{7}$$

as samples of a random variable D whose location we want to estimate robustly. We do so by discarding α of the largest d_i's using the one-sided trimmed mean,

$$\widehat{KL}_{\alpha\text{-trimmed}}\left(p\|q\right) = m\left(\frac{1}{m-k}\sum_{j=1}^{m-k} d_{(j)}\right) \tag{8}$$

with $d_{(j)}$ denoting the order statistics and $k = \lfloor \alpha m \rfloor$. We also tried the more traditional two-sided trimmed mean [18], but found it did not perform as well. We attributed this to the lack of negative outliers in the right-skewed distribution of D.

An example of the adaptation from the MNM to a normal segment is shown in Figure 1(a). A change in a few of the Gaussians lead to a divergence of 20.9 using Equation 6. Adaptation from the same MNM to an anomaly is shown in Figure 1(b). The resulting divergence was only 2.9 despite more of the Gaussians being affected by the adaptation. After discarding the four most divergent Gaussians (shown with thicker lines) in an unsupervised manner, the normal and anomalous segments had trimmed KL divergences of 0.2 and 1.3, respectively.

5. Experimental Results

5.1. Training Data Contamination

The aim of this work was to relax our previous assumption that a large amount of purely normal data was available for training. Obtaining data that is mostly normal is relatively inexpensive since it does not require any annotation. We investigated if any methods could robustly model the normal data, even when it was partially contaminated with anomalies. Figure 2 shows

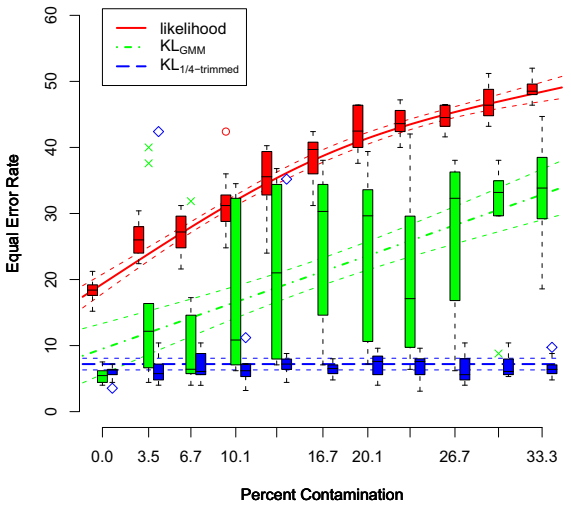


Figure 2: Box and spline plots of anomaly detection EER at various contamination levels using 16 Gaussians. The dotted lines around the splines indicate 95% confidence levels.

the EER for each of the investigated methods using mixtures of 16 Gaussians as contamination levels were varied from 0% to 33% in approximate increments of $3\frac{1}{3}\%$. To examine the relationship between performance and contamination, we first performed linear regression and then fit natural cubic splines to assess the linearity. Model selection for the splines was performed using the Bayes information criterion (BIC) [19].

Linear regression for the log likelihood method suggested that error rate increased with the amount of contamination (EER% = 22 + 0.87 per contamination percent, $r^2 = 0.87, P < 0.001$). However, spline fitting suggested that the relationship was slightly nonlinear ($df = 2, P < 0.001$). For purely normal training data, the KL divergence achieved the lowest EER (5.8%) with a median absolute deviation (MAD) of 1.3%. The difference between its performance and the $\frac{1}{4}$-trimmed KL divergence (6.3% EER, 0.9% MAD) was not significant ($P = 0.44$) using a Wilcoxon paired-sample signed rank test.

For contaminated training data, trimming one quarter of the Gaussians resulted in significantly better performance, with one exception at 6.7% contamination ($P = 0.19$). The variability of performance for the untrimmed KL divergence was not well accounted for with a linear model (EER% = 9.5 + 0.70 per contamination percent, $r^2 = 0.32, P < 0.001$), but the BIC suggested that higher order splines offered no better fit. The performance of the $\frac{1}{4}$-trimmed KL divergence did not show a significant dependence on the amount of contamination ($P = 0.53$) and a constant model resulted in a better fit (EER% = 7.6, $P < 0.001$). The median EER for all methods including $\frac{1}{2}$-trimming KL divergence are shown in Table 1 for select contamination levels.

5.2. Model Selection

We also evaluated the performance of each method as we varied the number of Gaussians from 1 to 16 when training on 33% contaminated data (Figure 3). Spline fitting suggested that all

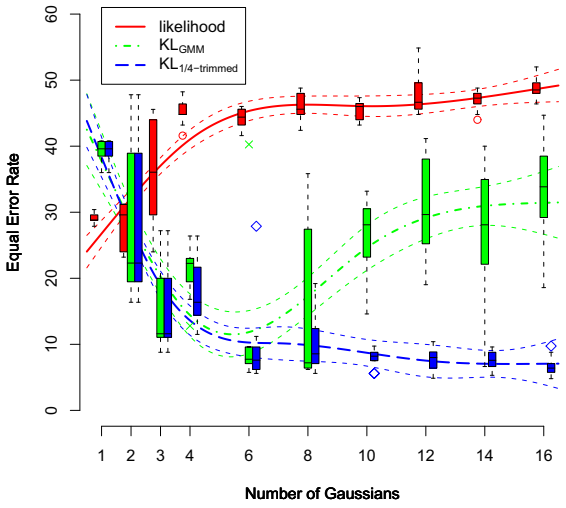


Figure 3: Box and spline plots of anomaly detection EER for various model complexities when training with 33% anomalous contamination. The dotted lines around the splines indicate 95% confidence intervals.

relationships were nonlinear, with three degrees of freedom for both the log likelihood method and KL divergence and four degrees of freedom for the $\frac{1}{4}$-trimmed KL divergence. The log likelihood method achieved its lowest EER of 28.8% (1.2% MAD) using a single Gaussian, although this was not consistent for other contamination levels. The KL divergence achieved its lowest EER of 7.7% (1.5% MAD) using 6 Gaussians. Both methods showed a dependence on model complexity that would require careful optimization for any new data set. In contrast, trimming one quarter of the Gaussians resulted in robust performance over a wide range of model complexities (6.4% EER, 1.0% MAD for 16 Gaussians). This performance was significantly better than the untrimmed KL divergence with a comparable model complexity ($P < 0.001$), but not when compared to the untrimmed divergence using 6 Gaussians ($P = 0.053$).

6. Conclusion

Using only unlabeled data, our goal was to develop a robust acoustic anomaly detector using two syllable rate features from a speech activity detector. When trained on purely normal data, we found that the KL divergence achieved the lowest EER of the three methods. When subjected to training contamination, the performance of the KL divergence suffered dramatically and optimizing its model complexity became extremely important. Seeing the merit in this approach, we wanted to improve its robustness to contamination.

After trimming one quarter of the most divergent Gaussians we found that the effect of contamination up to 33% was not significant. We experimented with other trimming ratios, but one quarter had the most consistent performance regardless of contamination level and model complexity. Such a detector could work in tandem with other speech processors, enabling the overall system to have a means of detecting anomalous audio at little additional cost.

Table 1: Percent Equal Error Rate (Median with n=10)

Contamination		0%	3.5%	10%	20%	33%
4 Gauss.	likelihood	19.2	22.4	22.8	24.4	44.8
	$\widehat{KL}_{\text{GMM}}$	4.8	8.8	10.1	12.0	22.3
	$\widehat{KL}_{\text{1/4-trim}}$	4.6	7.2	8.4	15.8	16.4
	$\widehat{KL}_{\text{1/2-trim}}$	5.9	11.8	15.1	27.1	24.8
8 Gauss.	likelihood	20.0	24.0	29.6	38.8	45.6
	$\widehat{KL}_{\text{GMM}}$	4.4	5.2	5.6	6.9	7.2
	$\widehat{KL}_{\text{1/4-trim}}$	4.4	4.2	5.9	6.8	8.6
	$\widehat{KL}_{\text{1/2-trim}}$	6.8	8.4	9.2	10.8	13.0
16 Gauss.	likelihood	18.4	26.0	31.2	42.5	48.5
	$\widehat{KL}_{\text{GMM}}$	5.5	12.2	10.8	29.6	33.8
	$\widehat{KL}_{\text{1/4-trim}}$	6.3	5.8	6.2	7.6	6.4
	$\widehat{KL}_{\text{1/2-trim}}$	11.6	12.8	12.2	12.4	11.2

7. References

[1] N. Borges and G. G. L. Meyer, "Unsupervised distributional anomaly detection for a self-diagnostic speech activity detector," in *CISS*, 2008.

[2] A. Lazarevic, L. Ertoz, V. Kumar, A. Ozgur, and J. Srivastava, "A comparative study of anomaly detection schemes in network intrusion detection," in *ICDM*, 2003.

[3] R. J. Bolton and D. J. Hand, "Unsupervised profiling methods for fraud detection," in *Proc. Credit Scoring Conference*, 2001.

[4] V. J. Hodge and J. Austin, "A survey of outlier detection methodologies," *Artificial Intelligence Review*, vol. 22, pp. 85–126, 2004.

[5] G. Box, "Non-normality and tests on variances," *Biometrika*, vol. 40, no. 3-4, pp. 318–335, 1953.

[6] J. Tukey, "A survey of sampling from contaminated distributions," *Contributions to probability and statistics*, pp. 448–485, 1960.

[7] P. Huber, "Robust estimation of a location parameter," *The Annals of Mathematical Statistics*, pp. 73–101, 1964.

[8] F. Hampel, "A general qualitative definition of robustness," *The Annals of Mathematical Statistics*, pp. 1887–1896, 1971.

[9] D. J. Nelson, D. C. Smith, and J. L. Townsend, "Voice activity detector," US Patent No. 6556967, April 2003.

[10] D. C. Smith, J. Townsend, D. J. Nelson, and D. Richman, "A multivariate speech actitivity detector based on the syllable rate," in *ICASSP*, 1999.

[11] Audacity: Plug-In Effects, *http://audacity.sourceforge.net/download/plugins*. January, 2009.

[12] MIDI Database, *http://www.mididb.com*. January, 2009.

[13] FindSounds, *http://www.findsounds.com*. January, 2009.

[14] A. Canavan, D. Graff, and G. Zipperlen, "CALLHOME American English Speech," Linguistic Data Consortium, Philadelpha, 1997.

[15] S. Kullback and R. A. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, vol. 22, no. 1, 1951.

[16] J. R. Hershey and P. A. Olsen, "Approximating the Kullback Leibler divergence between Gaussian mixture models," in *ICASSP*, 2007.

[17] J. Goldberger, S. Gordon, and H. Greenspan, "An efficient image similarity measure based on approximations of KL-divergence between two Gaussian mixtures," in *ICCV*, 2003.

[18] R. A. Maronna, D. R. Martin, and V. J. Yohai, *Robust Statistics: Theory and Methods*, ser. Probability and Statistics. Wiley, 2006.

[19] G. Schwarz, "Estimating the dimension of a model," *The Annals of Statistics*, pp. 461–464, 1978.