A discriminative reliability-aware classification model with applications to intelligibility classification in pathological speech

Naveen Kumar, Shrikanth S. Narayanan

Signal Analysis and Interpretation Lab,
Department of Electrical Engineering,
University of Southern California, Los Angeles.
komathnk@usc.edu, shri@sipi.usc.edu

Abstract

Many computational paralinguistic tasks need to work with noisy human annotations that are inherently challenging for the human annotator to provide. In this paper, we propose a discriminative model to account for the inherent heterogeneity in the reliability of annotations associated with a sample while training automatic classification models. Reliability is modeled as a latent factor that governs the dependence between the observed features and its corresponding annotated class label. We propose an expectation-maximization algorithm to learn the latent reliability scores using maximum entropy models in a mixture-of-experts like framework. In addition, two models - a feature dependent reliable model and a feature independent unreliable model are also learned. We test the proposed method on classifying the intelligibility of pathological speech. The results show that the method is able to exploit latent reliability information on feature sets that are noisy. Comparing against a baseline of reliability-blind maximum entropy model, we show that there is merit to reliability-aware classification when the feature set is unreliable.

Index Terms: reliability, mixture-of-experts, pathological speech, crowdsourcing

1. Introduction

Despite significant advances in machine learning, we find that certain pattern recognition problems are intrinsically more complex than others. For example, in a classification task certain classes may be more "noise-like" or difficult to model compared to others. The challenge in these tasks often lies in adequately modeling the variability in observations within a given dataset. Consider the binary classification task for intelligibility in pathological speech [1]. The non-intelligible (NI) class in this case often presents more variability than the intelligible (I) class. Standard machine learning approaches to model this variability consider models such as mixture of Gaussians with a large number of parameters or currently popular methods using deep learning [2, 3] to explain the variability of the NI class. However, the modeling power and complexity of such techniques comes at the risk of overfitting on noisy data. This problem is further compounded by data sparsity or increase in the number of feature dimensions. Moreover, the results obtained by such methods are often not directly interpretable, which makes it difficult to scale the method to other datasets. Thus, directly modeling the variability in feature distribution might not always be the best approach to deal with heterogeneity in class distributions.

Algorithms such as *boosting* [4] try to solve this problem by weighting data samples differently to sequentially train a cascade of weak learners. Misclassified samples are given more weight in subsequent stages of training such that the learners are complementary. In fact, boosting belongs to an increasingly popular class of machine learning algorithms referred to as ensemble methods [5, 6]. Ensemble methods such as random forests [7, 8] are also able to jointly identify the features that are useful for classification. Yet another variation of ensemble methods is known as mixture-of-experts [9, 10] in which each "expert" classifier is proficient in modeling certain aspects/ regions of the feature space.

We also observe that in most practical classification tasks, heterogeneity in the dataset often results from outliers. This usually occurs as a result of either the feature observations or their corresponding labels being noisy. The latter is more common in tasks requiring subjective or perceptual judgment from human annotators, common in computational paralinguistics [11, 12]. Hence, robustness to noisy samples or annotations is important for these algorithms. A well-known example of an algorithm that is robust to noisy features is the Support Vector Machine (SVM) [13] classifier. SVM maximizes the margin between the class boundary and the nearest sample from each class known as the *support vectors* [14]. Alternatively, the Random Sampling Consensus (RANSAC) [15] paradigm is also used for removing outliers from the dataset by finding the largest subset of the data that has consensus within itself with respect to the classification model and can be assumed to contain only inliers. Along these lines, we proposed a Bayesian network for feature fusion in [16] that accounts for the reliability of each sample and feature set for the classification task during training. Reliability is defined as a latent factor in the generative model that controls the dependence between features and the class label (discussed further in Section 2).

On the other hand, the issue of robustness to noisy annotations has also gathered significant attention. It has found particular importance in the context of *crowdsourcing* experiments [17] where it might not be always feasible to ensure the reliability of annotators beforehand. Popular methods for modeling annotator reliability often assume a *noisy channel* model that distorts the true class label in an annotator-dependent fashion [18]. An extension to this model was proposed in [19], by jointly learning a classifier that could exploit feature information in addition to multiple noisy annotations. In [20] the authors proposed a globally-variant locally-constant model to introduce data-dependent distorion models for each annotator.

Work supported by NSF, NIH and DoD

While these models can be used to estimate the true labels on a dataset, their results cannot be interpreted directly. They also assume the noisy labels to be a function of the true label and features, which might not be necessarily true in all cases. To deal with this problem , we propose a discriminative reliability-aware classification model in this paper, as an extension of the latent reliability model proposed in [16]. This proposed discriminative model allows us to overcome issues of data-sparsity and poor parameter estimation encountered by the generative model proposed in the previous work. We use a maximum-entropy (MaxEnt) model to parametrize the relation between class labels and features. The latent sample reliability is inferred using a logistic regression on the features, that is trained jointly with the MaxEnt classification model for predicting class labels. The proposed model can also be viewed as a mixture-of-experts where the two experts respectively capture reliable and unreliable characteristics of the data samples. We perform experiments on a binary intelligibility classification task for pathological speech [21]. Results are compared against a traditional logistic regression model that assumes all samples to be equally reliable.

2. Bayesian Reliability Aware Model

We proposed a Bayesian model in [16], for taking into account the latent reliability associated with each feature modality. The central assumption in this model was to introduce feature reliability as a latent *binary* random variable R, that controlled the dependence between features X and the class labels Y. When a feature is unreliable ($R = 0$) it is assumed to have been generated at random from a garbage model, irrespective of the class label. This assumption is formalized in Eq.(1) where $\mathbf{\Theta}$ represents the *reliable* model and $\mathbf{\Phi}$ represents the parameters corresponding to the *unreliable* model. We shall carry forward this notation for models throughout the rest of the paper.

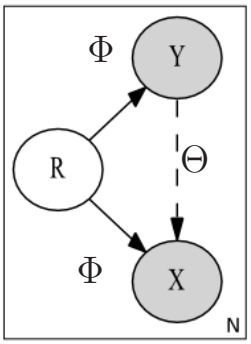

Figure 1: Bayesian model proposed in [16] for reliability-aware fusion of feature sets for object classification. The dashed line between features X and Y indicate, that the dependence between them is contingent on the reliability R. Latent variables are denoted by unshaded nodes.

$$Pr(X, Y|R) = Pr(X, Y; \mathbf{\Theta})^R (Pr(Y; \mathbf{\Phi}) Pr(X; \mathbf{\Phi}))^{1-R} \tag{1}$$

Generative Form

$$Pr(X|Y, R) = Pr(X|Y; \mathbf{\Theta})^R Pr(X; \mathbf{\Phi})^{1-R} \tag{2}$$

In [16], we used this reliability model in a generative form as shown in Eq.(2) to distinguish between unreliable and reliable features for an object classification task in underwater sonar images. The proposed Bayesian model (in Fig.1) was used to compare average reliabilities for features on different datasets.

While this model provides a good insight into reliability of different features, it suffers from the issue of data-sparsity, common to most generative models. A generative model must parametrically describe the distributions $Pr(X|Y; \mathbf{\Theta})$ and $Pr(X; \mathbf{\Phi})$ to generate features. As a result, parameter estimation becomes quickly infeasible with increase in number of dimensions because of lack of data samples. In comparison, discriminative models only learn parameters for conditional distributions of labels given features and hence are more efficient in using data during training. In the next section, we show how the notion of reliability introduced here can be extended to a discriminative model.

3. Discriminative Model for reliability

We modify our reliability assumption in the context of discriminative models to describe the reliability of annotated labels for each sample instead of their features as shown in Eq.(3)

Discriminative Form

$$Pr(Y, R|X) = \underbrace{Pr(Y|X; \mathbf{\Theta})^R}_{\text{reliable}} \underbrace{Pr(Y; \mathbf{\Phi})^{1-R}}_{\text{unreliable}} Pr(R|X) \tag{3}$$

The reliable model Θ now represents the chance that the annotator was careful in assigning the class label Y after examining the data sample with features X. The unreliable model Φ denotes the chance that, the annotator simply assigned the label at random. R still represents the binary random variable corresponding to reliability and is used to select either the reliable or the unreliable model to generate labels. This formulation allows us to write the complete data conditional likelihood in terms of the parameters of the models $\mathbf{\Theta}$ and $\mathbf{\Phi}$ which can be estimated by maximizing the data conditional log-likelihood. The sample reliability R is a latent variable in this model and hence we use Expectation-Maximization (EM) to optimize the data log-likelihood. This is described in detail next.

3.1. Formulation and Notation

We use the 1-of-K encoding to denote class labels in this paper, where K is the number of classes in the classification task. Accordingly, the class label Y_i for the i^{th} sample is represented using K binary values where $\sum_{k=1}^{K} Y_{ik} = 1$ and $Y_{ik} = 1$ indicates that the sample belongs to the k^{th} class. The corresponding D-dimensional feature vector for each of the N samples are denoted by $X_1, X_2, \ldots, X_N$.

As stated earlier, we assume that if a sample is reliable ($R_i = 1$) for the classification task, the observed class label Y_i for the sample is assumed to be dependent on its features X_i, according to the reliable model $\mathbf{\Theta}$. In this work, we assume $\mathbf{\Theta}$ to be a MaxEnt model parametrized by $\mathbf{W} \in \mathbb{R}^{D \times K}$ where $\{\mathbf{W}_1, \ldots, \mathbf{W}_K\}$ are the weight vectors for each class. According to the MaxEnt model, the probability of i^{th} sample belonging to class k given features X_i is given by the normalized exponential or softmax function as shown in Eq.(4) which we shall denote by Ψ_{ik}. In practice, any model that can be trained using soft labels can be used instead of the MaxEnt model, but we choose the MaxEnt model in this work because of its ease of training and the inherently probabilistic formulation. MaxEnt models are also otherwise known as multinomial logistic regression and popular in fields such as natural language processing [22].

When a sample is unreliable ($R_i = 0$) for the classification task, we assume its class label Y_i to have been generated at random by rolling a K faced die with a probability $\eta_k, k = \{1, \cdots, K\}$ for each face. This marginal distribution for Y acts as the garbage model $\mathbf{\Phi}$ as shown in Eq.(5). In addition, we assume that the reliability of a sample's label can be modeled by its location in the feature space. This assumption can be easily understood, by considering the fact that an annotator's reliability might depend on certain data-specific characteristics. Thus, an annotator's efficiency might be variable over different regions of the feature space. We capture this data-dependent relation for reliability by assuming that reliability $R_i = \{0, 1\}$ can be modeled by a logistic regression on the features X_i parametrized by $\mathbf{r} \in \mathbb{R}^D$ as shown in Eq.(6), where $\sigma(.)$ denotes the sigmoid function.

$$Pr(Y_{ik} = 1|X_i; \mathbf{\Theta}) = \frac{e^{\mathbf{w}_k{}^T X_i}}{\sum_{j=1}^{K} e^{\mathbf{w}_j{}^T X_i}} = \Psi_{ik}(\mathbf{W}) \quad (4)$$

$$Pr(Y_{ik} = 1; \mathbf{\Phi}) = \eta_k \quad (5)$$

$$Pr(R_i = 1|X_i) = \sigma(\mathbf{r}^T X_i) = \rho_i(\mathbf{r}) \quad (6)$$

4. ML parameter estimation

It can be shown that the model proposed above is a mixture-of-experts model. Maximum-likelihood (ML) parameter estimation for this model is well studied [9] and can be easily done using the EM algorithm. Since the proposed model is discriminative we try to maximize the conditional total data log-likelihood. Assuming that each sample is independently drawn from the distribution, we can factorize the total conditional likelihood of data $\mathcal{D} = \{X, Y, R\}$ given parameters $(\mathbf{W}, \eta, \mathbf{r})$ as shown in Eq. (7).

$$Pr[\mathcal{D}|\mathbf{W}, \eta, \mathbf{r}] = \prod_{i=1}^{N} Pr(Y_i, R_i|X_i; \mathbf{W}, \eta, \mathbf{r}) \quad (7)$$

Substituting from Eq. (3) leads to the following log-likelihood function using the notations shown earlier.

$$\begin{aligned}\mathcal{L} &= ln\ Pr[\mathcal{D}|\mathbf{W}, \eta, \mathbf{r}] \\ &= \sum_{i=1}^{N} R_i\ ln\ \rho_i + (1 - R_i)\ ln(1 - \rho_i) \\ &+ \sum_{i=1}^{N}\sum_{k=1}^{K} Y_{ik}\ [R_i\ ln\ \Psi_{ik} + (1 - R_i) ln\ \eta_k] \end{aligned} \quad (8)$$

To find the ML parameter estimates $(\mathbf{W}^*, \eta^*, \mathbf{r}^*)$ we maximize $\mathcal{L}$ in Eq. (8) using the EM algorithm as described next.

4.1. The EM algorithm

Note that R being a latent variable in this optimization problem, should not occur in the parameter estimation equations. The EM algorithm provides an efficient iterative way to perform ML estimation in such cases in the presence of missing data [23]. Instead of optimizing $\mathcal{L}$, the EM algorithm instead works with a lower bound approximation of $\mathcal{L}$ that is easier to optimize. This lower bound function is obtained by computing the expected value of $\mathcal{L}$ with respect to the posterior distribution of the latent variable R (for more details see [24]) . We shall represent this lower bound by $\mathcal{L}'$ in Eq. (9) where $\mathbb{E}_f[.]$ denotes the expectation operator with respect to distribution f.

$$\mathcal{L}' = \mathbb{E}_{f:Pr(R_i|X_i,Y_i)}\ [\mathcal{L}] \quad (9)$$

The EM algorithm then employs an iterative procedure in which we alternate between estimating the bound $\mathcal{L}'$ (Expectation step) and finding ML estimates of parameters by maximizing $\mathcal{L}'$ (Maximization step). Note that $\mathcal{L}'$ in Eq. (9) is no longer a function of the latent variable R thereby simplifying the optimization.

Expectation step
At any iteration, given the observation $\mathcal{D} = \{X, Y\}$ and parameters $\mathbf{W}^{old}, \eta^{old}, \mathbf{r}^{old}$ from the previous iteration, we first compute the expected reliability $\gamma_i = \mathbb{E}_f(R_i)$ for each sample with respect to the posterior distribution f for R_i as follows

$$\begin{aligned}\gamma_i &= Pr(R_i = 1|X_i, Y_i; \mathbf{W}^{old}, \eta^{old}, \mathbf{r}^{old}) \\ &= \frac{Pr(Y_i|R_i = 1, X_i)P(R_i = 1|X_i)}{Pr(Y_i, R_i = 1|X_i) + Pr(Y_i, R_i = 0|X_i)} \\ &= \frac{\rho_i \prod_{k=1}^{K} (\Psi_{ik})^{Y_{ik}}}{\rho_i \prod_{k=1}^{K} (\Psi_{ik})^{Y_{ik}} + (1 - \rho_i) \prod_{k=1}^{K} \eta_k^{Y_{ik}}} \end{aligned} \quad (10)$$

To compute $\mathcal{L}'$ we can simply substitute R_i with γ_i in Eq. (8) since none of the other terms are functions of R_i.

$$\mathcal{L}' = \sum_{i=1}^{N} \gamma_i\ ln\ \rho_i + (1 - \gamma_i)\ ln(1 - \rho_i) \quad (11)$$

$$+ \sum_{i=1}^{N}\sum_{k=1}^{K} Y_{ik}\ [\gamma_i\ ln\ \Psi_{ik} + (1 - \gamma_i) ln\ \eta_k] \quad (12)$$

Maximization step
Once the expected log-likelihood function $\mathcal{L}'$ has been computed in the E-step, we use the current value of γ_i to estimate the optimal parameter values for the current iteration $\mathbf{W}^{new}, \mathbf{r}^{new}, \eta^{new}$ that maximize $\mathcal{L}'$.

Estimating η: Obtaining an analytical update equation for η_k is straightforward by using the method of Lagrange multiplier with the additional constraint $\sum_{k=1}^{K} \eta_k = 1$. Then, setting $\partial L'/\partial \eta = 0$ yields the following parameter update equation for the unreliable model parameter η_k for each class.

$$\eta_k^{new} = \frac{\sum_{i=1}^{N} Y_{ik}(1 - \gamma_i)}{\sum_{i=1}^{N} Y_{ik}} \quad (13)$$

Estimating W and r: Although the objective functions for $\mathbf{W}$ and $\mathbf{r}$ are convex it is not possible to obtain closed form update equations for these parameters. Hence, we find $\mathbf{W}^{new}, \mathbf{r}^{new}$ by gradient ascent on $\mathcal{L}'$.

$$\mathcal{H}_r = \sum_{i=1}^{N} \gamma_i ln\ \sigma(\mathbf{r}^T X_i) + (1 - \gamma_i) ln\ [1 - \sigma(\mathbf{r}^T X_i)] \quad (14)$$

$$\mathcal{H}_W = \sum_{i=1}^{N} \gamma_i \sum_{k=1}^{K} Y_{ik} ln\ \Psi_{ik}$$

From Eqns. 14 we note that the objective function for parameter $\mathbf{W}$ corresponds to the multinomial logit cost function where each sample is weighted by γ_i while for parameter $\mathbf{r}$, the objective function corresponds to ordinary logistic regression using γ_i as soft labels instead. Hence, both the optimization problems are convex with a unique optimum. We maximize these

functions using the L-BFGS optimization algorithm [25, 26] implemented in the SciPy toolkit [27]. The L-BFGS algorithm is convenient because unlike Newton's method, it does not require direct computation of the Hessian matrix which can often turn out to be singular. Only required function computations are $\mathcal{H}_r, \mathcal{H}_W, \nabla_r \mathcal{H}_r, \nabla_W \mathcal{H}_W$.

At each iteration of the EM algorithm, we additionally compute the value of the log-likelihood function $\mathcal{L}'(\mathbf{W}^{new}, \mathbf{r}^{new}, \eta^{new})$ to check for convergence. The value of the log-likelihood function for the EM algorithm monotonically increases, and the algorithm is deemed to have converged when the change in value of $\mathcal{L}'$ is below a given threshold ϵ. Typically, the number of iterations required for convergence depends not only on the number of samples N but also on the number of classes K as it is tied to the number of parameters. Once the optimal parameter values $\mathbf{W}^*, \eta^*, \mathbf{r}^*$ have been learned we can infer the class posteriors for a new test sample Z as shown in Eq.(15)

$$Pr(Y_{ik} = 1|Z) = \Psi_k(\mathbf{W}^*, Z)\sigma(\mathbf{r}^{*T} Z) + \eta_k^* \left[1 - \sigma(\mathbf{r}^{*T} Z)\right] \tag{15}$$

5. Speech intelligibility classification experiment

We demonstrate the proposed algorithm on a speech-intelligibility classification experiment for pathological speech [21]. The term pathological speech here is used to refer to atypicalities in voice resulting from disease or surgery of the vocal tract. This often affects speech intelligibility and hence automatic assessment is of considerable interest. Moreover, the manual assessment of intelligibility being a subjective task can result in biased labels, making estimation of sample and label reliability an interesting yet challenging task.

For our experiments, we use the NKI CCRT Speech Corpus [28] consisting of 2385 sentence-level utterances from 55 speakers undergoing treatment for inoperable tumors of the head and neck. This dataset released at the Interspeech 2012 Speaker Trait Challenge [11] contains binary intelligibility annotations for each utterance where binary labels (I/NI) were created after thresholding EWE scores [29] obtained from multiple annotators. While the challenge focused mainly on obtaining a high accuracy of classification for intelligibility [21], the crowd-sourced and noisy aspect of annotations were not considered in any detail.

From each utterance, we extract features relating to the prosody, voice quality and pronunciation aspects of the speech signal. These features are popular in speech signal processing for studying paralinguistic information such as emotional state or for designing goodness of pronunciation measures. In [1] the authors studied in detail, these feature sets, in the context of pathological speech and selected a subset of 13 features (prosody: 6, pronunciation: 2, voice quality: 5) that were shown to improve the average classification accuracy on fusion. We use these selected feature sets for our experiments in this paper. Lack of intelligibility in pathological speech can depend on a variety of aspects of the speech signal where each factor contributes differently. This feature variability makes this domain particularly appealing for reliability-aware modeling. We perform experiments using 5-fold cross validation on the NKI CCRT dataset. For each data fold, we train the reliability-aware model using features and labels extracted from the remaining folds. This trained model is then used to infer class labels on the test fold. We compare our proposed approach against a logistic regression model which uses the same algorithm as the reliable model in our approach. However, the baseline algorithm is reliability-blind and assumes all samples to be equally reliable. The same feature sets are used to test both the algorithms. We perform separate experiments using each of the feature sets voice quality, prosody and pronunciation and also test a system using all features together. The results obtained on this dataset are shown in Table 1. We observe that reliability-aware modeling helps improve the classification accuracy on the feature sets pronunciation and voice quality. However, no improvement in classification accuracy is obtained for the prosody feature set, which also happens to be the most informative feature for predicting speech intelligibility.

Table 1: Results of reliability-aware binary classification of speech intelligibility for different feature sets. All accuracy figures are in %. The chance accuracy is 50.3%. Note that improvement in classification accuracy is obtained when the feature sets are more unreliable (highlighted).

Feature	Logistic	Proposed	$\rho_{\mathbf{avg}}$
voice quality	58.2	59.8	0.43
prosody	67.1	66.7	0.73
pronunciation	55.1	56.2	0.16
all	68.0	67.8	0.78

To better understand these results, we perform further analysis using average reliability scores as a diagnostic. The average reliability ρ_{avg} over the entire dataset can be computed by making a Bernoulli assumption on the latent variable R as $Pr(R_i = 1) = \rho_{avg}$. Simplifying the model enables us to analyze the reliability of each feature set in terms of an average reliability score ρ_{avg}. The results of ρ_{avg} obtained for each feature set are shown in Table 1 and seem to match with our hypothesis that the proposed reliability-aware classification method performs better when the feature set is inherently more unreliable (highlighted in Table 1). This might be the reason why our model is not able to extract any further performance by exploiting reliability information in experiments for the prosody feature set. We also note that for similar reasons feature fusion strategies with reliability-aware classification also do not help in improving the classification performance. This may be because we currently use a single reliability model $\mathbf{r}$ for all features together, whereas the experimental results clearly suggest that the degree of reliability of each feature set is different.

6. Conclusion

In this paper, we proposed a discriminative maximum entropy model to exploit sample reliability for classification in a dataset. Sample reliability is modeled as a latent factor R that dictates the dependence between features X and class labels Y. Our proposed approach is similar to the mixture-of-experts framework with two experts that learn reliable ($\mathbf{\Theta}$) and unreliable ($\mathbf{\Phi}$) models for predicting the class label given features. We propose an EM algorithm to jointly learn these models in a supervised fashion. In addition, we also learn a logistic regression model $\mathbf{r}$ to predict a soft reliability score using features.

Our experiments on the speech intelligibility classification task suggest that reliability-aware classification helps more when the feature set is noisy i.e. when more unreliable samples are present. In the future, we would like to jointly model labels from multiple annotators. We would also like to fuse reliability-aware models trained on different feature sets to be able to deal with their different reliability characteristics.

7. References

[1] J. Kim, N. Kumar, A. Tsiartas, M. Li, and S. S. Narayanan, "Automatic intelligibility classification of sentence-level pathological speech," *Computer speech & language*, vol. 29, no. 1, pp. 132–144, 2015.

[2] J. Schmidhuber, "Deep learning in neural networks: An overview," *arXiv preprint arXiv:1404.7828*, 2014.

[3] G. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[4] R. E. Schapire, "The strength of weak learnability," *Machine learning*, vol. 5, no. 2, pp. 197–227, 1990.

[5] H. Drucker, C. Cortes, L. D. Jackel, Y. LeCun, and V. Vapnik, "Boosting and other ensemble methods," *Neural Computation*, vol. 6, no. 6, pp. 1289–1301, 1994.

[6] T. G. Dietterich, "Ensemble methods in machine learning," in *Multiple classifier systems*. Springer, 2000, pp. 1–15.

[7] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[8] P. O. Gislason, J. A. Benediktsson, and J. R. Sveinsson, "Random forests for land cover classification," *Pattern Recognition Letters*, vol. 27, no. 4, pp. 294–300, 2006.

[9] M. I. Jordan and R. A. Jacobs, "Hierarchical mixtures of experts and the em algorithm," *Neural computation*, vol. 6, no. 2, pp. 181–214, 1994.

[10] D. J. Miller and H. S. Uyar, "A mixture of experts classifier with learning based on both labelled and unlabelled data," in *Advances in neural information processing systems*, 1997, pp. 571–577.

[11] B. Schuller, S. Steidl, A. Batliner, E. Nöth, A. Vinciarelli, F. Burkhardt, R. Van Son, F. Weninger, F. Eyben, T. Bocklet *et al.*, "The interspeech 2012 speaker trait challenge." in *INTERSPEECH*, 2012.

[12] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[13] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[14] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification*. John Wiley & Sons, 2012.

[15] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[16] N. Kumar, U. Mitra, and S. Narayanan, "Robust object classification in underwater sidescan sonar images by using reliability aware fusion of shadow features," in *IEEE Journal of Oceanic Engineering*, 2014.

[17] K. Audhkhasi and S. S. Narayanan, "Data-dependent evaluator modeling and its application to emotional valence classification from speech." in *INTERSPEECH*, 2010, pp. 2366–2369.

[18] A. P. Dawid and A. M. Skene, "Maximum likelihood estimation of observer error-rates using the em algorithm," *Applied statistics, JSTOR*, pp. 20–28, 1979.

[19] V. C. Raykar, S. Yu, L. H. Zhao, A. Jerebko, C. Florin, G. H. Valadez, L. Bogoni, and L. Moy, "Supervised learning from multiple experts: whom to trust when everyone lies a bit," in *Proceedings of the 26th Annual international conference on machine learning*. ACM, 2009, pp. 889–896.

[20] K. Audhkhasi and S. Narayanan, "A globally-variant locally-constant model for fusion of labels from multiple diverse experts without using reference labels," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 4, pp. 769–783, 2013.

[21] J. Kim, N. Kumar, A. Tsiartas, M. Li, and S. Narayanan, "Intelligibility classification of pathological speech using fusion of multiple subsystems," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[22] A. Ratnaparkhi, "A simple introduction to maximum entropy models for natural language processing," *IRCS Technical Reports Series*, p. 81, 1997.

[23] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 1–38, 1977.

[24] C. M. Bishop, *Pattern recognition and Machine learning*. Springer New York, 2006.

[25] D. C. Liu and J. Nocedal, "On the limited memory bfgs method for large scale optimization," *Mathematical programming*, vol. 45, no. 1-3, pp. 503–528, 1989.

[26] C. Zhu, R. H. Byrd, P. Lu, and J. Nocedal, "Algorithm 778: L-bfgs-b: Fortran subroutines for large-scale bound-constrained optimization," *ACM Transactions on Mathematical Software (TOMS)*, vol. 23, no. 4, pp. 550–560, 1997.

[27] E. Jones, T. Oliphant, and P. Peterson, "Scipy: Open source scientific tools for python," *http://www. scipy. org/*, 2001.

[28] R. P. Clapham, L. van der Molen, R. van Son, M. van den Brekel, and F. J. Hilgers, "NKI-CCRT corpus: Speech intelligibility before and after advanced head and neck cancer treated with concomitant chemoradiotherapy." European Language Resources Association (ELRA), 2012.

[29] M. Grimm and K. Kroschel, "Evaluation of natural emotions using self assessment manikins," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2005, pp. 381–385.