AUTOMATIC VOICE DISORDER DETECTION FROM A PRACTICAL PERSPECTIVE *

Jazmin Vidal[1], **Dayana Ribas**[2], **Cyntia Bonomi**[1], **Eduardo Lleida**[2], **Luciana Ferrer** [1], **Alfonso Ortega**[2]
[1] Instituto de Investigación de Ciencias de la Computación (ICC), Universidad de Buenos Aires-CONICET, Argentina
[2] ViVoLab, Aragón Institute for Engineering Research (I3A), University of Zaragoza, Spain
jvidal,gbonomi,lferrer@dc.uba.ar
dribas,ortega,lleida@unizar.es

ABSTRACT

Voice disorders, such as dysphonia, are common among the general population. These pathologies often remain untreated until they reach a high level of severity. Assisting the detection of voice disorders could facilitate early diagnosis and subsequent treatment. In this study, we address the practical aspects of Automatic Voice Disorders Detection (AVDD). In real world scenarios, data annotated for voice disorders is usually scarce due to various challenges involved in the collection and annotation of such data. However, some relatively large datasets are available for a reduced number of domains. In this context, we propose the use of a combination of out-of-domain and in-domain data for training a deep neural network-based AVDD system, and offer guidance on the minimum amount of in-domain data required to achieve acceptable performance.

Further, we propose the use of a cost-based metric, the normalized expected cost, to evaluate performance of AVDD systems in a way that closely reflects the needs of the application. As an added benefit, optimal decisions for the expected cost can be made in a principled way given by Bayes decision theory.

Finally, we argue that for medical applications like AVDD, the categorical decisions need to be accompanied by interpretable scores that reflect the confidence of the system. Even very accurate models often produce scores that are not suited for interpretation. Here, we show that such models can be easily improved by adding a calibration stage trained with just a few minutes of in-domain data. The outputs of the resulting calibrated system can then better support practitioners in their decision-making process.

***K*eywords** Automatic Voice Disorder Detection · Calibration · Self-supervised models · proper scoring rules · health applications

1 Introduction

Currently, voice disorders exhibit a relatively large prevalence within the general population. Prior studies [1, 2, 3] have documented a significant impact of voice problems on individuals, particularly for those whose livelihoods depend on their voice, such as teachers, telemarketers, TV presenters, and singers. Various factors contribute to individuals with voice problems delaying their visit to a doctor, allowing the condition to worsen over time. The emergence of remote health services presents an opportunity to leverage intelligent solutions in assisting doctors with the remote screening of patients. This advancement has the potential to facilitate early diagnosis and continuous monitoring of patients, reducing the need for hospital visits and alleviating the burden on the healthcare system.

Automatic systems for Voice Disorders Detection (AVDD) serve as the foundation for these remote services [4, 5]. AVDD systems can be implemented as machine learning models trained to discern pathological from healthy speech.

[*] *Cite as*: **Jazmin Vidal, Dayana Ribas, Cyntia Bonomi, Eduardo Lleida, Luciana Ferrer, Alfonso Ortega, Automatic Voice Disorder Detection from a Practical Perspective, Journal of Voice, 2024,ISSN 0892-1997, https://doi.org/10.1016/j.jvoice.2024.03.001. (https://www.sciencedirect.com/science/article/pii/S0892199724000729)**

A wide range of classification models, from statistical approaches to Deep Neural Networks (DNN) based models, have been proposed for this task [6, 7, 8, 9, 10, 5, 11, 12, 13, 14]. For instance, Kadiri et al. [15] explored suitable representations for this task, spanning spectral and cepstral features, voice quality, perturbation measures, and complexity measures. Colnago Contreras et al. [14] compared eight cepstral coefficient extraction techniques and evaluated their individual and combined effectiveness in representing speech signals against established classification methods to detect dysphonia in speech signals. Gupta et al. [16] introduced CNN-based approaches using Residual Network (ResNet) for categorizing the severity of dysarthric speech using short-duration speech segments of less than one second. More recently, Ribas et al. [17] assessed the effectiveness of using Self-Supervised (SS) representations for AVDD, exploring different combinations of upstream and downstream models to approach the task as a binary classification between healthy and pathological speech. Fonseca et al. [13], utilized signal energy (SE), zero-crossing rates (ZCRs), and signal entropy (SH) to construct a comprehensive time-frequency-information map that allowed them to tackled the classification of pathologies with similar phonic symptoms. A comprehensive review by Hegde et al. [18] describes an extensive list of AVDD approaches, detailing the methods employed for representation and classification, along with their corresponding performance.

Thus far, previous efforts have primarily concentrated on enhancing the accuracy of AVDD systems using the same publicly-available datasets for training. Systems trained on data from a given domain, consisting of specific acoustic conditions and phonetic content, often fail to generalize to other domains. Hence, the performance of AVDD systems on data from patients acquired in a hospital may be significantly worse than the performance estimated on the development data. While this problem can be solved by developing the system using data from the domain of interest, this is often infeasible or suboptimal due to the small size of such in-domain datasets.

Our previous study [17], explored the use of a large out-of-domain dataset to complement a smaller in-domain dataset for system development. The results demonstrated that incorporating this additional out-of-domain data significantly improves performance compared to using only the in-domain data. In this paper, we adopt this effective approach to train a deep neural network-based AVDD system, making use of both in-domain and out-of-domain data. This methodology allows us to draw meaningful and practical conclusions, providing valuable guidelines on the optimal amount of in-domain data necessary to achieve an acceptable performance level for the specific application. To simulate the target scenario, we use two publicly accessible datasets: the Saarbruecken Voice Database (SVD) [19] as the in-domain dataset and the Advanced Voice Function Assessment Database (AVFAD) [20] as the out-of-domain dataset. These corpus are noteworthy for being among the very few openly available databases that include speech samples from both patients with voice disorders and control speakers, as diagnosed by specialized otorhinolaryngology or speech pathologists experts.

Another crucial issue in practical application is the selection of relevant metrics for evaluating system performance. The accuracy, one of the most widely used metrics for AVDD, given by one minus the error rate, considers that all errors are equally severe. Yet, in the medical scenario, failing to detect that a person has a voice disorder should probably carry higher penalty than incorrectly diagnosing a healthy person, suggesting that accuracy is not an appropriate metric for this application. In this work, we propose to use a generalized version of the error rate called expected cost, where each type of error can be penalized with a different cost, appropriately selected for the task of interest.

In addition, in many applications, it is desirable for the system to provide not only categorical decisions but also scores that can be interpreted by the doctor as the posterior probability that the patient has a voice disorder given the provided speech sample. In those cases, systems should be assessed by metrics that reflect the quality of the posterior probabilities they generate, like the Cross-Entropy. Notably, we show that even the most discriminative systems do not produce good-quality posteriors making them unreliable for interpretation. This happens because the best performing systems in terms of discrimination are based on DNNs with a large number of parameters which makes them prone to overfitting the training data. The overfitting, in turn, results in the system producing overconfident posteriors. Fortunately, a simple solution exists for this scenario consisting of a calibration transformation of the scores added as a final stage to the system. Importantly, we show that this calibration transformation can be successfully trained with just a few minutes of in-domain data. Specifically, results show that with 30 to 40 minutes of in-domain data used to train the DNN classifier and the calibration stage, the AVDD system can provide close to optimal discrimination as well as interpretable scores.

The next section describes our proposed approaches for the development of AVDD systems from a practical perspective. Then, we describe the experimental design and results, highlighting the impact of the proposed approaches. Finally, we conclude with a discussion.

2 Assessment of AVDD Performance from a Practical Perspective

In this section we describe the aspects we believe to be most important for the development of AVDD systems from a practical perspective and describe our proposed approaches. As an overview, Figure 1 shows the various steps involved in the development and deployment of an AVDD system.

The "data selection" stage in this Figure depicts the first step of system development, where the data for training and evaluation is selected. In order for the estimated performance to reflect the one that will be observed in practice once the system is deployed, the evaluation needs to be done using in-domain data, i.e., data that perfectly reflects the conditions of the practical scenario for which the system will be used. For training, though, both in-domain and out-of-domain data can be used. The next step in the process, the "AVDD system" stage, is the development of the AVDD system itself, which takes an audio sample as input and generates a score that indicates the probability that the speaker in the audio has a voice disorder. While extremely important, the study of models for AVDD is covered in many prior publications as mentioned in the introduction, and is out of the scope of this paper. For the experiments, we use a previously proposed architecture [17].

Another essential aspect in development process is the assessment of the performance of the system that takes place in the "evaluation" stage. Given the evaluation dataset, performance metrics that reflect the value that the end-user will obtain from the system can be computed. Here, we assume that the end-user requires not only categorical decisions, but also a score that reflects the posterior probability of the patient having a voice disorder, given the input sample. To assess the quality of categorical decisions we propose to use the normalized expected cost (NEC), with costs set to penalize the false negatives (missed detections) more severely than the false positives (false alarms). Categorical decisions are made by thresholding the scores using the threshold that optimizes the metric of choice. For the NEC (and for accuracy), the optimal threshold can be determined theoretically using Bayes decision theory, as will be explained later in this section. To assess the quality of the scores, before decision-making, we use the cross-entropy, which provides an overall measure of the goodness of the scores as posterior probabilities.

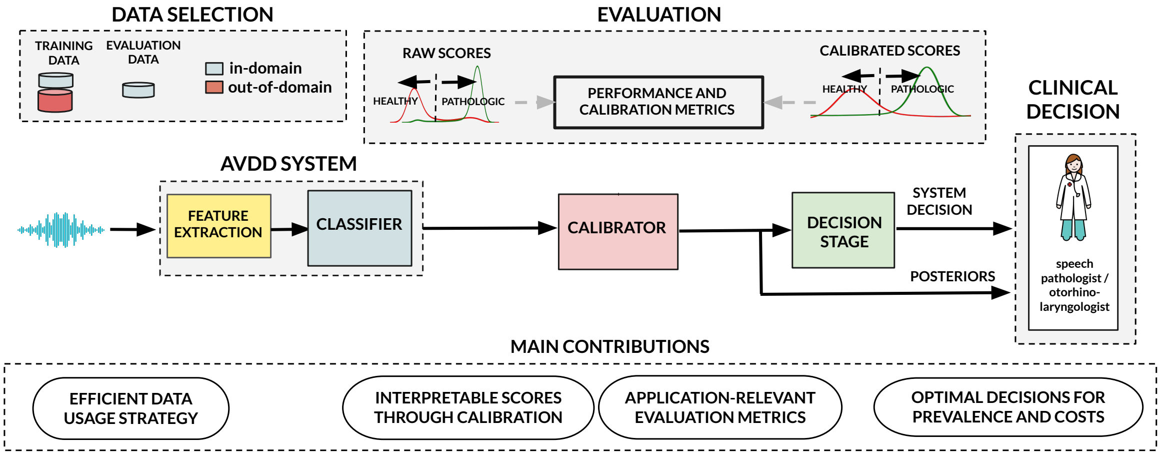

Figure 1: General scheme of an AVDD system and the development process, with a focus on the stages that have the most impact in a practical scenario: data selection, evaluation, calibration and decision-making.

At the "evaluation stage", calibration metrics may also be computed to assess whether a calibration stage needs to be added to the system to improve the quality of the posteriors which are provided to the end-user. For this, we propose to use the calibration loss metric, which is described in Section 2.4. If the calibration loss is large, then a calibration stage should be added to the system. The scores generated by the calibrator should be re-evaluated using the same metrics as before. Finally, the calibrated scores can be used to make categorical decisions and also provided directly to the end-user for interpretation in the "clinical decision" stage. The following sections explain the different stages in Figure 1 in detail.

2.1 Training and evaluation data

As in most machine learning applications, AVDD systems need to be trained on data that are representative of the target domain. For example, models should ideally be trained on the same demographic, acoustic conditions, language, and types of voice disorders for which they will later be used. The collection of such in-domain data for each new application is a slow and costly process since it requires the recruitment of subjects, recording of their speech, and careful labelling from a medical evaluation of the health status of the voice. For this reason, most AVDD datasets collected in the same scenario where the system would be deployed are relatively small.

A few relatively large datasets on voice disorders are publicly available. Such datasets are collected under a limited variety of recording conditions and do not necessarily include the full range of existing voice disorders. Yet, despite this issues, leveraging these larger – though potentially mismatched – datasets for system training can lead to improved performance on the target domain.

In this work, we simulate a realistic scenario where a large out-of-domain dataset and a much smaller in-domain dataset are available for development. We use the setup from our previous work [17] where the datasets are split as follows:

- For training, we use a large out-of-domain database with characteristics that are significantly different to those of the target domain plus a small in-domain dataset that is perfectly matched to the target domain.
- For evaluation, we use a small in-domain dataset from the target scenario.

We highlight the importance of evaluating on in-domain data. The goal of evaluation is to predict the performance that the system will have when deployed. If the evaluation data does not match exactly the conditions of the target domain, the estimated performance may not be a good predictor of the system performance, as will be perceived by the end-user on the data of interest.

2.2 Metrics for assessing the quality of categorical decisions

When making categorical decisions in binary classification, two types of errors exist: false positives also known as false alarms, samples labelled by the system as having a voice disorder but which were produced by healthy individuals; and false negative or misses, samples labelled as healthy but which were produced by a patient with a voice disorder. Given a certain evaluation dataset, a number of different classification metrics can be computed as a function of the number of false alarms and misses, and the number of samples of each class.

The most commonly used metrics in the AVDD literature are accuracy, unweighted average recall (UAR), among others [21, 15, 11, 22]. These metrics rely on error rates computed from the confusion matrix values given by the true and false positives and negatives (TP, FP, TN, FN). Despite their widespread use, these metrics do not necessarily reflect the needs of the application, since they do not take into account the cost that each type of error has in practice.

To address this problem, recent works on medical applications have proposed the use of metrics that consider the impact of each type of error on the use-case scenario. These metrics include the Net Benefit [23], its sibling, the expected cost (EC) [24, 25, 26], and Weighted Utility, a generalization of Net Benefit [27]. These metrics allow for the explicit specification of the cost (or utility) that each decision will have in the clinical setting given the true label of the sample. For example, in the case of AVDD, it is reasonable to assume that false alarms should be considered as less costly than misses. In the former case, a healthy patient may undergo unnecessary further testing or even treatment. In the latter case, the disorder would go untreated and may eventually develop into a more severe, perhaps incurable, pathology.

Given an evaluation dataset with a total of N samples, H samples from healthy subjects (true negatives), P samples from patients with a voice pathology (true positives), FP false positives and FN false negatives, the EC is given by:

$$\mathrm{EC} \quad = \quad \mathrm{C_{FN}}\,\mathrm{P_P}\,\mathrm{R_{FN}} + \mathrm{C_{FP}}\,\mathrm{P_H}\,\mathrm{R_{FP}}, \tag{1}$$

where $\mathrm{R_{FP}} = \mathrm{FP}/\mathrm{H}$ is the false positive rate, namely the conditional probability of deciding that a sample belongs to a patient with a voice disorder given that the patient was healthy; $\mathrm{R_{FN}} = \mathrm{FN}/\mathrm{H}$ is the false negative rate, i.e., conditional probability of deciding that the subject is healthy given that she/he presents a voice disorder; $\mathrm{P_P} = \mathrm{P}/\mathrm{N}$ is the proportion of samples that belongs to subjects with a voice disorder; $\mathrm{P_H} = \mathrm{H}/\mathrm{N}$ is the proportion of samples that belongs to healthy subjects. The parameters $\mathrm{C_{FP}}$ and $\mathrm{C_{FN}}$ determine how costly each type of error is deemed for the application of interest.

In this work we will use a normalized version of the EC, given by

$$\mathrm{NEC} \quad = \quad \frac{\mathrm{EC}}{\min(\mathrm{C_{FN}}\,\mathrm{P_D}, \mathrm{C_{FP}}\,\mathrm{P_H})}, \tag{2}$$

The denominator corresponds to the EC of the best naive system, one that does not have access to the input sample and always makes the same decision, the one that minimizes the EC. For an in-depth description of the EC and the NEC, we refer the reader to [26].

Note that accuracy and UAR are tightly related to special cases of the EC. Accuracy is obtained as one minus the EC when $\mathrm{C_{FP}} = \mathrm{C_{FN}} = 1$, while UAR is obtained as one minus the EC with $\mathrm{C_{FP}} = 1/\mathrm{P_H}$ and $\mathrm{C_{FN}} = 1/\mathrm{P_D}$. Note that

UAR implicitly assigns a cost to false alarms and misses that may not be appropriate for the task, depending on the values of $\mathrm{P_D}$ and $\mathrm{P_H}$. If, as is common in AVDD datasets, $\mathrm{P_D} > \mathrm{P_H}$, then $\mathrm{C_{FN}} < \mathrm{C_{FP}}$, which implies an assumption that false negatives (misses) are less costly than false positives (false alarms), something that may not reflect the needs of the clinical setting. In this work, we propose to use the EC metric with a higher cost for false negatives than for false positives, setting $\mathrm{C_{FN}} = 3$, and $\mathrm{C_{FP}} = 1$.

An advantage of the EC and its special cases like the accuracy and the UAR is that, when the system produces well-calibrated posteriors (see Section 2.4 for a definition of calibration), optimal decisions can be obtained using Bayes decision theory [28]. For our AVDD binary classification problem, Bayes decisions correspond to comparing the posterior for the class “patient has a voice disorder” to the following threshold:

$$\mathrm{thr} = \frac{\mathrm{C_{FP}}}{\mathrm{C_{FP}} + \mathrm{C_{FN}}}, \tag{3}$$

and deciding that the patient has a voice disorder whenever that posterior is larger than the threshold. For accuracy, the Bayes threshold is given by 0.5, resulting in the usual decision rule of choosing the class with the largest score. For UAR, the Bayes threshold is given by $\mathrm{P_D}$, the prevalence of voice disorders in the test data which, in the SVD dataset is larger than 0.5. Hence, fewer samples will be labelled as having a voice disorder when optimizing UAR than accuracy, resulting in more missed diagnoses, as required by the metric which effectively gives a higher weight to false positives than to false negatives. On the other hand, for our selected EC with $\mathrm{C_{FN}} = 3$ and $\mathrm{C_{FP}} = 1$, the threshold is given by $1/4$, resulting in fewer false negatives, at the cost of more false positives. This, we believe, is a more appropriate operating point for the AVDD application.

All metrics discussed in this section are designed to assess the quality of categorical decisions for a specific decision threshold. In contrast, the Area Under the Receiver Operating Curve (AUROC or AUC), also commonly used in AVDD literature, integrates the performance across the full range of possible decision thresholds, reflecting the performance of the system without committing to a single decision threshold. While useful during the early stages of systems development, the AUC does not reflect the actual performance of the categorical decisions as will be perceived by the end-user and, hence, should not be used to assess the goodness of the system. In particular, a system may be better than another in terms of AUC but worse in terms of NEC. In that case, the system that is better in terms of NEC should be selected for deployment since that is the metric that reflects the needs of the application.

2.3 Metric for assessing the interpretability of scores

When interpretable scores are required, a metric that reflects the quality of the scores as posteriors is needed for evaluation. A principled way to assess the quality of posteriors is through proper scoring rules (PSRs) [29]. PSRs assign a numerical cost to a posterior distribution based on how well it aligns with the true label of the sample. The expected value of a PSR is minimized when the posterior under evaluation coincides with the reference distribution with respect to which the expectation is taken. This means that a lower PSR value indicates that the posterior distribution is closer to this reference distribution. The expected value of a PSR (EPSR) over the data then reflects the quality of the posteriors.

One widely used EPSR is the cross-entropy. The (empirical) cross-entropy is given by:

$$\mathrm{XE} = -\frac{1}{\mathrm{N}} \sum_{i=1}^{N} \log P(\mathrm{c_i}|\mathrm{x_i}) \tag{4}$$

where $P(\mathrm{c_i}|\mathrm{x_i})$ is the posterior probability produced by the system for input sample $\mathrm{x_i}$ for class $\mathrm{c_i}$, the true class of the sample. The summation is taken over the N samples in the evaluation dataset. The cross-entropy measures the average surprise or information content of the true labels given the posteriors produced by the system. Lower cross-entropy indicates a better match between posteriors and true labels.

In this work, we will use normalized XE (NXE) as the main metric for assessing whether the system's output are useful for interpretation. The normalization is done by diving the XE of the system by the XE of a naive system that does not have access to the input samples and always outputs the prior probability of each class. Such a system is, from all systems that do not have access to the input, the one that minimizes the XE. The NXE should always be lower than 1.0 since a larger value indicates that the system is worse than the best naive system. Assuming we know the class prior probabilities, a system with NXE larger than 1.0 can be trivially improved by replacing its output with that of the best naive system. Hence, a NXE larger or close to 1.0 implies that the posteriors are poor and should be fixed. Even when NXE is lower than 1.0, the posteriors may sometimes be suboptimal, meaning that they can be improved by a post-hoc transformation, as described in the next section.

2.4 Assessing and fixing calibration

Classification systems designed to produce posterior probabilities are usually trained with XE as objective. Yet, despite being trained to minimize an objective that reflects the goodness of the scores as posteriors, these systems can sometimes end up producing extremely poor posteriors on the test data. This happens because the model overfits the training data leading to an unrealistically good XE value on that data, but a rather poor one on the test data. This is a common phenomenon in modern DNNs which are prone to overfitting due to their large number of parameters [30].

Fortunately, the effect of overfitting can usually be easily counteracted by adding a new stage in the system which is meant to *calibrate* the scores. Calibration refers to the process of transforming a set of scores into the best possible posteriors that can be obtained by this operation. The goal of doing calibration is to reduce the value of an EPSR, most commonly, the cross-entropy. If posteriors can be improved by calibration, then they are said to be miscalibrated. On the other hand, well-calibrated posteriors cannot be improved by adding a calibration stage to the system. A perfectly calibrated system produces optimal Bayes decision for the EC of choice, meaning that no other decision rule can result in better categorical decisions. For a discussion on calibration and many references on the topic, please see [26]. The specific calibration approach used in this work will be discussed in the next section as part of the system description.

Model calibration has been well-studied in epidemiology literature [31, 32] and more recently in medical imaging [33, 24]. However, to our knowledge, this topic has not been discussed in the context of voice pathology detection. Yet, AVDD systems are prone to overfitting and, hence, to miscalibration, particularly so because of the scarcity of data making the scores from these system suboptimal for interpretation unless a calibration stage is added to the system.

Diagnosing calibration problems in a system requires an explicit or implicit process of training a calibration transform and evaluating whether this transform improved the posteriors. Currently, the most standard calibration metric is the Expected Calibration Error (ECE) [34], a measure derived from reliability diagrams. These diagrams are constructed by binning the posteriors obtained on the evaluation dataset for one of the two classes into M uniform intervals. Then, the frequency of that class within the samples that fall in each bin is computed and plotted as the bar height in each bin (see Figure 8). These frequencies determine the calibrated posterior for all the samples within the corresponding bin. The ECE is then computed as the average absolute difference between average posterior within each bin B_m, called $\mathrm{avep}(B_m)$, and the corresponding calibrated posterior (the bin height), $\mathrm{frac}(B_m)$:

$$\mathrm{ECE} = \sum_{\mathrm{m}=1}^{\mathrm{M}} \frac{|\mathrm{B_m}|}{\mathrm{K}} |\mathrm{frac}(\mathrm{B_m}) - \mathrm{avep}(\mathrm{B_m})|. \tag{5}$$

For well-calibrated posteriors the reliability plot is close to the diagonal, and the ECE is close to 0. The ECE, while widely used in the literature, has a series of known problems [26, 35, 36]. For this reason, in this work, in addition to showing ECE for consistency with prior literature, we use the calibration loss (CalLoss). CalLoss is given by the difference between a *raw* EPSR, computed on the scores as they come out of the system, and a *minimum* EPSR, computed on the scores after calibration. In this work, we use NXE as the EPSR, and compute the relative CalLoss as:

$$\mathrm{RelCalLoss} = 100\,\frac{\mathrm{NXE_{raw}} - \mathrm{NXE_{min}}}{\mathrm{NXE_{raw}}}. \tag{6}$$

The value of $\mathrm{NXE_{min}}$ measures the performance that the system can achieve by optimally transforming the scores into better posteriors, i.e., the best performance the system could achieve after calibration. This is sometimes called the refinement or discrimination performance since it reflects the inherent ability of the system to separate the classes from each other. For binary classification, $\mathrm{NXE_{min}}$ can be obtained using the Pool Adjacent Violators (PAV) algorithm [37] which finds the monotonic transformation of the scores that minimizes an EPSR (nicely, optimizing any EPSR will lead to the same transform). The resulting non-parametric calibration transform gives us the best possible posteriors that can be obtained from the scores generated by the system on the evaluation dataset. These scores may be too optimistic since the transform is trained on the evaluation data itself, but it provides a lower bound for the EPSR. If the RelCalLoss is small, then we can be sure that the scores are well calibrated.

We would like to note that many works in recent machine learning literature use calibration metrics as a way to assess the interpretability of posteriors. We believe this is a misguided practice. The usefulness of the posteriors is assessed, by construction, by EPSRs. Calibration metrics, such as ECE or CalLoss, only measure part of the performance of the posteriors. In the extreme, a system can have perfect calibration but be completely uninformative. For instance, the best naive system described above, which always outputs the class priors, is a rather useless system since it does not carry any information about the input sample. Yet, its CalLoss and ECE are equal to 0. On the other hand, a system may have a CalLoss of 50% and still be better –more useful– than the best naive system if its NXE is lower than 1.0. In summary, while calibration metrics are useful for diagnosing potential calibration problems, they should not be used as an indication of the quality of the posteriors.

3 Experimental Design

In this section we describe the DNN-based system and the calibration approach used in the experiments, as well as the two databases used for training and testing.

3.1 Classification model

The classifier used in our experiments is based on the prior work [17], where we showed that using self-supervised (SS) representations for AVDD results in significant improvements over prior approaches where DNNs are trained from scratch on AVDD datasets. We use the HuBERT model [38] which generates 768-dimensional representations at a rate of 50 vectors per second. HuBERT is a transformer-based model trained with a self-supervised approach to predict a target clean speech signal from a noisy masked version of that signal. These representations have been used for various tasks including emotion recognition, pronunciation assessment or speech enhancement [39, 40, 41]. The HuBERT model is taken as a feature extractor and fine-tuned during training. We use the base model[2], which was pretrained with 960 hours of speech from the Librispeech corpus.

The sequence generated by this HuBERT model is processed by a downstream model consisting of a transformer with a class-token which summarizes the sequence into a single utterance-level token [42] using multi-head self-attention mechanisms. A final output layer with softmax activation function provides the posterior probabilities for each of the two classes. This model is trained using cross-entropy loss at utterance level. Details on the architecture and the training process can be found in [17]. The classification models are built in PyTorch using the S3prl toolkit.[3]

3.2 Calibration model

For a given audio sample, the output of the downstream model are the posterior probabilities for each of the two classes. As explained above, it may be possible to improve these posteriors by applying a calibration transformation. In this work we use linear logistic regression as calibration transformation, which, for the binary classification case is also known as Platt scaling [43]. This transformation has the following expression:

$$\tilde{s}_P = \text{sigmoid}(\alpha \log(s_P) + \beta) \tag{7}$$

where s_P is the posterior for the class “patient has a voice disorder” output by the downstream model, and α and β are two scalar parameters. These parameters are trained to minimize the cross entropy between the transformed scores $\tilde{s}_P$ and the binary labels. The functional form of this equation can be shown to be optimal when the classes have Gaussian distribution with the same covariance [44]. In practice, the transformation is often found to be quite effective even when these conditions are not met. The python package psrcal[4] was used for training and applying the calibration transformation in the experiments.

3.3 Databases

In this experiments we use two publicly available datasets: the Advanced Voice Function Assessment Database (AVFAD) and the Saarbruecken Voice Dataset (SVD).

AVFAD [20] is an open-access dataset in the Portuguese language with approximately 32 hours of speech recordings. It contains 363 recordings labeled as healthy by 250 females and 113 males, and 346 recordings labeled as pathological by 249 females and 97 males. There are 26 different labelled pathologies. Recordings contain sustained vowels and phrases. In this paper we used the phrases chunked into 4-second segments. The audio was originally recorded in 48 KHz and, for our experiments, was downsampled to 16 KHz. This corpus was used as the out-of-domain dataset and used only for training purposes.

SVD [19] is an open-access database in the German language with more than 2000 speakers recording phrases and sustained vowels. After keeping the phrases only, we are left with 1988 sessions from 1627 speakers adding up to a total of 72 minutes. There are 634 recordings labeled as healthy by 382 females and 252 males, and 1354 recordings labeled as pathological by 726 females and 628 males. Pathological samples may be labelled as having one or more pathologies. There are also available annotations of the GRBAS scales (grade, rough, breathy, asthenic, and strained) [45] that provides a perceptual appraisal of voice quality. For further analysis of results we use the Grade rating of overall dysphonia level as an indicator of the voice disorder severity. The database includes 71 pathologies with an imbalanced distribution. As for AVFAD, we only use the phrases from this dataset. In this case, though, the phrases are already

[2] `https://dl.fbaipublicfiles.com/hubert/hubert_base_ls960.pt`

[3] The code is available in `https://github.com/dayanavivolab/s3prl/tree/voicedisorder`

[4] `https://github.com/luferrer/psr-calibration`

short duration and, hence, are not chunked. This corpus is used both for training and testing using a cross-validation approach described below.

3.4 Data splitting for experiments

As explained above, the goal is to simulate a practical scenario where limited in-domain data is available for system development. We also assume that a relatively larger amount of out-of-domain data is available for the task. In this work, we take AVFAD as the large out-of-domain dataset and use it only for training purposes. We use SVD as the smaller in-domain dataset which needs to be wisely used for training the downstream model, the calibration transform, and for final performance evaluation. We explore the effect that varying the amount of in-domain data used for training the downstream model and the calibration transform has on performance. The goal of this exploration is to obtain a guideline on the data requirements. Given a new application, we might be faced with the need to collect some in-domain data. We want to answer how much data should we collect and how should we optimally use it for training the two modules in the system.

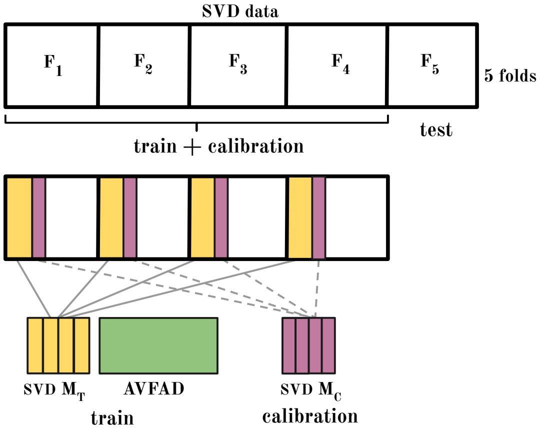

Figure 2: Data splitting strategy used for the experiments. The SVD data is split in 5 folds. Then, the following process is repeated for each fold $\mathrm{F_i}$: 1) a subset of size $\mathrm{D_T}$ (yellow blocks) from each of the folds that are not $\mathrm{F_i}$ is selected, 2) these subsets are pooled together to get $\mathrm{M_T} = 4\mathrm{D_T}$ minutes of data, merged with the AVFAD data and used to train the downstream model, 3) a subset of size $\mathrm{D_C}$ (violet blocks) is selected from each of the fold that are not $\mathrm{F_i}$ after excluding the data selected in step 1 above, 4) these subsets are pooled together to get $\mathrm{M_C}$ minutes of data, and the model generated in set 2 is used to generate scores for this data, 5) a calibration model is trained using these scores, 6) the downstream model from step 2 is used to generate scores on fold $\mathrm{F_i}$, 7) these scores are calibrated using the model obtained in step 5. Finally, the scores obtained in step 7 on each fold $\mathrm{F_i}$ are pooled together to compute the final metrics.

Figure 2 shows a diagram of the data splitting approach used in our experiments to address these questions. First, the SVD data is split into five folds with similar number of speakers with around 14 minutes and 400 speakers each. For each test fold, we select $\mathrm{M_T}/4$ minutes from each of the other four folds for training a downstream model. We set $\mathrm{M_T}$ to be 5, 10, 20, 30 or 40 minutes, corresponding to approximately 147, 292, 568, 838 and 1060 sessions, respectively. Then, to this selected data, we add the complete AVFAD set and use 90% of the resulting set to train the downstream DNN model using stochastic gradient descent, using the remaining 10% to select the best model across epochs. Finally, the selected model is used to generate scores on the test fold. This process is repeated for each fold and the resulting scores for every fold are concatenated. In the end, scores for the complete SVD dataset are available for computing performance metrics for each value of $\mathrm{M_T}$. We call these the uncalibrated or *raw* scores.

In a second set of experiments we explore the impact of adding a calibration stage after the classifier. The calibration transform is trained on a subset of the data from each fold that was not selected for training the downstream model. Since each fold contains around 14 minutes of speech and we use at most 10 of those for training the downstream model, we always have at least 4 minutes in each fold that are not used for training the downstream model. To explore the impact of the amount of calibration data in the performance of the posteriors, for each test fold, we select $\mathrm{M_C}/4$ minutes from each of the other four folds to train the calibration model. We set $\mathrm{M_C}$ to 2, 5, 10 and 15 minutes. The calibration model trained on this data is then applied to the test fold. This process is repeated for each fold and all resulting calibrated scores are pooled together to compute the performance metrics.

4 Results and discussion

This section presents the experimental results. First, we show the results on the raw scores, without a calibration stage, while increasing the amount $\mathrm{M_T}$ of in-domain data used to train the downstream model. As expected, the discrimination performance drastically improves as the amount of in-domain training data increases. On the other hand, the calibration performance is poor for all cases. Finally, we show how calibration can be easily fixed using a relatively small amount of calibration data, $\mathrm{M_C}$.

4.1 Discrimination performance of raw scores

Figure 3 shows two metrics that reflect the system's discrimination performance: the AUC and the minimum cross-entropy, $\mathrm{NXE_{min}}$ (see Section 2.4). These metrics are immune to calibration problems, assessing only the amount of information that a system provides about the class of a sample regardless of whether the scores are good or bad posterior probabilities. We can see that the discrimination performance drastically improves as the amount of in-domain training data increases. In particular, the $\mathrm{NXE_{min}}$ is close to 1.0 when only out-of-domain data is used for training ($\mathrm{M_T} = 0$). This value indicates that the system is close to a naive system that does not know anything about the input samples. As $\mathrm{M_T}$ increases, $\mathrm{NXE_{min}}$ reaches values under 0.5, indicating that the system contains useful information about the input samples. The value of AUC follows a similar trend, starting at around 0.65 when only out-of-domain data is used for training and reaching values over 0.90 when 40 minutes are used.

The largest value of $\mathrm{M_T}$, 40 minutes, contains speech from 1060 sessions. Obtaining an in-domain dataset with such a large number of speakers may be infeasible in many practical applications. Fortunately, even with a quarter of that data, 10 minutes of speech from 292 sessions, the performance is still significantly better than that of a random system and of a system trained only with out-of-domain data. We hypothesize that perhaps the number of speakers could be reduced further if each speaker provided a larger amount of speech, maintaining the total amount of speech at around 10 minutes. This hypothesis cannot be tested with the SVD dataset since each speaker recorded only one phrase. We will address this question in later work using a new dataset that is currently being collected.

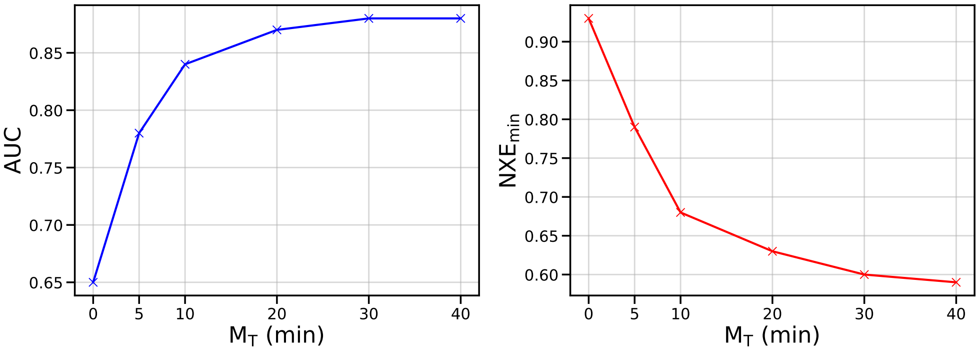

Figure 3: Discrimination metrics, AUC (higher is better) and $\mathrm{NXE_{min}}$ (lower is better), as a function of the total amount of data, $\mathrm{M_T}$, used to train and select the best epoch of the downstream model. No calibration stage is used for these results.

4.2 Calibration performance of raw scores

Figure 4 shows the calibration performance measured by ECE and RelCalLoss (see Section 2.4) as a function of the amount of training data for the downstream model. We can see that, while calibration performance improves as more in-domain data is used for training, the best system, with $\mathrm{M_T} = 40$, still has a high calibration error, with ECE values over 10% and RelCalLoss over 60%. That is, over 60% of the NXE of the system is due to miscalibration, even when $\mathrm{M_T}$ is relatively large. Notably, the trends indicate that increasing the amount of training data over 40 minutes would not result in significant further improvements in calibration. The downstream model does not produce well-calibrated posteriors, regardless of the amount of in-domain data used for training. This is likely due to the model overfitting the training data. Fortunately, this problem can be fixed by adding a simple post-hoc calibration stage to the system, as discussed in the next section.

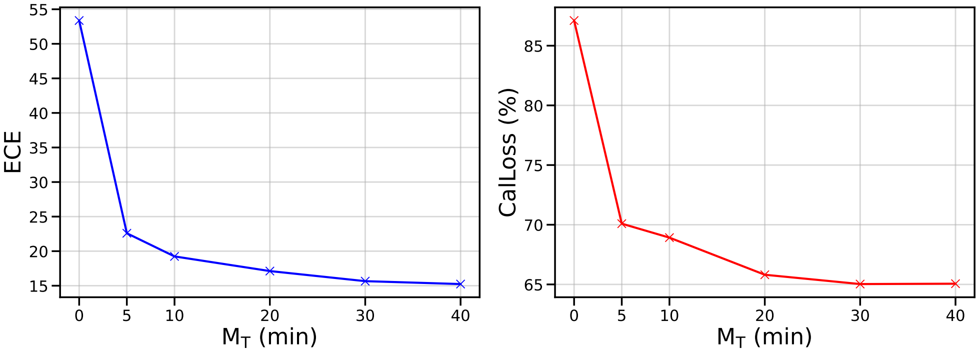

Figure 4: Calibration metrics, ECE and RelCalLoss, as a function of M_T, as in Figure 3. Also as in that figure, no calibration stage is used for these results.

4.3 Overall performance of raw and calibrated scores

Figure 5 shows two metrics that reflect the performance of categorical decisions at two different operating points: a NEC, using $C_{FN}/C_{FP} = 3$, and the standard UAR, which, for our data, corresponds to one minus the EC with $C_{FN}/C_{FP} = 0.47$. For both metrics, the decision threshold is the one determined by Bayes decisions. Bayes decisions improve as the scores are better calibrated, i.e., better posterior probabilities. We also show the NXE, which reflects the performance of the posteriors without committing to a specific operating point. Hence, the NXE reflects the overall goodness of the scores as posterior probabilities, indicating how useful these scores would be for a human attempting to interpret them as such.

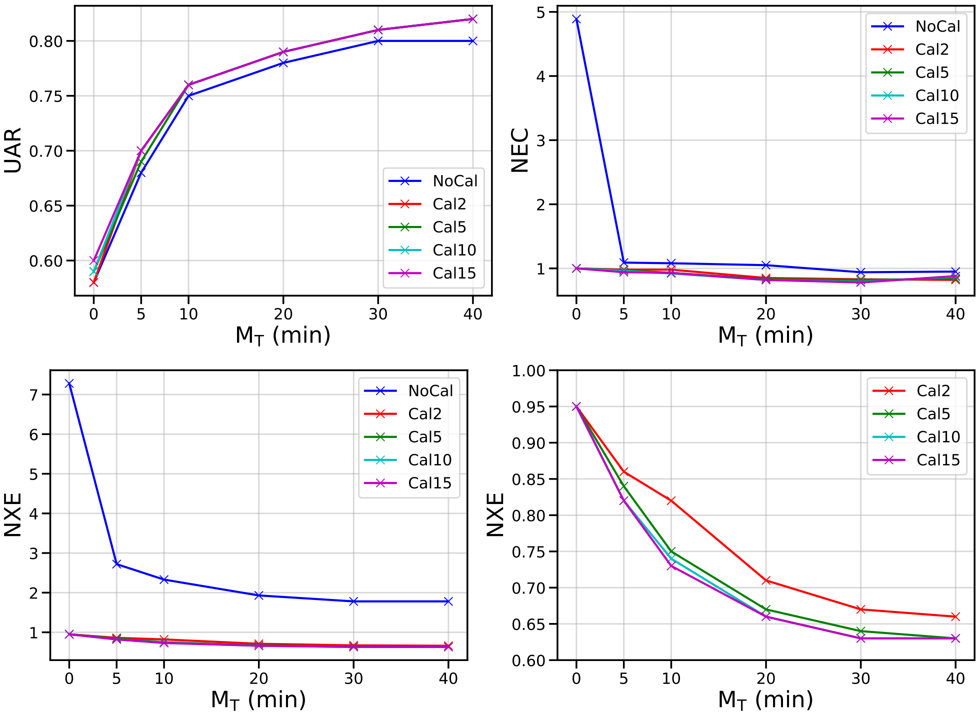

Figure 5: Overall performance metrics, UAR, NEC and NXE as a function of M_T in minutes, as in Figure 3. The two bottom plots show a zoomed in version of the top ones for NXE and NEC, without the NoCal curve.

The figure shows the performance as a function of M_T, indicated in the x-axis, and M_C, shown as different curves (see legend). We can see that the UAR and the NEC sharply improve (for UAR higher is better, for NEC lower is better) as M_T increases, in line with the observations in Section 4.1. On the other hand, except when $M_T = 0$ for NEC, the post-hoc calibration stage has very little effect on these metrics.

Turning to the NXE, we can see that the uncalibrated scores (blue lines) are very poor in terms of this metric. In particular, the NXE is larger than 1.0 for all values of M_T indicating that the scores are worse than those of a naive system and, hence, not useful for interpretation. Focusing on the zoomed in plot for the NXE on the bottom right we can see that, when calibration is performed, the NXE improves to values below 1.0, except for the model trained only on AVFAD data (i.e., when $\mathrm{M}_\mathrm{T} = 0$) for which discrimination performance is very poor, almost near 1.0. Further, we can see that the performance tends to improve as M_C increases, though only by a small margin. Hence, we can conclude that a small amount of data is sufficient for training the calibration model.

Table 1 presents an overview of the system's performance in terms of NEC, accuracy (ACC), and UAR, each computed using their corresponding optimal Bayes threshold. For each metric, we provide values of sensitivity (Sen), specificity (Spe), and precision (Pre) based on the respective thresholds used in their calculation. Additionally, normalized cross entropy (NXE) is included in the last column. The first two rows in the table show the performance of raw (uncalibrated) scores, with the downstream model trained with 0 minutes and 40 minutes of in-domain data as well as the AVFAD out-of-domain data. The last row corresponds to the system in the second row with an additional calibration stage trained with 15 minutes of in-domain data.

As seen in Figure 5, the NEC values show a significant decrease from 4.89 to 0.88 as M_T goes from 0 to 40 minutes. The same trends between those two systems can be observed for accuracy (ACC) which improves from 45.47% to 83.80%, and for UAR which improves from 58.27% to 82.61%. The sensitivity (Sen(%)) drastically increases for all three metrics with the introduction of in-domain data, showing the model's enhanced ability to correctly identify pathological instances when in-domain data is used for training. On the other hand, specificity (Spe(%)), the ability of the model to correctly identify healthy subjects as such, exhibits a general tendency to decrease as more in-domain data is added. This is due to the fact that the Bayes threshold does not provide the optimal trade off between sensitivity and specificity when the models are poorly calibrated, resulting in high specificity at the expense of low sensitivity. As the system's calibration is improved, the Bayes threshold results in a better trade off between both types of errors, indicated by the improvement in the corresponding overall metric.

Finally, we can see that normalized cross entropy (NXE) values decrease from 7.28 to 0.63 with the use of in-domain data for training the down-stream model and the calibrator, indicating in a drastic improvement in the quality of the posteriors.

		$\mathrm{Thr}_\mathrm{NEC} = 0.25$				$\mathrm{Thr}_\mathrm{ACC} = 0.5$				$\mathrm{Thr}_\mathrm{UAR} = 0.68$				
M_T	C_T	NEC	Sen(%)	Spe(%)	Pre(%)	ACC(%)	Sen(%)	Spe(%)	Pre(%)	UAR(%)	Sen(%)	Spe(%)	Pre(%)	NXE
0	0	4.89	24.59	93.53	89.03	45.47	22.74	94.00	89.01	58.27	22.23	94.32	89.31	7.28
40	0	0.95	89.66	70.97	86.83	83.60	88.84	72.39	87.30	80.66	88.62	72.39	87.30	1.78
40	15	0.88	92.02	62.46	83.96	83.80	89.06	72.55	87.39	82.61	85.89	79.33	89.87	0.63

Table 1: Metrics for assessing the quality of categorical decisions (NEC, ACC and UAR) and the quality of scores (NXE) for three systems: 1) a system that is not exposed to any in-domain data (first line, with $\mathrm{M}_\mathrm{T} = \mathrm{M}_\mathrm{C} = 0$), 2) a system that uses in-domain data only to train the downstream model (second line, with $\mathrm{M}_\mathrm{T} = 40$ and $\mathrm{M}_\mathrm{C} = 0$), and 3) a system that uses in-domain data for training the downstream model and a calibrator (third line, with $\mathrm{M}_\mathrm{T} = 40$ and $\mathrm{M}_\mathrm{C} = 15$). The threshold for each metric for categorical decisions (NEC, ACC, AUR) is indicated in the first row. For these three metrics, the sensitivity (Sen), specificity (Spe) and precision (Pre) at the corresponding threshold are also listed.

Figure 6 shows the variation of three metrics, one minus accuracy (1-ACC) and two NEC values, as a function of the decision threshold for the system in the last line in Table 1. Metrics are computed using scores that are log odds i.e., the logarithm of the ratio between the posterior for class P and the posterior for class H. The threshold is transformed to the log odds domain to improve visualization. That is, $\mathrm{logoddsthr} = \log(\mathrm{thr}/(1 - \mathrm{thr}))$, where thr is given in Equation (3), and is applied to the log odds of the posteriors rather than to the posterior for pathological speech. NEC_3 is the metric we have been calling NEC in previous figures and tables, obtained with $\mathrm{C}_\mathrm{FN} = 3$, and $\mathrm{C}_\mathrm{FP} = 1$. NEC_1 is a NEC with $\mathrm{C}_\mathrm{FN} = 1$, and $\mathrm{C}_\mathrm{FP} = 1$. As discussed in Section 2.2, the EC with these costs is equal to one minus accuracy. Hence, NEC_1 is simply a normalized version 1-ACC.

The dotted vertical lines represent the optimal Bayes thresholds calculated for each metric (the threshold for 1-ACC and for NEC_1 are the same). The blue line represents the Bayes threshold for NEC_3, and the red line represents the Bayes threshold for NEC_1 and 1-ACC. In this example, NEC_3 is not exactly optimized at the Bayes threshold due to random differences between the data used to train the calibration model and the test data.

The figure shows that selecting the threshold that optimizes NEC_1, which corresponds to $\mathrm{thr} = 0.5$ and $\mathrm{logoddsthr} = 0$, is suboptimal for NEC_3. Conversely, the threshold that optimizes NEC_3, which corresponds to $\mathrm{thr} = 0.25$ and

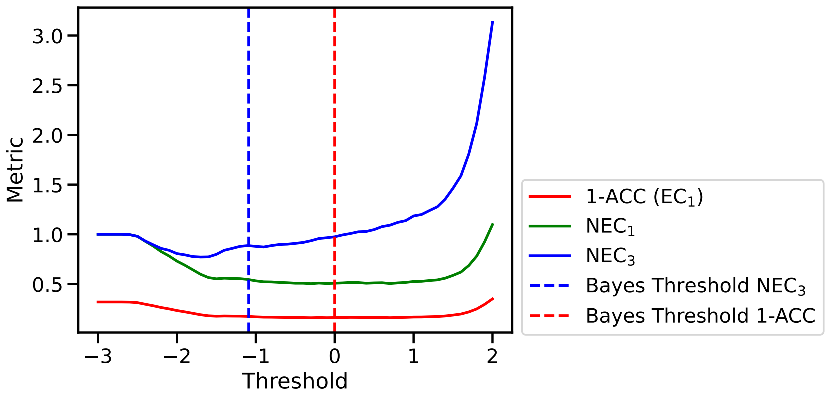

Figure 6: Variation of 1-accuracy (1-ACC) and two different NEC values (NEC_3 obtained with $\mathrm{C}_{\mathrm{FN}} = 3$, and NEC_1 obtained with $\mathrm{C}_{\mathrm{FP}} = 1$) as a function of the decision threshold for the system in the last line in Table 1. NEC_1 corresponds to the normalized version of 1-ACC. The dotted vertical lines represent the optimal Bayes thresholds calculated for each metric (the threshold for 1-ACC and for NEC_1 are the same).

$\mathrm{logoddsthr} = -1.09$, does not generally correspond to the best value of NEC_1. In this example, NEC_1 is quite flat in that region resulting in both thresholds giving very similar performance, but this may not be the case for other systems. In general, it is important to select the threshold that optimizes the metric of interest rather than always defaulting to using a logodds threshold of 0.0 (or posterior threshold of 0.5).

4.4 Error Analysis

As discussed in the previous section, the systems trained with 40 minutes of in-domain data as well as AVFAD data have a reasonable performance, reaching a sensitivity over 92% for a specificity over 62% for NEC. Yet, performance is not perfect. In this section we analyze the score distributions conditioned to the age of the subject and the overall severity of their pathology given by the Grade rating (G), uncovering interesting trends between these conditions and the performance of the system.

The left plot in Figure 7 shows the distribution of the log odds for each of the two classes for the system trained with 40 minutes of SVD data plus AVFAD before and after calibration with 15 minutes of SVD data. Figure 8 shows the reliability plots for the same two systems. The height of each bar corresponds to the ratio of pathological samples among those that had a posterior for class P within that bin. The green bars below the reliability plot show a histogram of the posterior probabilities for class P. We can see that, for the raw scores, all values accumulate in the left-most and right-most bins, indicating that the system is always very certain about the class of the sample, even when it is wrong. We can also see that among all samples for which the system outputs a posterior over 0.9, the center of the blue bar is below the diagonal showing that the system is overconfident and the predicted probabilities are, on average, larger than they should be. After calibration, we can see that the center of this bar is closer to the diagonal, showing an improvement in calibration, and that the scores are less concentrated around extreme values.

Going back to Figure 7, the middle plot shows the score distribution of the pathological samples dividing them by the G rating. We can see that the samples for which the system gives low scores correspond mostly to the lowest severity levels 0 and 1, indicating that perhaps those samples were labelled as pathological not based on the produced speech but rather based on some other diagnostic test. Hence, a system based only on the speech produced by the patient would not be able to detect a voice disorder. Most samples with a G rating of 3 have scores larger than the threshold, indicating that the system is quite successful in identifying these patients as having a voice pathology.

Finally, the right plot in Figure 7 shows the score distributions by age range. We can see that, for the younger age range (red curves), the distributions of the two classes are very well separated, indicating that these samples are easily classified by the system. On the other hand, for the older age ranges, a large fraction of the healthy subjects have large scores, indicating the the system believes those samples to be pathological. Focusing on the green dashed curve we can see that the system incorrectly detects a voice disorder in large fraction of the older patients. In future work we plan to explore approaches for solving this problem within the downstream model by conditioning the model with the age of the subject.

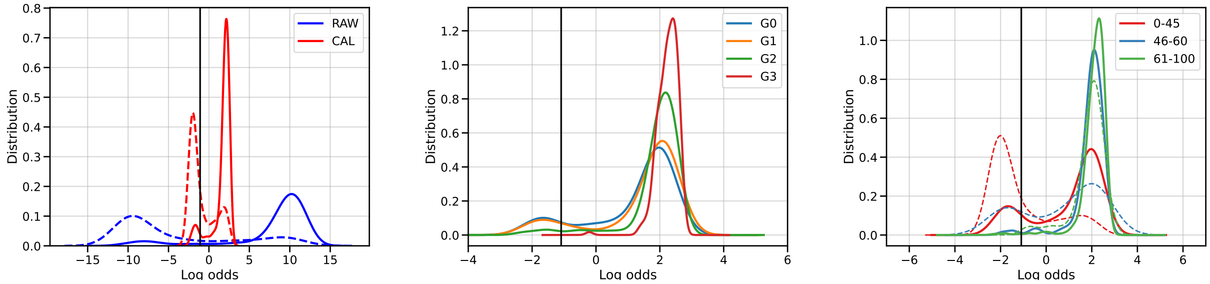

Figure 7: Score distributions by class (left), by severity level for the pathological speech (middle), and by class and age range (right). RAW stands for uncalibrated and CAL for calibrated scores. Solid lines correspond to class P (pathological), and dotted lines to class H (healthy). The vertical lines indicate the decision threshold for the selected NEC metric.

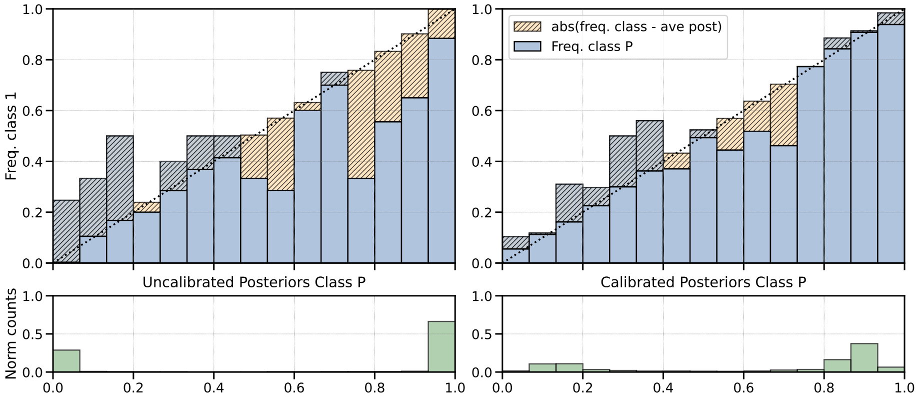

Figure 8: Reliability plots for the raw and calibrated scores for $\mathrm{M_T} = 40$ and $\mathrm{M_C} = 15$.

5 Conclusions

In this work, our focus has been on studying the different components of an AVDD system to better suit real-world scenarios. We built upon our prior research to investigate effective strategies for harnessing both in-domain and out-of-domain data to train a system, specifically for use in medical applications. Notably, our system has shown an impressive ability to distinguish between healthy and pathological speech, using a DNN-based classifier approach in conjunction with a pragmatic method for creating the training dataset. The approach was designed with resource limitations in mind, achieved by supplementing out-of-domain data from the out-domain dataset with varying amounts of in-domain data. The results are striking, with the system achieving an AUC value of 0.9 with just 30 to 40 minutes of in-domain data.

In practical applications it is essential that the evaluation metric reflects the needs of the use-case scenario. The use of cost-based metrics, as proposed in this paper and in some recent works dealing with other medical applications, allows the designer to explicitly specify the consequences that different errors (false alarms vs. misses) carry in terms of medical decisions. In this work, we propose the use of the normalized expected cost (NEC), with costs set to prioritize the detection of pathological cases, which we believe to be appropriate for the AVDD task where missing a diagnosis could be very costly to the patient and to the medical system, which may later have to treat a more severe pathology.

The NEC metric has the additional advantage of enabling the use of Bayes theory for optimal decision-making as long as the scores are good posterior probabilities. In this work, the quality of the posteriors provided by the system is measured with strictly proper scoring rules (SPSRs). In particular, we use the cross-entropy metric, a specific case of SPSR, which is widely used as DNN training objective but much more rarely used for evaluation. A low cross-entropy indicates that the scores are good posterior probabilities and, hence, can be used to make Bayes decisions. Further, and importantly, well-calibrated posteriors provide interpretable information to the end-user, something essential in medical applications.

Our results show that increasing the amount of in-domain data has a positive impact on NEC, AUC, UAR, and accuracy. However, when assessing the quality of the posteriors using cross-entropy, we see that, even with the largest amount of in-domain data, the quality of the posteriors is quite poor. That is, even though the system's posteriors can be used to discriminate both classes well, they are overconfident, having no value for interpretation and being suboptimal for Bayes decision-making. This problem can be solved by adding a post-hoc calibration stage to the system. We show that with only a few minutes of in-domain data for training a calibration stage, the system can provide dramatically better posteriors than the original DNN classifier.

We hope that the proposed approach can be applied generally to new databases. In other words, if a dataset from a new domain is taken and the model is adjusted using the techniques we propose, we expect to achieve similar improvements to those shown here. However, if it is expected that the model trained in a specific domain, such as AVFAD+SVD, will generalize directly to another domain, we do not anticipate that it will work effectively. In this case, it is likely that an adaptation stage to the new domain will be required before applying the proposed method.

In future work, we plan to extend our research to real-world scenarios, specifically those involving audio recordings in the noisy environments of clinics or hospital consultations. Additionally, we intend to explore how calibration models are influenced by channel variability introduced when audio is sourced from mobile devices designed for remote health services. Moreover, we plan to explore approaches for regularizing the DNN model to reduce overfitting and, as a consequence, improve the calibration performance of the system, in an effort to avoid the need for a post-hoc calibration stage.

Furthermore, we propose to introduce additional inputs in the model representing the different factors that we found to affect the performance of the system: the age and sex of the speaker, and severity of their voice disorder. The system would then be able to adapt to each population of speakers, hopefully improving its performance on those groups that currently suffer from poor performance.

Acknowledgements

This work was supported by MCIN/AEI/10.13039/501100011033 under Grant PID2021-126061OB-C44, and in part by the Government of Aragón (Grant Group T36 23R). Also, it was supported in part by the National Agency of Scientific and Technological from Argentina under Grant PICT-2019-01936 and by the National Scientific and Technical Research Council from Argentina under Grant PIP2021-2023-11220200101038. This project has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 101007666.

References

[1] Nelson Roy, Ray M. Merrill, Susan Thibeault, Rahul A. Parsa, Steven D. Gray, and Elaine M. Smith. Prevalence of Voice Disorders in Teachers and the General Population. *Journal of Speech Language and Hearing Research*, 47:281–293, 2004.

[2] Nelson Roy, Ray M. Merrill, and Steven D. Gray. Voice disorders in the general population: prevalence, risk factors, and occupational impact. *The Laryngoscope*, 115:1988–1995, 2005.

[3] Neil Bhattacharyya. The prevalence of voice problems among adults in the United States. *The Laryngoscope*, 124:2359–2362, 2014.

[4] Zulfiqar Ali, M. Alsulaiman, Ghulam Muhammad I. Elamvazuthi, A. Al-Nasheri, TA. Mesallam, M. Farahat, and KH. Malki. Intra- and Inter-Database Study for Arabic, English, and German Databases: Do Conventional Speech Features Detect Voice Pathology? *Journal of Voice*, 31:386.e1–386.e8, 2017.

[5] Laura Verde, Giuseppe De Pietro, Mubarak Alrashoud, Ahmed Ghoneim, Khaled N. Al-Mutib, and Giovanna Sannino. Leveraging artificial intelligence to improve voice disorder identification through the use of a reliable mobile app. *IEEE Access*, 7:124048–124054, 2019.

[6] David Martínez González, Eduardo Lleida, Alfonso Ortega, Antonio Miguel, and Jesús Antonio Villalba López. Voice pathology detection on the saarbrücken voice database with calibration and fusion of scores using multifocal toolkit. In *Advances in Speech and Language Technologies for Iberian Languages - IberSPEECH 2012 Conference, Madrid, Spain, November 21-23, 2012. Proceedings*, volume 328 of *Communications in Computer and Information Science*, pages 99–109. Springer, 2012.

[7] Huiyi Wu, John Soraghan, Anja Lowit, and Gaetano Di Caterina. A deep learning method for pathological voice detection using convolutional deep belief network. In *Proc. Interspeech 2018*, pages 446–450. ISCA, 2018.

[8] Yi-Te Hsu, Zining Zhu, Chi-Te Wang, Shih-Hau Fang, Frank Rudzicz, and Yu Tsao. Robustness against the channel effect in pathological voice detection. *ArXiv*, abs/1811.10376, 2018.

[9] Pavol Harár, Jesús Alonso, Jiri Mekyska, Zoltán Galáž, Radim Burget, and Zdenek Smekal. Voice pathology detection using deep learning: a preliminary study. pages 1–4, 07 2017.

[10] Musaed Alhussein and Ghulam Muhammad. Voice pathology detection using deep learning on mobile healthcare framework. *IEEE Access*, 6:41034–41041, 2018.

[11] Mazin Abed Mohammed, Karrar Hameed Abdulkareem, Salama A. Mostafa, Mohd Khanapi Abd Ghani, Mashael S. Maashi, Begonya Garcia-Zapirain, Ibon Oleagordia, Hosam Alhakami, and Fahad Taha AL-Dhief. Voice pathology detection and classification using convolutional neural network model. *Applied Sciences*, 10(11), 2020.

[12] Laura Verde, Nadia Brancati, Giuseppe De Pietro, Maria Frucci, and Giovanna Sannino. A deep learning approach for voice disorder detection for smart connected living environments. *ACM Trans. Internet Technol.*, 22(1), oct 2021.

[13] Everthon Silva Fonseca, Rodrigo Capobianco Guido, Sylvio Barbon Junior, Henrique Dezani, Rodrigo Rosseto Gati, and Denis César Mosconi Pereira. Acoustic investigation of speech pathologies based on the discriminative paraconsistent machine (dpm). *Biomedical Signal Processing and Control*, 55:101615, 2020.

[14] Rodrigo Colnago Contreras, Monique Simplicio Viana, Everthon Silva Fonseca, Francisco Lledo Dos Santos, Rodrigo Bruno Zanin, and Rodrigo Capobianco Guido. An experimental analysis on multicepstral projection representation strategies for dysphonia detection. *Sensors*, 23(11):5196, 2023.

[15] Sudarsana Reddy Kadiri and Paavo Alku. Analysis and Detection of Pathological Voice using Glottal Source Features. *Journal of Selected Topics in Signal Processing*, 14(2):367–379, 2020.

[16] Siddhant Gupta, Ankur T Patil, Mirali Purohit, Mihir Parmar, Maitreya Patel, Hemant A Patil, and Rodrigo Capobianco Guido. Residual neural network precisely quantifies dysarthria severity-level based on short-duration speech segments. *Neural Networks*, 139:105–117, 2021.

[17] Dayana Ribas, Miguel A. Pastor, Antonio Miguel, David Martínez, Alfonso Ortega, and Eduardo Lleida. Automatic voice disorder detection using self-supervised representations. *IEEE Access*, 11:14915–14927, 2023.

[18] Smitha Rai Sarika Hegde, Surendra Shetty and Thejaswi Dodderi. A Survey on Machine Learning Approaches for Automatic Detection of Voice Disorders. *Journal of Voice*, 33(6):947.e11–947.e33, 2019.

[19] Manfred Pützer and Jacques Koreman. A German database of pathological vocal fold vibration. pages 143–153, 1997.

[20] Luis M.T. Jesus, Inês Belo, Jessica Machado, and Andreia Hall. The Advanced Voice Function Assessment Databases (AVFAD): Tools for Voice Clinicians and Speech Research. In Fernanda Dreux M. Fernandes, editor, *Advances in Speech-language Pathology*, chapter 14. IntechOpen, Rijeka, 2017.

[21] Purva Barche, Krishna Gurugubelli, and Anil Kumar Vuppala. Towards automatic assessment of voice disorders: A clinical approach. In *Proc. Interspeech 2018*, pages 2537–2541, 2020.

[22] Mark Huckvale and Catinca Buciuleac. Automated Detection of Voice Disorder in the Saarbrücken Voice Database: Effects of Pathology Subset and Audio Materials. In *Interspeech*, pages 1399–1403, 2021.

[23] Andrew J. Vickers, Ben Van Calster, and Ewout W. Steyerberg. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *BMJ*, 352, 2016.

[24] Patrick Godau, Piotr Kalinowski, Evangelia Christodoulou, Annika Reinke, Minu Tizabi, Luciana Ferrer, Paul Jäger, and Lena Maier-Hein. Deployment of image analysis algorithms under prevalence shifts. *arxiv:2303.12540*, 2023.

[25] Lena Maier-Hein, Annika Reinke, Patrick Godau, and et.al. Metrics reloaded: Pitfalls and recommendations for image analysis validation. *arxiv:2206.01653*, 2022.

[26] Luciana Ferrer. Analysis and comparison of classification metrics. *arXiv:2209.05355*, 2022.

[27] Andrea Campagner, Federico Sternini, and Federico Cabitza. Decisions are not all equal—introducing a utility metric based on case-wise raters' perceptions. *Computer Methods and Programs in Biomedicine*, 221:106930, 2022.

[28] Christopher M Bishop. *Pattern Recognition and Machine Learning*, volume 4 of *Information science and statistics*. Springer, 2006.

[29] Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 2012.

[30] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR, 06–11 Aug 2017.

[31] Ben Van Calster, Daan Nieboer, Yvonne Vergouwe, Bavo De Cock, Michael J Pencina, and Ewout W Steyerberg. A calibration hierarchy for risk models was defined: from utopia to empirical data. *Journal of clinical epidemiology*, 74:167–176, 2016.

[32] Nan Van Geloven, Daniele Giardiello, Edouard F Bonneville, Lucy Teece, Chava L Ramspek, Maarten Van Smeden, Kym IE Snell, Ben Van Calster, Maja Pohar-Perme, Richard D Riley, et al. Validation of prediction models in the presence of competing risks: a guide through modern methods. *bmj*, 377, 2022.

[33] Candelaria Mosquera, Luciana Ferrer, Diego Milone, Daniel Luna, and Enzo Ferrante. Impact of class imbalance on chest x-ray classifiers: towards better evaluation practices for discrimination and calibration performance. *arXiv preprint arXiv:2112.12843*, 2021.

[34] Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[35] Teodora Popordanoska, Raphael Sayer, and Matthew B. Blaschko. A consistent and differentiable Lp canonical calibration error estimator. In *in Proc. of NeurIPS*, New Orleans, 12 2022.

[36] Jeremy Nixon, Michael W Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. Measuring calibration in deep learning. In *CVPR Workshops*, 2019.

[37] Niko Brümmer. Measuring, refining and calibrating speaker and language information extracted from speech. *University of Stellenbosch, Stellenbosch*, 2010.

[38] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. *CoRR*, abs/2106.07447, 2021.

[39] Miguel A Pastor, Dayana Ribas, Alfonso Ortega, Antonio Miguel, and Eduardo Lleida. Cross-corpus training strategy for speech emotion recognition using self-supervised representations. *Applied Sciences*, 13(16):9062, 2023.

[40] Jazmin Vidal, Pablo Riera, and Luciana Ferrer. Mispronunciation detection using self-supervised speech representations. *SLaTe*, 2023.

[41] Hyungchan Song, Sanyuan Chen, Zhuo Chen, Yu Wu, Takuya Yoshioka, Min Tang, Jong Won Shin, and Shujie Liu. Exploring wavlm on speech enhancement. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 451–457. IEEE, 2023.

[42] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[43] John Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

[44] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

[45] Minoru Hirano and Karen R McCormick. Clinical examination of voice by minoru hirano, 1986.